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 catastrophic forgetting for incremental learning algorithms. Furthermore, we find that in certain cases repetition allows for performance approaching 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, and 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 [7, 49, 21] has identified five key properties of infant's 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 [7]. Third, when infants hold and manipulate objects, they generate extended, *contiguous views* that may help in revealing 3D object shape [49, 20, 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]

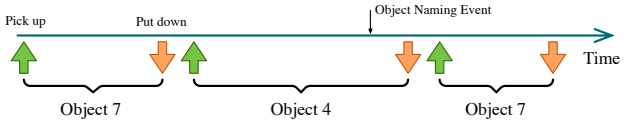


Figure 1: Schematic of incremental object learning based on infant play. Objects occur sequentially as *exposures* consisting of sets of frames with contiguous viewpoints. A sparse and noisy supervisory signal for category learning (naming events), accompanies the wealth of visual data.

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 exam-

[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, we will focus on the task of learning from visual experience in this paper.

	COIL [36]	NORB [27]	CORe50 [31]	ShapeNet [6]	Sculptures [45]	CRIB
Category/Instance	I	C	C	C	I	I
Synthetic	✗	✗	✗	✓	✓	✓
Pose Labelling	✓	✓	✗	✓	✗	✓
Unlimited View Granularity	✗	✗	N/A	✗	✗	✓
Data Generation API	✗	✗	✗	✗	✗	✓
Temporally Continuous	✗	✗	✓	✗	✗	✓
Bounding box annotation	✗	✗	✓[3]	✗	✗	✓

Table 1: Characteristics of different datasets of objects that may be used for incremental learning compared to CRIB. Below the horizontal line are characteristics especially relevant to developmentally plausible incremental learning.

ples [13, 16]. Recent works on incremental learning have developed methods using distillation loss [29] and exemplars [38, 5] 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 reproducability. CRIB comes with a novel dataset, Toys-200, consisting of 3D models of 200 diverse and developmentally-appropriate object instances. Our initial experiments with CRIB have already 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 instances and categories across a diversity of datasets (Toys-200, ShapeNet [6], and CIFAR[25]).

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 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 instance and category learning over time.
- 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 weakly-supervised incremental learning tasks.

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 [5]. In comparison to these works, we provide a novel learning environment (CRIB with Toys200) 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, 23, 28, 32, 38, 5] did not address instance learning or the use of 3D models to learn from contiguous viewpoints. They addressed only 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 [1, 2, 9]). 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 [12], and efforts to leverage commercial video games [24, 40, 39]. Multiple synthetic optical flow datsets [33, 4, 46] have led to performance improvements, as have generated 3D car assets from [35]. 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 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 [6] 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 Toys200 objects are very diverse in comparison to both Sculptures and ShapeNet and were designed to be 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 1.

This paper is also connected to a long-standing interest in developmentally-inspired approaches to robotics and learning [8]. Works such as Gepperth et. al. [14] and Kanan et. al. [22] connect to our interest in biologically-inspired incremental learning. Recent work by Haber et. al. [17] 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

Since our goal is to explore the behavior of incremental object recognition algorithms in a developmentally plausible setting, we require a visual learning environment that allows this. Such an environment requires the following:
Unlimited Data: The ability to efficiently generate unlimited visual data for relevant objects is critical. This is because infants receive vast amounts of visual information each day, and this ability would enable us to understand the relevance of the amount of data available.
Developmental relevance: The visual data produced has to model object exploration behaviors in early infancy. Objects available in the environment have to be objects that can be found in an infant's environment. The tasks enabled by the produced data should be relevant to infant learning.
Integration: To be generally useful, our learning environment has to be easy to integrate with the existing data loading mechanisms of 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, including toy objects. 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 achieves our goal of an visual learning environment.

3.1. CRIB Learning Environment

3.1.1 3D object models:

A highly diverse set of toy-like objects is central to generating developmentally plausible object instance data for visual recognition. We collect the Toys-200 dataset, consisting of 200 unique toy object models from the freely available library Blendswap [3] under CC licenses. To build a challenging and visually diverse (See Figure 2) dataset we initially collect synthetic replicas of 30 specific object instances used in research with infants [49] and supplement it with additional toy-like objects. The toy like criterion was decided by students collecting the data based on the characteristics of initial set of 30 objects. A specific material shader was developed to give Toys-200 a more toy-like, plastic appearance. To support category learning, we use the ShapeNet [6] dataset with the trade-off that ShapeNet contains objects foreign to an infant's environment.

3.1.2 Generating data:

CRIB generates learning exposures of object instances (See Figure 3). An object exposure is a sequence of contiguous views of a rotating object. Rendering object exposures requires initial user specifications in addition to choosing an object.

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 a tuple of (location, rotation, temperature, strength).

Object motion: To specify characteristics of the rotation of the objects, the user specifies the total number of frames, and the total number of random tuples $(x, y, z, s) =$

Figure 2: A rendering of approximately one third of the 3D models used by the CRIB data generator. These toy-like objects were curated from Blendswap, and adjusted in a way to appear visually diverse in shape and color.

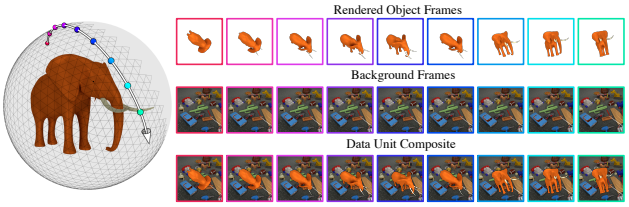


Figure 3: Illustrating the three steps of data generation: 1. object rendering, 2. background selection, 3. foreground and background compositing.

(azimuth, rotation, elevation, scale) that are used to interpolate on the object's viewing sphere and scale.

Preprocessing: Once the specification is made, the following process is fully automated. First an object is imported in Blender, to a scene where a camera is placed directly above the object. The imported geometry is joined to form a single object, 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 during the rotating motion around its center of mass and the change in scale, it will remain inside the camera field of view. Note that due to the various object storage formats, implementing correct automatic importing and pre-processing that will not result in unwanted artifacts such as missing materials, misaligned geometry is nontrivial.

Foreground rendering: The specified lights are placed above the object and the sequence of frames is rendered without a background. At this step, instance segmentations and bounding boxes are collected as well.

Background rendering: Backgrounds are image sequences of objects laid out on a floor in a cluttered manner. The camera above the objects moves slightly over time to loosely emulate head motion (see second row of Figure 3). Our aim in doing this is to include a dynamic background environment, cluttered with different objects from the one in focus. When rendering the background images, we make sure that the foreground object is not in the background. The last step is to composite the foreground and background, and add a small amount of pixel-wise noise.

Testing data: To evaluate whether an algorithm can recognize an object, CRIB generates views of the object at random rotations, elevations and scales with random lighting conditions and backgrounds.

Technical information for all steps of the data generation procedure is available in the supplementary materials.

3.1.3 Integration:

CRIB is implemented as a simple API. Once the user instantiates it for an object dataset before training, calls to the generating API can be made that then return images and ground truth information. Integration with CRIB is as straightforward as loading images and annotations from disk.

3.2. Learning Tasks in CRIB

Characteristics of CRIB such as the ability to exhaustively sample the appearance of an object, allow us to investigate multiple new incremental learning tasks. For all tasks only one object is shown to the algorithm at a time.

Supervised Single Exposure: This task is close to standard incremental learning, where class partitioned images come from an existing image collection, so that images from each partition are only used once for training. This is in contrast to standard batch learning in which images are drawn at random with replacement from the entire dataset. In this task algorithms have to learn with one labelled exposure per object instance.

Supervised Repeated Exposure: In this task, incremental learning algorithms can be repeatedly exposed to objects from previous exposures. To the best of our knowledge, prior work has not investigated the case where repetition is available. This task is enabled by CRIB, and allows for novel insight into deep incremental learning algorithms.

Unsupervised Repeated Exposure: In this task, incremental learning algorithms can also be repeatedly exposed to objects from previous exposures but no labels are provided for each exposure. This is similar to discriminative incremental clustering [15]. The study of this problem is intriguing since infants can learn with almost no explicit supervision as discussed in the introduction, leading us to take the first step in investigating whether incremental learning algorithms can do the same. Learning from unlabelled repeated exposures is significantly more challenging since the learned models are not guaranteed to have a consistent identity, and the algorithm is also responsible for novelty detection and re-identification of previously seen objects.

4. Experiments

In this section we introduce the baseline algorithms we use in our study, and show novel results in three incremental learning sub-tasks. We use incremental accuracy for object instance recognition, measured at the end of each object exposure as in [38]. At the end of each learning exposure, the model's classification accuracy is computed on test samples from the set of objects or categories seen until then. The average of the per object instance or per category recognition accuracy values over time produces an incremental accuracy curve and a final incremental accuracy value after the last contiguous view is seen. We use this metric for all experiments.

For each of the learning tasks generated using CRIB, contiguous views are 100 frames long and are generated by interpolating between three random points on the viewing sphere of an object, with scale smoothly varying from 0.3 to 1.1 of the original scale of the object, unless stated otherwise. As mentioned in §3 each light source is defined by a 4-tuple (location, rotation, temperature and strength). Over a data unit, location and rotation are slightly jittered, random values within a range that produce a favorable visual result are chosen for strength, and within 4000-6000K (indoor lighting range) are chosen for temperature. Testing is done on 100 frames of random object views, scale and lighting for each object that has been seen previously. Additional training details are included in the supplementary materials.

4.1. Incremental Learning Methods

We implement three recent deep network based incremental learning algorithms [29, 38, 5] with slight changes. All methods use ResNet-34 [18] as the backbone architecture.

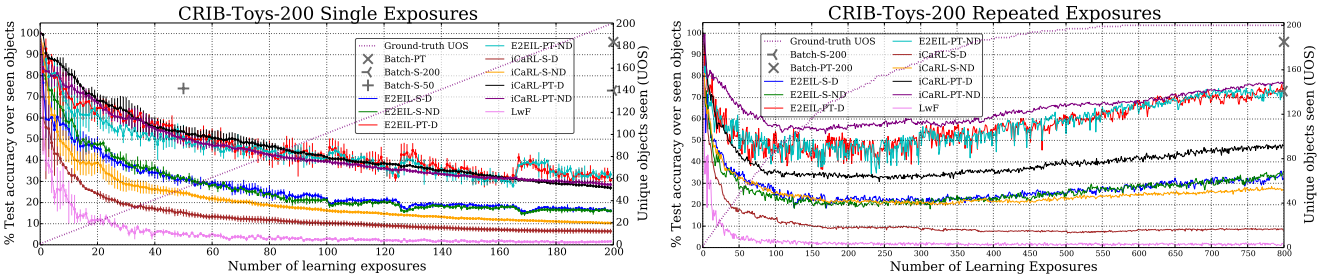
LwF [29] is a strategy around a standard CNN to address catastrophic forgetting for image classification. For all new classes at each time step, the fully connected layer of the CNN is expanded by adding output sigmoid units. Distillation loss is applied to the old outputs, to keep the network parameters from significantly changing because of gradient backpropagation from new data. Our LwF differs from [29] in using sigmoid units rather than a softmax layer for classification and uses additional data augmentation.

iCaRL [38] builds on LwF by including explicit memory in the form of an exemplar set that is managed as the algorithm encounters more exposures. iCaRL uses this exemplar set to perform nearest exemplar mean classification in feature space. The inference procedure consists of computing normalized exemplar mean features per class using the underlying CNN, and then classifying by determining the nearest exemplar mean from the normalized feature of each testing sample. Signal for backpropagation comes from the fully connected layer. The difference between our iCaRL and [38] is that we use additional data augmentation.

E2EIL [5] builds upon the previous two methods. The loss takes into account ground truth labels of the samples from old classes as well as new classes. The method is end-to-end since it uses the network outputs for classification. In addition to the training at each time step, E2EIL adds balanced fine-tuning which targets unbalanced training, when the number of samples from old classes is significantly lower than the number of samples for the new class. Exemplar set construction is similar to iCaRL, but is done twice: right after training and after balanced fine-tuning. Our E2EIL adapts a new distillation loss which is suitable for the case with only one object at a time, with a temperature-squared weighting [19] and different data augmentation. See supplemental materials for details on the changes to all algorithms.

We experiment with versions of iCaRL and E2EIL using a pre-trained ILSVRC-2014 [10] architecture, and not using distillation. Based on prior transfer learning results [44, 48, 11], it is expected that methods starting from a pretrained architecture should also perform better for incremental learning. We confirm these trends in our evaluation and perform extensive tests using pre-trained models.

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 method trained from a random initialization using distillation loss. In our experiments iCaRL without distillation can use labels from the exemplar set.



(a) (b)

Figure 4: (a) Performance of LwF, 4 variants of iCaRL and 4 variants of E2EIL when presented with a single exposure for each object instance from CRIB-Toys-200. Standard-deviation bars were computed over 3 runs for every experiment—each with different random orderings of objects. (b) shows performance of the same methods with repeated exposure.

4.2. Catastrophic Forgetting in CRIB

In this task, algorithms are trained on 200 exposures, each being one data unit of a unique toy object generated by CRIB. This differs from the task in [29, 38, 5] in that new instances are shown one at a time. For this experiment, explicit memory based methods (E2EIL, iCaRL and variants) are allowed an exemplar set size of 600 images, or 3% of the total data that the algorithm will be exposed to. In [38, 5] for CIFAR, the authors allow for 4% of the total data.

As evident in Figure 4a, all methods [4] have a general downward trend, which is similar to previous results on other datasets [38, 29, 5]. This trend is generally attributed to catastrophic forgetting of classes which were seen early on in the experiment, and the growing number of concepts the algorithm has to learn over time. The results for iCaRL-S-ND show that distillation is not as effective as training with exemplar labels, while the results for E2EIL-(S/PT)-ND and iCaRL-PT-ND show that distillation loss might not be necessary since the models trained with distillation loss do not perform much better if at all, as compared to their respective counterparts without distillation loss. Furthermore, the experiment with (iCaRL/E2EIL)-PT-(D/ND) shows that test accuracy can be easily improved by using a pre-trained model[5]. This finding aligns with [44, 48, 11], where pre-training allows for better performance. To verify our findings, and confirm the difficulty of data generated by CRIB we repeat this experiment with CIFAR-100 [26] with iCaRL-S-ND and iCaRL-PT-ND (see Figure 5).

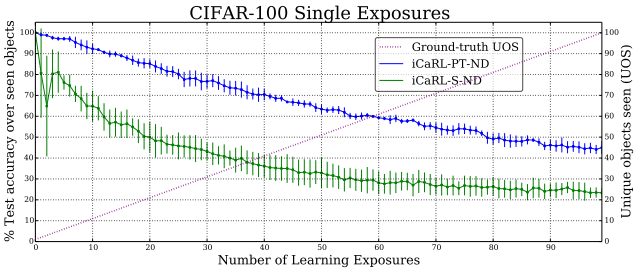


Figure 5: Performance of iCaRL-S-ND and iCaRL-PT-ND on CIFAR-100. The error bars have been plotted over 3 runs each with a different random order of categories exposed. Both learning algorithms have an exemplar set size of 2000

4.3. Repetition Reduces Catastrophic Forgetting

Through our experimental evaluation in this section we find that (1) the majority of tested incremental learning algorithms exhibit improvement in accuracy (2) repeated exposures allow incremental learning algorithms to eventually achieve an accuracy close to a pre-trained batch learning method. We include an exemplar sensitivity study by varying the number of exemplars, training new models and evaluating the incremental accuracy for the algorithms.

To initially investigate the behavior of algorithms when repetition is allowed, we perform experiments with 200 objects over 800 data units (each object appearing four times), with an explicit memory of 600 exemplars. For every experiment run, we generate a random sequence of object instances such that all methods experience the same number of objects by any time step, but not necessarily the same instances in the same order. Figure 4b shows the results for each method in this experimental task. All learning algorithms, except LwF and iCaRL-S-D, demonstrate some ability to leverage repeated exposures to improve the classifier's performance over time. The relative performance for the algorithms in this task aligns with the trend shown in the single exposure task. Note that the performance gap be-

[4] In our experiments we found that even when starting from a network with pre-trained weights, LwF performed worse than iCaRL or E2EIL methods that start from a randomly initialized network, and hence, have not included LwF-PT-D in the results.

[5] ResNet-34 [18] model pre-trained on ILSVRC-2014 classification dataset [10].

tween iCaRL-(S/PT)-D and iCaRL-(S/PT)-ND in this task is bigger compared to the single exposure task. This potentially indicates that distillation loss is hindering the ability to leverage repeated exposures for iCaRL and showing the advantage of simply storing exemplar labels.

We perform experiments using three datasets: CRIB-Toys, CRIB-ShapeNet [6] and CIFAR [25] 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 use an ImageNet pre-trained [10] backbone architecture for all algorithms and perform the following experiments:

1. **CRIB-Toys**: 50 objects over 500 data units (each object is shown 10 times).
2. **CRIB-ShapeNet**: 20 categories over 500 data units. (25 instances from each category are shown)
3. **CIFAR**: 20 categories over 1000 data units (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. To the best of our knowledge, previous work has not shown such increasing trends for incremental learning algorithms.

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 exposure.

CRIB-ShapeNet is a categorization task where repeated exposure 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 [6]. Data units 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 category which has been seen. The performance, seen in Figure 6d, demonstrates that repeated exposures to completely novel instances in the same category leads to improvements on a categorization task. Furthermore, this experiment extends our initial finding of improvement towards batch performance via repeated exposures to a different task and dataset.

For **CIFAR**, we sample with replacement 100 images in a category (500 images) for each learning exposure. This allows the algorithm to be exposed to images repeatedly during different exposures. In Figure 6c we can see that the performance on iCaRL-PT-ND decreases at first and starts to go up after all 20 objects have been seen. Coverage for each category is the portion of unique images seen within a category, with the mean over all categories shown on the plot. iCaRL-PT-ND improves faster before it reaches 100% average coverage. After that point, the improvements are at a slower pace. The performance at the end of 1000 exposures is on par with a batch algorithm. This result shows that gains 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 old instance it belongs to, prior to training. The algorithm has to construct its own object models in this manner.

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, the algorithm finds the distance in the unit normalized feature space of the images from an exposure from the exemplar means. This is followed by computing the mean of these distances over the images in the exposure, and using this as a score to classify the current exposure as one of the objects which the learning algorithm has a model for. If the minimum distance-score is more than a given threshold, the exposure is deemed to have come from a novel instance, otherwise, it gets classified as the old instance with the minimum distance-score. The threshold used was found as the optimal operating point from a Precision-Recall analysis over the binary classification problem of novelty detection.

We evaluate this baseline algorithm on different data unit length and repeated exposure tasks on CRIB. After any exposure, testing is done on random views of all objects (ground truth) seen up to that exposure. 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 shows ablation study results. For 100 data unit length, the algorithm's accuracy is constant or decreases with a greater number of objects and four repeated exposures. Also for the same data unit length, the proportion

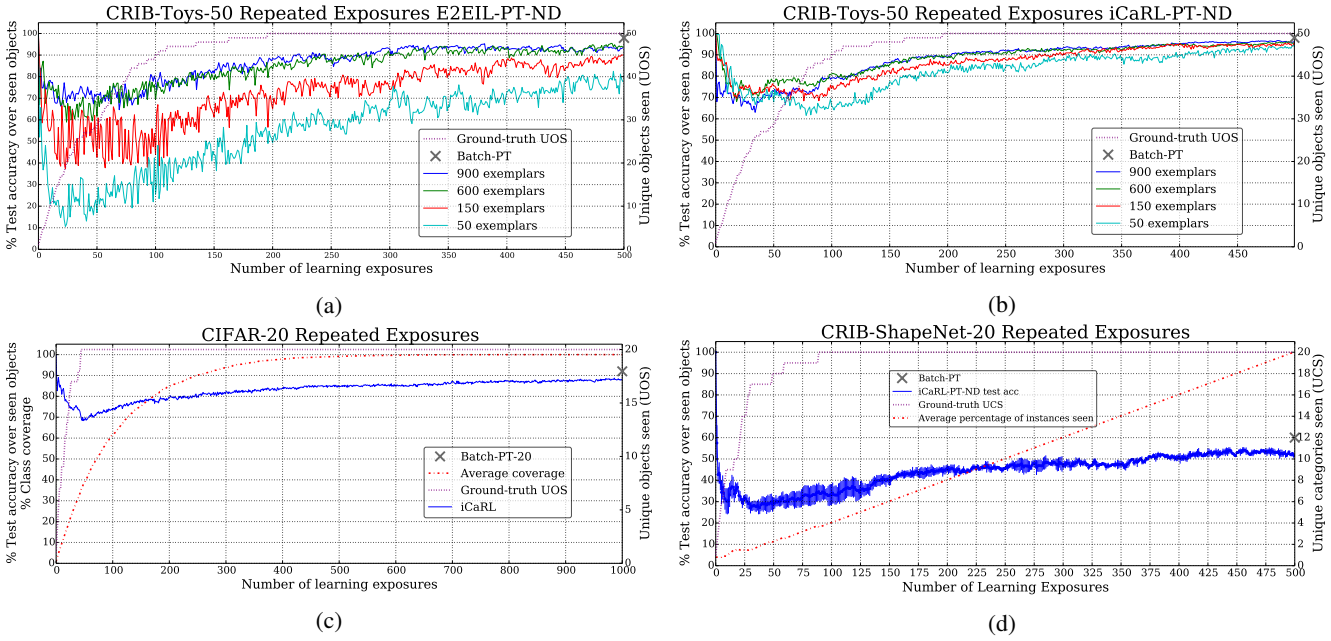


Figure 6: Top : Performance of (a) iCaRL-PT-ND and (b) E2EIL-PT-ND with different number of exemplars on 50 objects of CRIB-Toys. Bottom : (c) Performance of iCaRL-PT-ND (400 exemplars) on 20 categories of CIFAR (d) Performance of iCaRL-PT-ND (1500 exemplars) on 20 categories of CRIB-ShapeNet. All experiments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

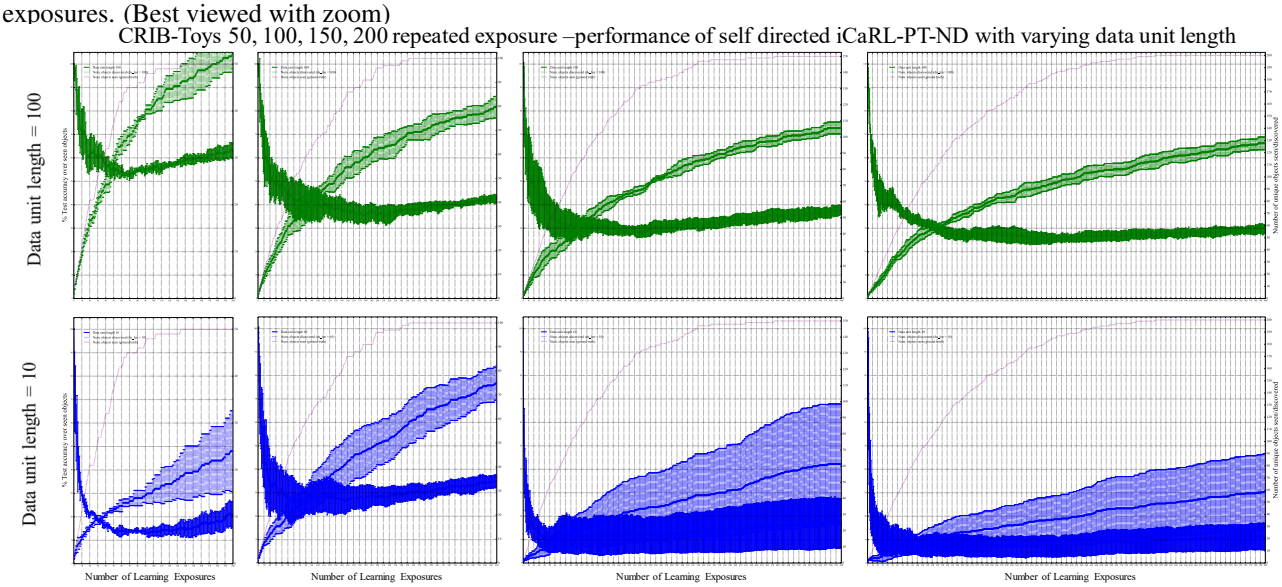


Figure 7: Performance of iCaRL-PT-ND (600 exemplars) on unsupervised repeated exposures of CRIB-Toys-50, 100, 150, 200 (left to right) with data unit lengths of 10 and 100. The 10 length data units are the first 10 frames of the 100 length data unit. (Best viewed with zoom)

of objects discovered as compared to the ground truth number of unique objects seen decreases as with greater object set sizes. Across all experiments with different total numbers of objects, there is a consistent trend that a smaller data unit length results in a lower final accuracy. Furthermore, a lower data unit length seems to result in higher variability in performance over multiple runs with different order of objects encountered.

5. Conclusion

We introduce CRIB, a novel visual learning environment for object recognition that models infant learning. Through empirical evaluation we determine the visual difficulty of CRIB. Further, via experiments on CRIB, the new Toys-200 dataset and other datasets we reach novel findings about the positive effects of repetition on catastrophic forgetting. Finally, we show intriguing results on a challenging new task that has yet to be successfully tackled by deep learning approaches. We hope that this work will enable and motivate the computer vision community to develop new incremental learning methodology.

References

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

[2] 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. 2

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

[4] 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

[5] 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

[6] 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, 7

[7] 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

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

[9] R. De Rosa, T. Mensink, and B. Caputo. Online open world recognition. *arXiv preprint arXiv:1604.02275*, 2016. 2

[10] 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, 6, 7

[11] 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

[12] 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

[13] R. M. French. Catastrophic Forgetting in Connectionist Networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. 2

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

[15] 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

[16] I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. 2

[17] N. Haber, D. Mrowca, L. Fei-Fei, and D. L. Yamins. Learning to play with intrinsically-motivated self-aware agents. *arXiv preprint arXiv:1802.07442*, 2018. 3

[18] 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, 6

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

[20] 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

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

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

[23] 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

[24] 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

[25] A. Krizhevsky, V. Nair, and G. Hinton. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html (vi sited on Mar. 1, 2016)*, 2009. 2, 7

[26] A. Krizhevsky, V. Nair, and G. Hinton. The CIFAR-100 Dataset. *online: https://www.cs.toronto.edu/ kriz/cifar.html*, 2014. 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. 2, 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. 2, 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. 2, 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)*, July 2017. 2, 5, 6

[39] S. R. Richter, Z. Hayder, and V. Koltun. Playing for benchmarks. 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. 2, 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