An Egocentric Perspective on Active Vision and Visual Object Learning in Toddlers

Sven Bambach[1], David J. Crandall[1,2], Linda B. Smith[2,3], Chen Yu[2,3]
[1]School of Informatics and Computing, [2]Cognitive Science Program, [3]Department of Psychological and Brain Sciences
Indiana University
Bloomington, IN 47405, USA
{sbambach, djcran, smith4, chenyu}@indiana.edu

***Abstract*—Toddlers quickly learn to recognize thousands of everyday objects despite the seemingly suboptimal training conditions of a visually cluttered world. One reason for this success may be that toddlers do not just passively perceive visual information, but actively explore and manipulate objects around them. The work in this paper is based on the idea that active viewing and exploration creates "clean" egocentric scenes that serve as high-quality training data for the visual system. We tested this idea by collecting first-person video data of free toy play between toddler-parent pairs. We use the raw frames from this data, weakly annotated with toy object labels, to train state-of-the-art machine learning models (Convolutional Neural Networks, or CNNs). Our results show that scenes captured by parents and toddlers have different properties, and that toddler scenes lead to models that learn more robust visual representations of the toy objects.**

I. Introduction

Visual object recognition is a fundamental skill, and even infants as young as 3-4 months are able to extract perceptual cues that allow categorical differentiations of visual stimuli [1], [2]. Two-year-old toddlers are easily able to recognize a variety of everyday objects, allowing them to rapidly learn word-to-object mappings [3] that build the developmental basis for more complex skills such as language learning. But how do toddlers become such efficient learners despite relying on visual input from an inherently cluttered and referentially ambiguous world, where objects are encountered under seemingly sub-optimal conditions, including extreme orientations and partial occlusions?

Many studies on early visual object recognition are based on experimental designs that passively present controlled visual stimuli, aiming to isolate the effects of various features on building visual representations of objects. While these paradigms are powerful, we know that they are very different from young children's everyday learning experiences: active toddlers do not just passively perceive visual information but instead generate manual actions to objects, actively selecting and creating the scenes that form the visual input they learn from [4], [5]. Recent studies show that this active exploration and manipulation of objects might be systematic in nature. Toddlers that manually explore 3-d objects tend to dwell on planar viewpoints, a bias which increases with age (12-36 months) [6]. Moreover, infants that are more interested in manually exploring objects also build more robust expectations about unseen viewpoints of 3-d objects [7].

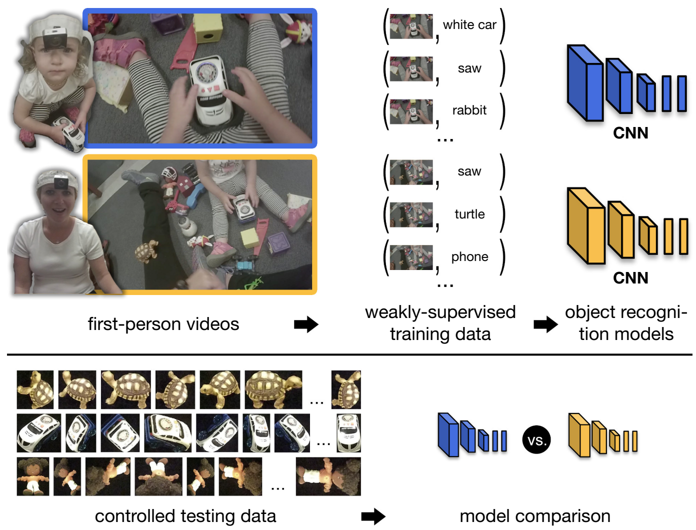


Fig. 1: *Overview of our experiments.* Using head-mounted cameras, we capture video data from toddlers and their parents during joint toy play. We use this data, weakly annotated with toy object labels, to train different object recognition models. We compare performance of models trained with toddlers versus parents using a separate, controlled test set.

A. Rationale for the Present Approach

The success of any learning system depends on the data that it is trained on. The overall hypothesis in the present study is that toddlers naturally create high-quality training data for object recognition by actively exploring and manipulating the world around them. To test this idea, we consider a context that is representative of a toddler's everyday experience: playing with toys. We use head-mounted cameras to collect first-person video data from a naturalistic environment in which parents and children were asked to jointly play with a set of toy objects. Figure 2 shows example frames of this data, contrasting the scenes generated by toddlers with those generated by their parents. We quantify different scene statistics, finding that toddlers generate scenes that contain fewer and larger objects compared to their adult counterparts. To study if and how a learning system can take advantage of these differences and build more robust representations of

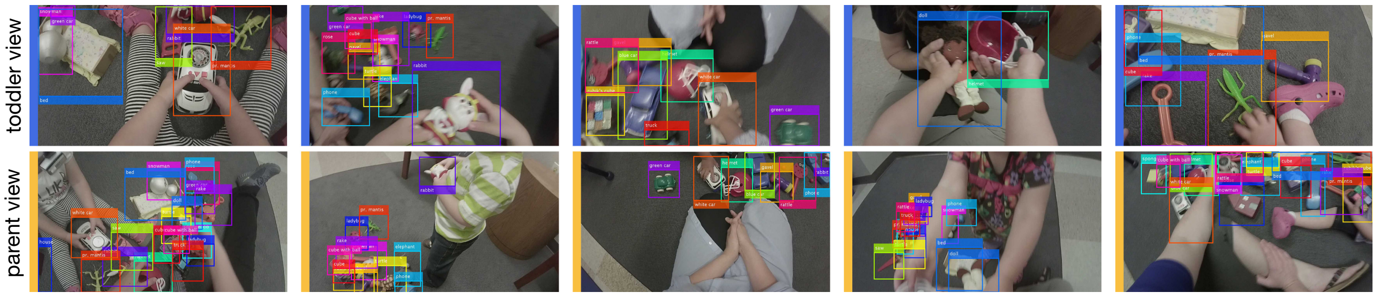


Fig. 2: *Sample first-person video frames* captured during joint toddler-parent toy play, contrasting toddler views (top) and parent views (bottom). Each column shows synchronized frames. Bounding box annotations for each type of toy are also shown.

the objects, we use the first-person data to train and compare different object recognition models. More specifically, we train Convolutional Neural Networks (CNNs), which are the current state-of-the-art for object recognition in the computer vision community [8], and are also increasingly used as "proxy models" by researchers who study human vision [9].

B. Relationship to Previous Work

Studying the effects of toddlers' active viewing behavior during visual object learning using first-person cameras and CNNs was recently introduced in Bambach *et al.* [10], which focused on how toddlers capture visually diverse instances of objects from many different viewpoints. Specifically, CNNs were trained in a fully-supervised manner using cropped-out images of the toys, thus ignoring the context of how and where objects appeared in the first-person view. In contrast, the current study draws inspiration from recent insights into weakly-supervised CNN training for object localization [11]. Rather than cropped-out instances of each toy, we directly feed the raw frames of the entire first-person scenes to the neural network model. This allows us to study differences in toddler and parent data at the scene level, and introduces referential uncertainty between object labels and objects in view. We investigate this uncertainty in terms of quality and quantity of the training data. Specifically, we manipulate quantity by annotating different amounts of the toys in view, and quality by annotating toys based on how large or how centered they appear in view. We compose a series of training simulations, finding that networks that were trained with toddler data sometimes drastically outperform their parent data counterparts, suggesting that toddlers create scenes that facilitate visual object learning.

II. Datasets

We use an existing dataset collected by our past related study [10] to test our hypothesis. This dataset consists of videos from head-mounted cameras that capture the first-person viewpoints of toddlers and parents jointly playing with a set of toys in a natural environment. It also includes a set of controlled close-up photographs of the same set of toys. We use the first-person viewpoints to train our CNN object models, and the controlled viewpoints to test their performance on an independent dataset.

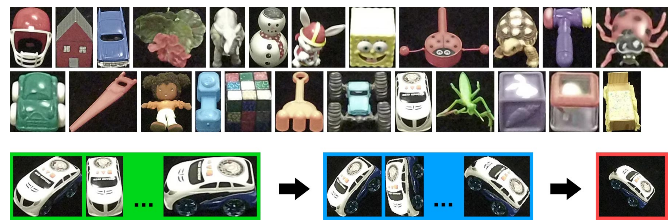


Fig. 3: *Sample images from the controlled test data.* To add viewpoint and scale variation, each toy was photographed from 8 different viewpoints (green), and then rotated 8 times (blue) and cropped at a lower zoom level (red).

We briefly summarize the datasets here; see [10] for details. For the first-person dataset, ten toddler-parent dyads (mean child age 22.6 months) were invited to engage in free-flowing play with a set of 24 toys (Figure 3) in a small lab outfitted as a play room. Each dyad was simply encouraged to play together and with the toys as they pleased. Parent and infant both wore head-mounted cameras to capture an approximation of their respective fields of view (Figure 2). About 8 minutes of video data were collected per dyad. One frame every five seconds was then extracted from each video, producing a set of 1,914 frames (957 each from toddlers and parents). Objects were manually annotated in each image with a bounding box (Figure 2). The controlled dataset (Figure 3) consists of 128 close-up photos of each of the same 24 objects, photographed on a turntable: 64 photos from each combination of the eight 45° orientations about the vertical axis and eight 45° orientations about the axis of the camera center, and a cropped and uncropped version of each image to simulate two levels of zoom.

III. Scene Statistics of the First-person Data

During joint play with a set of toys, toddlers and parents actively create many scenes within their self-selected fields of view. These scenes may contain toys in different quantities, scales, and levels of clutter (see Figure 2). From the perspective of a learning system that aims to build a stable visual representation of each type of toy, different scenes thus create different levels of ambiguity and difficulty. We are interested in whether the active viewing and visual exploration behavior of toddlers actually creates less ambiguous scenes and thus potentially higher quality data for visual object recognition. To substantiate this idea, we begin by studying different properties of toy objects in the fields of view (FOV) of toddlers and parents.

A. Object Size

Scenes might be more informative if the objects of interest dominate. We approximate the actual size of a toy object with the area of its bounding box, and measure the fraction of the field of view that is occupied by this box. Figure 4a contrasts the distributions of perceived object sizes between toddlers and parents. Toddlers create significantly larger object views with a mean size of 5.2% FOV versus 2.8% FOV for parents. For reference, the white car toy (orange bounding box) in the fist column of Figure 2 has a size of 13% FOV in the toddler view and 5% FOV in the parent view.

Even when only large objects are in view, there may be substantial referential ambiguity if all objects are roughly the same size. When toddlers actively select and manipulate toys, those toys should be visually dominant in comparison to the remaining toys in view. To examine this idea, we compute the fraction of the average size of the largest n toys in view over the average size of the remaining toys. As shown in Figure 4b, the relative size difference between large and small toys in view is consistently greater in the toddler data, suggesting that toddler views feature less ambiguity than parent views.

B. Object Centeredness

How centered an object appears within the field of view may also contribute to its visual importance considering the center-bias of eye gaze observed in head-mounted eye-tracking experiments [12]. To measure centeredness, we compute the distance from the center of an object bounding box to the center of the field of view. Figure 4c contrasts the distributions of object-to-center distances between toddlers and parents. We observe no significant difference (mean distance is 48.4% of the maximum possible distance for toddlers, 48.5% for parents), suggesting this is not actually a major differentiator between the views.

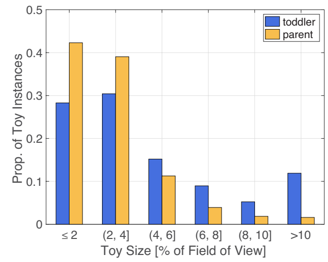
C. Number of Objects

Finally, the ambiguity of a scene also depends on how many objects appear in view at the same time. Figure 4d studies this, showing the number of objects that appear simultaneously in each frame. The results suggest that toddlers create scenes that contain significantly fewer toys in view compared to their parents (10.1 versus 11.8 on average). Moreover, the fraction of frames with a small number (fewer than 4) of objects is about 20% for toddlers but only 13% for parents. Conversely, parents are more likely to have almost all objects in view at once (24% with more than 17 objects for parents versus only 15% for infants).

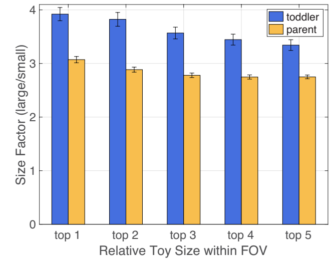
IV. Object Recognition with Deep Networks

A. Fully-supervised Object Recognition with CNNs

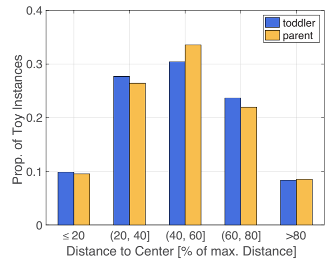
In the computer vision literature, object recognition algorithms are usually trained and evaluated on datasets that



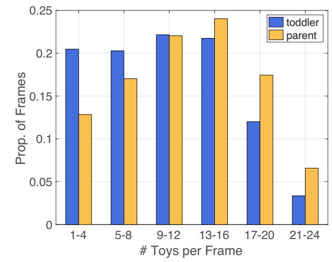
(a) Size of toy objects



(b) Relative toy size within FOV



(c) Toy distance to FOV center



(d) Number of toys per frame

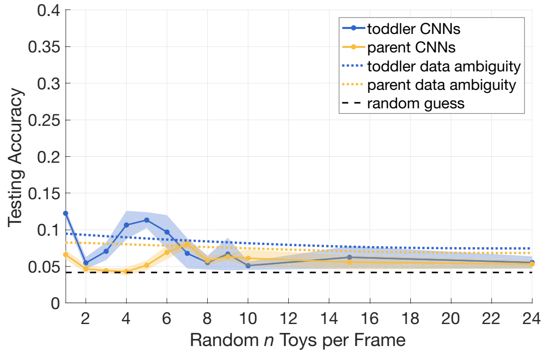
Fig. 4: *Toy object statistics of the first-person scenes.* A comparison of how toy objects appear in the fields of view of toddlers and parents, in terms of (a-b) object size, (c) object location in view, (d) number of objects in view.

contain a set of n predefined visual object classes [13]. As a result, most techniques use discriminative models that are trained to classify an image of an object into one of these (mutually exclusive) n classes, and each training image is assumed to contain an instance of exactly one class, and nothing else. State-of-the-art object recognition models like Convolutional Neural Networks (CNNs) explicitly encode this assumption into the loss function that is minimized during training. For example, the most common loss function for classification tasks is categorical cross-entropy, which encourages the network to output a probability distribution across classes that is very confident for exactly one class (low entropy) rather than multiple classes (high entropy).

B. Weakly-supervised Training with First-person Data

In the context of the naturalistic first-person data described in Sections II and III, the assumption that every scene contains exactly one object of one class is almost always violated: real-world scenes contain multiple objects, and the labeled object may not dominate the view. We are interested in studying (1) to what extend a standard CNN classifier (trained with crossentropy loss) can overcome these violations, and (2) differences between models that are trained with data collected by toddlers when compared to models trained on parent data.

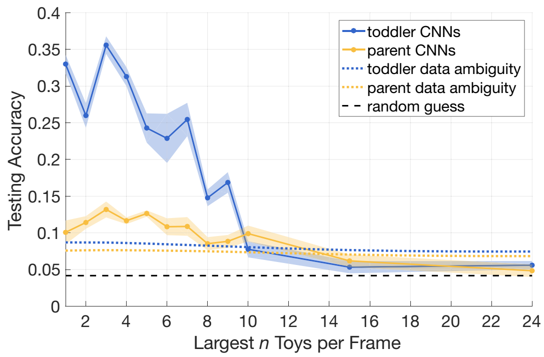
Towards these goals, we run various simulations where we train multiple CNN models under different "weakly-supervised" conditions. In each condition we label a specific subset of the toys that are present in the field of view under the following paradigm: Starting from a frame f that contains k toy objects ($1 \leq k \leq 24$), we generate up to k training exemplars where each exemplar consists of a pair



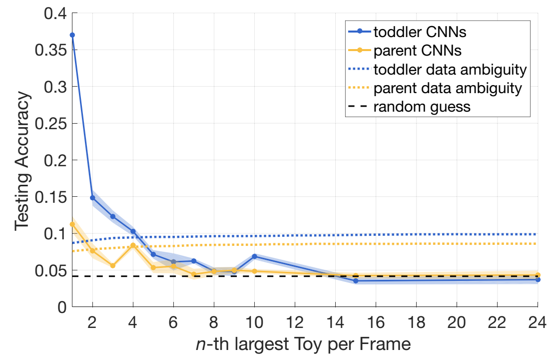
(a) Labeling n random objects per frame

top n	1	2	3	4	5	6	7	8	9	10	24
(a),(b),(c) # exemplars	917	1,794	2,632	3,430	4,191	4,905	5,582	6,207	6,774	7,285	9,646
	942	1,862	2,761	3,622	4,456	5,262	6,029	6,741	7,412	8,025	11,313
(b) avg. object size [%]	15.6	12.4	10.4	9.2	8.4	7.8	7.3	6.9	6.5	6.3	5.2
	7.1	5.9	5.2	4.7	4.4	4.1	3.9	3.7	3.5	3.4	2.8
(c) avg. object size [%]	10.2	8.7	7.8	7.3	6.9	6.6	6.3	6.1	6.0	5.8	5.2
	3.8	3.6	3.5	3.4	3.3	3.2	3.1	3.1	3.0	3.0	2.8
(b) avg. center distance [%]	38	41	43	44	45	46	47	47	47	48	48
	44	45	46	46	46	46	47	47	47	47	48
(c) avg. center distance [%]	20	25	29	32	34	37	38	40	42	43	48
	21	25	28	31	33	35	37	39	40	41	48

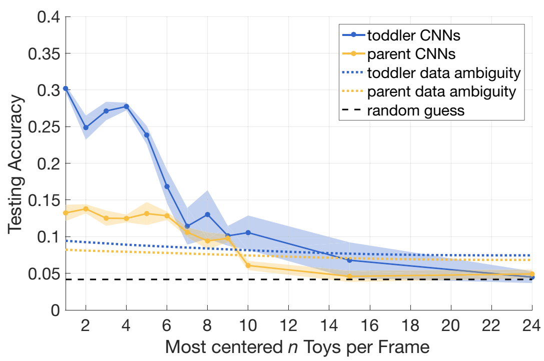
(d) Training data statistics



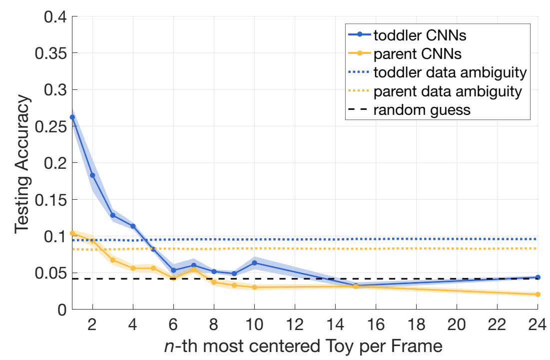
(b) Labeling the n largest toys per frame



(e) Labeling the n-th largest toy per frame



(c) Labeling the n most centered toys per frame



(f) Labeling the n-th most centered toy per frame

Fig. 5: *Object recognition accuracies for different training simulations.* (a-c, e-f) Solid lines depict the overall testing accuracy of CNN models based on the controlled test set of 24 toy objects. Every data point shows the average of five independently trained networks and the shaded areas depict the standard error. Dashed lines depict baselines. (d) Summary of the total number of training exemplars, the average size and average center distance of labeled objects across different training simulations.

of the same (repeated) frame and the toy object label l, i.e. $(f, l_1), ..., (f, l_k)$. Only generating training exemplars based on a subset of the toys in each frame lets us manipulate the overall amount of training data, while choosing which of the toy objects to label potentially affects the quality of the training data.

Since this paradigm creates simple image-label pairs, it allows us to train a discriminative CNN under the same conditions as described in Section IV-A. This is a difficult learning problem for two main reasons: (1) each training image shows the whole first-person view and is potentially referentially ambiguous with respect to the object label, and (2) part of the training data may even be contradictory since the model (falsely) assumes that each frame contains only one object.

Across all simulations, we train models using either the first-person data collected by toddlers, or the first-person data collected from parents, and compare their object classification accuracy on the controlled dataset of Section II.

C. Implementation Details

We use the well-established VGG16 [14] CNN architecture for all of our experiments. VGG16 has a fixed input layer of $224\times224\times3$ neurons, which means we resize all frames to 224×224 pixels. This input layer is followed by 14 convolutional layers, 2 fully-connected layers, and the output layer. The convolutional layers are divided into 5 blocks and each block is followed by a spatial max-pooling operation. All neurons have ReLU activation functions. A complete description of the architecture can be found in [14]. We adjust the output layer of the network to have 24 neurons to accommodate our 24-way object classification task. Following common protocol, we initialize the convolutional layers with weights pre-trained on the ImageNet dataset [13]. Each network is trained via backpropagation using batch-wise stochastic gradient descent and a categorical crossentropy loss function. The learning rate is 0.001, the momentum is 0.9, and the batch size is 64 images. We stop training each network after 20 epochs, after which the loss had converged consistently across different simulations.

V. Learning based on Frame-Specific Metrics

One basic question is whether CNNs can successfully learn object models from the first-person scenes at all. Since not all 24 toys occur simultaneously in every single frame, learning (in the sense of finding a mapping between toy objects and correct labels) should be possible in principle. Moreover, we expect the toddler data to be less ambiguous in that regard since the toddler scenes contain fewer toys on average. Recall that we create training data by generating up to k exemplars $((f,l_1),...,(f,l_k))$ from a single frame f that contains k toys. Thus we can compute the probability that an exemplar is labeled as toy t given that it contains t, $P(l=t|t\in f)$ by simply computing the fraction of training images that are labeled as t over the training images that contain t. One can think of the average probability across all object classes as a measure that captures the referential ambiguity between labels and objects (assuming each object in a scene is equally likely to be labeled). This probability would be 1 for perfectly clean training data, and $\frac{1}{24}$ if the data is completely ambiguous. We report this measure in our results as an additional baseline.

A. Learning from random Toys in View

In our first simulation, we generate training data by simply labeling a random subset of the toys in each training frame. Figure 5a shows the testing accuracies (on the controlled dataset described in Section II) of different CNNs as a function of the number of annotated toys per frame. The blue solid line depicts accuracies based on CNNs trained only on the toddler data while the orange line is based on CNNs trained only on the parent data. As CNN training is non-deterministic, each data point shows the mean testing accuracy across five independently trained networks.

The results show that both parent and toddler networks can achieve above chance accuracies. Also in both cases, the accuracy tends to decrease as n is increased, i.e. as more toys per frame are labeled. This suggests that training with

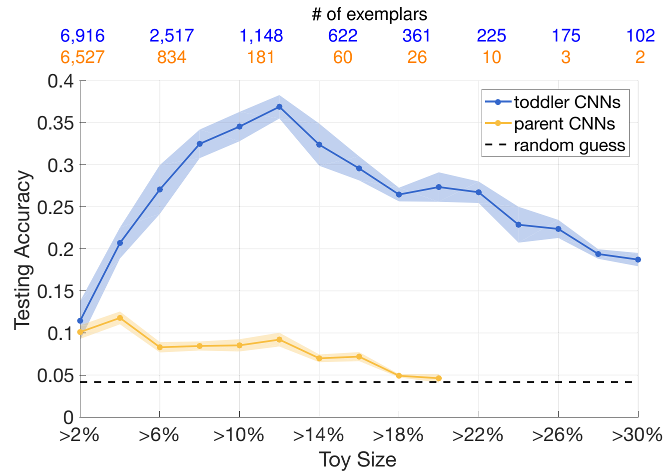


Fig. 6: *Object recognition accuracies* when only labeling toy objects with a minimum size. The table shows the total number of training exemplars for each condition.

fewer overall training exemplars facilitates learning compared to training with more (but potentially contradictory) exemplars. Overall, the toddler networks indeed perform better than the parent networks. This difference may be caused by two different factors: (1) toddlers see fewer objects in view (as indicated by the different baselines), and (2) toddlers create larger views of objects. We further investigate the effect of object size in the next simulation.

B. Learning from the largest Toys in View

From a teaching perspective, labeling a random toy in view is perhaps not the most effective strategy. If the size of objects matters we should see better learning overall and better learning for toddler data in particular if we instead label the subset of the n largest toys in each scene. The results of this simulation are summarized in Figure 5b. Indeed, both parent and toddler networks now outperform their baselines, indicating that the models were more likely to associate object labels with larger objects in view.

Overall, the toddler networks now drastically outperform the parent networks (top accuracy of 36% versus 13%), which further supports the idea that larger objects facilitate learning. For reference, when labeling only the largest toy in each frame its average size is 15.6% FOV in the toddler data, but only 7.1% FOV in the parent data.

Since we generate labels based on object size, generating more training data does not only result in more contradictory exemplars, but also lower quality exemplars. Consequently, we observe a more drastic drop-off in accuracy as n increases.

C. Learning from the most centered Toys in View

A different reasonable teaching strategy is to label the n most centered toys in each scene. Figure 5c summarizes the results of this simulation. Again, both parent and toddler models outperform their baselines, indicating that they successfully learned that more centered toys in view are more likely to be labeled. There is a positive correlation between object size and centeredness (0.23 in the toddler data; 0.16 for parents), so object size may still have an effect. However, the most centered toy in each frame is on average much smaller than the largest toy (10.2% FOV for toddlers, 3.8% for parents), yet the networks achieve overall comparable accuracies.

Again, toddler networks drastically outperform parent networks. Since there is no significant difference in object centeredness across the datasets, this difference is still likely driven by the overall difference in object size.

D. Learning from Misleading Exemplars

Another insightful training approach is to only label the n-th largest (or most centered) toy object in each frame rather than the top n objects. This approach controls for the total number of training exemplars (as it is independent of n) and avoids contradictorily labeled exemplars. At the same time, if centeredness or size are important, then increasing n creates increasingly misleading exemplars.

Figures 5e and 5f show the simulation results of training with the n-th largest and n-th most centered objects respectively. In both cases, only toddler networks achieve results that are significantly above the baselines. Compared to the previous simulations, overall recognition accuracies decrease much more sharply as n increases, highlighting the effect of the misleading exemplars. This drop-off is most drastic for the toddler networks trained on the largest versus second-largest toys in view. This implies that having a very large "distractor object" in view is particularity detrimental for learning, further highlighting the importance of object size.

VI. Learning based on Absolute Metrics

The results presented in Section V suggest that toddlers create scenes that facilitate visual object learning primarily by bringing a few objects dominantly into the field of view. To measure the effect of object size more directly, we run another set of training simulations. This time, we only label objects of a certain minimum absolute size, regardless of their relative size to other objects in view. This creates another quality versus quantity trade-off since increasing the minimum object size results in fewer training exemplars.

Results are summarized in Figure 6. Object recognition accuracy increases with object size in the toddler data, reaching its peak when training with ~800 frames in which the target object covers at least 12% of the FOV. Interestingly, while there is a quality versus quantity trade-off, the overall accuracy remains relatively high, indicating that CNNs can build relatively robust object models from just a few high-quality exemplars. Parents on the other hand did not generate enough high-quality exemplars to learn robust object representations.

VII. Summary and Conclusion

We used first-person video data captured during free toy play between toddlers and their parents to train different object recognition models (based on Convolutional Neural Networks). Our results show that (1) CNNs could successfully learn representations of the toy objects despite being trained only with raw frames from the first-person view, and (2) models trained with data from the toddlers' perspectives drastically outperformed parent-trained models in many conditions. These results, together with [10], demonstrate that a visual learning system can directly benefit from the active viewing behavior of toddlers. More specifically, toddlers tend to generate visually diverse viewpoints of the objects they interact with [10], and, as highlighted in the present study, toddlers tend to bring objects of interest largely and dominantly into view, thus creating visually and referentially unambiguous scenes.

Acknowledgments

This work was supported by the National Science Foundation (CAREER IIS-1253549, CNS-0521433, BCS-15233982) 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 and Computing through the Emerging Areas of Research Project "Learning: Brains, Machines, and Children." It used the Romeo FutureSystems Deep Learning facility, which is supported in part by Indiana University and NSF (RaPyDLI-1439007). We would like to thank Sam Dong, Steven Elmlinger, Seth Foster, and Charlene Tay for helping with the collection of the first-person toy play dataset.

References

[1] P. C. Quinn and P. D. Eimas, "Perceptual cues that permit categorical differentiation of animal species by infants," *J Exp Child Psychology*, vol. 63, no. 1, pp. 189–211, 1996.

[2] P. C. Quinn, P. D. Eimas, and M. J. Tarr, "Perceptual categorization of cat and dog silhouettes by 3-to 4-month-old infants," *Journal of experimental child psychology*, vol. 79, no. 1, pp. 78–94, 2001.

[3] L. Smith and C. Yu, "Infants rapidly learn word-referent mappings via cross-situational statistics," *Cognition*, vol. 106, no. 3, pp. 1558–1568, 2008.

[4] C. Yu, L. Smith, H. Shen, A. Pereira, and T. Smith, "Active information selection: Visual attention through the hands," *IEEE TAMD*, vol. 1, no. 2, pp. 141–151, 2009.

[5] L. B. Smith, C. Yu, and A. F. Pereira, "Not your mothers view: The dynamics of toddler visual experience," *Developmental science*, vol. 14, no. 1, pp. 9–17, 2011.

[6] A. Pereira, K. James, S. Jones, and L. Smith, "Early biases and developmental changes in self-generated object views," *Journal of vision*, vol. 10, no. 11, p. 22, 2010.

[7] K. C. Soska, K. E. Adolph, and S. P. Johnson, "Systems in development: motor skill acquisition facilitates three-dimensional object completion." *Dev Psychol*, vol. 46, no. 1, p. 129, 2010.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, 2016.

[9] I. Gauthier and M. J. Tarr, "Visual object recognition: Do we (finally) know more now than we did?" *Annual Review of Vision Science*, vol. 2, pp. 377–396, 2016.

[10] S. Bambach, D. J. Crandall, L. B. Smith, and C. Yu, "Active viewing in toddlers facilitates visual object learning: An egocentric vision approach," in *Proc. CogSci*, 2016, pp. 1631–1636.

[11] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, 2016.

[12] S. Bambach, L. B. Smith, D. J. Crandall, and C. Yu, "Objects in the center: How the infants body constrains infant scenes," in *Proc. ICDL-EPIROB*, 2016.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. CVPR*, 2009.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.