



Article

Unified Probabilistic Deep Continual Learning through Generative Replay and Open Set Recognition

Martin Mundt ^{1,*}, Iuliia Pliushch ¹, Sagnik Majumder ², Yongwon Hong ³ and Visvanathan Ramesh ¹

¹ Department of Computer Science and Mathematics, Goethe University, 60323 Frankfurt am Main, Germany; pliushch@em.uni-frankfurt.de (I.P.); vramesh@em.uni-frankfurt.de (V.R.)

² Department of Computer Science, The University of Texas at Austin, Austin, TX 78712, USA; sagnik@cs.utexas.edu

³ Department of Computer Science, Yonsei University, Seoul 03722, Korea; yhong@yonsei.ac.kr

* Correspondence: martin.mundt@tu-darmstadt.de

† Current affiliation: Department of Computer Science, TU Darmstadt, 64289 Darmstadt, Germany.

Abstract: Modern deep neural networks are well known to be brittle in the face of unknown data instances and recognition of the latter remains a challenge. Although it is inevitable for continual-learning systems to encounter such unseen concepts, the corresponding literature appears to nonetheless focus primarily on alleviating catastrophic interference with learned representations. In this work, we introduce a probabilistic approach that connects these perspectives based on variational inference in a single deep autoencoder model. Specifically, we propose to bound the approximate posterior by fitting regions of high density on the basis of correctly classified data points. These bounds are shown to serve a dual purpose: unseen unknown out-of-distribution data can be distinguished from already trained known tasks towards robust application. Simultaneously, to retain already acquired knowledge, a generative replay process can be narrowed to strictly in-distribution samples, in order to significantly alleviate catastrophic interference.

Keywords: continual deep learning; catastrophic forgetting; open-set recognition; variational inference; deep generative models



Citation: Mundt, M.; Pliushch, I.; Majumder, S.; Hong, Y.; Ramesh, V. Unified Probabilistic Deep Continual Learning through Generative Replay and Open Set Recognition. *J. Imaging* **2022**, *8*, 93. <https://doi.org/10.3390/jimaging8040093>

Academic Editors: Simone Palazzo and Carmelo Pino

Received: 23 January 2022

Accepted: 26 March 2022

Published: 31 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Consider an empirically optimized deep neural network for a particular task, for the sake of simplicity, say the classification of dogs and cats. Typically, such a system is trained in a closed world setting [1] according to an isolated learning paradigm [2]. That is, we assume the observable world to consist of a finite set of known instances of dogs and cats, where training and evaluation is limited to the same underlying statistical data population. The training process is treated in isolation, i.e., the model parameters are inferred from the entire existing dataset at all times. However, the real world requires dealing with sequentially arriving tasks and data originating from potentially unknown sources.

In particular, should we wish to apply and extend the system to an open world, where several other animals (and non animals) exist, there are two critical questions: (a) How can we prevent obvious mispredictions if the system encounters a new class? (b) How can we continue to incorporate this new concept into our present system without full retraining? With respect to the former question, it is well known that neural networks yield overconfident mispredictions in the face of unseen unknown concepts [3], a realization that has recently resurfaced in the context of various deep neural networks [4–6]. With respect to the latter question, it is similarly well known that neural networks, which are trained exclusively on newly arriving data, will overwrite their representations and thus forget encoded knowledge—a phenomenon referred to as catastrophic interference or catastrophic forgetting [7,8]. Although we have worded the above questions in a way that naturally exposes their connection: to identify what is new and think about how new

concepts can be incorporated, they are largely subject to separate treatment in the respective literature. While open-set recognition [1,9,10] aims to explicitly identify novel inputs that deviate with respect to already observed instances, the existing continual learning literature predominantly concentrates its efforts on finding mechanisms to alleviate catastrophic interference (see [11] for an algorithmic survey).

In particular, the indispensable system component to distinguish seen from unseen unknown data, both as a guarantee for robust application and to avoid the requirement of explicit task labels for prediction, is generally missing from recent continual-learning works. Inspired by this gap, we set out to connect open-set recognition and continual learning. The underlying connecting element is motivated from the prior work of Bendale and Boulton [12], who proposed to leverage extreme value theory (EVT) to address open-set detection in deep neural networks. The authors suggested to modify softmax prediction scores on the basis of feature space distances in blackbox discriminative models. Although this approach is promising, it alas comes with the substantial caveat that purely discriminative networks are prone to encode noise as features [13] or fall for a most simple discriminative solution that neglects meaningful features [14]. Inspired by these former insights, we set out to connect open-set recognition and continual learning, while overcoming present limitations through treatment from a generative modeling perspective.

Our specific contributions are that we propose to unify the prevention of catastrophic interference in continual learning with open-set recognition in a single model. Specifically, we extend prior EVT works [9,10,12] to a natural formulation on the basis of the aggregate posterior in variational inference with deep autoencoders [15,16]. By *identifying out-of-distribution instances* we can detect unseen unknown data and prevent false predictions; by explicitly *generating in-distribution samples* from areas of high probability density under the aggregate posterior, we can simultaneously circumvent rehearsal of ambiguous uninformative examples. This leads to robust application, while significantly reducing catastrophic interference. We empirically corroborated our approach in terms of improved out-of-distribution detection performance and simultaneously reduced the continual catastrophic interference. We further demonstrate benefits through recent deep generative modeling advances, such as autoregression [2,17,18] and introspection [19,20], validated by scaling to high-resolution color images.

1.1. Background and Related Work

1.1.1. Continual Learning

In isolated supervised learning, the core assumption is the presence of i.i.d. data at all times and training is conducted using a dataset $D \equiv \left\{ \left(x^{(n)}, y^{(n)} \right) \right\}_{n=1}^N$, consisting of N pairs of data instances $x^{(n)}$ and their corresponding labels $y^{(n)} \in \{1 \dots C\}$ for C classes. In contrast, in continual learning, data $D_t \equiv \left\{ \left(x_t^{(n)}, y_t^{(n)} \right) \right\}_{n=1}^{N_t}$ with $t = 1, \dots, T$ arrives sequentially for T disjoint sets, each with number of classes C_t .

It is assumed that only the data of the current task is available. Without additional mechanisms, tuning on such a sequence will lead to catastrophic interference [7,8], i.e., representations of former tasks being overwritten through present optimization. A recent review of many continual-learning algorithms to prevent said interference was provided by Parisi et al. [11]. Here, we present a brief summary of the key underlying principles.

Alleviating catastrophic interference is most prominently addressed from two angles. Regularization methods, such as synaptic intelligence (SI) [21] or elastic weight consolidation (EWC) [22] explicitly constrain the weights during continual learning to avoid drifting too far away from the previous tasks' solutions. In a related picture, learning without forgetting [23] uses knowledge distillation [24] to regularize the end-to-end functional.

Rehearsal methods on the other hand, store data subsets from distributions belonging to old tasks or generate samples in pseudo-rehearsal [25]. The central component of the latter is thus the selection of significant instances. For methods, such as incremental classifier and representation learning (iCarl) [26], it is therefore common to resort to aux-

iliary techniques, such as the nearest-mean classifier [27] or core sets [28]. Inspired by complementary learning systems [29], dual-model approaches sample data from a separate generative memory. In a bio-inspired incremental learning architecture (GeppNet) [30], long short-term memory [31] is used for storage, whereas generative replay [32] samples from an additional generative adversarial network (GAN) [33].

As detailed in Variational Generative Replay (VGR) [34,35], methods with a Bayesian perspective encompass a natural capability for continual learning by making use of the learned distribution. Existing works nevertheless fall into the above two categories and their combination: a prior-based approach using the former task's approximate posterior as the new task's prior [36] or estimating the likelihood of former data through generative replay or other forms of rehearsal [34,37]. Crucially, the success of many continual-learning techniques can be attributed primarily to the considered evaluation scenario. With the exception of VGR [34], the majority of above techniques train a separate classifier per task and thus either require the explicit storage of task labels or assume the presence of a task oracle during evaluation. This multi-head scenario prevents "cross-talk" between classifier units by not sharing them, which would otherwise rapidly decay the accuracy as newly introduced classes directly confuse existing concepts. While the latter is acceptable to limit catastrophic interference, it also signifies a major limitation in practical applications. Even though VGR [34] uses a single classifier, the researchers trained a separate generative model per task to avoid catastrophic interference in the generator.

Our approach builds upon these previous works and leverages variational inference in deep generative models. However, we propose to tie the prevention of catastrophic interference with open-set recognition through a natural mechanism based on the aggregate posterior in a single model.

1.1.2. Out-of-Distribution and Open Set Recognition

The above-mentioned literature focused their efforts predominantly on addressing catastrophic interference. Even though continual learning is the desideratum, the corresponding evaluation is thus conducted in a closed world setting, where instances that do not belong to the observed data distribution are not encountered. In reality, this is not guaranteed as users could provide arbitrary inputs or unknowingly present the system with novel inputs that deviate substantially from previously seen instances. Our models thus require the ability to identify unseen examples in the unconstrained open world and categorize them as either belonging to the already known set of classes or as presently being unknown. We provide a small overview of approaches that aim to address this question in deep neural networks. A comprehensive survey was provided by Boulton et al. [1].

As the most simple approach, the aim of calibration works is to separate a known and unknown input through prediction confidence, often by fine tuning or re-training an already existing model. In out-of-distribution detector for neural networks (ODIN) [38], this is addressed through perturbations and temperature scaling, while Lee et al. [39] used a separately trained GAN to generate out-of-distribution samples from low probability densities and explicitly reduced their confidence through the inclusion of an additional loss term. Similarly, the objectsphere loss [40] defines an objective that explicitly aims to maximize entropy for upfront available unknown inputs.

As we do not have access to future data a priori, by definition, a naive conditioning or calibration on unseen unknown data is infeasible. The commonly applied thresholding is insufficient as overconfident prediction values cannot be prevented [3]. Bayesian neural network models [41] could be believed to intrinsically be able to reject statistical outliers through model uncertainty [34] and overcome this limitation of overconfident prediction values. For use with deep neural networks, it was suggested that stochastic forward passes with Monte-Carlo Dropout (MCD) [42] can provide a suitable approximation. However, the closed-world assumption in training and evaluation still persists [1]. In addition, variational approximations in deep networks [15,34,37,43] and corresponding uncertainty estimates

suffer from similar overconfidence, and the distinction of unseen out-of-distribution data from already trained knowledge is known to be unsatisfactory [5,6].

A more formal approach was suggested in works based on open-set recognition [9]. The key here is to limit predictions originating from open space, that is, the area in obtained embeddings that is outside of a small radius around previously observed training examples. Without re-training, post hoc calibration or modifying loss functions, one approach to open-set recognition in deep networks is through extreme-value theory (EVT) [10,12]. Here, limiting the threat of overconfidence is based on monotonically decreasing the recognition function's probability with respect to increasing distance of instances to the feature embedding of known training points. The Weibull distribution, as one member of the family of extreme value distributions, has been empirically demonstrated to work well in conjunction with distances in the penultimate deep network layer as the underlying feature space. On the basis of extreme values to this layer's average activation values, the authors devised a procedure to revise the Softmax prediction values, referred to as OpenMax.

In a similar spirit, our work avoids relying on predictive values, while also moving away from empirically chosen deep neural network feature spaces. We instead propose to use EVT to bind the approximate posterior in variational inference. We thus directly operate on the underlying (lower-bound to the) data distribution and the generative factors. This additionally allows us to constrain the generative replay to distribution inliers, which further alleviates catastrophic interference.

2. Materials and Methods

2.1. Unifying Catastrophic Interference Prevention with Open Set Recognition

We first summarize the preliminaries on continual learning from a perspective of variational inference in deep generative models [15,43]. We then proceed by bridging the improved prevention of catastrophic interference in continual learning with the detection of unseen unknown data in open-set recognition.

2.1.1. Preliminaries: Learning Continually through Variational Auto-Encoding

We start with a problem scenario similar to the one introduced in "Auto-Encoding Variational Bayes" [15], i.e., we assume that there exists a data generation process responsible for the creation of the labeled data given some random latent variable z . We consider a model with a shared encoder with variational parameters θ , decoder and linear classifier with respective parameters ϕ and ξ . The joint probabilistic encoder learns an encoding to a latent variable z , over which a unit Gaussian prior $p(z)$ is placed.

Using variational inference, the encoder's purpose is to approximate the true posterior to $p_\phi(x, z)$ and $p_\xi(y, z)$. The probabilistic decoder $p_\phi(x|z)$ and probabilistic linear classifier $p_\xi(y|z)$ then return the conditional probability density of the input x and target y under the respective generative model given a sample z from the approximate posterior $q_\theta(z|x)$. This yields a generative model $p(x, y, z)$, for which we assume a factorization and generative process of the form $p(x, y, z) = p(x|z)p(y|z)p(z)$. For variational inference with this model, the sum over all elements in the dataset $n \in D$ in the following lower-bound is optimized:

$$p(x, y) \geq \mathbb{E}_{q_\theta(z|x)} [\log p_\phi(x|z) + \log p_\xi(y|z)] - \beta KL(q_\theta(z|x) || p(z)), \quad (1)$$

where KL denotes the Kullback-Leibler divergence. In other words, the right hand side of Equation (1) defines our loss $\mathcal{L}(x, y; \theta, \phi, \xi)$. This model can be seen as employing a variant of a (semi-)supervised variational auto-encoder (VAE) [16] with a β term [44], where, in addition to approximating the data distribution, the model learns to incorporate the class structure into the latent space. Without the blue terms, the original unsupervised VAE formulation [15] is recovered. This forms the basis for continual learning with open-set recognition as discussed in the subsequent section. An illustration of the model is shown in Figure 1.

Abstracting away from the mathematical detail and speaking informally about the intuition behind the model, we first encode a data input x and encode it into two vectors.

These vectors represent the mean and standard deviation of a Gaussian distribution. Using the reparametrization trick $\epsilon \cdot \sigma + \mu$, a sample from this distribution is then calculated. During training, the respective embedding, also referred to as the latent space, is encouraged to follow a unit Gaussian distribution through the minimization of the Kullback-Leibler divergence. A linear classifier that operates directly on this latent embedding to predict a class for a sample additionally ensures that the obtained distribution is clustered according to the classes.

Examples of such fits are shown in the later Figure 2. Finally, the decoder takes, as input, the latent variable and reconstructs the original data input during training. Once the model is finished training, we can also directly draw a sample from the Gaussian distribution, obtain a latent sample and generate a novel data point directly, without the need to compute the encoder first. A corresponding full and formal derivation of Equation (1), the lower-bound to the joint distribution $p(x, y)$ is supplied in Appendix A.1.

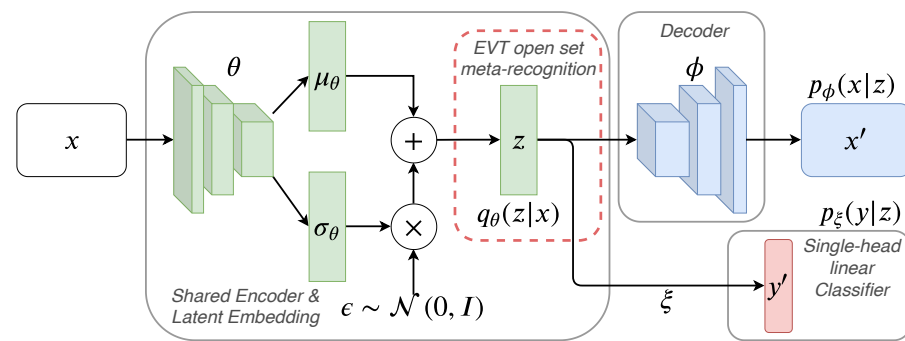


Figure 1. A joint continual-learning model consisting of a shared probabilistic encoder $q_{\theta}(z|x)$, probabilistic decoder $p_{\phi}(x, z)$ and probabilistic classifier $p_{\xi}(y, z)$. For open-set recognition and generative replay with outlier rejection, extreme-value theory (EVT) based bounds on the basis of the approximate posterior are established.

Without further constraints, one could continually train the above model by sequentially accumulating and optimizing Equation (1) over all currently present tasks $t = 1, \dots, T$. Being based on the accumulation of real data, this provides an upper bound to the achievable performance in continual learning. However, this form of continued training is generally infeasible if only the most recent task’s data is assumed to be available. Making use of the model’s generative nature, we can follow previous works [34,37] and estimate the likelihood of former data through generative replay:

$$\mathcal{L}_t(x, y; \theta, \phi, \xi) = \frac{1}{2} \frac{1}{N_t} \sum_{n=1}^{N_t} \mathcal{L}(x_t^{(n)}, y_t^{(n)}; \theta, \phi, \xi) + \frac{1}{2} \frac{1}{N'_t} \sum_{n=1}^{N'_t} \mathcal{L}(x'_t{}^{(n)}, y'_t{}^{(n)}; \theta, \phi, \xi) \quad (2)$$

where

$$x'_t \sim p_{\phi, t-1}(x|z); y'_t \sim p_{\xi, t-1}(y|z) \text{ and } z \sim p(z). \quad (3)$$

Here, x'_t is a sample from the generative model with its corresponding classifier label y'_t . N'_t is the number of instances of all previously seen tasks. In this way, the expectation of the log-likelihood for all previously seen tasks is estimated and the dataset at any point in time $\tilde{D}_t \equiv \left\{ (x_t^{(n)}, y_t^{(n)}) \right\}_{n=1}^{N_t} = \{ (x_t \cup x'_t, y_t \cup y'_t) \}$ is a concatenation of past data generations and the current task’s real data.

2.1.2. Open Set Recognition and Generative Replay with Statistical Outlier Rejection

Trained naively in the above fashion, our model will unfortunately suffer from accumulated errors with each successive iteration of generative replay, similar to the current literature approaches. To avoid this, we would alternatively require the training of multiple encoders to approximate each task’s posterior individually, as in variational continual

learning (VCL) [36], or train multiple generators, as in VGR [34]. We posit that the main challenge is how high-density areas under the prior $p(z)$ are not necessarily reflected in the structure of the aggregate posterior $q_{\theta,t}(z)$ [45]. The latter refers to the practically obtained encoding [46]:

$$q_{\theta,t}(z) = \mathbb{E}_{p_{\tilde{D}_t}(\tilde{x})}[q_{\theta,t}(z|\tilde{x})] \approx \frac{1}{\tilde{N}_t} \sum_{n=1}^{\tilde{N}_t} q_{\theta,t}(z|\tilde{x}^{(n)}) \tag{4}$$

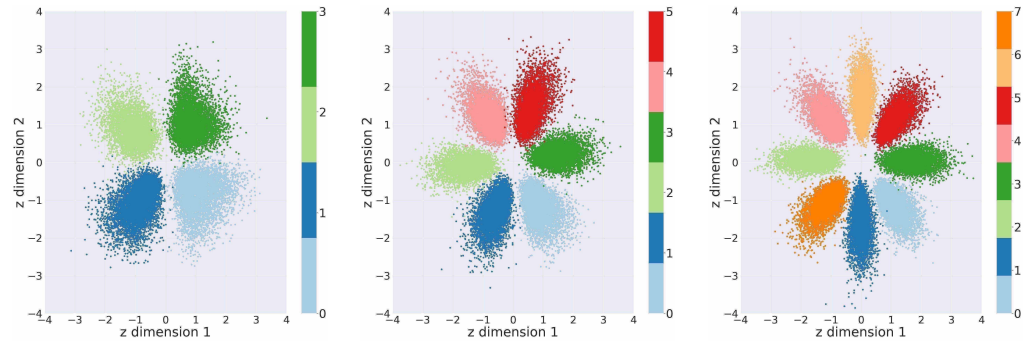


Figure 2. 2-D latent space aggregate posterior visualization for continually learned MNIST (Modified National Institute of Standards and Technology database). From left to right, the latent space for four, six and then eight classes are shown. This is best viewed in color.

To provide intuition, we illustrate this prior-posterior discrepancy on the obtained two-dimensional latent encodings for a continually trained supervised MNIST (Modified National Institute of Standards and Technology database) [47] model in Figure 2. Here, we can make two observations: to preserve the inherent data structure, the aggregate posterior deviates from the prior. In fact, this is further amplified by the imposed necessity for linear class separation and the beta term in Equation (1); however, we note that the discrepancy is desired even in completely unsupervised scenarios [45,46].

The underlying rationale is that there needs to be a balance in the effective latent encoding overlap [48], which can best be summarized with a direct quote from the recent work of Mathieu et al. [49]: “The overlap is perhaps best understood by considering extremes: with too little the latents effectively become a lookup table; too much, and the data and latents do not convey information about each other. In either case, meaningfulness of the latent encodings is lost.” (p. 4). Additional discussion on the role of beta can be found in Appendix A.2.

Thus, the generated data from low-density regions of the aggregate posterior do not generally correspond to the encountered data instances. Conversely, data instances that fall into high-density regions under the prior should not generally be considered as statistical inliers with respect to the observed data distribution; recall Figure 2. This boundary between low- and high-density regions forms the basis for a natural connection between open-set recognition and continual learning: generate from high-density regions and reject novel instances that fall into low-density regions.

Ideally, we could find a solution by replacing the prior in the KL divergence of Equation (1) with $q_{\theta,t}(z)$ and, respectively, sampling $z \sim q_{\theta,t-1}(z)$ in Equations (2) and (3). Even though using the aggregate posterior as a subsequent prior is the objective in multiple recent works, it can be challenging in high dimensions, lead to over-fitting or come at the expense of additional hyper-parameters [45,50,51]. To avoid finding an explicit representation for the multi-modal $q_{\theta,t}(z)$, we draw inspiration from the EVT-based OpenMax approach [12] in deep neural networks. However, instead of using knowledge about extreme distances in penultimate layer activations to modify a Softmax prediction, we now propose to apply EVT on the basis of the class conditional aggregate posterior.

In this view, any sample can be regarded as statistically outlying if its distance to the classes’ latent mean is extreme with respect to what has been observed for the majority of correctly predicted data instances, i.e., the sample falls into a region of low density under the aggregate posterior and is less likely to belong to $p_{\tilde{D}}(\tilde{x})$. For convenience, let us

introduce the indices of all correctly classified instances at the end of task t as $m = 1, \dots, \tilde{M}_t$. To obtain bounds on the aggregate posterior, we first define the mean latent vector for each class for all correctly predicted seen data instances $\bar{z}_{c,t}$ and the respective set of latent distances as

$$\Delta_{c,t} \equiv \left\{ f_d \left(\bar{z}_{c,t}, \mathbb{E}_{q_{\theta,t}(z|\bar{x}_t^{(m)})}[z] \right) \right\}_{m \in \tilde{M}_{c,t}} \quad \text{with} \quad \bar{z}_{c,t} = \frac{1}{|\tilde{M}_{c,t}|} \sum_{m \in \tilde{M}_{c,t}} \mathbb{E}_{q_{\theta,t}(z|\bar{x}_t^{(m)})}[z]. \quad (5)$$

Here, f_d signifies a choice of distance metric. We proceed to model this set of distances with a per class heavy-tail Weibull distribution $\rho_{c,t} = (\tau_{c,t}, \kappa_{c,t}, \lambda_{c,t})$ on $\Delta_{c,t}$ for a given tail-size η . As these distances are based on the class conditional approximate posterior, we can thus bound the latent space regions of high density. The tightness of the bound is characterized through η , that can be seen as a prior belief with respect to the outlier quantity assumed to be inherently present in the data distribution. The choice of f_d determines the nature and dimensionality of the obtained distance distribution. For our experiments, we find that the cosine distance and thus a univariate Weibull distance distribution per class seems to be sufficient. Using the cumulative distribution function of this Weibull model ρ_t we can now estimate any sample's outlier (or inlier) probability:

$$\omega_{\rho,t}(z) = \min \left(1 - \exp \left(- \frac{|f_d(\bar{z}_t, z) - \tau_t|}{\lambda_t} \right)^{\kappa_t} \right), \quad (6)$$

where the minimum returns the smallest outlier probability across all classes. If this outlier probability is larger than a prior rejection probability Ω_t , the instance can be considered as unknown. Such a formulation, which we term open variational auto-encoder (OpenVAE), now provides us with the means to learn continually and identify unknown data:

1. For a novel data instance, Equation (6) yields the outlier probability based on the probabilistic encoder $z \sim q_{\theta,t}(z|x)$, and a false overconfident classifier prediction can be avoided.
2. To mitigate catastrophic interference, Equation (6) can be used on top of $z \sim p(z)$ to constrain the generative replay (Equation (3)) to the aggregate posterior thus avoiding the need to sample it directly.

To give an illustration of the benefits, we show the generated MNIST [47] and larger resolution flower images [52] together with their outlier percentage in Figure 3. In practical application, we discard the ambiguous examples that are due to low-density regions and thus a high outlier probability. Even though we conduct sampling with rejection, note how this is computationally efficient, as we only need to calculate the heavy probabilistic decoder for accepted statistically inlying examples, and sampling from the prior with computation of Equation (6) is almost negligible in comparison.

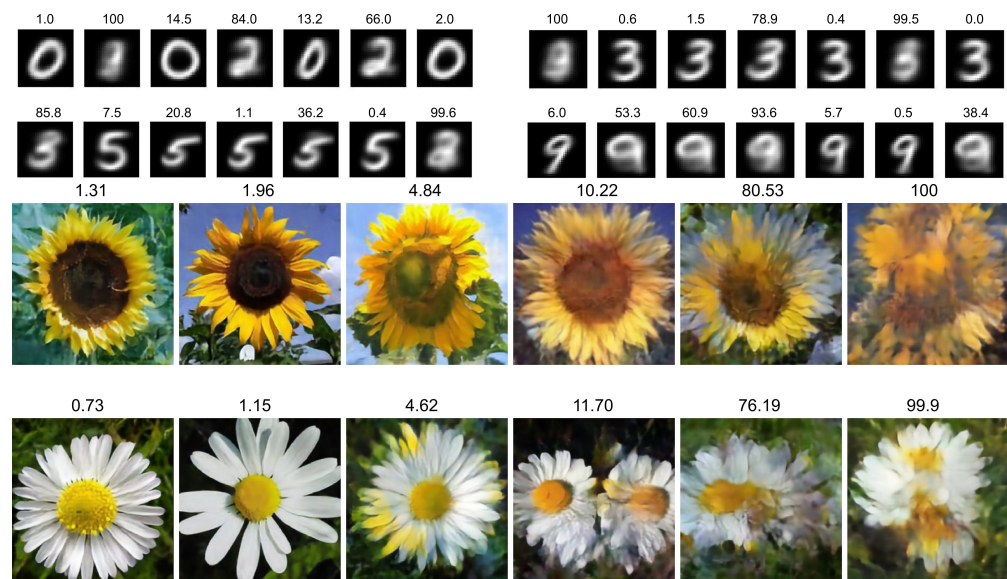


Figure 3. Generated images $x \sim p_{\phi,t}(x|z)$ with $z \sim p(z)$ and their corresponding class c obtained from the classifier $p_{\xi,t}(y|z)$ together with their open-set outlier percentage in our proposed open variational auto-encoder (OpenVAE). Image quality degradation and class ambiguity can be observed with the increasing outlier likelihood. Generated 28×28 MNIST images are from the 2-D latent space of Figure 2, classified as $c = 0$ (top left), $c = 3$ (top right), $c = 5$ (bottom left) and $c = 9$ (bottom right). Generated 256×256 resolution flower images are based on a 60-dimensional latent space of a model trained with introspection (see experiments and Appendix A.3), which are classified as “sunflower” (top) and “daisy” (bottom).

3. Results

Instead of presenting a single experiment for continual learning in the constant presence of outlying non-task data, we chose to empirically corroborate our proposed approach in two experimental parts. The first section is dedicated to out-of-distribution detection, where we demonstrate the advantages of EVT in our generative model formulation. We then proceed to showcase how catastrophic interference is also mitigated by confining generative replay to aggregate posterior inliers in class incremental learning.

We emphasize that whereas the sections are *presented individually*, our approach’s uniqueness lies in *using a core underlying mechanism to unify both challenges simultaneously*. The rationale behind choosing this form of presentation is to help readers better contextualize the contribution of OpenVAE with the existing literature as, to the best of our knowledge, there exists no present other work that yields adequate continual classification accuracy while being able to robustly recognize unknown data instances. As such, we will now see that existing continual-learning approaches provide no suitable mechanism to overcome the challenge of providing robust predictions when data outside the known benchmark set are included.

3.1. Open Set Recognition

We experimentally highlight OpenVAE’s ability to distinguish unknown task data from data belonging to known tasks to avoid overconfident false predictions.

Experimental Set-Up and Evaluation

In summary, our goal is two-fold. The typical goal is to train on an initial task and correctly classify the held-out or unseen test data for this task. That is, we desire a large average classification test accuracy. In addition to this, in order to ensure that this classification is robust to unknown data, we now additionally desire to have a large value for a second kind of accuracy. Our simultaneous goal is to consider all test data

of already trained tasks as inlying, while successfully identifying 100% of completely unknown datasets as outliers.

For this purpose, we evaluate OpenVAE’s and other models’ capability to distinguish the in-distribution test set of a respectively trained MNIST (Modified National Institute of Standards and Technology database) [47], FashionMNIST [53], AudioMNIST [54] from the other two and several unknown datasets: Kuzushiji-MNIST (KMNIST) [55], Street-View House Numbers (SVHN) [56] and Canadian Institute for Advanced Research (CIFAR) datasets (in both versions with 10 and 100 classes) [57]. Here, the (Fourier-transformed) audio data is included to highlight the extent of the challenge, as not even a different modality is easy to detect without our proposed approach. In practice, we evaluate three criteria according to which a decision of whether a data instance is an outlier can be made:

1. The classifier’s predictive entropy, as recently suggested to work surprisingly well in deep networks [58] but technically well known to be overconfident [3]. The intuition here is that the predictive entropy $-\sum_{y \in C} p(y|x) \log p(y|x)$ considers the probability of all other classes and is at a maximum if the distribution is uniform, i.e., when the confidence in the prediction is low.
2. The generative model’s obtained negative log-likelihood, to concur with previous findings [5,6] on overconfidence in generative models. On the basis of Equation (1), the intuition is that the negative log-likelihood should be much larger for unseen data.
3. Our suggested OpenVAE aggregate posterior-based EVT approach, according to the outlier likelihood introduced Equation (6).

Results

Figure 4 provides a qualitative intuition behind the three criteria and respective percentage of the total dataset being considered as outlying for FashionMNIST. Consistent with Nalisnick et al. [6], we can observe that the use of reconstruction loss can sometimes distinguish between the known tasks’ test data and unknown datasets but results in failure for others. In the case of the classifier predictive entropy, depending on the exact choice of entropy threshold, generally only a partial separation can be achieved. Furthermore, both of these criteria pose the additional challenge of the results being highly dependent on the choice of the precise cut-off value. In contrast, the test data from the known tasks is regarded as inlying across a wide range of rejection priors Ω_t for Equation (6), and the majority of other datasets is consistently regarded as outlying by our introduced OpenVAE approach.

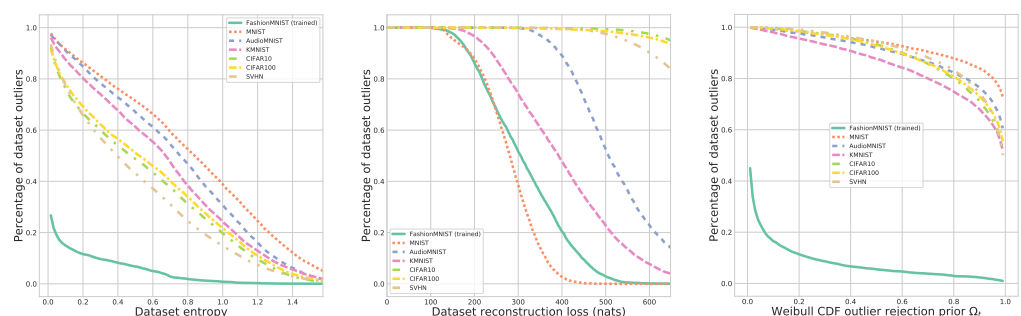


Figure 4. Model trained on FashionMNIST evaluated on unknown datasets. Robust classification of a known dataset (percentage of dataset outliers at 0%), while correctly flagging unknown datasets as outlying (percentage of dataset outliers at 100%), occurs when the solid green curve is separated from any of the colored dashed curves. **(Left)** Classifier entropy is insufficient to separate unknown from the known task’s test data. **(Center)** Reconstruction log-likelihood allows for a partial distinction. **(Right)** Our posterior-based EVT approach in OpenVAE considers the large majority of unknown data as statistical outliers across a wide range of rejection priors Ω_t .

Corresponding quantitative outlier detection accuracies are provided in Table 1. To find thresholds for the sensitive entropy and reconstruction curves, we used a 5% validation split to determine the respective value at which 95% of the validation data is considered as

inlying before using these priors to determine outlier counts for the known tasks' test set as well as other datasets. In an intuitive picture, we "trace" the solid green curve of Figure 4 for a validation set of the originally trained dataset, check where we intersect with the x -axis for a y -axis value of 5% and then fix the corresponding criterion's value at this point as an outlier rejection threshold for testing. We then report the percentage of the test set being considered as an outlier, together with the percentage for various unknown datasets. In the table, we additionally extend our intuition of Figure 4 to now further investigate what would happen if we had not trained a single VAE model that learned reconstruction and classification according to Equation (1) but separate models. For this purpose, we also investigate a dual model approach, i.e., a purely discriminative deep-neural-network-based classifier and a separate unsupervised VAE (Equation (1) without blue terms).

In this way, we can showcase the advantages of a generative modeling formulation that considers the joint distribution $p(x, y)$ in conjunction with EVT. For instance, we can compare our values with the purely discriminative OpenMax EVT approach [59]. At the same time, this provides a justification for why the existing continual-learning approaches of the next section, especially those relying on the maintenance of multiple models, are non-ideal, as they cannot seem to adequately solve the open-set challenge.

In terms of the obtained results, with the exception of MNIST, which appears to be an easy to identify dataset for all approaches, we can make two key observations:

1. Both EVT approaches generally outperform the other criteria, particularly for our suggested aggregate posterior-based OpenVAE variant, where a near perfect open-set detection can be achieved.
2. Even though EVT can be applied to purely discriminative models (as in OpenMax), the generative OpenVAE model trained with variational inference consistently exhibited more accurate outlier detection. We posit that this robustness is due to OpenVAE explicitly optimizing a variational lower bound that considers the data distribution $p(x)$ in addition to a pure optimization of features that maximize $p(y|x)$.

Open Set Recognition with Monte-Carlo Dropout Based Uncertainty

One might be tempted to assume that the trained weights of the individual deep neural network encoder layers are still deterministic and the failure of predictive entropy as a measure for unseen unknown data could thus primarily be attributed to uncertainty not being expressed adequately. Placing a distribution on the weights, akin to a fully Bayesian neural network, would then be expected to resolve this issue. For this purpose, we further repeat all of our experiments by treating the model weights as the random variable being marginalized through the use of Monte-Carlo Dropout (MCD) [42]. Accordingly, the models were re-trained with a Dropout probability of 0.2 in each layer. We then conducted 50 stochastic forward passes through the entire model for prediction. The obtained open-set recognition results are reported in Table 2.

Although MCD boosts the outlier detection accuracy, particularly for criteria, such as predictive entropy, the previous insights and drawn conclusions still hold. In summary, the joint generative model generally outperforms a purely discriminative model in terms of open-set recognition, independently of the used metric, and our proposed aggregate posterior-based EVT approach of OpenVAE yields an almost perfect separation of known and unseen unknown data. Interestingly, this was already achieved in the prior table without MCD. Resorting to the repeated model calculation of MCD thus appears to be without enough of an advantage to warrant the added computational complexity in the context of posterior-based open-set recognition, a further key advantage of OpenVAE.

3.2. Learning Classes Incrementally in Continual Learning

To showcase how our OpenVAE approach mitigates catastrophic interference in addition to successfully handling unknown data in robust prediction, we conduct an investigation of the test accuracy when learning classes incrementally.

Experimental Set-Up and Evaluation

We consider the incremental MNIST dataset (where classes arrive in groups of two) and the corresponding versions of the FashionMNIST and AudioMNIST datasets, similar to popular literature [11,21,22,32,34]. We re-emphasize that such a setting has a sole focus on mitigating catastrophic interference and does not account for the challenges presented in the previous open-set recognition section, which we detail in the prospective discussion section. For a flexible comparison, we report our aggregate posterior-based generative replay approach in OpenVAE on both a simple multi-layer perceptron (MLP), as well as a deep convolutional neural network (CNN) based on wide residual networks (WRN). For the former, we follow previous continual-learning studies and employ a two-hidden-layer and 400-unit multi-layer perceptron [60]. For the latter, we use both encoder and decoder architectures of 14-layer wide residual networks [61,62] with a latent dimensionality of 60 [2,18]. For our statistical outlier rejection, we use a rejection prior of $\Omega_t = 0.01$ and dynamically set tail-sizes to 5% of seen examples per class.

For our own experiments, we report the mean and standard deviation of the average classification test accuracy across five experimental repetitions. If our re-implementation of related works achieved a better than original value, we report this number, otherwise the work that reported the specific best value is cited next to it. The full training details, including details on hardware and code, are supplied in Appendix A.4.

Table 1. Outlier detection values of the joint model and separate discriminative and generative models (denoted as “CNN + VAE”; discriminative convolutional neural network and variational auto-encoder), when considering 95% of the known tasks’ validation data as inlying. The percentage of detected outliers is reported based on the classifier predictive entropy, reconstruction negative log-likelihood (NLL) and our posterior-based extreme-value theory approach. Note that larger values are better, except for the test data of the trained dataset, where ideally 0% should be considered as outlying. The outlier detection values have additionally been color coded, where *worse* results appear in red. A deeper shading thus indicates a method’s failure to robustly recognize unknown data as such. With this color coding, we can easily see how MNIST appears to be an easy to identify dataset for all approaches; however, we notice right away that our OpenVAE is the *only* method (row) that does *not* have a single red value for any dataset combination. In fact, the lowest outlier detection accuracy of OpenVAE is a very high 94.76%.

Outlier Detection at 95% Validation Inliers (%)				MNIST	Fashion	Audio	KMNIST	CIFAR10	CIFAR100	SVHN
Trained	Model	Test Acc.	Criterion							
MNIST	Dual, CNN + VAE	99.40	Class entropy	4.160	90.43	97.53	95.29	98.54	98.63	95.51
			Reconstruction NLL	5.522	99.98	99.97	99.98	99.99	99.96	99.98
			OpenMax	4.362	99.41	99.80	99.86	99.95	99.97	99.52
	Joint VAE	99.53	Class entropy	3.948	95.15	98.55	95.49	99.47	99.34	97.98
			Reconstruction NLL	5.083	99.50	99.98	99.91	99.97	99.99	99.98
			OpenVAE (ours)	4.361	99.78	99.67	99.73	99.96	99.93	99.70
FashionMNIST	Dual, CNN + VAE	90.48	Class entropy	74.71	5.461	69.65	77.85	24.91	28.76	36.64
			Reconstruction NLL	5.535	5.340	64.10	31.33	99.50	98.41	97.24
			OpenMax	96.22	5.138	93.00	91.51	71.82	72.08	73.85
	Joint VAE	90.92	Class Entropy	66.91	5.145	61.86	56.14	43.98	46.59	37.85
			Reconstruction NLL	0.601	5.483	63.00	28.69	99.67	98.91	98.56
			OpenVAE (ours)	96.23	5.216	94.76	96.07	96.15	95.94	96.84
AudioMNIST	Dual, CNN + VAE	98.53	Class entropy	97.63	57.64	5.066	95.53	66.49	65.25	54.91
			Reconstruction NLL	6.235	46.32	4.433	98.73	98.63	98.63	97.45
			OpenMax	99.82	78.74	5.038	99.47	93.44	92.76	88.73
	Joint VAE	98.57	Class entropy	99.23	89.33	5.731	99.15	92.31	91.06	85.77
			Reconstruction NLL	0.614	38.50	3.966	36.05	98.62	98.54	96.99
			OpenVAE (ours)	99.91	99.53	5.089	99.81	100.0	99.99	99.98

Table 2. Outlier detection values of the joint model and separate discriminative and generative models (denoted as “CNN + VAE”; discriminative convolutional neural network and variational auto-encoder), when considering 95% of known tasks’ validation data is inlying. The percentage of detected outliers is reported based on classifier predictive entropy, reconstruction negative log-likelihood (NLL) and our posterior-based EVT approach. In contrast to Table 1, the results are now averaged over 50 Monte-Carlo dropout samples, with $p_{dropout} = 0.2$ for each layer, per data-point, respectively, to assess the model uncertainty. Note that larger values are better, except for the test data of the trained dataset, where ideally 0% should be considered as outlying. The color coding is analogous to Table 1.

Outlier Detection at 95% Validation Inliers (%)				MNIST	Fashion	Audio	KMNIST	CIFAR10	CIFAR100	SVHN
Trained	Model	Test Acc.	Criterion							
MNIST	Dual, CNN + VAE	99.41	Class entropy	4.276	91.88	96.50	96.65	95.84	97.37	98.58
			Reconstruction	4.829	99.99	100.0	99.90	100.0	100.0	100.0
			OpenMax	4.088	87.84	98.06	95.79	97.34	98.30	95.74
	Joint, VAE	99.54	Class entropy	4.801	97.63	99.38	98.01	99.16	99.39	98.90
			Reconstruction	5.264	99.98	100.0	100.0	100.0	100.0	100.0
			OpenVAE (ours)	4.978	99.99	100.0	99.94	99.96	99.95	99.68
FashionMNIST	Dual, CNN + VAE	90.58	Class entropy	75.50	5.366	70.78	74.41	49.42	49.17	38.84
			Reconstruction NLL	55.45	5.048	59.99	99.83	99.35	99.35	99.62
			OpenMax	77.03	4.920	55.48	70.23	58.73	57.06	44.54
	Joint, AE	91.50	Class Entropy	85.05	4.740	67.90	78.04	63.89	66.11	59.42
			Reconstruction	1.227	5.422	85.85	39.76	99.94	99.72	99.99
			OpenVAE (ours)	95.83	4.516	94.56	96.04	96.81	96.66	96.28
AudioMNIST	Dual, CNN + VAE	98.76	Class entropy	99.97	61.26	4.996	96.77	63.78	65.76	59.38
			Reconstruction NLL	7.334	52.37	5.100	98.19	99.97	99.90	99.96
			OpenMax	92.74	67.18	5.073	90.41	90.56	90.97	89.58
	Joint, VAE	98.85	Class entropy	99.39	89.50	5.333	99.16	94.66	95.12	97.13
			Reconstruction NLL	15.81	53.83	4.837	41.89	99.90	99.82	99.95
			OpenVAE (ours)	99.50	99.27	5.136	99.75	99.71	99.59	99.91

Results

In Table 3, we report the final accuracy after having trained on each of the five increments. For an overall reference, we provide the achievable upper-bound continual-learning performance, i.e., accumulating all data over time and optimizing Equation (1). We can observe that our proposed OpenVAE approach provides significant improvement over generative replay with a conventional supervised VAE. In comparison with the immediately related works, our approach surpasses variational continual learning (VCL) [36], an approach that employs a full Bayesian neural network (BNN), with the additional benefit that our approach scales trivially to complex network architectures.

In contrast to variational generative replay (VGR) [34], OpenVAE initially appears to fall short. This is not surprising as VGR trains a separate GAN on each task’s aggregate posterior, an apples to oranges comparison considering that we only use a single model. Nevertheless, even in a single model, we can surpass the multi-model VGR by leveraging recent advancements in generative modeling, e.g., by making the neural architecture more complex or augmenting our decoder with autoregressive sampling [2,18] (a complementary technique to OpenVAE, often also called PixelVAE and summarized in Appendix A.3).

At the bottom of Table 3, we can see that this significantly improves upon the previously obtained accuracy. The full accuracies, along with other metrics per dataset for all intermediate steps can be found in Appendix A.6.

Table 3. The accuracy α_T at the end of the last increment $T = 5$ for class incremental learning approaches averaged over five runs. For a fair comparison, if our re-implementation of related works achieved a better than original value, we report our number, otherwise the work that reported the specific best value is cited right next to the result. Intermediate results can be found in Appendix A.6.

Method	Final Accuracy $\alpha_T(T = 5)$ [%]		
	MNIST	FashionMNIST	AudioMNIST
MLP upper bound	98.84	87.35	96.43
WRN upper bound	99.29	89.24	97.87
EWC [22]	55.80 [63]	24.48 \pm 2.86	20.48 \pm 1.73
DGR [32]	75.47 [64]	63.21 \pm 1.96	48.42 \pm 2.81
VCL [36]	72.30 [35]	32.60 [35]	-
VGR [35]	92.22 [35]	79.10 [35]	-
Supervised VAE	60.88 \pm 3.31	62.72 \pm 1.38	69.76 \pm 1.37
OpenVAE—MLP	87.31 \pm 1.22	66.14 \pm 0.50	81.84 \pm 1.44
OpenVAE—WRN	93.24 \pm 3.74	69.88 \pm 1.71	87.72 \pm 1.59
OpenPixelVAE	96.84 \pm 0.35	80.85 \pm 0.72	90.23 \pm 1.14

High-Resolution Flower Images

While the main goal of this paper is not to push the achievable boundaries of generation, we take this argument one step further and provide empirical evidence that our suggested aggregate posterior-based EVT sampling provides similar benefits when scaling to higher resolution color images. For this purpose, we consider the additional flowers dataset [52] at a resolution of 256×256 , investigated with five classes and increments of one class per step [65,66].

In addition to autoregressive sampling, we also include a second complementary generative modeling improvement here, called VAEs with introspection (IntroVAE) [19]. A technical description of PixelVAE and IntroVAE is detailed in Appendix A.3. For each generative modeling variant, including autoregression and introspection, we report the degradation of accuracy over time in Figure 5 and demonstrate how their respective open-set-aware version provides substantial improvements. Intuitively, this improvement is due to an increase in the visual generation quality; see the examples in the earlier Figure 3.

First, it is apparent how every OpenVAE variant improves upon its non open-set aware counterpart. We further observe that the best version, OpenIntroVAE, appears to be in the same ballpark as complex recent GAN approaches [65,66], even though they do not solve the open-set recognition challenge and conduct a simplified evaluation. The latter works use a lower resolution of 128×128 (we were unable to scale to satisfying results at higher resolution) with additional distillation mechanisms, a continuously trained generator but a classifier that is trained and assessed only once at the end. We nevertheless report the respective values for intuition. We conclude that the obtained final accuracy can be competitive and is remarkably close to the achievable upper bound. A suspected initial VAEs generation quality limitation appears to be lifted with modern extensions and our proposed sampling scheme.

We also support our quantitative statements visually with a few selected generated images for the various generative variants in Figure 6. We emphasize that these examples are supposed to primarily provide visual intuition in support of the existing quantitative results, as it is difficult to draw conclusions from a perceived subjective quality from a few images alone. From a qualitative viewpoint, the OpenVAE without generative modeling extensions appears to suffer from the limitations of a traditional VAE and generates blurry images.

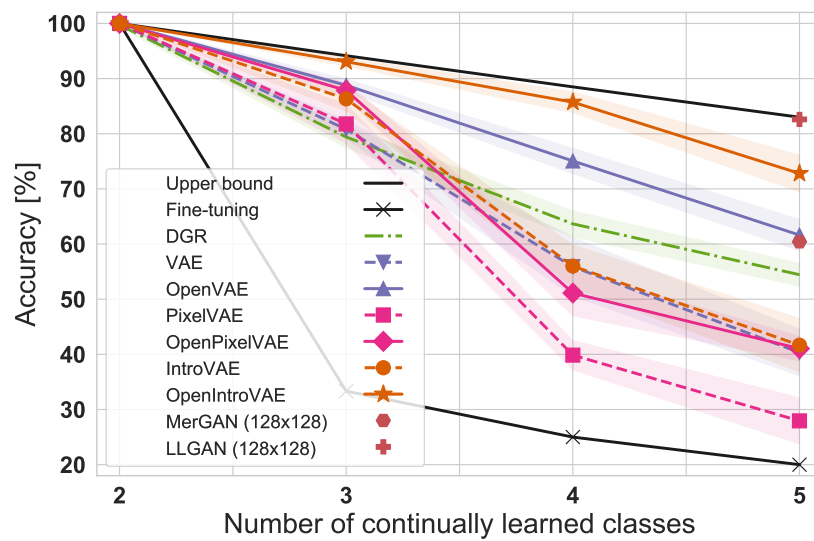


Figure 5. Classification accuracy over five runs for continually learned flowers at 256×256 resolution to demonstrate how generative modeling advances draw similar benefits from our proposed aggregate posterior constrained generative replay (solid lines) over the open-set-unaware baselines (dashed counterparts).

However, our open-set approach nevertheless provides a clearer disambiguation of classes, particularly already at the stage of task 2. The addition of introspection significantly increases the image detail, albeit still degrades considerably due to ambiguous interpolations in samples from low-density areas outside the aggregate posterior. This is again resolved by combining introspection with our proposed posterior-based EVT approach, where image quality is retained across multiple generative replay steps. From a purely visual perspective it is clear why this model outperforms the other approaches significantly in terms of quantitative accuracy values.

Interestingly, our visual inspection also hints at why the PixelVAE and its open-set variant perform much worse than perhaps initially expected. As the caveat is the same in both PixelVAE and OpenPixelVAE, we only show generated instances for the latter. From these samples, we can hypothesize why the initial performance is competitive but rapidly declines. It appears that the autoregression suffers from forgetting in terms of its long-range pixel dependency.

Whereas at the beginning, the information is locally consistent across the entire image, in each consecutive step, a further portion of subsequent pixels for old tasks is progressively replaced with uncorrelated noise. The conditioning thus appears to primarily be captured on new tasks only, resulting in interference effects. We continue this discussion alongside potential other general limitations of generative modeling variant choices in Appendix A.5.



Figure 6. Generated 256×256 flower images for various continually trained models. Images were selected to provide a qualitative intuition behind the quantitative results of Figure 5. Images are compressed for a side-by-side view.

4. Discussion

As a final piece of discussion, we would like to recall and emphasize a few important points of how our results should be interpreted and contextualized.

4.1. Presence of Unknown Data and Current Benchmarks

Perhaps most importantly, we re-iterate that OpenVAE is unique in that it provides a grounded basis to conduct continual learning in the presence of unknown data. However, as evidenced from the quantitative open-set recognition results, the inclusion of unknown data instances into continual learning would immediately result in the failure of the present continual-learning approaches at this point, simply because they lack a principled mechanism to provide robust predictions. For this reason, we show traditional incremental classification results as a proxy to assess our improved aggregate posterior-based generation quality.

Our class incremental accuracy reports in this paper should thus be interpreted with caution as they represent only a part of OpenVAE’s capability, similar to a typical ablation

study. We nevertheless provided this type of comparison, in order to situate OpenVAE with respect to some existing generative continual-learning methods in terms of catastrophic forgetting, rather than presenting OpenVAE in isolation in a more realistic new setting.

4.2. State of the Art in Class Incremental Learning and Exemplar Rehearsal

Following the above subsection, we note that a fair comparison of realistic class incremental learning is further complicated due to various involved factors. In fact, multiple related works make various additional assumptions on the extra storage of explicit data subsets and the use of multiple generative models per task or even multiple classifiers. We do not make these assumptions here in favor of generality. In this spirit, we focused our evaluation on our contributions' relevant novelty with respect to combining the detection of unknown data with the prevention of catastrophic forgetting in generative models.

The introduced OpenVAE shows that both are achievable simultaneously. At the same time, the reader familiar with the recent continual-learning literature will likely notice that some modern approaches that are attributed with state of the art in class incremental learning have not been included in our comparison. These approaches all fall into the category of exemplar rehearsal. We would like to emphasize that this is deliberate and not out of ignorance, as we see these works as purely complementary. We nevertheless wish to give deserved credit to these works and provide an outlook to one future research direction.

The primary reason for omitting a direct comparison with state of the art works in continual learning that employ exemplar rehearsal is that we believe such a comparison would be misleading. In fact, contrasting our OpenVAE against these works would imply that these methods are somehow competing. In reality, exemplar rehearsal, or the so called extraction of core sets, is an auxiliary mechanism that can be applied out-of-the-box to our experimental set-up in this work. The main premise here is that catastrophic forgetting in continual learning can be reduced by retaining an explicit subset of the original data and subsequently continuously interleaving this stored data into the training process.

Early works, such as iCarl [26] show that performance is then a function of two key aspects: the data selection technique and the memory buffer size. The former, selection of an appropriate data subset, essentially boils down to a non-continual-learning question, i.e., how to approximate the entire distribution through only a few instances. Exemplar rehearsal works thus make use of existing techniques here, such as core sets [28], herding [67], nearest mean-classifiers [27] or simply picking data samples uniformly at random [68].

The second question, on memory buffer size, has an almost trivial answer. The larger the memory buffer size, the better the performance. This is intuitive, yet also makes comparison challenging, as a memory buffer of the size of the entire dataset is analogous to what we referred to as "incremental upper bound" in our experiments. If we were to simply store the complete dataset, then catastrophic forgetting would be avoided entirely. Modern class incremental learning works make heavy use of this fact and store large portions of the original data, showing that the more data is stored, the higher the performance goes.

Primary examples include the recent works on Mnemonics Training [69], Contrastive Continual Learning (Co2L) [70] or Dark Experience Replay (DER) [71]. We do not wish to dive into a discussion here of whether or not such data storage is realistic or what size of a memory buffer should be assumed. A respective reference that questions and discusses whether storing of original data is synonymous with progress in continual learning is Greedy Sampler and Dumb Learner (GDumb) [68], where it is shown that the amount of extracted data alone amounts to a significant portion of "state-of-the-art" performance.

Primarily, we point out that the latter works all show that a larger memory buffer shows "better" class incremental learning performance, i.e., less forgetting. However, most importantly, extracting and storing parts of the original data into a separate memory buffer is an auxiliary process that is entirely complementary to our propositions of OpenVAE. As such, each of the methods referenced in this subsection is straightforward to combine with our work. Although we see such a combination as important prospective work, we leave detailed experimentation up to future investigations.

The rationale behind this choice is that inclusion of a memory buffer will inevitably additionally boost the performances of the results of Table 3, yet provide no additional insights to our main hypothesis and contribution: the proposition of OpenVAE to show that detection of unknown data for robust prediction can effectively be achieved alongside reduction of catastrophic forgetting in continual learning.

5. Conclusions

We proposed an approach to unify the prevention of catastrophic interference in continual learning with open-set recognition based on variational inference in deep generative models. As a common denominator, we introduced EVT-based bounds to the aggregate posterior. The correspondingly named OpenVAE was shown to achieve compelling results in being able to distinguish known from unknown data, while boosting the generation quality in continual learning with generative replay.

We believe that our demonstrated benefits from recent generative modeling techniques in the context of high-resolution flower images with OpenVAE provide a natural synergy to be explored in a range of future applications. We envision prospective works to employ OpenVAE as a baseline when relaxing the closed-world assumption in continual learning and allowing unknown data to appear in the investigated benchmark streams at all times in the move to a more realistic evaluation.

Author Contributions: The authors contributed to this work in the following ways: Conceptualization, M.M. and V.R.; methodology, M.M.; software, M.M., I.P. and S.M.; validation, M.M., S.M. and Y.H.; formal analysis, M.M. and I.P.; investigation, M.M.; resources, V.R.; writing—original draft preparation, M.M. and I.P.; writing—review and editing, M.M., I.P. and V.R.; visualization, M.M., I.P. and Y.H.; supervision, V.R.; project administration, M.M.; funding acquisition, V.R. All authors have read and agreed to the published version of the manuscript.

Funding: We acknowledge the influence of various projects in shaping this paper. The projects include EU H2020 Project AEROBI Grant number 687384, EU H2020 Project RESIST Grant number 769066, BMBF project AISEL funding number 01IS19062. Additional financial support from Goethe University was instrumental in concluding the research.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All investigated datasets in this paper are publicly accessible popular benchmarks. Respective citations to the original works are provided in the main body upon first mention of each dataset. For convenience, we briefly list all datasets here and provide the public URL: Modified National Institute of Standards and Technology database (MNIST) [47] (<http://yann.lecun.com/exdb/mnist/>, accessed on 22 January 2022), FashionMNIST [53] (<https://github.com/zalandoresearch/fashion-mnist>, accessed on 22 January 2022), AudioMNIST [54] (<https://github.com/soerenab/AudioMNIST>, accessed on 22 January 2022), Kuzushiji-MNIST (KMNIST) [55] (<http://codh.rois.ac.jp/kmnist/index.html.en>, accessed on 22 January 2022), Street View House Numbers (SVHN) [56] (<http://ufldl.stanford.edu/housenumbers/>, accessed on 22 January 2022), Canadian Institute for Advanced Research (CIFAR) datasets CIFAR10 & CIFAR100 [57] (<https://www.cs.toronto.edu/~kriz/cifar.html>, accessed on 22 January 2022), Oxford Flowers [52] (<https://www.robots.ox.ac.uk/~vgg/data/flowers/>, accessed on 22 January 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

Our appendix provides further details for the material presented in the main body. We first present more in-depth explanations and discussions on the introduced general concepts. At the end of the appendix, we then follow up with a full set of experimental results to complement the investigation of the experimental section. Specifically, the structure is as follows:

- Appendix A.1 Derivation of the lower-bound, Equation (1) of the main body.
- Appendix A.2 Extended discussion, qualitative and quantitative examples for the role of β .
- Appendix A.3 Description of generative model extensions: autoregression and introspection.
- Appendix A.4 The full specification of the training procedure and hyper-parameters, including exact architecture definitions.
- Appendix A.5 Discussion of limitations.

This is followed by full sets of quantitative continual-learning results for all task increments, including reconstruction losses, in part Appendix A.6.

Appendix A.1. Lower-Bound Derivation

As mentioned in the main body of the paper, in supervised continual learning, we are confronted with a dataset $D \equiv \left\{ \left(x^{(n)}, y^{(n)} \right) \right\}_{n=1}^N$, consisting of N pairs of data instances $x^{(n)}$ and their corresponding labels $y^{(n)} \in \{1 \dots C\}$ for C classes. We consider a problem scenario similar to the one introduced in “Auto-Encoding Variational Bayes” [15], i.e., we assume that there exists a data generation process responsible for the creation of the labeled data given some random latent variable z . For simplicity, we follow the authors’ derivation for our model with the additional inclusion of data labels, however, without the β term that is present in the main body. We point to the next section for a discussion of β .

Ideally, we would like to maximize $p(x, y) = \int p(z)p(x, y|z)dz$, where the integral and the true posterior density

$$p(z|x, y) = \frac{p(x, y|z)p(z)}{p(x, y)} \tag{A1}$$

are intractable. We thus follow the standard practice of using variational Bayesian inference and introducing an approximation to the posterior $q(z)$, for which we will specify the exact form later. Making use of the properties of logarithms and applying the above Bayes rule, we can now write:

$$\log p(x, y) = \int q(z)[\log p(x, y|z) + \log p(z) - \log p(z|x, y) + \log q(z) - \log q(z)]dz, \tag{A2}$$

as the left-hand side is independent of z and $\int q(z)dz = 1$. Using the definition of the Kullback–Leibler divergence (KLD) $KL(q || p) = - \int q(x) \log(p(x)/q(x))$ we can rewrite this as:

$$\log p(x, y) - KL(q(z) || p(z|x, y)) = \mathbb{E}_{q(z)}[\log p(x, y|z)] - KL(q(z) || p(z)). \tag{A3}$$

Here, the right hand side forms a variational lower-bound to the joint distribution $p(x, y)$, as the KL divergence between the approximate and true posterior on the left hand side is strictly positive.

At this point, we make two choices that deviate from prior works that made use of labeled data in the context of generative models for semi-supervised learning [72]. We assume a factorization of the generative process of the form $p(x, y, z) = p(x|z)p(y|z)p(z)$ and introduce a dependency of $q(z)$ on x but not explicitly on y , i.e., $q(z|x)$. In contrast to class-conditional generation, this dependency essentially assumes that all information about the label can be captured by the latent z , and there is thus no additional benefit in explicitly providing the label when estimating the data likelihood $p(x|z)$.

This is crucial as our probabilistic encoder should be able to predict labels without requiring it as input to our model, i.e., $q(z|x)$ instead of the intuitive choice of $q(z|x, y)$. However, we would like the label to nevertheless be directly inferable from the latent z . In order for the latter to be achievable, we require the corresponding classifier that learns to predict $p(y|z)$ to be linear in nature. This guarantees linear separability of the classes in latent space, which can in turn then be used for open-set recognition and the generation of specific classes as shown in the main body.

Appendix A.2. Further Discussion on the Role of β

In the main body, the role of the β term [44] in our model’s loss function is summarized briefly. Here, we delve into further detail with qualitative and quantitative examples to support the arguments made by prior works [46,48]. To facilitate the discussion, we repeat Equation (1) of the main body:

$$\mathcal{L}(x^{(n)}, y^{(n)}; \theta, \phi, \xi) = \mathbb{E}_{q_{\theta}(z|x^{(n)})} [\log p_{\phi}(x^{(n)}|z) + \log p_{\xi}(y^{(n)}|z)] - \beta KL(q_{\theta}(z|x^{(n)}) || p(z)) \tag{A4}$$

The β term weights the strength of the regularization by the prior through the Kullback-Leibler (KL) divergence. The selection of this strength is necessary to control the information bottleneck of the latent space and regulate the effective latent encoding overlap. To repeat the main body and previous arguments by [46,48]: too large β values (typically $\gg 1$) will result in a collapse of any structure present in the aggregate posterior. Too small β values (typically $\ll 1$) lead to the latent space being a lookup table. In either case, there is no meaningful information between the latents. This effect is particularly relevant to our objective of linear class separability, which requires the formation of an aggregate latent encoding that is disentangled with respect to the different classes.

To visualize the effect of beta, we trained multiple models with different β values on the MNIST dataset, in an isolated fashion with all data present at all times to focus on the effect of β . The corresponding two-dimensional aggregate encodings at the end of training are shown in Figure A1. Here, we can empirically observe above described phenomenon. With a beta of one and larger, the aggregate posterior’s structure starts to collapse and the aggregate encoding converges to a normal distribution.

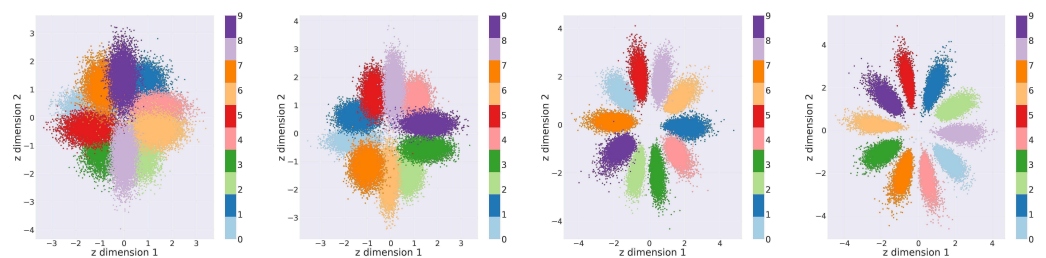


Figure A1. 2-D MNIST latent space visualization with different β values. From left to right, visualization of $\beta = 1.0, 0.5, 0.1, 0.05$.

While this minimizes the distributional mismatch with respect to the prior, the separability of classes is also lost and an accurate classification cannot be achieved. On the other hand, if the beta value becomes ever smaller, there is insufficient regularization present, and the aggregate posterior no longer follows a normal distribution at all. The latter does not only render sampling for generative replay difficult, it also challenges the assumption of distances to each class’ latent mean being Weibull-distributed, as the latter can essentially be seen as a skewed normal.

At this point, it is important to make the following note. Whereas the interpretation of beta always follows the mentioned reasoning, the precise quantitative values of beta can depend heavily on the way losses are treated in practice. In particular, we emphasize that many coding environments, such as PyTorch and TensorFlow (<https://pytorch.org> and <https://www.tensorflow.org>, accessed on 22 January 2022), tend to average losses by default, e.g., normalization of the reconstruction loss term by spatial image dimensionality $I \times I$ and respective normalization of the KLD by the model’s latent dimensionality.

Arguably, the former factor leads to a much larger division than the latter. As such, the natural scale on which the individual reconstruction and KLD losses operate, with the KLD usually being a much much smaller regularization term, can easily be altered. We thus emphasize that the quantitative value of beta should always be regarded in its exact empirical context. In order to provide a quantitative intuition for the role of loss

normalization and its connection to beta, we show examples for the models trained with different β with 2-D latent spaces and 60-D latent spaces in Tables A1 and A2, respectively. In both cases, the losses were normalized by the respective spatial image size and chosen latent dimension, i.e., nats per dimension.

For reference, the un-normalized nats quantities are reported in brackets. We observe that decreasing the value of beta below one is necessary to improve the classification accuracy when the losses are normalized, as well as the overall variational lower bound. Taking the 60 dimensional case as a specific example, we can also observe that reducing the beta value too far and decreasing it from, e.g., 0.1 to 0.05 leads to deterioration of the variational lower bound, from 119.596 to 121.101 natural units, while the classification accuracy by itself does not improve further.

We can see that this is due to the KL divergence residing on the same scale as the normalized reconstruction loss, whereas the latter would typically be much greater when scaled by the image size. Although this may initially appear to render the interpretation of beta more complicated than advocated in the initial work [44], we noticed that a normalized loss appears to come at the advantage of the same beta value of 0.1 consistently yielding the best results across all of our experiments. That is, we can use the same value of beta independently of whether the 28×28 MNIST or the 256×256 flower images are investigated, always with a latent dimensionality of 60.

Table A1. Losses obtained for different β values for MNIST with a 2-D latent space. Training conducted in isolated fashion to quantitatively showcase the role of β . Un-normalized values in nats are reported in brackets for reference purposes.

In Nats per Dimension (Nats in Brackets)					
2-D Latent	Beta	KLD	Recon Loss	Class Loss	Accuracy [%]
train	1.0	1.039 (2.078)	0.237 (185.8)	0.539 (5.39)	79.87
test		1.030 (2.060)	0.235 (184.3)	0.596 (5.96)	78.30
train	0.5	1.406 (2.812)	0.230 (180.4)	0.221 (2.21)	93.88
test		1.382 (2.764)	0.228 (178.8)	0.305 (3.05)	92.07
train	0.1	2.055 (4.110)	0.214 (167.8)	0.042 (0.42)	99.68
test		2.071 (4.142)	0.212 (166.3)	0.116 (1.16)	98.73
train	0.05	2.395 (4.790)	0.208 (163.1)	0.025 (0.25)	99.83
test		2.382 (4.764)	0.206 (161.6)	0.159 (1.59)	98.79

Table A2. Losses obtained for different β values for MNIST with a 60-D latent space. Training conducted in isolated fashion to quantitatively showcase the role of β . Un-normalized values in nats are reported in brackets for reference purposes.

In Nats per Dimension (Nats in Brackets)					
60-D Latent	Beta	KLD	Recon Loss	Class Loss	Accuracy [%]
train	1.0	0.108 (6.480)	0.184 (144.3)	0.0110 (0.110)	99.71
test		0.110 (6.600)	0.181 (142.0)	0.0457 (0.457)	99.03
train	0.5	0.151 (9.060)	0.162 (127.1)	0.0052 (0.052)	99.87
test		0.156 (9.360)	0.159 (124.7)	0.0451 (0.451)	99.14
train	0.1	0.346 (20.76)	0.124 (97.22)	0.0022 (0.022)	99.95
test		0.342 (20.52)	0.126 (98.79)	0.0286 (0.286)	99.38
train	0.05	0.476 (28.56)	0.115 (90.16)	0.0018 (0.018)	99.95
test		0.471 (28.26)	0.118 (92.53)	0.0311 (0.311)	99.34

Appendix A.3. Complementary Generative Modelling Advances

At the time of their initial introduction, it was notorious that variational autoencoders produced blurry examples and were associated with an inability to scale to more complex

high-resolution color images. This is in contrast to their prominent generative counterparts, the generative adversarial network [33]. Although this stigma perhaps still holds until today, there have been many successful recent efforts to address this challenge.

In our final outlook in the main body, we thus empirically showcased the impact of generative modeling advances with their optional improvements in two promising research directions: autoregression [2,17,18] and introspection [19,20]. The commonality between these approaches is their aim to overcome the limitations of independent pixel-wise reconstructions. In this appendix section, we briefly summarize the foundation of these generative extensions.

Appendix A.3.1. Improvements through Autoregressive Decoding

In essence, autoregressive models improve the probabilistic decoder through a spatial conditioning of each scalar output value on the previous ones, in addition to conditioning on the latent variable:

$$p(x|z) = \prod_i p(x_i|x_1, \dots, x_{i-1}, z) \tag{A5}$$

In an image, generation thus needs to proceed pixel by pixel and is commonly referred to as PixelVAE [18]. This conditioning is generally achieved by providing the input to the decoder during training, i.e., including a skip path that bypasses the probabilistic encoding. A concurrent introduction of autoregressive VAEs thus coined this model “lossy” [2]. This is because local information can now be increasingly modeled without access to the latent variable, and the encoding of z can focus on the global information.

Although the main body’s accuracies of generative replay with autoregression are assuring in the MNIST scenario, autoregressive sampling comes with a major caveat. When attempting to operate on larger data, the computational complexity of the pixel by pixel data creation procedure grows in direct proportion to the input dimensionality. With increasing input size, the repeated calculation of the autoregressive decoder layers can thus rapidly render the generation practically infeasible.

Appendix A.3.2. Introspection and Adversarial Training

A promising alternative perspective towards autoencoding beyond pixel similarities is to leverage the insights obtained from generative adversarial networks (GAN). To this matter, Larsen et al. [73] proposed a hybrid model called VAE-GAN. Here, the crucial idea is to append a GAN style adversarial discriminator to the variational autoencoder. This yields a model that promises to overcome a conventional GAN’s mode collapse issues, as the VAE is responsible for the rich encoding, while letting the added discriminator judge the decoder’s output based on perceptual criteria rather than individual pixel values.

The more recent IntroVAE [19] and adversarial encoder generator networks [20] have subsequently come to the realization that this does not necessarily require the auxiliary real-fake discriminator, as the VAE itself already provides strong means for discrimination, namely its probabilistic encoder. We leverage this idea of introspection for our framework, as it does not require any architectural or structural changes beyond an additional term in the loss function.

For sake of brevity, we denote the probabilistic encoder through the parameters ϕ and decoder θ in the following equations. Training our model with introspection is then equivalent to adding the following two terms to our previously formulated loss function:

$$\mathcal{L}_{IntroVAE_Enc} = \mathcal{L}_{VAE} - \beta[m - KL(\theta(\phi(z)) || p(z))]^+ \tag{A6}$$

and

$$\mathcal{L}_{IntroVAE_Dec} = \mathcal{L}_{Rec} - \beta KL(\theta(\phi(z)) || p(z)). \tag{A7}$$

Here, \mathcal{L}_{VAE} corresponds to the full loss of the main body’s Equation (1) and \mathcal{L}_{Rec} corresponds to the reconstruction loss portion: $\mathbb{E}_{q_{\theta}(z|x^{(n)})}[\log p_{\phi}(x^{(n)}|z)]$. In the above equations, we followed the original authors’ proposal to include a positive margin m ,

with $[\cdot]$ denoting $\max(0, \cdot)$. This hinge loss formulation serves the purpose of empirically limiting the encoder's reward to avoid a too massive gap in a min–max game of above competing KL terms.

Aside from the regular loss that encourages the encoder to match the approximate posterior to the prior for real data, the encoder is now further driven to maximize the deviation from the posterior to the prior for generated images. Conversely, the decoder is encouraged to “fool” the encoder into producing a posterior distribution that matches the prior for these generated images. The optimization is conducted jointly. In comparison with a traditional VAE, this can thus be seen as training in an adversarial-like manner, without necessitating additional discriminative models. As such, introspection fits naturally into our OpenVAE, and no further changes are required.

Appendix A.4. Training Hyper-Parameters and Architecture Definitions

In this section, we provide a full specification of hyper-parameters, model architectures and the training procedure used in the main body.

Architecture

For our MNIST style continual-learning experiments, we report both a simple multi-layer perceptron architecture, as well as a deeper wide residual network variant. For the former, we follow previous continual-learning studies and employ a two-hidden-layer and 400-unit multi-layer perceptron [60]. For the latter, we base our encoder and decoder architecture on 14-layer wide residual networks [61,62] with a latent dimensionality of 60 to demonstrate scalability to high-dimensions and as used in lossy auto-encoders [2,18]. Our main body's reported out-of-distribution detection experiments are all based on this WRN architecture.

For a common frame of reference, all methods share the same underlying WRN architecture, including the investigated separate classifiers (for OpenMax) and generative models of the reported dual model approaches. All hidden layers include batch-normalization [74] with a value of 10^{-5} and use rectified linear unit (ReLU) activations. A detailed list of the architectural components is provided in Tables A3 and A4. For the higher resolution 256×256 flower images, we used a deeper 26-layer WRN version, in analogy to previous works [2,18]. Here, the last encoder and first decoder blocks are repeated an extra three times, resulting in an additional three stages of down- and up-sampling by a factor of two. The encoder's spatial output dimensionality is thus equivalent to the 14-layer architecture applied to the eight-times lower resolution images of the simpler datasets.

Autoregression

For the autoregressive variant, we set the number of output channels of the decoder to 60 and append three additional pixel decoder layers, each with a kernel size of 7×7 . We use 60 channels in each autoregressive layer for the MNIST dataset and 256 for the more complex flower data [2,18]. Whereas we report reconstruction log-likelihoods in natural units (nats) in the upcoming detailed supplementary results (recall that we have only shown quantitative validation of the model through the proxy of continual classification in the main body), these models are practically formulated as a classification problem with a 256-way Softmax. The corresponding loss is in bits per dimension.

We converted these values to have a better comparison; however, in order to do so, we need to sample from the pixel decoder's multinomial distribution to calculate a binary cross-entropy on reconstructed images. We further note that all losses are normalized with respect to the spatial and latent dimensions, as mentioned in the prior appendix section, which explained normalization in the context of the role of beta.

Table A3. A 14-layer wide residual network (WRN) encoder with a widen factor of 10. Convolutional layers (conv) are parametrized by a quadratic filter size followed by the amount of filters. p and s represent zero padding and stride, respectively. If no padding or stride is specified, then $p = 0$ and $s = 1$. Skip connections are an additional operation at a layer, with the layer to be skipped specified in brackets. Convolutional layers are followed by batch-normalization and a rectified linear unit (ReLU) activation. The probabilistic encoder ends on fully-connected layers for μ and σ that depend on the chosen latent space dimensionality and the data's spatial size.

Layer Type	WRN Encoder	
Layer 1	conv 3×3 —48, $p = 1$	
Block 1	conv 3×3 —160, $p = 1$; conv 3×3 —160, $p = 1$; conv 3×3 —160, $p = 1$; conv 3×3 —160, $p = 1$	conv 1×1 —160 (skip next layer) shortcut (skip next layer)
Block 2	conv 3×3 —320, $s = 2$, $p = 1$; conv 3×3 —320, $p = 1$; conv 3×3 —320, $p = 1$; conv 3×3 —320, $p = 1$	conv 1×1 —320, $s = 2$ (skip next layer) shortcut (skip next layer)
Block 3	conv 3×3 —640, $s = 2$, $p = 1$; conv 3×3 —640, $p = 1$; conv 3×3 —640, $p = 1$; conv 3×3 —640, $p = 1$	conv 1×1 —640, $s = 2$ (skip next layer) shortcut (skip next layer)

Table A4. A 14-layer WRN decoder with a widen factor of 10. P_w and P_h refer to the input's spatial dimension. Convolutional (conv) and transposed convolutional (conv_t) layers are parametrized by a quadratic filter size followed by the amount of filters. p and s represent zero padding and stride, respectively. If no padding or stride is specified, then $p = 0$ and $s = 1$. Skip connections are an additional operation at a layer, with the layer to be skipped specified in brackets. Every convolutional and fully-connected (FC) layer is followed by batch-normalization and a rectified linear unit (ReLU) activation function. The model ends on a Sigmoid function.

Layer Type	WRN Decoder	
Layer 1	FC $640 \times \lfloor P_w/4 \rfloor \times \lfloor P_h/4 \rfloor$	
Block 1	conv_t 3×3 —320, $p = 1$; conv 3×3 —320, $p = 1$; conv 3×3 —320, $p = 1$; conv 3×3 —320, $p = 1$; upsample $\times 2$	conv_t 1×1 —320 (skip next layer) shortcut (skip next layer)
Block 2	conv_t 3×3 —160, $p = 1$; conv 3×3 —160, $p = 1$; conv 3×3 —160, $p = 1$; conv 3×3 —160, $p = 1$; upsample $\times 2$	conv_t 1×1 —160 (skip next layer) shortcut (skip next layer)
Block 3	conv_t 3×3 —48, $p = 1$; conv 3×3 —48, $p = 1$; conv 3×3 —48, $p = 1$; conv 3×3 —48, $p = 1$	conv_t 1×1 —48 (skip next layer) shortcut (skip next layer)
Layer 2	conv 3×3 —3, $p = 1$	

Introspection

For the introspective model variant, we note that the original authors of IntroVAE introduced additional weighting terms in front of the reconstruction loss, in order to drastically lower its magnitude and the added KL divergence. We observed that the former is simply due to lack of normalization with respect to input width and height and hence

the reconstruction loss growing proportionally with the spatial input size, whereas the *KL* divergence typically does not reflect this behavior for a fixed-size latent space.

Given that we average the loss over the image size in our practical experimentation, we found this additional hyper-parameter to be unnecessary. The other hyper-parameter to weight the added adversarial *KL* divergence term is essentially equivalent to the already introduced beta, alas without our motivation in earlier sections but simply as a heuristic so as to not overpower the reconstruction loss. The introspective variant of our OpenVAE thus does not introduce any additional hyper-parameters or architectural modifications beyond the additional loss term, as summarized in the previous section.

Stochastic Gradient Descent (SGD)

Optimization parameters were used in consistence with the literature [2,18]. Accordingly, all models are optimized using stochastic gradient descent with a mini-batch size of 128 and Adam [16] with a learning rate of 0.001 and first and second momenta equal to 0.9 and 0.999. For MNIST, FashionMNIST and AudioMNIST, no data augmentation or preprocessing is applied. For the flower experiments, images are stochastically flipped horizontally with a 50% chance and the mini-batch size is reduced to 32. We initialize all weights according to He et al. [75]. All class incremental models were trained for 120 epochs per task, except for the flower experiment. While our investigated single model exhibits representational transfer due to weight sharing and need not necessarily be trained for the entire amount of epochs for each subsequent task, this guarantees convergence and a fair comparison of results with respect to the achievable accuracy of other methods.

Due to the much smaller dataset size, architectures were trained for 2000 epochs on the flower images, in order to obtain a similar amount of update iterations. For the generative replay with statistical outlier rejection, we used an aggressive rejection rate of $\Omega_t = 0.01$ (although we obtain almost analogous results with 0.05) and dynamically set tail-sizes to 5% of seen examples per class. As mentioned in the main body, the used open-set distance measure was the cosine distance.

To enable our out-of-distribution detection experiments in the main body—which includes comparison with datasets, such as CIFAR—we resized all images to 32×32 . Recall that we also made use of the AudioMNIST dataset in the main body to showcase the challenge in open-set recognition, where most approaches fail to recognize Audio data as out-of-distribution, even though its form is entirely different from the commonly observed object-centric images. To make this comparison possible, we followed the original dataset's authors and use the described Fourier transform on the audio data to obtain frequency images.

De-Quantization, Overfitting and Data Augmentation

As the autoregressive model variants require a de-quantization of the input (to transform the discrete eight-unit input into a continuous distribution) [2,18], we employed a denoising procedure on the input. Specifically, for our MNIST-like gray-scale datasets, we add noise to the input, sampled from a normal distribution with mean 0 and variance 0.25. As typical for denoising autoencoders, the reconstruction loss nevertheless aims to recover the unperturbed original input.

As the latter could be argued to serve an additional data augmentation effect (we always observed an improvement), we adopted the denoising procedure for all models, even if no autoregression was used. In this way, a fair comparison was enabled. For the colored high-resolution flower images, such gray-scale Gaussian noise seems less meaningful. Here, we realize that our primary interest lies in maintaining the discriminative performance of our model and less so on the visual quality of the generated data.

We can thus take advantage of the de-quantization perturbation distribution as means to encode our prior knowledge of common generative pitfalls. In our specific context, it is well known that a traditional VAE without further advances commonly fails to generate non-blurry, crisp images. However, we can include and work around this belief by letting

the denoising assume the form of de-blurring, e.g., by stochastically adding a varying Gaussian blur to inputs (as done in [4]).

Even though the decoder is ultimately still encouraged to remove this blur and reconstruct the original clean image, the encoder is now inherently required to learn how to manage blurry input. It is encouraged to build up a natural invariance to our choice of perturbation. In the context of maintaining a classifier with generative replay, to an extent, it should then no longer be a strict requirement to replay locally detailed crisp images, as long as the information required for discrimination is present.

Hardware and Software

All models were trained on single GeForce GTX 1080 (Nvidia, Santa Clara, CA, USA) graphics processing units (GPU), with the exception of the high-resolution flower image experiments, where we used a single V100 GPU (Nvidia, Santa Clara, CA, USA) per experiment. Our implementation is based on *PyTorch* (<https://pytorch.org>, last accessed on 22 January 2022), including data loading functionality for the majority of the investigated public datasets through the *torchvision* library. The AudioMNIST data was preprocessed using the *librosa* Python library (<https://librosa.org/doc/latest/index.html>, last accessed on 22 January 2022), following the setting of the original AudioMNIST dataset authors [54]. Our code will be publicly available.

Elastic Weight Consolidation (EWC)

Recall that for related work approaches, we reported quantitative values found in the literature if our reproduction matched or did not surpass this number or, conversely, our obtained value if it turned out to be better. Primarily, the latter discrepancy was the case for our reproduction of the EWC experiments on FashionMNIST, where we obtained marginally improved results. Here, the number of Fisher samples was fixed to the total number of data points from all the previously seen tasks. A suitable Fisher multiplier value λ was determined by conducting a grid search over a set of five values: 50, 100, 500, 1000 and 5000 on held-out validation data for the first two tasks in sequence. We observed exploding gradients if λ was too high. However, a very small λ lead to excessive drift in the weight distribution across subsequent tasks that further resulted in catastrophic interference. Empirically, $\lambda = 500$ seemed to provide the best balance.

Appendix A.5. Limitations

We believe that there are three main sources for limitations of our work. Some of these are general to the considered continual-learning scenarios, whereas others are more directly related to our proposed approach. Specifically, we can group the limitations into:

1. Limitations of our proposed aggregate posterior-based EVT approach and its use for open-set recognition and generative replay.
2. Limitations of the employed generative model variant, i.e., caveats of autoregression or introspection.
3. Limitations in terms of obtainable insights from investigated scenarios, i.e., the specific continual-learning set-up.

As only the first point is concerned with immediate limitations of our method, we provide the most detailed discussion of these aspects first, before providing a small overview of conceivable general limitations that are less specific to the contributions of our paper but, nevertheless, of potential interest.

Appendix A.5.1. Aggregate Posterior-Based EVT Limitations

There are several imaginable caveats to our method, some of which are of theoretical nature and have not yet been observed in our experiments and some of which the reader should be aware of when reproducing our experiments. The two main caveats we surmise are: the assumed unimodality of the latent space distance distribution that forms the basis

for the Weibull based EVT approach and the necessity for a “burn-in” phase at the start of training.

Distance distribution unimodality: Recall that we make use of extreme value theory by fitting a Weibull distribution based on the distances to the practically obtained aggregate posterior of our joint model. This was motivated from a direct mathematical expression for the aggregate posterior being cumbersome to obtain, as the latter can in principle be arbitrarily complex. To circumvent this challenge, we imposed a linear separation of classes through the use of a linear classifier on the latent variable z and subsequent treatment of the aggregate posterior-based on each individual class. As such, we obtained the distances to the mean of the aggregated encoding for each class. As a result, a Weibull distribution for statistical outlier rejection was crafted, with one mode per class. While this Weibull distribution can be multivariate depending on the choice of distance distribution, e.g., a cosine distance would collapse latent vectors into scalar distances and a euclidean distance could preserve dimensionality, each class is nevertheless assumed to have a single distance mean and thus a single mode of the Weibull distribution. This forms a theoretical limitation of our approach, although not presently observed to hinder our practical application.

Concretely, we assume that the distances to the mean can be described by a single mean of the distribution, a single variance and a shift parameter. Intuitively, this limits the applicability of the approach, should the obtained aggregate posterior per class remain more complex in terms of forming multiple clusters within a class. However, we also note that a unimodality in terms of distance distribution is not analogous to the existence of a single well-formed cluster in the aggregate posterior. For instance, if a class were to be symmetrically distributed around a low-density region, think of a donut for example, the three parameter distribution on the distances to the center would still capture this through a single mode. A limitation would arise if multiple clusters within a class were to form without the presence of such symmetry. We have not yet observed the latter in practice; however, we note that it presently marks a theoretical limitation of our specified approach.

“Burn-in” phase: Our obtained Weibull distance distribution based on the aggregate posterior is obtained as a “meta-recognition” module. That is, the Weibull distribution is not trained but is a derived quantity of the aggregate posterior. In order to form the basis for statistical rejection of outliers and constraining generation to inliers an initial “burn-in” phase needs to exist to first obtain some meaningful approximate posterior estimate. At this point, it could be argued whether this is a limitation or not. In principle, any deep neural network model arguably has to undergo an initial stage of training before its representations can be leveraged. As one of the main contributions in our paper is improved open-set recognition, we believe this aspect is nevertheless important to mention. Consequently, our trained model and aggregate posterior-based open-set-recognition mechanism enable robust application of a trained model or, as demonstrated, its continual learning. In the very first epochs of training, it is however presently expected that the data reflects the true task and does not contain potentially corrupted inputs, as a notion of “in-distribution” first needs to be built up. Even though we mention this as a limitation, we also note that we are unaware of a deep-neural-network-based approach that would not have an initial training phase as a necessity in supervised learning.

Appendix A.5.2. Limitations of the Employed Generative Model Variants

We investigated modern generative modeling advances that build on top of the conventional VAE. The primary purpose was to demonstrate that the notorious blurriness of VAE generations can easily be overcome based on recent insights. As such, our proposed approach was shown in the context of more complex high-resolution color images, with recent generative modeling advances being shown to draw similar benefits from our proposed open-set mechanism. Two investigated variants were autoregression (PixelVAE) and introspection (IntroVAE). Although neither of these formulations is our contribution, nor a key to our proposed formulation, we provide a brief description of their limitations in a continual-learning context.

Autoregression: During training, autoregression does not initially appear to come with significant caveats beyond the perhaps added computational overhead of using larger sized convolutions to capture more local context (e.g., 7×7 convolutions in contrast to typically employed stacks of 3×3 kernels). The conditioning on pixel values during training is usually practically achieved through masking operations, enabling training on a similar time scale to non-autoregressive counterparts. However, for the generation of actual images, pixel values need to be sampled sequentially, much in contrast to a typical VAE calculating the entire decoder in a single-shot pass. As such, the time it takes to continually learn on later tasks, where old information is rehearsed based on generative examples, comes with an unfortunate increase in required computation time taken up for autoregressive sampling. For plain VAEs, or the IntroVAE variant, this is not usually a problem as generation typically takes significantly less time than training with backpropagation. For autoregressive sampling in its sequential formulation, this is quickly no longer the case.

Introspection: Fortunately, introspection does not come with a similar computational caveat as autoregression. In contrast to a conventional VAE, the computational overhead lies in one additional pass through the encoder for each update to compute the adversarial term of generated examples. As such, this computational increase is a less severe drawback. A perhaps more significant caveat is the presence of the min–max objective potentially rendering the training more difficult in terms of finding a good point of convergence. In other words, because the losses are balanced in the adversarial game, finding a satisfactory end point can often be subject to subjectively being satisfied with the visual perceived generation quality. Whereas the IntroVAE benefits greatly from stability (in contrast to the often observed collapse in pure GANs) and is thus typically trained for extended periods of time, this could also mark a conceivably difficult trade-off when deciding when to continue optimizing the next task in continual learning while simultaneously minimizing total training time.

Appendix A.5.3. Limitations of the Investigated Scenarios

The key contribution of our paper is in showing how a principled single mechanism can be used in a single model to unify the prevention of catastrophic interference in continual learning with open-set recognition. The present formulation of our paper investigates these aspects in two experimental subsections, to adequately showcase the benefits from each perspective. In retrospect, while this is the initially stated motivation, it is clear that the investigated challenges are actually part of a greater theme towards robust continual learning.

As there is no immediate literature to compare with, we decided that the presented empirical analysis of the main body would provide the most immediate benefit to the reader. We do however also note that this is a more general limitation of predominant investigation in the literature. Here, our work provides the first steps in the direction of a more meaningful investigation of continual learning, for instance, where task scenarios are not pre-defined to contain clear-cut boundaries. Our approach has demonstrated that it is possible to accurately identify when the distribution experiences a major disruption in the process of learning continually. Future investigation should thus lift persisting limitations of the rigid investigation and consider scenarios where tasks are not always introduced at a known point in time.

Appendix A.6. Full Continual Learning Results for All Intermediate Steps

In the main body, we reported the final classification accuracy at the end of multiple task increments in continual learning. In this section, we provide a full list of intermediate results and a two-fold extension to the reporting.

First, instead of purely reporting the final obtained accuracy value, we follow prior work [60] and report more nuanced accuracy metrics: the base task's accuracy over time $\alpha_{t,base}$, the new task's accuracy $\alpha_{t,new}$ and the overall accuracy at any point in time $\alpha_{t,all}$. The first of these, the "base" metric, reports the accuracy of the initial first tasks, e.g., digits

0 and 1 in MNIST and their accuracy degradation over time when the data are no longer available when subsequent tasks arrive. Conversely, the “new” metric always portrays only the accuracy of the most recent task, in independence of the other existing classes.

Finally, the “all” accuracy is the accuracy averaged over all presently existing tasks. For instance, the final accuracy reported in the main body is thus this overall metric at the end of observing all tasks. This is a more appropriate way to evaluate the quality of the model over time. Given that the employed mechanism to avoid catastrophic interference in continual learning is generative replay, it also gives us further insight into whether an accuracy degradation is due to old tasks being forgotten, i.e., catastrophic interference occurring because the decoder-sampled data will no longer resemble the instances of the observed data distribution or the encoder not being able to encode further new knowledge.

Second, we report respective three metric variants for all intermediate steps of our models for the data negative log-likelihood (NLL). Here, we note that it is particularly important to see the new, base and all metrics in conjunction, as each individual task does not have the same level of difficulty and measures up differently in terms of quantitative NLL values. Nevertheless, the initial task’s degradation can similarly be monitored and the overall value at any point in time gives us a direct mean for comparison across models.

In addition to the values reported in the main body, we report the detailed full set of intermediate results for the five task steps of the class incremental scenarios in Tables A5–A7. The upper bound (UB) and fine tuning (FT) (tuning on only the most recent task) are again reported for reference. We have now also included a variant of DGR, i.e., a dual model approach with separate generative and separate discriminative model, where the generative model is based on the autoregressive PixelVAE. We omitted the latter result from the main body as the insight is analogous to that of the non-autoregressive comparison (and comparison to a GAN as the generative model).

The joint open-set model variant appears to have the upper hand, in addition to being able to solve the open-set recognition task. In general, once more, we can observe the increased effect of error accumulation due to unconstrained generative sampling from the prior in comparison to the open-set counterpart that limits sampling to the aggregate posterior. The statistical deviations across experiment repetitions in the base and the overall classification accuracies are higher and are generally decreased by the open-set models. For example, in Table A5 the MNIST base and overall accuracy deviations of a naive supervised variational auto-encoder (SupVAE) are higher than the respective values for OpenVAE, starting already from the second task increment.

Correspondingly, the accuracy values themselves experience larger decline for SupVAE than for OpenVAE with progressive increments. This difference is not as pronounced at the end of the first task increment because the models have not been trained on any of their own generated data yet. Successful literature approaches, such as the variational generative replay proposed by [34], thus avoid repeated learning based on previous generated examples and simply store and retain a separate generative model for each task.

The strength of our model is that, instead of storing a trained model for each task increment, we are able to continually keep training our joint model with data generated for all previously seen tasks by filtering out ambiguous samples from low-density areas of the posterior. Similar trends can also be observed for the respective pixel models. Interestingly, the audio dataset also appears to be a prime example to advocate the necessity of a single joint model, rather than maintenance of multiple models as proposed in DGR [32] or VGR [34]. If we look carefully at the averaged “all” accuracy values of our OpenVAE model, we can see that the accuracy between tasks 2 and 3 and, similarly, tasks 4 and 5, first decreases and then increases again. In other words, due to the shared nature of the representations, learning later tasks brings benefits to the solution of already learned former tasks.

Such a form of “backward transfer” is difficult to obtain, if not even impossible, in approaches that maintain multiple separated models or even regularization approaches that discourage retrospective change of older representations. We believe the possibility to retrospectively improve older representations to be an additional strength of our approach,

where the benefits of a shared representation single model of generative nature become even more evident.

With respect to the obtained negative log-likelihoods we can make two observations. First, by themselves, the small relative improvements between models should be interpreted with caution as they do not directly translate to maintaining continual-learning accuracy. Second, we can also observe that, at every increment for all $\gamma_{t,all}$ and respective quantities for only the new task $\gamma_{t,new}$, negative log-likelihoods are more difficult to interpret compared with the accuracy counterpart. While the latter is normalized between zero and unity, the NLL of different tasks is expected to fluctuate largely according to the task's images' complexity.

To give a concrete example, it is rather straightforward to come to the conclusion that a model suffers from limited capacity or lack of complexity if a single newly arriving class cannot be classified well. In the case of NLL, it is common to observe either a large decrease for the newly arriving class or a large increase depending on the specifically introduced class. As such, these values are naturally comparable between models but are challenging to interpret across time steps without also analyzing the underlying nature of the introduced class.

The exception is formed by the base task's $\gamma_{t,base}$. In analogy to base classification accuracy, this quantity still measures the amount of catastrophic interference across time. However, in all tables we can observe that catastrophic interference is almost imperceptible. As this is not at all reflected in the respective accuracy over time, it further underlines our previous arguments that NLL is not necessarily the best metric to monitor in the presented continual-learning scenario, with the classification proxy seemingly providing a better indicator of continual generative model degradation.

Table A5. The results for class incremental continual-learning approaches averaged over five runs, baselines and the reference isolated learning scenario for MNIST at the end of every task increment. This is an extension of Table 3 in the main body. Here, in addition to the accuracy α_t , γ_t also indicates the respective negative log-likelihood (NLL) at the end of every task increment t .

MNIST	t	UB	FT	SupVAE	OpenVAE	PixelVAE DGR	SupPixelVAE	OpenPixelVAE
$\alpha_{base,t}$ (%)	1	100.0	100.0	99.97 ± 0.029	99.98 ± 0.018	99.97 ± 0.002	99.97 ± 0.026	99.86 ± 0.084
	2	99.82	00.00	97.28 ± 3.184	99.30 ± 0.100	99.54 ± 0.285	96.90 ± 2.907	99.64 ± 0.095
	3	99.80	00.00	87.66 ± 8.765	96.69 ± 2.173	99.16 ± 0.611	90.12 ± 5.846	98.88 ± 0.491
	4	99.85	00.00	54.70 ± 22.84	94.71 ± 1.792	98.33 ± 1.119	76.84 ± 9.095	98.11 ± 0.797
	5	99.57	00.00	19.86 ± 7.396	92.53 ± 4.485	98.04 ± 1.397	56.53 ± 4.032	97.44 ± 0.785
$\alpha_{new,t}$ (%)	1	100.0	100.0	99.97 ± 0.029	99.98 ± 0.018	99.97 ± 0.002	99.97 ± 0.026	99.86 ± 0.084
	2	99.80	99.85	99.75 ± 0.127	99.80 ± 0.126	99.71 ± 0.122	99.74 ± 0.052	99.82 ± 0.027
	3	99.67	99.94	99.63 ± 0.172	99.61 ± 0.055	99.41 ± 0.084	99.22 ± 0.082	99.56 ± 0.092
	4	99.49	100.0	99.05 ± 0.470	99.15 ± 0.032	98.61 ± 0.312	97.84 ± 0.180	98.80 ± 0.292
	5	99.10	99.86	99.00 ± 0.100	99.06 ± 0.171	97.31 ± 0.575	96.77 ± 0.337	98.63 ± 0.430
$\alpha_{all,t}$ (%)	1	100.0	100.0	99.97 ± 0.029	99.98 ± 0.018	99.97 ± 0.002	99.97 ± 0.026	99.86 ± 0.084
	2	99.81	49.92	98.54 ± 1.638	99.55 ± 0.036	99.60 ± 0.142	98.37 ± 1.448	99.69 ± 0.051
	3	99.72	31.35	95.01 ± 3.162	98.46 ± 0.903	98.93 ± 0.291	96.14 ± 1.836	99.20 ± 0.057
	4	99.50	24.82	81.50 ± 9.369	97.06 ± 1.069	98.22 ± 0.560	91.25 ± 0.992	98.13 ± 0.281
	5	99.29	20.16	64.34 ± 4.903	93.24 ± 3.742	96.52 ± 0.658	83.61 ± 0.927	96.84 ± 0.346
$\gamma_{base,t}$ (nats)	1	63.18	62.08	64.34 ± 2.054	62.53 ± 1.166	90.52 ± 0.263	100.0 ± 1.572	99.77 ± 2.768
	2	62.85	126.8	74.41 ± 10.89	65.68 ± 1.166	91.27 ± 0.789	100.4 ± 1.964	101.2 ± 3.601
	3	63.36	160.4	81.89 ± 10.09	69.29 ± 1.541	91.92 ± 0.991	100.3 ± 4.562	101.1 ± 4.014
	4	64.25	126.9	90.62 ± 10.08	71.69 ± 1.379	91.75 ± 1.136	102.7 ± 7.134	101.0 ± 4.573
	5	64.99	123.2	101.6 ± 8.347	77.16 ± 1.104	92.05 ± 1.212	102.4 ± 6.195	100.5 ± 4.942

Table A5. Cont.

MNIST	t	UB	FT	SupVAE	OpenVAE	PixelVAE DGR	SupPixelVAE	OpenPixelVAE
$\gamma_{new,t}$ (nats)	1	63.18	62.08	64.34 ± 2.054	62.53 ± 1.166	90.52 ± 0.263	100.0 ± 1.572	99.77 ± 2.768
	2	88.75	87.93	89.91 ± 0.107	89.64 ± 3.709	115.8 ± 0.805	125.7 ± 2.413	124.6 ± 3.822
	3	82.53	87.22	87.65 ± 0.530	85.37 ± 1.725	107.7 ± 0.600	118.3 ± 3.523	116.5 ± 2.219
	4	72.68	74.61	79.49 ± 0.489	74.75 ± 0.777	100.9 ± 0.659	107.1 ± 5.316	102.3 ± 1.844
	5	85.88	92.00	93.55 ± 0.391	89.68 ± 0.618	113.4 ± 0.820	118.2 ± 1.572	113.3 ± 0.755
$\gamma_{all,t}$ (nats)	1	63.18	62.08	64.34 ± 2.054	62.53 ± 1.166	90.52 ± 0.263	100.0 ± 1.572	99.77 ± 2.768
	2	75.97	107.3	82.02 ± 5.488	76.62 ± 1.695	102.9 ± 0.408	111.9 ± 2.627	112.7 ± 3.300
	3	79.58	172.3	89.88 ± 3.172	82.95 ± 1.878	104.8 ± 1.114	114.9 ± 4.590	114.6 ± 4.788
	4	79.72	203.1	95.83 ± 2.747	85.30 ± 1.524	103.9 ± 0.759	114.3 ± 3.963	112.1 ± 2.150
	5	81.97	163.7	107.6 ± 1.724	92.92 ± 2.283	106.1 ± 0.868	118.7 ± 5.320	111.9 ± 2.663

Table A6. The results for class incremental continual-learning approaches averaged over five runs, baselines and the reference isolated learning scenario for FashionMNIST at the end of every task increment. This is an extension of Table 3 in the main body. Here, in addition to the accuracy α_t , γ_t also indicates the respective NLL at the end of every task increment t .

Fashion	t	UB	FT	SupVAE	OpenVAE	PixelVAE DGR	SupPixelVAE	OpenPixelVAE
$\alpha_{base,t}$ (%)	1	99.65	99.60	99.55 ± 0.035	99.59 ± 0.082	99.57 ± 0.091	99.58 ± 0.076	99.54 ± 0.079
	2	96.70	00.00	92.02 ± 1.175	92.36 ± 2.092	82.40 ± 6.688	90.06 ± 1.782	88.60 ± 1.998
	3	95.95	00.00	79.26 ± 4.170	83.90 ± 2.310	78.55 ± 3.964	83.70 ± 3.571	87.66 ± 0.375
	4	91.35	00.00	50.16 ± 6.658	64.70 ± 2.580	54.69 ± 3.853	50.23 ± 7.004	68.31 ± 3.308
	5	92.20	00.00	39.51 ± 7.173	60.63 ± 12.16	60.04 ± 5.151	47.83 ± 13.41	74.45 ± 2.889
$\alpha_{new,t}$ (%)	1	99.65	99.60	99.55 ± 0.035	99.59 ± 0.082	99.57 ± 0.091	99.58 ± 0.076	99.54 ± 0.079
	2	95.55	97.95	90.98 ± 0.626	92.64 ± 2.302	97.73 ± 1.113	96.47 ± 0.596	97.31 ± 0.475
	3	93.35	99.95	90.26 ± 1.435	83.40 ± 3.089	99.09 ± 0.367	97.33 ± 0.725	96.88 ± 1.156
	4	84.75	99.90	85.65 ± 2.127	84.18 ± 2.715	97.55 ± 0.588	96.12 ± 0.675	95.47 ± 1.332
	5	97.50	99.80	96.92 ± 0.774	96.51 ± 0.707	98.85 ± 0.141	97.91 ± 0.596	98.63 ± 0.176
$\alpha_{all,t}$ (%)	1	99.65	99.60	99.55 ± 0.035	99.59 ± 0.082	99.57 ± 0.091	99.58 ± 0.076	99.54 ± 0.079
	2	95.75	48.97	91.83 ± 0.730	92.31 ± 1.163	86.22 ± 3.704	92.93 ± 0.160	92.17 ± 1.425
	3	93.02	33.33	83.35 ± 1.597	86.93 ± 0.870	76.77 ± 4.378	84.07 ± 1.069	87.30 ± 0.322
	4	87.51	25.00	64.66 ± 3.204	76.05 ± 1.391	62.93 ± 3.738	64.42 ± 1.837	76.36 ± 1.267
	5	89.24	19.97	58.82 ± 2.521	69.88 ± 1.712	72.41 ± 2.941	63.05 ± 1.826	80.85 ± 0.721
$\gamma_{base,t}$ (nats)	1	209.7	209.8	208.9 ± 1.213	209.7 ± 3.655	267.8 ± 1.246	230.8 ± 3.024	232.0 ± 2.159
	2	207.4	240.7	212.7 ± 0.579	212.1 ± 0.937	273.6 ± 0.631	232.5 ± 1.582	231.8 ± 0.416
	3	207.6	258.7	219.5 ± 1.376	216.9 ± 1.208	274.0 ± 0.552	235.6 ± 2.784	231.6 ± 0.832
	4	207.7	243.6	223.8 ± 0.837	217.1 ± 0.979	273.7 ± 0.504	236.4 ± 3.157	231.4 ± 2.550
	5	208.4	306.5	232.8 ± 5.048	222.8 ± 1.632	274.1 ± 0.349	241.1 ± 1.747	234.1 ± 1.498
$\gamma_{new,t}$ (nats)	1	209.7	209.8	208.9 ± 1.213	209.7 ± 3.655	267.8 ± 1.246	230.8 ± 3.024	232.0 ± 2.159
	2	241.1	240.2	241.8 ± 0.502	241.9 ± 0.960	313.4 ± 1.006	275.8 ± 1.888	275.3 ± 1.473
	3	213.6	211.8	215.4 ± 0.501	213.0 ± 0.635	269.1 ± 0.616	268.3 ± 3.852	262.9 ± 1.893
	4	220.5	219.7	223.6 ± 0.381	220.9 ± 0.522	282.4 ± 0.321	259.1 ± 1.305	259.6 ± 2.050
	5	246.2	242.0	248.8 ± 0.398	244.0 ± 0.646	305.8 ± 0.286	283.2 ± 2.150	283.5 ± 2.458
$\gamma_{all,t}$ (nats)	1	209.7	209.8	208.9 ± 1.213	209.7 ± 3.655	267.8 ± 1.246	230.8 ± 3.024	232.0 ± 2.159
	2	224.2	240.4	226.6 ± 2.31	226.9 ± 0.918	293.8 ± 0.349	254.3 ± 1.513	255.8 ± 0.436
	3	220.7	246.1	227.2 ± 0.606	224.9 ± 0.642	285.7 ± 0.510	261.5 ± 2.970	259.1 ± 0.929
	4	220.4	238.7	230.4 ± 0.524	226.1 ± 0.560	284.9 ± 0.703	263.2 ± 2.259	259.5 ± 3.218
	5	226.2	275.1	242.2 ± 0.754	234.6 ± 0.823	289.5 ± 0.396	271.7 ± 2.117	267.2 ± 0.586

Table A7. The results for class incremental continual-learning approaches averaged over five runs, baselines and the reference isolated learning scenario for AudioMNIST at the end of every task increment. This is an extension of Table 3 in the main body. Here, in addition to the accuracy α_t , γ_t also indicates the respective NLL at the end of every task increment t .

Audio	t	UB	LB	SupVAE	OpenVAE	PixelVAE DGR	SupPixelVAE	OpenPixelVAE
$\alpha_{base,t}$ (%)	1	99.99	100.0	99.21 ± 0.568	99.95 ± 0.035	100.0 ± 0.000	99.71 ± 0.218	99.27 ± 0.410
	2	99.92	00.00	98.98 ± 0.766	98.61 ± 0.490	99.52 ± 0.273	97.86 ± 0.799	97.88 ± 2.478
	3	100.0	00.00	92.44 ± 1.306	95.12 ± 2.248	93.15 ± 3.062	81.38 ± 5.433	95.82 ± 3.602
	4	99.92	00.00	76.43 ± 4.715	86.37 ± 5.63	81.55 ± 8.468	50.58 ± 14.60	91.56 ± 5.640
	5	98.42	00.00	59.36 ± 7.147	79.73 ± 4.070	64.60 ± 8.739	29.94 ± 18.47	75.25 ± 10.18
$\alpha_{new,t}$ (%)	1	99.99	100.0	99.21 ± 0.568	99.95 ± 0.035	100.0 ± 0.000	99.71 ± 0.218	99.27 ± 0.410
	2	99.75	100.0	91.82 ± 4.577	89.23 ± 7.384	99.71 ± 0.043	99.78 ± 0.128	99.81 ± 0.189
	3	98.92	99.58	95.20 ± 1.495	94.43 ± 3.030	98.23 ± 1.092	98.41 ± 0.507	99.30 ± 0.550
	4	97.33	98.67	53.02 ± 6.132	72.22 ± 8.493	95.31 ± 0.868	94.30 ± 0.914	97.87 ± 0.293
	5	98.67	100.0	84.93 ± 6.297	89.52 ± 6.586	98.18 ± 0.885	97.00 ± 0.520	99.43 ± 0.495
$\alpha_{all,t}$ (%)	1	99.99	100.0	99.21 ± 0.568	99.95 ± 0.035	100.0 ± 0.000	99.71 ± 0.218	99.27 ± 0.410
	2	99.83	50.00	93.84 ± 2.558	93.93 ± 3.756	99.50 ± 0.157	98.64 ± 0.875	99.67 ± 0.033
	3	99.56	33.19	94.26 ± 1.669	95.70 ± 1.524	95.37 ± 1.750	90.10 ± 1.431	97.77 ± 1.017
	4	98.60	24.58	77.90 ± 4.210	85.59 ± 3.930	86.97 ± 2.797	75.55 ± 3.891	95.41 ± 1.345
	5	97.87	20.02	81.49 ± 1.944	87.72 ± 1.594	75.50 ± 3.032	63.44 ± 5.252	90.23 ± 1.139
$\gamma_{base,t}$ (nats)	1	433.7	423.2	435.2 ± 15.69	424.2 ± 2.511	434.2 ± 1.068	432.6 ± 0.321	433.8 ± 0.370
	2	422.5	439.4	423.9 ± 0.517	425.2 ± 1.402	434.4 ± 1.082	432.5 ± 0.551	433.5 ± 1.464
	3	420.7	429.2	422.7 ± 0.690	423.8 ± 1.148	434.6 ± 0.785	432.9 ± 0.723	433.1 ± 1.269
	4	419.9	428.5	422.8 ± 0.367	423.5 ± 0.937	434.2 ± 1.209	433.0 ± 0.781	433.0 ± 1.283
	5	418.4	432.9	422.7 ± 0.182	423.5 ± 0.586	435.1 ± 1.915	431.4 ± 0.666	432.3 ± 0.189
$\gamma_{new,t}$ (nats)	1	433.7	423.2	435.2 ± 15.69	424.2 ± 2.511	434.2 ± 1.068	432.6 ± 0.321	433.8 ± 0.370
	2	381.2	384.1	382.5 ± 1.355	385.3 ± 12.56	390.4 ± 0.694	389.4 ± 0.208	389.4 ± 1.304
	3	435.9	436.7	436.3 ± 0.639	436.9 ± 0.688	444.7 ± 0.545	442.7 ± 0.513	442.4 ± 0.275
	4	485.9	487.1	486.7 ± 0.385	486.5 ± 0.701	497.4 ± 0.740	494.4 ± 0.700	494.8 ± 0.386
	5	421.3	425.2	423.9 ± 0.681	422.9 ± 0.537	431.9 ± 1.032	428.0 ± 0.851	429.7 ± 1.223
$\gamma_{all,t}$ (nats)	1	433.7	423.2	435.2 ± 15.69	424.2 ± 2.511	435.2 ± 15.69	432.6 ± 0.321	433.8 ± 0.370
	2	401.9	411.8	403.2 ± 0.831	403.5 ± 1.274	412.4 ± 0.871	410.9 ± 0.351	411.5 ± 1.406
	3	412.1	418.9	413.6 ± 0.410	413.8 ± 0.573	423.3 ± 0.618	421.0 ± 1.026	421.9 ± 0.661
	4	430.3	438.4	432.4 ± 0.436	432.6 ± 0.862	441.6 ± 0.420	439.8 ± 0.833	439.8 ± 0.718
	5	427.2	440.4	431.4 ± 0.255	430.9 ± 0.541	440.3 ± 1.297	436.9 ± 0.751	437.7 ± 0.432

References

- Boult, T.E.; Cruz, S.; Dhamija, A.R.; Gunther, M.; Henrydoss, J.; Scheirer, W.J. Learning and the Unknown: Surveying Steps Toward Open World Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Honolulu, HI, USA, 27 January–1 February 2019.
- Chen, Z.; Liu, B. Lifelong Machine Learning. In *Synthesis Lectures on Artificial Intelligence and Machine Learning*; Brachman, R., Rossi, F., Stone, P., Eds.; Morgan & Claypool Publishers LLC.: San Rafael, CA, USA, 2016; Volume 10, pp. 1–145.
- Matan, O.; Kiang, R.; Stenard, C.E.; Boser, B.E.; Denker, J.; Henderson, D.; Hubbard, W.; Jackel, L.; LeCun, Y. Handwritten Character Recognition Using Neural Network Architectures. In Proceedings of the 4th United States Postal Service Advanced Technology Conference, Washington, DC, USA, 5–7 December 1990; pp. 1003–1012.
- Hendrycks, D.; Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
- Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.V.; Lakshminarayanan, B.; Snoek, J. Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift. In Proceedings of the Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019.
- Nalisnick, E.; Matsukawa, A.; Teh, Y.W.; Gorur, D.; Lakshminarayanan, B. Do Deep Generative Models Know What They Don’t Know? In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
- McCloskey, M.; Cohen, N.J. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychol. Learn. Motiv.-Adv. Res. Theory* **1989**, *24*, 109–165.

8. Ratcliff, R. Connectionist Models of Recognition Memory: Constraints Imposed by Learning and Forgetting Functions. *Psychol. Rev.* **1990**, *97*, 285–308. [[CrossRef](#)] [[PubMed](#)]
9. Scheirer, W.J.; Rocha, A.; Sapkota, A.; Boulton, T.E. Towards Open Set Recognition. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2013**, *35*, 1757–1772. [[CrossRef](#)] [[PubMed](#)]
10. Scheirer, W.J.; Jain, L.P.; Boulton, T.E. Probability Models For Open Set Recognition. *IEEE Trans. Pattern Anal. Mach. Intell. (TPAMI)* **2014**, *36*, 2317–2324. [[CrossRef](#)] [[PubMed](#)]
11. Parisi, G.I.; Kemker, R.; Part, J.L.; Kanan, C.; Wermter, S. Continual Lifelong Learning with Neural Networks: A Review. *Neural Netw.* **2019**, *113*, 54–71. [[CrossRef](#)] [[PubMed](#)]
12. Bendale, A.; Boulton, T.E. Towards Open Set Deep Networks. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
13. Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; Madry, A. Adversarial Examples are not Bugs, they are Features. In Proceedings of the Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019.
14. Shah, H.; Tamuly, K.; Raghunathan, A.; Jain, P.; Netrapalli, P. The Pitfalls of Simplicity Bias in Neural Networks. In Proceedings of the Neural Information Processing Systems (NeurIPS), Online, 6–12 December 2020.
15. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. In Proceedings of the International Conference on Learning Representations (ICLR), Scottsdale, AZ, USA, 2–4 May 2013.
16. Kingma, D.P.; Ba, J.L. Adam: A Method for Stochastic Optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
17. van den Oord, A.; Kalchbrenner, N.; Kavukcuoglu, K. Pixel Recurrent Neural Networks. In Proceedings of the 33rd International Conference on Machine Learning (ICML 2016), New York, NY, USA, 19–24 June 2016.
18. Gulrajani, I.; Kumar, K.; Faruk, A.; Taiga, A.A.; Visin, F.; Vazquez, D.; Courville, A. PixelVAE: A Latent Variable Model for Natural Images. In Proceedings of the 5th International Conference on Learning Representations (ICLR 2017), Toulon, France, 24–26 April 2017.
19. Huang, H.; Li, Z.; He, R.; Sun, Z.; Tan, T. Introvae: Introspective variational autoencoders for photographic image synthesis. In Proceedings of the Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 2–8 December 2018.
20. Ulyanov, D.; Vedaldi, A.; Lempitsky, V. It Takes (Only) Two: Adversarial Generator-Encoder Networks. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New Orleans, LA, USA, 2–7 February 2018.
21. Zenke, F.; Poole, B.; Ganguli, S. Continual Learning Through Synaptic Intelligence. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017.
22. Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A.A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. Overcoming catastrophic forgetting in neural networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 3521–3526. [[CrossRef](#)] [[PubMed](#)]
23. Li, Z.; Hoiem, D. Learning without forgetting. In Proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 8–16 October 2016.
24. Hinton, G.E.; Vinyals, O.; Dean, J. Distilling the Knowledge in a Neural Network. In Proceedings of the Neural Information Processing Systems (NeurIPS), Deep Learning Workshop, Montreal, QC, Canada, 8–13 December 2014.
25. Robins, A. Catastrophic Forgetting, Rehearsal and Pseudorehearsal. *Connect. Sci.* **1995**, *7*, 123–146. [[CrossRef](#)]
26. Rebuffi, S.A.; Kolesnikov, A.; Sperl, G.; Lampert, C.H. iCaRL: Incremental Classifier and Representation Learning, *arXiv* **2017**, arXiv:1611.07725.
27. Mensink, T.; Verbeek, J.; Perronnin, F.; Csurka, G.; Mensink, T.; Verbeek, J.; Perronnin, F.; Csurka, G. Metric Learning for Large Scale Image Classification: Generalizing to New Classes at Near-Zero Cost. In Proceedings of the European Conference on Computer Vision (ECCV), Florence, Italy, 7–13 October 2012.
28. Bachem, O.; Lucic, M.; Krause, A. Coresets for Nonparametric Estimation—The Case of DP-Means. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015.
29. O’Reilly, R.C.; Norman, K.A. Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends Cogn. Sci.* **2003**, *6*, 505–510. [[CrossRef](#)]
30. Gepperth, A.; Karaoguz, C. A Bio-Inspired Incremental Learning Architecture for Applied Perceptual Problems. *Cogn. Comput.* **2016**, *8*, 924–934. [[CrossRef](#)]
31. Hochreiter, S.; Schmidhuber, J. Long Short-Term Memory. *Neural Comput.* **1997**, *9*, 1735–1780. [[CrossRef](#)] [[PubMed](#)]
32. Shin, H.; Lee, J.K.; Kim, J.J.; Kim, J. Continual Learning with Deep Generative Replay. In Proceedings of the Neural Information Processing Systems (NeurIPS), Long Beach, CA, USA, 4–9 December 2017.
33. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative Adversarial Nets. In Proceedings of the Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 8–13 December 2014.
34. Farquhar, S.; Gal, Y. A Unifying Bayesian View of Continual Learning. In Proceedings of the Neural Information Processing Systems (NeurIPS), Bayesian Deep Learning Workshop, Montreal, QC, Canada, 2–8 December 2018.
35. Farquhar, S.; Gal, Y. Towards Robust Evaluations of Continual Learning. In Proceedings of the International Conference on Machine Learning (ICML), Lifelong Learning: A Reinforcement Learning Approach Workshop, Stockholm, Sweden, 10–15 June 2018.
36. Nguyen, C.V.; Li, Y.; Bui, T.D.; Turner, R.E. Variational Continual Learning. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.

37. Achille, A.; Eccles, T.; Matthey, L.; Burgess, C.P.; Watters, N.; Lerchner, A.; Higgins, I. Life-Long Disentangled Representation Learning with Cross-Domain Latent Homologies. In Proceedings of the Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 2–8 December 2018.
38. Liang, S.; Li, Y.; Srikant, R. Enhancing the Reliability of Out-of-distribution Image Detection in Neural Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
39. Lee, K.; Lee, H.; Lee, K.; Shin, J. Training Confidence-Calibrated Classifiers for Detecting Out-of-Distribution Samples. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
40. Dharmija, A.R.; Günther, M.; Boulton, T.E. Reducing Network Agnostophobia. In Proceedings of the Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 2–8 December 2018.
41. MacKay, D.J.C. A Practical Bayesian Framework. *Neural Comput.* **1992**, *472*, 448–472. [[CrossRef](#)]
42. Gal, Y.; Ghahramani, Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015.
43. Graves, A. Practical variational inference for neural networks. In Proceedings of the Neural Information Processing Systems (NeurIPS), Granada, Spain, 12–17 December 2011.
44. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.P.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
45. Tomczak, J.M.; Welling, M. VAE with a vampprior. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), Canary Islands, Spain, 9–11 April 2018.
46. Hoffman, M.D.; Johnson, M.J. ELBO surgery: Yet another way to carve up the variational evidence lower bound. In Proceedings of the Neural Information Processing Systems (NeurIPS), Advances in Approximate Bayesian Inference Workshop, Barcelona, Spain, 5–10 December 2016.
47. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2323. [[CrossRef](#)]
48. Burgess, C.P.; Higgins, I.; Pal, A.; Matthey, L.; Watters, N.; Desjardins, G.; Lerchner, A. Understanding disentangling in beta-VAE. In Proceedings of the Neural Information Processing Systems (NeurIPS), Workshop on Learning Disentangled Representations, Long Beach, CA, USA, 4–9 December 2017.
49. Mathieu, E.; Rainforth, T.; Siddharth, N.; Teh, Y.W. Disentangling disentanglement in variational autoencoders. In Proceedings of the International Conference on Machine Learning (ICML), Long Beach, CA, USA, 10–15 June 2019; pp. 7744–7754.
50. Bauer, M.; Mnih, A. Resampled Priors for Variational Autoencoders. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), Naha, Okinawa, Japan, 16–18 April 2019.
51. Takahashi, H.; Iwata, T.; Yamanaka, Y.; Yamada, M.; Yagi, S. Variational Autoencoder with Implicit Optimal Priors. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), Honolulu, HI, USA, 27 January–1 February 2019.
52. Nilsback, M.E.; Zisserman, A. A Visual Vocabulary For Flower Classification. In Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), Washington, DC, USA, 17–22 June 2006; pp. 1447–1454.
53. Xiao, H.; Rasul, K.; Vollgraf, R. Fashion-MNIST: A Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv* **2017**, arXiv:1708.07747.
54. Becker, S.; Ackermann, M.; Lopuschkin, S.; Müller, K.R.; Samek, W. Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals. *arXiv* **2018**, arXiv:1807.03418.
55. Clanuwat, T.; Bober-Irizar, M.; Kitamoto, A.; Lamb, A.; Yamamoto, K.; Ha, D. Deep Learning for Classical Japanese Literature. In Proceedings of the Neural Information Processing Systems (NeurIPS), Workshop on Machine Learning for Creativity and Design, Conference, Montreal, QC, Canada, 2–8 December 2018.
56. Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; Ng, A.Y. Reading Digits in Natural Images with Unsupervised Feature Learning. In Proceedings of the Neural Information Processing Systems (NeurIPS), Granada, Spain, 12–17 December 2011.
57. Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images; Technical Report, Toronto. 2009. Available online: <https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf> (accessed on 22 January 2022).
58. Hendrycks, D.; Gimpel, K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In Proceedings of the International Conference on Learning Representations (ICLR), Toulon, France, 24–26 April 2017.
59. Bendale, A.; Boulton, T.E. Towards Open World Recognition. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
60. Kemker, R.; McClure, M.; Abitino, A.; Hayes, T.; Kanan, C. Measuring Catastrophic Forgetting in Neural Networks. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), New Orleans, LA, USA, 2–7 February 2018.
61. Zagoruyko, S.; Komodakis, N. Wide Residual Networks. In Proceedings of the In Proceedings of the British Machine Vision Conference (BMVC), York, UK, 19–22 September 2016.
62. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016.
63. Chaudhry, A.; Dokania, P.K.; Ajanthan, T.; Torr, P.H.S. Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.

64. Hu, W.; Lin, Z.; Liu, B.; Tao, C.; Tao, Z.; Zhao, D.; Ma, J.; Yan, R. Overcoming catastrophic forgetting for continual learning via model adaptation. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
65. Wu, C.; Herranz, L.; Liu, X.; Wang, Y.; van de Weijjer, J.; Raducanu, B. Memory Replay GANs: Learning to generate images from new categories without forgetting. In Proceedings of the Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 2–8 December 2018.
66. Zhai, M.; Chen, L.; Tung, F.; He, J.; Nawhal, M.; Mori, G. Lifelong GAN: Continual Learning for Conditional Image Generation. In Proceedings of the International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2019.
67. Welling, M. Herding dynamical weights to learn. In Proceedings of the International Conference on Machine Learning (ICML), Montreal, QC, Canada, 14–18 June 2009.
68. Prabhu, A.; Torr, P.; Dokania, P. GDumb: A Simple Approach that Questions Our Progress in Continual Learning. In Proceedings of the European Conference on Computer Vision (ECCV), Online, 23–28 August 2020.
69. Liu, Y.; Su, Y.; Liu, A.A.; Schiele, B.; Sun, Q. Mnemonics Training: Multi-Class Incremental Learning without Forgetting. In Proceedings of the Computer Vision Pattern Recognition (CVPR), Online, 14–19 June 2020.
70. Cha, H.; Lee, J.; Shin, J. Co2L: Contrastive Continual Learning. In Proceedings of the International Conference on Computer Vision (ICCV), online, 11–17 October 2021.
71. Buzzega, P.; Boschini, M.; Porrello, A.; Abati, D.; Calderara, S. Dark Experience for General Continual Learning: A Strong, Simple Baseline. In Proceedings of the Neural Information Processing Systems (NeurIPS), Online, 6–12 December 2020.
72. Kingma, D.P.; Rezende, D.J.; Mohamed, S.; Welling, M. Semi-Supervised Learning with Deep Generative Models. In Proceedings of the Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 8–13 December 2014.
73. Larsen, A.B.L.; Sonderby, S.K.; Larochelle, H.; Winther, O. Autoencoding beyond pixels using a learned similarity metric. In Proceedings of the International Conference on Machine Learning (ICML), New York, NY, USA, 19–24 June 2016.
74. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International Conference on Machine Learning (ICML), Lille, France, 6–11 July 2015.
75. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.