Incremental Object Learning from Contiguous Views

Stefan Stojanov[1], Samarth Mishra[*1], Ngoc Anh Thai[*1], Nikhil Dhanda[1], Ahmad Humayun[1],
Chen Yu[2], Linda B. Smith[2], James M. Rehg[1]
Georgia Institute of Technology[1]
Indiana University Bloomington[2]
{sstojanov, smishra, athai6, nn3, ahmadh, rehg}@gatech.edu
{chenyu, smith4}@indiana.edu

Abstract

In this work, we present CRIB (Continual Recognition Inspired by Babies), a synthetic incremental object learning environment that can produce data that models visual imagery produced by object exploration in early infancy. CRIB is coupled with a new 3D object dataset, Toys-200, that contains 200 unique toy-like object instances, and is also compatible with existing 3D datasets. Through extensive empirical evaluation of state-of-the-art incremental learning algorithms, we find the novel empirical result that repetition can significantly ameliorate the effects of catastrophic forgetting. Furthermore, we find that in certain cases repetition allows for performance approaching that of batch learning algorithms. Finally, we propose an unsupervised incremental learning task with intriguing baseline results.

1. Introduction

Children are amazing learning machines.[1] Infants acquire extensive object knowledge through self-directed play with minimal supervision, a fact which is remarkable in contrast to the quantity of labeled data required by current deep learning methods. During play, infants pick up, examine, and put down toys of their own volition. The moments in which a supervisory signal is available, for example when an adult names an object, are extremely rare in comparison to the huge volume of unlabeled perceptual inputs. See Fig. 1 for a schematic of this play process.

Research in child development [8, 49, 22] has identified five key properties of infants' play experiences. First, while infants become experts at object categorization, the bulk of their early visual experience involves *object instances*, in the form of toys and everyday objects. Second, their exposure to object instances is highly *repetitive*, with many objects (e.g. a favorite sippy cup) recurring over and over again [8]. Third, when infants hold and manipulate objects, they generate extended, *contiguous views* that may help in revealing 3D object shape [49, 21, 37]. Fourth, infant learning is fundamentally *incremental*, as objects are held and examined in sequence, and once an object has been put down, its imagery is no longer available for learning. Fifth, infants must provide their own supervision when learning about instances, and leverage a sparse, noisy, and unsynchronized supervisory signal when learning object names.[2]

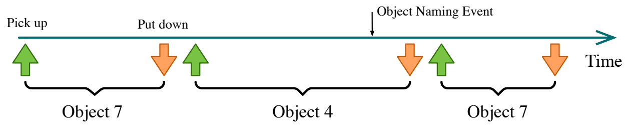


Figure 1: Schematic of incremental object learning based on infant play. Objects occur sequentially as *exposures* consisting of sets of frames with contiguous viewpoints. A sparse and noisy supervisory signal for category learning (naming events), accompanies the wealth of visual data.

These properties of the infant learning environment stand in stark contrast to current methods for object learning in computer vision, which are based on processing mini-batches of randomly-sampled, labelled frames that cover a significant subset of the label space. This approach ensures that gradient updates do not favor one class over another in moving collectively towards higher accuracy. However, when data is processed *incrementally* in standard deep learning architectures, the result is *catastrophic forgetting*, in which object representations developed early in training are forgotten at the expense of more recent examples [14, 17]. Recent works on incremental learning have

[*] Equal contribution.

[1] In the domain of word learning, for example, children acquire an average of 8 to 10 new words per day and reach a vocabulary of 60,000 words by adulthood [34].

[2] While there is a debate in developmental science about the extent to which children's knowledge is innate versus learned, in this paper we focus on the task of learning from visual experience.

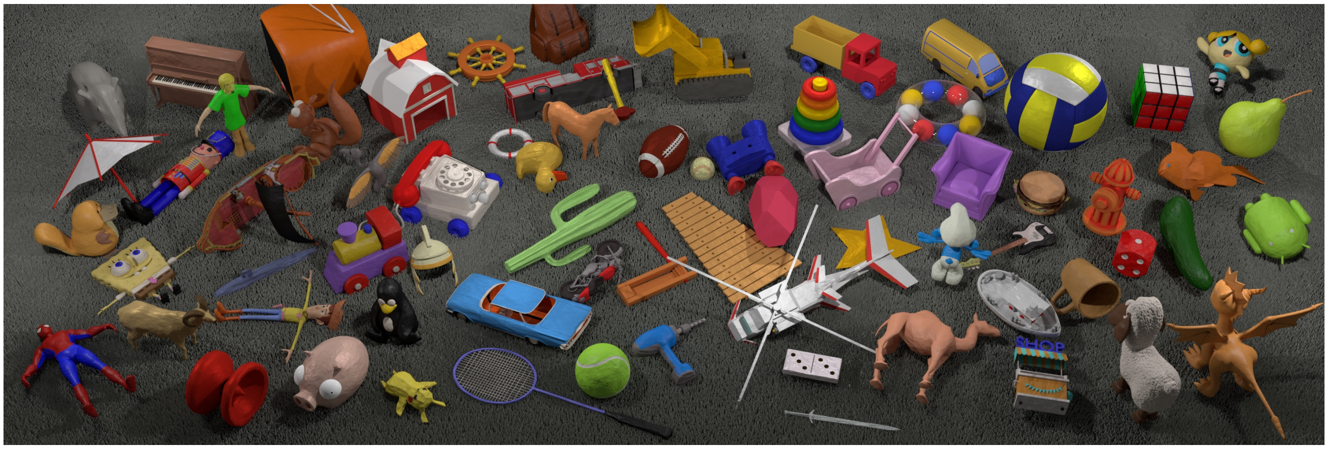


Figure 2: A rendering of approximately one third of the 3D models in Toys-200.

developed methods using distillation loss [29] and exemplars [38, 6] to address the catastrophic forgetting problem, and they represent a valuable point of contact with infant learning. Crucially, however, these prior works have not incorporated repetition, which we will demonstrate to be critical for effective incremental learning (see Sec. 4.3).

This paper introduces a developmentally-motivated environment for object learning known as *CRIB* (Continual Recognition Inspired by Babies), which supports incremental learning of object instances (and categories) from contiguous views with repetition, in both supervised and unsupervised settings. CRIB is an ideal testbed for research in incremental learning, as it provides convenient access to unlimited data with the ability to precisely manipulate key dimensions of the learning task and ensure reproducibility. CRIB comes with a novel dataset, Toys-200, consisting of 3D models of 200 diverse and developmentally-appropriate object instances. Our experiments with CRIB have uncovered some intriguing empirical properties of incremental learning tasks which have not been observed in prior work. Specifically, we show that in incremental learning with repetition it is possible to ameliorate the effects of catastrophic forgetting, with the performance of pre-trained models approaching that of batch-learning. These findings hold for both instance and category learning across a diversity of datasets (Toys-200, ShapeNet [7], and CIFAR[26]).

CRIB is implemented as an API that can easily be incorporated into data loaders for standard deep learning frameworks like PyTorch and TensorFlow, and will be made freely-available to the research community. It supports the paradigm illustrated in Figure 1, in which the learner receives a sequence of *object learning exposures*, each one consisting of a set of frames corresponding to a contiguous sequence of views of a particular object instance. CRIB supports three different incremental learning tasks, and we provide baseline results and extensive experimental results for each in Sec. 4. We hope that CRIB will enable new lines of attack on both incremental learning and developmentally-motivated object learning problems. In summary, this work makes the following contributions:

- The CRIB environment for developmentally-inspired object learning along with the Toys-200 dataset of developmentally-plausible 3D object instances
- A freely-available data generator which integrates into standard deep learning platforms, supports existing 3D datasets, and is capable of generating unlimited data for incremental instance and category learning
- The identification of incremental learning with repetition as a key learning task which makes it possible to ameliorate the effects of catastrophic forgetting
- An extensive evaluation of the effects of distillation loss, explicit exemplar memory and repetitions on both supervised and unsupervised incremental learning tasks[3]

2. Related Work

This paper is most closely related to prior work on *incremental learning* using deep models, and our experiments leverage existing algorithms for learning without forgetting [29], iCARL [38], and E2EIL [6]. In comparison to these works, we provide a novel learning environment (CRIB with Toys-200) and several novel tasks, as well as extensive experiments on multiple datasets that illuminate important aspects of incremental learning approaches, such as the role of repeated exposures, distillation loss, and the impact of exemplar set size, on incremental learning performance. In contrast, prior works [29, 24, 28, 32, 38, 6] did not address instance learning or the use of 3D models to learn from contiguous viewpoints. They addressed only

[3] All resources needed to reproduce the experimental results in this paper and any subsequent releases of software and data will be available at https://iolfcv.github.io/

the single exposure paradigm for category learning using existing image datasets of a fixed size.

Another related body of work is *open world recognition* (of which representative citations are [2, 3, 10]). It is relevant due to its emphasis on self-supervision. Our experiments on weakly-supervised learning from sequential object exposures in Sec. 4.4 are a point of contact with this literature, although our specific paradigm and methods differ from this prior work.

Our development of CRIB is part of an on-going effort to explore the use of *computer graphics rendering and simulation environments* to investigate machine learning topics in controlled settings and address the large scale data requirements of deep learning. Examples include purpose-built autonomous driving simulators such as TORCS [47] and CARLA [13], and efforts to leverage commercial video games [25, 40, 39]. Multiple synthetic optical flow datsets [33, 5, 46] have led to performance improvements, as have generated 3D car assets from [35]. Although the Active Vision Dataset [1] is not synthetic, it is a dense collection of RGB-D images of real scenes that can simulate the visual information perceived by a robot moving through an environment. We are not aware of any prior work on simulation environments which specifically target the learning tasks or synthetic data generation goals addressed by CRIB.

Our work on Toys-200 is related to other efforts in curating *datasets of objects* for recognition tasks. Prior work on collecting real image datasets of 3D objects, such as NORB [27], COIL [36], and more recently, CORe50 [31], are less relevant to this work. More closely-related are works that created synthetic 3D object datasets, such as ShapeNet [7] and Sculptures [45], which have led to significant progress in the domain [41]. In comparison, Toys-200 contains fewer instances (307 for Sculptures and 51K for ShapeNet). However, it occupies a sweet-spot in terms of size and diversity, as the Toys-200 objects are highly diverse in comparison to both Sculptures and ShapeNet and were designed to reflect the types of toys and everyday objects that infants would be likely to encounter. In conjunction with CRIB, we can support a much wider range of data generation approaches than any prior works, as summarized in Table **??** of the Appendix.

This paper is also connected to a long line of research on developmentally-inspired approaches to robotics and learning (e.g. venues such as [9]). Works such as Gepperth et. al. [15] and Kanan et. al. [23] connect to our interest in biologically-inspired incremental learning. Recent work by Haber et. al. [18] shares our interest in play behavior. Other works have developed specific computational models for children's cognitive processes (see [30] for a recent example). None of these works address the specific tasks or settings which characterize our paper.

3. Approach

In order to achieve our goal of exploring the behavior of incremental object recognition algorithms in a developmentally plausible setting, we require a visual learning environment with the following characteristics:

Unlimited Data: The ability to efficiently generate unlimited visual data for each of our objects is critical because it allows us to vary the amount of repetition and generate arbitrarily long experimental runs while ensuring that the learning algorithm continues to receive novel inputs.

Developmental relevance: Our goal is to generate visual data which simulates the object exploration behaviors in early infancy. This requires the use of developmentally-plausible object sets and the ability to generate sequences of contiguous object views.

Integration: To facilitate rapid experimentation, it must be easy to integrate our learning environment with existing data loading mechanisms in modern deep learning frameworks.

We develop **CRIB** (Continual Recognition Inspired by Babies)—a synthetic visual learning environment that fulfills these requirements. CRIB can generate unlimited learning exposures in the form of contiguous views of object instances. Since CRIB is implemented as a Python API it is directly compatible with all popular deep learning frameworks. CRIB is built using the free and cross platform 3D graphics software Blender and uses the Cycles ray tracing engine for rendering. The following section describes how CRIB provides a novel environment for incremental learning experiments.

3.1. CRIB Learning Environment

In this section we describe the process by which we created the Toys-200 dataset and the details of the object rendering approach in CRIB.

3.1.1 3D Object Models for Toys-200

A highly diverse set of toy-like objects is central to generating developmentally plausible object instance data for visual recognition. We collected the Toys-200 dataset of 200 unique toy object models from Blendswap [4], selecting models that were freely-available under a CC license. We began by targeting a core set of 30 specific object categories [49] that are frequently used in research with infants, identifying the best 3D model instance for each one. In order to build a challenging and visually-diverse (See Figure 2) dataset, we supplemented this initial set with additional toy-like objects. The criterion of ”toy-like” was implemented by selecting objects which were similar to the core objects in terms of their level of detail in shape and appearance, and their plausibility for being a child's toy. A specific material shader was developed to give Toys-200 by

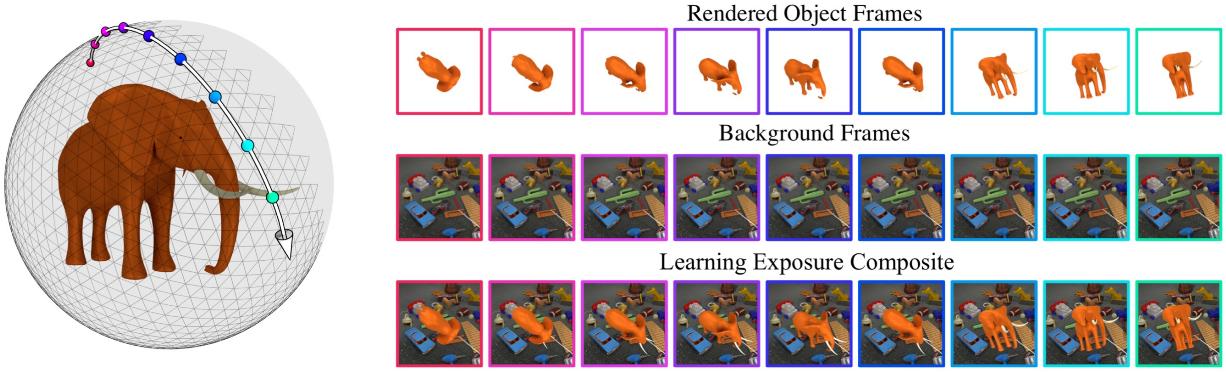


Figure 3: Steps in generating visual exposures using CRIB (top to bottom): 1. Foreground object rendering, 2. background scene selection, 3. foreground and background compositing. A visual exposure consists of multiple clips corresponding to arcs on view sphere.

combining basic Blender material shaders, set up to give the objects surface texture and reflectance properties of plastic, toy-like objects. For our experiments involving category learning, we used the well-known ShapeNet [7] dataset, with appropriate modifications to incorporate it into CRIB. Refer to the Appendix for more details.

3.1.2 Generating Learning Exposures

A learning exposure is a sequence of images obtained by rotating an object of interest relative to a fixed camera, designed to simulate the kind of object views that children are known to generate during object play [21]. An exposure consists of a sequence of short video clips, where each clip is generated by rendering a sequence of images of the object as its pose is linearly interpolated between two fixed poses. See Figure 3 for an illustration of one sequence. The final object pose for one sequence is the starting pose for the next, and the images from each sequence are concatenated into a single contiguous sequence to form the learning exposure.

CRIB generates learning exposures from 3D object models using a set of user-specified API parameters. Below we discuss key aspects of the rendering process, and a detailed technical description of all the steps to generate a learning exposure can be found in the Appendix.

Lighting: In the API, the user specifies a lighting setting of either four point or three rod light sources placed above the object. Further characteristics are defined by parameters for pose, temperature and intensity.

Object rotation: Object rotation in CRIB is generated by linearly interpolating between object poses (azimuth, rotation, elevation, scale) over a number of frames. To specify the qualitative characteristics of object rotation in the learning exposure, the user specifies parameters for the total number of frames in the learning exposure and the total number of different object poses for interpolation.

Preprocessing: Once the API parameters are specified, the following fully-automated process proceeds: The target object is imported into Blender, its center of mass is estimated and it is positioned in the center of the camera frame. The object is then appropriately scaled so that it remains remains inside the camera field of view during the rotating motion around its center of mass and the change in scale.

Foreground rendering: The specified light sources are instantiated and the sequence of frames is rendered without a background. At this step, instance segmentations and bounding boxes are collected for the foreground object.

Background rendering: Backgrounds are image sequences of objects from Toys-200, which are distributed over the floor to create a cluttered background. The camera above the objects moves slightly over time to emulate head motion (such background frames are illustrated in Figure 3). This results in a dynamic, cluttered background environment which makes the recognition task more challenging and simulates real-life play scenarios in which a child interacts with a set of toys (e.g. dumped from a toybox). We ensure that the foreground object is not also present in the background. Once the background sequence has been rendered, the final step is to composite the foreground and background layers in each frame, and add a small amount of pixel-wise noise.

Testing image generation: For evaluation purposes, CRIB can also generate single images of a target object at a random rotation, elevation and scale, with random lighting conditions and backgrounds.

3.2. Learning Tasks in CRIB

CRIB supports three different incremental learning scenarios, two of which are novel. In each case, CRIB provides learning exposures which are combined with stored past images in forming minibatches which are used for training. The details vary with the task and the architecture, and are detailed in the following sections.

Supervised Single Exposure: This is the standard incremental learning task, in which classes or instances are presented sequentially to the learner. The key property is that the learner sees each object exactly once. This leads to catastrophic forgetting in all of the cases that we evaluated.

Supervised Repeated Exposure: In this novel task, classes or instances are presented to the learner sequentially *with repetition*. At random, the learner is given new learning exposures for previously-seen objects. Our experiments demonstrate that allowing a limited amount of repetition (e.g. 10 exposures each for 200 object instances) allows existing algorithms to approach the batch performance.[4]

Unsupervised Repeated Exposure: In this task, learners receive repeated exposures to a sequence of objects, but no labels are provided. This is similar to discriminative incremental clustering [16]. This very challenging task requires the learner to identify learning instances corresponding to novel objects, and re-identify previously-seen objects. It mirrors the challenge infants face during play, as most of their learning exposures will not be accompanied by an object name.

4. Experiments

In this section, we introduce the baseline algorithms in our study, and present novel experimental results for the three incremental learning tasks from Sec. 3.2. Performance is measured using incremental accuracy as in [38]: Following training on each learning exposure, the classification accuracy is computed on unseen test samples from all instances or categories the learner has seen up to that point.

All learning exposures generated with CRIB are 100 frames long and interpolate between three randomly-chosen points on the view sphere, with scale smoothly varying from 0.3 to 1.1. Light source position is jittered at random and light intensity is randomly-sampled from 4000-6000K (indoor lighting temperature range). 100 random testing frames are generated for each object. Refer to the appendix for additional details.

4.1. Incremental Learning Methods

We produced our own implementations of three recent CNN-based incremental learning algorithms [29, 38, 6]. Differences in our implementation from the original are described below and in the appendix. All methods use ResNet-34 [19] as the backbone architecture.

LwF [29] addresses catastrophic forgetting by modifying the loss function used to train a standard CNN incrementally. Each time a new class is introduced, the fully connected layer of the CNN is expanded by adding an output sigmoid unit for the new class. Distillation loss [20] is applied to the outputs for the other classes in attempt to preserve the information they encode and prevent significant changes due to backprop on the current class. Our LwF differs from [29] in using sigmoid units rather than a softmax layer for classification, and in performing additional data augmentation.

iCaRL [38] builds on LwF by including explicit memory in the form of an exemplar set managed by the learning algorithm. The exemplars are used to perform nearest exemplar mean classification in feature space. The inference procedure consists of computing normalized exemplar mean features per class using the CNN, and then classifying by determining the nearest exemplar mean from the normalized features of each testing sample. Training follows LwF when distillation loss is used, otherwise it is standard CNN training. Our iCaRL [38] implementation uses additional data augmentation.

E2EIL [6] builds on the previous two methods. During training, the loss takes into account ground truth labels of the samples from the other classes as well as the current class. Unlike iCaRL, training is end-to-end since the network outputs are used for classification. E2EIL adds balanced fine-tuning which targets the case when the number of samples from the other classes is significantly lower than the number of samples for the current class. Exemplar set construction follows iCaRL, but is done twice: after training and after balanced fine-tuning. Our implementation adopts a temperature-squared weighting [20] for distillation loss, computes distillation loss over all seen classes, and uses a different data augmentation scheme.

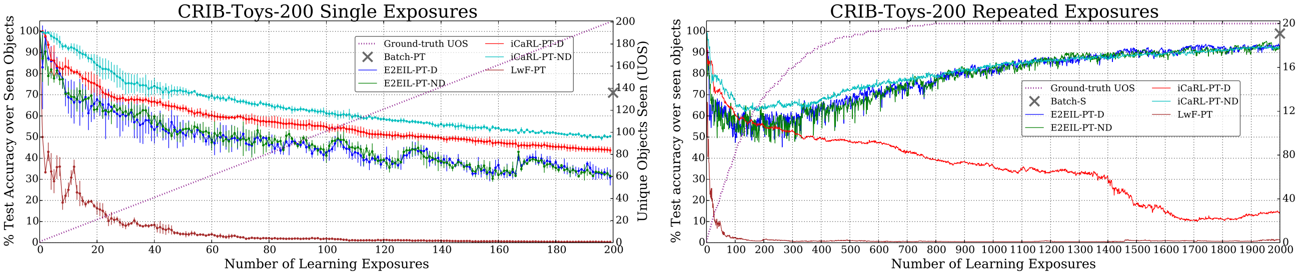
Our experiments include both training from scratch and initializing weights from a pre-trained ILSVRC-2014 [11] architecture. Based on prior transfer learning results [44, 48, 12], we would expect that starting from a pretrained architecture should yield better performance, and our results confirm this. We also train with and without distillation loss, in order to quantify its benefit. Note that when distillation loss is not used, we apply the classification loss to all output nodes and use the exemplar labels.

Our naming convention: in iCaRL-PT-ND, PT indicates starting with a pre-trained backbone architecture, and ND means that distillation loss is not used, whereas iCaRL-S-D refers to training from scratch (i.e. random weight initialization) and using distillation loss.

4.2. Single Exposure Yields Catastrophic Forgetting

In this section we demonstrate that the single exposure task leads to catastrophic forgetting for all of our baseline methods in two datasets: Toys-200 and CIFAR-100. In comparison to prior work [29, 38, 6], our Toys-200 experiments are the first demonstration of catastrophic forgetting in instance learning, and our CIFAR-100 experiments differ in that we present classes one-at-a-time rather than two or

[4] Note that in experiments with 3D rendered data, such as Toys-200, it is still the case that each rendered image is used only once, as each new learning instance will correspond to a new trajectory on the view sphere.



(a)

(b)

Figure 4: (a) Performance of iCaRL, E2EIL and LwF when presented with a single exposure for each object instance from CRIB-Toys-200. (b) shows performance of the same methods with repeated exposure.

more. For the Toys-200 experiment, we use 200 exposures, one for each object instance. E2EIL and iCaRL use an exemplar set size of 600 images, or 3% of the total data (as compared to 4% in [38, 6]). We computed standard error bars by repeating each experiment 3 times.

As evident in Figure 4a, all results for Toys-200 have a general downward trend, which is similar to the results in [38, 29, 6] and is attributed to catastrophic forgetting of the instances which were seen early in the sequence. The results for iCaRL-PT-(D/ND) show that training with distillation is not favorable in this task, while the results for E2EIL-PT-(D/ND) show that distillation loss does not make a difference. We find that test accuracy can be easily improved by using a pre-trained model, aligning with [44, 48, 12]. In addition, we tested on CIFAR-100 [26] with iCaRL-S-ND and iCaRL-PT-ND (see Figure 5) with single classes presented sequentially. Further experiments using random initialization are included in the Appendix.

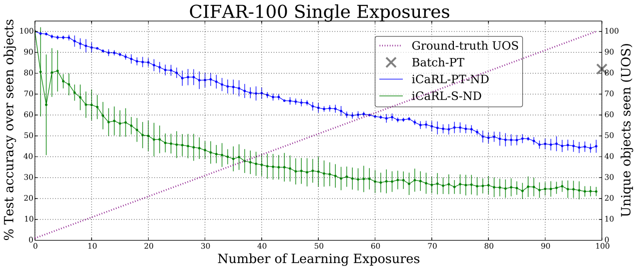


Figure 5: Performance of iCaRL-S-ND and iCaRL-PT-ND on CIFAR-100 confirm catastrophic forgetting. Both algorithms have an exemplar set size of 2000.

4.3. Repetition Reduces Catastrophic Forgetting

In this section, we demonstrate that introducing repetitions during incremental learning ameliorates the effects of catastrophic forgetting, resulting in improvements in accuracy and enabling the majority of tested algorithms to eventually approach the performance of a pre-trained batch learning method. We also examine the effect of the number of exemplars.

Our first experiments with repetition are with Toys-200, using 2000 learning exposures (each object appearing ten times), with an explicit memory of 600 exemplars. For every experimental run, we generate a random sequence of object instances such that all methods experience the same number of objects by each time step, but not the same instances in the same order. Figure 4b shows the results. E2EIL and iCaRL-PT-ND achieve an accuracy close to the batch learning algorithm, whereas iCaRL-PT-D does not show an improvement. Note that the performance gap between iCaRL-PT-D and iCaRL-PT-ND in this task is larger in comparison to the single exposure task. This potentially indicates that distillation loss is hindering the ability to leverage repeated exposures for iCaRL, and highlights the advantage of simply using the exemplar labels. Results for algorithms trained from random initialization in the Appendix further confirm this finding.

We perform further experiments using three datasets: CRIB-Toys, CRIB-ShapeNet [7] and CIFAR [26] to (1) evaluate whether incremental learning with repeated exposures can allow incremental algorithms to get close to the performance of batch algorithms beyond an instance learning task (2) evaluate the importance of the number of exemplars on the accuracy gains from repeated exposure. We perform the following experiments:

1. **CRIB-Toys-50**: 50 objects over 500 learning exposures (each object is shown 10 times).
2. **CRIB-ShapeNet-20**: 20 categories over 500 learning exposures. (25 instances from each category are shown)
3. **CIFAR-20**: 20 categories over 1000 learning exposures (each category is shown 50 times).

Figure 6 contains the results for this experiment. It is evident that the same trend applies to all three datasets: the performance of the algorithms declines at first before increasing as they get more repeated exposures of previously seen objects, and towards the end gets close to the performance of a batch learning algorithm. We believe these are the first findings for learning with repetitions in incremental learning.

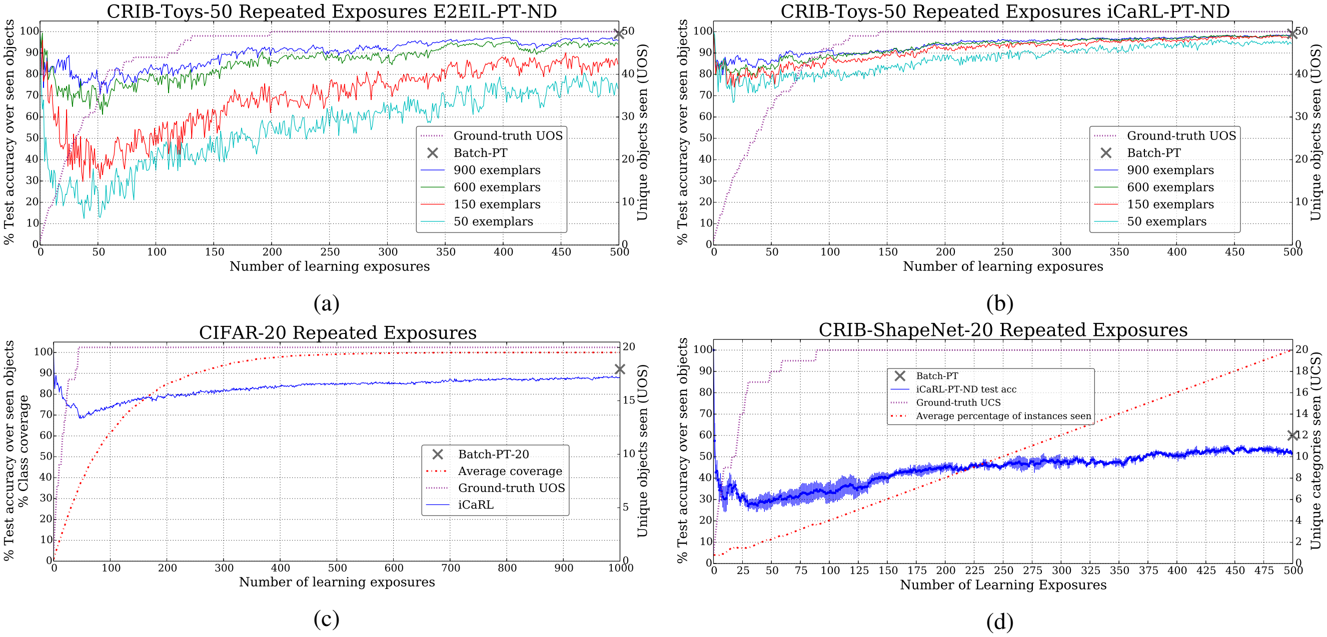


Figure 6: Top : Performance of (a) iCaRL-PT-ND and (b) E2EIL-PT-ND with different number of exemplars on 50 objects of CRIB-Toys. Bottom : (c) Performance of iCaRL-PT-ND (400 exemplars) on 20 categories of CIFAR (d) Performance of iCaRL-PT-ND (1500 exemplars) on 20 categories of CRIB-ShapeNet. (Best viewed with zoom)

For **CRIB-Toys-50**, as evident from Figures 6a and 6b, both iCaRL-PT-ND and E2EIL-PT-ND maintain the upward trend first observed in Figure 4b. Additionally, this experiment demonstrates that for an instance task, regardless of the total number of exemplars (18%, 12%, 3% or 1% of the total data), given sufficient repetition, the accuracy of iCaRL-PT-ND is close to pre-trained batch performance. While E2EIL-PT-ND maintains an increasing trend, the variants trained with small numbers of exemplars are not as close to the pre-trained batch model after 500 exposures.

CRIB-ShapeNet-20 is a categorization task where repeated exposures are different instances from the same category. 25 instances for training and 15 for testing are chosen randomly from 20 categories of the ShapeNet Core55 dataset [7]. Learning exposures generated with CRIB for each instance are provided over 500 exposures and testing is done on 100 frames of random object views, scale and lighting for each instance in the test set for a seen category. The performance (Figure 6d), shows that repeated exposures to new instances in the same category leads to improvements for categorization. This result extends our initial finding of improvement towards batch performance via repeated exposures to a different task and dataset.

For **CIFAR-20**, we sample with replacement 100 images from a total of 500 images per category for each learning exposure. This exposes the algorithm to images repeatedly during different exposures. The performance on iCaRL-PT-ND decreases at first and starts to go up after all 20 objects have been seen (Figure 6c). Coverage for each category is the portion of unique images seen within a category, with the mean over all categories shown on the plot. The accuracy improvement rate decreases after 100% of the data is shown to iCaRL-PT-ND. The final incremental accuracy is on par with a batch algorithm, showing that improvements due to repetition of concepts are not unique to CRIB.

4.4. Incremental Learning Without Supervision

In this task a learning algorithm needs to do novelty detection—to determine whether or not an exposure comes from a novel instance, and recognition—to determine which previously-seen instance it belongs to, prior to updating parameters.

Prior work on open set and open world recognition [42, 43] tackles the subproblem of novelty detection by thresholding on known class scores to detect whether a new data point belongs to a class that has not been encountered. Drawing from these works, we use the following algorithm for our straightforward baseline based on iCaRL-PT-ND. At any given exposure, in unit normalized feature space, the algorithm first finds the distance of the images from a learning exposure to all the exemplar means. This is followed by finding the mean distance of the images in the exposure, and using it as a score to determine whether the current object has been previously seen. If the minimum distance-score is more than a given threshold, the exposure labeled as coming from a new object instance, otherwise it gets classified as the previously seen instance with the minimum distance-score. The threshold used was found as the optimal oper-

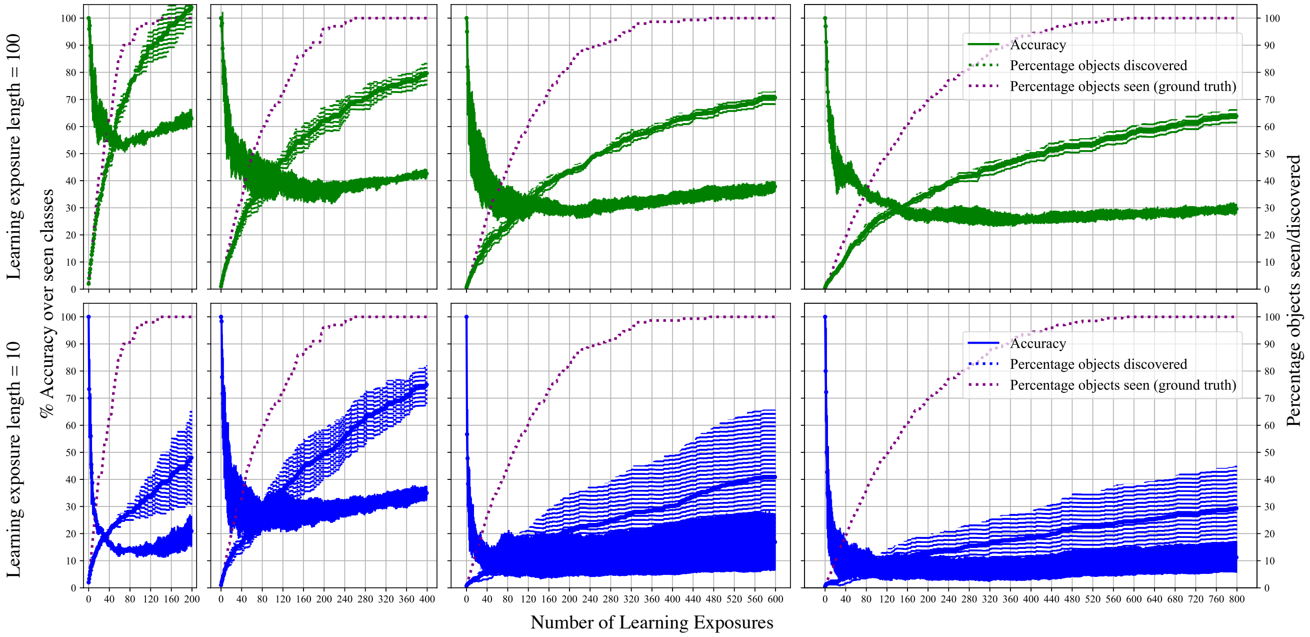


Figure 7: Performance of iCaRL-PT-ND (600 exemplars) on unsupervised repeated exposures of CRIB-Toys-50, 100, 150, 200 (left to right) with learning exposure lengths of 10 and 100. The 10 length learning exposures are the first 10 frames of the 100 length learning exposure. (Best viewed with zoom)

ating point from a precision-recall analysis over the binary classification problem of novelty detection.

We evaluate this baseline algorithm on different learning exposure lengths and number of repeated exposure tasks with CRIB. After each learning exposure, testing is done on random views of ground truth seen objects. Since the labels given by a learning algorithm in this task need not have any correspondence with the ground truth labels, a one-to-one correspondence is first established between these sets of labels based on a maximum accuracy matching, and then the learning algorithm's test accuracy is computed.

Figure 7 contains the results of this study. For 100 learning exposure length, the algorithm's accuracy is constant or decreases with a greater number of objects and four repeated exposures. Further, for the same learning exposure length, the proportion of objects discovered compared to the ground truth number of unique objects seen decreases with greater object set sizes. Across all experiments with different total numbers of objects, there is a consistent trend that a smaller learning exposure length results in a lower final accuracy. Furthermore, a lower learning exposure length results in higher variability in performance over multiple runs with different order of objects encountered.

5. Conclusion

We introduce CRIB, a novel environment for generating unlimited training data for incremental learning of object categories and instances, based on rendering learning exposures from 3D object models. CRIB models the kinds of object views generated by infants during play. We introduce a novel instance learning dataset called Toys-200. We use CRIB to study three incremental learning scenarios and demonstrate that allowing repeated exposures dramatically improves the performance of state-of-the-art methods, allowing them to converge to the batch learning accuracy in many cases. Finally, we show intriguing results on the challenging new task of incremental learning without supervision. Our CRIB enviornment and data is freely-available, and we hope that this work will enable and motivate the development of new incremental learning methodology.

6. Acknowledgement

We would like to thank all the reviewers and Qian Shao for his early contributions to the incremental learning experiments. This work was supported by NSF Awards BCS-1524565 and BCS-1523982. In addition, this work was partially supported by the Indiana University Areas of Emergent Research initiative in Learning: Brains, Machines, Children.

References

[1] P. Ammirato, A. C. Berg, and J. Kosecka. Active vision dataset benchmark. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition Workshops, pages 2046–2049, 2018. 3

[2] A. Bendale and T. Boult. Towards open world recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1893–1902, 2015. 3

[3] A. Bendale and T. E. Boult. Towards open set deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1563–1572, 2016. 3

[4] blendswap.com. https://blendswap.com. 3

[5] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision*, pages 611–625. Springer, 2012. 3

[6] F. M. Castro, M. J. Marin-Jimenez, N. Guil, C. Schmid, and K. Alahari. End-to-end incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 233–248, 2018. 2, 5, 6

[7] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2, 3, 4, 6, 7

[8] E. M. Clerkin, E. Hart, J. M. Rehg, C. Yu, and L. B. Smith. Real-world visual statistics and infants' first-learned object names. *Phil. Trans. R. Soc. B*, 372(1711):20160055, 2017. 1

[9] Conference. ICDL-EPIROB. http://www.icdl-epirob.org/. 3

[10] R. De Rosa, T. Mensink, and B. Caputo. Online open world recognition. *arXiv preprint arXiv:1604.02275*, 2016. 3

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 5

[12] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *International Conference on Machine Learning*, pages 647–655, 2014. 5, 6

[13] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 3

[14] R. M. French. Catastrophic Forgetting in Connectionist Networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. 1

[15] A. Gepperth and C. Karaoguz. A Bio-inspired Incremental Learning Architecture for Applied Perceptual Problems. *Cognitive Computation*, 8(5):924–934, 2016. 3

[16] R. Gomes, M. Welling, and P. Perona. Incremental learning of nonparametric bayesian mixture models. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008. 5

[17] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio. An empiritcal investigation of catastrophic forgetting in gradient-based neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 1

[18] N. Haber, D. Mrowca, S. Wang, L. F. Fei-Fei, and D. L. Yamins. Learning to play with intrinsically-motivated, self-aware agents. In *Advances in Neural Information Processing Systems*, pages 8398–8409, 2018. 3

[19] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[20] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. 5

[21] K. H. James, S. S. Jones, L. B. Smith, and S. N. Swain. Young children's self-generated object views and object recognition. *Journal of Cognition and Development*, 15(3):393–401, 2014. 1, 4

[22] P. J. Kellman and M. E. Arterberry. *The cradle of knowledge: Development of perception in infancy*. MIT press, 2000. 1

[23] R. Kemker and C. Kanan. FearNet: Brain-Inspired Model for Incremental Learning. In *International Conference on Learning Representations (ICLR)*, Apr 2018. 3

[24] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming Catastrophic Forgetting in Neural Networks. *Proceedings of the National Academy of Sciences*, page 201611835, 2017. 2

[25] P. Krähenbühl. Free supervision from video games. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 18)*, pages 2955–2964, 2018. 3

[26] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009. 2, 6

[27] Y. LeCun, F. J. Huang, and L. Bottou. Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages II–104, 2004. 3

[28] S.-W. Lee, J.-H. Kim, J. Jun, J.-W. Ha, and B.-T. Zhang. Overcoming Catastrophic Forgetting by Incremental Moment Matching. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 4652–4662. 2017. 2

[29] Z. Li and D. Hoiem. Learning without Forgetting. In *European Conference on Computer Vision (ECCV)*, pages 614–629, 2016. 2, 5, 6

[30] S. Liu, T. D. Ullman, J. B. Tenenbaum, and E. S. Spelke. Ten-month-old infants infer the value of goals from the costs of actions. *Science*, 358(6366):1038–1041, 2017. 3

[31] V. Lomonaco and D. Maltoni. CORe50: a New Dataset and Benchmark for Continuous Object Recognition. In *Proceedings of the 1st Annual Conference on Robot Learning*, 2017. 3

[32] D. Lopez-Paz and M. A. Ranzato. Gradient Episodic Memory for Continual Learning. In *Advances in Neural Information Processing Systems 30 (NIPS)*, pages 6467–6476. 2017. 2

[33] N. Mayer, E. Ilg, P. Hausser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 4040–4048, 2016. 3

[34] B. McMurray. Defusing the Childhood Vocabulary Explosion. *Science*, 317(5838):631–631, 2007. 1

[35] Y. Movshovitz-Attias, T. Kanade, and Y. Sheikh. How Useful is Photo-realistic Rendering for Visual Learning? In *European Conference on Computer Vision (ECCV)*, pages 202–217, 2016. 3

[36] S. Nayar, S. Nene, and H. Murase. Columbia Object Image Library (COIL 100). *Department of Comp. Science, Columbia University, Tech. Rep. CUCS-006-96*, 1996. 3

[37] A. F. Pereira, K. H. James, S. S. Jones, and L. B. Smith. Early biases and developmental changes in self-generated object views. *Journal of vision*, 10(11):22–22, 2010. 1

[38] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. iCaRL: Incremental Classifier and Representation Learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5533–5542, July 2017. 2, 5, 6

[39] S. R. Richter, Z. Hayder, and V. Koltun. Playing for benchmarks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2213–2222, 2017. 3

[40] S. R. Richter, V. Vineet, S. Roth, and V. Koltun. Playing for data: Ground truth from computer games. In *European Conference on Computer Vision*, pages 102–118, 2016. 3

[41] M. Savva, F. Yu, H. Su, A. Kanezaki, T. Furuya, R. Ohbuchi, Z. Zhou, R. Yu, S. Bai, X. Bai, et al. Shrec17 track large-scale 3d shape retrieval from shapenet core55. 3

[42] W. J. Scheirer, A. de Rezende Rocha, A. Sapkota, and T. E. Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2013. 7

[43] W. J. Scheirer, L. P. Jain, and T. E. Boult. Probability models for open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2317–2324, 2014. 7

[44] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014. 5, 6

[45] O. Wiles and A. Zisserman. SilNet : Single- and Multi-View Reconstruction by Learning from Silhouettes. In *BMVC*. BMVA Press, 2017. 3

[46] J. Wulff, D. J. Butler, G. B. Stanley, and M. J. Black. Lessons and insights from creating a synthetic optical flow benchmark. In *European Conference on Computer Vision*, pages 168–177. Springer, 2012. 3

[47] B. Wymann. TORCS - The Open Racing Car Simulator. 3

[48] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014. 5, 6

[49] D. Yurovsky, L. B. Smith, and C. Yu. Statistical Word Learning at Scale: The Baby's View is Better. *Developmental Science*, 16(6):959–966, 2013. 1, 3