CHAMELEON SAMPLING: DIVERSE AND PURE EXAMPLE SELECTION FOR ONLINE CONTINUAL LEARNING WITH NOISY LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Learning under continuously changing data distribution with noisy labels is a desirable real-world problem yet challenging. Although many proposals have addressed each problem of continual learning or noisy label separately, tackling a combined challenge is underexplored. To address the task of *online continual learning with noisy labels*, we argue the importance of both diversity and purity of examples in the episodic memory of continual learning models. To balance diversity and purity in the episodic memory, we propose to combine a novel memory management strategy and robust learning. Specifically, we first propose to balance the trade-off between diversity and purity in the episodic memory with noisy labels. We then refurbish or apply unsupervised learning by splitting noisy examples into multiple groups using the Gaussian mixture model for addressing label noise. We empirically validate our approach on four real-world or synthetic benchmark datasets, including CIFAR10 and 100, Clothing1M, and mini-WebVision. Our method significantly outperforms prior arts in this challenging task set-up.

1 INTRODUCTION

Continual learning (CL) addresses learning methodologies on a continuous and online stream of annotated data. It is regarded as a realistic and practical learning set-up (Lopez-Paz & Ranzato, 2017; Prabhu et al., 2020; Bang et al., 2021). The CL methods are known to suffer from continuously changing data distribution and stability-plasticity dilemma due to catastrophic interference (Goodfellow et al., 2013; McCloskey & Cohen, 1989). In addition, when the CL model is deployed in real-world such as e-commerce applications, the annotated labels are often unreliable due to less controlled data curation process, *e.g.*, crowd-sourcing (Xu et al., 2021). Although these two issues occur simultaneously, they have been studied separately in each literature (Kirkpatrick et al., 2017; Prabhu et al., 2020; Li et al., 2019; Song et al., 2019). In the CL literature, many studies have used a small amount of data or model knowledge from the previous tasks, such as momentum matching (Lee et al., 2017), regularization on parameters (Kirkpatrick et al., 2017; Chaudhry et al., 2018), and sampling-based memory management (Prabhu et al., 2020; Bang et al., 2021). The learning with noisy labels literature includes sample selection methods (Jiang et al., 2018; Yu et al., 2019; Han et al., 2018), re-labelling (Song et al., 2019; Reed et al., 2015), and semi-supervised learning approach (Li et al., 2019; Nguyen et al., 2019; Zhou et al., 2020).

Motivated by the significance to address both challenges in a single framework, we design a new task, *online continual learning with noisy labels*, to expedite AI deployment in more realistic and practical scenarios. In recent literature for continual learning (Kirkpatrick et al., 2017; Prabhu et al., 2020; Bang et al., 2021), arguably the sample selection criterion is of importance for episodic memory-based CL methods, specifically, diversity-aware sample selection policy is shown to be effective in online learning scenarios. However, under noisy label data stream, diversity-based memory construction turns out to contain many examples with corrupted labels, leading to poor generalization of the model owing to the overfitting to noisy labels. The noisy data stream makes the methods less applicable to the real world scenario.

To address the issue in a single framework, we propose a memory sampling policy with robust learning on the sampled examples. We call our method Chameleon Sampling (ChamS) and illustrate

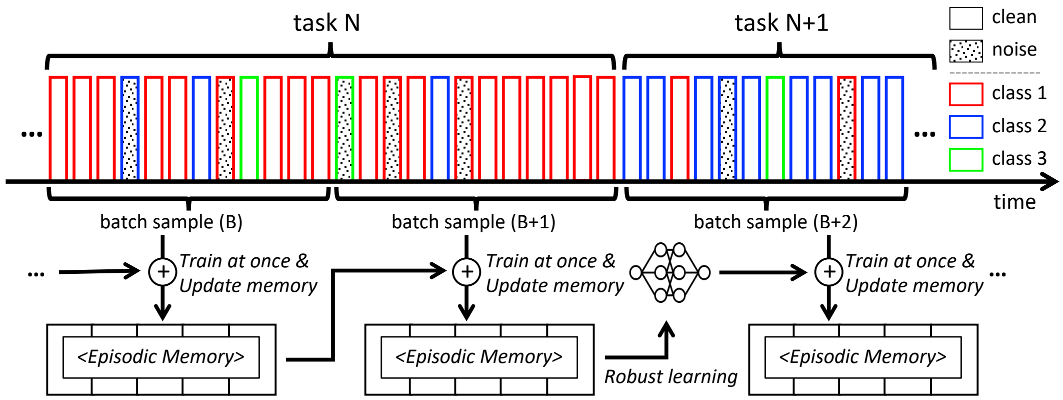


Figure 1: Online blurry continual learning with noisy labels setup and overview of proposed Chameleon Sampling (ChamS). The task share classes similar to blurry setup (Prabhu et al., 2020), examples might be falsely labeled, and the class distribution changes over tasks. The proposed method of updating memory balances between diversity and purity, so it avoids both the disturbance from noisy labels by low purity and overfitting by training with similar data. We further employ a novel robust learning method with a memory so that the model cannot be interfered by noisy labels in the memory, and the robust and well-trained model helps select examples at the next task.

the model overview in Figure 1. Specifically, ChamS constructs a robust episodic memory that preserves a set of training examples that are *diverse* and *pure*:

1. **Diversity:** In continual learning, deep neural networks (DNNs) are typically trained only with a small amount of training data in the episodic memory. Thus, the memory should maintain examples with high diversity. Otherwise, the DNNs easily overfit to similar examples and then lead to performance degradation (Sener & Savarese, 2018; Bang et al., 2021).
2. **Purity:** DNNs have extremely high capacity to fit all the examples with corrupted labels, leading to poor generalization on unseen data (Zhang et al., 2021). Thus, maintaining clean examples in the memory is essential for robust continual learning.

However, these two factors are in *trade-off* relationship in general; clean examples mostly exhibit smaller losses than noise examples due to the memorization effect of DNNs (Arpit et al., 2017; Song et al., 2021). In other words, emphasizing purity in memory sampling does not promote to include abundant difficult examples with large-loss, while emphasizing diversity does not include a sufficient number of clean (small-loss) examples. Thus, it is inevitable to contain a certain number of noise examples in the memory for the best trade-off between diversity and purity, which makes our setup non-trivial; the diversity measured on noisy examples would be misleading.

To alleviate the issue, we incorporate a robust learning scheme in our framework to selectively train the examples in the memory. Specifically, we identify mislabeled examples incorrectly included in the memory and splits them into two subsets; one to be re-labeled with high confidence, the other to be used for unsupervised learning for its low confidence. By the unified learning framework of diversity-aware sampling and robust learning with the memory by the hybrid learning of re-labeling and unsupervised learning, we construct the memory with high diversity and purity and complement the potential risk of memory sampling caused by the trade-off between diversity and purity.

Our contributions are summarized, as follows:

- Proposing the first noisy online continual learning set-up, which is realistic and challenging.
- Proposing a unified framework of both memory update considering both diversity and purity and handling noisy labels in the memory.
- Proposing an adaptive annealing scheme for a coefficient to balance with diversity and purity when updating memory for various noise ratio and complexity of data distribution.

2 Related Work

2.1 Task-free Blurry online Continual Learning

There have been many studies on task-free online continual learning, which are divided into two categories; *1) rehearsal-based approaches* (Bang et al., 2021; Prabhu et al., 2020; Riemer et al., 2018), where episodic memory stores a few examples of old tasks, *2) regularization-based approaches* (Kirkpatrick et al., 2017; Chaudhry et al., 2018), which exploit the information of old tasks using the parameters of models instead of the episodic memeory. Since rehearsal-based approaches generally have better performance, we focus on improving memory update methods.

Blurry Setup. Recently, some works (Bang et al., 2021; Prabhu et al., 2020; Aljundi et al., 2019) have tackled that the disjoint setup is an unsubstantial setup, thus conducting experiments on a blurry online continual learning setup. Unlike the disjoint setup where each task has different classes, each task shares the same classes with different class distribution in the blurry setup. For making a more realistic continual learning setup, we introduce blurry online continual learning with noisy labels.

Episodic Memory Updates. Although there have been many strategies for data sampling, most of them are not suitable for the online continual learning setup. Riemer et al. (2018) applied *reservoir sampling* for updating memory in online continual learning, and it can approximately optimize over the stationary distribution of all seen examples. Prabhu et al. (2020) proposed *greedy balancing sampler* such that it focuses on balancing the number of examples for each class. In contrast, Bang et al. (2021) proposed *rainbow memory* to select diversified examples with constant interval of uncertainty. However, this family of methods overlooks the problem of noisy labels. Therefore, this calls for a new memory sampling approach for robust continual learning with noisy labels.

2.2 Learning with Noisy Labels

Numerous approaches have been introduced to prevent DNNs from overfitting to noisy labels. They are categorized into three groups along with our proposed ChamS.

Re-labeling. Reed et al. (2015) developed a more coherent network that improves its ability to evaluate the consistency of noisy labels with label confidence via cross-validation. Song et al. (2019) regarded the examples with consistent label predictions as refurbishable ones, considering that the consistent predictions coincide with true labels with a high probability. Chen et al. (2021) averaged the model's softmax output on each example over the whole training, and then re-trained the model using the averaged soft labels. However, they still have a risk of overfitting to false labels, especially when the number of classes or noise rate is large due to the false correction.

Sample Selection. Jiang et al. (2018) proposed to guide the training of a student network with a pre-trained mentor network in a collaborative manner. Following the general pipeline of sample selection, the mentor network provides the student network with small-loss examples whose labels are likely to be correct. Similarly, Han et al. (2018) and Yu et al. (2019) utilized two models, but each model selects a certain number of small-loss examples and feeds them to its peer model for further training. Furthermore, Yu et al. (2019) proposed to use a disagreement strategy in conjunction with the Han et al. (2018)'s work.

Semi-Supervised Learning. Nguyen et al. (2019) combined supervised learning with sample selection such that a semi-supervised learning approach progressively filters out false labeled examples from noisy data; the examples with incorrect labels are progressively removed by using self-ensemble predictions. Li et al. (2019) fit two-component and one-dimensional Gaussian mixture models (GMMs) to the loss of training examples for sample selection, and then applied a semi-supervised learning technique called MixMatch (Berthelot et al., 2019). Meanwhile, Zhou et al. (2020) employed a two-phase learning strategy, which alternates between supervised training on selected clean examples and semi-supervised learning on re-labeled noisy examples.

2.3 Robust Continual Learning with Noisy Labels

Online learning and robust learning with noisy labels have been studied in difference research communities. Hence, a straightforward approach for continual learning with noisy labels is combining a memory sampling method for online learning with a robust learning method for handling noisy labels. However, this naive integration is difficult to resolve the challenge because of the trade-off issue between example diversity and purity. In this paper, we tackle the trade-off relationship by constructing the episodic memory with high diversity and purity, and mitigate the potential risk of memory sampling by using a complementary robust learning approach.

3 CHAMELEON SAMPLING (CHAMS)

We propose a novel problem setup of continual learning with noisy labels in Section 3.1. Then, we describe the two main components of ChamS: (1) the construction of robust episodic memory in Section 3.2 and (2) robust learning with the episodic memory in Section 3.3. Overall algorithm on ChamS is described in Appendix A.

3.1 PRELIMINARIES: BLURRY ONLINE CONTINUAL LEARNING WITH NOISY LABELS

In the blurry continual learning setup, the class distributions change according to the data stream, *i.e.*, tasks. That is, the major class of a certain task is likely to be a minor class of other tasks and, as such, most of the tasks have imbalanced class distribution between them.

Let C and T be the number of classes and tasks, respectively. Then, at each task t, the set of classes are split into a set of *major classes* M_t and a set of *minor classes* m_t. Following the most recent literature (Bang et al., 2021) for the blurry setup, each class should be a major class only in a certain task and be a minor class in other tasks, thus satisfying the following conditions:

$$\cup_{t=1}^{T} M_t = \{1, ..., C\} \text{ and } \cap_{t=1}^{T} M_t = \varnothing \text{ and } M_t \cup m_t = \{1, ..., C\}. \tag{1}$$

Next, the training data $\mathcal{D}_c$ for a class c is splitted into multiple subsets assigned for different tasks provided as the input stream to the model, $\mathcal{D}_c = \{D_{(c,t)} : t \in \{1, ..., T\}\}$ such that $\cap_{t=1}^{T} D_{(c,i)} = \varnothing$. Here, if the class c is one of the major classes for the task t (*i.e.*, $c \in M_t$), then $|D_{(c,t)}| \,/\, |\mathcal{D}_c| = L$, where L is the specified ratio of data for a major class in a task t. Therefore, the stream dataset $\mathcal{S}_t$ at a task t is defined as $\mathcal{S}_t = \cup_{i=1}^{C} D_{(i,t)}$. Furthermore, in the presence of noisy labels, training data $\mathcal{S}_t$ for a task t involves many mislabeled examples, thus $\mathcal{S}_t = \{(x, \tilde{y})\}_{i=1}^{|\mathcal{S}_t|}$ where x is an example and $\tilde{y}$ is a noisy label which may be corrupted from the ground-truth label y.

3.2 ROBUST EPISODIC MEMORY CONSTRUCTION

For the episodic memory, selecting diversified examples is arguably one of the crucial steps to improve model performances (Bang et al., 2021). However, in our label noise setup, this approach rather expedites overfitting to noisy labels because many false labeled examples are misclassified as the ones with high diversity. Contrary, the purity-based approach based on the small-loss trick only selects the examples with low diversity. To handle this trade-off relationship between purity and diversity, we introduce a score function that considers both aspects of the training example. Specifically, the proposed score function is the combination of two factors: *(i)* the training loss of an example to measure the purity based on the fact that small-loss examples are prone to be the clean ones, and *(ii)* the representation similarity to measure the diversity, where only the relevant representation (in Definition 1) participates in calculation.

Definition 1 *Let $f(x) \in \mathbb{R}^n$ be the representation of an example x obtained by a classifier f. Then, a **relevant representation** $f_{rel}(x; j)$ for a class j is defined as a subset of elements $e \in f(x)$ such that $f_{rel}(x; j) = \{e \in f(x) \mid w_{(e,j)} > 1/|C| \sum_{c=1}^{C} w_{(e,c)}\}$, where $w_{(e,c)}$ is the weight parameter of the classification layer associated with the element e and class j.*

Thus, the score function for an example x with its noisy label $\tilde{y}$ is finally formulated by

$$\text{Score}(x, \tilde{y}) = (1 - \alpha)\, \overbrace{l(x, y)}^{\text{purity}} + \alpha\, \overbrace{\big(1/|\mathcal{M}[\tilde{y}]|\big) \sum_{\hat{x} \in \mathcal{M}[\tilde{y}]} \text{Cosine}\big(f_{rel}(x; \tilde{y}), f_{rel}(\hat{x}; \tilde{y})\big)}^{\text{diversity}}, \tag{2}$$

where $\mathcal{M}$ is the episodic memory at the current moment, and $\mathcal{M}[\tilde{y}]$ returns all the examples annotated with $\tilde{y}$ in the memory. In addition, we use the cosine similarity to measure the similarity of two representations because of its symmetricity; and α is the balancing coefficient of diversity and purity. Thus, we can achieve the best trade-off between the two aspects by controlling the α value.

Adaptive Balancing Coefficient. It turns out that the optimal balancing coefficient α could be varying depending on the ratio of label noise or the type of datasets (please see Section 4.3 for more details). In this sense, we present a data- nad noise-agnostic approach that automatically adjusts the coefficient value during training. By the fact that easy or clean examples are favored in the early stage of training with DNNs, we conclude that diversified examples are more important at the later stage of training for better generalization; this learning tendency has also been witenessed in curriculum learning (Graves et al., 2017). Thus, we design an adaptive rule, which judges the best balancing coefficient based on the training loss of examples in the current batch, being formulated as $\alpha = (1/2)\min(1/l(\mathbf{x},\mathbf{y}), 1)$, where $l(\mathbf{x},\mathbf{y})$ is the average loss for all examples in the current batch. Accordingly, the coefficient value increases gradually to 1 since the training loss is monotonically decreasing via optimization. Therefore, ChamS gives a higher weight to the purity at the early stage, while to the diversity at the late stage.

3.3 Robust Learning with Episodic Memory

Even with the best trade-off between diversity and purity, the episodic memory may include noisy examples; mostly, they are the examples with high diversity, contributing to better generalization of the model. Hence, we present a robust hybrid learning of re-labeling and unsupervised learning. Note that the unification of robust learning on top of diversity-aware memory sampling allows to fully exploit all the examples in the episodic memory.

As a first step, we divide the episodic memory into clean and noisy subsets based on the small-loss trick (Li et al., 2019; Arazo et al., 2019). Thus, we fit GMMs to the training loss of all examples in the memory such that it estimates the probability of an example being clean. Given a noisy example x with its noisy label $\tilde{y}$, its clean probability is obtained through the posterior probability of GMMs,

$$\begin{aligned} &p_{gmm}\big(g|\ell(x,\tilde{y};\Theta)\big) = p(g)p\big(\ell(x,\tilde{y};\Theta)|g\big)/p\big(\ell(x,\tilde{y};\Theta)\big), \\ &\text{where } \ \ell(x,\tilde{y};\Theta) = -\sum_{c=1}^{C}\mathbb{1}(c == \tilde{y})\log\big(p_{model}(x,c;\Theta)\big), \end{aligned} \tag{3}$$

where g denotes the Gaussian component for the small-loss examples; and $p_{model}(x,c)$ is the softmax output of the model for the class c. Then, the clean set $\mathcal{C}$ and noisy set $\mathcal{N}$ is obtained by

$$\mathcal{C} := \big\{(x,\tilde{y}) \in \mathcal{M} : p_{gmm}\big(g|\ell(x,\tilde{y};\Theta)\big) \geq 0.5\big\} \text{ and } \mathcal{N} := \big\{(x,\tilde{y}) \in \mathcal{M} : p_{gmm}\big(g|\ell(x,\tilde{y};\Theta)\big) < 0.5\big\}. \tag{4}$$

Typically, only the clean set is used to train the model in a supervised manner in modern sample selection approaches (Han et al., 2018; Yu et al., 2019; Song et al., 2021), while the noisy set is discarded to pursue robust learning. However, in continual learning, the full use of episodic memory is crucial as the model should be managed by only using a small number of examples in the memory.

In this regard, we split the noisy set into two subsets, one to be re-labeled with high confidence and the other to be used for unsupervised learning for its low confidence; the examples with low confidence for re-labeling rather expedite the model to overfit if they are incorrectly re-labeled due to the perceptual consistency of DNNs (Reed et al., 2015). As such, we use a re-labeling approach for the former set while applying an unsupervised training loss, called consistency regularization, for the latter set. To divide the noise set into the two sets, we introduce *predictive uncertainty* formulated by

$$\mathrm{U}(\mathrm{x}) = 1.0 - \max_{c}\big(p_{model}(x,c;\theta)\big) \ \text{ where } \ c \in \{1,2,\dots C\}. \tag{5}$$

According to our analysis (see Figure 4(b)), the uncertainty distribution of the two set follows *bi-modal* distributions similar to the loss distribution of clean and noise examples. Therefore, the GMMs is also fitted to the uncertainty of the (expected) noisy examples in the noise set $\mathcal{N}$. Similar to Eq. (4), the noise set is split into the re-labeling set $\mathcal{R}$ and the unlabeled set $\mathcal{U}$,

$$\mathcal{R} := \big\{(x,\tilde{y}) \in \mathcal{N} : p_{gmm}\big(u|\mathrm{U}(x)\big) \geq 0.5\big\} \ \text{ and } \ \mathcal{U} := \big\{(x,\tilde{y}) \in \mathcal{N} : p_{gmm}\big(u|\mathrm{U}(x)\big) < 0.5\big\}, \tag{6}$$

where u is the guassian component for the low-uncertainty examples. Because the examples with high uncertainty are difficult to re-label, they are treated as the unlabeled set without relying on their possibly unreliable labels.

Re-labeling with $\mathcal{R}$. The model's prediction magnifies the useful underlying information in training data before overfitting. Accordingly, we refurbish the example with low uncertainty (*i.e.*, high confidence) by mixing its given noisy label $\tilde{y}$ with the model's prediction $p_{model}(x)$ according to the model's confidence to prediction, which is the posterior probability of the mixture component u,

$$\hat{y} = p_{gmm}(u|\mathrm{U}(x)) \cdot p_{model}(x) + \big(1.0 - p_{gmm}(u|\mathrm{U}(x))\big) \cdot \tilde{y}. \tag{7}$$

This is a soft re-labeling approach to progressively refine the given noisy label based on the model's prediction evolving during training.

Unsupervised Learning with $\mathcal{U}$. We exploit even the unlabeled set $\mathcal{U}$ in an unsupervised manner. We adopt consistency regularization, which is widely used in the unsupervised and semi-supervised learning literature (Berthelot et al., 2019; Tarvainen & Valpola, 2017). This regularization effectively helps learn useful knowledge from the examples without knowing their ground-truth labels by penalizing the prediction difference between the two examples augmented from the same one,

$$\ell_{reg}(x) = \frac{1}{|\mathcal{U}|} \sum_{x \in \mathcal{U}} \big|\big| p_{model}\big(s(x);\Theta\big) - p_{model}\big(w(x);\Theta\big) \big|\big|_2, \tag{8}$$

where $s(\cdot)$ and $w(\cdot)$ are strong and weak augmentation functions for an example x. We use AutoAugment (Cubuk et al., 2019) and random horizontal flipping as the two augmentors, respectively.

Overall, ChamS trains the model robustly with the three meticulously designed subsets, namely a clean set $\mathcal{C}$, a re-labeled set $\mathcal{R}$, an unlabeled set $\mathcal{U}$, and $\mathcal{M} = \mathcal{C} \cup \mathcal{R} \cup \mathcal{U}$.

4 EXPERIMENTS

4.1 EXPERIMENTAL SET-UP

Datasets and Noise Injection. To validate our method, we perform an image classification task on four benchmark datasets: CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), a subset of 80 million categorical images; Clothing1M (Xiao et al., 2015), a real-world noisy data of large-scale crawled clothing images; WebVision (Li et al., 2017), a subset of real-world noisy data consisting of large-scale web images. As the data labels in CIFARs are almost clean, we inject two widely used synthetic noise in the literature, namely symmetric (SYM) and asymmetric (ASYM) label noise (Han et al., 2018). Symmetric noise flips the ground-truth label into other possible labels with equal probability, while asymmetric noise flips the ground-truth one into a certain label with high probability. For a thorough evaluation, we adjust the ratio of label noise from 20% to 60%. We do not inject any label noise into Clothing1M and WebVision because they contain real label noise whose ratio is estimated at 38.5% and 20.0%, respectively (Song et al., 2020). The detailed experimental configuration for online continual learning with noisy labels is provided in Appendix B.

Baselines. We compare ChamS with the combination of representative memory sampling methods for continual learning and robust learning methods for handling noisy labels. As summarized in Section 2, there are two representative memory sampling approaches, *reservoir sampling* (RSV) (Riemer et al., 2018) and *greedy balancing sampler* (GBS) (Prabhu et al., 2020). As for the robust learning approaches, there are three representative approaches, a re-labeling approach *SELFIE* (Song et al., 2019), a sample selection approach *Co-teaching* (Han et al., 2018), and a semi-supervised learning approach *DivideMix* (Li et al., 2019). Thus, a total of eight combinations are compared with ChamS; two only using the memory sampling approaches and six combining them with the three robust learning approaches.

Metrics. We report the *last* test (or validation) accuracy, which is the most widely used metric in continual learning and robust learning (Chaudhry et al., 2018; Han et al., 2018), where "last" refers to the value measured after all tasks are completed. For robustness in continual learning, the two factors, purity and diversity, for the episodic memory $\mathcal{M}$ are another important metrics. Hence, we introduce the two metrics of measuring its purity and diversity,

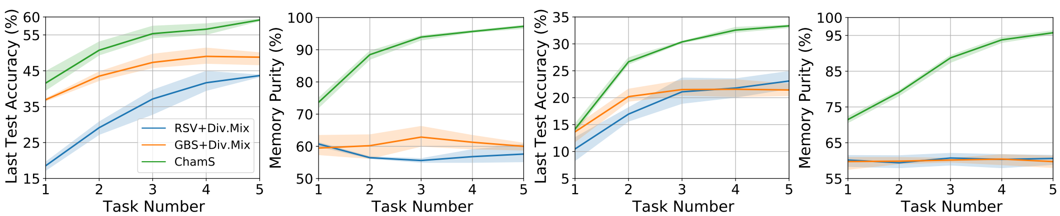
$$\text{Purity} = \frac{\sum_{(x,\tilde{y}) \in \mathcal{M}} \mathbb{1}(\tilde{y} == y)}{|\mathcal{M}|}, \;\; \text{Diversity} = \frac{1}{C} \sum_{c=1}^{C} \frac{2}{|\mathcal{M}_c|(|\mathcal{M}_c|-1)} \sum_{i \in \mathcal{M}_c} \sum_{j \in \mathcal{M}_c, i \neq j} ||f(x_i) - f(x_j)||_2, \tag{9}$$

where $f(x)$ is the representation of an example x by a fully trained model, and $\mathcal{M}_c$ is a subset of $\mathcal{M}$ consisting of training examples whose ground-truth label is c; and C is the number of classes. Purity measures how many clean examples are correctly included in the memory, while Diversity measures how the examples in the memory are spread out in the representation space per class.

Implementation Details. We use ResNet (He et al., 2016) as a classifier for all compared algorithms; ResNet18, ResNet32, ResNet34 and ResNet34 are used for CIFAR-10, CIFAR-100, Clothing1M, and WebVision, respectively. For CIFARs, we train ResNet using an initial learning rate

Table 1: Comparison of the last test accuracy over benchmarks on CIFAR-10 and CIFAR-100 with SYM-$\{20\%, 40\%, 60\%\}$ and ASYM-$\{20\%. 40\%\}$. K is the size of episodic memory.

Methods	CIFAR-10 (K=500)					CIFAR-100 (K=2,000)				
	Sym.			Asym.		Sym.			Asym.	
	20	40	60	20	40	20	40	60	20	40
RSV (Riemer et al., 2018)	54.53	39.24	28.65	53.58	39.99	29.37	19.33	10.48	29.79	20.26
+ SELFIE (Song et al., 2019)	54.54	39.17	28.80	51.77	40.41	28.93	19.45	10.48	29.76	20.05
+ Co-teaching (Han et al., 2018)	56.12	39.84	30.53	53.47	38.68	30.40	22.01	13.29	30.84	20.33
+ DivideMix (Li et al., 2019)	56.05	43.66	35.09	56.13	42.27	30.50	23.10	13.14	30.18	20.11
GBS (Prabhu et al., 2020)	54.82	41.79	27.30	54.19	40.38	27.96	18.82	10.16	29.48	20.25
+ SELFIE (Song et al., 2019)	55.43	41.56	27.76	51.30	40.68	27.95	18.43	9.85	29.56	19.82
+ Co-teaching (Han et al., 2018)	55.09	42.72	31.37	53.99	39.52	29.51	20.89	12.83	28.99	20.47
+ DivideMix (Li et al., 2019)	57.79	48.79	34.33	57.40	44.57	29.57	21.43	13.16	28.73	19.62
ChamS (ours)	**61.33**	**59.17**	**52.43**	**61.56**	**47.10**	**35.63**	**33.35**	**28.78**	**34.56**	**25.66**



(a) CIFAR-10 (b) CIFAR-100

Figure 2: Last test accuracy and memory purity changes as the task number increases on CIFAR-10/100 with SYM-40%. The figures with various noise ratio are described in Appendix C.

of 0.05 with cosine annealing with a batch size of 16 and the training epochs of 256 following the literature (Bang et al., 2021). For Clothing1M and WebVision, all the training hyper-parameters are the same as for CIFARs except for the training epochs of 128. In all experiments, CutMix (Yun et al., 2019) and AutoAugment (Cubuk et al., 2019) are used because they signifcantly improves the performance in continual learning (Bang et al., 2021). For reliability, we repeat all the experiments trice with a single NVIDIA P40 GPU and report the average performance.

4.2 Robustness Comparison

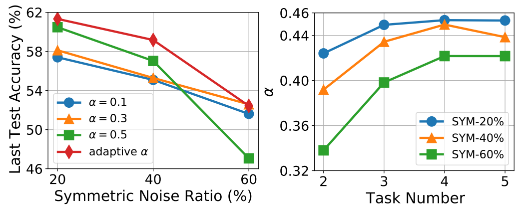
Highlights. Table 1 compares ChamS with eight carefully combined methods on the two CIFARs datasets with varying symmetric and asymmetric noise. Overall, ChamS consistently outperforms all other compared methods. In particular, the performance gain becomes large as the noise ratio increases or the difficulty of training data increases from CIFAR-10 to CIFAR-100. Quantitatively, the test accuracy significantly increases with ChamS by 4%–15% compared with other methods.

Details. SELFIE does not provide any performance gain even when combined with the two memory sampling approaches. Since SELFIE requires an abundant number of training data for re-labeling, it rather produces many false corrections in the online learning setup, where the episodic memory contains a small number of training examples. In addition, despite the use of multiple networks, the performance improvement by Co-teaching is negligible compared to the performance improvement ChamS achieves. Unlike SELFIE and Co-teaching, DivideMix exhibits considerable performance improvement in CIFAR-10, but it is observed to perform poorly in CIFAR-100 with asymmetric noise, which is a more realistic label noise scenario.

Purity of Episodic Memory. The performance dominance of ChamS is attributed to the high purity of the episodic memory. Figure 2 shows the convergence curves of the last test accuracy and the memory purity across the tasks; ChamS is compared with the two variants of DivideMix respectively combined with RSV and GBS, showing the best performance among baselines. Unlike the two baselines, ChamS exhibits significantly higher test accuracy and memory purity for every task. Particularly, the performance gap between ChamS and other methods tends to gradually increase as the task number increases due to the synergistic effect of alternating robust episodic memory construction (Section 3.2) and robust learning with the memory (Section 3.3). Moreover, the performance

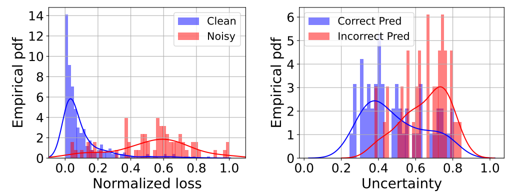
Table 2: Comparison of the last test accuracy over various static α on CIFAR-10 and CIFAR-100 with SYM-$\{20\%, 40\%, 60\%\}$ and ASYM-$\{20\%, 40\%\}$. K is the size of episodic memory.

	CIFAR-10 (K=500)					CIFAR-100 (K=2,000)				
	Sym.			Asym.		Sym.			Asym.	
α	20	40	60	20	40	20	40	60	20	40
0.1	57.40	55.09	51.61	58.91	47.42	36.07	33.59	**28.10**	33.95	**25.60**
0.3	58.12	55.29	**52.59**	58.40	**48.95**	35.52	**33.60**	27.31	34.61	25.47
0.5	**60.48**	**57.03**	47.06	**60.36**	46.67	**37.01**	32.31	23.00	**36.61**	25.51
0.6	29.50	30.33	24.44	36.04	33.67	25.33	18.89	13.51	26.66	18.16



(a) (b)

Figure 3: Accuracy by different α on CIFAR-10 with SYM-{20%,40%,60%}; (a) Comparison of the last test accuracy w. and w.o. the adaptive strategy, (b) Comparison of the mean of α for each task with the adaptive strategy.



(a) (b)

Figure 4: Distribution of examples in the episodic memory training on CIFAR-10 with SYM-20%. GMMs effectively distinguish between (a) clean and noisy labels by loss, and (b) correct and incorrect predictions by uncertainty.

gain of ChamS is consistent even with the increase in data difficulty. That is, the purity of other baselines gets drastically worse in CIFAR-100, while ChamS maintains its significant dominance. These gains are still valid on various noise ratio, and details are described in Appendix C.

4.3 Ablation Study on Adaptive Balancing Coefficient

The balancing coefficient α determines the contribution of purity and diversity for memory sampling, as described in Section 3.2. In this section, we provide an ablation study on the balancing term by contrasting the performance change according to different α values, as summarized in Table 2. It turns out that the test accuracy decreases drastically if $\alpha > 0.5$ because diversity is dominantly considered in memory sampling rather than memory purity. In the presence of label noise, this suggests that memory purity must be considered with high priority, as opposed to the conventional continual learning that mainly emphasizes example diversity. In addition, it is noteworthy that the best α value differs depending on the noise ratio and datasets. In general, with the increase in noise ratios, a smaller α value is prone to achieve higher test accuracy because memory purity is far more important than memory diversity when label noise is heavy. The changes of diversity and purity by α are described in Appendix D.

Efficacy of Adaptive Strategy. Figure 3a compares the performance of ChamS with and without using the proposed adaptive balancing coefficient strategy. ChamS with the adaptive strategy always shows higher test accuracy. In addition, Figure 3b shows the α values authomatically determined by ChamS with the adaptive strategy for three difference noise scenarios. The α tends to increase as the task number increases except for the initial task. In other words, small-loss (clean) examples are favored at the earlier stage of training for robust learning, while diverse examples are favored at the later stage of training for better generalization. Therefore, this curriculum learning scheme allows ChamS to exploit the best α value for the diverse scenario of learning with noisy labels.

4.4 Component Analysis of Robust Learning with ChamS.

We analyze the effect of each component in our proposed robust learning approach, which exploits the clean set $\mathcal{C}$ together with the re-labeled set $\mathcal{R}$ and the unlabeled set $\mathcal{U}$. Figure 4a shows the loss distribution of all training examples in the memory $\mathcal{M}$, splitting $\mathcal{M}$ into the clean set $\mathcal{C}$ and noise set $\mathcal{N}$, while Figure 4b shows the uncertainty distribution of the examples in the noise set $\mathcal{N}$, splitting

Table 3: Ablation study for ChamS on CIFAR-10 and CIFAR-100 with various noise ratios and types. SYM and ASYM refer to the symmetric and asymmetric label noise, respectively.

Robust Learning		CIFAR-10 (K=500)					CIFAR-100 (K=2,000)				
		SYM			ASYM		SYM			ASYM	
Re-label	Consistency	20	40	60	20	40	20	40	60	20	40
		61.77	55.40	46.88	60.60	46.37	33.11	26.84	18.59	31.52	21.33
✓		57.72	55.36	44.11	61.00	46.59	34.73	31.52	26.17	33.21	22.67
	✓	48.63	46.20	31.26	52.12	36.68	31.71	24.31	18.09	29.85	20.31
✓	✓	61.33	**59.17**	**52.43**	**61.56**	**47.10**	**35.63**	**33.35**	**28.78**	**34.56**	**25.66**

$\mathcal{N}$ into the re-labeled set $\mathcal{R}$ and the unlabeled set $\mathcal{U}$. First of all, the loss distributions in Figure 4a are bi-modal for clean and noisy examples, thus clearly separating them by fitting GMMs. Likewise, the uncertainty distributions in Figure 4b are bi-model as well for the correctness of model's predictions. Hence, re-labeling using the model's prediction for the left Guassian component (*i.e.*, low-uncertainty examples) ensures that the proposed re-labeling approach performs with high precision.

Next, Table 3 summarizes the contribution to the additional use of the re-labeling and consistency regularization in ChamS. Using either re-labeling or consistency regularization for unsupervised learning does not bring out performance improvement in many cases; in most cases, each technique rather degrades the performance because entirely re-labeling the noisy set makes many false correction, while treating the entire noisy set as an unlabeled set for consistency regularization utilizes only a small number of examples in the clean set $\mathcal{C}$. However, an opposite trend is observed when using the two techniques altogether. That is, training the model for the clean and re-labeled set $\mathcal{C} \cup \mathcal{R}$ in conjunction with the consistency loss for the unlabeled set $\mathcal{U}$ makes great synergistic effect, improving the test accuracy by up to 10.2% compared to when neither technique is used.

4.5 Results with Real-world Noisy Data

We empirically validate the robustness of seven robust approaches including ChamS in the two real-world noisy datasets, WebVision and Clothing1M in Table 4. As the Clothing-1M dataset has the class imbalance problem, we report the macro-accuracy, which is the average of the last accuracy over classes. Similar to the results with synthetic noise in Table 1, ChamS maintains its dominance over the six other compared robust methods even in real label noise. Quantitatively, ChamS outperforms the other robust methods by 5.75% and 0.59% for WebVision and Clothing1M, respectively. It implies that the proposed ChamS significantly improve the robustness against label noise for online continual learning.

Table 4: Last validation cccuracy on WebVision and Clothing1M with real-world noise

Methods	WebVision (K=1000)	Clothing1M (K=700)
RSV+ SELFIE	19.08	28.69
RSV+ Co-teaching	16.45	26.27
RSV+ DivideMix	11.73	18.99
GBS + SELFIE	20.01	29.92
GBS + Co-teaching	17.77	28.83
GBS + DivideMix	13.98	22.48
ChamS (ours)	**25.76**	**30.51**

5 Conclusion

We address online and blurry continual learning with noisy labels, which can be occurred frequently in practical AI deployment in real-world. To handle this challenging setup, we propose a method of balancing diversity and purity in episodic memory management, followed by a complementary robust learning approach against noisy labels. Specifically, we define the score function that considers the training loss (purity) and features similarity (diversity) simultaneously. In addition, a data- and noise-agnostic approach is proposed to automatically adjust the balancing coefficient during training. Last, a robust learning method is presented to handle several false labeled examples possibly included in the memory. We verify that ChamS not only improves model performance but also helps select more informative examples when updating memory. ChamS outperforms other robust methods in multiple synthetic or real-world noisy datasets in the CL setting, but it is still possible to further improve the performance due to the large gap when compared with the state-of-the-art robust method under the non-CL setting. To reduce this performance gap, we will investigate a more powerful robust method to overcome the limitation of CL.

Ethics Statement

All continual learning (CL) methods including the proposed one would adapt and extend the already trained AI model to recognize more and better with the streamed data. The CL methods will expedite the deployment of AI systems to help humans by its versatility of adapting to a new environment out of the factory or research labs. As all CL methods, however, would suffer from adversarial streamed data as well as data bias, which may cause ethnic, gender or biased gender issues, the proposed method would not be an exception. But we expect that our method could contribute discriminate adversarial or ethically controversial data by regarding them as noise and it is a great future research avenue. Although the proposed method has *no intention* to allow such problematic cases, the method may be exposed to such threats. Relentless efforts should be made to develop mechanisms to prevent such usage cases in order to make the continuously updating machine learning models safer and enjoyable to be used by humans.

Reproducibility Statement

We describe the datasets and baselines compared in details in the paragraph titled ‘Datasets’ and ‘Baselines and Metrics’ in Sec. 4.1. The details of model architecture and implementation details including hyperparameter setups are presented in the paragraph titled ‘Implementation Details’ in Sec. 4.1. We take the reproducibility of the research very seriously and solemnly promise to release all codes, containers (*e.g.*, Docker) that includes running environments and learned models of pretraining and downstream tasks in a public repository.

References

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *NeurIPS*, 32:11816–11825, 2019.

Eric Arazo, Diego Ortego, Paul Albert, Noel E O’Connor, and Kevin McGuinness. Unsupervised label noise modeling and loss correction. In *ICML*, 2019.

Devansh Arpit, Stanisław Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *ICML*, pp. 233–242, 2017.

Jihwan Bang, Heesu Kim, YoungJoon Yoo, Jung-Woo Ha, and Jonghyun Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *CVPR*, pp. 8218–8227, 2021.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel. Mixmatch: A holistic approach to semi-supervised learning. *NeurIPS*, 32, 2019.

Arslan Chaudhry, Puneet K. Dokania, Thalaiyasingam Ajanthan, and Philip H. S. Torr. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *ECCV*, 2018.

Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *AAAI*, volume 35, pp. 11442–11450, 2021.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *CVPR*, pp. 113–123, 2019.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *international conference on machine learning*, pp. 1311–1320. PMLR, 2017.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W Tsang, and Masashi Sugiyama. Co-teaching: robust training of deep neural networks with extremely noisy labels. In *NeurIPS*, pp. 8536–8546, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, pp. 2304–2313. PMLR, 2018.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13):3521–3526, 2017.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Sang-Woo Lee, Jin-Hwa Kim, Jaehyun Jun, Jung-Woo Ha, and Byoung-Tak Zhang. Overcoming catastrophic forgetting by incremental moment matching. In *NeurIPS*, pp. 4655–4665, 2017.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *ICLR*, 2019.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*, 2017.

David Lopez-Paz and Marc'Aurelio Ranzato. Gradient episodic memory for continual learning. *NeurIPS*, 30:6467–6476, 2017.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Duc Tam Nguyen, Chaithanya Kumar Mummadi, Thi Phuong Nhung Ngo, Thi Hoai Phuong Nguyen, Laura Beggel, and Thomas Brox. Self: Learning to filter noisy labels with self-ensembling. In *ICLR*, 2019.

Ameya Prabhu, Philip HS Torr, and Puneet K Dokania. Gdumb: A simple approach that questions our progress in continual learning. In *ECCV*, pp. 524–540. Springer, 2020.

Scott E Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *ICLR (Workshop)*, 2015.

Matthew Riemer, Ignacio Cases, Robert Ajemian, Miao Liu, Irina Rish, Yuhai Tu, and Gerald Tesauro. Learning to learn without forgetting by maximizing transfer and minimizing interference. In *ICLR*, 2018.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *ICLR*, 2018.

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *ICML*, pp. 5907–5915. PMLR, 2019.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *arXiv preprint arXiv:2007.08199*, 2020.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Robust learning by self-transition for handling noisy labels. In *KDD*, 2021.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Neurips*, pp. 1195–1204, 2017.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *CVPR*, pp. 2691–2699, 2015.

Wenqiang Xu, Liangxiao Jiang, and Chaoqun Li. Improving data and model quality in crowdsourcing using cross-entropy-based noise correction. *Information Sciences*, 546:803–814, 2021.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor W Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In *ICML*, 2019.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pp. 6023–6032, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Tianyi Zhou, Shengjie Wang, and Jeff Bilmes. Robust curriculum learning: From clean label detection to noisy label self-correction. In *ICLR*, 2020.

A DETAILED ALGORITHM

Algorithm 1 Chameleon Sampling (ChamS)

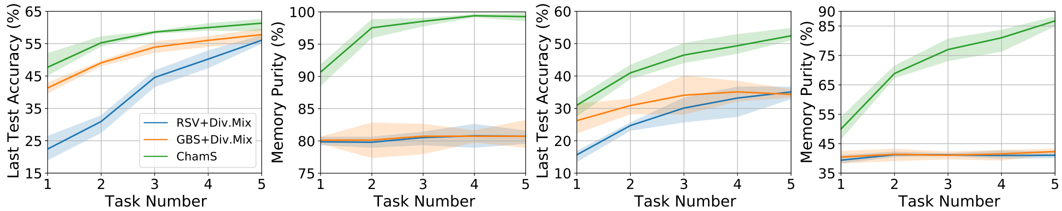
Input: $\mathcal{S}_t$ denotes stream data at task t, $\mathcal{M}$ denotes exemplars stored in a episodic memory.
for t in tasks **do**
  for $\mathbf{x}, \tilde{\mathbf{y}}$ in $\mathcal{S}_i$ **do**
    $l = CrossEntropyLoss(\mathbf{x}, \tilde{\mathbf{y}})$
    $\theta = SGD(l, \theta)$ with learning rate 0.001 ▷ Online train from stream data
    for $x, \tilde{y}$ in zip($\mathbf{x}, \tilde{\mathbf{y}}$) **do** ▷ Update memory
      calculate $\alpha = \frac{1}{2}\min(1/l, 1)$ ▷ Calculate adaptive α
      calculate score(x) by Equation 2
      sort $\mathcal{M}$ by score
      $\mathcal{M}$.pop()
    end for
  end for
  while $e <$ MaxEpoch **do** ▷ Robust train using memory
    $\mathcal{C}, \mathcal{N}$ = GMM($\mathcal{M}, \theta$) by loss
    $\mathcal{R}, \mathcal{U}$ = GMM($\mathcal{N}, \theta$) by uncertainty
    for i = 1 **to** num_iter **do**
      mini-batch $(\mathbf{x}, \tilde{\mathbf{y}}) = \{(x_i, \tilde{y_i}) | i = 1, 2, ..., B\}$ from $\mathcal{C}$
      mini-batch $(\mathbf{x}, \hat{\mathbf{y}}) = \{(x_i, \hat{y_i}) | i = 1, 2, ..., B\frac{|\mathcal{R}|}{|\mathcal{C}|}\}$ from $\mathcal{R}$
      mini-batch $(\mathbf{s}, \mathbf{w}) = \{(s(x_i), w(x_i)) | i = 1, 2, ...B\frac{|\mathcal{U}|}{|\mathcal{C}|}\}$ from $\mathcal{U}$
      $l_c = CrossEntropyLoss(\mathbf{x}, \tilde{\mathbf{y}})$
      $l_r = CrossEntropyLoss(\mathbf{x}, \hat{\mathbf{y}})$, where $\hat{\mathbf{y}}$ from Equation 7
      get l_{reg} from Equation 8 using $(\mathbf{s}, \mathbf{w})$
      $l_{tot} = l_c + l_r + l_{reg}$
      $\theta = SGD(l_{tot}, \theta)$ with cosine annealing scheduler
    end for
  end while
end for

B DETAILED CONFIGURATION FOR ONLINE CONTINUAL LEARNING

To split dataset into several tasks, we follow Bang et al. (2021) to make *blurry-CL*. *Blurry-CL* contains two types of classes, major and minor classes. We set $L = 0.9$, so the major classes are exclusive for each task, and assign 90% of the examples in the corresponding task. After assigning examples of major classes into all the tasks, the minor classes, which are not major classes in the current task, assign the rest of the examples into the tasks evenly. Since some labels might be incorrect, real class distribution of each task might be different. We configure CIFAR-10/100, clothing1M and WebVision as 5, 5, 7 and 10 tasks, respectively. Furthermore, we consider online-CL which the incoming samples are presented to a model only once except for the examples from the memory.

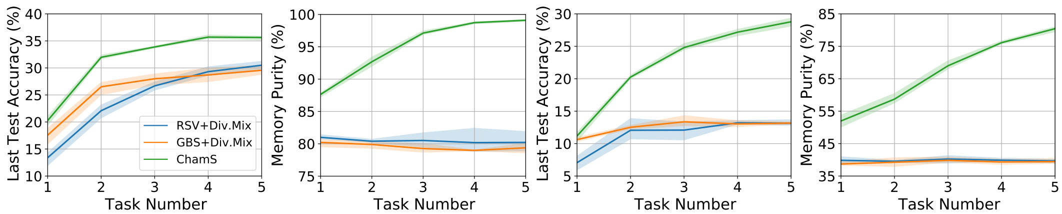
C ACCURACY AND MEMORY PURITY BY VARIOUS NOISE RATIOS

We add experimental results of the last accuracy and memory purity of ChamS on CIFAR-10/100 with the symmetric noise 20% and 60% in Figure 5 and 6. We can find consistent results as in Figure 2. On different scenarios, ChamS outperforms the other baselines for both accuracy and memory purity over the entire task stream. Especially, in memory purity, ChamS greatly outperforms the other baselines, and it shows that our method can effectively find clean labels from corrupted data for the online continual learning setup. Since we believe that memory sampling and robust learning from ChamS are the relationship to help each other, so the last accuracy and memory purity increases and have larger gap as the task number increases.



(a) CIFAR-10 with SYM-20% (b) CIFAR-10 with SYM-60%

Figure 5: Illustration of last accuracy and memory purity changes as the task number increases on CIFAR-10 with SYM-$\{20\%, 60\%\}$.

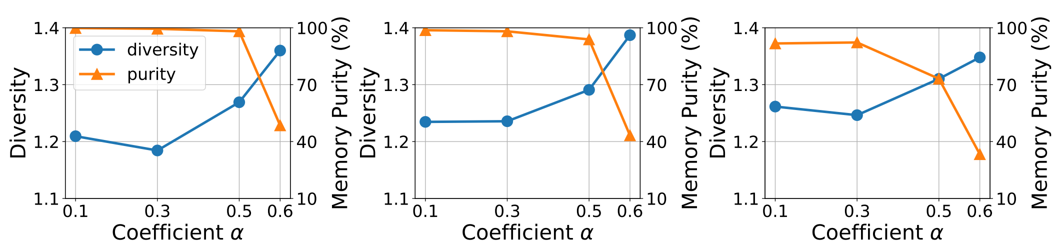


(a) CIFAR-100 with SYM-20% (b) CIFAR-100 with SYM-60%

Figure 6: Illustration of last accuracy and memory purity changes as the task number increases on CIFAR-100 with SYM-$\{20\%, 60\%\}$.

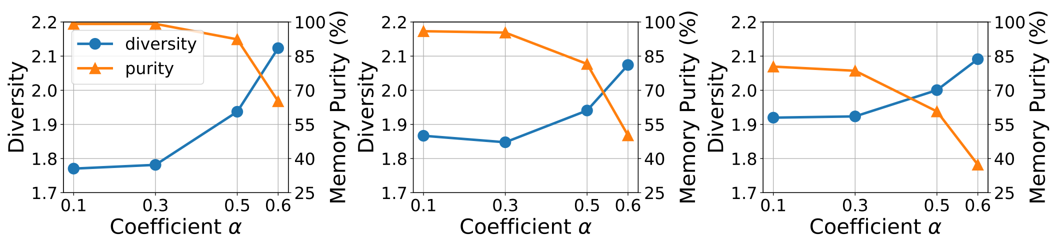
D DIVERSITY VS. PURITY

To analyze how diversity and purity in a memory change according to coefficient α, we plot the diversity and purity metrics in Figure 7 and 8 for various noise ratios. As hyper-parameter α increases, the diversity score increases while the purity score decreases in all figures. When α is increased beyond a certain value, it can be seen that purity is sacrificed for the sample diversity. If there are too many noisy labels in a memory, a trained model would inevitably learn the noisy labels, which can lead to performance degradation. Therefore, the best strategy is to set the α before the abrupt decrease in the purity score. For example, in Figure 7a and 7b, memory purity is over 95% until $\alpha = 0.5$, and then falls to less than 50% when α is set to 0.6. Thus, the best strategy for balancing the diversity and memory purity can be obtained at $\alpha = 0.5$. Meanwhile, in the case of Figure 7c, memory purity drops to 70% at $\alpha = 0.5$, so it can be expected that the best performance is obtained at $\alpha = 0.3$, before memory purity is too much degraded. This is because memory purity becomes more important as the ratio of noisy labels contained in memory increases. These results are consistent with the results in Table 2.



(a) (b) (c)

Figure 7: Illustration of diversity and purity in the memory as the coefficient α changes on CIFAR-10 with (a) SYM-20%, (b) SYM-40%, and (c) SYM-60%.



(a)

(b)

(c)

Figure 8: Illustration of diversity and purity in the memory as the coefficient α changes on CIFAR-100 with (a) SYM-20%, (b) SYM-40%, and (c) SYM-60%.