Does Continual Learning = Catastrophic Forgetting?
Anh Thai, Stefan Stojanov, Zixuan Huang, Isaac Rehg, James M. Rehg
Georgia Institute of Technology
{atha6,sstojanov,zixuanh,isaacrehg,rehg}@gatech.edu

Abstract

Continual learning is known for suffering from catastrophic forgetting, a phenomenon where earlier learned concepts are forgotten at the expense of more recent samples. In this work, we challenge the assumption that continual learning is inevitably associated with catastrophic forgetting by presenting a set of tasks that surprisingly do not suffer from catastrophic forgetting when learned continually. We provide evidence that these reconstruction-type tasks exhibit positive forward transfer and that single-view 3D shape reconstruction improves the performance on learned and novel categories over time. We provide the novel analysis of knowledge transfer ability by looking at the output distribution shift across sequential learning tasks. Finally, we show that the robustness of these tasks leads to the potential of having a proxy representation learning task for continual classification. The codebase, dataset and pre-trained models released with this article can be found at https://github.com/rehg-lab/CLRec.

1. Introduction

In continual learning (CL), a stream of incrementally-arriving inputs is processed without access to past data. A key challenge is to avoid catastrophic forgetting [30]—large negative backward transfer (BWT) [22], which arises if previously-learned representations are degraded by more recent exposures. Substantial effort has been made to combat forgetting [10, 12, 21, 52], and it has come to exemplify continual learning. However, past works have explored a surprisingly limited set of tasks, with an almost exclusive focus on classification.

In this work, we demonstrate the surprising finding that a broad set of continual reconstruction tasks, including 3D shape reconstruction, 2.5D sketch estimation and 2D image reconstruction, do not exhibit catastrophic forgetting (Sec. 4). In fact, we show that these tasks exhibit minimal negative backward transfer and even positive backward transfer without the use of any heuristics or strategies to prevent forgetting. To the best of our knowledge, we are the first to provide an extensive study of CL for reconstruction-type tasks. Our findings suggest that the challenge of catastrophic forgetting, established by prior work on CL classification, does not in fact characterize all CL problems.

In addition to mitigating negative BWT, another essential characteristic of successful CL models is achieving positive forward transfer (FWT). Positive FWT [22] arises when a model’s representation, trained on a sequence of prior tasks, is beneficial for a future learning task. This is an important property because it can enable CL methods to leverage shared representations, a common property of batch learning methods (e.g. learning low-level visual features in image classification) and a goal of transfer learning. Collectively, BWT and FWT characterize the effectiveness of CL methods in evolving feature representations incrementally.

Recently, the GDumb (classifier) baseline [26] has called the role of FWT in CL into question. GDumb is an episodic representation learner, which maintains a set of evolving exemplar memory but reinitializes the feature representation and trains it from scratch during each learning exposure, eliminating the possibility of FWT. GDumb was shown to out-perform other state-of-the-art methods which were designed to achieve positive FWT, highlighting the tension between the goals of achieving positive FWT and avoiding negative BWT. These prior findings for classification beg the question of whether achieving positive FWT is beneficial for CL reconstruction. We demonstrate that continuously-updated representations lead to improved performance and positive FWT for a variety of reconstruction tasks (see Sec. 5).

A limitation of prior work on knowledge transfer (i.e. FWT and BWT) is that the findings depend upon the use of a specific CL method. A basic question is whether we can find algorithm-agnostic characterizations of CL tasks that might shed light on the surprising behavior of continual reconstruction. We define a measure of the output distribution shift across learning exposures and demonstrate that small distribution shifts across sequential tasks lead to better knowledge transfer and improvements in BWT and FWT.

1We use the term learning exposure to refer to each new increment of data, i.e. the learner’s next “exposure” to the concepts being learned.
Reconstruction Tasks

- Single-view Depth Pred. (Fig. 2a)
- Single-object Pointcloud 3D Shape Rec. (Fig. 1c)
- Image auto-encoding (Fig. 2c)
- Single-view Image 3D Shape Rec. (Figs. 1a, 1b)
- Single-view Depth 3D Shape Rec. (Figs. 1a, 1b)

We are the first to investigate and demonstrate that a set of CL tasks is intrinsically robust to catastrophic forgetting. While most prior CL works have addressed image classification, a few prior works have addressed various other tasks: Aljundi et al. [3] studied the problem of actor face tracking in video, while [1,7,23,25] explored image segmentation. Shmelkov et al. [34] and Liu et al. [20,40] investigated incremental object detection while [18,44] learned image generation. Elhoseiny et al. [11] examined continual fact learning by utilizing a visual-semantic embedding. Wang et al. [41] studied CL of camera localization given input RGB image while Cai et al. [5] explored online CL of geolocation with natural distribution shift in the input that occurs over real time. Others [2,13,48] focused on reinforcement learning task.

Most closely related to our work is Yan et al. [50] that investigated continual learning of scene reconstruction. Similar to our work, they employed implicit shape representation (signed-distance-field) to represent 3D scenes. In contrast, this work aimed to continually reconstruct the input scene given a stream of depth images from different views. The input distribution shift in this setting is the shift between one view of the scene to another and the objective is to produce a smooth representation of the same input scene observed over time. Our work on the other hand, explores CL of reconstruction task in the context of visual classes, which is more challenging since the underlying semantic in the inputs changes over time. Note that all of these CL works reported challenges with catastrophic forgetting commensurate with the classification setting.

**Analysis of Continual Learning.** Our analysis of the behavior of CL tasks is most closely related to the body of works that analyzes general dynamics of CL [14,39]. While Verwimp et al. [39] examined the benefits and drawbacks of rehearsal methods in CL, Knoblauch [14] showed that optimal CL algorithms solve an NP-HARD problem and require perfect memory. Specifically, optimal parameters \( \theta_t \) for each new task must lie in the intersection of SAT of all tasks learned up to \( t \). Perfect memory refers to the ability to approximate the parameters that optimize all seen tasks. This approach explains the merit of employing memory replay in CL instead of regularization-based approaches. While [17] discussed the different concept drift in CL, our analysis focuses more on the output distribution shift that can be used as a means to understand the knowledge transfer ability of various CL tasks.

**Generalization in Batch-Mode 3D Shape Reconstruction.** Our analysis of the generalization ability of CL 3D single-view shape reconstruction task in Sec. 5 is based on prior works that investigate the ability of single image 3D shape reconstruction models to generalize to unseen shape categories in batch mode [33,37,53]. We are the first to provide generalization analysis of these models in the CL setting, utilizing the 3-DOF VC approach which has been

| Input Rep. \( \rightarrow \) Output Rep. | Reconstruction Tasks |
|-----------------------------------------|----------------------|
| 2D \( \rightarrow \) 2D                 | Image auto-encoding (Fig. 2c) |
|                                         | Single-view Silhouette Pred. (Fig. 2b) |
| 2D \( \rightarrow \) 2.5D                | Single-view Surface Normals Pred. (Fig. 2a) |
| 2D \( \rightarrow \) 3D                 | Single-view Image 3D Shape Rec. (Figs. 1a, 1b) |
| 2.5D \( \rightarrow \) 3D               | Single-view Depth 3D Shape Rec. (Figs. 1a, 1b) |
| 3D \( \rightarrow \) 3D                 | Single-object Pointcloud 3D Shape Rec. (Fig. 1c) |

Table 1. Summary of the reconstruction tasks we evaluate that demonstrate robustness to catastrophic forgetting. There are 5 types of tasks based on the input to output representation mapping. (Sec. 6). This gives us the ability to “forecast” the performance of a supervised CL task in a way that is agnostic to the algorithm and backbone architecture design. We believe these are the first results to demonstrate the feasibility of approximating the difficulty of a CL task without performing computationally expensive model training.

As a means to further investigate the relationship between continual reconstruction and categorization, we demonstrate that continual single-view 3D shape reconstruction can serve as an effective proxy task for classification. Specifically, a continuously-trained shape representation (without any class label supervision) is effective for continual image classification given only a small exemplar budget (Sec. 7). In summary, this paper makes the following contributions:

- The novel finding that some continual reconstruction tasks (Tbl. 1) do not suffer from catastrophic forgetting (Sec. 4).
- That these continual reconstruction tasks demonstrate positive forward transfer and the ability of single-view 3D shape reconstruction to generalize to novel classes unseen during training (Sec. 5).
- Novel analysis of knowledge transfer ability in CL demonstrates that smaller output distribution shift across learning exposures leads to better knowledge transfer in CL (Sec. 6).
- Using single-view 3D shape reconstruction as a proxy task for classification results in a competitive CL method given a limited exemplar budget (Sec. 7).

**2. Related Work**

Our work is most closely-related to four bodies of prior work: 1) CL works outside of the image classification paradigm (relevant to our findings on CL for reconstruction), 2) Analysis of CL (relevant to our output distribution shift analysis), 3) Generalization ability of models for single image 3D shape reconstruction (relevant to our investigation of generalization ability of CL single-view 3D shape reconstruction models), and 4) CL for classification task (relevant to our proxy representation task findings).

**CL of Non-Classification Tasks.** We are the first to investigate and demonstrate that a set of CL tasks is intrinsically robust to catastrophic forgetting. While most prior CL works have addressed image classification, a few prior works have addressed various other tasks: Aljundi et al. [3] studied the problem of actor face tracking in video, while [1,7,23,25] explored image segmentation. Shmelkov et al. [34] and Liu et al. [20,40] investigated incremental object detection while [18,44] learned image generation. Elhoseiny et al. [11] examined continual fact learning by utilizing a visual-semantic embedding. Wang et al. [41] studied CL of camera localization given input RGB image while Cai et al. [5] explored online CL of geolocation with natural distribution shift in the input that occurs over real time. Others [2,13,48] focused on reinforcement learning task.
shown to learn a more general shape representation than the object-centered (OC) approach.

**CL for Classification.** Our work on a reconstruction-based proxy task for CL classification (Sec. 7) is unique, but it is peripherally-related to other CL works which explore alternative classification losses or forms of supervision. We share with Yu et al. [51] the use of the nearest-class-mean (NCM) classification rule. We use NCM for classification based on a latent shape representation trained without class supervision, while Yu et al. use NCM for classification in an embedding layer which is trained with ground-truth class labels. Another related work by Rao et al. [29] performs unsupervised CL in a multi-task setting where the boundaries between tasks are unknown. In contrast, our unsupervised training paradigm utilizes single-view 3D shape reconstruction as a proxy task.

### 3. Problem Formulation

**Supervised Reconstruction Tasks.** The objective of these tasks is to learn the mapping function \( f_\theta : \mathcal{X} \rightarrow \mathcal{Y} \) over the observed data \( \{(x_i, y_i)\}_{i=1}^N \sim D \). For example, single-view 3D shape reconstruction aims to output the 3D shape of the object represented in the input image while depth map reconstruction predicts the depth values of the scene/object given in the input. Note that \( x \) and \( y \) can be different depending on the specific reconstruction task considered. Generally, the desired function \( f \) is a composition of two functions \( f = D \circ E \). The encoder \( E \) extracts the feature representation from the input \( x \), followed by the decoder \( D \) that produces the output \( y \) from the encoded feature representation.

**Continual Learning of Reconstruction.** In this setting, at each learning exposure \( t \), the learning model observes the data \( \{(x_i^{(t)}, y_i^{(t)})\}_{i=1}^{N_t} \sim D_t \) indexed by \( t \in \{1, 2, \ldots, T\} \) and learns to optimize the parameters \( \theta_t \) of the function \( f_{\theta_t} : \mathcal{X}_t \rightarrow \mathcal{Y}_t \) by minimizing the supervised loss \( \mathcal{L}(\theta_t) = \mathbb{E}_{D_t}[\ell(y^{(t)}, f_{\theta_t}(x^{(t)}))] \) where \( \ell(\cdot, \cdot) \) is some loss function associated with the specific reconstruction task. We employ the notion of single exposure to refer to the standard continual learning paradigm where data is introduced sequentially and never revisited while repeated exposures refers to the paradigm introduced in Stojanov et al. [35] where data can be revisited after being learned. In this setting, each visual class occurs a fixed number of times (e.g. 10 repetitions) in random order\(^2\). Note that in this work, we assume that each \( D_t \) is defined over a set of \( M_t \) visual categories \( \{C_k^{(t)}\}_{k=1}^{C_t} \).

**Training.** During training, the learning model does not have access to previously seen data \( D_{1:t-1} \). We optimize the parameters \( \theta_t \) of the function \( f \) continuously at each learning exposure upon observing the data stream \( D_t \). Specifically, the learned parameters \( \theta_{t-1} \) at exposure \( t-1 \) serve as the initialization parameters for the model at exposure \( t \), which we referred to as continuous representation learning. This is the standard SGD training that has been shown to suffer from catastrophic forgetting in prior works. Without any further heuristics such as additional losses, external memory or other methods employed, this technique is referred to as fine-tuning strategy [19].

**Evaluation.** At test time, the model is evaluated on the test split of all known categories. Specifically, at each learning exposure \( t \) we compute the average accuracy of all classes seen up to \( t \). Specifically, \( \text{Acc}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \text{acc}_i^{(t)} \) where \( N_t \) is the number of classes seen up to exposure \( t \) and \( \text{acc}_i^{(t)} \) is the accuracy of class \( i \) after learning exposure \( t \). Plotting the average accuracy at all learning exposures results in the learning curve of the CL model (e.g. Fig. 1a). Note that accuracy metrics reported for all the tasks are in range \([0, 1]\).

We further report backward and forward transfer metrics [22] in addition to the average performance curve at each learning exposure. Specifically, backward transfer (BWT) measures the average change in performance in the last learning exposure w.r.t when the concepts are first introduced and forward transfer (FWT) indicates the average change in performance between the random initialization and the performance of the learning exposure right before the concepts are introduced. Note that while BWT is bounded in \([-1, 1]\), FWT depends on the random initialization performance on each dataset. A more successful CL learner will demonstrate higher BWT and FWT.

**Positioning CL Reconstruction.** Considering the three continual learning scenarios [38], CL reconstruction is most
closely related to Domain-IL scenario. In both settings, the learning objective (e.g., depth value prediction) is the same across learning exposures, but the input data distribution changes over time (from one set of visual classes to another). This breaks the i.i.d.-sampling assumption present in the SGD optimization procedure where each training minibatch is sampled i.i.d. from the entire data distribution. This presents the same main challenges faced by other CL tasks such as classification, object detection, and segmentation.

4. Reconstruction Tasks Do Not Suffer from Catastrophic Forgetting

We identify 5 types of reconstruction tasks based on their input and output properties, as listed in Tbl. 1. Our key finding is that CL tasks of each of these five types do not suffer from catastrophic forgetting. It is important to emphasize that the “continual learning” algorithm used in this section is the simple finetuning strategy specified in Sec. 3 that is known to perform poorly for CL classification task. Specifically, we do not need to utilize additional losses, external memory, or other methods to achieve good continual learning performance.

Note that different categories of shapes exhibit significant domain shift that poses significant challenges to continual learning. For example, the categories “chair” and “bowl” in ShapeNet define very different 3D data distributions with no parts in common. From this point of view, it is quite surprising that we do not observe forgetting for such continual reconstruction tasks. We therefore organize shapes by category in constructing our learning exposures, so that the category label is a means to characterize the domain shift between successive exposures.

Our findings for learning 3D shape reconstruction, 2.5D prediction, and 2D reconstruction are presented in Secs. 4.1, 4.2, and 4.3 respectively. We report the average accuracy at each learning exposure as described in Sec. 3 and backward transfer for all the experiments.

4.1. Single Object 3D Shape Reconstruction

We first present reconstruction tasks where the output representation is in 3D. Specifically, the goal of the desired function $f$ is to produce a 3D surface representation of the single object present in the input. We examine implicit continuous 3D representations such as signed-distance-fields (SDF) and continuous occupancies since they were identified to achieve superior performance in the batch setting [24,37,49].

**Approach.** We utilize SDFNet [37] and OccNet [24] as backbone architectures for CL. We train both methods with the 3-DOF VC representation (varying in azimuth, elevation and camera tilt) from [37], which was shown to give the best generalization performance. We also train with OC representation for SDF representation, in which the model is trained to output the shape in the canonical pose. We examine the behavior of these models given different input representations: 2D where inputs are single-view RGB images, 2.5D where inputs are ground truth depth and normals maps and 3D where inputs are sparse 3D pointclouds.

**Datasets & Metric.** We train on all 55 classes of ShapeNetCore.v2 [8] (52K instances) with 5 classes per exposure for the single exposure case, and on the largest 13 classes of ShapeNetCore.v2 (40K meshes), denoted as ShapeNet13, with 2 classes per exposure for the repeated exposure case. Note that ShapeNetCore.v2 is currently the largest shape dataset with category labels and ShapeNet13 is the standard split for 3D shape reconstruction. Each exposure is generated from all of the samples from the training split of each category currently present. Following prior works in shape reconstruction [36,37,49] we report the average FS@1 at each learning exposure. We use SDFNet as the batch reference. All models are trained from random initialization.

**Results.** The results are shown in Figs. 1a, 1b and 1c for single and repeated exposures on all single object 3D shape reconstruction settings (last 3 rows of Tbl. 1). For single exposure with 2D and 2.5D inputs (Fig. 1a), all algorithms maintain their accuracy over time and even exhibit a slight upward trend of increasing accuracy while for 3D inputs (Fig. 1c) the performance significantly increases over time and is on par with batch. Note that we conducted 3 runs and the results converge to the same conclusion with an average std of 0.02 at each learning exposure. All models including the model trained with OC representation do not suffer from catastrophic forgetting as evidenced by the minimal nega-
tive and even positive backward transfer. This is surprising since we are not taking any steps to ameliorate catastrophic forgetting and each learning exposure presents a significant domain shift, as the learner must incorporate information about the shape of a new class of objects. Since SDFNet and OccNet differ significantly in shape representation and are trained with different losses ($L_1$ loss for SDFNet and BCE loss for OccNet) this finding possibly reflects a basic property of the shape reconstruction problem rather than the inductive biases of a particular model.

In the repeated exposures setting (Fig. 1b), the performance of both architectures when trained with 3-DOF VC improves significantly over time, and eventually performs on par with the batch learner. These models achieve significant positive BWT which indicates that catastrophic forgetting is mitigated. Unlike the experiments in [35], which showed similar asymptotic behavior for classification accuracy, these results were obtained without exemplar memory or other heuristics. Note that SDFNet trained with OC does not show a significant increase as 3-DOF VC over time. This complements the finding in [37] that training with 3-DOF VC results in a more robust feature representation.

4.2. Single-view 2.5D Sketches Prediction

The task in Sec. 4.1 requires the model to infer the global 3D structure of each object. In this section we investigate the related task of estimating depth and surface normals (2.5D) from RGB input images in the single exposure case (Tbl. 1, second row). We adopt the U-ResNet18-based MarrNet [45] architecture, with an ILSVRC-2014 [32] pre-trained ResNet18 for the image encoder. We evaluate depth prediction using the commonly used thresholding accuracy [15, 28]. For normals prediction, we report the accuracy based on the cosine distance threshold between the predicted and ground truth surface normals [43]. Fig. 2a demonstrates that single exposure 2.5D prediction does not suffer catastrophic forgetting as the accuracy increases over time. These findings further extend the 3D shape reconstruction results from Fig. 1.

4.3. 2D Reconstruction

The tasks in Secs. 4.1 and 4.2 require the model to solve a challenging 2D to 3D inference problem. We conduct additional experiments on continual 2D to 2D mapping that includes learning to segment foreground/background given an RGB input image and image autoencoding. For silhouette prediction, we utilize U-ResNet18-based MarrNet [45] architecture train with BCE loss. We report the average IoU at each learning exposure as in Fig. 2b which demonstrates that single exposure silhouette prediction does not suffer from significant forgetting (minimal negative backward transfer). In fact we observe that the IoU increases over time.

For image autoencoding, we present results in Fig. 2c. We use a randomly initialized shallow architecture with 4 conv. layers followed by max pooling for the encoder where the bottleneck feature vector has dimension $16 	imes 2 	imes 2$. We experiment on CIFAR-100 [16] (size $32 	imes 32$) with one class per exposure and use SSIM [42] scaled to range $[0, 1]$ as the accuracy metric. SSIM increases over time and eventually reaches batch performance. This is yet more evidence for the robustness of continual reconstruction.

4.4. Discussion of CL Reconstruction & Limitations

We have identified (for the first time) a set of continual reconstruction tasks that do not suffer from forgetting, as evidenced by the results in Figs. 1 and 2. These models were trained with standard SGD, without exemplar memory or other heuristics. One point of contact between classification and reconstruction is that both sets of tasks benefit significantly from repeated exposures (see Fig. 1b). In Sec. 5 we demonstrate the value of continuously-updated representations on both seen and novel classes.

We now briefly discuss two potential limitations of our work. First, our reconstruction experiments, with the exception of image autoencoding, all use synthetic 3D object models as opposed to real-world images. However, we point out that 3D shape reconstruction on synthetic images (with 2D and 2.5D inputs) is still a very challenging computational problem, e.g. the SOTA result on ShapeNet13 is an FS@1 of 0.5 out of a maximum of 1.0 [37]. This in turn raises the second possible limitation, that the lack of forgetting may be tied in part to the fact that the models are not yet able to achieve very high accuracy. More accurate models might be more closely tuned to the data distribution in each exposure, increasing the potential for domain shift. While this might be true for 3D shape reconstruction from 2D and 2.5D inputs, the 3D shape reconstruction from 3D inputs, 2.5D and 2D reconstruction tasks achieve a high level of accuracy, which provides a counterpoint to the argument. In Sec. 6 we provide one explanation for these observations. We hope this work will encourage the community to conduct additional investigations into this intriguing phenomenon.

Negative societal impacts. Training CL models is computationally expensive since the models are trained to convergence for multiple learning exposures. This can be mitigated by algorithm improvements that allow models to learn faster. Keeping exemplar memory for replay might violate the privacy policies for sensitive data, which can be addressed by generative methods.
5. Forward Transfer in CL Single-view 3D Reconstruction

In this section, we discuss the ability of the learning model to propagate useful representations learned in the past to current and future learning exposures (positive FWT). We focus our analysis on the challenging problem of single-view 3D shape reconstruction. We first demonstrate that continuous representation learning is beneficial as we observe significantly stronger performance compared to episodic representation learning for this task. We further note that positive FWT is obtained, as evidenced by the accuracy improvement on seen and novel classes over time. While generalization to unseen classes has been studied extensively in the batch setting of single-view 3D shape reconstruction, and has been identified to be a significantly challenging problem [37,53], we are the first to analyze this behavior in a continual learning setting.

GDumb [26] is an episodic representation learner, designed to test the hypothesis that there is no value in continuous representation learning. Specifically, at each learning exposure, the model is randomly reinitialized and trained from scratch on the exemplar set which ensures that a subset of data from all previous learning exposures is available. This approach surprisingly achieves competitive performance at classification. We hypothesize that in contrast to this observation, continuous representation learning improves the performance in single-view 3D shape reconstruction, and has been identified to be a significantly challenging problem [37,53], we are the first to analyze this behavior in a continual learning setting.

We conduct our experiments on ShapeNet13 with single exposure and 1 shape class per learning exposure. We choose $K = 1000$ (3.7% of total training data) to be the exemplar set size and evaluate the performance of the models on all learned classes (Sec. 3). In Fig. 3a we observe that the performance of GSmart improves over time and eventually exceeds that of GDumb by 0.15 FS@1. This significant gap highlights the benefit of continuous representation learning across each learning exposure.

We further investigate the ability of these models to generalize to novel categories. We evaluate GDumb, GSmart and continual SDFNet (C-SDFNet) (Sec. 4.1) on a held out set of 42 classes of ShapeNetCore.v2 with 50 instances for each category (Fig. 3b). All algorithms perform poorly on the unseen classes after the initial learning exposures, which demonstrates that it is significantly challenging to generalize to novel categories after learning on only a few classes. However, the performance of C-SDFNet and GSmart improves over time as more classes are learned while GDumb remains low. This illustrates benefit of continuous representation learning as a useful feature that aids generalization and improves the performance on novel classes over time. We note that positive FWT is also observed in other reconstruction tasks (see Tbl. 2).

6. Analysis of Knowledge Transfer Ability

Our findings in Secs. 4 and 5 have highlighted the significance of knowledge transfer in CL reconstruction. While BWT and FWT quantify the knowledge transfer during CL, they require training and evaluating computationally expensive CL models. Furthermore, these measures only reflect the performance of specific CL algorithms and do not speak to a CL task in general. In this section, we attempt to gain more insight into knowledge transfer given a task and a dataset in an algorithm-agnostic manner, by focusing on changes in the output distribution. We use this approach to further analyze the benefit of exemplar memory in classification. We begin by stating a key hypothesis connecting the output distribution to CL task knowledge transfer ability.

**Hypothesis:** Given a supervised task $T$ with input and output distributions $X$ and $Y$, at each learning exposure $t$, the training and evaluating 3D shape reconstruction from 3D inputs on ShapeNetCore.v2 takes 3 days on two NVIDIA GeForce RTX 2080Ti GPUs. On the other hand, computing output distribution distance only takes $\approx 45$ minutes which is two orders of magnitude more efficient.
model observes $X_t \sim \mathcal{X}$ and $Y_t \sim \mathcal{Y}$. We hypothesize that if the distance of the conditional distribution $Y_t | X_t$ between each learning exposure becomes smaller, knowledge (backward and forward) transfer increases for any CL method.

We now present the intuition behind our formulation. Let $\mathcal{D}$ be some dataset consisting of two parts $\mathcal{D}_1$ and $\mathcal{D}_2$ that are independently generated. During batch training we optimize the parameters $\theta$ by minimizing the loss

$$\theta = \arg \min_{\theta} L(\theta) = -\log p(D|\theta) \quad (1)$$

Since $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2$ and $\mathcal{D}_1$ and $\mathcal{D}_2$ are independent, Eq. 1 becomes

$$\theta = \arg \min_{\theta} L(\theta) = -\log p(D_1|\theta) - \log p(D_2|\theta)$$

$$= -\log p(Y_1|X_1, \theta) - \log p(Y_2|X_2, \theta)$$

During continual learning when $\mathcal{D}_1$ and $\mathcal{D}_2$ are learned sequentially, we optimize $L_1(\theta_1) = -\log p(D_1|\theta_1)$ and $L_2(\theta_2) = -\log p(D_2|\theta_2)$ separately, which leads to a sub-optimal solution for $L(\theta)$. When the distance between the conditional distributions $Y_1 | X_1$ and $Y_2 | X_2$ is small, it is more likely that the optimal parameters $\theta_1$ for $L_1$ coincides with the optimal parameters $\theta_2$ for $L_2$ and hence the joint parameters $\theta$ that optimize the batch training model.

**Analysis.** In this section we demonstrate the empirical evidence for the earlier hypothesis. Note that in all of the following analyses, the input $X_t$ is defined to be a visual object category.

**Distribution Distance Metric.** We use the first Wasserstein distance metric (EMD) to quantify the distance between two output distributions. EMD was introduced by Rubner et al. [31] to measure the structural similarity between distributions. In contrast to other statistical measurements like KL divergence or Chi-squared statistics, EMD can be used to measure the similarity between both continuous and discrete distributions with different supports. Given distributions $u$ and $v$, we define

$$d(u, v) = \inf_{\pi \in \Gamma(u, v)} \int_{\mathbb{R} \times \mathbb{R}} |x - y|d\pi(x, y)$$

and express the distance between two learning exposures $t$ and $t'$ as

$$D(t, t') = \frac{1}{|S|} \int_{s \in S} d(u_s, u_{s'})ds \quad (2)$$

| Task            | Mean Dist. | BWT | FWT |
|-----------------|------------|-----|-----|
| Sil Pred.       | 0.075      | -0.003 | 0.836 |
| VC 3D Shape Rec.| 0.077      | -0.123 | 0.105 |
| Depth Pred.     | 0.084      | -0.136 | 0.094 |
| OC 3D Shape Rec.| 0.116      | -0.220 | 0.090 |
| Classification  | 1          | -1   | -0.077 |

Table 2. The relationship between the mean output distribution distance across learning exposures and BWT and FWT for different CL tasks. Lower is better for distribution distance while higher is better for BWT and FWT. Lower distance leads to better knowledge transfer (higher BWT and FWT).

Figure 4. The relationship between the output distribution shift across learning exposures, the number of exemplars per class and BWT on CIFAR-100. Experiments are run for 0, 20, 40, 60, 80, 100 exemplars/class. Larger exemplar set size leads to smaller output distribution distance and higher BWT.

where $u_t$ and $u_{t'}$ are the output distributions at exposures $t$ and $t'$ respectively and $S$ is the support set of $u_t$ and $u_{t'}$. We now analyze the output distribution shift for different continual learning tasks.

**3D Shape Reconstruction.** In this setting, the output $Y_t^{SDF}$ represents the ground truth SDF values for the support set $S$ consisting of 3D coordinates. We first select 1000 3D points uniformly in a unit grid of resolution $128^3$. For each shape class, we randomly sample 1000 objects. Each 3D point $q_i$ defines a distribution of SDF values within a shape class $P(q_i) = \mathbb{P}(Y_t^{SDF} | q_i, X_t)$. From Eq. 2, the final output distribution distance between each shape class is

$$D(t, t') = \frac{1}{N_q} \sum_{i=1}^{N_q} d(P^{(t)}_{q_i}, P^{(t')}_{q_i})$$

where $N_q$ is the number of 3D points. We present the results for both coordinate representations (OC and 3-DOF VC) described in Sec. 4.1.

**2D Depth Prediction and 2D Silhouette Prediction.** In this setting, $Y_t^{pix}$ represents the value of each pixel of the input $X_t$ (depth value and binary value for depth and silhouette pred. respectively). The support set $S$ is the set of 2D pixel coordinates. Each pixel $p_i$ then defines a distribution of pixel values within a class $P^{(t)}(p_i) = \mathbb{P}(Y_t^{pix} | p_i, X_t)$. The output distribution distance between each class is

$$D(t, t') = \frac{1}{N_p} \sum_{i=1}^{N_p} d(P^{(t)}_{p_i}, P^{(t')}_{p_i})$$

where $N_p$ is the number of pixels. For depth prediction, we first center crop the input images. For each class we randomly sample 800 objects and for each image sample 1000 pixels uniformly.

**Classification.** The output $Y_t$ represents the ground truth class labels of the input $X_t$. The output distribution for each class is then $P^{(t)} = \mathbb{P}(Y_t = c | X_t)$. Different from the reconstruction tasks, the output distribution between each learning exposure does not share a support set. We assume that the class labels are sequentially incremented integers for each new class in the sequence. The final output distribution distance is computed as $D(t, t') = d(P^{(t)}_t, P^{(t')}_t)$

Small output distribution shift is associated with improved knowledge transfer. We first compute the output
7. Proxy Task for Continual Classification

The robustness of representation learning and the ability to transfer knowledge between learning exposures in single-view 3D shape reconstruction begs the question of whether it could be used as a proxy task to improve CL classification [38]. We test this hypothesis here via a simple approach: We train a 3D reconstruction model, SDFNet VC on RGB images continually as in Sec. 4.1, and at inference time we extract the feature from its image encoder with a forward pass. We maintain an exemplar set of 20 images/class with class labels randomly sampled from the training dataset. We do not use the labels for training. Instead, we use the extracted representation to do nearest-class-mean (NCM) classification with the exemplars at testing time. Specifically, the mean feature of each class is first computed from the exemplar set. Then test samples are assigned the label of the closest mean feature via cosine distance. We decide to utilize NCM as a classifier instead of training a fully-connected layer with cross-entropy loss, due to the fact that the exemplar set size is small (< 1% of the training data) and it has been shown that linear classifier trained with CE loss tends to overfit significantly when the dataset is imbalanced [6, 46].

We conduct experiments with ShapeNet13 with one class per exposure. We first show that the feature representation learned by the single-view 3D shape reconstruction task is discriminative despite not having access to ground truth labels during training. We compare the performance of the proxy classifier against an ImageNet pretrained feature representation model. Specifically, we extract the feature from the ImageNet pretrained ResNet18 via a forward pass and use NCM as the classifier with the same exemplar set size as the proxy classifier. Fig. 5 shows evidence that shape features are more beneficial for continual classification than the rich discriminative feature representation from ImageNet. We further compare the proxy classifier against two classification baselines: GDumb [26] and a standard classifier trained continually with cross entropy loss and the same exemplar set, denoted as Classifier with Exemplars. Fig. 5 shows that the 3D shape proxy classifier outperforms the GDumb and Classifier with Exemplars on ShapeNet13. This demonstrates that a significant amount of discriminative information is encoded in the continual shape representation and suggests that it may be beneficial to explore other proxy tasks as a means to improve CL classification.

8. Conclusion

We have identified a set of CL reconstruction tasks, including 3D shape reconstruction, 2.5D sketch estimation, and 2D image reconstruction, that do not exhibit catastrophic forgetting. Hence, the answer to the question we pose in our title “Does continual learning = catastrophic forgetting?” is in general “No,” despite the central role of forgetting in prior research on continual classification. We further show that reconstruction tasks benefit from continuous representation training and exhibit positive forward transfer. In addition, the continual version of the challenging single-view 3D shape reconstruction task demonstrates improvements in performance over time on both seen and novel categories. We are the first to investigate the generalization ability of single-view 3D shape reconstruction models in the context of CL.

We provide a novel algorithm-agnostic means to characterize the knowledge transfer performance of CL tasks via output distribution shift analysis. We show that reduction in shift is associated with increased knowledge transfer. We link reconstruction and classification by showing that feature representations from reconstruction can be effective for CL of classification under a limited exemplar budget. Our
findings point to a need to enlarge the space of CL tasks and develop algorithm-agnostic measures of CL performance.

9. Acknowledgement

We would like to thank Miao Liu and Meera Hahn for the helpful discussion. This work was supported by NIH R01-MH114999 and NSF Award 1936970. This paper is dedicated to the memory of Chengming (Julian) Gu.

References

[1] Plop: Learning without forgetting for continual semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 2
[2] David Abel, Dilip Arumugam, Lucas Lehnert, and Michael Littman. State abstractions for lifelong reinforcement learning. In International Conference on Machine Learning, pages 10–19, 2018. 2
[3] Rahaf Aljundi, Klaas Kelchtermans, and Tinne Tuytelaars. Task-free continual learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 2
[4] Blender Online Community. Blender - a 3D modelling and rendering package. Blender Foundation, Blender Institute, Amsterdam. 12
[5] Zhipeng Cai, Ozan Sener, and Vladlen Koltun. Online continual learning with natural distribution shifts: An empirical study with visual data. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 8281–8290, October 2021. 2
[6] Francisco M Castro, Manuel J Marin-Jimenez, Nicolas Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In Proceedings of the European Conference on Computer Vision (ECCV), pages 233–248, 2018. 8
[7] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for continual learning in semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. 2
[8] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012, 2015. 4
[9] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In European conference on computer vision, pages 628–644. Springer, 2016. 12
[10] Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. arXiv preprint arXiv:1909.08383, 2019. 1
[11] Mohamed Elhoseiny, Francesca Babiloni, Rahaf Aljundi, Marcus Rohrbach, Manohar Paluri, and Tinne Tuytelaars. Exploring the challenges towards lifelong fact learning. In Asian Conference on Computer Vision, pages 66–84. Springer, 2018. 2
[12] Xisen Jin, Junyi Du, and Xiang Ren. Gradient based memory editing for task-free continual learning. arXiv preprint arXiv:2006.15294, 2020. 1
[13] Christos Kaplanis, Murray Shanahan, and Claudia Clopath. Continual reinforcement learning with complex synapses. arXiv preprint arXiv:1802.07239, 2018. 2
[14] Jeremias Knoblauch, Hisham Husain, and Tom Diethe. Optimal continual learning has perfect memory and is np-hard. In International Conference on Machine Learning, pages 5327–5337. PMLR, 2020. 2
[15] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Komer. Evaluation of cnn-based single-image depth estimation methods. In Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pages 0–0, 2018. 5, 13
[16] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. The CIFAR-10 Dataset. online: https://www.cs.toronto.edu/~kriz/cifar.html, 2014. 5
[17] Timothée Lesort, Massimo Caccia, and Irina Rish. Understanding continual learning settings with data distribution drift analysis. arXiv preprint arXiv:2104.01678, 2021. 2
[18] Timothée Lesort, Hugo Caselles-Dupré, Michael Garcia-Ortiz, Andrei Stoian, and David Filliat. Generative models from the perspective of continual learning. In 2019 International Joint Conference on Neural Networks (IJCNN), pages 1–8. IEEE, 2019. 2
[19] Zhizhong Li and Derek Hoiem. Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence, 40(12):2935–2947, 2017. 3
[20] Xiailei Liu, Hao Yang, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Multi-task incremental learning for object detection, 2020. 2
[21] Yaoyao Liu, Yuting Su, An-An Liu, Bernt Schiele, and Qianru Sun. Mnemonics training: Multi-class incremental learning without forgetting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12245–12254, 2020. 1
[22] David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. In Advances in neural information processing systems, pages 6467–6476, 2017. 1, 3
[23] Andrea Maracani, Umberto Michieli, Marco Toldo, and Pietro Zanuttigh. Recall: Replay-based continual learning in semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 7026–7035, October 2021. 2
[24] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 4460–4470, 2019. 4, 12, 13
[25] Umberto Michieli and Pietro Zanuttigh. Incremental learning techniques for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Oct 2019. 2
[26] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In ECCV, 2020. 1, 6, 8, 12, 14

[27] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 652–660, 2017. 12

[28] Michael Ramamonjisoa and Vincent Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, Oct 2019. 5

[29] Dushyant Rao, Francesco Visin, Andrei Rusu, Razvan Pascanu, Yee Whye Teh, and Raia Hadsell. Continual unsupervised representation learning. In Advances in Neural Information Processing Systems, pages 7647–7657, 2019. 3

[30] Anthony Robins. Catastrophic Forgetting, Rehearsal and Pseudorehearsal. Connection Science, 7(2):123–146, 1995. 1

[31] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. International journal of computer vision, 40(2):99–121, 2000. 7

[32] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3):211–252, 2015. 5

[33] Daeyun Shin, Charless C Fowlkes, and Derek Hoiem. Pixels, voxels, and views: A study of shape representations for single view 3d object shape prediction. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 3061–3069, 2018. 2

[34] Konstantin Shmelkov, Cordelia Schmid, and Karteek Alahari. Incremental learning of object detectors without catastrophic forgetting. In Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, pages 8777–8786, 2019. (Oral, Best Paper Finalist). 3, 5

[35] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3405–3419, 2018. 7

[36] Anh Thai, Stefan Stojanova, Vijay Upadhy, and James M Rehg. Incremental object learning from contiguous views. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 8777–8786, 2019. (Oral, Best Paper Finalist). 3, 5

[37] Maxim Tatarchenko, Stefan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3d reconstruction networks learn? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3405–3419, 2019. 4, 13

[38] Anh Thai, Stefan Stojanova, Vijay Upadhy, and James M Rehg. 3d reconstruction of novel object shapes from single images. arXiv preprint arXiv:2006.07752, 2020. 2, 4, 5, 6, 12, 13

[39] Gido M Van de Ven and Andreas S Tolias. Three scenarios for continual learning. arXiv preprint arXiv:1904.07734, 2019. 3, 8

[40] Jianren Wang, Xin Wang, Yue Shang-Guan, and Abhinav Gupta. Wanderlust: Online continual object detection in the real world. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 10829–10838, October 2021. 2

[41] Shuzhe Wang, Zakaria Laskar, Iaroslav Melekhov, Xiaotian Li, and Juho Kannala. Continual learning for image-based camera localization. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 3252–3262, 2021. 2

[42] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004. 5

[43] Ziyun Wang, Volkan Isler, and Daniel D Lee. Surface hof: Surface reconstruction from a single image using higher order function networks. In 2020 IEEE International Conference on Image Processing (ICIP), pages 2666–2670. IEEE, 2020. 5

[44] Chenshen Wu, Luis Herranz, Xiailei Liu, Joost van de Weijer, Bogdan Raducanu, et al. Memory replay gans: Learning to generate new categories without forgetting. In Advances in Neural Information Processing Systems, pages 5962–5972, 2018. 2

[45] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Murrnet: 3d shape reconstruction via 2.5 d sketches. In Advances in neural information processing systems, pages 540–550, 2017. 5, 13, 14

[46] Yue Wu, Yinpeng Chen, Lijuan Wang, Yuancheng Ye, Zicheng Liu, Yandong Guo, and Yun Fu. Large scale incremental learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 374–382, 2019. 8

[47] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pages 3485–3492. IEEE, 2010. 12

[48] Ju Xu and Zhanxing Zhu. Reinforced continual learning. In Advances in Neural Information Processing Systems, pages 899–908, 2018. 2

[49] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. 2019. 4

[50] Zike Yan, Yuxin Tian, Xuesong Shi, Ping Guo, Peng Wang, and Hongbin Zha. Continual neural mapping: Learning an implicit scene representation from sequential observations. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 15782–15792, October 2021. 2

[51] Lu Yu, Bartlomiej Twardowski, Xiailei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. Semantic drift compensation for class-incremental learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. 3
[52] Junting Zhang, Jie Zhang, Shalini Ghosh, Dawei Li, Serafettin Tasci, Larry Heck, Heming Zhang, and C-C Jay Kuo. Class-incremental learning via deep model consolidation. In The IEEE Winter Conference on Applications of Computer Vision, pages 1131–1140, 2020. 1

[53] Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Joshua B Tenenbaum, William T Freeman, and Jiajun Wu. Learning to Reconstruct Shapes from Unseen Classes. In Advances in Neural Information Processing Systems (NeurIPS), 2018. 2, 6
This supplementary material document is structured as follows: In Sec. A we describe the training data in more detail; In Sec. B we provide details on the CL algorithms used in the paper, their training implementation details and evaluation metrics; In Section C we further explain the repeated exposures setting.

A. Datasets

A.1. ShapeNetCore.v2

**Datasets**: ShapeNetCore.v2 consists of 55 categories with 52K CAD models. This is the current largest 3D shape dataset with category labels. Many prior works in 3D shape reconstruction [9, 24] utilized a subset of 13 largest categories—ShapeNet13, which consists of approximately 40K 3D instances. Tbl. 3 lists the 13 categories and the number of samples in each category. For ShapeNet13, we use the standard train/val/test split from prior shape reconstruction works [9, 24]. We sample 100 objects/category from the test split for evaluation in the repeated exposures case. For the remaining 42 classes in ShapeNetCore.v2, we split randomly with proportion 0.7/0.1/0.2 for train/val/test splits. In the single exposure case on all classes of ShapeNetCore.v2, we randomly sample 30 objects/category for testing. For evaluating novel category generalization ability, we sample 50 objects from the 42 classes. The license of ShapeNetCore.v2 is specified in https://shapenet.org/terms.

**Rendering**: We render 25 views of RGB images, ground truth silhouette, depth and surface normal maps with resolution $256 \times 256$ for each object. Following [37], we generate data using Cycles ray-tracing engine in Blender [4] with 3 degree-of-freedom, varying camera azimuth $\theta \in [0, 360^\circ]$, elevation $\phi \in [-50^\circ, 50^\circ]$ and tilt. For experiments with RGB images as inputs, we render with varying light, specular surface reflectance and random backgrounds from SUN Scenes [47].

**SDF Point Sampling Strategy**: For 3D shape reconstruction, training 3D points are sampled more densely close to the surface of the mesh. In the single exposure case on all classes of ShapeNetCore.v2, we sample half of the training points within a distance of 0.03 to the surface, 30% with distance in the range $[0.03, 0.1]$ and 20% in the range $[0.1, 1.1]$. To train and evaluate OccNet, we obtain mesh occupancy values by binary masking $I\{sdf \leq i\}$ where $i$ is the isosurface value.

A.2. CIFAR-100

This is a standard image dataset consisting of 100 categories with 500 training and 100 testing samples for each category. Each image is of resolution $32 \times 32$. In our experiment for classification baselines with repeated exposures, 60 categories are chosen randomly from 100 categories, which we denote as CIFAR-60.

| ID      | Name      | Num samples |
|---------|-----------|-------------|
| 02691156| airplane  | 4045        |
| 02828884| bench     | 1813        |
| 02933112| cabinet   | 1571        |
| 02958343| car       | 3532        |
| 03001627| chair     | 6778        |
| 03211117| display   | 1093        |
| 03636649| lamp      | 2318        |
| 03691459| loudspeaker| 1597       |
| 04090263| rifle     | 2373        |
| 04256520| sofa      | 3173        |
| 04379243| table     | 8436        |
| 04401088| telephone | 1089        |
| 04530566| watercraft| 1939        |

**Total**: 39,757

Table 3. Statistics of ShapeNet13.

B. Description of Algorithms

B.1. Single Object 3D Shape Reconstruction (Secs. 4,5)

**Architecture**: We adapt SDFNet [37] and OccNet [24] with ResNet-18 encoder for continual training with 2D and 2.5D inputs and SDFNet with PointNet [27] encoder for 3D input. Specifically, the architecture consists of an encoder initialized with random weights and a point module which are multiple blocks of fully-connected layers with ReLU activation. Conditional Batch Normalization is used as applying an affine transformation on the output of the point module, conditioned on the feature vector produced by the encoder.

**GDumb For CL 3D Shape**. We employ SDFNet with ResNet-18 encoder as the backbone architecture and follow the training procedure of GDumb for classification task [26]. Specifically, we randomly select an exemplar set of size $K = 1000$ ($\approx 3.7\%$ of the training data), equally divided for all the seen categories at each learning exposure. We initialize the learning model randomly to train from scratch on the selected exemplar set at each learning exposure.

**GSmart**. Different from GDumb for CL 3D shape, we continuously update the representation at each learning exposure. Please see Algs. 1,2,3 for the pseudo code of CL algorithms evaluated in Sec. 5 in the main text.

**Loss function**: SDFNet uses $L_1$ loss as the loss function, with high weights for points close to the surface. Specifically,

$$L(s, \hat{s}) = \begin{cases} 
|s - \hat{s}|, & \text{if } |s| > 0.01 \\
4|s - \hat{s}|, & \text{otherwise}
\end{cases}$$

where $s$ is the ground truth SDF value and $\hat{s}$ is the predicted SDF value.
archically determines the voxels that contain the mesh to
in the cube, MISE starts from a lower resolution and hier-
SDF/occupancy values for all the points uniformly sampled
generate the predicted mesh. Instead of generating the
archically extracts the mesh isosurface introduced by [24]

Exemplar set: \( \theta \)

\[
C \leftarrow \text{SELECT_RANDOM}(C \cup D_{\text{train}}^t)
\]

\( \theta_t, \text{acc}_t \leftarrow \text{SDFNet}(\theta, C, D_{\text{val}}^t) \)

\text{Result:} (\text{acc}_1, \text{acc}_2, \ldots, \text{acc}_T)

Algorithm 3: C-SDFNet

\text{Input:} Batch training procedure
\text{SDFNet}(\theta, D_{\text{train}}^t, D_{\text{val}}^t) that returns the
trained parameters \( \theta \) and the performance of the trained model on \( D_{\text{val}}^t \)

\text{Data:} (RGB image, 3D coordinates, SDF values)
pair datasets \( D_{\text{train}}^t = \bigcup_{i=1}^T D_{\text{train}}^i \),
\( D_{\text{val}}^t = \bigcup_{i=1}^T D_{\text{val}}^i \)

\text{Define:} \( \ell : \text{weighted } L_1 \text{ loss} \)

1 \text{ init}
2 \quad \text{Exemplar set: } C = \{ \}
3 \text{ foreach learning exposure } t \text{ in } 1, 2, \ldots, T \text{ do}
4 \quad \theta \leftarrow \text{RANDOM_INIT}(\theta)
5 \quad C \leftarrow \text{SELECT_RANDOM}(C \cup D_{\text{train}}^t)
6 \quad \theta_t, \text{acc}_t \leftarrow \text{SDFNet}(\theta, C, D_{\text{val}}^t)
7 \text{ end}
8 \text{ Result:} (\text{acc}_1, \text{acc}_2, \ldots, \text{acc}_T)

MISE to work on both SDF and occupancy values.

\text{Metric: Following [36, 37], we use F-Score at 1\% as our main evaluation metric. We first sample 300K and 100K points respectively on the surface of the predicted mesh (S_1) and ground truth mesh (S_2). The metric is computed as the following}

\[
FS_{@1} = \frac{2 \cdot \text{prec}_{@1} \cdot \text{rec}_{@1}}{\text{prec}_{@1} + \text{rec}_{@1}}
\]

where \( \text{prec}_{@1} \) is the precision at 1\%, which measures the portion of points from \( S_1 \) that lie within a threshold 0.01 to the points from \( S_2 \) (in the case where the mesh is normalized to fit in a unit cube) and \( \text{rec}_{@1} \) is the recall at 1\%, which measures the portion of points from \( S_2 \) that lie within a threshold 0.01 to the points from \( S_1 \).

B.2. Single-view 2.5D Sketches Prediction (Sec. 4)

\text{Architecture: We adapt the 2.5D sketch estimation from MarrNet [45] for continual training. The backbone architecture for MarrNet is a U-ResNet18 with the ResNet18 image encoder initialized with ILSVRC-2014 pre-trained weights.}

\text{Loss functions: We use MSE as the loss function for depth and normals prediction. Specifically,}

\[
MSE(I, \hat{I}) = \frac{1}{K} \sum_{i,j}^K || I(i, j) - \hat{I}(i, j) ||_2^2
\]

where \( I \) and \( \hat{I} \) are the ground truth and predicted images respectively.

\text{Metrics: For depth prediction, we report threshold accuracy: percentage of } y_i \text{ such that}

\[
\max \left( \frac{y_i}{y_i^*} \frac{y_i^*}{y_i} \right) < \sigma
\]

where \( y_i \) and \( y_i^* \) are the predicted and ground truth depth values at pixel \( i \) and \( \sigma \) is the threshold. In our evaluation, we use \( \sigma = 1.25 \) as in [15].
For normals, we report cosine distance threshold as the main metric. We first convert the RGB values of the normal map into 3D vectors

\[ n = 2 \left( \begin{bmatrix} c - 0.5 \\ 0.5 \\ 0.5 \end{bmatrix} \right) \]

where \( n \) and \( c = \begin{bmatrix} r \\ g \\ b \end{bmatrix} \) are the normal and color vectors respectively. Cosine distance threshold accuracy is then computed as

\[ \langle \frac{n}{\|n\|_2}, \frac{n^*}{\|n^*\|_2} \rangle > \sigma \]

where \( n \) and \( n^* \) are predicted and ground truth normals. We set \( \sigma = 0.9 \).

**B.3.2D Reconstruction (Sec. 4)**

**B.3.1 Silhouette Prediction**

We use MarrNet [45] as the backbone architecture for continual training with BCE loss function. We report Intersection-over-Union for silhouette prediction as the metric. Specifically,

\[ IoU(I, \hat{I}) = \frac{|I \cap \hat{I}|}{|I \cup \hat{I}|} \]

**B.3.2 Image Autoencoding**

**Architecture:** We implement a shallow network with 4 conv. layers, each followed by a max pooling layer which we termed ConvAutoEncoder. Each conv. layer has 16 channels and the dimension of the bottle-neck feature vector is \( 16 \times 2 \times 2 \). The network is randomly initialized.

**Loss function:** We train ConvAutoEncoder with MSE loss for each pixel, defined as

\[ \mathcal{L}(I, \hat{I}) = \frac{1}{K \times K \times 3} \sum_{i=1}^{K} \sum_{j=1}^{K} \sum_{c=1}^{3} \|I(i,j,c) - \hat{I}(i,j,c)\|_2^2 \]

where \( K \) is the size of the input image and \( c = \{1, 2, 3\} \) is the 3 input channels (red, green, blue).

**Metric:** We use SSIM scaled to range \([0, 1]\) as the main evaluation metric for the image autoencoding experiment. Specifically, given two image windows \( x \) and \( y \) of the same size \( N \times N \) the original SSIM metric is computed as

\[ SSIM(x, y) = \frac{(2\mu_x \mu_y + c_1)(2\sigma_{xy} + c_2)}{\mu_x^2 + \mu_y^2 + c_1(\sigma_x^2 + \sigma_y^2 + c_2)} \]

with \( \mu_x, \mu_y \) be the averages of \( x \) and \( y \) respectively, \( \sigma_x^2, \sigma_y^2, \sigma_{xy} \) are the variances of \( x \) and \( y \) respectively, \( c_1, c_2 \) are constants to avoid dividing by 0 in the denominator.

**B.4. Classification Baselines (Sec. 7)**

**GDumb** [26] is an algorithm that randomly selects exemplars and performs training on the exemplar set only. At each learning exposure, the model is trained from scratch on the exemplar set, in which each category is represented with the same number of samples. GDumb utilizes the standard cross-entropy loss and classifies using the network outputs. We used our PyTorch implementation of GDumb with ResNet18 initialized randomly as the feature extractor.

**Classifier with Exemplars** is a simple baseline where we train a standard classifier with cross-entropy loss continually. At each learning exposure, the learning model is trained on the current training data combined with the randomly selected exemplar set without any further heuristics. Similar to GDumb, we use randomly initialized ResNet18 as the feature extractor.

**ImageNet Pretrained** is the baseline we use to highlight that the feature space learned by CL single-view 3D shape model from RGB image without ground truth label is discriminative. For each new class, we randomly select the exemplar set from the training data. At test time, we first extract the feature representation from the ILSVRC-2014 pretrained ResNet18 for each test sample. We then perform NCM to predict the label using the exemplar set.

**C. Further Explanation for Repeated Exposures Setting**

In the repeated exposure setting, each class occurs a fixed number of times (e.g. 10 repetitions) in random order. For example, in the case of 50 classes repeated 10 times, we would first generate 500 learning exposures, and then perform a random permutation to obtain the order seen by the learner. As a result, classes repeat in complex and highly-variable patterns. Note that even though classes repeat, each learning exposure still contains only a single class (or a small number), thereby preserving the domain shift between exposures that makes CL challenging.