Expert Systems With Applications 283 (2025) 127774

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

CEDL+: Exploiting evidential deep learning for continual out-of-distribution detection

Eduardo Aguilar [a,c,*], Bogdan Raducanu [b], Petia Radeva [c], Joost van de Weijer [b]

[a] Departamento de Ingeniería de Sistemas y Computación, Universidad Católica del Norte, Antofagasta, 1270709, Chile
[b] Computer Vision Center, Universitat Autònoma de Barcelona, Barcelona, 08193, Spain
[c] Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Barcelona, 08007, Spain

ARTICLE INFO

Dataset link: https://github.com/Continvvm/continuum/tree/master/continuum/datasets

Keywords:
OOD detection
Object recognition
Continual learning
Uncertainty quantification

ABSTRACT

The current deep learning paradigm is generally based on two main assumptions that are not met in many real-world applications: (i) all the data is jointly available for training (allowing for IID training); and (ii) at inference time, we only have data belonging to the classes seen during training (closed-world assumption). In this paper, we study the more realistic scenario, where we have to learn from a non-stationary data stream and in addition we should assess the certainty of the predictions for application in open-world settings. Therefore, we endow a continual learning method with the ability to quantify uncertainty, thus improving its reliability and robustness. To this end, Evidential Deep Learning is integrated into a continual learning framework to efficiently perform continual out-of-distribution (OOD) data detection as the model increases its knowledge. The new approach has been validated on three public datasets and in several continual learning settings, clearly outperforming the existing state-of-the-art methods.

1. Introduction

Deep learning-based methods have achieved remarkable results across a broad spectrum of computer vision applications (Poyser & Breckon, 2024). Most deep learning methods are based on two underlying assumptions: (i) the learning is performed on a stationary data stream (all training data is jointly available from the beginning); and (ii) evaluation of the method is only done with data belonging to the classes seen during training (the closed-world assumption). However, many real-world applications face non-stationary data streams, in which case the algorithms have to learn from a sequence of data, accumulating the knowledge incrementally. In addition, algorithms should be capable of detecting out-of-distribution data (data of classes that have not been considered during training). Therefore, our research focuses on *continual out-of-distribution detection* where a model is trained sequentially with data from a subset of classes, which can be extended over time. In this case, the method should be able to discern between the seen and unseen classes (out-of-distribution data) to avoid providing unreliable predictions.

On one side, continual learning emerged as a line of research that addresses the problem of training from non-stationary data, allowing sequential training of the model. Unfortunately, a highly undesirable effect associated with sequential learning is represented by catastrophic forgetting, where mainly the knowledge belonging to the latest learning stage is remembered. In consequence, several strategies have been proposed to retain previously learned knowledge (stability) while incorporating new knowledge (plasticity) into the model (Masana et al., 2023): regularization-based methods, rehearsal-based methods and architecture-based methods.

On the other side, it is of utmost importance to provide the model with the ability and robustness to detect out-of-distribution (OOD) samples to prevent wrong predictions or to be vulnerable to adversarial attacks (Qian et al., 2022). To this end, uncertainty-aware methods provide a good opportunity to ensure the robustness of the method for OOD detection. There are four main families of methods that can quantify uncertainty in predictions (Gawlikowski et al., 2023): Bayesian-based methods, ensemble-based methods, test time augmentation methods, and single deterministic neural networks. The first three methods are computationally expensive during training, inference, or both. The latter is more suitable for an open-world setting. Among the methods to estimate the uncertainty of neural networks, *evidential deep learning* (Sensoy et al., 2018) stands out because of its ability to pinpoint the various sources of uncertainty: the method can distinguish between a lack of confidence (as measured by *vacuity*) and conflicting evidence (as measured by *dissonance*) (Hu et al., 2021; Zhao et al.,

* Corresponding author at: Departamento de Ingeniería de Sistemas y Computación, Universidad Católica del Norte, Antofagasta, 1270709, Chile.
E-mail addresses: eduardo.aguilar@ub.edu (E. Aguilar), bogdan@cvc.uab.es (B. Raducanu), petia.ivanova@ub.edu (P. Radeva), joost@cvc.uab.es (J. van de Weijer).

https://doi.org/10.1016/j.eswa.2025.127774
Received 17 June 2024; Received in revised form 30 October 2024; Accepted 15 April 2025
Available online 27 April 2025

2019). Within this approach, the network's output is used to set the parameters of a Dirichlet distribution on the class probabilities. These parameters can then be used to assess the uncertainty of the network. The method has been successfully evaluated for the task of OOD detection (Sensoy et al., 2018), but has not yet been combined with continual learning methods to extend the application to non-stationary training data streams. Overall, the problem of OOD detection in a continual learning framework is a very recent research topic (Aljundi et al., 2022).

In this paper, we focus on posthoc methods capable of being used at inference time without requiring additional training to detect OOD samples. Most of these methods are used at the top of the logit layer of classical neural network which cannot provide the degree of certainty about the prediction. Integrating evidential deep learning into a continual learning framework makes it possible to provide reliable predictions and leverage quantified uncertainty to detect OOD samples efficiently. Our main contributions are:

- We are the first to integrate evidential deep learning into a continual learning approach to simultaneously perform incremental object recognition and OOD detection.
- A new loss function that makes it possible to distill knowledge and thus avoids catastrophic forgetting by explicitly regulating the Dirichlet distribution, thereby preventing newly acquired knowledge from interfering with the uncertainty estimates of previous tasks.
- Comparison of the proposed CEDL+ with the state-of-the-art on three public datasets and various incremental settings with clear outperformance.
- Analysis of three combinations of vacuity and dissonance measures evaluating their usability for detecting data in old classes, current classes and unseen classes (OOD data).

To the best of our knowledge, no approach has been proposed to detect OOD samples in a continual learning framework using an uncertainty-aware method. This work extends our previous workshop paper (Aguilar et al., 2023). We extend our previous work with an improved regularization loss function that is tailored to evidential learning. In addition, further evaluation of performance has been evaluated on two extra datasets (TinyImageNet and Food101) and an extra continual learning setting has been considered. Furthermore, we considered a cross-dataset scenario for OOD detection. Finally, the results obtained demonstrate the benefit of our proposal not only for OOD detection (as in the previous version) but also for performing object recognition.

The paper is structured as follows: in the next section, we present a brief review of the works most related to our research. In Section 3, we describe the proposed method for continual out-of-distribution detection. In Section 4, we present the experimental setup. In Section 5, we discuss the results obtained. Finally, we present conclusions and future directions.

2. Related work

We briefly review the most relevant works in uncertainty estimation and continual learning.

Continual Learning: Continual Learning has reached a certain degree of maturity nowadays, being the focus of intensive research over the recent years (Masana et al., 2023). Two strategies to mitigate catastrophic forgetting in continual learning have been considered in this paper: regularization-based methods and rehearsal-based methods.

Regularization-based methods are a category of continual learning techniques that focuses on preventing the model from forgetting past knowledge by penalizing significant changes to its parameters during the learning process. Weight regularization methods consist of computing prior importance for all parameters in the network and selectively penalizing its changes (Akyürek et al., 2022; Aljundi et al., 2018; Kirkpatrick et al., 2017; Wang et al., 2021). Another group of works addresses the *stability-plasticity dilemma* by introducing an auxiliary network (Kim, Noci, et al., 2023), which together with the main network acts as a regularizer or by expanding and optimizing the parameters only of the new task (Wang et al., 2021), while a forgetting factor regulates a penalty to merge the main network parameters with the expanded ones selectively. In Jung et al. (2023), they propose an approach to enhance stability and plasticity in parallel by leveraging multi-scale feature maps, which are constructed by projecting raw images into meaningful subspaces. On the other hand, functional regularization methods (Cermelli et al., 2020; Cha et al., 2021) apply a regularization on the network output (class probabilities) or the intermediate features (Douillard et al., 2020).

Rehearsal-based methods rely on replaying data belonging to past tasks while learning a new one, in order to avoid catastrophic forgetting. The replayed data could be real images (also called exemplars, stored in a buffer) (Buzzega et al., 2020; Rebuffi et al., 2017) or features (Hayes et al., 2020; Oh et al., 2022; Tiwari et al., 2022; Zheng et al., 2024). Alternatively, other replay methods do not store representations of real data, but they use a generative mechanism to recreate synthetic exemplars. Early approaches were using GANs to generate synthetic data (Mundt et al., 2022; Wu et al., 2018; Zhai et al., 2019), but more recent ones are diffusion-driven (Gao & Liu, 2023; Kim et al., 2024; Liang et al., 2024). Several exemplar selection strategies have been proposed to find the most representative data to be stored, most of them based on a singular criterion (e.g., distance-based (Rebuffi et al., 2017)) and recently based on multiple criteria (e.g, by combining distance, cluster variance and classifier loss (Zhuang et al., 2022)). The effectiveness of these methods is evident in small buffer sizes, with moderate or large sizes, the random selection of exemplars remains a strong baseline (Wiewel et al., 2022).

OOD Detection in Continual Learning: While OOD detection is well-established in classical machine learning, its application to continual learning, which is a more realistic setting, only recently has started to be studied. A related problem is continual class discovery (Roy et al., 2022; Zhao & Mac Aodha, 2023), however, although both handle data from classes not seen during training, continual OOD detection focuses on identifying data with a significantly different distribution from the training set, marking such samples as unreliable or rejecting them without learning or categorizing them. One of the first works that pioneered this research direction is presented in Aljundi et al. (2022), where they established benchmarks for the continual OOD detection problem. A different approach is adopted in Rios et al. (2022), where they introduce a self-supervised continual novelty detector, which builds incrementally a statistical model over the space of intermediate features produced by a deep network and utilizes feature reconstruction errors as uncertainty scores to guide the detection of novel samples. In Kim, Esmaeilpour, et al. (2022), a unified approach for both class-incremental and task-incremental settings is proposed, using a dual mechanism based on task masking. In Kim, Xiao, et al. (2022), the authors presented a theoretical study according to which an improved CIL performance is strongly correlated with a good OOD detector. This theoretical study has been further extended in Kim, Xiao, et al. (2023), where they show that CIL is indeed learnable.

Recently, a meta-learning approach, based on neural processes for uncertainty estimation (including OOD detection) in continual learning, is proposed in Jha et al. (2023). The neural processes encode tasks into probabilistic distributions, allowing for reliable uncertainty estimates. The approach uses task-specific modules in a hierarchical model and applies regularizers to reduce forgetting. Another approach (Zhu, Cheng, et al., 2024) fine-tunes a given model with reliable weight consolidation and weight space interpolation. This allows for simultaneous detection of both misclassified and OOD samples in a continual learning setting. Finally, in Zhu, Yi, and Zhang (2024), they propose a new boundary-unknown continual learning scenario. To identify task

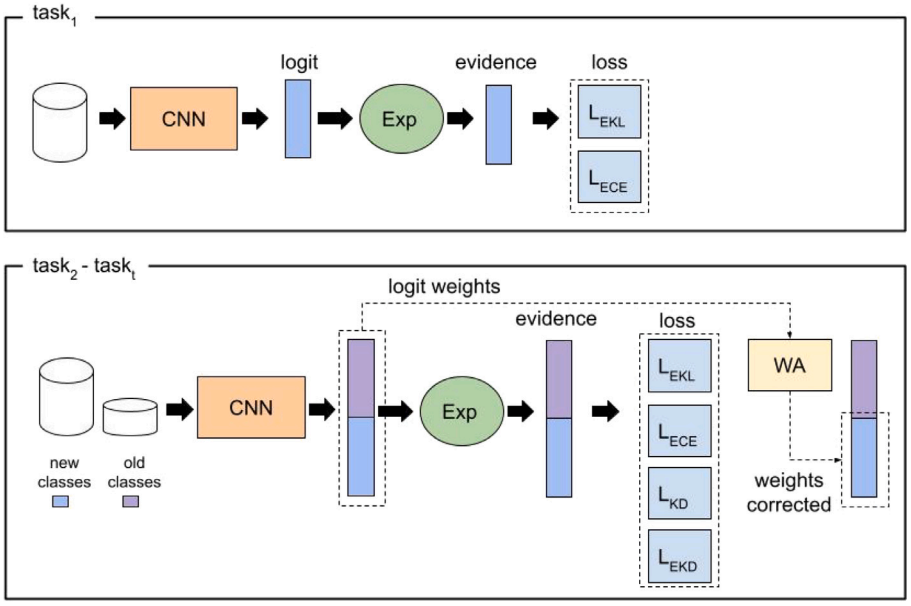

Fig. 1. Overview of the CEDL+ approach illustrated for t incremental steps. Evidential Deep Learning is used for the first task (top), and CEDL+ which considers the proposed losses L_{EKL} and L_{EKD} is used for the rest (bottom). WA denotes Weight Alignment.

boundaries, they design a continual OOD detection method based on softmax probabilities, which can detect OOD samples for the latest learned task.

Uncertainty Estimation in Continual Learning: Uncertainty can be defined as unknowing something that could happen. Estimating uncertainty allows us to quantify this degree of unknown and can therefore be useful to improve decision-making and avoid unexpected behavior of the model. Uncertainty occurs mainly in two types (Kendall & Gal, 2017): epistemic uncertainty and aleatoric uncertainty. Epistemic uncertainty, also known as model uncertainty, occurs due to the lack of sufficient training data to represent its distribution. As for aleatoric uncertainty, also known as data uncertainty, it occurs due to random noise (e.g., errors in the measuring instrument, incorrect labels, etc.) that can alter the collected data. To estimate uncertainty in deep learning, Evidential Deep Learning (EDL) (Sensoy et al., 2018) method stands out for its solid theory, easy implementation and optimal use of resources. In closed-world assumption, EDL has been used as a posthoc method to perform OOD sample detection (Hu et al., 2021; Zhao et al., 2019). Open-set recognition has been addressed in Bao et al. (2021) considering an adapted EDL method that replaces the regularization term by a calibration uncertainty approach to mitigate overconfident prediction in a numerically stable manner. Vacuity was used to detect samples of novel classes. Other EDL-based approaches have analyzed the usefulness of Vacuity and Dissonance measures for OOD sample detection demonstrating outstanding results (Hu et al., 2021; Zhao et al., 2019). On the other hand, within the framework of continual learning, an EDL approach was recently proposed for the class-based incremental semantic segmentation problem (Holmquist et al., 2023). Here the authors propose to use the estimated uncertainty as a probability measure for the background class, justifying that this class shifts after each increment (past and future classes are correlated).

This paper proposes a novel continual learning method based on rehearsal and knowledge distillation to detect OOD samples. To this end, unlike previous work, we propose for the first time the integration of evidential deep learning into a continual learning framework. In this way, we endow the model with the ability to estimate predictive uncertainty (vacuity and dissonance) and use this information as a posthoc method to discriminate between in-distribution (IND) and OOD samples. In doing so, we will provide a reliable and scalable model for open-world visual recognition.

3. Continual evidential deep learning

This section first presents the class-incremental learning setup used as the base method for the proposal. Next, evidential deep learning is explained. Finally, the combination of both methodologies for continual out-of-distribution detection with evidential learning is described (Fig. 1 shows the resulting CEDL+ method).

3.1. Class-incremental learning preliminaries

In this subsection, we formally define the class-incremental learning problem. The method described here serves as a basis for our proposal from which the adaptation is made for the integration with evidential deep learning. In the experimental section, we refer to this method as a 'baseline'.

In a CIL setting, data arrives sequentially split into n tasks $T = \{T_1, \ldots, T_n\}$, where each task consists of a training set $D_i^S = (x_i, y_i)$ with x_i representing the data, y_i its associated label and S is the subset of the data. Following the general continual learning acceptance, the class labels of each task are disjoint, i.e. $y_i \cap y_j = \emptyset$, for $i \neq j$. During the learning phase, a model M is trained n times to learn a subset of the data, D^S, incrementally.

To mitigate the catastrophic forgetting effect, we use a rehearsal-based approach by replaying a small number of exemplars from previously seen tasks stored in a buffer. Additionally, we use knowledge distillation (Hinton et al., 2014) at the logits level, to maintain the classifier's performance for the previous tasks. The distillation is performed using a teacher–student approach, in which the teacher, representing the model after learning $n-1$ tasks (M^{n-1}), transfers the knowledge from the learned classes to the student, representing the model after learning n tasks (M^n), by minimizing the Kullback–Leibler (KL) divergence of the output probabilities. Although exemplars' replay and knowledge distillation be effective in preventing catastrophic forgetting, the model may still have biases towards the new classes (Zhao et al., 2020). Based on the observation that the norm of the weight vectors for new classes tends to be larger than that for the old classes and, therefore, the output logits also tend to be larger, biasing the prediction towards the new classes. In Zhao et al. (2020), a Weight Aligning (WA) approach was proposed to correct the biases in the Fully Connected (FC) layer. This bias correction is performed before starting the training of each incremental step from T_2 to T_n.

In conclusion, the baseline method considered in our work contemplates these three techniques discussed above (exemplars replay, knowledge distillation and weight aligning) to counteract the catastrophic forgetting problem.

Regarding the model prediction, let us consider $f(x) = \mathrm{W}^{\intercal} \cdot g(x)$ be the output logits for the input image x, $g(.)$ the feature extracted from the layer before the logits one, W the logit weights. Then, the model prediction is given by:

$$o(z,\tau)_i = \frac{e^{z_i/\tau}}{\sum_{i \neq j} e^{z_j/\tau}} \tag{1}$$

where τ is a temperature used to smooth the probability distribution over the classes. It is usually set to 1 to provide model prediction and to more than 1 to perform knowledge distillation.

As for the loss function, there are two terms: cross-entropy loss (L_{CE}) and knowledge distillation loss (L_{KD}). The first term is standard for all tasks and is formally defined as follows:

$$L_{CE} = -\sum_{c=1}^{C} y_c \cdot \log(o(f(x),1)_c) \tag{2}$$

where C is the number of object classes present in $\{D^1, \ldots, D^n\}$ subsets where n correspond to the nth task, and y_i corresponds to the value of the ith position of the binary vector that represents Ground Truth (GT) label y in one-hot encoding format.

After the first task, the method must cope with catastrophic forgetting. In this case, the three techniques described above are used. For the memory-replay case, the exemplars stored in the buffer are added to the current task and the method is retrained with the goals of extracting features and learning to classify both old and new classes. For the case of knowledge distillation, a new term is contemplated in the loss function to retain the knowledge of old classes while the model learns new classes. This loss function is formally defined as follows:

$$L_{KD} = \mathrm{KL}[o(f(x),2)^{st} \| o(f(x),2)^{tch}] = \sum_{c=1}^{C_{old}} o(f(x),2)_c^{tch} \cdot (\log(o(f(x),2)_c^{tch}) - \log(o(f(x),2)_c^{st}))$$

where st is the model training with the current task (student), tch is the model training with the previous task (teacher) and KL[.‖.] is the Kullback–Leibler divergence. Finally, in the WA case, the logits wights are corrected before starting the next incremental step as follows:

$$f'(x) = (f(x)_{1,\ldots,c_{old}}, \gamma \cdot f(x)_{c_{old}+1,\ldots,c_{old}+C_{new}})$$
$$\gamma = \frac{mean([\|W_1\|, \|W_2\|, \ldots, \|W_{c_{old}}\|])}{mean([\|W_{c_{old}+1}\|, \|W_{c_{old}+2}\|, \ldots, \|W_{c_{old}+C_{new}}\|])}$$

where c_{old} are the cth classes considering the total classes from D^1 to D^{n-1}, C_{new} is the number of classes in the D^n and $\|W_c\|$ corresponds to the norm of the vector corresponding to the cth class.

In summary, the loss function corresponding to the baseline model can be defined in a continual learning framework as follows:

$$L = \begin{cases} L_{CE}, & \text{if } n = 1, \\ 0.5 \cdot (L_{CE} + L_{KD}), & \text{otherwise.} \end{cases} \tag{3}$$

3.2. Evidential deep learning

In open-world image recognition, a robust model in the face of uncertain data should be mandatory, particularly for detecting OOD samples. Although deep ensemble-based methods and Bayesian-based methods have proven effective in detecting OOD samples (Lakshminarayanan et al., 2017), both approaches are computationally expensive. As an alternative, EDL is a deterministic deep learning approaches capable of quantifying uncertainty without the need for additional computational resources. EDL stands out for its ease of implementation, its remarkable performance in close-world image recognition and, in particular, its good ability to detect OOD data (Sensoy et al., 2018).

Instead of classical deep learning approaches to perform object recognition, which provides class probabilities through a softmax activation in the output layer, EDL is based on subjective logit theory and through a nonlinear activation in the output layer provides an evidence score related to each class. EDL formulates learning as a process of evidence acquisition and estimates prediction uncertainty by placing a Dirichlet distribution on the class probabilities.

Let us consider the evidence $e = exp(f(x))$, the Dirichlet parameters $\alpha = e + 1$, Dirichlet strength $S = \sum_{c=1}^{C} \alpha_c$ and the probability mass function $p = \frac{\alpha}{S}$, then, the learning of an EDL considers the optimization of a loss function composed of two terms: a loss function to learn the patterns to form the multi-nomial opinions that support the evidence related to the classes and a loss function that acts as a regularizer that penalizes high evidence in poorly predicted samples. For the former, typically the Type II Maximum Likelihood loss is used we call Evidential cross-entropy loss L_{ECE}, because of its similarity to the cross-entropy loss when an exponential activation is used to generate the evidence (which is our case). As for the latter, a KL-divergence which we denote L_{EKL}, is used. Formally, both loss functions are defined as follows:

$$L_{ECE} = \sum_{j=1}^{C} y_{ij}(log(S_i) - log(\alpha_{ij})), \tag{4}$$

$$L_{EKL} = \mathrm{KL}[\mathrm{Dir}(p_i|\tilde{\alpha_i})||Dir(p_i|\mathbf{1})] = log\left(\frac{\Gamma(\sum_{c=1}^{C}\tilde{\alpha_{ic}})}{\Gamma(C)\prod_{c=1}^{C}\Gamma(\tilde{\alpha_{ic}})}\right) + \sum_{c=1}^{C}(\tilde{\alpha_{ic}} - 1)\left[F(\tilde{\alpha_{ic}}) - F(\sum_{j=1}^{C}\tilde{\alpha_{ij}})\right], \tag{5}$$

$$\mathrm{Dir}(p,\alpha) = \frac{\Gamma(S)}{\prod_{c=1}^{C}\Gamma(\alpha_c)}\prod_{c=1}^{C} p_c^{\alpha_c - 1}; \alpha_c > 0. \tag{6}$$

where Dir(.,.) is the prior Dirichlet distribution for drawing the class probability for the ith sample, $\Gamma(.)$ is the gamma function, $F(.)$ is the digamma function, $\tilde{\alpha_i} = (\alpha_i - 1) \cdot (1 - y_i) + 1$ is the Dirichlet parameters that act as a mask to focus the penalty on the evidence acquired for erroneous classes, y_i is the GT label in one-hot encoding and $\mathbf{1}$ corresponds to a vector of ones representing non-evidence.

3.3. Integration of CL and EDL

Improved Continual Evidential Deep Learning (CEDL+) is a method that integrates the EDL approach in a Continual Learning framework to perform simultaneously object recognition and OOD sample detection. CEDL+ adapts the baseline method described in Section 3.1 to quantify both predicted classes and uncertainty. This requires some changes to the baseline method in terms of the activation function used in the output layer and the loss function. Motivated by Bao et al. (2021), the softmax activation was replaced by the exponential activation and L_{CE} was replaced by L_{ECE}. As mentioned in Sensoy et al. (2018), L_{ECE} may overfit during training and lose its ability to acquire the evidence correctly. To avoid this, the authors in Sensoy et al. (2018) recommend adding L_{EKL} regularization in the loss function. With this in mind, the loss function can be defined as follows:

$$L_1 = \lambda_1 \cdot L_{ECE} + (1 - \lambda_1) \cdot L_{EKL}, \tag{7}$$

where λ_1 is a hyperparameter used to weight both terms of the loss function.

So far, L_1 works in the closed-world assumption, i.e. when all classes are seen by the model during training. However, in a continual learning framework this does not happen. During each incremental step, the model must cope with the change of data distribution produced, because the new data belong to different classes than those previously seen. In our previous work (Aguilar et al., 2023), we observed that L_{KD} is compatible with a continual evidential learning approach to preserve knowledge about previously learned classes. Although L_{KD} showed good results in preventing catastrophic forgetting,

in this work we exploit the distillation of the entire Dirichlet distribution to provide an even better ability of the model to preserve evidence and thus provide better results. To this end, we propose two components for the knowledge distillation loss function: the current L_{KD} and the proposed L_{EKD}. This loss function is formally defined as:

$$L_2 = \lambda_2 \cdot L_{KD} + (1-\lambda_2) \cdot L_{EKD}, \tag{8}$$

$$\begin{aligned} L_{EKD} &= \mathrm{KL}[\mathrm{Dir}(p_i^{st}|\alpha_i^{st})||\mathrm{Dir}(p_i^{tch}|\alpha_i^{tch})] = \\ &\log\frac{\Gamma(\sum_{c=1}^{C_{old}}\alpha_{ic}^{st})}{\Gamma(\sum_{c=1}^{C_{old}}\alpha_{ic}^{tch})} + \\ &\sum_{c=1}^{C_{old}}\log\frac{\Gamma(\alpha_{ic}^{tch})}{\Gamma(\alpha_{ic}^{st})} + \\ &\sum_{c=1}^{C_{old}}(\alpha_{ic}^{st}-\alpha_{ic}^{tch})\times[\digamma(\alpha_{ic}^{st}) - \digamma(\sum_{c=1}^{C_{old}}\alpha_{ic}^{st})]. \end{aligned} \tag{9}$$

Considering that the knowledge of the old classes is distilled to the student from the teacher, it is not necessary that the regularizer, L_{EKL} is reapplied to the old classes. For this reason, the L_{EKL} is redefined to focus only on the new classes. The new formulation of L_{EKL} is as follows:

$$\begin{aligned} L_{EKL} &= \mathrm{KL}[\mathrm{Dir}(p_i|\tilde{\alpha_i})||Dir(p_i|\mathbf{1})] = \\ &log\left(\frac{\Gamma(\sum_{c=C_{old}+1}^{C_{new}}\tilde{\alpha_{ic}})}{\Gamma(C)\prod_{c=C_{old}+1}^{C_{new}}\Gamma(\tilde{\alpha_{ic}})}\right) + \\ &\sum_{c=C_{old}+1}^{C_{new}}(\tilde{\alpha_{ic}}-1)\left[\digamma(\tilde{\alpha_{ic}}) - \digamma(\sum_{j=C_{old}+1}^{C_{new}}\tilde{\alpha_{ij}})\right] \end{aligned} \tag{10}$$

Finally, the proposed loss function for the CIL model CEDL+ is defined as follows:

$$L = \begin{cases} L_1, & \text{if } n = 1, \\ 0.5\cdot(L_1+L_2), & \text{otherwise.} \end{cases} \tag{11}$$

Bias Correction: During inference time, a bias correction technique (He & Zhu, 2022) was used considering that the prediction from T_2 to T_n may be biased to the new classes learned. Specifically, the bias correction is defined as follows:

$$f(\hat{x})_c = f(x)_c / \|W_c\|, c \in \{1,2,\ldots,C\} \tag{12}$$

where C corresponds to all the classes seen so far. Then, the refined prediction is given by $o(f(\hat{x}),1)$.

3.4. Uncertainty quantification for OOD detection

The proposed CELD+ performs the OOD detection based on the uncertainty related to the model prediction of the input data. A high and low uncertainty means that the data are probably OOD and IND, respectively. In EDL-based methods, two measures of uncertainty can be estimated: Vacuity and Dissonance (Hu et al., 2021; Zhao et al., 2019). Vacuity measures the lack of evidence and, therefore, can be useful for detecting samples of classes other than those used for training. Dissonance measures the conflict of evidence, which is useful for detecting samples at the decision boundary. In a continual learning framework, vacuity can be used to detect OOD data and dissonance to help identify data belonging to old classes (Aguilar et al., 2023). The latter holds, because after each incremental step, new classes may conflict with previous ones. These uncertainty measures are detailed as follows:

$$\mathrm{Vac}(\alpha) = \frac{C}{S} \tag{13}$$

$$\mathrm{Diss}(\alpha) = \sum_{c=1}^{C} b_c \frac{\sum_{i\neq c} b_i \mathrm{Bal}(b_i, b_c)}{\sum_{i\neq c} b_i} \tag{14}$$

$$\mathrm{Bal}(b_i, b_c) = \begin{cases} 1 - \frac{|b_i - b_c|}{b_i + b_c} & \text{if } b_i b_c \neq 0 \\ 0 & \text{else} \end{cases}$$

where $b = \frac{e}{S}$ is the belief mass.

4. Validation

This section describes three public datasets and the comparison methods for the evaluation of OOD detection. Next, the implementation setting and class-incremental learning settings are detailed. Finally, evaluation metrics for both object recognition and OOD data detection are presented.

4.1. Datasets

The method for continual OOD detection has been validated on three public datasets:

CIFAR-100 (Krizhevsky et al., 2009) is a popular object recognition dataset consisting of 20 superclasses which in turn have 5 classes each. A total of 60,000 low-resolution images (32×32) are evenly distributed over the 100 object classes, of which 50,000 are used for training and the remaining 10,000 for testing.

TinyImageNet is a subset of the ImageNet dataset (Deng et al., 2009). TinyImageNet consists of 200 classes with a total of 120,000 images of 64×64 pixels. The images are distributed as follows: 100,000 for training, 10,000 for validation and the remaining 10,000 for testing, keeping the same ratio of images for each class.

Food101(Bossard et al., 2014) is one of the most popular food recognition datasets. It consists of 101 international dishes with 1,000 images for each dish. Image resolution varies from image to image, with a maximum side of 512 pixels. The dataset is divided as follows: 75% of images for training and the remaining 25% for testing, keeping the same ratio of images for each class.

4.2. Evaluation protocol

Following the protocol used in He and Zhu (2022) for OOD detection evaluation, we compare the proposed approach with several posthoc methods such as: MSP (Hendrycks & Gimpel, 2016), ODIN (Liang et al., 2018), Energy Score (Liu et al., 2020), Entropy (Kuan & Mueller, 2022), MSP-BC-CE (He & Zhu, 2022). For comparison purposes, all of them are applied on top of the baseline method. Moreover, the proposed CEDL+ method was also compared with our previous CEDL method. Fig. 2 exemplifies the evaluation protocol for the proposed CEDL+ when N tasks are considered in a CIL framework. After each incremental learning, OOD detection is evaluated by considering the test data for the classes seen as IND and for the classes of the immediate next task as OOD.

In addition, CEDL+ was evaluated in a cross-dataset scenario. In this case, the IND also corresponds to the test data of the seen classes. However, the OOD corresponds to the data of 20% of the classes from a different dataset than the one used for training.

4.3. Implementation

All experiments have been performed using the same hyperparameters regardless of the dataset. The number of training epochs was 120 for each incremental step, the batch size was 128, an SGD optimizer with a momentum of 0.9 was used, the initial learning rate was $1e-1$, reduced by following a cosine annealing schedule, and the weight decay was $5e-4$. Regarding backbone and image size, we used the following configurations: ResNet32 (He et al., 2016) for CIFAR-100 and image size of 32×32; ResNet18 (He et al., 2016) for TinyImageNet and image size of 64×64; and ResNet18 for Food101 and image size of 224×224.

Regarding the number of exemplars stored in the memory-replay buffer, we have selected 20 samples per class for CIFAR-100 and

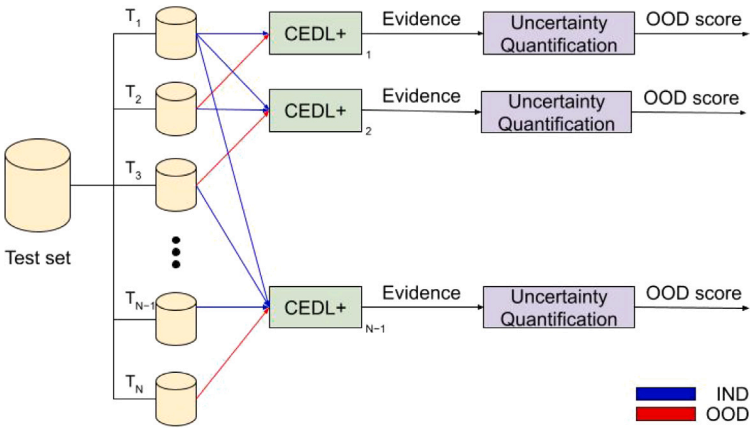

Fig. 2. Schematic to illustrate the evaluation of the proposed CEDL+ method for OOD detection in a CIL framework. For IND, data from the classes seen (see blue arrows) and for OOD, data from the classes of the immediately next task (see red arrows) are used.

Table 1
Settings used to validate the proposed method on the three datasets.

Dataset	Settings	Tasks	Base classes	Step size
CIFAR-100	S_1	10	10	10
	S_2	5	20	20
	S_3	6	50	10
Food101	S_1	10	11	10
	S_2	5	21	20
	S_3	6	51	10
TinyImageNet	S_1	10	20	20
	S_2	5	40	40
	S_3	6	100	20

TinyImageNet, and 50 samples per class in the case of Food101. Finally, to weight the contribution of the different parts of the loss function $\lambda_1 = 0.9$ and $\lambda_2 = 0.9$ were empirically fixed based on CIFAR-100 object recognition performance and were maintained for the rest of the datasets.

The same data preprocessing was applied to all selected datasets. Image normalization is a popular approach using the mean and standard deviation from the large-scale object recognition dataset ImageNet (Krizhevsky et al., 2012). In addition, during the training phase, the RandAugment (Cubuk et al., 2020) method was used to increase the variability of the data and thus avoid the overfitting problem.

All methods were implemented using the machine learning framework Pytorch and the experiments were performed on a computer with an NVIDIA RTX 2080 TI graphics card.

4.4. Class-incremental learning settings

We adopted two popular settings to validate the proposed method for the CIL problem. One consists of evenly distributing the classes over several tasks (Aljundi et al., 2022; He & Zhu, 2022) and the other one uses 50% of the classes for the first task and then evenly distributes the rest (Wiewel et al., 2022). Table 1 provides a summary of the incremental learning settings: S_1 and S_2 correspond to the first setting, while S_3 corresponds to the second setting.

For each dataset, to define the tasks according to the settings mentioned above, we have previously shuffled the classes according to Wiewel et al. (2022).

4.5. Evaluation metrics

To validate the proposed method, classical metrics for both object recognition and OOD sample detection were selected.

4.5.1. Object recognition

The evaluation of object recognition performance is usually performed using two metrics: Average Classification Accuracy (ACA), which evaluates the performance of the model at the end of performing all increments, and Average Incremental Accuracy (AIA), which evaluates the performance after each increment and then averages all individual evaluations. The formal definition of these metrics is as follows:

$$\text{ACA} = \text{ACC}_T, \quad \text{AIA} = \frac{1}{T} \cdot \sum_{n=1}^{T} \text{ACC}_n \tag{15}$$

4.5.2. OOD sample detection

To perform OOD sample detection, the dataset was divided into four subsets:

- IND_c corresponds to the data belonging to the current classes used for training on the tth task.
- IND_p corresponds to the data belonging to the forgotten classes, i.e. the data used in the previous task.
- IND corresponds to the data belonging to the seen classes.
- OOD corresponding to the data belonging to the next task (unseen classes).

For comparative purposes, the classical metrics for evaluating OOD detection are selected. These are: the Area Under the Receiver Operating Characteristic curve (AUROC) and the False Positive Rate at least 95% of the true positive rate (FPR95). AUROC provides a score between 0 to 1 to determine the chance that a pair of IND and OOD samples are correctly distinguished. An AUROC equal to 0 means that the samples have been completely misclassified, 1 is completely well-classified and 0.5 that the classification has been random. As for FPR95, it can be interpreted as the probability that a negative sample (OOD) can be misclassified as a positive sample (IND), when recall is as high as 95%. Therefore, a higher AUROC value and a lower FPR95 value indicate better detection performance. Note that to evaluate the performance of OOD detection in terms of AUROC and FPR95, OOD subset is considered as a negative class and the rest are considered as positive classes, except when IND_c is compared to IND_p, in which case IND_p is treated as a negative class.

5. Results

In this section, the results of object recognition and OOD sample detection are presented and discussed. In addition, an analysis of the model performance in a cross-dataset scheme and an analysis of the usefulness of the vacuity and dissonance measure for discriminating between IND_c, IND_p and OOD are performed. Finally, an ablation study for the loss function is performed.

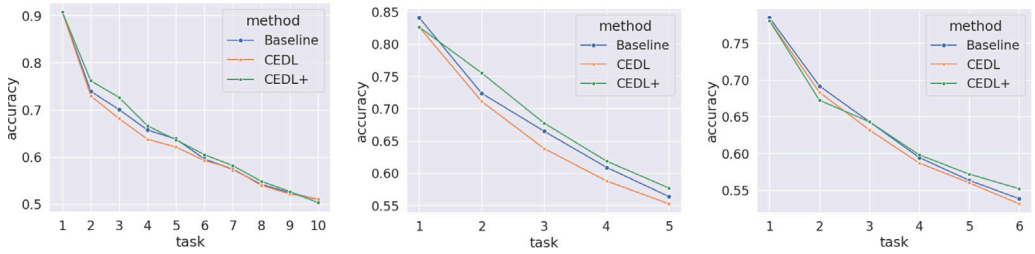

Fig. 3. Overall accuracy obtained in CIFAR-100 by the baseline, CEDL and CEDL+ methods after each incremental step corresponding to three incremental settings.

Table 2
Classification results in terms of ACA and AIA.

		S_1		S_2		S_3	
Dataset	Method	ACA	AIA	ACA	AIA	ACA	AIA
CIFAR-100	Baseline	50.96%	63.87%	**56.36%**	**68.02%**	53.85%	63.59%
	CEDL	**51.07%**	63.16%	55.61%	66.27%	53.15%	62.90%
	CEDL+	50.31%	**64.62%**	**56.36%**	67.53%	**55.21%**	**63.63%**
TinyImageNet	Baseline	32.94%	44.80%	38.75%	49.87%	34.23%	43.58%
	CEDL	32.68%	45.43%	38.73%	50.79%	34.16%	44.11%
	CEDL+	**35.27%**	**48.96%**	**40.90%**	**53.40%**	**37.62%**	**45.69%**
Food101	Baseline	59.78%	68.69%	68.63%	74.94%	63.64%	72.10%
	CEDL	63.34%	73.58%	70.65%	77.89%	64.24%	**72.36%**
	CEDL+	**64.80%**	**75.17%**	**71.36%**	**78.68%**	**65.56%**	71.66%

5.1. Continual object recognition

The results obtained for the three datasets using the evaluation methods to perform object recognition in several class-incremental settings are presented in Table 2. Here, the baseline method is equivalent to the one proposed in Zhao et al. (2020), CEDL corresponds to the previous work published in Aguilar et al. (2023) and CEDL+ is the proposed extended version of the CEDL method. The maximum response given in the output layer by the methods is used to determine the predicted class. In the case of the CIFAR-100 dataset, it can be noticed that the proposed CEDL+ method improves by more than 1% almost all the results obtained by CEDL and provides comparable results even better than the baseline. For more complex datasets, such as TinyImageNet and Food101, the improvement is more noticeable, outperforming the results obtained by the baseline and CEDL by a wide margin in three different settings. Moreover, specifically for Food101, it can be noticed that EDL-based approaches provide much better performance than the baseline, mainly for settings S_1 and S_2. As for S_3, unlike the other settings, the first task contains 50% of the classes and the remaining classes are equally distributed in the remaining 5 tasks. It can be noticed that this setting is more challenging, showing in the experiments the production of the highest knowledge loss after training task 2. This has an impact on a lower performance improvement compared to the other settings.

In addition to the ACA and AIA metrics which provide an overview of the performance of the model, the accuracy after each task in CIFAR-100 is shown in Fig. 3 for S_1, S_2 and S_3 settings. As can be seen, the proposed approach improved the performance in comparison with the rest of the models, even though in some cases for the first task it performed worse. Particularly in the S_3 setting, it can be seen that the proposed method provides less performance than the rest after the second task. In this case, the high-class difference between the first two tasks influences more the proposed model than the rest which then stabilizes and manages anyway to provide a better performance.

The role of the buffer size: In the replay-based method, the buffer size and sample selection strategy may affect the stability of the model, i.e. the ability to retain prior knowledge while acquiring new knowledge. Buffer size was selected in the experiments based on the most commonly used value in the CL methods validated in CIFAR-100 (Wiewel et al., 2022), TinyImageNet (Aljundi et al., 2022) and Food101 (Raghavan et al., 2024), as indicated in Section 4.3.

Table 3
AIA obtained by CEDL+ on CIFAR-100 using different buffer sizes and the incremental setting S_2.

Buffer size	AIA
200	59.84%
500	62.48%
1000	65.10%
2000	67.53%

As is expected, the larger the buffer size is, the higher the performance of the model. In Table 3, object recognition performance was evaluated as a function of buffer size for CIFAR-100 using the S_2 incremental setting. As expected, performance was directly proportional to the buffer size used. If we compare the results obtained with those reported in Wiewel et al. (2022), the proposed method tends to be more sensitive when a very small buffer size (<= 500) is applied. However, for larger buffer sizes (>= 1000), the method provides better performance. This suggests that, to improve the quality of the OOD detection provided by the model in a continual learning framework, it is important to increase the buffer size where the exemplars are stored.

5.2. Continual OOD sample detection

After each incremental step, the ability of the baseline, CEDL and CEDL+ methods to detect OOD samples was evaluated on CIFAR-100 (see Fig. 4). For comparison purposes, five popular posthoc methods were considered at the top of the baseline model. In the case of CEDL and CEDL+, the vacuity measure was selected for OOD detection due to its better performance compared with dissonance, as demonstrated in Aguilar et al. (2023). The results obtained in terms of AUROC show that the proposed CEDL+ outperforms the state-of-the-art by a wide margin in all scenarios. Specifically for S_2 and S_3 settings, the improvement is clearly noticeable after the second task showing a better ability to prevent catastrophic forgetting.

Table 4 shows the average results obtained for OOD detection, in terms of AUROC and FPR95, using different incremental settings, datasets and methods. Here, all test data belonging to the seen classes (including the current task) are considered IND and test data belonging to the classes after the current task are considered OOD. Interestingly, in all settings and for all datasets, the CEDL+ method, which uses vacuity as a measure to differentiate the IND sample from the OOD, is superior to the rest in both AUROC and FPR95. Only in one case, CEDL+ does perform slightly worse considering the metric AUROC (0.44% less), specifically, when compared to MSP-BC-CE using the S_3 setting in the Food101 dataset. Despite the latter case, considering both measures (AUROC and FPR95), the CEDL+ method is still better, achieving a comparable result in AUROC and an improvement of about 5% in terms of FPR95.

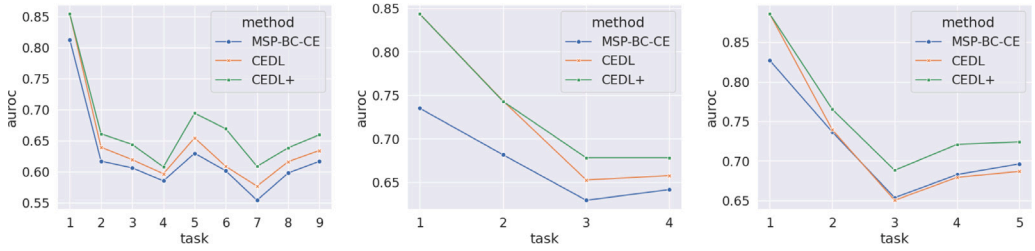

Fig. 4. AUROC obtained in CIFAR-100 by the MSP-BC-CE, CEDL and CEDL+ methods after each incremental step corresponding to three incremental settings.

Table 4
OOD detection: Average AUROC and FPR95 on CIFAR-100, TinyImageNet and Food101 over all tasks within incremental settings S_1, S_2 and S_3 by several state-of-the-art posthoc methods, CEDL and CEDL+. The best results are shown in bold.

CIFAR-100						
Method	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
MSP	61.14%	91.91%	66.34%	90.09%	66.80%	90.02%
ODIN	60.23%	92.43%	65.39%	90.73%	65.69%	91.02%
Energy Score	60.85%	91.29%	65.03%	90.76%	64.50%	91.20%
Entropy	61.72%	91.05%	66.71%	89.74%	66.43%	90.52%
MSP-BC-CE	62.47%	91.14%	67.18%	90.00%	71.92%	83.06%
CEDL	64.44%	85.87%	72.42%	76.09%	72.82%	77.08%
CEDL+	**67.08%**	**83.96%**	**73.55%**	**73.90%**	**75.67%**	**72.14%**
TinyImageNet						
Method	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
MSP	60.27%	92.18%	62.61%	92.56%	61.01%	92.60%
ODIN	60.06%	92.74%	62.42%	92.59%	60.77%	92.80%
Energy Score	58.72%	93.04%	60.46%	92.66%	59.02%	93.28%
Entropy	59.78%	92.43%	61.90%	92.28%	60.48%	92.76%
MSP-BC-CE	63.86%	88.06%	67.59%	86.43%	64.70%	88.58%
CEDL	67.37%	82.31%	73.95%	73.74%	67.10%	80.38%
CEDL+	**70.94%**	**78.44%**	**76.32%**	**69.86%**	**73.39%**	**73.74%**
Food101						
Method	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑	FPR95 ↓
MSP	60.27%	92.18%	70.20%	88.46%	72.09%	86.62%
ODIN	60.06%	92.74%	69.02%	89.62%	70.67%	88.95%
Energy Score	58.72%	93.04%	67.05%	89.69%	68.87%	88.95%
Entropy	59.78%	92.43%	69.07%	89.04%	70.85%	88.06%
MSP-BC-CE	63.86%	88.06%	77.03%	77.24%	**77.85%**	75.76%
CEDL	70.63%	81.74%	79.17%	71.13%	75.62%	74.31%
CEDL+	**72.40%**	**78.87%**	**80.37%**	**66.99%**	77.41%	**70.86%**
	S_1		S_2		S_3	

5.3. Cross-dataset evaluation

OOD detection within the same dataset is a complex task, because it is possible that classes still belonging to different tasks could be visually similar (sharing some features). In a continual learning framework, it is also natural to consider the scenario in which data belong to multiple domains (in this case, one domain will represent one task). For this reason, the capability of the proposed method was evaluated considering a cross-dataset scenario in which IND is the data from one dataset and OOD is the data from another one. Specifically, the S_2 incremental setting was used, where after each task, the test samples of the classes seen from the dataset used for training were considered as IND and the samples corresponding to 20% of the classes from the other dataset as OOD data. The mean performance of the individual results for each task in terms of AUROC and FPR95 is shown in Table 5. The first column lists the datasets with the IND samples and columns two through four list the datasets with the OOD samples. Compared to the results in Table 4, it can be observed that the model can detect OOD samples better when the data come from different domains. The more significant the difference between domains is, the greater the ability to detect OOD samples. For example, in the case when using CIFAR-100 as IND and Food101 as OOD, a success rate above 90% for detecting OOD data is achieved.

Table 5
OOD detection results obtained by CEDL+ in terms of AUROC and FPR95 using a cross-dataset evaluation. Five incremental steps were considered and the average AUROC and FPR95 were reported.

	AUROC		
	CIFAR-100	TinyImageNet	Food101
CIFAR-100	–	79.20%	90.71%
TinyImageNet	79.86%	–	89.26%
Food101	91.20%	91.42%	–
	FPR95		
	CIFAR-100	TinyImageNet	Food101
CIFAR-100	–	64.96%	41.40%
TinyImageNet	64.36%	–	44.66%
Food101	37.15%	35.69%	–

5.4. Influence of catastrophic forgetting on OOD detection

In addition to standard evaluation metrics for OOD detection, it is interesting to analyze how catastrophic forgetting affects OOD detection. We argue that catastrophic forgetting may negatively affect OOD detection in a continual learning framework because model embeddings may not reflect the characteristics of old classes and decision boundaries may become less reliable. Specifically, during OOD detection, in class-incremental learning, we want to identify different data from those corresponding to the classes learned during training. If during the next incremental learning, the previous classes are forgotten, these classes will be unknown to the model and could be misclassified as OOD data.

To illustrate the influence of catastrophic forgetting on OOD detection in the proposed CEDL+ one can look at Fig. 5. In the left column of Fig. 5, one can see how the object recognition and OOD detection performance changes after each incremental step. The reported results contemplate a particular evaluation protocol to highlight the forgetting and be able to analyze how the forgetting affects the OOD detection. For this purpose, the test data in the first task is considered IND, and the test data for the last task is OOD. It can be observed that the accuracy of the first task data (ACC-1st) decreases after each incremental step. This is due to both (1) the addition of new classes to the model and (2) the ability to retain knowledge about previously learned classes. In particular, it is observed that a large decrease in performance occurs in the evaluation after the second incremental step. Similarly, forgetting affects AUROC performance mainly in the early tasks, and after task 7 tends to be stable. On the other hand, focusing on the OOD detection (see the second column of Fig. 5), also a large decrease happens in the second incremental step. Moreover, it is observed that the good performance achieved when comparing the first task test data (the orange bar) with the OOD data in the first incremental step decreases in the following ones, being even worse compared to the performance of all seen class data (the blue bar). This demonstrates the effect of catastrophic forgetting on OOD detection, where older classes (and forgotten to some extent) tend to provide more errors in OOD detection than newer classes learned.

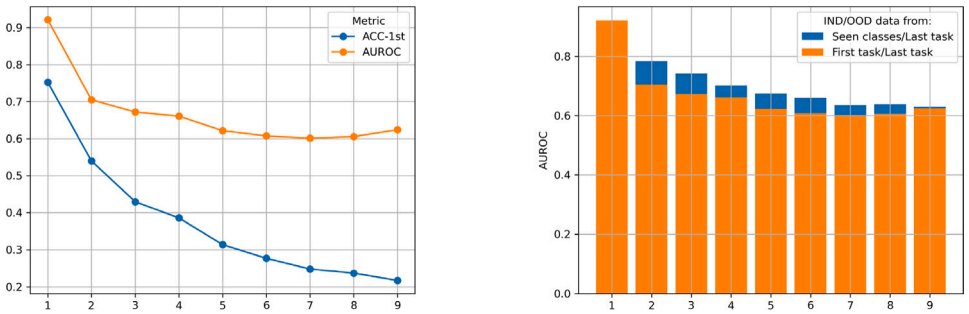

Fig. 5. Object recognition and OOD data detection performance on TinyImageNet using the class-incremental setting S_1. The left column shows the ACC-1st and AUROC, which are determined after each incremental step. For AUROC the test data of the classes corresponding to the first task are considered IND and for the last task OOD. The right column shows that AUROC considers two cases: (a) First tasks data as IND (orange bar); and (b) Data from seen classes as IND (blue bar). In both cases, OOD corresponds to the data from the last task.

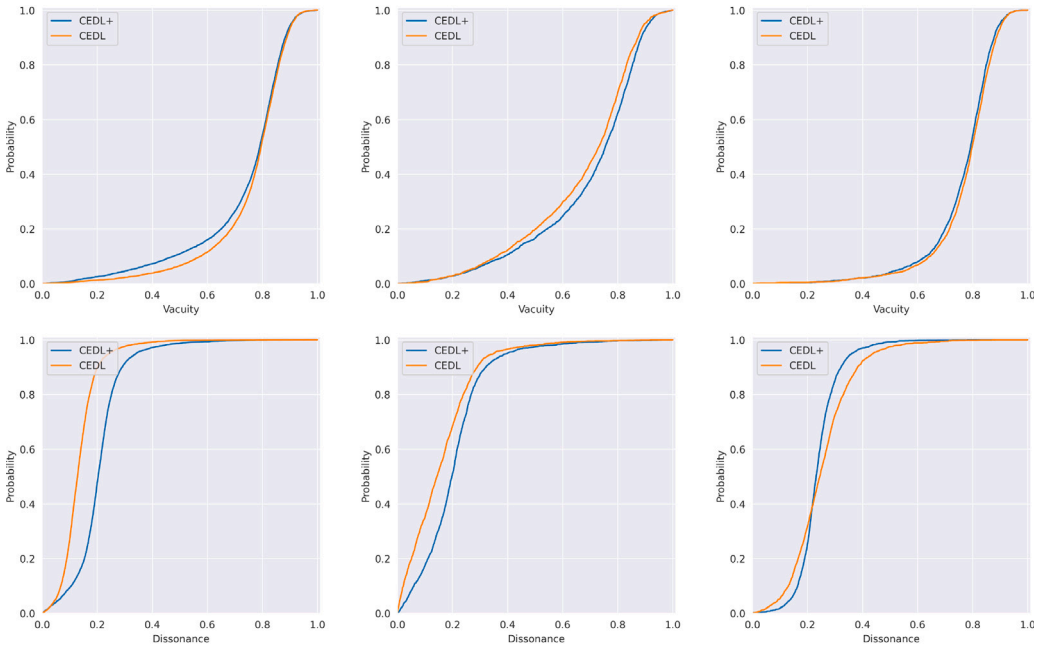

Fig. 6. CDF of Vacuity (top) and Dissonance (bottom), using the model predictions on CIFAR-100 after task 4 in the incremental setting S_2. From left to right, in the CDF of the data one can be seen corresponding to the previous classes (left), the current classes (center) and OOD data (right).

5.5. Vacuity and dissonance analysis

Vacuity and Dissonance are two measures of uncertainty that can be estimated in Evidential Deep Learning. Both measures are analyzed in the Continual Learning framework to determine why CEDL+ performs better than CEDL and also to understand the usefulness of these measures for interpreting results and identifying the type of target data. For a given sample, a high vacuity means that the model has no evidence to determine the predicted class and, therefore, it is likely that the class corresponds to one that the model does not know. On the other hand, a high dissonance means that the model has evidence for more than one known class, which is since different classes share some extracted features.

In Fig. 6, the Cumulative Distribution Function (CDF) for the Vacuity and Dissonance provided by CEDL+ and CEDL can be observed on the three subsets: IND_p (left), IND_c (middle) and OOD (right). For this purpose, the model trained after task 4 was used. From each plot we can interpret the probability of the samples obtaining a vacuity or dissonance lower than a predetermined value in the interval from 0 to 1. Considering the meaning of vacuity and dissonance, it is expected to have low vacuity and low dissonance in the classes seen, low vacuity and high dissonance in the old classes, and high vacuity and low dissonance in the unseen classes. Looking at the vacuity, both CEDL+ and CEDL methods provide almost the same CDF for the OOD subset, where it can be observed that they have a higher chance of having high vacuity than IND_p and IND_c subsets. As for IND_c, CEDL provides a slightly better vacuity estimate than CEDL+, because it is expected to have a low vacuity on data relating to classes seen, especially if these data correspond to the most recently learned classes. However, for IND_p CEDL performs worse than CEDL+ demonstrated by a distribution close to OOD. At best, the CDFs of IND_p and IND_c should be almost the same and completely different from those of OOD. The former two tend to lie to the left of the graph and the latter to the right. This behavior is close to that provided by the proposed CEDL+ and that is why this method can better detect the data corresponding to the classes seen in comparison with the OOD samples.

On the other hand, looking at the dissonance, one can observe again a better behavior of the CEDL+, where the dissonance is lower for the IND_c, higher for the OOD and in between for the IND_p. More details on dissonance and vacuity at different degrees of change can be seen in Table 6. It is interesting to note that when considering a chance of 0.5 in the CDF of the dissonance, the CEDL method provides an unexpected behavior when analyzing the score obtained in the different subsets in comparison with that obtained at lower chance values. Specifically,

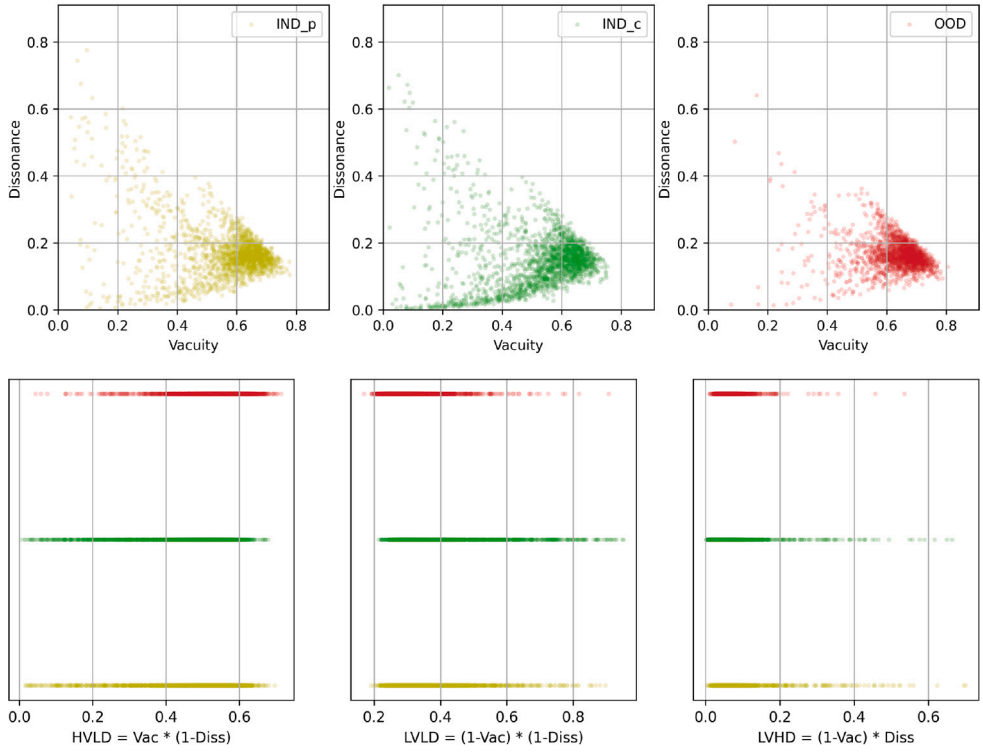

Fig. 7. Scatter plots with vacuity and dissonance obtained by CEDL+ on CIFAR-100 after task 4 considering S_2 incremental step. Data corresponding to previous classes, current classes and unseen classes are shown in yellow, green and red, respectively.

Table 6
Vacuity and Dissonance at different degrees of chance based on the CDF of the predictive distributions.

		Vacuity			Dissonance		
Method	Subset	0.10	0.25	0.50	0.10	0.25	0.50
CEDL+	IND_p	0.4775	0.6896	0.7909	0.1054	0.1657	0.2037
	IND_c	0.3894	0.6036	0.7558	0.0601	0.1311	0.1982
	OOD	0.6286	0.7227	0.7938	0.1652	0.2002	0.2322
CEDL	IND_p	0.5752	0.7182	0.7956	0.0680	0.0974	0.1287
	IND_c	0.3624	0.5586	0.7317	0.0220	0.0661	0.1441
	OOD	0.6456	0.7367	0.7988	0.1311	0.1832	0.2452

in this case, the dissonance of IND_c is higher than that of IND_p. This unexpected behavior can be detrimental in the case of combining both measures.

The combination of both measures is analyzed qualitatively taking into account the results of the proposed CEDL+ method. A scatter plot for the vacuity and dissonance obtained for the IND_p, IND_c and OOD subsets is presented in Fig. 7 at the top. Specifically, in the case of IND_p only the data from the directly preceding task are used to consider the same number of points in all of them. Three possible combinations of vacuity and dissonance were identified from the scatter plot: High Vacuity and Low Dissonance (HVLD); Low Vacuity and Low Dissonance (LVLD); and Low Vacuity and High Dissonance (LVHD). With these combinations, a score was obtained and is shown in the lower part of Fig. 7. Here, it can be seen that the OOD data tend to be those with the highest vacuity and the lowest dissonance. IND_p and IND_c have almost the same behavior, but with slightly higher vacuity and dissonance for IND_p. In addition, when analyzing the combination of the measures, we can observe that the values for HVLD, LVLD and LVHV of the data corresponding to OOD, IND_c and IND_p tend to be higher, respectively. This suggests that the combination of vacuity and dissonance may provide us with some clues to differentiate the data in these three subsets and devise specific strategies for each of them.

Fig. 8 shows some samples obtained considering the test data belonging to the CIFAR-100 with HVLD, LVLD and LVHD maxima together with the respective Dirichlet plot on simplex. The CEDL+ model used was the one trained after task 4. Surprisingly, in this case all samples belong to different subsets, as explained above. The highest values of HVLD, LVLD and LVHD correspond to the OOD samples, IND_c and IND_p, respectively. For the OOD samples, it can be observed that the model provides a lack of evidence and provides a random prediction for classes that are completely different in comparison with the target image. In the case of IND_c, the model provides a correct and very confident prediction. Finally, for IND_p data, the model provides a conflict of evidence and in this case all classes are related to the target image.

5.6. Ablation study

Role of L_{KD} and L_{EKD} for the model improvement. When comparing CEDL+ and CEDL, the main difference occurs in the formulation of the loss function. Specifically, CEDL uses only L_{KD} and CEDL+ considers additionally the incorporation of L_{EKD} in the knowledge distillation part of the loss function. In CEDL+, it is not only considered distilling the evidence from the teacher model to the student model, but also the entire Dirichlet distribution. As noted above, this change produces a marked performance improvement, outperforming results across all metrics (ACA, AUROC and FPR95), datasets (CIFAR-100, TinyImageNet and Food101) and scenarios (S_2 and S_3) evaluated. To analyze the importance of both components, L_{KD} and L_{EKD}, L_2 is adapted by removing one or both of them. The evaluation is performed on four possible losses: (1) If both L_{KD} and L_{EKD} are eliminated, the new loss function will consider $L_2 = 0$; (2) If L_{EKD} is eliminated, the new loss function will be $L_2 = \lambda_2 \cdot L_{KD}$; (3) If L_{KD} is eliminated from L_2, the new loss function will be $L_2 = (1-\lambda_2) \cdot L_{EKD}$; and (4) If no loss is eliminated from L_2, the loss function is as proposed in CEDL+.

The results are reported in Table 7 for both object recognition and OOD sample detection. For the analysis, settings S_2 and S_3 were

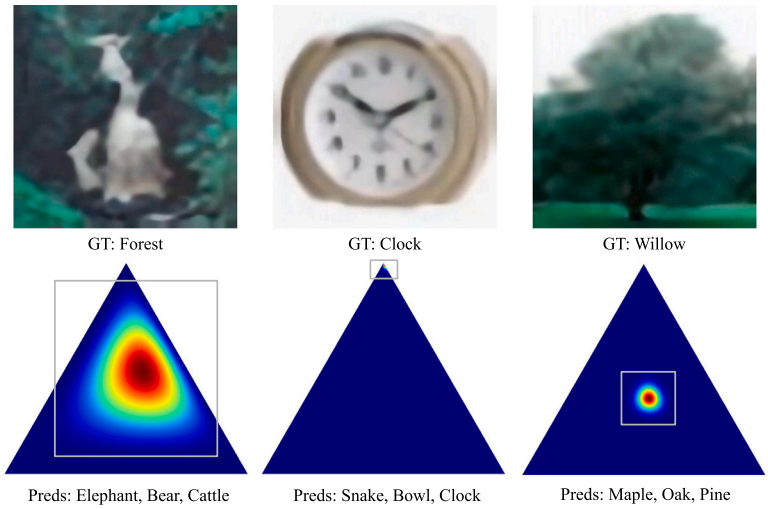

Fig. 8. Samples of images based on the maximum score obtained by HVLD (left), LVLD (middle) and LVHD (right) formulas. The vacuity and dissonance were obtained by CEDL+ on CIFAR-100 after task 4 considering the S_2 incremental setting.

Table 7
Object recognition and OOD detection results obtained by CEDL+ when some of the components of the loss function were removed.

CIFAR-100							
L_{KD}	L_{EKD}	ACA	AUROC ↑	FPR95 ↓	ACA	AUROC ↑	FPR95 ↓
✗	✗	47.6%	69.1%	79.3%	45.6%	69.6%	80.5%
✓	✗	55.6%	72.4%	76.1%	53.2%	72.8%	77.1%
✗	✓	**57.7%**	**75.9%**	**70.2%**	54.7%	**76.6%**	72.2%
✓	✓	56.4%	73.6%	73.9%	**55.2%**	75.7%	**72.1%**
TinyImageNet							
L_{KD}	L_{EKD}	ACA	AUROC ↑	FPR95 ↓	ACA	AUROC ↑	FPR95 ↓
✗	✗	30.3%	70.3%	76.0%	29.7%	66.2%	80.6%
✓	✗	38.7%	74.0%	73.7%	34.2%	67.1%	80.4%
✗	✓	38.6%	74.9%	71.3%	32.1%	71.7%	75.6%
✓	✓	**40.9%**	**76.3%**	**69.9%**	**37.6%**	**73.4%**	**73.7%**
Food101							
L_{KD}	L_{EKD}	ACA	AUROC ↑	FPR95 ↓	ACA	AUROC ↑	FPR95 ↓
✗	✗	64.2%	75.0%	75.2%	59.2%	73.0%	76.8%
✓	✗	70.7%	79.2%	71.1%	64.2%	75.6%	74.3%
✗	✓	69.2%	79.9%	67.2%	64.1%	77.1%	**70.7%**
✓	✓	**71.4%**	**80.4%**	**67.0%**	**71.7%**	**77.4%**	70.9%
		S_2			S_3		

considered, because they consist of a similar number of tasks and cover two completely different settings; one with classes evenly distributed in the tasks and the other with a large difference in the number of classes contained in the first task in comparison with the rest. The entire combination of the two components of the loss function was evaluated on the three public datasets. As expected, when none of the losses are taken into account for knowledge distillation, the model is strongly affected by catastrophic forgetting, resulting in poor performance in both object recognition and OOD sample detection. On the other hand, when only using L_{EKD} in L_2 the model presents a better OOD detection capability compared to only using L_{KD}, even when the object recognition performance is worse. This suggests that the distillation of the entire Dirichlet distribution contributes to the model generating a better quality of uncertainty to differentiate IND vs. OOD data. When only L_{KD} is used in L_2, the model usually provides better object recognition performance compared to using L_{EKD}. Furthermore, it can be observed that S_3 affects the model capacity more when L_{EKD} is used alone and not in conjunction with L_{KD}. Finally, it can be observed that a better balance between object recognition and OOD detection performance can be achieved when both components are used in the loss function.

5.7. Limitations

The proposed CEDL+ method presents a remarkable performance for OOD detection compared to the rest of the evaluated methods. However, unlike the other posthoc methods, CEDL+ requires small changes in the model design to learn the evidence and changes in the loss function to preserve it. CEDL+ can be applied to several CL methods that include knowledge distillation in their approach, but may be more challenging for other CL techniques.

6. Conclusions

In this paper, we proposed a new method for continual out-of-distribution detection. Therefore, we extended evidential deep learning to the non-stationary setting. In addition, we introduced a new component in the loss function that distills knowledge taking into account a full Dirichlet distribution. The estimated vacuity is then used as an uncertainty measure during inference. The proposed CEDL+ method outperformed several posthoc methods by a wide margin on three public datasets, using several incremental learning settings for the OOD detection problem. Furthermore, CEDL+ also outperformed the results in object recognition in terms of ACA and AIA. On the other hand, the ability of the proposed method to detect OOD data was also validated in a cross-dataset scenario, achieving in some cases a very high detection performance (over 90% in terms of AUROC). Finally, measures of vacuity and dissonance were analyzed. From this analysis, it could be observed that different combinations of them can provide some insights to characterize the data into IND_p, IND_c and OOD. In future work, we will evaluate the integration of other deterministic methods to quantify uncertainty in a continual learning framework to evaluate its ability to detect OOD and provide a still more reliable solution for open-world visual recognition.

CRediT authorship contribution statement

Eduardo Aguilar: Conceptualization, Methodology, Software, Investigation, Writing – original Draft, Visualization, Funding acquisition. **Bogdan Raducanu:** Investigation, Writing – original Draft, Supervision, Funding acquisition. **Petia Radeva:** Conceptualization, Writing – review & editing, Funding acquisition. **Joost van de Weijer:** Methodology, Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is funded by ANID (No. FONDECYT DE INICIACIÓN 11230262), the Horizon EU project MUSAE (No. 01070421), 2021-SGR-01094 (AGAUR), Icrea Academia'2022 (Generalitat de Catalunya), Robo STEAM (2022-1-BG01-KA220-VET-000089434, Erasmus+ EU), DeepSense (ACE053/22/000029, ACCIÓ), DeepFoodVol (AEI-MICINN, PDC2022-133642-I00), PID2022-141566NB-I00 (AEI-MICINN), CERCA Programme / Generalitat de Catalunya, and Grants TED2021-132513B-I00, PID2022-143257NB-I00 by MCIN/AEI/10.13039/501100011033, by the European Union NextGenerationEU/PRTR, and by ERDF A Way of Making Europa.

Data availability

The datasets used during the current study are available at the following URL: https://github.com/Continvvm/continuum/tree/master/continuum/datasets.

References

Aguilar, E., Raducanu, B., Radeva, P., & Van de Weijer, J. (2023). Continual evidential deep learning for out-of-distribution detection. In *Proceedings of the IEEE ICCV workshops* (pp. 3444–3454).

Akyürek, A. F., Akyürek, E., Wijaya, D. T., & Andreas, J. (2022). Subspace regularizers for few-shot class incremental learning. In *ICLR*.

Aljundi, R., Babiloni, F., Elhoseiny, M., Rohrbach, M., & Tuytelaars, T. (2018). Memory aware synapses: Learning what (not) to forget. In *ECCV*.

Aljundi, R., Reino, D. O., Chumerin, N., & Turner, R. E. (2022). Continual novelty detection. In *Conference on lifelong learning agents* (pp. 1004–1025). PMLR.

Bao, W., Yu, Q., & Kong, Y. (2021). Evidential deep learning for open set action recognition. In *Proceedings of the IEEE ICCV* (pp. 13349–13358).

Bossard, L., Guillaumin, M., & Van Gool, L. (2014). Food-101–mining discriminative components with random forests. In *ECCV*.

Buzzega, P., Boschini, M., Porrello, A., Abati, D., & Calderara, S. (2020). Dark experience for general continual learning: a strong, simple baseline. In *NeurIPS*.

Cermelli, F., Massimiliano, M., Rota Bulo, S., Ricci, E., & Caputo, B. (2020). Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE CVPR* (pp. 9233–9242).

Cha, S., Hsu, H., Hwang, T., Calmon, F. P., & Moon, T. (2021). CPR: classifier-projection regularization for continual learning. In *ICLR*.

Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE CVPR workshops* (pp. 702–703).

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE CVPR* (pp. 248–255). IEEE.

Douillard, A., Cord, M., Ollion, C., Robert, T., & Valle, E. (2020). PODNet: Pooled outputs distillation for small-tasks incremental learning. In *ECCV*.

Gao, R., & Liu, W. (2023). DDGR: Continual learning with deep diffusion-based generative replay. In *ICML, vol. 202* (pp. 10744–10763).

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. (2023). A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, *56*(Suppl 1), 1513–1589.

Hayes, T. L., Kafle, K., Shrestha, R., Acharya, M., & Kanan, C. (2020). Remind your network to prevent catastrophic forgetting. In *ECCV*.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE CVPR* (pp. 770–778).

He, J., & Zhu, F. (2022). Out-of-distribution detection in unsupervised continual learning. In *2022 IEEE CVPR workshops* (pp. 3849–3854).

Hendrycks, D., & Gimpel, K. (2016). A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*.

Hinton, G., Vinyals, O., & Dean, J. (2014). Distilling the knowledge in a neural network. In *NeurIPS deep learning workshop*.

Holmquist, K., Klasén, L., & Felsberg, M. (2023). Evidential deep learning for class-incremental semantic segmentation. In *Scandinavian conference on image analysis* (pp. 32–48). Springer.

Hu, Y., Ou, Y., Zhao, X., Cho, J. H., & Chen, F. (2021). Multidimensional uncertainty-aware evidential neural networks. In *Proceedings of the AAAI conference on artificial intelligence, vol. 35, no. 9* (pp. 7815–7822).

Jha, S., Gong, D., Zhao, H., & Yao, L. (2023). NPCL: Neural processes for uncertainty-aware continual learning. In *NeurIPS*.

Jung, D., Lee, D., Hong, S., Jang, H., Bae, H., & Yoon, S. (2023). New insights for the stability-plasticity dilemma in online continual learning. In *ICLR*.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS, vol. 30*.

Kim, J., Cho, H., Kim, J., Tiruneh, Y. Y., & Baek, S. (2024). SDDGR: Stable diffusion-based deep generative replay for class incremental object detection. In *Proceedings of the IEEE CVPR* (pp. 28772–28781).

Kim, G., Esmaeilpour, S., Xiao, C., & Liu, B. (2022). Continual learning based on OOD detection and task masking. In *Proceedings of the IEEE ICCV workshops*.

Kim, S., Noci, L., Orvieto, A., & Hofmann, T. (2023). Achieving a better stability-plasticity trade-off via auxiliary networks in continual learning. In *Proceedings of the IEEE CVPR* (pp. 11930–11939).

Kim, G., Xiao, C., Konishi, T., & Liu, B. (2023). Learnability and algorithm for continual learning. In *ICML* (pp. 16877–16896).

Kim, G. K., Xiao, C., Konishiy, T., Ke, Z., & Liu, B. (2022). A theoretical study on solving continual learning. In *NeurIPS*.

Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, *114*(13), 3521–3526.

Krizhevsky, A., Hinton, G., et al. (2009). *Learning multiple layers of features from tiny images*: *Tech report*, Toronto, ON, Canada.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *NeurIPS, vol. 25*.

Kuan, J., & Mueller, J. (2022). Back to the basics: Revisiting out-of-distribution detection baselines. In *ICML workshop on principles of distribution shift*.

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS, vol. 30*.

Liang, S., Li, Y., & Srikant, R. (2018). Enhancing the reliability of out-of-distribution image detection in neural networks. In *ICLR*.

Liang, J., Zhong, J., Gu, H., Lu, Z., Tang, X., Dai, G., Huang, S., Fan, L., & Yang, Q. (2024). Diffusion-driven data replay: A novel approach to combat forgetting in federated class continual learning. In *ECCV*.

Liu, W., Wang, X., Owens, J., & Li, Y. (2020). Energy-based out-of-distribution detection. *NeurIPS*, *33*, 21464–21475.

Masana, M., Liu, X., Twardowski, B., Menta, M., Bagdanov, A. D., & Van De Weijer, J. (2023). Class-incremental learning: survey and performance evaluation on image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(5), 5513–5533.

Mundt, M., Pliushch, I., Majumder, S., Hong, Y., & Ramesh, V. (2022). Unified probabilistic deep continual learning through generative replay and open set recognition. *Journal of Imaging*, *8*(4).

Oh, Y., Baek, D., & Ham, B. (2022). ALIFE: Adaptive logit regularizer and feature replay for incremental semantic segmentation. In *NeurIPS*.

Poyser, M., & Breckon, T. P. (2024). Neural architecture search: A contemporary literature review for computer vision applications. *Pattern Recognition*, *147*, Article 110052.

Qian, Z., Huang, K., Wang, Q. F., & Zhang, X. Y. (2022). A survey of robust adversarial training in pattern recognition: Fundamental, theory, and methodologies. *Pattern Recognition*, *131*, Article 108889.

Raghavan, S., He, J., & Zhu, F. (2024). Online class-incremental learning for real-world food image classification. In *Proceedings of the IEEE WACV* (pp. 8195–8204).

Rebuffi, S. A., Kolesnikov, A., Sperl, G., & Lampert, C. H. (2017). icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE CVPR* (pp. 2001–2010).

Rios, A., Ahuja, N., Ndiour, I., Genc, U., Itti, L., & Tickoo, O. (2022). incDFM: Incremental deep feature modeling for continual novelty detection. In *ECCV*.

Roy, S., Liu, M., Zhong, Z., Sebe, N., & Ricci, E. (2022). Class-incremental novel class discovery. In *ECCV* (pp. 317–333). Springer.

Sensoy, M., Kaplan, L., & Kandemir, M. (2018). Evidential deep learning to quantify classification uncertainty. *NeurIPS*, *31*.

Tiwari, R., Killamsetty, K., Iyer, R., & Shendy, P. (2022). GCR: Gradient coreset based replay buffer selection for continual learning. In *Proceedings of the IEEE CVPR* (pp. 100–108).

Wang, L., Zhang, M., Jia, Z., Li, Q., Ma, K., Bao, C., Zhu, J., & Zhong, Y. (2021). AFEC: Active forgetting of negative transfer in continual learning. In *NeurIPS*.

Wiewel, F., Bartler, A., & Yang, B. (2022). Dirichlet prior networks for continual learning. In *IJCNN* (pp. 1–8). IEEE.

Wu, C., Herranz, L., Liu, X., Wang, Y., van de Weijer, J., & Raducanu, B. (2018). Memory replay GANs: learning to generate images from new categories without forgetting. In *NeurIPS*.

Zhai, M., Chen, L., Tung, F., He, J., Nawhal, M., & Mori, G. (2019). Lifelong GAN: Continual learning for conditional image generation. In *Proceedings of the IEEE ICCV* (pp. 2759–2768).

Zhao, B., & Mac Aodha, O. (2023). Incremental generalized category discovery. In *Proceedings of the IEEE ICCV* (pp. 19137–19147).

Zhao, X., Ou, Y., Kaplan, L., Chen, F., & Cho, J. H. (2019). Quantifying classification uncertainty using regularized evidential neural networks. In *AAAI FSS*.

Zhao, B., Xiao, X., Gan, G., Zhang, B., & Xia, S. T. (2020). Maintaining discrimination and fairness in class incremental learning. In *Proceedings of the IEEE CVPR* (pp. 13208–13217).

Zheng, B., Zhou, D. W., Ye, H. J., & Zhan, D. C. (2024). Multi-layer rehearsal feature augmentation for class-incremental learning. In *ICML, vol. 235* (pp. 61649–61663).

Zhu, F., Cheng, Z., Zhang, X. Y., Liu, C. L., & Zhang, Z. (2024). RCL: Reliable continual learning for unified failure detection. In *Proceedings of the IEEE CVPR* (pp. 12140–12150).

Zhu, X., Yi, J., & Zhang, L. (2024). Continual learning with unknown task boundary. *IEEE Transactions on Neural Networks Learning System*, 1–13. http://dx.doi.org/10.1109/TNNLS.2024.3412934.

Zhuang, C., Huang, S., Cheng, G., & Ning, J. (2022). Multi-criteria selection of rehearsal samples for continual learning. *Pattern Recognition*, *132*, Article 108907.