The Surprising Positive Knowledge Transfer in Continual 3D Object Shape Reconstruction

Anh Thai, Stefan Stojanov, Zixuan Huang, James M. Rehg
Georgia Institute of Technology
{athai6, sstojanov, zixuanh, rehg}@gatech.edu

Abstract

Continual learning has been extensively studied for classification tasks with methods developed to primarily avoid catastrophic forgetting, a phenomenon where earlier learned concepts are forgotten at the expense of more recent samples. In this work, we present a set of continual 3D object shape reconstruction tasks, including complete 3D shape reconstruction from different input modalities, as well as visible surface (2.5D) reconstruction which, surprisingly demonstrate positive knowledge (backward and forward) transfer when training with solely standard SGD and without additional heuristics. We provide evidence that continually updated representation learning of single-view 3D shape reconstruction improves the performance on learned and novel categories over time. We provide a 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 pretrained models released with this article can be found at https://github.com/rehg-lab/CLRec

1. Introduction

Various applications in domains like AR/VR, autonomous driving, and robotics where 3D reconstruction is essential require continually learning and processing streams of input data. For instance, a home robot assistant learns about the newly installed bathtub after being familiar with known household objects like chairs and tables. While many important properties of 3D object shape reconstruction methods such as generalization ability and large-scale batch training have been studied extensively in prior works [34, 23, 36, 46], the feasibility of this task in a continual learning setting has not been investigated.

The goal of continual learning (CL) is to train models incrementally to solve a sequence of tasks without access to past data. The learner receives a sequence of learning exposures,1 each containing a subset of the overall data distribution and comprising a task (e.g., in image classification a learning exposure might contain samples from two ImageNet classes.) Note that this setting is in stark contrast to the batch training setting where the model is optimized upon observing the entire training data distribution. The fundamental challenge of CL is backward and forward knowledge transfer [21]. Backward transfer (BWT) refers to the effectiveness of the current representation in solving previously-learned tasks. Large negative BWT results in catastrophic forgetting, the phenomenon where the representations learned in previous learning exposures degrade significantly over time at the expense of more recent data. For example, learning classification on 10 tasks with 20 classes/task sequentially on Tiny-ImageNet [30] with solely vanilla SGD training leads to only 7.92% average accuracy at the end, when tested on all classes. On the contrary, batch training obtains 60% [4]. Tackling catastrophic forgetting has been attempted by a large number of prior works [9, 49, 20, 24] by employing multiple complex train-
ing heuristics and has come to characterize continual learning for many different tasks (e.g., classification, segmentation, detection, etc.) Also important is forward transfer (FWT), which refers to the utility of the learned representation for unseen future tasks. Positive FWT enables CL methods to leverage shared representations across tasks, so that training on new tasks is more effective than training from scratch. Past works have largely focused on classification tasks [28, 6, 21, 25], with a few exceptions [5, 47]. A common theme of these efforts is the difficulty of avoiding negative BTW and achieving positive FWT. Please see Fig. 1 for an illustration of the standard CL setting.

In this work, we demonstrate that continual object shape reconstruction tasks exhibit surprisingly effective knowledge transfer using standard deep architectures and vanilla SGD, without any of the special losses, exemplars, or other approaches routinely used in CL to overcome forgetting. This is illustrated in Fig. 2 for the challenging task of single-view 3D shape reconstruction [34, 36, 46], in which the learner predict the 3D shape of an object given a single input image. Each learning exposure contains samples from a subset of object classes, and we test generalization to both seen and unseen classes of objects. Fig. 2a illustrates the BTW performance of our CL reconstruction approach. The shape reconstructions rendered in the second column were produced after the model received its first (and only) learning exposure containing that object class, resulting in good reconstruction performance. In contrast, the reconstructions in the third column were obtained at the end of CL after all learning exposures had been introduced. Note that the model received only one exposure to each object class. Surprisingly, the quality of the reconstruction produced by the final model slightly improves relative to the first exposure, which is evidence for the lack of negative backward transfer. Fig. 2b illustrates FWT performance. While the model was never trained on these unseen classes, the quality of the 3D reconstructions improves steadily as learning progresses, proving strong and surprising evidence for positive FWT and the ability to leverage a shared representation between tasks using only fine-tuning via vanilla SGD. We believe that our novel findings provide crucial insights into the feasibility of systems that require continual learning of object shape.

In summary, this paper makes the following contributions: 1) Formulation of continual object shape reconstruction tasks (Tbl. 1), including complete 3D shape reconstruction from different input modalities and visible 3D surface (2.5D) reconstruction (Sec. 3); 2) The surprising finding that these tasks exhibit lack of negative backward transfer and catastrophic forgetting (Sec. 4); 3) Evidence for improved generalization ability which is indicative of positive forward transfer (Sec. 5); 4) Novel output distribution shift measurement which demonstrates that smaller output distribution shift across learning exposures leads to better knowledge transfer in continual learning (Sec. 6); 5) Using single-view 3D shape reconstruction as a proxy task for classification is effective 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 (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 [24, 7, 22, 1] explored image segmentation. Some works [32, 19, 39] investigated incremental object detection while [17, 42] learned image generation. Elhoseiny et al. [10] examined continual fact learning by utilizing a visual-semantic embedding. Wang et al. [40] studied CL of camera localization given an input RGB image while [5] explored online CL of geolocalization with natural distribution shift in the input that occurs over real time. Others [2, 13, 45] focused on reinforcement learning.

Most closely related to our work is Yan et al. [47] 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

| Input Rep. | Output Rep. | Reconstruction Tasks |
|-----------|-------------|----------------------|
| 2D → 3D  | Single-view Image 3D Shape Rec. (Figs. 3a, 3b) |
| 2.5D → 3D | Single-view Surface Normals Pred. (Fig. 3d) |
| 3D → 3D  | Single-Object Pointcloud 3D Shape Rec. (Fig. 3e) |
| 2D → 2.5D| Single-view Depth Pred. (Fig. 3f) |
| 2D → 2D  | Image auto-encoding (Sup.) | Single-view Silhouette Pred. (Sup.) |

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.

2While there is nothing inherently categorical about shape reconstruction, categories are routinely-used to identify sets of similar shapes for training and evaluation purposes, e.g. in testing generalization to unseen categories of shapes [34, 36, 50].
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 semantics in the inputs change 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, 38]. While [38] examined the benefits and drawbacks of rehearsal methods in CL, [14] showed that optimal CL algorithms solve an NP-HARD problem and require the ability to approximate the parameters that optimize all seen tasks. While [16] discussed the different concept drifts 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 [36, 50, 31]. We are the first to provide generalization analysis of these models in the CL setting, utilizing the 3-DOF VC approach which was 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. [48] 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. [27] 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

Continual Learning of Reconstruction. 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 \( t \in \{1, 2, \ldots, T\} \). For example, single-view 3D shape reconstruction aims to output the 3D shape of the object represented in the input image. The model 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 [33] 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\(^3\). Note that in this work, we assume that each \( D_t \) is defined over a set of \( M_t \) visual categories.\(^4\)

Training. During training, the learning model does not have access to previously seen data \( D_{t-1} \). We optimize the parameters \( \theta_t \) of the function \( f \) continuously at each learning exposure upon observing only 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 refer 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 [18].

Evaluation. At test time we consider the following metrics at each learning exposure: 1) \( \text{Acc}^+_t \): accuracy on all

\(^3\)Details discussed in the Supp.

\(^4\)Organizing shapes into categories allows us to characterize how new shape concepts are introduced during learning.
Figure 3: (a) Performance of shape reconstruction methods with 2D and 2.5D inputs when presented with a single exposure for each category from all 55 categories of ShapeNetCore.v2, 5 classes/exposure (b) repeated exposures case on ShapeNet13 with 10 repeated exposures, 2 classes/exposure (c) single exposure case on ShapeNetCore.v2 of shape methods with 3D inputs. (d) Results for 2.5D estimation. Performance in terms of thresholding accuracy ($\delta = 1.25$) for depth prediction and thresholding cosine distance ($\delta = 0.9$) for surface normals. Backward transfer is reported in parenthesis. Catastrophic forgetting does not happen to any of the algorithms in any case.

known categories (Secs. 4, 7) and 2) $Acc^2$: accuracy on a fixed, held out set of unseen classes that are never explicitly learned (Sec. 5). Plotting the average accuracy at all learning exposures results in the learning curve of the CL model. All accuracy metrics reported are in range $[0, 1]$.

We further report backward and forward transfer metrics [21] 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.

4. Single Object Shape Reconstruction Does Not Suffer from Catastrophic Forgetting

Tbl. 1 lists the five types of reconstruction tasks that we evaluate in this work, which include 3D, 2.5D, and 2D output domains. Our key finding is that CL tasks from each of these five types do not suffer from catastrophic forgetting. It is important to emphasize that the “continual learning” algorithm specified in Sec. 3, that is known to perform very poorly for classification tasks. 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 and 2.5D prediction are presented in Secs. 4.1 and 4.2 respectively. We additionally conduct experiments on 2D reconstruction tasks in the Sup. In Sec. 4.3 we present two possible simple explanations for the lack of catastrophic forgetting and provide empirical evidence that rejects these hypotheses. We report $Acc^2$ 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, given a single image or sparse pointcloud as the input, the goal of the desired function $f$ is to produce a 3D surface representation of the object present in the input. We focus our analysis on signed-distance-fields (SDF) since it was identified to achieve superior performance in the batch setting [36, 46]. The SDF value of a point in 3D space indicates the distance to the closest surface from that point, with the sign encoding whether the point is inside (negative) or outside (positive) of
the watertight object surface. Thus, the 3D surface is represented as a zero-level set where all the points lying on the surface of the object have SDF value 0.

**Approach.** We utilize SDFNet \[36\] and OccNet\[^5\] \[23\] as backbone architectures for CL with 2D and 2.5D input representations where inputs are single-view RGB images and ground truth depth and normal maps respectively. We train both methods with the 3-DOF VC representation (varying in azimuth, elevation and camera tilt) from \[36\], which was shown to give the best generalization performance.\[^6\] We also train with object-centered (OC) representation for SDF representation, in which the model is trained to output the shape in the canonical pose. For 3D input representations where inputs are sparse 3D pointclouds, we further examine a variant of ConvOccNet \[12\] that outputs SDFs instead of continuous occupancies (ConvSDFNet). In the Supp, we additionally show results on a standard pointcloud autoencoder following in \[11\].

**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.\[^7\] Following prior works in shape reconstruction \[36, 46, 34\] we report the average FS@1 at each learning exposure. We use SDFNet as the batch reference for 2D and 2.5D inputs. For 3D inputs we include ConvSDFNet batch performance. All models are trained from random initialization.

**Results.** The results are shown in Figs. 3a, 3b and 3c 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. 3a), all algorithms maintain their accuracy over time and even exhibit a slight upward trend of increasing accuracy while for 3D inputs (Fig. 3c) the performance increases more consistently 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 negative 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 object class. Since our findings hold on various model architectures with different input/output representations, this 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. 3b), the performance of both SDFNet and OccNet when trained with 3-DOF VC improves significantly over time, and eventually performs on par with batch.\[^8\] These models achieve significant positive BWT which indicates that catastrophic forgetting is mitigated. Unlike the experiments in \[33\], 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 \[36\] that training with 3-DOF VC results in a more robust feature representation.

### 4.2. Single-view 2.5D Sketch 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 \[43\] architecture, with an ILSVRC-2014 \[30\] pre-trained ResNet18 for the image encoder. We evaluate depth prediction using the commonly used thresholding accuracy \[15, 26\]. For normals prediction, we report the accuracy based on the cosine distance threshold between the predicted and ground truth surface normals \[41\].\[^9\] Fig. 3d 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. While the performance of some CL models learned with single exposure when all data has been seen does not reach batch (for 2D→3D, 2.5D→3D, and 2D→2.5D tasks), we note that these tasks are sufficiently challenging (even in the batch setting where data is iid) and emphasize that the surprising positive trend of the curves has never been shown in prior CL works.

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. We refer to the Supp. for details.

### 4.3. Discussion of CL Object Shape Reconstruction

We have identified (for the first time) a set of continual object shape reconstruction tasks that do not suffer from catastrophic forgetting (see Fig. 3) when models are trained\[^8\]. The 65 learning exposures (x axis) in Fig. 3b result from 13 ShapeNet classes divided by 2 classes per exposure with 10 repetitions each.\[^9\]

---

\[^5\]OccNet utilizes continuous occupancies as the 3D representation.

\[^6\]3-DOF VC SDFNet is the current SOTA for single image 3D shape reconstruction.

\[^7\]For instance, if chair and table are present in the current learning exposure, the model will be trained on all chairs and tables in the respective training splits.

\[^8\]The 65 learning exposures (x axis) in Fig. 3b result from 13 ShapeNet classes divided by 2 classes per exposure with 10 repetitions each.

\[^9\]More details in Supp.
using standard SGD without any heuristics. A key question is why this is happening. We examine two possible simple explanations for the CL performance of single-view 3D shape reconstruction: 1) The learner encodes “low-level” features of the inputs that are present for all object classes and facilitate easy generalization, and 2) the domain shift between consecutive learning exposures is small, making the CL problem almost trivial. We find that neither of these hypotheses is supported by our findings, suggesting that the behavior we have discovered is nontrivial, which can motivate for future research and investigation.

Low-level Features. Are there some low-level visual properties shared by all 3D shapes that the learner can index on to solve CL? This seems implausible, as single image reconstruction is a challenging task that requires learning mid- to high-level properties of classes of shapes (e.g., concavities in bowls and tubs, protrusions in chairs and tables) in order to learn to reconstruct occluded surfaces. Since shape reconstruction losses penalize the entire 3D shape (including occluded surfaces), good performance on unseen classes requires nontrivial generalization. We also demonstrate that learned shape representations encode categorical information: We fit a linear classifier on top of the shape features extracted from SDFNet trained on ShapeNetCore.v2 (all 55 classes) and we find that it obtains 65% accuracy, compared to 16% for random features and 42% for ImageNet pre-trained features. This shows that the learner is encoding complex properties of 3D shape in solving the task.

Domain Shift. In Fig. 4, we present quantitative evidence that continual shape reconstruction is characterized by significant class-based domain shift: The per-class reconstruction performance for three representative classes is low before each class is learned (introduced in the training data) and then rises significantly after. It’s clear that the learned representation is responding to the properties of each class, and yet there is very little forgetting. We present additional analysis of domain shift in Sec. 6, to shed further light on this phenomenon. In summary, we argue that CL object shape reconstruction is solving a nontrivial task which requires a complex generalization ability, and therefore merits further investigations in future work using the framework we have provided.

5. Generalization of CL 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 (FWT). We focus our analysis on the challenging problem of single-view 3D shape reconstruction. 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 [50, 36], we are the first to analyze this behavior in a continual learning setting. In this section, we report Accg.

We conduct our experiments on ShapeNet13 with single exposure and 1 shape class per learning exposure on continual SDFNet (C-SDFNet) (Sec. 4.1). We evaluate C-SDFNet on a held out set of 42 classes of ShapeNetCore.v2 with 50 instances for each category (Fig. 5). The model performs 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 improves over time as more classes are learned. This illustrates benefit of continuous representation learning as a useful feature that aids generalization and improves the performance on novel classes over time. In Fig. 2 we show qualitative results that demonstrate positive knowledge transfer ability of single-view 3D shape reconstruction task.

In the Supp. we provide further evidence that continuous representation training is beneficial for CL of single-image 3D shape reconstruction by comparing with an episodic training approach that was shown to achieve competitive performance in CL classification. We additionally present a simple yet competitive CL classification baseline that employs continuous representation update strategy in [35].

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 expen-
Figure 6: Visualization of our output distribution shift measurement. We use the first Wasserstein distance metric (EMD) which measures the least amount of required work to transform one distribution into another to obtain the output distribution shift between different exposures $t$ and $t'$. 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 (see details in Supp.). We first state the hypothesis connecting the output distribution to CL task knowledge transfer ability.

**Hypothesis:** When the distance of the output distribution between each learning exposure becomes smaller, backward and forward transfer increase for any CL method.

We now present the intuition behind our formulation.

Let $D$ be some dataset consisting of two parts $D_1$ and $D_2$ that are independently generated. During batch training we optimize the parameters $\theta \in \Gamma$ where $\Gamma$ is the model parameter space by minimizing the negative likelihood. Since $D = D_1 \cup D_2$ and $D_1$ and $D_2$ are independent, the negative likelihood reduces to $-\log p(D_1|\theta) - \log p(D_2|\theta) = -\log p(Y_1|X_1, \theta) - \log p(Y_2|X_2, \theta)$ where $X_i \sim X$ and $Y_i \sim Y$ are the inputs and outputs respectively. During continual learning when $D_1$ and $D_2$ 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 where $\theta_1, \theta_2 \in \Gamma$ are model parameters, which leads to a suboptimal solution for $L(\theta)$. When the distance between the conditional distributions $Y_i|X_i$ 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.** We now demonstrate the empirical evidence for the earlier hypothesis. Note that in all of the following analyses, the input $X_i$ 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. [29] 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 $d(u, v) = \inf_{\pi \in \Pi(u, v)} \int_{x \times y} |x - y|d\pi(x, y)$ and express the distance between two learning exposures $t$ and $t'$ as

$$D(t, t') = \frac{1}{|\mathcal{Y}|} \int_{s \in \mathcal{S}} d(u_t, u_{t'})ds$$

where $u_t$ and $u_{t'}$ are the output distributions at exposures $t$ and $t'$ respectively and $\mathcal{S}$ is the support set of $u_t$ and $u_{t'}$ (please see Fig. 6 for a visual illustration). We now analyze the output distribution shift for different CL tasks. Note that we normalize the distribution shift by the range of the output values so that they are defined over a support set of the same length.

**3D Shape Reconstruction.** In this setting, the output $Y_{t, t'}^{\text{SDF}}$ represents the ground truth SDF values for the support set $\mathcal{S}$ 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 $P_{q_i}^{(t)} = \mathbb{P}(Y_{t, t'}^{\text{SDF}}|q_i, X_i)$. From Eq. 1, the final output distribution distance between each shape class is $D(t, t') = \frac{1}{N_q} \sum_{i=1}^{N_q} d(P_{q_i}^{(t)}, P_{q_i}^{(t')})$ where $N_q$ is the number of 3D points. We present the results for both OC and 3-DOF VC representations described in Sec. 4.1.

**2.5D Depth Prediction and 2D Silhouette Prediction.** In this setting, $Y_{t, t'}^{\text{pix}}$ represents the value of each pixel of the input $X_i$ (depth value and binary value for depth and silhouette pred. respectively). The support set $\mathcal{S}$ is the set of 2D pixel coordinates. Each pixel $p_i$ then defines a distribution of pixel values within a class $P_{p_i}^{(t)} = \mathbb{P}(Y_{t, t'}^{\text{pix}}|p_i, X_i)$. The output distribution distance between each class is $D(t, t') = \frac{1}{N_p} \sum_{i=1}^{N_p} d(P_{p_i}^{(t)}, P_{p_i}^{(t')})$ 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.

We first compute the output distribution distance as described above for each task and compare it with the resulting BWT and FWT. To verify the effectiveness of the proposed method and to ensure fairness we continually train each task using the fine-tuning strategy on ShapeNet13 from 2D RGB input images with 1 class per learning exposure and report the average output distribution distance and the BWT and

| Task                  | Normalized Mean Dist. | BWT ↑ | FWT ↑ |
|-----------------------|-----------------------|-------|-------|
| Sil Pred.             | 0.075                 | -0.003| 0.836 |
| VC 3D Shape Rec.      | 0.076                 | -0.123| 0.105 |
| Depth Pred.           | 0.084                 | -0.136| 0.094 |
| OC 3D Shape Rec.      | 0.105                 | -0.220| 0.090 |
Figure 7: (a) Our approach to CL classification using reconstruction as a proxy task: We extract the feature representations of the exemplars and test data via a forward pass on the trained 3D reconstruction model. Classification is done via Nearest Class Mean. (b) CL performance (Proxy Rep Learning) is shown for ShapeNet13 with RGB input. Given a limited exemplar budget, we outperform ImageNet pretrained features and classification baselines.

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 for class-IL classification [37]. We test that 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 (Fig. 7a). 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 [44, 6].

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. 7b 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 [25] and a standard classifier trained continually with cross entropy loss and the same exemplar set, denoted as Classifier with Exemplars. Fig. 7b 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. Note that our goal in this section is to show that the unsupervised pretrained shape features give surprisingly high performance despite not being trained to perform classification or use any heuristics. Therefore, we do not compare our approach extensively to existing SOTA CL classification methods and do not attempt to make SOTA claims against these methods.

8. Conclusion

We have identified that CL object 3D shape reconstruction from various modalities exhibit lack of negative backward transfer. In addition, we show that the challenging single-view 3D shape reconstruction task exhibits positive knowledge transfer by investigating the generalization ability of single-view 3D shape reconstruction models in the context of CL for the first time. As a means to characterize the knowledge transfer performance of CL tasks, we provide a novel algorithm-agnostic approach that analyzes output distribution shift. We show that reduction in shift is associated with increased knowledge transfer. We further demonstrate that single-view 3D shape reconstruction task can serve as a promising proxy task for CL classification. We hope that our findings will encourage the community to investigate the intriguing phenomenon observed in CL object shape reconstruction tasks.

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.
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] Pietro Buzzega, Matteo Boschin, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, Advances in Neural Information Processing Systems, volume 33, pages 15920–15930. Curran Associates, Inc., 2020. 1
[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. 2, 8
[7] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental 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. 5
[9] 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
[10] 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
[11] Achiropetas et al. Learning representations and generative models for 3d point clouds. In International conference on machine learning, pages 40–49. PMLR, 2018. 5
[12] Peng et al. Convolutional occupancy networks. In European Conference on Computer Vision, pages 523–540. Springer, 2020. 5
[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 Huisan, and Tom Diethe. Optimal continual learning has perfect memory and is np-hard. In International Conference on Machine Learning, pages 5327–5337. PMLR, 2020. 3
[15] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. 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
[16] Timothée Lesort, Massimo Caccia, and Irina Rish. Understanding continual learning settings with data distribution drift analysis. arXiv preprint arXiv:2104.01678, 2021. 3
[17] 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
[18] Zhizhong Li and Derek Hoiem. Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence, 40(12):2935–2947, 2017. 3
[19] Xialei Liu, Hao Yang, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto. Multi-task incremental learning for object detection, 2020. 2
[20] Yao Yao 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
[21] 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, 2, 4
[22] 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
[23] 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. 1, 5
[24] 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. 1, 2
[25] Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In ECCV, 2020. 2, 8
[26] 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
[27] 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
[28] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental Classifier and Representation Learning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5533–5542, July 2017.

[29] 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.

[30] 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.

[31] 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.

[32] 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, pages 3400–3409, 2017.

[33] Stefan Stojanov, Samarth Mishra, Ngoc Anh Thai, Nikhil Dhandha, Ahmad Humayun, Chen Yu, Linda B Smith, 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).

[34] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhunwei 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 3405–3414, 2019.

[35] Anh Thai, Stefan Stojanov, Zixuan Huang, Isaac Rehg, and James M Rehg. The surprising positive knowledge transfer in continual 3d object shape reconstruction. arXiv preprint arXiv:2101.07295, 2021.

[36] Anh Thai, Stefan Stojanov, Vijay Upadhyia, and James M Rehg. 3d reconstruction of novel object shapes from single images. arXiv preprint arXiv:2006.07752, 2020.

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

[38] Eli Verwimp, Matthias De Lange, and Tinne Tuytelaars. Rehearsal revealed: The limits and merits of revisiting samples in continual learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 9385–9394, October 2021.

[39] 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.

[40] 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.

[41] 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.

[42] Chenshen Wu, Luis Herranz, Xiaolei 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.

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

[44] 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.

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

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

[47] 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.

[48] Lu Yu, Bartlomiej Twardowski, Xiaolei 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.

[49] Jinming 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.

[50] Xiuming Zhang, Zhoutong Zhang, Shuzhe Wang, Donglei 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.

[51] Xiaoming 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.