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

Martin Mundt¹ Sagnik Majumder¹ Iuliia Pliushch¹ Visvanathan Ramesh¹

Abstract

We introduce a probabilistic approach to unify deep continual learning with open set recognition, based on variational Bayesian inference. Our single model combines a joint probabilistic encoder with a generative model and a linear classifier that get shared across sequentially arriving tasks. In order to successfully distinguish unseen unknown data from trained known tasks, we propose to bound the class specific approximate posterior by fitting regions of high density on the basis of correctly classified data points. These bounds are further used to significantly alleviate catastrophic forgetting by avoiding samples from low density areas in generative replay. Our approach requires no storing of old- or upfront knowledge of future data and is empirically validated on visual and audio tasks in class incremental, as well as cross-dataset scenarios across modalities.

1. Introduction

Modern machine learning systems are typically trained in a closed world setting according to an isolated learning paradigm. They take on the assumption that data is available at all times and data inputs encountered during application of the learned model come from the same statistical population as the training data. However, the real world requires dealing with sequentially arriving tasks and data coming from potentially yet unknown sources. A neural network that is trained exclusively on such newly arriving data will overwrite its representations and thus forget knowledge of past tasks, an early identified phenomenon coined catastrophic forgetting (McCloskey & Cohen, 1989). Moreover, when confronting the learned model with unseen concepts, misclassification is bound to occur (Matan et al., 1990).

Existing continual learning literature predominantly concen-

trates its efforts on finding mechanisms to alleviate catastrophic forgetting (Parisi et al., 2019) and the term continual learning is not necessarily used in a wider sense. Specifically, the aforementioned crucial 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. A naive conditioning on unseen unknown data through inclusion of a "background" class is infeasible as by definition we do not have access to it a priori. Commonly applied thresholding of prediction values is veritably insufficient as resulting large confidences cannot be prevented (Matan et al., 1990). Arguably this also includes variational methods (Kingma & Welling, 2013; Farquhar & Gal, 2018; Achille et al., 2018) to gauge neural network uncertainty, since the closed world assumption also holds true for Bayesian methods (Boult et al., 2019). Recently, Bendale & Boult (2016) have proposed extreme value theory (EVT) based meta-recognition to address open set detection on the basis of softmax predictions in conventional feed-forward deep neural networks. Inspired by this work, we propose a probabilistic approach to unify open set recognition with continual learning in a single deep model. Our specific contributions are:

- We introduce a single model for continual learning that combines a joint probabilistic encoder with a generative model and a linear classifier. This architecture enables a natural formulation to address open set recognition on the basis of EVT bounds to the class conditional approximate posterior in Bayesian inference.
- Apart from using EVT for detection of unseen unknown data, we show that generated samples from areas of low probability density under the aggregate posterior can be excluded in generative replay for continual learning. This leads to significantly reduced catastrophic forgetting without storing real data.
- Empirically, we show that our model can incrementally learn the classes of two image and one audio dataset, as well as cross-dataset scenarios across modalities, while being able to successfully distinguish various unseen datasets from data belonging to known tasks.

¹Center for Cognition and Computation, Goethe University, Frankfurt, Germany. Correspondence to: Martin Mundt <mmundt@em.uni-frankfurt.de>.

2. Background and Related Work

2.1. Continual Learning

In isolated supervised machine learning the core assumption is the presence of i.i.d. data at all times and training is conducted using a dataset $\mathcal{D} \equiv \left\{ (\mathbf{x}^{(n)}, y^{(n)}) \right\}_{n=1}^N$, consisting of N pairs of data instances $\mathbf{x}^{(n)}$ and their corresponding labels $y^{(n)} \in \{1 \dots C\}$ for C classes. In contrast, in continual learning task data $\mathcal{D}_t \equiv \left\{ (\mathbf{x}_t^{(n)}, y_t^{(n)}) \right\}_{n=1}^{N_t}$ with $t = 1, \dots, T$ arrives sequentially for T disjoint datasets, each with number of classes C_t . It is assumed that only the data of the current task is available. Different methods in the literature have been identified to prevent a model from forgetting past knowledge, either explicitly, through regularization or freezing of weights, or implicitly, through rehearsal of data by sampling retained subsets or sampling from a generative memory. A recent review of many continual learning methods is provided by [Parisi et al. \(2019\)](#). Here, we present a brief summary of particular related works.

Regularization and weight freezing: Regularization methods such as synaptic intelligence (SI) ([Zenke et al., 2017](#)) or elastic weight consolidation (EWC) ([Kirkpatrick et al., 2017](#)) explicitly constrain the weights during continual learning to avoid drifting too far away from previous tasks' solutions. In a related picture, learning without forgetting ([Li & Hoiem, 2016](#)) uses knowledge distillation ([Hinton et al., 2014](#)) to regularize the end-to-end functional. Further methods employ dynamically expandable neural networks ([Yoon et al., 2018](#)) or progressive networks ([Rusu et al., 2016](#)), that expand the capacity of the neural network, while freezing or regularizing existing representations.

Rehearsal: These methods store and rehearse data from distributions belonging to old tasks or generate samples in pseudo-rehearsal ([Robins, 1995](#)). The central component of the former is thus the selection of significant instances. For methods such as iCarl ([Rebuffi et al., 2017](#)) it is therefore common to resort to auxiliary techniques such as a nearest-mean classifier ([Mensink et al., 2012](#)) or coresets ([Bachem et al., 2015](#)). Inspired by complementary learning systems theory ([O'Reilly & Norman, 2003](#)), dual-model approaches sample data from a separate generative memory. In GeppNet ([Gepperth & Karaoguz, 2016](#)) an additional long-short term memory ([Hochreiter & Schmidhuber, 1997](#)) is used for storage, whereas generative replay ([Shin et al., 2017](#)) samples form a separately trained generative adversarial network (GAN) ([Goodfellow et al., 2014](#)).

Bayesian methods: As detailed by [Farquhar & Gal \(2018\)](#), Bayesian methods provide natural capability for continual learning by making use of the learned distribution in e.g. a variationally trained neural network ([Kingma &](#)

[Welling, 2013](#)). Existing works nevertheless fall into the above two categories: a prior-based approach using the former task's approximate posterior as the new task's prior ([Nguyen et al., 2018](#)) or estimating the likelihood of former data through generative replay or other forms of rehearsal ([Farquhar & Gal, 2018; Achille et al., 2018](#)).

Evaluation assumptions and multiple model heads: The success of many of these techniques can be attributed mainly to the considered evaluation scenario. With the exception of [Farquhar & Gal \(2018\)](#), all above techniques train a separate classifier per task and thus either require explicit storage of task labels, or assume the presence of a task oracle during evaluation. This multi-head classifier 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 gauge catastrophic forgetting, it also signifies a major limitation in practical application. Even though [Farquhar & Gal \(2018\)](#) use a single classifier, they train a separate generative model per task to avoid catastrophic forgetting of the generator.

Our approach builds upon these previous works by proposing a single model with single classifier head with a natural mechanism for open set recognition and improved generative replay from a Bayesian perspective.

2.2. Out-of-distribution and open set recognition

The above mentioned literature focuses their continual learning efforts predominantly on addressing catastrophic forgetting. 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 and our models need the ability to identify unknown examples in an open world. We provide a small overview of relevant approaches to address the latter. A more comprehensive review of recent methods is provided by [Boult et al. \(2019\)](#).

Bayesian uncertainty: Bayesian deep neural network models ([Kingma & Welling, 2013](#)) could be argued to intrinsically be able to reject statistical outliers. Intuitively, one could estimate a model's uncertainty through Monte-Carlo dropout ([Gal & Ghahramani, 2015](#)) or other variational approximations ([Farquhar & Gal, 2018](#)). However, this is generally insufficient as uncertain inputs are not necessarily unknown and vice versa unknowns do not necessarily have to appear as uncertain ([Boult et al., 2019](#)).

Calibration: The aim of these works is to separate a known and unknown input through prediction confidence, often by fine-tuning or re-training an already existing model. In ODIN ([Liang et al., 2018](#)) this is addressed through per-

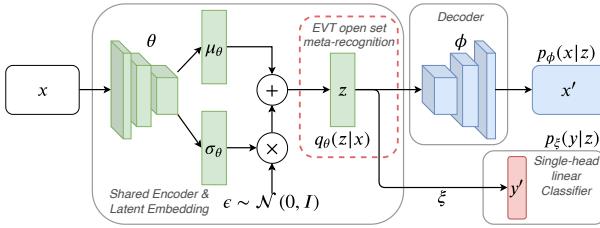


Figure 1. Joint continual learning model consisting of a shared probabilistic encoder $q_\theta(z|x)$, probabilistic decoder $p_\phi(x, z)$ and probabilistic classifier $p_\xi(y, z)$. For open set recognition and generative replay with outlier rejection, EVT based bounds on the basis of the approximate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

turbations and temperature scaling, while Lee et al. (2018) use a separately trained GAN to generate out-of-distribution samples from low probability densities and explicitly reduce their confidence through inclusion of an additional loss term. Similarly, Dhamija et al. (2018) define a loss function that aims to maximize entropy for unknown inputs.

Extreme value theory: One approach to open set recognition in deep neural networks is through extreme-value theory (EVT) based meta-recognition (Thomas et al., 2014; Bendale & Boult, 2016), i.e. without re-training or modifying loss functions by assuming upfront presence of unknown data. The goal here is to bound the open space on the basis of already seen data instances. Bendale & Boult (2016) have proposed OpenMax to modify a neural network’s softmax prediction values on the basis of extreme values of the penultimate layer’s activation values.

Our work extends these approaches by moving away from predictive values and instead uses EVT to bound the approximate posterior. In contrast to predictive values such as reconstruction losses, where differences in reconstructed images do not necessarily have to reflect the outcome with respect to our task’s target, we thus directly operate on the underlying (lower-bound to the) data distribution, and the generative factors. This allows us to also constrain generative replay to distribution inliers, which further alleviates catastrophic forgetting in continual learning. While we can still leverage variational inference to gauge model uncertainty, the need to rely on classifier entropy or confidence, that are known to be overconfident and can never be calibrated for all unknown inputs, is circumvented.

3. Unifying Continual Learning with Open Set Recognition

We consider the continual learning scenario with awareness of an open world from a perspective of variational Bayesian inference in deep neural networks (Kingma &

Welling, 2013). Our model consists of 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 is placed. Using variational inference, the encoder’s purpose is to approximate the true posterior to both $p_\phi(x, z)$ 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 $n \in D$ of the following loss thus needs to be optimized:

$$\begin{aligned} \mathcal{L}(\mathbf{x}^{(n)}, \mathbf{y}^{(n)}; \theta, \phi, \xi) &= -\beta KL(q_\theta(z|\mathbf{x}^{(n)}) || p(z)) \\ &+ \mathbb{E}_{q_\theta(z|\mathbf{x}^{(n)})} [\log p_\phi(\mathbf{x}^{(n)}|z) + \log p_\xi(\mathbf{y}^{(n)}|z)] \end{aligned} \quad (1)$$

This model can be seen as a variant of β -VAE (Higgins et al., 2017), where in addition to approximating the data distribution the model learns to incorporate the class structure into the latent space. It forms the basis for continual learning with open set recognition and respective improvements to generative replay, which will be discussed in subsequent sections. An illustration of the model is shown in figure 1.

3.1. Learning continually through generative replay

Without further constraints, one could continually train above model by sequentially accumulating and optimizing equation 1 over all currently present tasks $t = 1, \dots, T$:

$$\mathcal{L}_t^{\text{UB}}(\mathbf{x}, \mathbf{y}; \theta, \phi, \xi) = \sum_{\tau=1}^t \frac{1}{N_\tau} \sum_{n=1}^{N_\tau} \mathcal{L}(\mathbf{x}_\tau^{(n)}, \mathbf{y}_\tau^{(n)}; \theta, \phi, \xi) \quad (2)$$

Being based on the accumulation of real data, this equation provides an upper-bound to 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 generative nature of our model, we follow previous works (Farquhar & Gal, 2018; Achille et al., 2018) and estimate the likelihood of former data through generative replay:

$$\begin{aligned} \mathcal{L}_t(\mathbf{x}, \mathbf{y}; \theta, \phi, \xi) &= \frac{1}{N_t} \sum_{n=1}^{N_t} \mathcal{L}(\mathbf{x}_t^{(n)}, \mathbf{y}_t^{(n)}; \theta, \phi, \xi) \\ &+ \frac{1}{N'_t} \sum_{n=1}^{N'_t} \mathcal{L}(\mathbf{x}'_t^{(n)}, \mathbf{y}'_t^{(n)}; \theta, \phi, \xi) \end{aligned} \quad (3)$$

where,

$$\mathbf{x}'_t \sim p_{\phi, t-1}(\mathbf{x}|z); \mathbf{y}'_t \sim p_{\xi, t-1}(y|z) \text{ and } z \sim p(z) \quad (4)$$

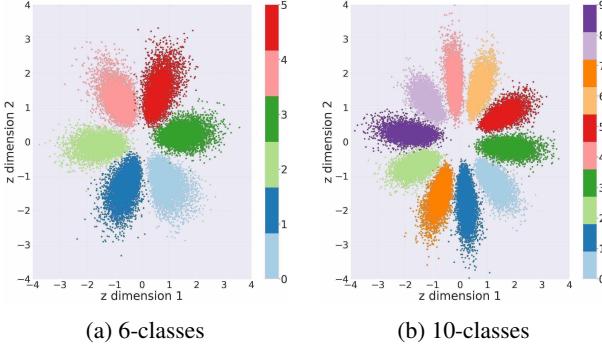


Figure 2. (a) 2-D latent space visualization for continually learned MNIST after inclusion of six classes and at the end of training.

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

3.2. Linear classifier expansion and the role of β

In contrast to prior works based on multiple models, our approach of using equation 3 to continually train a single model has two implications. With every encounter of an additional class: 1. a new classifier unit and corresponding weights need to be added. 2: the latent encoding needs to adjust to accommodate the additional class under the constraint of the classifier requirement of linear separability.

The first implication can be addressed by expanding the existing classifier weight tensor and only initializing the newly added weights. If the distribution from which the newly added weights are drawn is independent of the number of classes and only depends on the input dimensionality, such as the initialization scheme proposed by (He et al., 2015), the initialization scheme remains constant throughout training. While the addition itself will temporarily confuse existing units, this should make sure that newly added parameters are on the same scale as existing weights and thus trained in practice. Note that in principle, during the optimization of a task the weight distribution could shift significantly from its initial state. However, we do not encounter this potential issue in empirical experiments. Nevertheless, we point out that this currently under-explored topic requires separate future investigation in the context of model expansion.

For the second implication, the β term of equation 1 is crucial. Here, the role of beta is to control the capacity of

the information bottleneck and regulate the effective latent encoding overlap (Burgess et al., 2017), which can best be summarized with a direct quote from the recent work of Mathieu et al. (2019): “*The overlap factor is perhaps best understood by considering extremes: too little, and 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). This can be seen as under- or over-regularization by the prior of what is typically referred to as the aggregate posterior (Hoffman & Johnson, 2016):

$$q_{\theta,t}(\mathbf{z}) = \mathbb{E}_{p_{\tilde{\mathcal{D}}_t}(\tilde{\mathbf{x}})} [q_{\theta,t}(\mathbf{z}|\tilde{\mathbf{x}})] \approx \frac{1}{\tilde{N}_t} \sum_{n=1}^{\tilde{N}_t} q_{\theta,t}(\mathbf{z}|\tilde{\mathbf{x}}^{(n)}) \quad (5)$$

As an extension of this argument to our model, the necessity of linear class separation given \mathbf{z} requires a suitable level of encoding overlap. This forms the basis for our open set recognition and respective improved generative replay for continual learning, which will be discussed in the following paragraphs. Example two-dimensional latent encodings for a continually trained MNIST (LeCun et al., 1998) model with appropriate β are shown in figure 2. Here, we can see that the classes are cleanly separated in latent space, as enforced by the linear classification objective, and new classes can be accommodated continually. Further discussion on the choice of β can be found in the supplementary material.

3.3. Open set recognition and generative replay with statistical outlier rejection

Trained naively in above fashion, our model would suffer from accumulated errors with each successive iteration of generative replay, similar to current literature approaches. The main challenge is that high density areas under the prior $p(\mathbf{z})$ are not necessarily reflected in the structure of the aggregate posterior $q_{\theta,t}(\mathbf{z})$ (Tomczak & Welling, 2018). Thus, generated data from low density regions of the latter does not generally correspond to encountered data instances. Vice-versa, data instances that fall into high density regions under the prior should not generally be considered as statistical outliers with respect to the observed data distribution.

Ideally, this challenge would be solved by modifying equations 1 and 2 by replacing the Gaussian prior in the KL-divergence with $q_{\theta,t}(\mathbf{z})$ and respectively sampling $\mathbf{z} \sim q_{\theta,t-1}(\mathbf{z})$ for generative replay in equations 3 and 4. Even though using the aggregate posterior as the prior is the objective in multiple recent works, it can be challenging in high dimensions, lead to over-fitting and often comes at the expense of additional hyper-parameters (Tomczak & Welling, 2018; Bauer & Mnih, 2019; Takahashi et al., 2019). To avoid finding an explicit representation for the multi-modal $q_{\theta,t}(\mathbf{z})$, we leverage our model’s class disentanglement and draw inspiration from the EVT based OpenMax approach

(Bendale & Boult, 2016). However, instead of using knowledge about extreme distance values in penultimate layer activations to modify a Softmax prediction’s confidence, we 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 means 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 data instances at the end of task t as $m = 1, \dots, \tilde{M}_t$. To construct a statistical meta-recognition model, we first obtain each class’ mean latent vector for all correctly predicted seen data instances:

$$\bar{z}_{c,t} = \frac{1}{|\tilde{M}_{c,t}|} \sum_{m \in \tilde{M}_{c,t}} \mathbb{E}_{q_{\theta,t}(z|\tilde{x}_t^{(m)})} [z] \quad (6)$$

and define the respective set of latent distances as:

$$\Delta_{c,t} \equiv \left\{ f_d \left(\bar{z}_{c,t}, \mathbb{E}_{q_{\theta,t}(z|\tilde{x}_t^{(m)})} [z] \right) \right\}_{m \in \tilde{M}_{c,t}} \quad (7)$$

Here, f_d signifies a choice of distance metric. We proceed to fit 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 the distances are based on the class conditional approximate posterior, we can thus bound the latent space regions of high density. The tightness of the bounds 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 estimate any sample’s outlier probability:

$$\omega_{\rho,t}(z) = \min \left(1 - \exp \left(-\frac{|f_d(\bar{z}_t, z) - \tau_t|}{\lambda_t} \right) \right)^{\kappa_t} \quad (8)$$

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 as it is far away from all known classes. For a novel data instance, the outlier probability can be based on computation of the probabilistic encoder $z \sim q_{\theta,t}(z|x)$ and a false overconfident classifier prediction avoided. Analogously, for the generative model, equation 8 can be used with $z \sim p(z)$ and the probabilistic decoder only calculated for samples that are considered to be statistically inlying. This way, we can constrain the naive generative replay of equation 4 to the aggregate posterior, while avoiding the

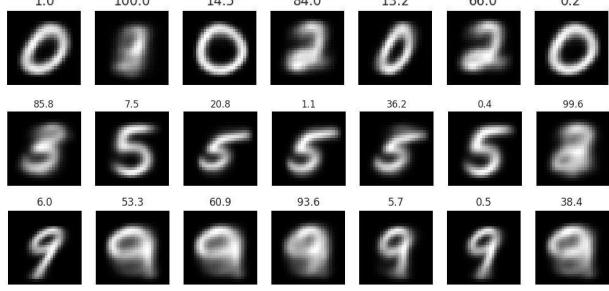


Figure 3. Generated MNIST images $x \sim p_{\phi,t}(x|z)$ with $z \sim p(z)$ and their corresponding class c obtained from the classifier $p_{\xi,t}(y|z)$ for $c = 0$ (top row), $c = 5$ (middle row) and $c = 9$ (bottom row) together with their open set outlier percentage.

need to sample $z \sim q_{\theta,t}(z)$ directly. Although this may sound detrimental to our method, it comes with the advantage of scalability to high dimensions. We further argue that the computational overhead for generative replay, both from sampling from the prior $z \sim p(z)$ in large parallelized batches and computation of equation 8, is negligible in contrast to the much more computationally heavy deep probabilistic decoder or even the linear classifier, as the latter only need to be calculated for accepted samples. To give a visual illustration, we show examples of generated MNIST images together with their outlier percentage in figure 3.

4. Experiments

Similar to recent literature (Zenke et al., 2017; Kirkpatrick et al., 2017; Farquhar & Gal, 2018; Shin et al., 2017; Parisi et al., 2019), we consider the incremental MNIST (LeCun et al., 1998) dataset, where classes arrive in groups of two, and corresponding versions of the FashionMNIST (Xiao et al., 2017) and AudioMNIST dataset (Becker et al., 2018). For the latter we follow the authors’ procedure of converting the audio recordings into spectrograms. In addition to this class incremental setting, we also evaluate cross-dataset scenarios, where datasets are sequentially added with all of their classes and the model has to learn across modalities.

For a common frame of reference, we base both encoder and decoder architectures on 14-layer wide residual networks with a latent dimensionality of 60 (He et al., 2016; Zagoruyko & Komodakis, 2016; Gulrajani et al., 2017; Chen et al., 2017). For the generative replay with statistical outlier rejection, we use an aggressive rejection rate of $\Omega_t = 0.01$ and dynamically set tail-sizes to 5% of seen examples per class. To avoid over-fitting, we add noise sampled from $\mathcal{N}(0, 0.25)$ to each input. This is preferable to weight regularization as it doesn’t entail unrecoverable units that are needed to encode later tasks. We thus refer to our proposed model as Open-set Classifying Denoising Variational Auto-Encoder (OCDVAE), for which we have found a value of

$\beta = 0.1$ to consistently work well, see discussion in the appendix. We empirically compare the following methods:

Dual Model: separate generative and discriminative variational models in analogy to [Shin et al. \(2017\)](#).

EWC: elastic weight consolidation ([Kirkpatrick et al., 2017](#)) for a purely discriminative model.

OCDVAE (ours): our proposed joint model with posterior based open set recognition and resulting statistical outlier rejection in generative replay.

CDVAE: the naive approach of generating from the prior distribution in our joint model. We include these results to highlight the effect of aggregate posterior to prior mismatch.

ISO: isolated learning, where all data is always present.

UB: upper-bound on achievable model performance by sequentially accumulating all data, given by equation 2.

LB: lower-bound on model performance when only the current task’s data is available. No additional mechanism is in place and full catastrophic forgetting occurs.

Our evaluation metrics are inspired by previously proposed continual learning measures ([Lopez-Paz & Ranzato, 2017](#); [Kemker et al., 2018](#)). In addition to overall accuracy $\alpha_{t,all}$, these metrics monitor forgetting by computing a base accuracy $\alpha_{t,base}$ for the initial task at increment t , while also gauging the amount of new knowledge that can be encoded by monitoring the accuracy for the most recent increment $\alpha_{t,new}$. We evaluate the quality of the generative models through classification accuracy as it depends on generated replay and a direct evaluation of pixel-wise reconstruction losses is not necessarily coupled to classification accuracy or retention thereof. However, we provide a detailed analysis of reconstruction losses for all tasks, as well as KL divergences for all experiments in the supplementary material.

To provide a fair comparison of achievable accuracy, all above approaches are trained to converge on each task using the Adam optimizer ([Kingma & Ba, 2015](#)). We repeat all experiments 5 times to gauge statistical consistency. The full hyper-parameter specification can be found in the supplementary material. All models were trained on a single GTX 1080 GPU and we make our code publicly available at <https://github.com/MrtnMndt/OCDVAEContinualLearning>

4.1. Learning across datasets in an open world

Achieved accuracies for continual learning across datasets are summarized in table 1. In general the upper-bound values are almost identical to isolated learning. Similarly, the new task’s metrics are negligibly close, as the WRN architecture ensures enough capacity to encode new knowledge. In contrast to EWC that is universally unable to maintain knowledge, as also previously observed by [Kemker et al. \(2018\)](#); [Parisi et al. \(2019\)](#), approaches based on generative replay are able to partially retain information. Yet they accumulate errors due to samples generated from low density

Table 1. Results for continual learning across datasets averaged over 5 runs, baselines and the reference isolated learning scenario for FashionMNIST (F) → MNIST (M) → AudioMNIST (A) and the reverse order. α_T indicates the respective accuracy at the end of the last increment $T = 3$.

	Cross-dataset	$\alpha_T(\%) (T=3)$		
		base	new	all
F-M-A	CDVAE ISO			94.95
	CDVAE UB	89.10	97.88	95.00
	CDVAE LB	00.00	98.12	22.70
	EWC	22.85 ± 0.294	93.31 ± 0.138	43.42 ± 0.063
	Dual Model	81.89 ± 0.104	96.78 ± 0.067	91.75 ± 0.064
	CDVAE	57.70 ± 4.480	96.73 ± 0.235	81.10 ± 1.769
	OCDVAE	80.11 ± 2.922	97.63 ± 0.042	91.13 ± 1.045
A-M-F	CDVAE ISO			94.95
	CDVAE UB	97.17	89.16	94.91
	CDVAE LB	00.00	89.72	34.51
	EWC	3.420 ± 0.026	87.54 ± 0.214	45.42 ± 0.731
	Dual Model	66.82 ± 0.337	89.15 ± 0.050	87.70 ± 0.102
	CDVAE	79.74 ± 2.431	88.50 ± 0.126	89.46 ± 0.600
	OCDVAE	94.53 ± 0.283	89.53 ± 0.367	94.06 ± 0.156

regions. This is noticeable for both the dual model approach with a separate VAE and discriminative model, and more heavily so for the naive CDVAE where the structure of $q_{\theta,t}(z)$ is further affected by the discriminator. However, our proposed OCDVAE model overcomes this issue to a considerable degree, rivalling and improving upon the performance of training separate models.

Apart from these classification accuracies, we also quantitatively analyze the models’ ability to distinguish unknown tasks’ data from data belonging to known tasks. Here, the challenge is to consider all unseen test data of already trained tasks as inlying, while successfully identifying 100 % of unknown datasets as outliers. For this purpose, we evaluate models after training on one dataset on its respective test set, the remaining tasks’ datasets and additionally the KMNIST ([Clanuwat et al., 2018](#)), SVHN ([Netzer et al., 2011](#)) and CIFAR ([Krizhevsky, 2009](#)) datasets.

We compare and contrast three criteria that could be used for open set recognition: classifier predictive entropy, reconstruction loss and our proposed latent based EVT approach. Naively one might expect the Bayesian approach to handle unknown data through uncertainty. We thus approximate the expectation with 100 variational samples from the approximate posterior per data point. Figure 4 shows the three criteria and respective percentage of the total dataset being considered as outlying for the OCDVAE model trained on FashionMNIST. In consistence with ([Nalisnick et al., 2019](#)), we can observe that use of reconstruction loss can only sometimes distinguish between the known tasks’ test data and unknown datasets. In the case of classifier predictive

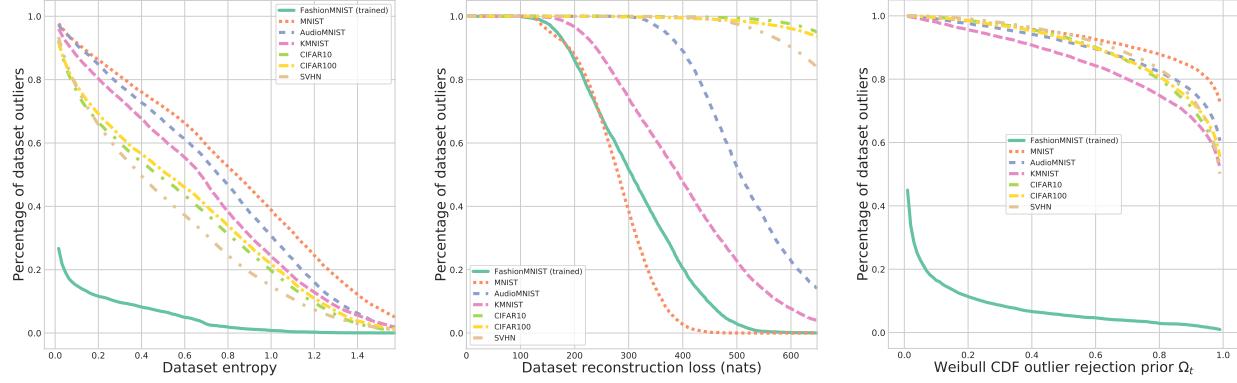


Figure 4. Trained FashionMNIST OCDVAE evaluated on unknown datasets. All metrics are averaged over 100 approximate posterior samples per data point. (Left) Classifier entropy values are insufficient to separate most of unknown from the known task’s test data. (Center) Reconstruction loss allows for a partial distinction. (Right) Our posterior based open set recognition considers the large majority of unknown data as statistical outliers across a wide range of rejection priors Ω_t .

Table 2. Test accuracies and outlier detection values of the joint OCDVAE and dual model (VAE and separate deep classifier) approaches when considering 95 % of known tasks’ validation data is inlying. Percentage of detected outliers is reported based on classifier predictive entropy, reconstruction loss and our posterior based EVT approach, averaged over 100 $z \sim q_\theta(z|x)$ samples per data-point respectively. Note that larger values are better, except for the test data of the trained dataset, where ideally 0% should be considered as outlying.

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

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 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 and the majority of other datasets is consistently regarded as outlying by our proposed open set mechanism.

We provide quantitative outlier detection accuracies in ta-

ble 2. Here, a 5% validation split is used 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. We provide this evaluation for both our joint model, as well as separate discriminative and generative models. While MNIST seems to be an easy to identify dataset for all approaches, we can make two major observations:

1. The latent based EVT approach generally outperforms

Table 3. Results for class incremental continual learning approaches averaged over 5 runs, baselines and the reference isolated learning scenario for the three datasets. α_T indicates the respective accuracy at the end of the last increment $T = 5$.

Class-incremental		$\alpha_T(\%)$ (T=5)		
		base	new	all
FashionMNIST	CDVAE ISO		89.54	
	CDVAE UB	92.20	97.50	89.24
	CDVAE LB	00.00	99.80	19.97
	EWC	00.17 \pm 0.076	99.60 \pm 0.023	20.06 \pm 0.059
	Dual Model	94.26 \pm 0.192	93.55 \pm 0.708	63.21 \pm 1.957
	CDVAE	39.51 \pm 7.173	96.92 \pm 0.774	58.82 \pm 2.521
	OCDVAE	60.63 \pm 12.16	96.51 \pm 0.707	69.88 \pm 1.712
MNIST	CDVAE ISO		99.45	
	CDVAE UB	99.57	99.10	99.29
	CDVAE LB	00.00	99.85	20.16
	EWC	00.45 \pm 0.059	99.58 \pm 0.052	20.26 \pm 0.027
	Dual Model	97.31 \pm 0.489	98.59 \pm 0.106	96.64 \pm 0.079
	CDVAE	19.86 \pm 7.396	99.00 \pm 0.100	64.34 \pm 4.903
	OCDVAE	92.35 \pm 4.485	99.06 \pm 0.171	93.24 \pm 3.742
AudioMNIST	CDVAE ISO		97.75	
	CDVAE UB	98.42	98.67	97.87
	CDVAE LB	00.00	100.0	20.02
	EWC	00.11 \pm 0.007	99.41 \pm 0.207	19.98 \pm 0.032
	Dual Model	61.58 \pm 0.747	89.41 \pm 0.691	47.42 \pm 1.447
	CDVAE	59.36 \pm 7.147	84.93 \pm 6.297	81.49 \pm 1.944
	OCDVAE	79.73 \pm 4.070	89.52 \pm 6.586	87.72 \pm 1.594

the other criteria, particularly for the OCDVAE where a near perfect open set detection can be achieved.

- Even though we can apply EVT to a purely discriminative model, the joint OCDVAE model consistently exhibits more accurate outlier detection. We hypothesize that this is due to the joint model also optimizing a variational lower bound to the data distribution $p(\mathbf{x})$.

We provide figures similar to figure 4 for all models reported in table 2 in the supplementary material.

4.2. Learning classes incrementally

We show results in analogy to table 1 for the class incremental scenario in table 3. With the exception of MNIST, where the dual model approach fares well, a similar pattern as before can be observed in this more challenging scenario and our proposed OCDVAE approach significantly outperforms all other methods. Interestingly, as a result of using a single model across tasks, we observe backward transfer in some experiments. This is particularly apparent for AudioMNIST, where addition of the second increment first decays and inclusion of later tasks improves the second task's accuracy. We provide a detailed account of all intermediate results and examples of generated images for all increments $t = 1, \dots, 5$ in the supplementary material.

Table 4. Results for PixelVAE based continual learning approaches averaged over 5 runs in analogy to tables 1 and 3.

		Class-incremental		
		base	new	all
Fashion	Dual Pix Model	60.04 \pm 5.151	98.85 \pm 0.141	72.41 \pm 2.941
	PixCDVAE	47.83 \pm 13.41	97.91 \pm 0.596	63.05 \pm 1.826
	PixOCDVAE	74.45 \pm 2.889	98.63 \pm 0.176	80.85 \pm 0.721
MNIST	Dual Pix Model	98.04 \pm 1.397	97.31 \pm 0.575	96.52 \pm 0.658
	PixCDVAE	56.53 \pm 4.032	96.77 \pm 0.337	83.61 \pm 0.927
	PixOCDVAE	97.44 \pm 0.785	98.63 \pm 0.430	96.84 \pm 0.346
Audio	Dual Pix Model	64.60 \pm 8.739	98.18 \pm 0.885	75.50 \pm 3.032
	PixCDVAE	29.94 \pm 18.47	97.00 \pm 0.520	63.44 \pm 5.252
	PixOCDVAE	75.25 \pm 10.18	99.43 \pm 0.495	90.23 \pm 1.139
		Cross-dataset		
		base	new	all
F-M-A	Dual Pix Model	82.88 \pm 0.116	97.23 \pm 0.212	92.16 \pm 0.061
	PixCDVAE	56.44 \pm 1.831	97.50 \pm 0.184	80.76 \pm 0.842
	PixOCDVAE	81.84 \pm 0.212	97.75 \pm 0.169	91.76 \pm 0.212
A-M-F	Dual Pix Model	71.58 \pm 2.536	88.76 \pm 0.255	88.61 \pm 0.547
	PixCDVAE	49.38 \pm 2.256	88.54 \pm 0.042	82.18 \pm 0.672
	PixOCDVAE	91.90 \pm 0.282	89.91 \pm 0.177	93.82 \pm 0.354

4.3. Improving the generative model

In a final empirical evaluation, we investigate the choice of generative model and optionally enhance the probabilistic decoder with an autoregressive variant, where generation of a pixel's value is spatially conditioned on previous pixels (van den Oord et al., 2016; Gulrajani et al., 2017; Chen et al., 2017). We show results corresponding to tables 1 and 3 for pixel models in table 4. While we can observe that this generally further alleviates catastrophic forgetting induced by multiple successive iterations of generative replay, it does significantly more so for our proposed approach.

5. Summary and outlook

We have proposed a probabilistic approach to unify continual deep learning with open set recognition based on Bayesian inference. Using a single model that combines a shared probabilistic encoder with a generative model and an expanding linear classifier, we have introduced EVT based bounds to the approximate posterior. The derived open set recognition and corresponding generative replay with statistical outlier rejection have been shown to achieve compelling results in both task incremental as well as cross-dataset continual learning across image and audio modalities, while being able to distinguish seen from unseen data. Our approach readily benefits from recent advances such as autoregressive models (Gulrajani et al., 2017; Chen et al., 2017) and we therefore expect future application to extend to more complicated data such as larger scale color images.

References

- Achille, A., Eccles, T., Matthey, L., Burgess, C. P., Watters, N., Lerchner, A., and Higgins, I. Life-Long Disentangled Representation Learning with Cross-Domain Latent Homologies. *Neural Information Processing Systems (NeurIPS)*, 2018.
- Bachem, O., Lucic, M., and Krause, A. Coresets for Non-parametric Estimation - the Case of DP-Means. *International Conference on Machine Learning (ICML)*, 37: 209–217, 2015.
- Bauer, M. and Mnih, A. Resampled Priors for Variational Autoencoders. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 89, 2019.
- Becker, S., Ackermann, M., Lapuschkin, S., Müller, K.-R., and Samek, W. Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals. *arXiv preprint arXiv: 1807.03418*, 2018.
- Bendale, A. and Boult, T. E. Towards Open Set Deep Networks. *Computer Vision and Pattern Recognition (CVPR)*, 2016.
- Boult, T. E., Cruz, S., Dhamija, A., Gunther, M., Henry-doss, J., and Scheirer, W. Learning and the Unknown : Surveying Steps Toward Open World Recognition. *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.
- Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in beta-VAE. *Neural Information Processing Systems (NeurIPS), Workshop on Learning Disentangled Representations*, 2017.
- Chen, X., Kingma, D. P., Salimans, T., Duan, Y., Dhariwal, P., Schulman, J., Sutskever, I., and Abbeel, P. Variational Lossy Autoencoder. *International Conference on Learning Representations (ICLR)*, 2017.
- Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., and Ha, D. Deep Learning for Classical Japanese Literature. *Neural Information Processing Systems (NeurIPS), Workshop on Machine Learning for Creativity and Design*, 2018.
- Dhamija, A. R., Günther, M., and Boult, T. E. Reducing Network Agnostophobia. *Neural Information Processing Systems (NeurIPS)*, 2018.
- Farquhar, S. and Gal, Y. A Unifying Bayesian View of Continual Learning. *Neural Information Processing Systems (NeurIPS) Bayesian Deep Learning Workshop*, 2018.
- Gal, Y. and Ghahramani, Z. Dropout as a Bayesian Approximation : Representing Model Uncertainty in Deep Learning. *International Conference on Machine Learning (ICML)*, 48, 2015.
- Gepperth, A. and Karaoguz, C. A Bio-Inspired Incremental Learning Architecture for Applied Perceptual Problems. *Cognitive Computation*, 8(5):924–934, 2016.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative Adversarial Nets. *Neural Information Processing Systems (NeurIPS)*, 2014.
- Gulrajani, I., Kumar, K., Faruk, A., Taiga, A. A., Visin, F., Vazquez, D., and Courville, A. PixelVAE: a Latent Variable Model for Natural Images. *International Conference on Learning Representations (ICLR)*, 2017.
- He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *International Conference on Computer Vision (ICCV)*, 2015.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep Residual Learning for Image Recognition. *Computer Vision and Pattern Recognition (CVPR)*, 2016.
- Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *International Conference on Learning Representations (ICLR)*, 2017.
- Hinton, G. E., Vinyals, O., and Dean, J. Distilling the Knowledge in a Neural Network. *NeurIPS Deep Learning Workshop*, 2014.
- Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 1997.
- Hoffman, M. D. and Johnson, M. J. ELBO surgery: yet another way to carve up the variational evidence lower bound. *Neural Information Processing Systems (NeurIPS), Advances in Approximate Bayesian Inference Workshop*, 2016.
- Ioffe, S. and Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *International Conference on Machine Learning (ICML)*, 2015.
- Kemker, R., McClure, M., Abitino, A., Hayes, T., and Kanan, C. Measuring Catastrophic Forgetting in Neural Networks. *AAAI Conference on Artificial Intelligence (AAAI)*, 2018.
- Kingma, D. P. and Ba, J. L. Adam: a Method for Stochastic Optimization. *International Conference on Learning Representations (ICLR)*, 2015.
- Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. *International Conference on Learning Representations (ICLR)*, 2013.

- Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., and Hadsell, R. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences (PNAS)*, 114(13):3521–3526, 2017.
- Krizhevsky, A. Learning Multiple Layers of Features from Tiny Images. Technical report, Toronto, 2009.
- LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998.
- Lee, K., Lee, H., Lee, K., and Shin, J. Training Confidence-Calibrated Classifiers for Detecting Out-of-Distribution Samples. *International Conference on Learning Representations (ICLR)*, 2018.
- Li, Z. and Hoiem, D. Learning without forgetting. *European Conference on Computer Vision (ECCV)*, 2016.
- Liang, S., Li, Y., and Srikanth, R. Enhancing the Reliability of Out-of-distribution Image Detection in Neural Networks. *International Conference on Learning Representations (ICLR)*, 2018.
- Lopez-Paz, D. and Ranzato, M. A. Gradient Episodic Memory for Continual Learning. *Neural Information Processing Systems (NeurIPS)*, 2017.
- Matan, O., Kiang, R., Stenard, C. E., and Boser, B. E. Handwritten Character Recognition Using Neural Network Architectures. *4th USPS Advanced Technology Conference*, 2(5):1003–1011, 1990.
- Mathieu, E., Rainforth, T., Siddharth, N., and Teh, Y. W. Disentangling disentanglement in variational autoencoders. *International Conference on Machine Learning (ICML)*, pp. 7744–7754, 2019.
- McCloskey, M. and Cohen, N. J. Catastrophic Interference in Connectionist Networks : The Sequential Learning Problem. *Psychology of Learning and Motivation - Advances in Research and Theory*, 24(C):109–165, 1989.
- Mensink, T., Verbeek, J., Perronnin, F., Csurka, G., Mensink, T., Verbeek, J., Perronnin, F., and Csurka, G. Metric Learning for Large Scale Image Classification : Generalizing to New Classes at Near-Zero Cost. *European Conference on Computer Vision (ECCV)*, 2012.
- Nalisnick, E., Matsukawa, A., Teh, Y. W., Gorur, D., and Lakshminarayanan, B. Do Deep Generative Models Know What They Don't Know? *International Conference on Learning Representations (ICLR)*, 2019.
- Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading Digits in Natural Images with Unsupervised Feature Learning. *Neural Information Processing Systems (NeurIPS), Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.
- Nguyen, C. V., Li, Y., Bui, T. D., and Turner, R. E. Variational Continual Learning. *International Conference on Learning Representations (ICLR)*, 2018.
- O'Reilly, R. C. and Norman, K. A. Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in Cognitive Sciences*, 6(12):505–510, 2003.
- Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. Continual Lifelong Learning with Neural Networks: A Review. *Neural Networks*, 113:54–71, 2019.
- Rebuffi, S. A., Kolesnikov, A., Sperl, G., and Lampert, C. H. iCaRL: Incremental classifier and representation learning. *Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Robins, A. Catastrophic Forgetting, Rehearsal and Pseudorehearsal. *Connection Science*, 7(2):123–146, 1995.
- Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., and Hadsell, R. Progressive Neural Networks. *arXiv preprint arXiv: 1606.04671*, 2016.
- Shin, H., Lee, J. K., and Kim, J. J. Continual Learning with Deep Generative Replay. *Neural Information Processing Systems (NeurIPS)*, 2017.
- Takahashi, H., Iwata, T., Yamanaka, Y., Yamada, M., and Yagi, S. Variational Autoencoder with Implicit Optimal Priors. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:5066–5073, 2019.
- Thomas, M. R. P., Ahrens, J., and Tashev, I. Probability Models For Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014.
- Tomczak, J. M. and Welling, M. VAE with a vampprior. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 84, 2018.
- van den Oord, A., Kalchbrenner, N., and Kavukcuoglu, K. Pixel Recurrent Neural Networks. *International Conference on Machine Learning (ICML)*, 48:1747–1756, 2016.
- Xiao, H., Rasul, K., and Vollgraf, R. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv: 1708.07747*, 2017.

Yoon, J., Yang, E., Lee, J., and Hwang, S. J. Lifelong Learning with Dynamically Expandable Networks. *International Conference on Learning Representations (ICLR)*, 2018.

Zagoruyko, S. and Komodakis, N. Wide Residual Networks. *British Machine Vision Conference (BMVC)*, 2016.

Zenke, F., Poole, B., and Ganguli, S. Continual Learning Through Synaptic Intelligence. *International Conference on Machine Learning (ICML)*, 70:3987–3995, 2017.

Supplementary material

The supplementary material provides further details for the material presented in the main body. Specifically, the structure is as follows:

- A.** Full specification of training procedure and hyper-parameters, including exact architecture definitions.
- B.** Extended discussion, qualitative and quantitative examples for the role of β . Further visualization of continually learned 2-D latent encodings as an extension to figure 2 in the main body.
- C.** Additional visualization of open set detection for all quantitatively evaluated models considered in table 2 of the main body.
- D.** Full continual learning results for all task increments, including all reconstruction losses and KL divergences.
- E.** Visualization of generative replay examples for all models and task increments.

A. Training hyper-parameters and architecture definitions

We provide a full specification of hyper-parameters, model architectures and the training procedure in this section.

We base our encoder and decoder architecture on 14-layer wide residual networks (He et al., 2016; Zagoruyko & Komodakis, 2016) as used in lossy auto-encoders (Gulrajani et al., 2017; Chen et al., 2017), with a latent dimensionality of 60 to demonstrate scalability to high-dimensions and deep networks. These architectures are shown in detail in tables 5 and 6. Hidden layers include batch-normalization (Ioffe & Szegedy, 2015) with a value of 10^{-5} and use ReLU activations. For a common frame of reference, all methods' share the same underlying WRN architecture. For the autoregressive addition to our joint model, we set the number of output channels of the decoder to 60 and append 3 additional pixel decoder layers, each with a kernel size of 7×7 and 60 channels. While we will report reconstruction log-likelihoods in nats, these models are practically formulated as a classification problem with a 256-way Softmax. The corresponding loss is in bits per dimension. We have converted these values to have a better comparison, but 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 dimensions.

We use hyper-parameters consistent with the literature (Gulrajani et al., 2017; Chen et al., 2017). Accordingly, all models are optimized using stochastic gradient descent with a mini-batch size of 128 and Adam (Kingma & Ba, 2015)

Table 5. 14-layer 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. Every convolutional layer is followed by batch-normalization and a ReLU activation function. 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 $3 \times 3 - 48$, p = 1	
	conv $3 \times 3 - 160$, p = 1;	conv $1 \times 1 - 160$ (skip next layer)
	conv $3 \times 3 - 160$, p = 1;	shortcut (skip next layer)
	conv $3 \times 3 - 160$, p = 1	
Block 2	conv $3 \times 3 - 320$, s = 2, p = 1;	conv $1 \times 1 - 320$, s = 2 (skip next layer)
	conv $3 \times 3 - 320$, p = 1	
	conv $3 \times 3 - 320$, p = 1;	shortcut (skip next layer)
	conv $3 \times 3 - 320$, p = 1	
Block 3	conv $3 \times 3 - 640$, s = 2, p = 1;	conv $1 \times 1 - 640$, s = 2 (skip next layer)
	conv $3 \times 3 - 640$, p = 1	
	conv $3 \times 3 - 640$, p = 1;	shortcut (skip next layer)
	conv $3 \times 3 - 640$, p = 1	

with a learning rate of 0.001 and first and second momenta equal to 0.9 and 0.999. As detailed in the main body, we add noise sampled from $\mathcal{N}(0, 0.25)$ to the input to avoid over-fitting. Due to the inevitable data augmentation effect, we train all approaches in this denoising fashion. No further data augmentation or preprocessing is applied. We initialize all weights according to He et al. (2015).

All class incremental models are trained for 120 epochs per task on MNIST and FashionMNIST and 150 epochs on AudioMNIST. Complementary incremental cross-dataset models are trained for 200 epochs per task on data resized to 32×32 . While our proposed model exhibits forward 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 achievable accuracy of other methods. Isolated models are trained for 200 and 300 epochs until convergence respectively. For the generative replay with statistical outlier rejection, we use an aggressive rejection rate of $\Omega_t = 0.01$ (with 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 is the cosine distance.

For EWC, the number of Fisher samples is fixed to the total number of data points from all the previously seen tasks. A suitable Fisher multiplier value λ has been 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 observe exploding gradients if λ is too high. However, a very small λ leads to excessive drift in the weight distribution across subsequent tasks that

Table 6. 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 are followed by batch-normalization and a ReLU activation function. The model ends on a Sigmoid function.

Layer type	WRN decoder	
Layer 1	$\text{FC} 640 \times [P_w/4] \times [P_h/4]$	
Block 1	conv_t 3 × 3 - 320, p = 1;	conv_t 1 × 1 - 320 (skip next layer)
	conv 3 × 3 - 320, p = 1	
	conv 3 × 3 - 320, p = 1;	shortcut (skip next layer)
	conv 3 × 3 - 320, p = 1	
	upsample × 2	
Block 2	conv_t 3 × 3 - 160, p = 1;	conv_t 1 × 1 - 160 (skip next layer)
	conv 3 × 3 - 160, p = 1	
	conv 3 × 3 - 160, p = 1;	shortcut (skip next layer)
	conv 3 × 3 - 160, p = 1	
	upsample × 2	
Block 3	conv_t 3 × 3 - 48, p = 1;	conv_t 1 × 1 - 48 (skip next layer)
	conv 3 × 3 - 48, p = 1	
	conv 3 × 3 - 48, p = 1;	shortcut (skip next layer)
	conv 3 × 3 - 48, p = 1	
	Layer 2	conv 3 × 3 - 3, p = 1

further results in catastrophic inference. Empirically, $\lambda = 500$ in the class-incremental scenario and $\lambda = 1000$ in the cross-dataset setting seem to provide the best balance.

B. Further discussion on the role of β

In the main body the role of the β term (Higgins et al., 2017) in our model's loss function is pointed out. Here, we delve into further detail with qualitative and quantitative examples to support the arguments. To facilitate the discussion, we repeat equation 1 of the main body:

$$\begin{aligned} \mathcal{L}(\mathbf{x}^{(n)}, \mathbf{y}^{(n)}; \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\xi}) &= -\beta KL(q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^{(n)}) \parallel p(\mathbf{z})) \\ &+ \mathbb{E}_{q_{\boldsymbol{\theta}}(\mathbf{z}|\mathbf{x}^{(n)})} [\log p_{\boldsymbol{\phi}}(\mathbf{x}^{(n)}|\mathbf{z}) + \log p_{\boldsymbol{\xi}}(\mathbf{y}^{(n)}|\mathbf{z})] \end{aligned}$$

The β term weights the strength of the regularization by the prior through the KL divergence. 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 Hoffman & Johnson (2016) and Burgess et al. (2017): too large β values (typically $>> 1$) will result in a collapse of any structure present in the aggregate posterior. Too small β values (typically $<< 1$) lead to the latent space being a lookup table. In either case, there is no meaningful information between the latents. This is particularly relevant to our objective of linear class separability, that requires formation of an aggregate latent encoding that is disentangled with respect to the different classes. To visualize this, we

Table 7. Losses obtained for different β values for MNIST using the WRN architecture with 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.

2-D latent	Beta	In nats per dimension (nats in brackets)			Accuracy [%]
		KLD	Recon loss	Class Loss	
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 8. Losses obtained for different β values for MNIST using the WRN architecture with 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.

60-D latent	Beta	In nats per dimension (nats in brackets)			Accuracy [%]
		KLD	Recon loss	Class Loss	
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

have 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 aggregate encodings at the end of training are shown in figure 5. Here, we can empirically observe above points. With a beta of one and larger, the aggregate posterior's structure starts to collapse and the aggregate encoding converges to a Normal distribution. 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 gets ever smaller there is insufficient regularization present and the aggregate posterior no longer follows a Normal distribution. 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.

It is important to note that the we normalize losses with respect to dimensions, and the value of β should thus also be seen as a normalized quantity. While the relative effect of increasing or decreasing beta stays the same, the absolute value of β can be subject to any normalization.

To provide corresponding quantitative examples for the models trained with different β with 2-D latent spaces and 60-D latent spaces in tables 7 and 8 respectively. In both cases we

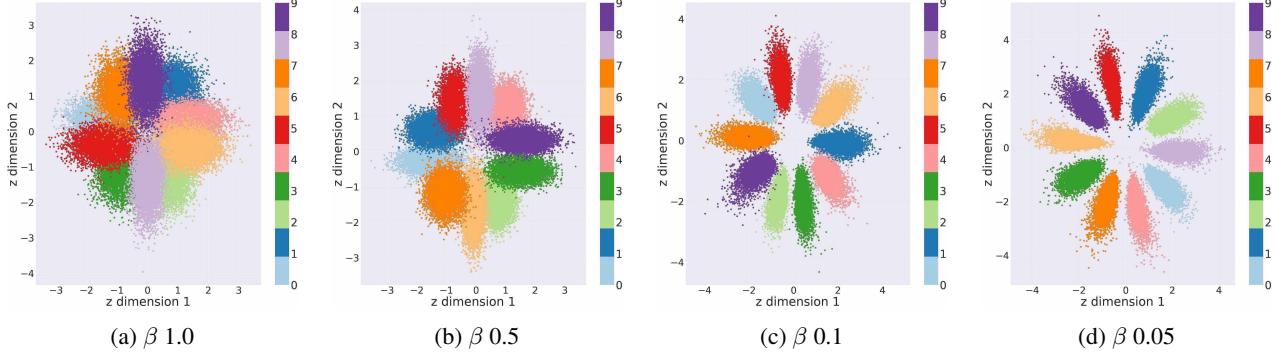


Figure 5. 2-D MNIST latent space visualization with different β values for the used WRN architecture.

observe that decreasing the value of beta below one is necessary to improve classification accuracy, 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.

To conclude our discussion on the role of beta, we show the full set of time steps to complete figure 2 of the main body. Here, a model with 2-D latent space and $\beta = 1$ was trained continually on MNIST and the aggregate encoding is visualized for each additionally arriving set of two classes in figure 6. In the figure's last panel, we further show the aggregate latent encoding at the last time step for the model trained in an autoregressive fashion, i.e. a PixelVAE (van den Oord et al., 2016; Gulrajani et al., 2017). Also coined "lossy autoencoder" by Chen et al. (2017), the authors argue that this model leaves the encoding of local structure to the autoregressive pixel decoder, with a focus on global structure in the latent encoding. In our visualization, this seems to be reflected in a change of the aggregate posterior's structure.

C. Additional open set recognition visualization

As we point out in section 4 of the main paper, our posterior based open set recognition considers almost all of the unknown datasets as statistical outliers, while at the same time regarding unseen test data from the originally trained tasks as distribution inliers across a wide range of rejection priors. In addition to the outlier rejection curves for FashionMNIST and the quantitative results presented in the main body, we also show the full outlier rejection curves for the remaining datasets, as well as all dual model approaches in figures 7, 8 and 9. These figures visually support the quantitative findings described in the main body and respective conclusions. In summary, the joint OCDVAE performs better at open set recognition in direct comparison to the dual model setting, particularly when using the EVT based criterion. Apart from the MNIST dataset, where reconstruction loss can be a sufficient metric for open set detection, the latent based approach also exhibits less dependency on the outlier rejection prior and consistently improves the ability to discern unknown data.

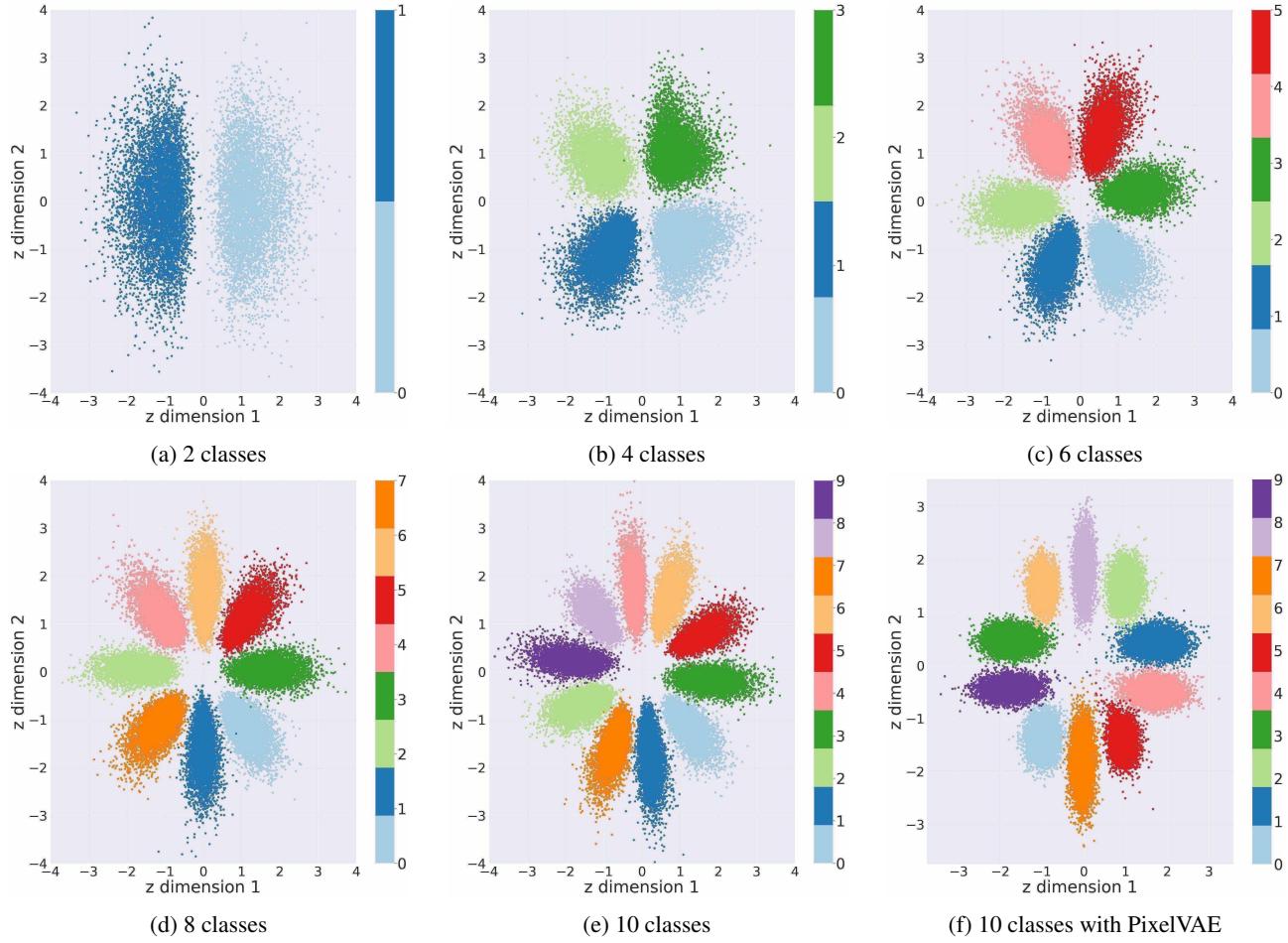
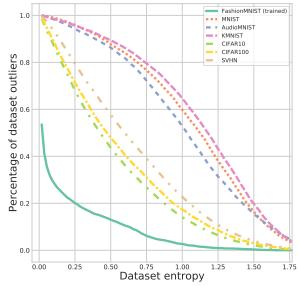
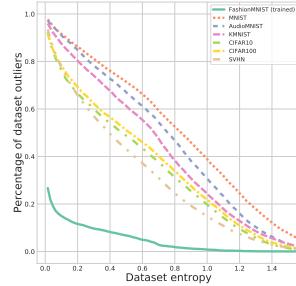


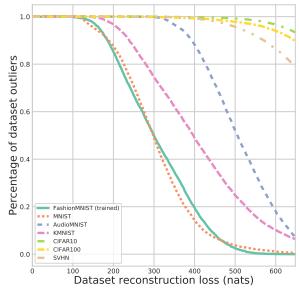
Figure 6. 2-D latent space visualization for continually learned MNIST at the end of every task increment as an extension of figure 2 in the main body (a-e). Panel (f) shows corresponding visualization at the end of training for all task increments with an autoregressive decoder.



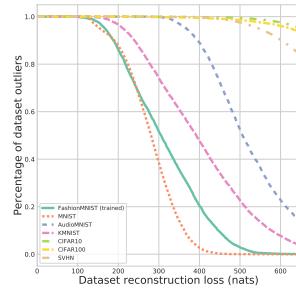
(a) Dual model classifier entropy based OSR



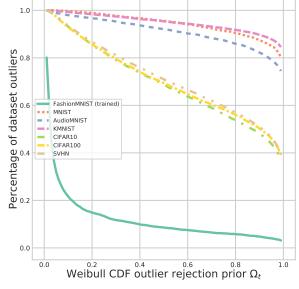
(b) OCDVAE classifier entropy based OSR



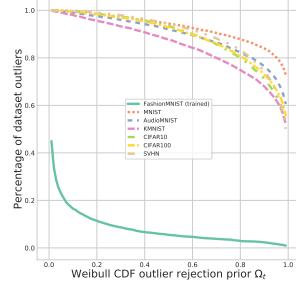
(c) Dual model reconstruction loss based OSR



(d) OCDVAE reconstruction loss based OSR

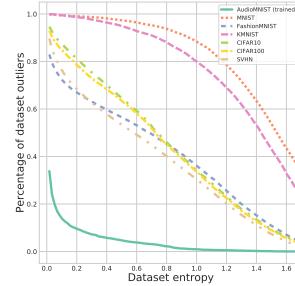


(e) Dual model posterior EVT based OSR

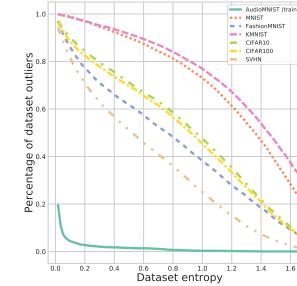


(f) OCDVAE posterior EVT based OSR

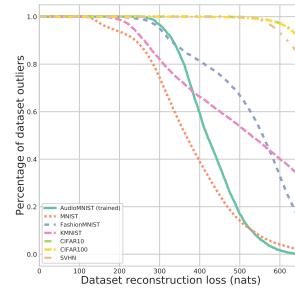
Figure 7. Dual model and OCDVAE trained on FashionMNIST evaluated on unseen datasets. Pairs of panels show the contrast between the approaches. Left panels correspond to the dual model, right panels show the joint OCDVAE model. (a+b) The classifier entropy values by itself are insufficient to separate most of unknown from the known task’s test data. (c+d) Reconstruction loss allows for a partial distinction. (e+f) Our posterior-based open set recognition considers the large majority of unknown data as statistical outliers across a wide range of rejection priors Ω_t , significantly more so in the OCDVAE model. All metrics are reported as the mean over 100 approximate posterior samples per data point.



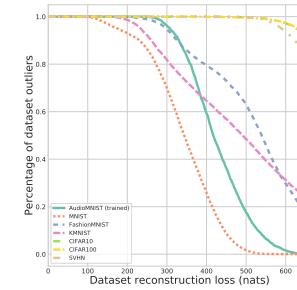
(a) Dual model classifier entropy based OSR



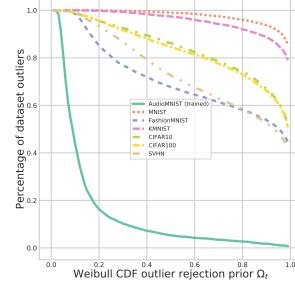
(b) OCDVAE classifier entropy based OSR



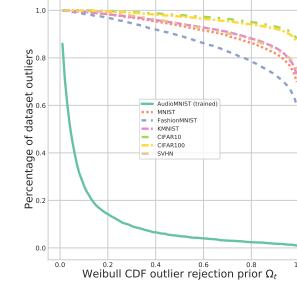
(c) Dual model reconstruction loss based OSR



(d) OCDVAE reconstruction loss based OSR



(e) Dual model posterior EVT based OSR



(f) OCDVAE posterior EVT based OSR

Figure 8. Dual model and OCDVAE trained on AudioMNIST evaluated on unseen datasets. Pairs of panels show the contrast between the approaches. Left panels correspond to the dual model, right panels show the joint OCDVAE model. (a+b) The classifier entropy values by itself are insufficient to separate most of unknown from the known task’s test data. (c+d) Reconstruction loss allows for a partial distinction. (e+f) Our posterior-based open set recognition considers the large majority of unknown data as statistical outliers across a wide range of rejection priors Ω_t , significantly more so in the OCDVAE model. All metrics are reported as the mean over 100 approximate posterior samples per data point.

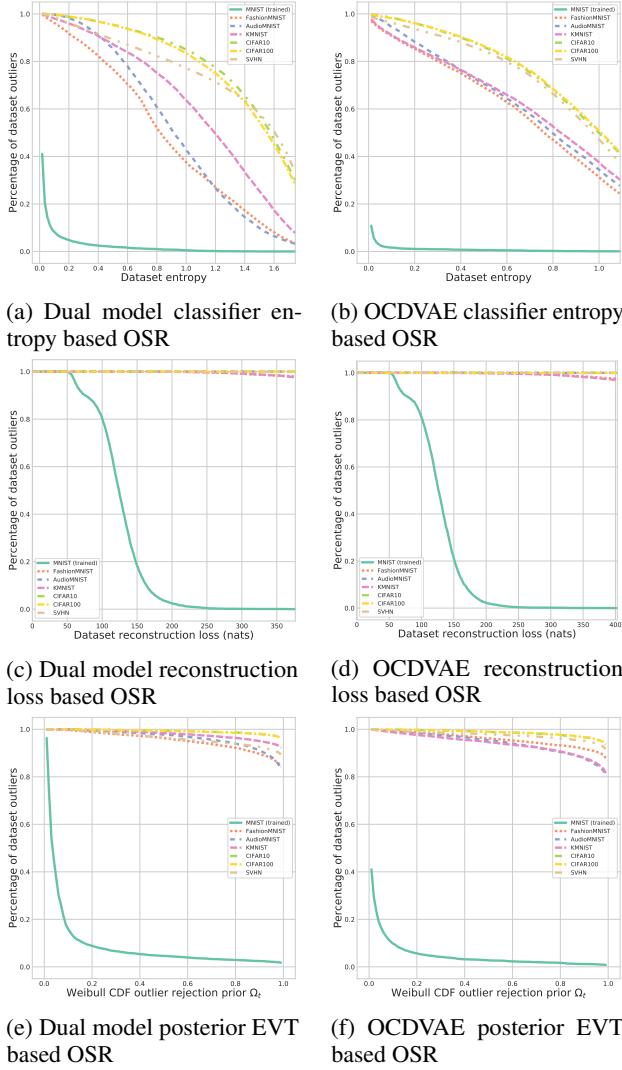


Figure 9. Dual model and OCDVAE trained on MNIST evaluated on unseen datasets. Pairs of panels show the contrast between the approaches. Left panels correspond to the dual model, right panels show the joint OCDVAE model. (a+b) The classifier entropy values by itself can achieve a partial separation between unknown and the known task’s test data. (c+d) Reconstruction loss allows for distinction if the cut-off is chosen correctly. (e+f) Our posterior-based open set recognition considers the large majority of unknown data as statistical outliers across a wide range of rejection priors Ω_t . All metrics are reported as the mean over 100 approximate posterior samples per data point. While the OCDVAE shows improvement upon the dual model approach, particularly if using classifier entropies for OSR, both models trained on MNIST perform well in OSR. In direct contrast with models trained on Fashion- or AudioMNIST and respective figures 7 and 8, this shows that evaluation on MNIST alone is generally insufficient.

D. Detailed results

In the main body we have reported three metrics for our continual learning experiments based on classification accuracy: 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 $\alpha_{t,all}$. This is an appropriate measure to evaluate the quality of the generative model over time given that the employed mechanism to avoid catastrophic inference in continual learning is generative replay. On the one hand, if catastrophic inference occurs in the decoder the sampled data will no longer resemble the instances of the observed data distribution. This will in turn degrade the encoder during continued training and thus the classification accuracy. On the other hand, this proxy measure for the generation quality avoids the common pitfalls of pixel-wise reconstruction metrics. The information necessary to maintain respective knowledge of the data distribution through the variational approximation in the probabilistic encoder does not necessarily rely on correctly reconstructing data’s local information. To take an example, if a model were to reconstruct all images perfectly but with some degree of spatial translation or rotation, then the negative log likelihood (NLL) would arguably be worse than that of a model which reconstructs local details correctly on a pixel level for a fraction of the image. As this could be details in e.g. the background or other class unspecific areas, training on corresponding generations does not have to prevent loss of encoder knowledge with respect to the classification task.

As such, a similar argument can be conjured for the KL divergence. On the one hand, monitoring the KL divergence as a regularization term by itself over the course of continual learning is meaningless without regarding the data’s NLL. On the other hand, for our OCDVAE model the exact value of the KL divergence does not immediately reflect the quality of the generated data. This is because we do not sample merely from the prior, but as explained in the main body employ a rejection mechanism to draw samples that belong to the aggregate posterior.

Nevertheless, for the purpose of completeness and in addition to the results provided in the experimental section of the main body, we provide the reconstruction losses and KL divergences for all applicable models in this supplementary material section. Analogous to the three metrics for classification accuracy of base, new and all tasks, we define the respective reconstruction losses $\gamma_{t,base}$, $\gamma_{t,new}$ and $\gamma_{t,all}$. The KL divergence KL_t always measures the deviation from the prior $p(z)$ at any point in time, as the prior remains the same throughout continual training. Following the above discussion, we argue that these values should be regarded with caution and should not be interpreted separately.

D.1. Full cross dataset results

We show the full cross dataset results in table 9 in extension to table 1 in the main body. An analogous table for the presented autoregressive models can be found in table 10. Similar to the accuracy values, we can observe that the mismatch between aggregate posterior and prior as expressed through the KL divergence is greater in a naive joint model (naive CDVAE) in comparison to a dual model approach with separate generative and discriminative models. Our proposed OCDVAE model, with respective rejection sampling scheme that takes into account the structure of the aggregate posterior, alleviates this to a large degree. The reconstruction losses of both the dual model and the joint OCDVAE approach show only negligible deviation with respect to the achievable upper bound and only limited catastrophic inference of the decoder occurs. However, we can also observe that by itself these quantities are not indicative of maintaining encoder knowledge with respect to representations required for classification. This is particularly visible in the tables’ second experiment, where we first train Audio data and then proceed with the two image datasets. Here, the KL divergence and reconstruction loss are both better for the dual model, whereas a much higher accuracy over time is maintained in the OCDVAE model. Naturally, this is because a significant mismatch between aggregate posterior and prior is also present in a purely unsupervised generative model and naively sampling from the prior will result in generated instances that do not resemble those present in the observed data distribution. While weaker in effect, this is similar to the naive CDVAE approach. Without the presence of the linear discriminator on the latents in the purely unsupervised generative model, there is however no straightforward mechanism to disentangle the latent space according to classes. Our proposed open set approach and the resulting constraint to samples from the aggregate posterior as presented in the OCDVAE is thus not trivially applicable.

D.2. Full class incremental results

In addition to reconstruction losses and KL divergences, we also report the detailed full set of intermediate results for the five task steps of the class incremental scenario. We thus extend table 3 in the main body with results for all task increments $t = 1, \dots, 5$ and a complete list of losses in tables 11, 12 and 13 for the three datasets respectively. The corresponding results for autoregressive models are presented in tables 14, 15 and 16.

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

Table 9. Results for incremental cross-dataset continual learning approaches averaged over 5 runs, baselines and the reference isolated learning scenario for FashionMNIST (F) → MNIST (M) → AudioMNIST (A) and the reverse order. Extension of table 1 in the main body. Here, in addition to the accuracy α_T , γ_T and KL_T also indicate the respective NLL reconstruction metrics and corresponding KL divergences at the end of the last increment $T = 3$.

Cross-dataset		$\alpha_T(\%)$			$\gamma_T(\text{nats})$			$KL_T(\text{nats})$
		base	new	all	base	new	all	all
F-M-A	CDVAE ISO			94.95			269.6	24.97
	CDVAE UB	89.10	97.88	95.00	311.2	434.3	269.7	25.20
	CDVAE LB	00.00	98.12	22.70	689.7	341.0	511.7	98.74
	EWC	22.85 ± 0.294	93.31 ± 0.138	43.42 ± 0.063				
	Dual Model	81.89 ± 0.104	96.78 ± 0.067	91.75 ± 0.064	320.0 ± 1.275	431.1 ± 1.474	273.7 ± 1.174	12.80 ± 0.060
	CDVAE	57.70 ± 4.480	96.73 ± 0.235	81.10 ± 1.769	360.9 ± 20.15	432.1 ± 0.231	296.4 ± 7.966	44.29 ± 4.047
	OCDVAE	80.11 ± 2.922	97.63 ± 0.042	91.13 ± 1.045	345.1 ± 7.446	430.7 ± 0.600	280.2 ± 1.069	25.42 ± 1.876
A-M-F	CDVAE ISO			94.95			269.6	24.97
	CDVAE UB	97.17	89.16	94.91	428.8	311.9	268.2	23.91
	CDVAE LB	00.00	89.72	34.51	506.6	311.0	351.1	34.13
	EWC	3.420 ± 0.026	87.54 ± 0.214	45.42 ± 0.731				
	Dual Model	66.82 ± 0.337	89.15 ± 0.050	87.70 ± 0.102	447.3 ± 6.700	308.5 ± 0.599	270.9 ± 1.299	12.89 ± 0.109
	CDVAE	79.74 ± 2.431	88.50 ± 0.126	89.46 ± 0.600	448.6 ± 5.187	315.1 ± 1.305	281.6 ± 3.205	33.38 ± 0.898
	OCDVAE	94.53 ± 0.283	89.53 ± 0.367	94.06 ± 0.156	433.4 ± 0.424	311.6 ± 0.353	271.2 ± 0.424	23.16 ± 0.121

Table 10. Results for PixelVAE based cross-dataset continual learning approaches averaged over 5 runs in analogy to table 9. Extension of table 4 in the main body. Here, in addition to the accuracy α_T , γ_T and KL_T also indicate the respective NLL reconstruction metrics and corresponding KL divergences at the end of the last increment $T = 3$.

Cross-dataset		$\alpha_T(\%)$			$\gamma_T(\text{nats})$			$KL_T(\text{nats})$
		base	new	all	base	new	all	all
F-M-A	Dual Pix Model	82.88 ± 0.116	97.23 ± 0.212	92.16 ± 0.061	288.5 ± 0.723	437.7 ± 0.404	251.6 ± 0.231	9.025 ± 1.378
	PixCDVAE	56.44 ± 1.831	97.50 ± 0.184	80.76 ± 0.842	289.8 ± 1.283	438.1 ± 0.990	252.6 ± 1.424	29.99 ± 0.629
	PixOCDVAE	81.84 ± 0.212	97.75 ± 0.169	91.76 ± 0.212	288.8 ± 0.141	437.1 ± 0.725	251.8 ± 0.636	21.07 ± 0.248
A-M-F	Dual Pix Model	71.58 ± 2.536	88.76 ± 0.255	88.61 ± 0.547	445.8 ± 1.601	290.4 ± 0.603	255.0 ± 0.533	9.164 ± 1.312
	PixCDVAE	49.38 ± 2.256	88.54 ± 0.042	82.18 ± 0.672	441.4 ± 0.495	287.0 ± 0.212	252.5 ± 0.201	30.60 ± 1.556
	PixOCDVAE	91.90 ± 0.282	89.91 ± 0.177	93.82 ± 0.354	438.5 ± 1.626	289.4 ± 0.356	251.3 ± 0.354	20.35 ± 0.424

overall classification accuracies are higher and are generally decreased by the open set models. For example, in table 11 the MNIST base and overall accuracy deviations of a naive CDVAE are higher than the respective values for OCDVAE starting already from the second task increment. Correspondingly, the accuracy values themselves experience larger decline for CDVAE than for OCDVAE with progressive increments. This difference is not as pronounced at the end of the first task increment because the models haven't been trained on any of their own generated data yet. Successful literature approaches such as the variational generative replay proposed by Farquhar & Gal (2018) 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.

We also see that regularization approaches such as EWC already fail at the first increment. In contrast to the success that has been reported in prior literature (Kirkpatrick et al., 2017; Kemker et al., 2018), this is due to the use of a single classification head. This is intuitive because introduction of new units, as described in the main body, directly confuses the existing classification. Regularization approaches by definition are challenged in this scenario because the weights are not allowed to drift too far away from previous values. For emphasis we repeat that however this scenario is much more practical and realistic than a multi-head scenario with a separate classifier per task. While regularization approaches are largely successful in the latter setting, it is not only restricted to the closed world, but further requires an

oracle at prediction stage to chose the correct classification head. In contrast, our proposed approach requires no knowledge of task labels for prediction and is robust in an open world.

With respect to KL divergences and reconstruction losses we can make two observations. First, the arguments of the previous section hold and by itself 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 reconstruction losses at every increment for all $\gamma_{t,all}$ and respective negative log likelihoods for only the new task $\gamma_{t,new}$ are harder to interpret than the accuracy counterpart. While the latter is normalized between zero and unity, the reconstruction loss of different tasks is expected to fluctuate largely according to the task’s images’ reconstruction 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 reconstruction it is common to observe either a large decrease in negative log likelihood for the newly arriving class, or a big increase depending on the specific 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 reconstruction loss $\gamma_{t,base}$. In analogy to base classification accuracy, this quantity still measures the amount of catastrophic forgetting across time. However, in all tables we can observe that catastrophic forgetting of the decoder as measured by the base reconstruction loss is almost imperceivable. As this is not at all reflected in the respective accuracy over time, it further underlines our previous arguments that reconstruction loss is not necessarily the best metric to monitor in the presented continual learning scenario.

Table 11. Results for class incremental continual learning approaches averaged over 5 runs, baselines and the reference isolated learning scenario for MNIST at the end of every task increment. Extension of table 3 in the main body. Here, in addition to the accuracy α_t , γ_t and KL_t also indicate the respective NLL reconstruction metrics and corresponding KL divergences at the end of every task increment t .

MNIST	t	CDVAE ISO	CDVAE UB	CDVAE LB	EWC	Dual Model	CDVAE	OCDVAE
$\alpha_{base,t}$	1		100.0	100.0	99.88 ± 0.010	99.98 ± 0.023	99.97 ± 0.029	99.98 ± 0.018
	2		99.82	00.00	00.61 ± 0.057	99.77 ± 0.032	97.28 ± 3.184	99.30 ± 0.100
	3		99.80	00.00	00.17 ± 0.045	99.51 ± 0.094	87.66 ± 8.765	96.69 ± 2.173
	(%) 4		99.85	00.00	00.49 ± 0.017	98.90 ± 0.207	54.70 ± 22.84	94.71 ± 1.792
	5		99.57	00.00	00.45 ± 0.059	97.31 ± 0.489	19.86 ± 7.396	92.53 ± 4.485
$\alpha_{new,t}$	1		100.0	100.0	99.88 ± 0.010	99.98 ± 0.023	99.97 ± 0.029	99.98 ± 0.018
	2		99.80	99.85	99.70 ± 0.013	99.81 ± 0.062	99.75 ± 0.127	99.80 ± 0.126
	3		99.67	99.94	99.94 ± 0.002	99.48 ± 0.294	99.63 ± 0.172	99.61 ± 0.055
	(%) 4		99.49	100.0	99.87 ± 0.015	99.46 ± 0.315	99.05 ± 0.470	99.15 ± 0.032
	5		99.10	99.86	99.58 ± 0.052	98.59 ± 0.106	99.00 ± 0.100	99.06 ± 0.171
$\alpha_{all,t}$	1		100.0	100.0	99.88 ± 0.010	99.98 ± 0.023	99.97 ± 0.029	99.98 ± 0.018
	2		99.81	49.92	50.16 ± 0.029	99.79 ± 0.049	98.54 ± 1.638	99.55 ± 0.036
	3		99.72	31.35	33.42 ± 0.027	99.32 ± 0.057	95.01 ± 3.162	98.46 ± 0.903
	(%) 4		99.50	24.82	25.36 ± 0.025	98.56 ± 0.021	81.50 ± 9.369	97.06 ± 1.069
	5	99.45	99.29	20.16	20.26 ± 0.027	96.64 ± 0.079	64.34 ± 4.903	93.24 ± 3.742
$\gamma_{base,t}$	1		63.18	62.08		62.17 ± 0.979	64.34 ± 2.054	62.53 ± 1.166
	2		62.85	126.8		63.69 ± 0.576	74.41 ± 10.89	65.68 ± 1.166
	3		63.36	160.4		67.34 ± 0.445	81.89 ± 10.09	69.29 ± 1.541
	(nats) 4		64.25	126.9		70.41 ± 0.436	90.62 ± 10.08	71.69 ± 1.379
	5		64.99	123.2		75.08 ± 0.623	101.6 ± 8.347	77.16 ± 1.104
$\gamma_{new,t}$	1		63.18	62.08		62.17 ± 0.979	64.34 ± 2.054	62.53 ± 1.166
	2		88.75	87.93		88.03 ± 0.664	89.91 ± 0.107	89.64 ± 3.709
	3		82.53	87.22		83.46 ± 0.992	87.65 ± 0.530	85.37 ± 1.725
	(nats) 4		72.68	74.61		73.23 ± 0.280	79.49 ± 0.489	74.75 ± 0.777
	5		85.88	92.00		89.32 ± 0.626	93.55 ± 0.391	89.68 ± 0.618
$\gamma_{all,t}$	1		63.18	62.08		62.17 ± 0.979	64.34 ± 2.054	62.53 ± 1.166
	2		75.97	107.3		75.64 ± 0.600	82.02 ± 5.488	76.62 ± 1.695
	3		79.58	172.3		81.24 ± 0.262	89.88 ± 3.172	82.95 ± 1.878
	(nats) 4		79.72	203.1		82.92 ± 0.489	95.83 ± 2.747	85.30 ± 1.524
	5	78.12	81.97	163.7		88.29 ± 0.363	107.6 ± 1.724	92.92 ± 2.283
$KL_{all,t}$	1		12.55	13.08		11.81 ± 0.123	13.00 ± 0.897	13.68 ± 0.785
	2		18.50	25.84		16.15 ± 0.149	20.20 ± 1.188	18.01 ± 0.154
	3		20.16	24.28		16.46 ± 0.122	24.24 ± 1.974	20.02 ± 0.161
	(nats) 4		20.48	26.32		16.09 ± 0.177	27.01 ± 1.851	20.26 ± 0.186
	5	22.12	21.02	24.87		16.13 ± 0.225	30.61 ± 1.240	21.02 ± 0.717

Table 12. Results for class incremental continual learning approaches averaged over 5 runs, baselines and the reference isolated learning scenario for FashionMNIST at the end of every task increment. Extension of table 3 in the main body. Here, in addition to the accuracy α_t , γ_t and KL_t also indicate the respective NLL reconstruction metrics and corresponding KL divergences at the end of every task increment t .

Fashion	t	CDVAE ISO	CDVAE UB	CDVAE LB	EWC	Dual Model	CDVAE	OCDVAE
$\alpha_{base,t}$	1		99.65	99.60	99.17 ± 0.037	99.58 ± 0.062	99.55 ± 0.035	99.59 ± 0.082
	2		96.70	00.00	02.40 ± 0.122	94.50 ± 0.389	92.02 ± 1.175	92.36 ± 2.092
	3		95.95	00.00	01.63 ± 0.032	94.88 ± 0.432	79.26 ± 4.170	83.90 ± 2.310
	(%) 4		91.35	00.00	00.33 ± 0.097	82.25 ± 4.782	50.16 ± 6.658	64.70 ± 2.580
	5		92.20	00.00	00.17 ± 0.076	94.26 ± 0.192	39.51 ± 7.173	60.63 ± 12.16
$\alpha_{new,t}$	1		99.65	99.60	99.17 ± 0.037	99.58 ± 0.062	99.55 ± 0.035	99.59 ± 0.082
	2		95.55	97.95	96.09 ± 0.260	89.31 ± 0.311	90.98 ± 0.626	92.64 ± 2.302
	3		93.35	99.95	99.92 ± 0.012	86.06 ± 2.801	90.26 ± 1.435	83.40 ± 3.089
	(%) 4		84.75	99.90	99.95 ± 0.060	73.63 ± 3.861	85.65 ± 2.127	84.18 ± 2.715
	5		97.50	99.80	99.60 ± 0.023	93.55 ± 0.708	96.92 ± 0.774	96.51 ± 0.707
$\alpha_{all,t}$	1		99.65	99.60	99.17 ± 0.037	99.58 ± 0.062	99.55 ± 0.035	99.59 ± 0.082
	2		95.75	48.97	49.28 ± 0.242	91.91 ± 0.043	91.83 ± 0.730	92.31 ± 1.163
	3		93.02	33.33	34.34 ± 0.009	79.98 ± 0.634	83.35 ± 1.597	86.93 ± 0.870
	(%) 4		87.51	25.00	25.21 ± 0.100	64.37 ± 0.707	64.66 ± 3.204	76.05 ± 1.391
	5	89.54	89.24	19.97	20.06 ± 0.059	63.21 ± 1.957	58.82 ± 2.521	69.88 ± 1.712
$\gamma_{base,t}$	1		209.7	209.8		207.7 ± 1.558	208.9 ± 1.213	209.7 ± 3.655
	2		207.4	240.7		209.0 ± 0.731	212.7 ± 0.579	212.1 ± 0.937
	3		207.6	258.7		213.0 ± 1.854	219.5 ± 1.376	216.9 ± 1.208
	(nats) 4		207.7	243.6		213.6 ± 0.509	223.8 ± 0.837	217.1 ± 0.979
	5		208.4	306.5		217.7 ± 1.510	232.8 ± 5.048	222.8 ± 1.632
$\gamma_{new,t}$	1		209.7	209.8		207.7 ± 1.558	208.9 ± 1.213	209.7 ± 3.655
	2		241.1	240.2		238.7 ± 0.081	241.8 ± 0.502	241.9 ± 0.960
	3		213.6	211.8		211.6 ± 0.543	215.4 ± 0.501	213.0 ± 0.635
	(nats) 4		220.5	219.7		219.5 ± 0.216	223.6 ± 0.381	220.9 ± 0.522
	5		246.2	242.0		242.8 ± 0.898	248.8 ± 0.398	244.0 ± 0.646
$\gamma_{all,t}$	1		209.7	209.8		207.7 ± 1.558	208.9 ± 1.213	209.7 ± 3.655
	2		224.2	240.4		223.8 ± 0.402	226.6 ± 2.31	226.9 ± 0.918
	3		220.7	246.1		221.9 ± 0.648	227.2 ± 0.606	224.9 ± 0.642
	(nats) 4		220.4	238.7		225.1 ± 3.629	230.4 ± 0.524	226.1 ± 0.560
	5	224.8	226.2	275.1		230.5 ± 1.543	242.2 ± 0.754	234.6 ± 0.823
$KL_{all,t}$	1		12.17	12.20		9.710 ± 0.345	13.21 ± 0.635	13.28 ± 0.644
	2		16.54	17.47		10.65 ± 0.101	17.60 ± 0.755	15.56 ± 0.696
	3		18.84	19.34		11.34 ± 0.057	21.25 ± 0.872	17.35 ± 0.307
	(nats) 4		20.06	17.31		10.96 ± 0.106	25.21 ± 0.929	19.81 ± 0.462
	5	23.27	20.27	21.61		11.45 ± 0.228	26.68 ± 0.859	20.47 ± 0.742

Table 13. Results for class incremental continual learning approaches averaged over 5 runs, baselines and the reference isolated learning scenario for AudioMNIST at the end of every task increment. Extension of table 3 in the main body. Here, in addition to the accuracy α_t , γ_t and KL_t also indicate the respective NLL reconstruction metrics and corresponding KL divergences at the end of every task increment t .

Audio	t	CDVAE ISO	CDVAE UB	CDVAE LB	EWC	Dual Model	CDVAE	OCDVAE
$\alpha_{base,t}$	1		99.99	100.0	100.0 ± 0.000	100.0 ± 0.000	99.21 ± 0.568	99.95 ± 0.035
	2		99.92	00.00	00.16 ± 0.040	93.08 ± 5.854	98.98 ± 0.766	98.61 ± 0.490
	3		100.0	00.00	00.29 ± 0.029	83.25 ± 6.844	92.44 ± 1.306	95.12 ± 2.248
	(%) 4		99.92	00.00	00.31 ± 0.015	72.02 ± 0.677	76.43 ± 4.715	86.37 ± 5.63
	5		98.42	00.00	00.11 ± 0.007	61.57 ± 0.747	59.36 ± 7.147	79.73 ± 4.070
$\alpha_{new,t}$	1		99.99	100.0	100.0 ± 0.000	100.0 ± 0.000	99.21 ± 0.568	99.95 ± 0.035
	2		99.75	100.0	99.78 ± 0.019	86.25 ± 8.956	91.82 ± 4.577	89.23 ± 7.384
	3		98.92	99.58	99.25 ± 0.054	95.16 ± 1.490	95.20 ± 1.495	94.43 ± 3.030
	(%) 4		97.33	98.67	97.03 ± 0.019	62.52 ± 4.022	53.02 ± 6.132	72.22 ± 8.493
	5		98.67	100.0	99.41 ± 0.207	89.41 ± 0.691	84.93 ± 6.297	89.52 ± 6.586
$\alpha_{all,t}$	1		99.99	100.0	100.0 ± 0.000	100.0 ± 0.000	99.21 ± 0.568	99.95 ± 0.035
	2		99.83	50.00	50.16 ± 0.119	89.67 ± 1.763	93.84 ± 2.558	93.93 ± 3.756
	3		99.56	33.19	33.28 ± 0.022	78.24 ± 3.315	94.26 ± 1.669	95.70 ± 1.524
	(%) 4		98.60	24.58	24.50 ± 0.017	60.43 ± 4.209	77.90 ± 4.210	85.59 ± 3.930
	5	97.75	97.87	20.02	19.98 ± 0.032	47.42 ± 1.447	81.49 ± 1.944	87.72 ± 1.594
$\gamma_{base,t}$	1		433.7	423.2		422.3 ± 0.573	435.2 ± 15.69	424.2 ± 2.511
	2		422.5	439.4		426.6 ± 2.840	423.9 ± 0.517	425.2 ± 1.402
	3		420.7	429.2		425.0 ± 0.339	422.7 ± 0.690	423.8 ± 1.148
	(nats) 4		419.9	428.5		425.4 ± 0.081	422.8 ± 0.367	423.5 ± 0.937
	5		418.4	432.9		425.2 ± 0.244	422.7 ± 0.182	423.5 ± 0.586
$\gamma_{new,t}$	1		433.7	423.2		422.3 ± 0.573	435.2 ± 15.69	424.2 ± 2.511
	2		381.2	384.1		381.3 ± 2.039	382.5 ± 1.355	385.3 ± 12.56
	3		435.9	436.7		436.8 ± 0.188	436.3 ± 0.639	436.9 ± 0.688
	(nats) 4		485.9	487.1		486.5 ± 0.432	486.7 ± 0.385	486.5 ± 0.701
	5		421.3	425.2		422.4 ± 0.784	423.9 ± 0.681	422.9 ± 0.537
$\gamma_{all,t}$	1		433.7	423.2		422.3 ± 0.573	435.2 ± 15.69	424.2 ± 2.511
	2		401.9	411.8		404.0 ± 2.407	403.2 ± 0.831	403.5 ± 1.274
	3		412.1	418.9		414.4 ± 0.385	413.6 ± 0.410	413.8 ± 0.573
	(nats) 4		430.3	438.4		433.9 ± 0.374	432.4 ± 0.436	432.6 ± 0.862
	5	429.7	427.2	440.4		432.7 ± 0.385	431.4 ± 0.255	430.9 ± 0.541
$KL_{all,t}$	1		11.65	11.20		4.639 ± 0.107	11.78 ± 1.478	11.16 ± 0.713
	2		11.78	13.61		5.135 ± 0.127	15.13 ± 1.128	14.06 ± 1.140
	3		13.40	17.09		5.427 ± 0.105	18.18 ± 1.140	13.61 ± 0.901
	(nats) 4		13.61	14.41		5.243 ± 0.135	22.93 ± 1.134	17.58 ± 1.102
	5	17.89	15.15	14.52		5.470 ± 0.055	22.96 ± 0.912	18.52 ± 1.131

Table 14. Results for PixelVAE based class incremental continual learning approaches averaged over 5 runs, baselines and the reference isolated learning scenario for MNIST at the end of every task increment in analogy to table 11. Extension of table 4 in the main body. Here, in addition to the accuracy α_t , γ_t and KL_t also indicate the respective NLL reconstruction metrics and corresponding KL divergences at the end of every task increment t .

MNIST	t	Dual Pix Model	PixCDVAE	PixOCDVAE
$\alpha_{base,t}$	1	99.97 \pm 0.002	99.97 \pm 0.026	99.86 \pm 0.084
	2	99.54 \pm 0.285	96.90 \pm 2.907	99.64 \pm 0.095
	3	99.16 \pm 0.611	90.12 \pm 5.846	98.88 \pm 0.491
	(%)	4 98.33 \pm 1.119	76.84 \pm 0.905	98.11 \pm 0.797
	5	98.04 \pm 1.397	56.53 \pm 4.032	97.44 \pm 0.785
$\alpha_{new,t}$	1	99.97 \pm 0.002	99.97 \pm 0.026	99.86 \pm 0.084
	2	99.71 \pm 0.122	99.74 \pm 0.052	99.82 \pm 0.027
	3	99.41 \pm 0.084	99.22 \pm 0.082	99.56 \pm 0.092
	(%)	4 98.61 \pm 0.312	97.84 \pm 0.180	98.80 \pm 0.292
	5	97.31 \pm 0.575	96.77 \pm 0.337	98.63 \pm 0.430
$\alpha_{all,t}$	1	99.97 \pm 0.002	99.97 \pm 0.026	99.86 \pm 0.084
	2	99.60 \pm 0.142	98.37 \pm 1.448	99.69 \pm 0.051
	3	98.93 \pm 0.291	96.14 \pm 1.836	99.20 \pm 0.057
	(%)	4 98.22 \pm 0.560	91.25 \pm 0.992	98.13 \pm 0.281
	5	96.52 \pm 0.658	83.61 \pm 0.927	96.84 \pm 0.346
$\gamma_{base,t}$	1	90.52 \pm 0.263	100.0 \pm 1.572	99.77 \pm 2.768
	2	91.27 \pm 0.789	100.4 \pm 1.964	101.2 \pm 3.601
	3	91.92 \pm 0.991	100.3 \pm 4.562	101.1 \pm 4.014
	(nats)	4 91.75 \pm 1.136	102.7 \pm 7.134	101.0 \pm 4.573
	5	92.05 \pm 1.212	102.4 \pm 6.195	100.5 \pm 4.942
$\gamma_{new,t}$	1	90.52 \pm 0.263	100.0 \pm 1.572	99.77 \pm 2.768
	2	115.8 \pm 0.805	125.7 \pm 2.413	124.6 \pm 3.822
	3	107.7 \pm 0.600	118.3 \pm 3.523	116.5 \pm 2.219
	(nats)	4 100.9 \pm 0.659	107.1 \pm 5.316	102.3 \pm 1.844
	5	113.4 \pm 0.820	118.2 \pm 1.572	113.3 \pm 0.755
$\gamma_{all,t}$	1	90.52 \pm 0.263	100.0 \pm 1.572	99.77 \pm 2.768
	2	102.9 \pm 0.408	111.9 \pm 2.627	112.7 \pm 3.300
	3	104.8 \pm 1.114	114.9 \pm 4.590	114.6 \pm 4.788
	(nats)	4 103.9 \pm 0.759	114.3 \pm 3.963	112.1 \pm 2.150
	5	106.1 \pm 0.868	118.7 \pm 5.320	111.9 \pm 2.663
$KL_{all,t}$	1	1.410 \pm 0.181	5.629 \pm 3.749	5.635 \pm 3.739
	2	3.177 \pm 0.702	9.238 \pm 0.674	7.495 \pm 0.738
	3	4.923 \pm 1.085	12.13 \pm 0.977	10.17 \pm 1.528
	(nats)	4 5.603 \pm 1.250	14.32 \pm 1.040	11.66 \pm 1.004
	5	9.296 \pm 1.346	16.37 \pm 0.970	12.49 \pm 0.551

Table 15. Results for PixelVAE based class incremental continual learning approaches averaged over 5 runs, baselines and the reference isolated learning scenario for FashionMNIST at the end of every task increment in analogy to table 12. Extension of table 4 in the main body. Here, in addition to the accuracy α_t , γ_t and KL_t also indicate the respective NLL reconstruction metrics and corresponding KL divergences at the end of every task increment t .

Fashion	t	Dual Pix Model	PixCDVAE	PixOCDVAE
$\alpha_{base,t}$	1	99.57 \pm 0.091	99.58 \pm 0.076	99.54 \pm 0.079
	2	82.40 \pm 6.688	90.06 \pm 1.782	88.60 \pm 1.998
	3	78.55 \pm 3.964	83.70 \pm 3.571	87.66 \pm 0.375
	(%)	4 54.69 \pm 3.853	50.23 \pm 7.004	68.31 \pm 3.308
	5	60.04 \pm 5.151	47.83 \pm 13.41	74.45 \pm 2.889
$\alpha_{new,t}$	1	99.57 \pm 0.091	99.58 \pm 0.076	99.54 \pm 0.079
	2	97.73 \pm 1.113	96.47 \pm 0.596	97.31 \pm 0.475
	3	99.09 \pm 0.367	97.33 \pm 0.725	96.88 \pm 1.156
	(%)	4 97.55 \pm 0.588	96.12 \pm 0.675	95.47 \pm 1.332
	5	98.85 \pm 0.141	97.91 \pm 0.596	98.63 \pm 0.176
$\alpha_{all,t}$	1	99.57 \pm 0.091	99.58 \pm 0.076	99.54 \pm 0.079
	2	86.22 \pm 3.704	92.93 \pm 0.160	92.17 \pm 1.425
	3	76.77 \pm 4.378	84.07 \pm 1.069	87.30 \pm 0.322
	(%)	4 62.93 \pm 3.738	64.42 \pm 1.837	76.36 \pm 1.267
	5	72.41 \pm 2.941	63.05 \pm 1.826	80.85 \pm 0.721
$\gamma_{base,t}$	1	267.8 \pm 1.246	230.8 \pm 3.024	232.0 \pm 2.159
	2	273.6 \pm 0.631	232.5 \pm 1.582	231.8 \pm 0.416
	3	274.0 \pm 0.552	235.6 \pm 2.784	231.6 \pm 0.832
	(nats)	4 273.7 \pm 0.504	236.4 \pm 3.157	231.4 \pm 2.550
	5	274.1 \pm 0.349	241.1 \pm 1.747	234.1 \pm 1.498
$\gamma_{new,t}$	1	267.8 \pm 1.246	230.8 \pm 3.024	232.0 \pm 2.159
	2	313.4 \pm 1.006	275.8 \pm 1.888	275.3 \pm 1.473
	3	269.1 \pm 0.616	268.3 \pm 3.852	262.9 \pm 1.893
	(nats)	4 282.4 \pm 0.321	259.1 \pm 1.305	259.6 \pm 2.050
	5	305.8 \pm 0.286	283.2 \pm 2.150	283.5 \pm 2.458
$\gamma_{all,t}$	1	267.8 \pm 1.246	230.8 \pm 3.024	232.0 \pm 2.159
	2	293.8 \pm 0.349	254.3 \pm 1.513	255.8 \pm 0.436
	3	285.7 \pm 0.510	261.5 \pm 2.970	259.1 \pm 0.929
	(nats)	4 284.9 \pm 0.703	263.2 \pm 2.259	259.5 \pm 3.218
	5	289.5 \pm 0.396	271.7 \pm 2.117	267.2 \pm 0.586
$KL_{all,t}$	1	3.610 \pm 0.856	7.164 \pm 0.759	7.809 \pm 1.255
	2	6.247 \pm 0.710	13.79 \pm 0.282	12.23 \pm 0.287
	3	7.811 \pm 0.799	18.26 \pm 0.818	15.36 \pm 0.530
	(nats)	4 8.982 \pm 0.812	21.75 \pm 0.561	18.31 \pm 0.333
	5	9.781 \pm 1.068	22.14 \pm 0.377	17.93 \pm 0.360

Table 16. Results for PixelVAE based class incremental continual learning approaches averaged over 5 runs, baselines and the reference isolated learning scenario for AudioMNIST at the end of every task increment in analogy to table 13. Extension of table 4 in the main body. Here, in addition to the accuracy α_t , γ_t and KL_t also indicate the respective NLL reconstruction metrics and corresponding KL divergences at the end of every task increment t .

Audio	t	Dual Pix Model	PixCDVAE	PixOCDVAE
$\alpha_{base,t}$ (%)	1	100.0 \pm 0.000	99.71 \pm 0.218	99.27 \pm 0.410
	2	99.52 \pm 0.273	97.86 \pm 0.799	97.88 \pm 2.478
	3	93.15 \pm 3.062	81.38 \pm 5.433	95.82 \pm 3.602
	4	81.55 \pm 8.468	50.58 \pm 14.60	91.56 \pm 5.640
	5	64.60 \pm 8.739	29.94 \pm 18.47	75.25 \pm 10.18
$\alpha_{new,t}$ (%)	1	100.0 \pm 0.000	99.71 \pm 0.218	99.27 \pm 0.410
	2	99.71 \pm 0.043	99.78 \pm 0.128	99.81 \pm 0.189
	3	98.23 \pm 1.092	98.41 \pm 0.507	99.30 \pm 0.550
	4	95.31 \pm 0.868	94.30 \pm 0.914	97.87 \pm 0.293
	5	98.18 \pm 0.885	97.00 \pm 0.520	99.43 \pm 0.495
$\alpha_{all,t}$ (%)	1	100.0 \pm 0.000	99.71 \pm 0.218	99.27 \pm 0.410
	2	99.50 \pm 0.157	98.64 \pm 0.875	99.67 \pm 0.033
	3	95.37 \pm 1.750	90.10 \pm 1.431	97.77 \pm 1.017
	4	86.97 \pm 2.797	75.55 \pm 3.891	95.41 \pm 1.345
	5	75.50 \pm 3.032	63.44 \pm 5.252	90.23 \pm 1.139
$\gamma_{base,t}$ (nats)	1	434.2 \pm 1.068	432.6 \pm 0.321	433.8 \pm 0.370
	2	434.4 \pm 1.082	432.5 \pm 0.551	433.5 \pm 1.464
	3	434.6 \pm 0.785	432.9 \pm 0.723	433.1 \pm 1.269
	4	434.2 \pm 1.209	433.0 \pm 0.781	433.0 \pm 1.283
	5	435.1 \pm 1.915	431.4 \pm 0.666	432.3 \pm 0.189
$\gamma_{new,t}$ (nats)	1	434.2 \pm 1.068	432.6 \pm 0.321	433.8 \pm 0.370
	2	390.4 \pm 0.694	389.4 \pm 0.208	389.4 \pm 1.304
	3	444.7 \pm 0.545	442.7 \pm 0.513	442.4 \pm 0.275
	4	497.4 \pm 0.740	494.4 \pm 0.700	494.8 \pm 0.386
	5	431.9 \pm 1.032	428.0 \pm 0.851	429.7 \pm 1.223
$\gamma_{all,t}$ (nats)	1	435.2 \pm 15.69	432.6 \pm 0.321	433.8 \pm 0.370
	2	412.4 \pm 0.871	410.9 \pm 0.351	411.5 \pm 1.406
	3	423.3 \pm 0.618	421.0 \pm 1.026	421.9 \pm 0.661
	4	441.6 \pm 0.420	439.8 \pm 0.833	439.8 \pm 0.718
	5	440.3 \pm 1.297	436.9 \pm 0.751	437.7 \pm 0.432
$KL_{all,t}$ (nats)	1	4.361 \pm 0.671	9.293 \pm 0.943	11.87 \pm 1.504
	2	5.130 \pm 0.636	14.00 \pm 0.748	12.40 \pm 0.719
	3	5.399 \pm 0.724	20.28 \pm 0.774	14.41 \pm 0.461
	4	5.817 \pm 1.038	24.91 \pm 0.845	16.00 \pm 0.505
	5	6.031 \pm 0.832	27.14 \pm 1.139	17.45 \pm 0.835

D.3. Backward transfer

As our model’s weights are shared fully across all tasks this opens up the scope for both forward and backward transfer of knowledge. For clarification, the former concept refers to existing tasks’ representations aiding in acquisition of a new task’s information, e.g. by speeding up training. The latter concept describes the reverse phenomenon, where introduction of a new task leads to learning of representations that retrospectively improve former tasks, even if its real data is no longer present. Figure 10 highlights an interesting case of backward transfer for class-incremental learning with our OCDVAE model on the AudioMNIST dataset, as quantitatively presented in tables 13 and 16. The addition of two new classes (four and five) at the end of the second increment leads to an improvement in the classification performance on class two, as indicated by the confusion matrices. We point out that this is a desirable continual learning property that can only emerge from having a single model with a single classification head.

E. Generative replay examples with CDVAE and OCDVAE

In this section we provide visualization of data instances that are produced during generative replay at the end of each task increment. In particular, we qualitatively illustrate the effect of constraining sampling to the aggregate posterior in contrast to naively sampling from the prior without statistical outlier rejection for low density regions. Figures 11, 12 and 13 illustrate generated images for MNIST, FashionMNIST and AudioMNIST respectively. For both a naive CDVAE as well as the autoregressive PixCDVAE we observe significant confusion with respect to classes. As the generative model needs to learn how to replay old tasks’ data based on its own former generations, ambiguity and blurry interpolations accumulate and are rapidly amplified. This is not the case for OCDVAE and PixOCDVAE, where the generative model is capable of maintaining higher visual fidelity throughout continual training and misclassification is scarce.

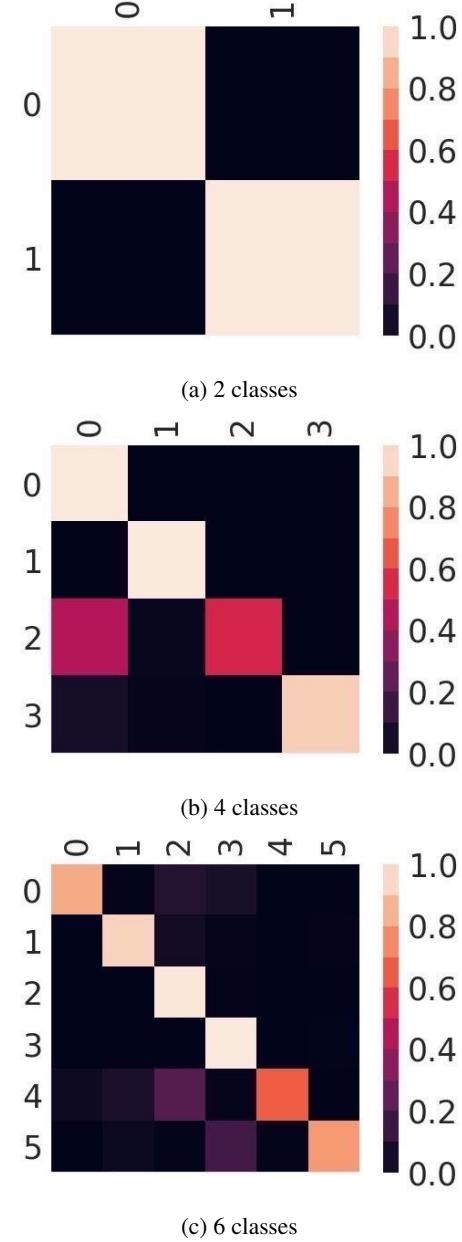


Figure 10. AudioMNIST confusion matrices for incrementally learned classes with the OCDVAE model. When adding classes two and three the model experiences difficulty in classification, however is able to overcome this challenge by exhibiting backward transfer when later learning classes four and five. It is also observable how forgetting of the initial classes is limited.

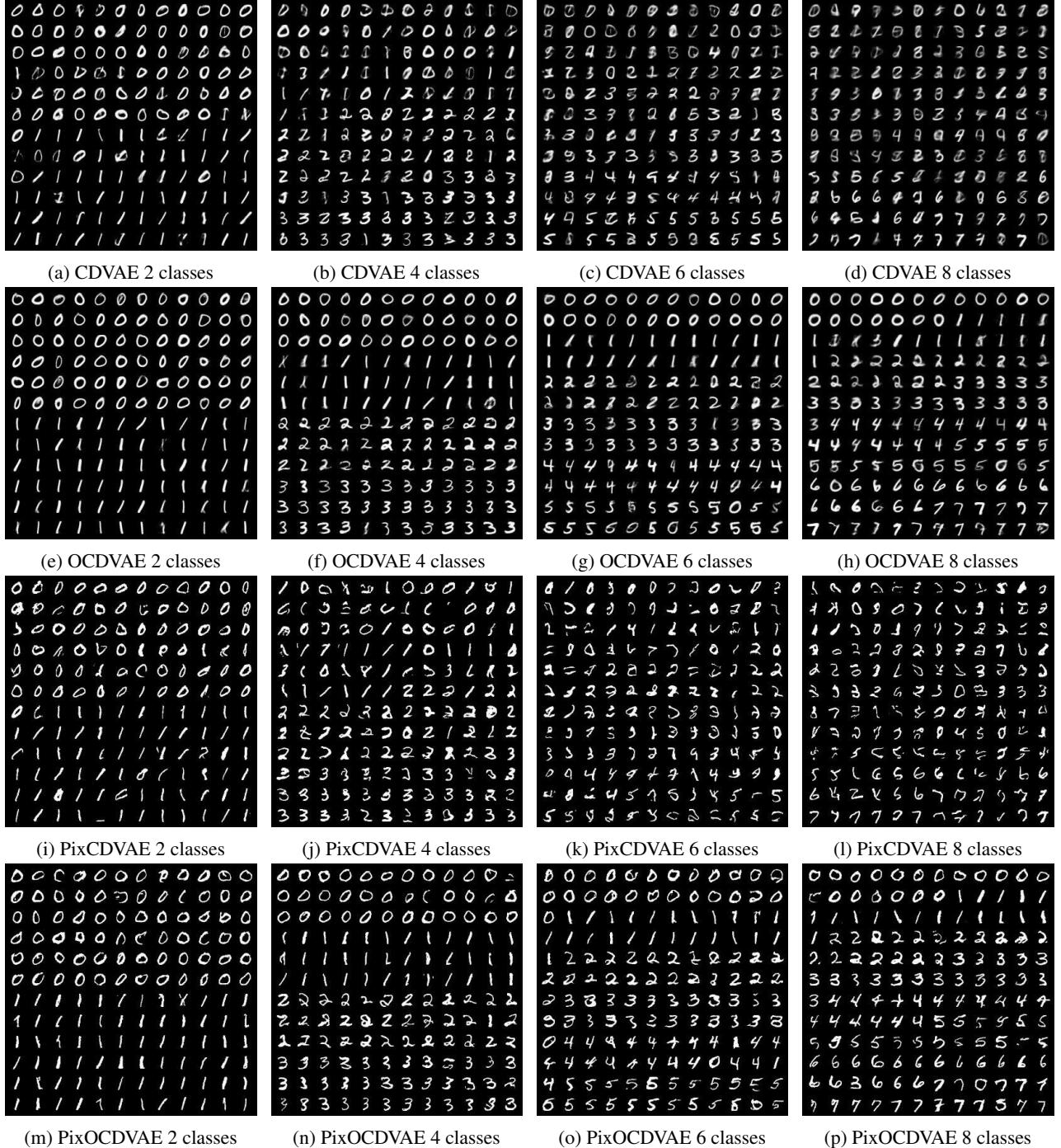


Figure 11. Generated images for continually learned incremental MNIST at the end of task increments for CDVAE (a-d), OCDVAE (e-h), PixCDVAE (i-l) and PixOCDVAE (m-p). Each individual grid is sorted according to the class label that is predicted by the classifier.

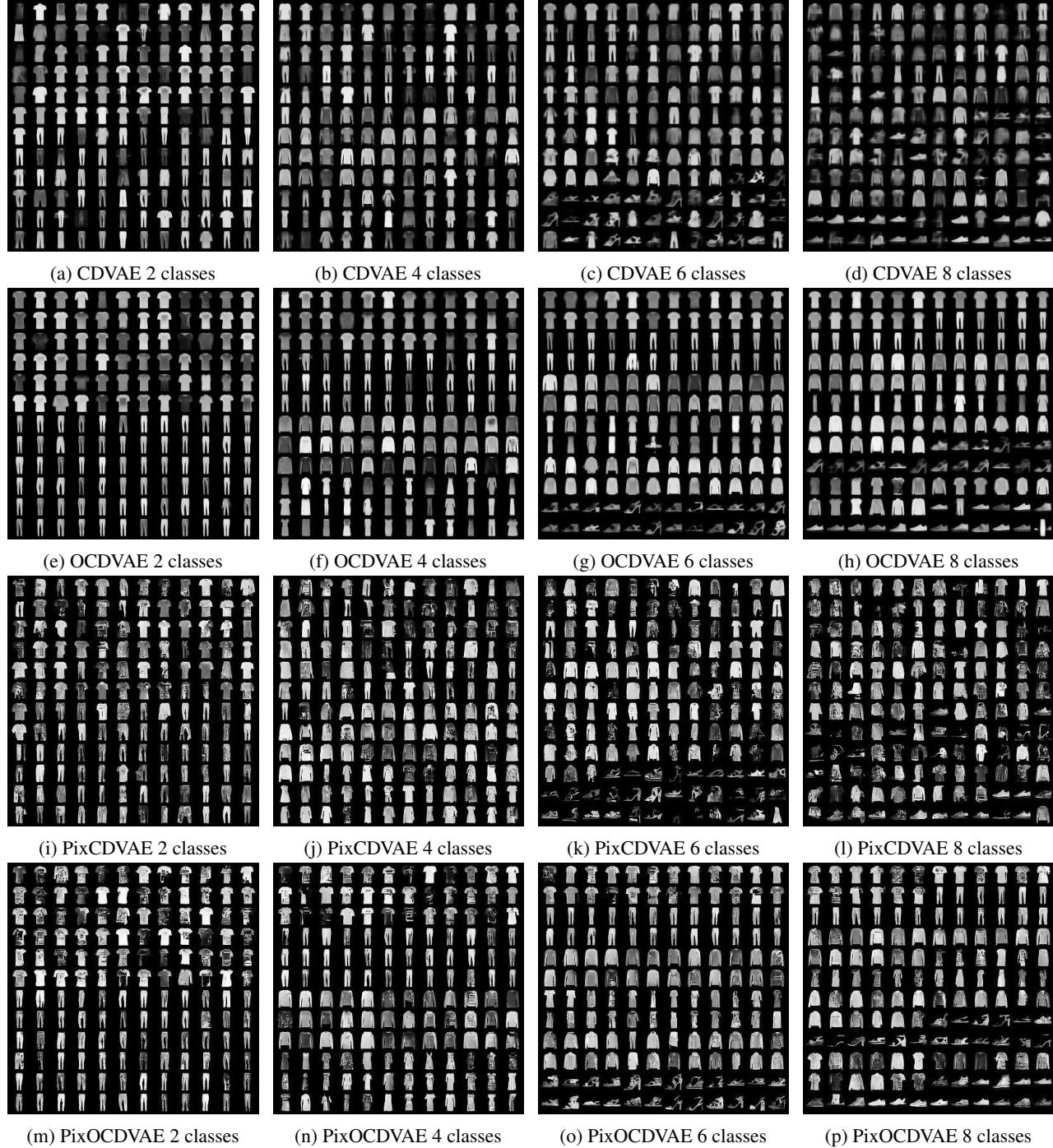


Figure 12. Generated images for continually learned incremental FashionMNIST at the end of task increments for CDVAE (a-d), OCDVAE (e-h), PixCDVAE (i-l) and PixOCDVAE (m-p). Each individual grid is sorted according to the class label that is predicted by the classifier.

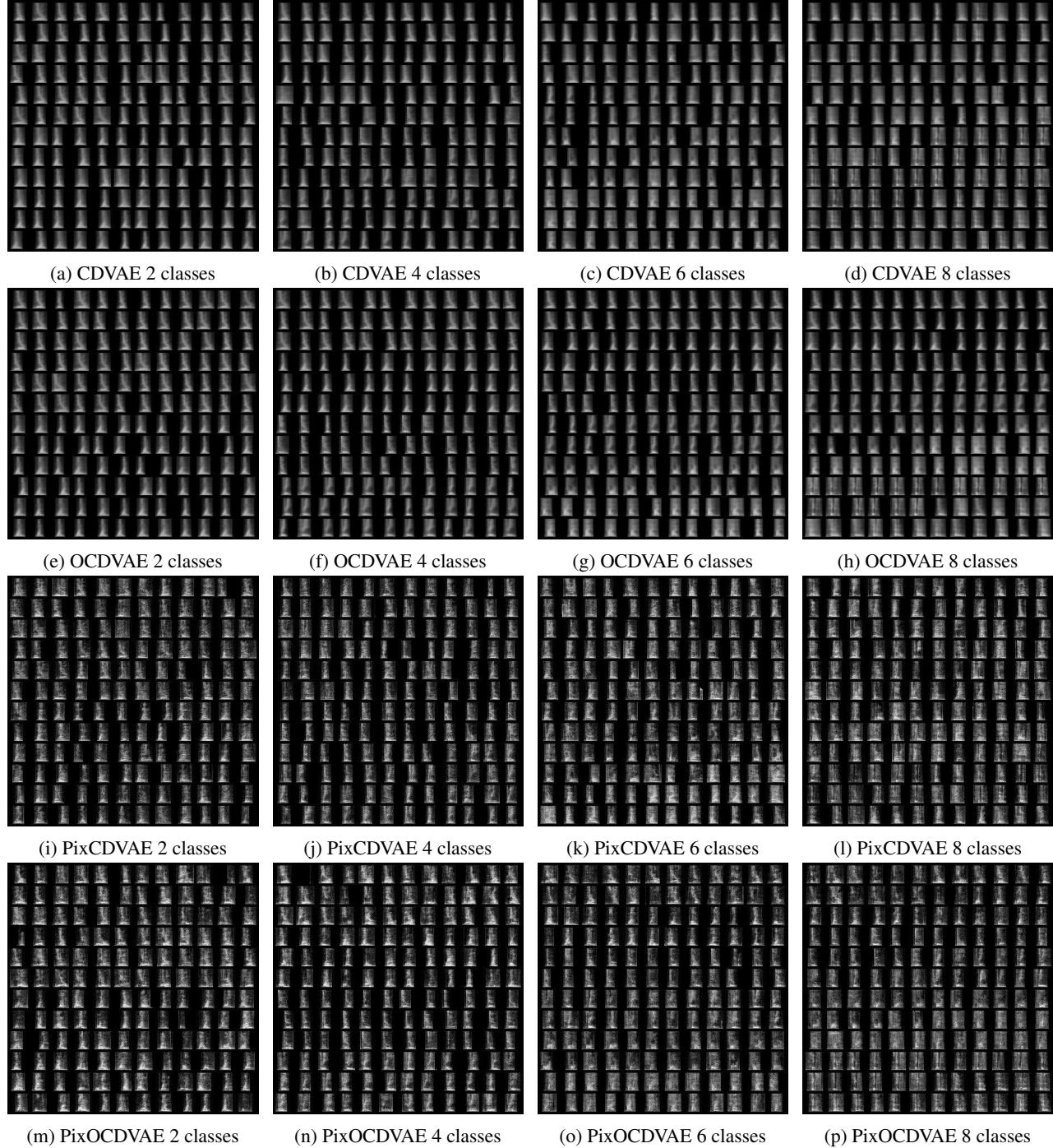


Figure 13. Generated images for continually learned incremental AudioMNIST at the end of task increments for CDVAE (a-d), OCDVAE (e-h), PixCDVAE (i-l) and PixOCDVAE (m-p). Each individual grid is sorted according to the class label that is predicted by the classifier.