Bridging Inter-task Gap of Continual Self-supervised Learning with External Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent research on Self-Supervised Learning (SSL) has demonstrated its ability to extract high-quality representations from unlabeled samples. However, in continual learning scenarios where training data arrives sequentially, SSL's performance tends to deteriorate. This study focuses on Continual Contrastive Self-Supervised Learning (CCSSL) and highlights that the absence of contrastive learning on inter-task data, due to the unavailability of historical samples, leads to a significant drop in performance. To tackle this issue, we introduce a simple and effective method called BGE, which **B**ridges the inter-task **G**ap of CCSSL using **E**xternal data from publicly available datasets. BGE enables the contrastive learning of each task data with external data, allowing relationships between them to be passed along the tasks, thereby facilitating *implicit* inter-task data comparisons. To overcome the limitation of the external data selection and maintain its effectiveness, we further propose the One-Propose-One algorithm to collect more relevant and diverse high-quality samples from the chosen external data while filtering out distractions from the out-of-distribution data. Experiments show that BGE can generate better discriminative representation in CCSSL, especially for inter-task data, and improve classification results with various external data compositions. Additionally, the proposed method can be seamlessly integrated into existing continual learning methods yielding significant performance improvement.

1 Introduction

In recent years, deep neural networks [13, 22, 35] have achieved great success, but plenty of works are under the assumption that all data are available simultaneously for training. In practical scenarios, acquiring the entire dataset at once is often challenging due to data being constantly updated. In this case, training the network continually suffers from catastrophic forgetting [38], meaning that the network severely forgets old task knowledge after learning the new one. Hence, continual learning investigates methods to train networks incrementally while mitigating catastrophic forgetting.

Although continual learning has been widely studied and numerous effective methods [32, 36, 40] have been proposed, most existing research remains focused on supervised learning, with Continual Contrastive Self-Supervised Learning (CCSSL) receiving relatively little attention. However, studying CCSSL is equally significant.

To prevent catastrophic forgetting, prior CCSSL works CaSSLe [16], PFR [18], and POCON [19] use knowledge distillation, while CPPF [11] incorporates prototype clustering. In this paper, we highlight an important but generally overlooked issue in these works: *Comparisons of inter-task data are absent.* Specifically, a widely accepted opinion in continual learning is that if the sum of each task's loss is minimized, then continual learning's performance reaches its upper bound: *joint learning.* However, in CCSSL, even if each task's loss is minimized, there is still a gap between joint

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

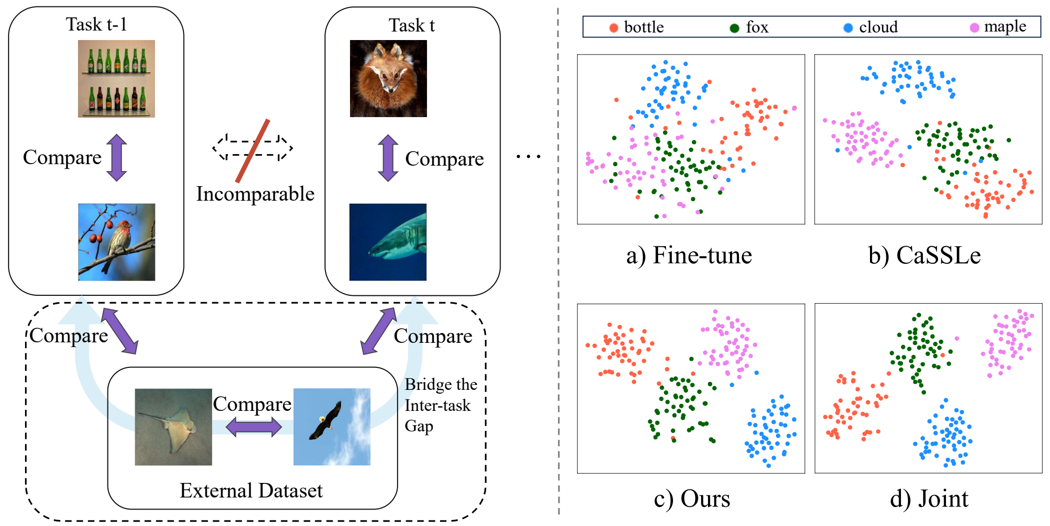

Figure 1: **Left:** Overview of our method BGE. In typical CCSSL methods, the inter-task data pairs are incomparable. We employ an external dataset to complement these missing comparisons, effectively bridging the inter-task gap. **Right:** t-SNE [47] visualization of four classes belonging to different tasks in continual learning. Compared to prior methods Fine-tune and CaSSLe [16], we make the inter-task data more separable.

learning. Because joint learning requires any sample pair in the entire dataset to participate in the contrastive loss computation. In contrast, in continual learning, inter-task data are unavailable to each other, meaning this aspect of the contrastive loss is never computed and optimized. This omission increases the likelihood of inter-task class confusion, as illustrated in Figure 1 Right, despite classes from four different tasks having distinctly different semantics, they still show confusion in prior methods Fine-tune and CaSSLe [16]. In contrast, our method and joint training consider inter-task comparisons and can better distinguish them.

Since we could not directly use data from other tasks for inter-task comparisons, we would like to compensate for these comparisons with the help of external data. Some prior works [31, 52, 56] have explored using external data for continual learning. GD [31] and ZSCL [56] use external data for distillation to stabilize the feature space, while requiring extensive external data and high computational costs. ST [52] employs external data as additional training data, but as a supervised method, it requires pseudo-labels, making it less robust to out-of-distribution (OOD) data. Tang et al. [45] enhance exemplar diversity with external data. Existing methods focus on using external data in supervised learning, but given that CCSSL does not require labels for training, we propose using external data in CCSSL, which avoids the need for pseudo-labels and is more generalizable and robust to OOD data. Besides, our motivation is to improve feature space by compensating for absent comparisons rather than merely stabilizing it, and it does not require extensive external data.

In summary, we propose incorporating publicly available external data into training to compensate for the absent inter-task comparisons, as shown in Figure 1 Left. When the external dataset is sufficiently large, it is reasonable to assume a high probability that some external data share similar features with the task data, even if they are in different classes. By incorporating these high-quality external data into CCSSL, the data from each task can be compared with them. enables the inter-data relationship to be passed along the tasks, thereby constructing implicit inter-task comparisons. Further, considering that external data in open-world scenarios may contain extensive OOD data that is not beneficial for task training, we propose the One-Propose-One (OPO) sampling algorithm, to sample high-quality external data that are relevant to tasks and sufficiently diverse without any hyperparameters.

Experiments demonstrate that BGE can be seamlessly integrated into existing methods, resulting in significant performance improvement. We also point out that although it may seem unsurprising that network performance improves with more training data, this improvement is not due to richer input features, because when we add equal external data into joint training, the performance doesn't improve even sometimes decreases. Instead, BGE compensates for the absent comparisons caused by inter-task data unavailability, which is much more meaningful in continual learning. Our contributions can be summarized as follows:

- We point out that existing methods overlook the issue of inter-task data comparisons, and propose BGE to incorporate external data into training to address this gap.
- We propose the One-Propose-One (OPO) sampling algorithm to sample external data that are relevant to tasks and sufficiently diverse, while also filtering out OOD data that are not beneficial for learning.
- Experiments show that BGE can be seamlessly integrated into existing CCSSL methods and consistently yields significant improvement.

2 Related work

Self-Supervised Learning (SSL) SSL trains the network without the need for supervised signals. One of the prominent branches is contrastive learning [5, 8–10, 21, 23, 53]. The objective of contrastive learning can be roughly explained as reducing the distance between positive pairs while enlarging it between negative pairs. SimCLR [8] simply follows this objective but requires a large batch size. MoCo [10, 23] introduces a momentum encoder and a negative sample dictionary to solve this problem. SwAV [5] and Barlow Twins [53] introduces prototype comparisons and cross-decorrelation loss, respectively. Then BYOL [21] and SimSiam [9] can conduct contrastive learning without negative samples. However, all these methods assume that a large dataset is available for pre-training, which is often impractical in real-world scenarios where data acquisition is incremental. Therefore, we research a continual method, which is more practical.

Since no labeling requirement, incorporating external data into SSL is straightforward. Prior long-tailed SSL works [3, 28] leverage external data to balance head and tail classes. Instead, we extend the exploration to continual learning, aiming to use external data to compensate for the absent inter-task comparisons while further preventing catastrophic forgetting.

Continual learning Continual learning allows the network to learn from sequentially arriving data and prevent catastrophic forgetting. Existing continual learning methods can be categorized into three groups, which are 1) Regularization-based methods [1, 14, 29, 32, 34, 50, 54] add additional regularization constraints such as knowledge distillation [14, 32, 50] or limiting important parameters update [1, 29, 34, 54] to network training. 2) Replay-based methods [4, 26, 40, 43, 55] save few representative data from old tasks called exemplars to recover the distribution of old data when the new task is trained. 3) Architecture-based methods [15, 36, 37, 41, 51], which adjust the architecture or parameters of the network during each task training. Currently, most continual learning methods still focus on supervised learning. While some of them [6, 33, 44] draw on the idea of contrastive learning, there are still few works consider continual learning without any supervision. Among them, CaSSLe [16], PFR[18], and POCON[19] use distillation, and CPPF[11] adds clustering to form a more complete framework. Sy-CON [7] also reveals the distinction between CCSSL and joint training, but it only additionally passes current task data into the old network to get more diverse intra-task negative features, which still fails to provide effective inter-task comparisons. Thus it underperforms in most contrastive learning frameworks. Compared to them, we introduce external data to facilitate implicit inter-task comparisons to solve the problem of absent inter-task comparisons.

3 Proposed method

3.1 Preliminary

Contrastive Self-Supervised Learning (CSSL) In Self-Supervised Learning (SSL), the dataset D contains only n image inputs $\{x_1, x_2, ..., x_n\}$ without labels. SSL trains a network f_θ parameterized by θ to map these inputs to embeddings $\{z_1, z_2, ..., z_n\}$. Many well-known SSL works [5, 8, 21, 23, 53] use contrastive learning framework. In contrastive learning, a random augmentation function A is pre-designed. Given an input x, two augmented views (x_a, x_b) are obtained by applying A twice. Subsequently, embeddings $z_a = f_\theta(x_a)$ and $z_b = f_\theta(x_b)$ are passed through a projector $h_{\theta'}$ parameterized by θ' to get $z'_a = h_{\theta'}(z_a)$, $z'_b = h_{\theta'}(z_b)$, which are involved in $\mathcal{L}_{SSL}$. In essence,

$\mathcal{L}_{SSL}$ expects the network to output similar embeddings for two views of the same input (i.e. positive pair), while ensuring that embeddings from views of different inputs (i.e. negative pair) are dissimilar.

Continual CSSL (CCSSL) In CCSSL setting, The overall dataset D is divided into multiple tasks. Assuming that T tasks $\{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_T\}$ are to be learned, D can be divided into $\{D_1, D_2, ..., D_T\}$, where $D_i \cap D_j = \emptyset, \forall i, j \in \{1 : T\}$. Also as SSL, for each task $\mathcal{T}_t$, D_t is only composed of n_t images $\{x_1, x_2, ..., x_{n_t}\}$ without labels. Continual learning requires the network to learn knowledge as each task's data arrives sequentially, with dataset D_i only available at $\mathcal{T}_i$. The optimization objective is to continually train the network parameter θ to satisfy every task, which is defined as:

$$\underset{\theta}{argmin} \sum_{t=1}^{T} \mathbb{E}_{(x_a,x_b)\sim A(D_t)} \mathcal{L}_{SSL}(h_{\theta'}(f_\theta(x_a)), h_{\theta'}(f_\theta(x_b))) \tag{1}$$

3.2 Revising and improving CCSSL via external data

Typical contrastive learning paradigms [8, 23, 53] can be generalized as reducing distances between positive pairs and enlarging them between negative pairs on feature hyperspheres. Adjusting the interrelationships of sample pairs in this way enables the network to effectively represent features [27, 49]. However, in CCSSL, the data is divided by tasks. During the learning process of task $\mathcal{T}_t$, data from other tasks are unavailable. This prevents adequate tuning of inter-sample relationships, resulting in suboptimal network training. We identify two reasons for this suboptimality: **1) The network rapidly forgets knowledge about old data due to catastrophic forgetting**, so their features cannot be well extracted in subsequent tasks. **2) Insufficient learning about each task occurs** because data from one task cannot act as negative samples for another task. While prior works address problem 1 through techniques like distillation [16, 18, 19] and clustering [11], problem 2 remains underexplored. However, we argue that this is unreasonable, and solving problem 2 is equally important.

Prior works [20, 32] widely agree that in the ideal case, continual learning can perform up to joint learning, wherein no forgetting occurs and each task reaches optimality. However, in CSSL, even if no forgetting occurs, there is still an optimization gap between continual and joint learning due to the absence of inter-task data comparisons in the training objective. Unlike supervised learning which guides the network through labels, CSSL relies on data interactions for network learning. When data is incomplete, the training objective also becomes incomplete. For better comprehension, we can decompose the joint training contrastive loss into two terms as in Eq. 2, representing the comparisons of intra-task and inter-task data, denoted as $\mathcal{L}_{intra}$ and $\mathcal{L}_{inter}$, respectively. $\mathcal{L}_{intra}$ is the training objective of the conventional CCSSL, also referred to as $\mathcal{L}_{continual}$. However, for input $x \in D_t$ in task $\mathcal{T}_t$, negative samples come exclusively from D_t rather than the overall dataset D, making direct comparisons between inter-task data infeasible. Consequently, $\mathcal{L}_{inter}$ can not be computed and optimized in continual learning forever, resulting in a $\mathcal{L}_{inter}$ gap between $\mathcal{L}_{continual}$ and $\mathcal{L}_{joint}$.

$$\begin{aligned}\mathcal{L}_{joint} = \frac{1}{T}\sum_{t=1}^{T}\Big(&\overbrace{\mathbb{E}_{(x_a,x_b)\sim A(D_t)}\mathcal{L}_{SSL}\left(h_{\theta'}\left(f_\theta\left(x_a\right)\right), h_{\theta'}\left(f_\theta\left(x_b\right)\right)\right)}^{\mathcal{L}_{intra} = \mathcal{L}_{continual}} \\ &\underbrace{+\,\mathbb{E}_{\substack{x_a\sim A(D_t),\\ x_b\sim A(D-D_t)}} \mathcal{L}_{SSL}\left(h_{\theta'}\left(f_\theta\left(x_a\right)\right), h_{\theta'}\left(f_\theta\left(x_b\right)\right)\right)}_{\mathcal{L}_{inter}}\Big)\end{aligned} \tag{2}$$

We argue that the lack of optimization for $\mathcal{L}_{inter}$ leads to confusion between inter-task data. Figure 1 Right compares the t-SNE visualizations of features from 4 CIFAR100 classes under joint and 10 tasks continual training (4 classes belong to different tasks during continual training). Compared to the joint-trained network, the continually trained network shows poor clustering and severe class boundary confusion. More experiments about inter-task confusion can be found at Appendix A.2.1. Despite CaSSLe [16] employing distillation to consolidate old knowledge, the issue of inter-task class boundary confusion remains. To address the overlooked problem of $\mathcal{L}_{inter}$, a straightforward idea is to save exemplars for each task. However, this may raise serious privacy concerns. We therefore explore an alternative method to optimize $\mathcal{L}_{inter}$ without exemplars and protect the discriminative class boundaries. Figure 1c shows the feature distribution of our method, with all 4 inter-task classes better distinguished, and the overall distribution closer to joint training.

To compensate for $\mathcal{L}_{inter}$, bridging the gap of inter-task comparisons is essential. This requires introducing additional comparisons into each task, implying extra data incorporation. Under the constraints of continual learning, simultaneous access to data from multiple tasks is infeasible. Therefore, the idea emerges to incorporate publicly available external data into CCSSL to address the lack of inter-task comparisons. Each task's data can be directly compared with external data, enabling relationships between data to be passed along the task sequence. Moreover, using external data better protects privacy, and the costs of obtaining unlabeled data from public data sources are extremely low. We thus propose our method BGE, meaning **B**ridging the inter-task comparison **G**ap with **E**xternal data, as shown in Figure 1 Left. BGE incorporates external data into each task's training except the first one, and resamples part of them after each task using our sampling algorithm (detailed in Section 3.3). This external data acts as a bridge for inter-task comparisons, constructing implicit comparisons for inter-task data. For task $\mathcal{T}_t$, with D_e^{t-1} as the external data sampled after task $\mathcal{T}_{t-1}$, the training objective is defined as:

$$\mathcal{L}_t = \mathbb{E}_{(x_a,x_b)\sim A\left(D_t \cup D_e^{t-1}\right)} \mathcal{L}_{SSL}\left(h_{\theta'}\left(f_\theta\left(x_a\right)\right), h_{\theta'}\left(f_\theta\left(x_b\right)\right)\right) \tag{3}$$

Incorporating external data aligns the optimization objective of continual learning more closely with Eq. 2, enhancing the mutual understanding of inter-task classes.

3.3 One-Propose-One (OPO) sampling

While abundant external data features generally cover in-task data comprehensively, incorporating all external data into continual learning is impractical due to computational constraints. Additionally, open-world external data may include substantial task-irrelevant out-of-distribution (OOD) data, which is unhelpful for training. Therefore, a sampling algorithm is needed to select high-quality external data. We observe that $\mathcal{L}_{inter}$ includes comparisons of current task data D_t with both old task data $D_{1:t-1}$ and future task data $D_{t+1:T}$. So sampled external data should ideally proxy for both old and future task data. To represent old data, sampled data should have similar features to them, while representing future data requires imaginative sampling. Therefore, our sampling algorithm is based on both proximity and diversity considerations, and integrates these two aspects into a single objective without any hyperparameters. We noted that prior sampling algorithms [3, 28] for long-tailed learning also consider proximity and diversity, but they require hyperparameters selection.

We measure proximity using the cosine distance between sample features. On the other hand, prior work [49] indicates that to avoid collapse, contrastive learning methods tend to map all inputs to a uniform distribution within the feature hypersphere (i.e. uniformity). Thus we assume that the entire distribution of the current task data approximately covers the hypersphere, ensuring diversity. Based on the above, we propose a sampling algorithm called *One-Propose-One (OPO)* as depicted in Algorithm 1. After training each task $\mathcal{T}_t$, OPO constructs the external dataset D_e^t, which is then incorporated in training task $\mathcal{T}_{t+1}$. Specifically, OPO considers that each in-task sample can equally propose an external sample with the closest feature distance to itself and has not been proposed. Given the current task budget K_t, we collect all proposed samples as a candidate set D_c, and select the K_t minimum distance samples to be added to the external dataset D_e^t. We follow iCaRL [40]'s exemplar update algorithm, maintaining an equal budget for each task within the total budget K. OPO ensures proximity and diversity without hyperparameters, maintaining similarity to old data and adequate coverage of future data features.

4 Experiments

4.1 Experimental setup

Dataset setup We conduct experiments with the following datasets: 1) **CIFAR100** [30], which contains 100 classes, each with 500 train images and 100 test images. Each image is 32×32 pixels. We follow the class incremental learning setting to split the classes equally by the number of tasks. Experiments are conducted under 4 tasks and 10 tasks settings, wherein each task contains 25 classes

Algorithm 1 *One-Propose-One(OPO)* Sampling Algorithm

Input: current task ID t, current task dataset D_t, entire external dataset D_{out}, last task sampled external dataset D_e^{t-1}, model f, total budget K, cosine distance metric $cos(\cdot,\cdot)$
Output: sampled external dataset D_e^t

Calculate current task budget $K_t = \frac{K}{t}$, Adjust $D_e^{t-1} = \text{REDUCEDATA}(D_e^{t-1}, K_t)$ [40]
Create candidate set $D_c = \{\}$
while $\mid D_c \mid < K_t$ **do**
  for each $x \in D_t$ **do**
    $u = argmin_{x' \in (D_{out} - D_e^{t-1})} cos(f(x), f(x')),\ d_u = min_{x_i \in D_t} cos(f(x_i), f(u))$
    $D_c = D_c \cup \{u\},\ D_{out} = D_{out} - \{u\}$
  end for
end while
$D'_c = \text{SORT}(D_c, key = d_u)\,[: K_t],\ D_e^t = D_e^{t-1} \cup D'_c$
return D_e^t

and 10 classes. 2) **ImageNet100** [46], which consists of 100 classes selected from ImageNet [12], with a total of 130K images of 224×224 pixels. It is equally split under 5 tasks and 10 tasks settings.

External dataset setup For CIFAR100, the selected external datasets include **CIFAR10**, **Places365**$_{test}$ (the test set of Places365 [57]) and **ImageNet-R** [24], among them, Places365$_{test}$ and ImageNet-R are OOD for CIFAR100. CIFAR10 contains 50,000 images with 32×32 pixels in 10 classes. Places365 is a scene recognition dataset with its test set containing 328,500 images of various scenes. ImageNet-R contains 24,000 images featuring art, cartoons, and other styles. We resize both Places365$_{test}$ and ImageNet-R to 32×32 pixels. We consider three compositions of external datasets, **CIFAR** (CIFAR10), **CP** (CIFAR10+Places365$_{test}$) and **CPI** (CIFAR10+Places365$_{test}$+ImageNet-R)

For ImageNet100, the external datasets include **ImageNet900**, **Places365** and **DomainNet** [39]. ImageNet900 is all data in ImageNet excluding ImageNet100, totaling 1.1 million images. Places365 contains 1.8 million images, and DomainNet contains 0.6 million images of 6 domains. They are also used here as OOD data. All data are 224×224 pixels. We consider three compositions of external datasets, **IN** (ImageNet-900), **INP** (ImageNet900+Places365) and **IND** (ImageNet900+DomainNet).

Baselines We compare the original performance of existing exemplar-free CCSSL methods to their performance when with BGE. The methods we compare include 1) **Fine-Tune (FT)**: Sequentially training the network with data from each task without additional prevention of catastrophic forgetting. 2) **CaSSLe** [16]: Introducing a distillation loss between the current model and the old model in the form of contrastive loss. 3) **PFR** [18]: Addressing catastrophic forgetting based on functional regularization [17]. We slightly optimized its network structure and training procedure.

Training and evaluation setup Unless specified otherwise, all experiments employ Barlow Twins [53] as the contrastive learning framework and Resnet18 [22] as the backbone. The sampling budget is uniformly set at 10K. For evaluation, we follow [16, 18, 19] to report the linear evaluation accuracy of the final network across all classes as the evaluation metric. For other setups see Appendix A.1.

4.2 Results

Performance improvement on prior methods We compare the performance improvement BGE yields to the base methods when using different external data compositions. Table 1 shows that on CIFAR100, BGE can consistently and significantly improve base methods. It is worth noting that as the number of tasks increases, BGE yields even greater improvement, with improvement of 1.5%-3.5% for 4 tasks and 2.5%-7% for 10 tasks. This is also in line with our motivation, as an increasing number of tasks results in more missing inter-task data comparisons.

Moreover, across different external dataset compositions, we observe that CIFAR yields the most significant improvement. This is attributed to the CIFAR10 dataset best matches the distribution of CIFAR100, thereby offering highly relevant features, even if their classes do not intersect. When incorporating datasets like Places365 or ImageNet-R, which are OOD for CIFAR100, the improvement decreases. Thanks to our OPO sampling algorithm can well resist the harm of OOD data (detailed in

Table 1: Comparison of BGE's performance improvement on CIFAR100. CIFAR, CP, and CPI are different external dataset compositions. Performance was evaluated by linear evaluation accuracy of the final network. We equally divided classes into 4 tasks and 10 tasks. BGE consistently improves base methods across different external dataset compositions. As for Joint training, ED represents adding equivalent external data, which does not improve the performance.

Methods	CIFAR		CP		CPI	
	4tasks	10tasks	4tasks	10tasks	4tasks	10tasks
FT	56.19	49.36	56.19	49.36	56.19	49.36
FT+*BGE*	59.49(+3.30)	56.62(+7.26)	58.69(+2.50)	55.14(+5.78)	58.71(+2.52)	55.74(+6.38)
CaSSLe [16]	60.04	53.89	60.04	53.89	60.04	53.89
CaSSLe+*BGE*	62.38(+2.34)	58.14(+4.25)	61.72(+1.68)	56.92(+3.03)	61.51(+1.47)	56.36(+2.47)
PFR [18]	60.92	55.57	60.92	55.57	60.92	55.57
PFR+*BGE*	64.37(+3.45)	61.02(+5.45)	63.15(+2.23)	60.31(+4.74)	62.88(+1.96)	59.99(+4.42)
Joint Acc						
Joint	68.09		68.09		68.09	
Joint+*ED*	68.15(+0.06)		67.11(-0.98)		68.19(+0.10)	

Table 2: Performance improvement yielded by BGE on ImageNet100. IN, INP, and IND are different external dataset compositions. ED represents adding equivalent external data in joint training.

Methods	IN		INP		IND	
	5tasks	10tasks	5tasks	10tasks	5tasks	10tasks
FT	64.02	56.72	64.02	56.72	64.02	56.72
FT+*BGE*	68.20(+4.18)	64.16(+7.44)	67.84(+3.82)	64.08(+7.36)	69.06(+5.04)	65.00(+8.28)
CaSSLe [16]	70.02	60.68	70.02	60.68	70.02	60.68
CaSSLe+*BGE*	72.46(+2.44)	66.80(+6.12)	71.44(+1.42)	65.94(+5.26)	72.68(+2.66)	67.10(+6.42)
PFR [18]	70.14	63.12	70.14	63.12	70.14	63.12
PFR+*BGE*	72.52(+2.38)	69.28(+6.16)	72.94(+2.80)	68.40(+5.28)	72.60(+2.46)	68.94(+5.82)
Joint Acc						
Joint	80.44		80.44		80.44	
Joint+*ED*	80.24(-0.20)		79.70(-0.74)		78.88(-1.56)	

Section 4.3). On ImageNet100, the performance improvement is shown in Table 2, showcasing a similar improvement regularity to that observed on CIFAR100. BGE achieves 1.5%-4% improvement for 5 tasks and 5%-7.5% improvement for 10 tasks. More experiments see Appendix A.2.7.

We also emphasize that although it might seem intuitive that network performance would improve with richer data because of richer features, BGE yielded improvement does not simply stem from using more data. In Table 1 and Table 2, we incorporate an equal amount of external data into joint training. However, the results do not improve, and may even decrease when the external data contains OOD samples. We believe this is because incorporating irrelevant external data into the training process causes the model to allocate some capacity to learning these unrelated data, thereby weakening its focus on the in-task data. Hence, the learning of external data can not directly contribute to the learning of in-task data.

Long task sequence experiments We conduct experiments with 100 tasks on CIFAR100, which means one task only contains one class, to verify the effectiveness of BGE on long task sequences. We set the sampling budget to 1000. Figure 2 shows the performance of different base methods with or without BGE as the learned tasks increase. On one hand, BGE improves the final network performance, especially evident in FT and PFR. On the other hand, the network's performance increases even more rapidly with BGE, indicating that the network's generalization ability to unseen

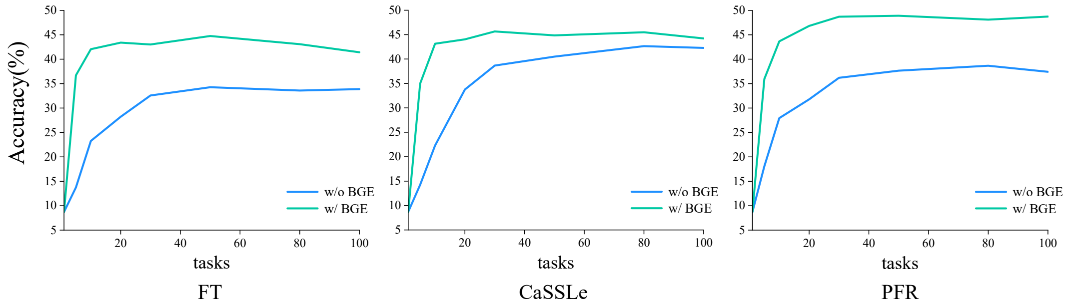

Figure 2: Performance improvement of BGE at CIFAR100 100 tasks setting.

Table 3: Accuracy on CIFAR100 and ImageNet100 with different sampling algorithms. **Bold** indicates better performance.

	CIFAR100 FT				CIFAR100 PFR			
External dataset	CP		CPI		CP		CPI	
Sampling algorithm	4tasks	10tasks	4tasks	10tasks	4tasks	10tasks	4tasks	10tasks
random	57.41	52.78	57.22	52.56	62.57	59.33	62.58	58.45
OPO	**58.69**	**55.14**	**58.71**	**55.74**	**63.15**	**60.31**	**62.88**	**59.99**
	ImageNet100 FT				ImageNet100 PFR			
External dataset	INP		IND		INP		IND	
Sampling algorithm	4tasks	10tasks	4tasks	10tasks	4tasks	10tasks	4tasks	10tasks
random	66.50	61.90	66.90	61.90	71.36	67.26	72.56	67.98
OPO	**67.84**	**64.08**	**69.06**	**65.00**	**72.94**	**68.40**	**72.60**	**68.94**

tasks is higher. This stems from BGE can both overcome catastrophic forgetting and compare with future tasks it guessed, thus accumulating more knowledge in the early training stages.

4.3 Ablation study

Sampling algorithm Table 3 shows the effect of OPO sampling compared to random sampling for FT and PFR improvement when external datasets contain OOD data. OPO algorithm consistently provides more improvement than random sampling. However, we also observed that when all external data are in-distribution (ID), the improvement from OPO algorithm is not stable. This suggests that external data quality is sufficiently high, making random sampling sufficient for our needs. To validate this, we calculated the Fréchet Inception Distance (FID) scores [25] between the in-task dataset and external datasets obtained by different sampling algorithms under CIFAR and CPI compositions, as shown in Figure 3. A lower FID score indicates greater similarity between two datasets, and vice versa. Figure 3 shows that with the CIFAR composition, the FID score is lower, and the effect of the OPO algorithm is little, indicating that this dataset is already of high quality. In contrast, under CPI, the FID score is higher when random sampling, while shows a significant decrease when OPO sampling. It indicates that the OPO algorithm adjusts the distribution of the external dataset considerably to make it more compatible with the in-task dataset. Therefore OPO algorithm will have more advantages when the external dataset contains OOD data.

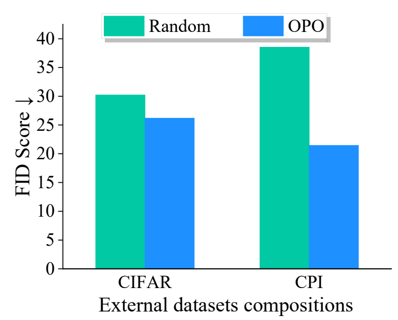

Figure 3: FID score of different sampling algorithms when CIFAR and CPI as external data.

Besides, we observed that the advantage of OPO sampling algorithm is more significant on the ImageNet100 dataset. We believe this can be attributed to two factors: 1) Higher image pixels contain more information, and fewer images will satisfy the proximity. 2) With a larger quantity of external data, there are more potentially high-quality data, facilitating better sampling.

Effect of additional positive and negative pairs We further investigate whether additional positive or negative pairs provided by BGE contribute more to performance improvement. We conduct experiments based on CaSSLe [16] on the CIFAR100 4 tasks setting. Because this experiment requires explicitly calculating the loss incurred by each positive and negative pair, we convert the framework to SimCLR [8]. We masked the additional positive or negative pairs in Table 4. The results show that both types of pairs improve performance individually, and negative pairs yield more significant improvement, supporting our emphasis that the impact of absent inter-task comparisons is severe but neglected. But positive pairs also yield performance improvement, which is because high-quality external data have feature intersections with in-task data, proving that external data can prevent catastrophic forgetting as well. With the synergistic effect of both, the improvement reaches the highest.

Table 4: Comparison of additional positive and negative pairs' effects.

Negative	Positive	Acc
		52.79
	✓	53.40
✓		55.61
✓	✓	56.21

Experiments with only OOD external data In the experiments presented in Table 1 and Table 2, all external data contain some amount of ID data. To assess BGE's performance without any ID data in the external dataset, we conduct experiments on CIFAR100 4 tasks based on PFR, as shown in Table 5. The external dataset is only composed of ImageNet-R or Places365$_{test}$. In joint training, these data are detrimental. While in continual training, BGE consistently improves the base method by nearly 2%, regardless of the composition of OOD data used. It indicates that the performance improvement from BGE does not only come from imitating in-task data features, but also from introducing similar additional comparisons into each task itself, which is beneficial for constructing implicit inter-task comparisons. Even if the external data has few recognizable similar features to the in-task data, the network can still try its best to mine valuable knowledge from external data to compensate for inter-task comparisons.

Table 5: Effectiveness of BGE when external data are totally OOD.

External dataset compositions		PFR	+BGE	Joint	Joint+ED
ImageNet-R	Places365$_{test}$				
✓		60.92	62.85(+1.93)	68.09	68.03(-0.06)
	✓	60.92	62.81(+1.89)	68.09	67.75(-0.34)
✓	✓	60.92	62.88(+1.96)	68.09	67.15(-0.94)

BGE with more types of datasets We validate the effectiveness of BGE across more aspects of external datasets. Table 6 presents the results when using GenImage [58], a dataset of generated images; CC3M [42], a dataset sourced from the Internet; and CUB200 [48], a fine-grained bird dataset as external dataset. Experiments with GenImage and CC3M demonstrate BGE's effectiveness with both model-generated and real-world Internet data, demonstrating its practical value. Since CUB200 is fine-grained and lacking in diversity, it is extremely unfriendly to BGE, yet BGE can still improve the base method.

Table 6: Performance of BGE when choosing more types of datasets.

External datasets	Acc
N/A	60.92
GenImage [58]	64.37
CC3M [42]	63.53
CUB200 [48]	62.42

5 Conclusion

In this paper, we address a commonly overlooked but severe issue in Continual Contrastive Self-Supervised Learning (CCSSL): the lack of inter-task comparisons. To tackle this, we propose our method BGE to incorporate external data into training, bridging the inter-task gap and facilitating implicit inter-task data comparisons. We also design the One-Propose-One sampling algorithm to select high-quality external data and filter out irrelevant OOD data. BGE can be seamlessly integrated into existing methods and yield significant improvement.

References

[1] R. Aljundi, F. Babiloni, M. Elhoseiny, M. Rohrbach, and T. Tuytelaars. Memory aware synapses: Learning what (not) to forget. In *Proceedings of the European conference on computer vision (ECCV)*, pages 139–154, 2018.

[2] S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021.

[3] J. Bai, Z. Liu, H. Wang, J. Hao, Y. Feng, H. Chu, and H. Hu. On the effectiveness of out-of-distribution data in self-supervised long-tail learning. *arXiv preprint arXiv:2306.04934*, 2023.

[4] J. Bang, H. Kim, Y. Yoo, J.-W. Ha, and J. Choi. Rainbow memory: Continual learning with a memory of diverse samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8218–8227, 2021.

[5] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33:9912–9924, 2020.

[6] H. Cha, J. Lee, and J. Shin. Co2l: Contrastive continual learning. In *Proceedings of the IEEE/CVF International conference on computer vision*, pages 9516–9525, 2021.

[7] S. Cha and T. Moon. Sy-con: Symmetric contrastive loss for continual self-supervised representation learning. *arXiv preprint arXiv:2306.05101*, 2023.

[8] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[9] X. Chen and K. He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

[10] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

[11] X. Chen, Z. Sun, K. Yan, S. Ding, and H. Lu. Combining past, present and future: A self-supervised approach for class incremental learning. *arXiv preprint arXiv:2311.08764*, 2023.

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[14] A. Douillard, M. Cord, C. Ollion, T. Robert, and E. Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XX 16*, pages 86–102. Springer, 2020.

[15] C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, and D. Wierstra. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*, 2017.

[16] E. Fini, V. G. T. Da Costa, X. Alameda-Pineda, E. Ricci, K. Alahari, and J. Mairal. Self-supervised models are continual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2022.

[17] S. Garg and Y. Liang. Functional regularization for representation learning: A unified theoretical perspective. *Advances in Neural Information Processing Systems*, 33:17187–17199, 2020.

[18] A. Gomez-Villa, B. Twardowski, L. Yu, A. D. Bagdanov, and J. Van de Weijer. Continually learning self-supervised representations with projected functional regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3867–3877, 2022.

[19] A. Gomez-Villa, B. Twardowski, K. Wang, and J. van de Weijer. Plasticity-optimized complementary networks for unsupervised continual learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1690–1700, 2024.

[20] D. Goswami, Y. Liu, B. Twardowski, and J. van de Weijer. Fecam: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[21] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33: 21271–21284, 2020.

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[23] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[24] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021.

[25] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[26] S. Hou, X. Pan, C. C. Loy, Z. Wang, and D. Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 831–839, 2019.

[27] W. Huang, M. Yi, X. Zhao, and Z. Jiang. Towards the generalization of contrastive self-supervised learning. *arXiv preprint arXiv:2111.00743*, 2021.

[28] Z. Jiang, T. Chen, T. Chen, and Z. Wang. Improving contrastive learning on imbalanced data via open-world sampling. *Advances in Neural Information Processing Systems*, 34:5997–6009, 2021.

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

[30] A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[31] K. Lee, K. Lee, J. Shin, and H. Lee. Overcoming catastrophic forgetting with unlabeled data in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 312–321, 2019.

[32] Z. Li and D. Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[33] H. Lin, B. Zhang, S. Feng, X. Li, and Y. Ye. Pcr: Proxy-based contrastive replay for online class-incremental continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24246–24255, 2023.

[34] X. Liu, M. Masana, L. Herranz, J. Van de Weijer, A. M. Lopez, and A. D. Bagdanov. Rotate your networks: Better weight consolidation and less catastrophic forgetting. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2262–2268. IEEE, 2018.

[35] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[36] A. Mallya and S. Lazebnik. Packnet: Adding multiple tasks to a single network by iterative pruning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 7765–7773, 2018.

[37] A. Mallya, D. Davis, and S. Lazebnik. Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In *Proceedings of the European conference on computer vision (ECCV)*, pages 67–82, 2018.

[38] M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier, 1989.

[39] X. Peng, Q. Bai, X. Xia, Z. Huang, K. Saenko, and B. Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019.

[40] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 2001–2010, 2017.

[41] J. Serra, D. Suris, M. Miron, and A. Karatzoglou. Overcoming catastrophic forgetting with hard attention to the task. In *International conference on machine learning*, pages 4548–4557. PMLR, 2018.

[42] P. Sharma, N. Ding, S. Goodman, and R. Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[43] H. Shin, J. K. Lee, J. Kim, and J. Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

[44] Z. Song, Y. Zhao, Y. Shi, P. Peng, L. Yuan, and Y. Tian. Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24183–24192, 2023.

[45] Y.-M. Tang, Y.-X. Peng, and W.-S. Zheng. Learning to imagine: Diversify memory for incremental learning using unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9549–9558, 2022.

[46] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020.

[47] L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[48] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[49] T. Wang and P. Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020.

[50] Y. Wu, Y. Chen, L. Wang, Y. Ye, Z. Liu, Y. Guo, and Y. Fu. Large scale incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 374–382, 2019.

[51] S. Yan, J. Xie, and X. He. Der: Dynamically expandable representation for class incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3014–3023, 2021.

[52] L. Yu, X. Liu, and J. Van de Weijer. Self-training for class-incremental semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[53] J. Zbontar, L. Jing, I. Misra, Y. LeCun, and S. Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pages 12310–12320. PMLR, 2021.

[54] F. Zenke, B. Poole, and S. Ganguli. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR, 2017.

[55] M. Zhai, L. Chen, and G. Mori. Hyper-lifelonggan: Scalable lifelong learning for image conditioned generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2246–2255, 2021.

[56] Z. Zheng, M. Ma, K. Wang, Z. Qin, X. Yue, and Y. You. Preventing zero-shot transfer degradation in continual learning of vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19125–19136, 2023.

[57] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[58] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36, 2024.

A Appendix / supplemental material

A.1 Experimental details

We use SGD optimizer with warmup cosine scheduler to train the network with batchsize of 256. For CIFAR100, we train 500 epochs per task with a learning rate of 0.3 and weight decay of 1e-4 for FT and CaSSLe[16]. For PFR[18], we use the learning rate as 0.4. For ImageNet100, we train 400 epochs per task with a learning rate of 0.4 and weight decay of 1e-4.

We use one RTX 3090 for CIFAR100 experiments and one A40 for ImageNet100 experiments. For CIFAR100 experiments, it takes about 5 hours in 4 tasks setting and 8 hours in 10 tasks setting. For ImageNet100 experiments, it takes about 17 hours in 5 tasks setting and 27 hours in 10 tasks setting.

A.2 More experiments

A.2.1 BGE's improvement to inter-task confusion

We categorize the results of classification errors into two types, inter-task confusion (the wrong prediction belongs to a different task than the target) and intra-task confusion (the wrong prediction belongs to the same task as the target). Under the CIFAR100 4 tasks setting, we compare the probability of each of the two types of confusion occurring for the class contained in the last task for the three baseline methods, as shown in Table 7. Ideally, the ratio of intra-task confusion to inter-task confusion should be 1:3, since the ratio of the number of current task classes to the total number of previous task classes is 1:3. However, the inter-task confusion in Table 7 is 5 to 7 times higher than the intra-task confusion, suggesting that the lack of $\mathcal{L}_{inter}$ optimization has a severe impact on performance, while BGE improves this and decreases inter-task confusion.

Table 7: Comparison of intra-task confusion and inter-task confusion. ↓ means the value is the lower the better.

Method	Intra-task confusion↓	Inter-task confusion↓
FT	4.56%	33.48%
FT+BGE	4.60%(+0.04%)	30.12%(-3.36%)
CaSSLe	6.84%	32.08%
CaSSLe+BGE	6.08%(-0.76%)	28.52%(-3.56%)
PFR	6.32%	29.64%
PFR+BGE	6.44%(+0.12%)	27.36%(-2.28%)

A.2.2 Experiments on the method without negative samples

While the results in Table 4 indicate that the effectiveness of BGE mainly stems from additional negative samples, we conducted experiments using the contrastive learning framework BYOL, which calculates contrastive loss without the need of negative samples, as shown in Table 8. The results indicate that our method still achieves improvement, demonstrating its applicability even in methods without negative samples.

Table 8: Performance improvement yielded by BGE in BYOL.

Methods	CIFAR		CP	
	4tasks	10tasks	4tasks	10tasks
FT	52.36	47.97	52.36	47.97
FT+*BGE*	56.88(+4.52)	49.42(+1.45)	56.37(+4.01)	49.22(+1.25)
CaSSLe	57.46	52.61	57.46	52.61
CaSSLe+*BGE*	59.20(+1.78)	56.16(+3.55)	58.92(+1.46)	55.22(+2.61)

A.2.3 Visualization of sample algorithm

We visualize the relationship between external and in-task samples obtained by different sampling algorithms under CIFAR and CPI compositions, as shown in Figure 4. When CIFAR10 as external data, the distributions of random and OPO samples are similar, both covering the entire area effectively. While in the CPI setting, random sampling fails to cover the entire area, in contrast, the OPO algorithm achieves superior proximity and diversity, consequently leading to greater performance improvement. This observation corroborates our discussion about the sampling algorithm in Section 4.3.

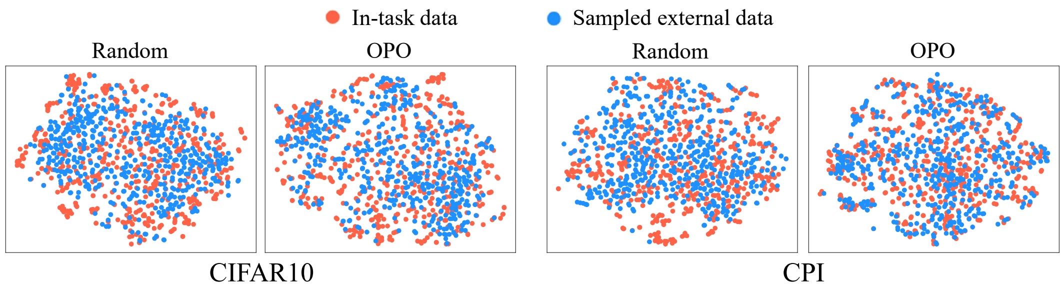

Figure 4: Comparison of external data sampled by different algorithms. When the entire external data quality is high (CIFAR), there is little difference between random and OPO sampling. When the data contains many OOD data (CPI), OPO outperforms random in sampling relevant and diverse samples.

A.2.4 Self-supervised learning feature characteristics

Previous work [2] points out that self-supervised trained networks map inputs together according to feature characteristics rather than according to labels as supervised trained networks tend to do. Inspired by them, we validate that we adopted network also has such characteristics. Table 9 shows the average number of one sample's k-nearest neighbors belonging to the class of this sample for networks trained in the supervised or self-supervised manner. It is evident that supervised networks consistently have more same-class neighbors, indicating that they cluster images based on labels. In contrast, self-supervised networks are less influenced by image classes, which is advantageous for incorporating external data.

Table 9: Statistics on how many of the k-nearest neighbors of a sample belong to the same class as this sample in self-supervised and supervised networks.

k	3	5	10	20	30	50	100	Acc
Supervised	1.76	2.93	5.58	10.87	15.63	24.38	40.86	71.64
Self-supervised	1.36	2.25	4.14	7.24	9.96	14.53	22.00	68.09

Table 10 presents the class statistics of the top 100 nearest neighbors of the "willow tree" class on the CIFAR100 dataset, as learned by self-supervised and supervised networks. Self-supervised learning results in a lower proportion of same-class neighbors, indicating less influence from class labels. Additionally, the neighbors of other classes in the self-supervised network exhibit features more similar to the "willow tree" class.

This insight suggests that external data, despite having different actual classes with in-task data, can proxy for the in-task data in self-supervised learning due to shared features. Thus giving us confidence that using external data in self-supervised learning as in BGE can yield good results and justify our cosine distance based sampling algorithm.

A.2.5 Fairness alignment

Introducing external data incurs additional iterations and new knowledge. To ensure fairness, we train the base method PFR for more epochs and use pre-training with external data to initialize the weights for in-task data training. Experimental results, as shown in Table 11, reveal that training

Table 10: The class name and average number of the top 5 classes with the highest number of the top 100 neighbors of the "willow tree" class.

Supervised learning		Self-supervised learning	
Neighbor class	Avg number	Neighbor class	Avg number
willow tree	48.59	willow tree	18.68
mushroom	7.85	oak tree	18.47
girl	4.19	maple tree	16.45
butterfly	3.05	pine tree	8.48
bus	2.94	forest	8.10

for more epochs and pre-training with external data do not lead to performance improvement. This highlights the effectiveness of BGE under fairer conditions.

Table 11: Comparison of the performance improvement of BGE and other factors to ensure fairness.

Methods	Acc
Base	60.92
Train more epochs	61.21
Use external data to pre-train	61.28
Ours	64.37

A.2.6 Experiment statistical significance

Due to limited computational resources, we report the mean and standard deviation of three random trials for only the primary experiments in Tables 12 and 13. The performance of the BGE on the three base methods when using CIFAR and CPI as external dataset compositions under the CIFAR100 4 tasks and 10 tasks setting is shown in Table 12. Table 13 shows the performance of BGE using different sampling algorithms with CPI as the external dataset, also in the CIFAR100 4 tasks and 10 tasks setting, across the same three baseline methods.

Table 12: Results with multiple runs.

Methods	CIFAR		CPI	
	4tasks	10tasks	4tasks	10tasks
FT	59.80±0.27	56.92±0.29	59.06±0.39	55.18±0.51
CaSSLe	62.39±0.41	57.99±0.28	61.86±0.36	56.52±0.21
PFR	64.13±0.24	60.01±0.02	63.12±0.33	59.94±0.05

Table 13: Results with multiple runs.

Methods	4tasks		10tasks	
	random	OPO	random	OPO
FT	57.61±0.42	59.06±0.39	52.81±0.23	55.18±0.51
CaSSLe	61.59±0.25	61.86±0.36	55.50±0.23	56.52±0.21
PFR	62.50±0.11	63.12±0.33	58.66±0.27	59.94±0.05

A.2.7 Full experiments

We present here the full set of experiments, encompassing various base methods, sampling budgets, sampling methods, and compositions of external datasets, demonstrating the performance improvement of BGE on CIFAR100 (Table 14) and ImageNet100 (Table 15).

Table 14: Full experiment results on CIFAR100 dataset.

Methods	External Dataset Budget	Sample method	CIFAR10 4tasks	CIFAR10 10tasks	CP 4tasks	CP 10tasks	CPI 4tasks	CPI 10tasks
FT	0	-	56.19	49.36	56.19	49.36	56.19	49.36
	5K	*random*	58.65(+2.46)	54.78(+5.42)	57.54(+1.35)	52.09(+2.73)	56.95(+0.76)	52.3(+2.94)
		OPO	58.51(+2.32)	54.39(+5.03)	57.56(+1.37)	54.59(+5.23)	58.3(+2.11)	53.15(+3.79)
	10K	*random*	60.01(+3.82)	56.56(+7.20)	57.41(+1.22)	52.78(+3.42)	57.22(+1.03)	52.56(+3.20)
		OPO	59.49(+3.30)	56.62(+7.26)	58.69(+2.50)	55.14(+5.78)	58.71(+2.52)	55.74(+6.38)
CaSSLe	0	-	60.04	53.89	60.04	53.89	60.04	53.89
	5K	*random*	61.26(+1.22)	56.72(+2.83)	60.86(+0.82)	54.47(+0.58)	61.06(+1.02)	54.52(+0.63)
		OPO	61.35(+1.31)	56.63(+2.74)	61.39(+1.35)	55.24(+1.35)	61.30(+1.26)	55.77(+1.88)
	10K	*random*	62.49(+2.45)	57.49(+3.60)	60.98(+0.94)	55.48(+1.59)	61.44(+1.40)	55.40(+1.51)
		OPO	62.38(+2.34)	58.14(+4.25)	61.72(+1.68)	56.92(+3.03)	61.51(+1.47)	56.36(+2.47)
PFR	0	-	60.92	55.57	60.92	55.57	60.92	55.57
	5K	*random*	62.84(+1.92)	60.01(+4.44)	62.39+(1.47)	58.49(+2.92)	62.16(+1.24)	57.78(+2.21)
		OPO	62.79(+1.87)	59.66(+4.09)	62.16(+1.24)	59.29(+3.72)	62.87(+1.95)	58.41(+2.84)
	10K	*random*	63.51(+2.59)	61.58(+6.01)	62.57(+1.65)	59.33(+3.76)	62.58(+1.66)	58.45(+2.88)
		OPO	64.37(+3.45)	61.02(+5.45)	63.15(+2.23)	60.31(+4.74)	62.88(+1.96)	59.99(+4.42)

Table 15: Full experiment results on ImageNet100 dataset.

Methods	External Dataset Budget	Sample method	IN 5tasks	IN 10tasks	INP 5tasks	INP 10tasks	IND 5tasks	IND 10tasks
FT	0	-	64.02	56.72	64.02	56.72	64.02	56.72
	10K	*random*	67.66(+3.64)	63.02(+6.30)	66.50(+2.48)	61.90(+5.18)	66.90(+2.88)	61.90(+5.18)
		OPO	68.20(+4.18)	64.16(+7.44)	67.84(+3.82)	64.08(+7.36)	69.06(+5.04)	65.00(+8.28)
CaSSLe	0	-	70.02	60.68	70.02	60.68	70.02	60.68
	10K	*random*	71.52(+1.50)	65.02(+4.34)	71.04(+1.02)	64.34(+3.66)	70.98(+0.96)	65.44(+4.76)
		OPO	72.46(+2.44)	66.80(+6.12)	71.44(+1.42)	65.94(+5.26)	72.68(+2.66)	67.10(+6.42)
PFR	0	-	70.14	63.12	70.14	63.12	70.14	63.12
	10K	*random*	72.82(+2.68)	68.20(+5.08)	71.36(+1.22)	67.26(+4.14)	72.56(+2.42)	67.98(+4.86)
		OPO	72.52(+2.38)	69.28(+6.16)	72.94(+2.80)	68.40(+5.28)	72.60(+2.46)	68.94(+5.82)

A.3 Limitations and future directions

There are still limitations to BGE, such as increased data volume for training, leading to additional computational costs. For future directions, we believe BGE can inspire further research into continual learning from the perspective of inter-task data relationships. Additionally, BGE's use of external data instead of exemplars to compensate for inter-task comparisons enhances privacy preservation, offering a pathway for future work to address privacy concerns associated with using exemplars. We research methods to allow the network to learn continually, which have no negative impact on society, and at the same time, we proposed method facilitates privacy protection and has a positive impact on society.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction in Section 1 accurately reflect our contributions in continual contrastive self-supervised learning.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Appendix A.3.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We do not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We realease our code to prove reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We release our code, and related information can be found at README.md in our code supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We specify all the training and test details in section 4.1 and Appendix A.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: we report error bars in Appendix A.2.6.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide sufficient information on the computer resources in Appendix A.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the NeurIPS Code of Ethics, and conduct research with it.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the societal impacts in Appendix A.3.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All existing assets we use are cited in section 4

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.