

DiffClass: Diffusion-Based Class Incremental Learning

Zichong Meng¹, Jie Zhang², Changdi Yang¹, Zheng Zhan¹, Pu Zhao^{*1}, and Yanzhi Wang^{*1}

¹ Northeastern University, Boston MA 02115, USA

² ETH Zürich, 8092 Zürich, Switzerland

Abstract. Class Incremental Learning (CIL) is challenging due to catastrophic forgetting. On top of that, exemplar-free CIL is even more challenging due to forbidden access to data of previous tasks. Recent exemplar-free CIL methods attempt to mitigate catastrophic forgetting by synthesizing previous task data. However, they fail to overcome the catastrophic forgetting due to the inability to deal with the significant domain gap between real and synthetic data. To overcome these issues, we propose a novel exemplar-free CIL method. Our method adopts multi-distribution matching (MDM) diffusion models to align quality of synthetic data and bridge domain gaps among all domains of training data. Moreover, our approach integrates selective synthetic image augmentation (SSIA) to expand the distribution of the training data, thereby improving the model’s plasticity and reinforcing the performance of our multi-domain adaptation (MDA) technique. With the proposed integrations, our method then reformulates exemplar-free CIL into a multi-domain adaptation problem to implicitly address the domain gap problem and enhance model stability during incremental training. Extensive experiments on benchmark CIL datasets and settings demonstrate that our method excels previous exemplar-free CIL methods with non-marginal improvements and achieves state-of-the-art performance. Our project page is available at <https://cr8br0ze.github.io/DiffClass>.

Keywords: Class Incremental Learning · Exemplar Free · Diffusion Model

1 Introduction

Although recent deep learning (DL) models have achieved superior performance even better than humans in various tasks, catastrophic forgetting [9] remains a challenging problem that limits the continual learning capabilities of DL models. Unlike humans, DL models are unable to learn multiple tasks sequentially, which forget the previous learned knowledge after learning new tasks. To address this, Class Incremental Learning (CIL) extensively investigates how to

* Corresponding Author

learn the information of new classes without forgetting past knowledge of previous classes. Various CIL works [3, 12, 25, 38, 39] try to untangle catastrophic forgetting through saving a small proportion of previous training data as exemplars in memory and retraining with them in new tasks. However, these methods suffer from privacy and legality issues of utilizing past training data, as well as memory constraints on devices. Different from previous exemplar-based CIL, Exemplar-Free CIL [7, 8, 33] has gained increasing popularity where DL models incrementally learn new knowledge without storing previous data as exemplars.

To counteract forgetting knowledge of past tasks, the most recent exemplar-free CIL works [7, 8, 33, 47] propose to synthesize previous data instead of using real data. The synthetic data of previous tasks are generated through either model inversion [45] with knowledge distillation or denoising diffusion models [11]. However, these methods suffer from significant domain gaps between synthetic data and real data especially when the number of incremental tasks is large (*i.e.* long-term CIL), which inevitably misleads the decision boundaries between new and previous classes. The obtained models favor plasticity over stability, meaning they tend to learn new knowledge but without keeping previous knowledge in mind as demonstrated in Sec. 3. Therefore, how to exhibit both stability and plasticity in exemplar-free CIL remains a crucial challenge.

To address these problems, we propose a novel exemplar-free CIL approach that bridges the crucial domain gaps and balances stability and plasticity. Our method incorporates a multi-distribution-matching (MDM) technique to fine-tune diffusion models resulting in closer distributions between not only synthetic and real data but also among synthetic data through all incremental training phases. Our method also reformulates exemplar-free CIL as task-agnostic multi-domain adaptation (MDA) problems to further deal with domain gaps between real and synthetic data, with selective synthetic image augmentation (SSIA) to enhance each incremental task learning with current task synthetic data.

We summarize our contributions as follows:

- We introduce a novel exemplar-free CIL method that explicitly mitigates forgetting and balances stability & plasticity by adopting MDM diffusion models and enhancing the dataset with SSIA to address domain gaps in exemplar-free CIL settings.
- We propose an innovative approach to reformulate exemplar-free CIL as task-agnostic MDA problems. This groundbreaking step implicitly manages domain gaps during CIL training, better addressing catastrophic forgetting in exemplar-free CIL.
- Extensive experiments on CIFAR100 [16] and ImageNet100 [28] demonstrate that our method effectively mitigates catastrophic forgetting in different exemplar-free CIL settings, surpassing SOTA methods with significant improvements.

2 Related Work

Class Incremental Learning (CIL) has merged as a challenging problem focusing on how models can incrementally learn new classes without forgetting previously acquired knowledge. To overcome the catastrophic forgetting, recent successful approaches [3, 12, 25, 38, 39, 41, 42, 49–52] store training data from previously learned classes as exemplars and replay them while learning new tasks. Exemplars are indeed helpful for reviewing past task knowledge and thus benefit the incremental learning process. However, due to privacy and legality issues, and memory constraints on devices, it may be unachievable in practice.

Thus, exemplar-free CIL gains increasing popularity among researchers. In recent years, instead of using exemplars, several exemplar-free CIL methods [2, 15, 17, 32, 43, 44] synthesize images of previously learned classes as a review instead of storing real images to mitigate forgetting. However, most of these methods suffer from significant performance degradation due to large domain gaps between synthetic and real data. Later methods [7, 33] propose to utilize modified knowledge distillation techniques to constrain the domain gaps. These methods fail as knowledge distillation tends to attribute the suboptimal performance from the previous task to the model’s learning capabilities in each current task, especially with the domain gaps in data. Different from previous works, our framework specifically aims to bridge domain gaps in exemplar-free CIL leading to a model with robustness in both stability and plasticity.

Diffusion model. Diffusion models [11, 34, 36] generate images through stochastic differential equations by progressively denoising them. This technique involves two primary stages: a forward diffusion process that incrementally introduces Gaussian noise into the input data, and a reverse diffusion process that is trained to gradually reverse this procedure, effectively removing the noise from noised input data.

Subsequent research has focused on enhancing the quality of generated outputs. There are various methods, including scaling the model [11, 21, 24, 26, 30, 48], and refining the training and sampling processes [18, 20, 35]. Among them, the Latent Diffusion Model (LDM) stands out for exhibiting great text-to-image generation quality due to scaling up the diffusion model by conducting both forward and backward diffusion processes in latent space.

In recent developments, diffusion models have also shown remarkable versatility beyond image generation. They are successfully applied in various domains, including other computer vision tasks [19, 40], audio processing [31], and even text-to-3D [23].

How to customize efficiently customized diffusion models also stands out as a recent heated topic. Various works have been proposed including techniques involving altering text embeddings [4, 5], altering text embeddings [4, 5], associating special words with small number of example images [27], or inserting a small number of new weights [13],

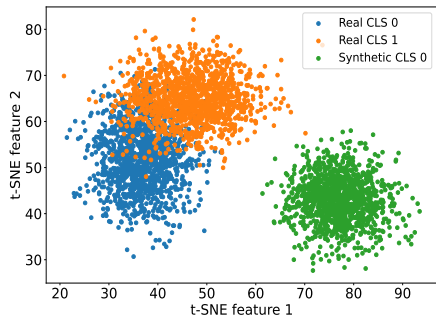


Fig. 1: Domain Gaps in Exemplar-Free CIL. The distribution of real classes is closer to each other while domain gaps exist between real class 0 and synthetic class 0.

In this work, we employ the Stable Diffusion model, *i.e.* a variant of LDM, and tailor it specifically for our method using LoRA [13] that can effectively address domain gaps in exemplar-free CIL.

3 Diagnosis: Domain Gaps in Exemplar-Free CIL

Although recent advancements in generative artificial intelligence can generate realistic images, we notice that the distributions of the generated synthetic images are still different from those of real images with domain gaps, leading to low accuracy in the classes trained with synthetic data in exemplar-free CIL settings. We also further dig into the low accuracy and find that the reason may be the model’s preference for domains over classes after training, *i.e.* the model classifies whether the image is real or synthetic rather than its true label.

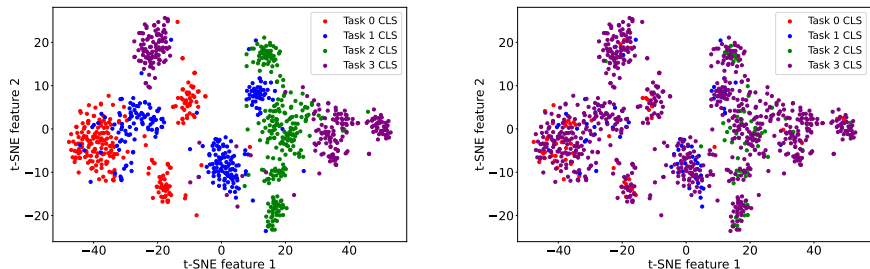
In Fig. 1, a t-SNE visualization is performed to compare real data of class 0 and 1 from ImageNet100 [12] with synthetic data of class 0 generated by the pretrained stable diffusion V1.5 model [26]. The visualization reveals that the distributions of the real classes are more closely aligned, while a significant domain gap is evident between the synthetic data of class 0 and its real counterpart.

These domain gaps can potentially effect model’s performance after model training with real and synthetic data, since the decision boundary can be significantly distorted by synthetic data, as it may treat the real class 0 and class 1 (with a smaller distribution discrepancy) as the same class in testing.

We also conduct an experiment in a class incremental setting to further verify. In specific, we train a model with only a ResNet [10] backbone and a linear classifier for the first four tasks (each with 5 classes) in a 20-task CIL setting on the ImageNet100 dataset (refer to Sec. 5 for more details). From the second to the fourth tasks, aside from the real data of the current task, we also train with synthetic data of the previous tasks generated by the pre-trained SD V1.5 model. We additionally train another model with entirely real data for the four tasks as a reference for how well the model can perform with real data.

Table 1: Diagnosis experiment accuracy result (in %) of incremental training the model with synthetic previous task data and real data of current task *vs.* training model with all real data for first four tasks of twenty-task incremental setting on ImageNet100.

Training Data Domain	CLS 0-4	CLS 5-9	CLS 10-14	CLS 15-19	Total Classes
Synthetic & Real	47.67	48.39	51.11	89.31	59.37
Real Data Only	85.97	80.11	83.54	81.27	82.72



(a) Feature Embedding with Ground Truth Label (b) Feature Embedding with Prediction Label

Fig. 2: t-SNE Visualization of Test Data’s Feature Embedding. Most of the previous task test data in incremental task 3 are misclassified as one of the task 3 classes.

In Tab. 1, we present the accuracy on the real test dataset at the end of task 4. As observed, the model performs significantly better on the classes of the new task (*i.e.* class 15-19, trained with real data) than previous tasks (*i.e.* class 0-14, trained with both real in previous task and synthetic data in the current task), demonstrating the model’s preference for plasticity over stability.

In Fig. 2, we further use t-SNE visualization for the feature embeddings of test data extracted from the incrementally trained ResNet18 backbone. As observed from Fig. 2, most of misclassified test data from classes of previous tasks are labeled to new classes of the most recent task, indicating the model’s labeling preference of domain over class, *i.e.* the model labels whether it is real or synthetic, rather than its true class.

Inspired by the diagnosis experiments, our method tries to mitigate the domain gaps and balance plasticity & stability.

4 Methodology

4.1 Framework

Following previous works [7, 25, 33], CIL contains N incremental learning phases or tasks. In the i^{th} incremental phase (or interchangeably \mathcal{T}_i) $0 \leq i < N$, our framework mainly consists of the following three steps.

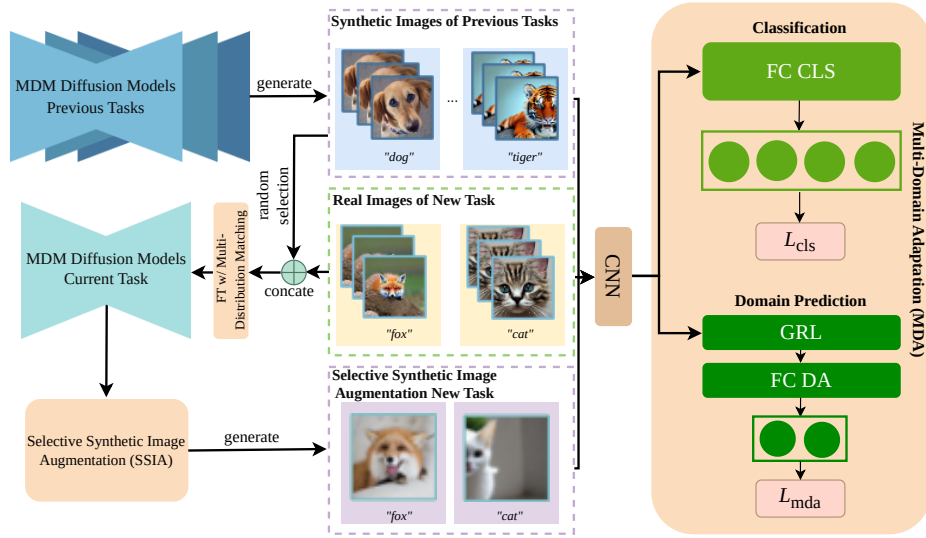


Fig. 3: Model Framework Overview learning on current task \mathcal{T}_{i+1} . previous MDM diffusion models $J_{0:i}$ are used to generated Synthetic Data of previous tasks $\mathcal{D}_{0:i}^{\text{syn}}$. MDM diffusion model of current task is then finetuned using MDM technique using Real current task Data $\mathcal{D}_i^{\text{real}}$ and randomly sampled small batch of $\mathcal{D}_{0:i}^{\text{syn}}$. $J_{0:i}$ is subsequently used to obtain $\mathcal{D}_i^{\text{aug}}$ by SSIA. The model trains with MDA on the combined dataset.

Finetuning Multi-Distribution Matching Diffusion Model with LoRA.

In the i^{th} incremental task, the real data of the current task $\mathcal{D}_i^{\text{real}}$ and the synthetic data of the previous tasks $\mathcal{D}_{0:i}^{\text{syn}}$ (notation 0:i means integers from 0 up to but not including i) generated by fine-tuned diffusion models $J_{0:i}$ is available. We use $\mathcal{D}_i^{\text{real}}$ and randomly sampled a small batch of $\mathcal{D}_{0:i}^{\text{syn}}$ to fine-tune a multi-distribution matching diffusion model J_i using LoRA. The finetuned diffusion model J_i can be used to generate synthetic data. Based on LoRA, the cost to finetune and store diffusion models is relatively small.

Forming Training Dataset for Current Task. The training dataset $\mathcal{D}_i^{\text{train}}$ for the current task consists of three parts, (1) the synthetic data of the previous tasks $\mathcal{D}_{0:i}^{\text{syn}}$ synthesized by fine-tuned diffusion models $J_{0:i}$, (2) the real data of the current task $\mathcal{D}_i^{\text{real}}$, and (3) the image augmentation data $\mathcal{D}_i^{\text{aug}}$ generated from J_i . For $i = 0$, the synthetic data are ignored. The model can then start training by randomly sampling training batches $(x^{\text{train}}, y^{\text{train}})$ from the newly-formed training dataset.

Training with Multi-Domain Adaptation. For each batch of training data, we adopt the training method with multi-domain adaptation. Specifically, after feature extraction with a CNN backbone defined as $F_i : \mathbb{R}^{h \times w \times 3} \rightarrow \mathbb{R}^d$, the extracted features go through two branches: a linear classifier $G_i : \mathbb{R}^d \rightarrow \mathbb{R}^c$,

and a gradient reverse layer (GRL) followed by a linear classifier $K_i : \mathbb{R}^d \rightarrow \mathbb{R}^2$. During training, G_i learns to classify representations of new classes in new tasks without forgetting previous classes, while K_i acquires the knowledge of boundaries between diffusion-generated synthetic data and real data.

The details and advantages of the three stages in each incremental learning phase are specified below.

4.2 Finetuning Multi-Distribution Matching Diffusion Model with LoRA

Previous exemplar-free CIL works either use or alter the sampling pre-trained diffusion models to synthesize data of previous tasks [8, 14]. However, these methods fail to generate realistic data with evident domain gaps (or distribution discrepancies) for the classes in the same incremental task or keep consistent generation quality across different incremental tasks. These bottlenecks affect the model’s robustness in stability as shown previously in Sec. 3.

Multi-Distribution Matching. To address this significant limitation in exemplar-free CIL, inspired by the recent work on training data synthesis [46] with an additional synthetic-to-real distribution-matching technique to enclose the gap between synthetic and real data distributions, we propose a multi-distribution matching (MDM) technique to fine-tune the diffusion model that best fit our exemplar-free CIL setting. In specific, when finetuning a diffusion model, we not only match the distributions of the synthetic and real data for the current task but also align the distributions of synthetic data in the current task with that in all previous tasks. With MDM, the diffusion models can be finetuned by optimizing the following loss:

$$\begin{aligned} \mathcal{L}_{MDM} &= \left\| \frac{1}{|\mathcal{D}_i^{\text{real}}| + |Z(\mathcal{D}_{0:i}^{\text{syn}})|} \sum_{j=1}^{|\mathcal{D}_i^{\text{real}}| + |Z(\mathcal{D}_{0:i}^{\text{syn}})|} (\epsilon - \epsilon_\theta(x'_t, t)) \right\|_{\mathcal{H}}^2 \\ &\leq \frac{1}{|\mathcal{D}_i^{\text{real}}|} \sum_{j=1}^{|\mathcal{D}_i^{\text{real}}|} \|\epsilon - \epsilon_\theta(x_t, t)\|_{\mathcal{H}}^2 = \mathcal{L}_{diff}. \end{aligned} \quad (1)$$

where $x' \in R_i + Z(S_{0:i})$ and $x \in R_i$. Here Z is a random selection function to incorporate only a small portion of synthetic data of past tasks $S_{0:i}$ for multi-distribution matching purposes. ϵ_θ is the noise predictor for latent space noised x_t with noise ϵ . And \mathcal{H} denotes it’s in the universal Reproducing Kernel Hilbert Space. The Loss is further constraint by the original stable diffusion loss on only $\mathcal{D}_i^{\text{real}}$ to emphasize while MDM is focused on multi-distribution matching crossing all training phase data, it should not compromise the fundamental denoising or data generation ability of the model of current real task classes. We also provide detailed deduction and proof for this equation in the Appendix.

In this way, the synthetic images generated using the diffusion models with the proposed MDM are of uniform quality in different classes and tasks. More

importantly, the distribution discrepancies or domain gaps between synthetic and real images become smaller, which fundamentally alleviates the potential domain bias problems and achieves better CIL performance.

4.3 Forming Current Task Training Dataset

Synthetic Data Augmentation has proven to enhance the model performance on various computer vision tasks due to its ability to enlarge training data distribution [1, 37]. In exemplar-free CIL, various image augmentation techniques [14, 53, 54] are frequently adopted. Therefore, when structuring the current task training dataset, aside from synthetic previous-task data generated by diffusion models $J_{0:i}$, and the real data of the current task, we further incorporate data augmentation $\mathcal{D}_i^{\text{aug}}$ with synthetic data of current task from J_i . However, in enhancing and aligning our method, we propose a different data augmentation technique, *i.e.* selective synthetic image augmentation (SSIA), to obtain $\mathcal{D}_i^{\text{aug}}$. In specific, rather than finetuning and utilizing generative models after each training phase [7, 8, 33], at the beginning phase of each task i , we finetune a MDM diffusion model J_i using LoRA as proposed in Sec. 4.2.

We generate twice the number of synthetic data as real data for the current task i and filter out the same number (or less) of distributional representative synthetic images as real data. It includes the following key steps.

- Calculate each generated class mean and create covariance matrices.

$$\mu_{cn}^{\text{gen}} = \frac{1}{|\mathcal{D}_{cn}^{\text{gen}}|} \sum_{\mathbf{x} \in \mathcal{D}_{cn}^{\text{gen}}} \mathbf{x}, \text{ where } cn \in \mathcal{CN}, \quad (2)$$

$$\text{Cov}_{cn}^{\text{gen}} = \frac{1}{|\mathcal{D}_{cn}^{\text{gen}}| - 1} \sum_{\mathbf{x} \in \mathcal{D}_{cn}^{\text{gen}}} (\mathbf{x} - \mu_{cn}^{\text{gen}})(\mathbf{x} - \mu_{cn}^{\text{gen}})^T, \quad (3)$$

where \mathcal{CN} denotes all classes in the current task.

- Sample the generated images for each current task class

$$\mathbf{x}_i^{cn} \sim \mathcal{N}(\mu_{cn}^{\text{gen}}, \text{Cov}_{cn}^{\text{gen}}), \quad (4)$$

- Calculate a selected threshold for synthetic image selection and construct the image augmentation dataset.

$$\tau_{cn}^{\text{gen}} = k \cdot \sqrt{\text{diag}(\text{Cov}_{cn}^{\text{gen}})}, \quad (5)$$

$$\mathcal{D}_i^{\text{aug}} = \bigcup_{cn=1}^{\mathcal{CN}} \{\mathbf{x}_i^{cn} \mid \|\mathbf{x}_i^{cn} - \mu_{cn}^{\text{gen}}\| \leq \tau_{cn}^{\text{gen}}\}. \quad (6)$$

With SSIA, our method can benefit for multiple reasons. MDM mitigates the domain gaps between synthetic data in different tasks and the diffusion models can generate more realistic high-quality images for SSIA. This helps to enhance

the model’s stability since domain-aligned training data can contribute to preventing feature embedding domain bias problems in exemplar-free CIL settings. SSIA can enable the model to better build knowledge for new classes. The model is capable of learning from the classes of current task trained with broader data distributions. The quality of images in SSIA is strong and representative since the synthetic images are selected from clusters around the class mean and span a calculated range with a broader class distribution. Moreover, the current task training dataset consists of both real and synthetic domains, which fortifies the multi-domain adaptation capabilities in our framework later discussed in Sec. 4.4.

4.4 Training with Multi-Domain Adaptation

Even with the multi-distribution matching technique, we still notice a nontrivial domain gap between synthetic data and real data in the training dataset. This domain gap will inevitably affect the model performance on classifying previous-task images during incremental learning, as shown in Sec. 3. Previous exemplar-free CIL works mainly adopt knowledge distillation techniques [7,33] to implicitly avoid the model favoring domains over classes, *i.e.* aiming to enable the model to classify whether it is real or synthetic rather than its true labels. However, knowledge distillation still fails to address the domain gap problem with low classification performance in CIL and a high computation complexity.

Multi-Domain Adaptation. To deal with these problems, we propose to reformulate exemplar-free CIL as a task-agnostic multi-domain adaption problem. Inspired by domain-adversarial training [6], for each task \mathcal{T}_i , after the original CNN backbone, besides the original linear classifier $G_i : \mathbb{R}^d \rightarrow \mathbb{R}^c$ for class label classification, we further construct an additional branch with a gradient reverse layer followed by another linear classifier $K_i : \mathbb{R}^d \rightarrow \mathbb{R}^2$ for domain prediction.

Hence we can formulate our exemplar-free CIL training approach in each task \mathcal{T}_i as optimizing the following:

$$\mathcal{L}_i^{\text{train}} = \mathcal{L}_i^{\text{cls}} + \mathcal{L}_i^{\text{da}}. \quad (7)$$

where

$$\mathcal{L}_i^{\text{cls}} = -\frac{1}{|\mathcal{D}_i^{\text{train}}|} \sum_{\mathbf{x} \in \mathcal{D}_i^{\text{train}}} y_c \log(G_i(F_i(\mathbf{x}))), \quad (8)$$

and

$$\mathcal{L}_i^{\text{da}} = -\frac{1}{|\mathcal{D}_i^{\text{train}}|} \sum_{\mathbf{x} \in \mathcal{D}_i^{\text{train}}} [y_d \log(K_i(F_i(\mathbf{x}))) + (1 - y_d) \log(1 - (K_i(F_i(\mathbf{x}))))]. \quad (9)$$

Here y_c represents the ground truth label for class c , and y_d represents the ground truth domain label d . The model needs to not only learn to classify the image but also distinguish whether it is real or synthetic.

Different from traditional domain-adversarial training with a focus on single target domain (real) data only, in our exemplar-free CIL setting, our model benefits from training both classification and domain branches using both target (real) and source (synthetic) domain data in each incremental task \mathcal{T}_i . For learning classification knowledge in \mathcal{T}_i , synthetic data is a crucial key for reviewing previous knowledge while real data contributes to gaining new knowledge. For learning multi-domain adaptation knowledge, adopting a mixture of data from both domains can contribute to differentiating and adapting to the distinct characteristics of each domain.

By reforming exemplar-free CIL as a straightforward task-agnostic multi-domain adaption problem, our method enjoys the following advantages. (i) Our model framework keeps simple without any cumbersome parts, which benefits incremental training efficiency. (ii) More importantly, our model is robust in both stability and plasticity since it is fully capable of learning important feature knowledge from both label classification and domain classification (synthetic *vs.* real) in each task. (iii) Our proposed method can not only perform well on a test dataset consisting of entirely real data but also elaborate to perform well on entirely synthetic test data and combined image groups (see Appendix), which better simulates the continual learning scenarios in real-world settings.

5 Experiment

5.1 Datasets and Evaluation Protocol

Datasets. To accurately and fairly evaluate our method in comparison with baselines, we use two representative datasets CIFAR100 [16] and ImageNet100 [12], which are widely adopted in CIL. CIFAR100 consists of 100 classes, each containing 500 training and 100 test images with the resolution $32 \times 32 \times 3$. ImageNet100 is a randomly sampled subset of ImageNet1000 [28], consisting of 100 classes each with 1300 training and 50 test images of various sizes.

Incremental Settings. Following prior works [7, 29, 33], for CIFAR100 and ImageNet100 datasets, we split the classes equally into $N = 5, 10,$ or 20 tasks (*e.g.*, each task has 5 classes if $N = 20$). For all approaches, we use the same random seed to randomly shuffle class orders for all datasets. Following previous works [7, 22, 29, 33, 53–55], the classification accuracy is defined as

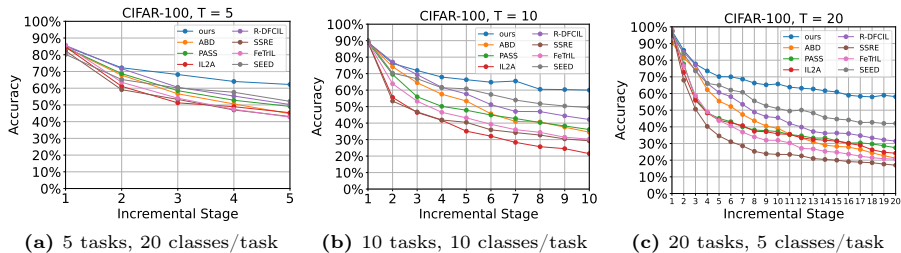
$$Acc_i = \frac{1}{|\mathcal{D}_{0:i+1}^{test}|} \sum_{(\mathbf{x}, y) \in \mathcal{D}_{0:i+1}^{test}} \mathbf{1}(\hat{y} = y), \text{ where } \hat{y} = \arg \max_{j \in \mathcal{C}_i} G_i^{(j)}(F_i(\mathbf{x})). \quad (10)$$

We report both the final accuracy from the last task Acc_L and the average incremental accuracy averaged over all incremental tasks $Acc_{avg} = \frac{1}{N} \sum_{i=0}^{N-1} Acc_i$.

Implementation Details. For a fair comparison, for CIFAR100, following previous works [7, 33], we use a modified 32-layer ResNet [10] as the backbone for all approaches. For our model, we train with SGD optimizer for 120 epochs. The learning rate is initially set to 0.1 with a decay factor of 0.1 after 100 epochs. The weight decay is set to 0.0002 and batch size of 128. For ImageNet100, we

Table 2: Evaluation results on CIFAR100 with protocol that equally split 100 classes into N tasks. The best results are in bold.

Approach	$N = 5$		$N = 10$		$N = 20$	
	Acc_{avg}	Acc_L	Acc_{avg}	Acc_L	Acc_{avg}	Acc_L
Upper Bound		70.67		70.67		70.67
ABD [33] (ICCV 2021)	60.78	44.74	54.00	34.48	43.32	21.18
PASS [54] (CVPR 2021)	63.31	49.11	52.01	36.08	41.84	27.45
IL2A [53] (NeurIPS 2021)	58.67	45.34	43.28	24.49	40.54	21.15
R-DFCIL [7] (ECCV 2022)	64.67	50.24	59.18	42.17	49.74	31.46
SSRE [55] (CVPR 2022)	56.96	43.05	43.41	29.25	31.07	16.99
FeTril [22] (WACV 2023)	58.68	42.67	47.14	30.28	37.25	20.62
SEED [29] (ICLR 2024)	63.05	52.14	62.04	51.42	57.42	42.87
Ours	69.77	62.21	68.05	58.40	67.10	57.11

**Fig. 4: Classification Accuracy of Each Incremental Task on CIFAR100.** Our method greatly outperforms all data-free CIL baselines in all incremental settings.

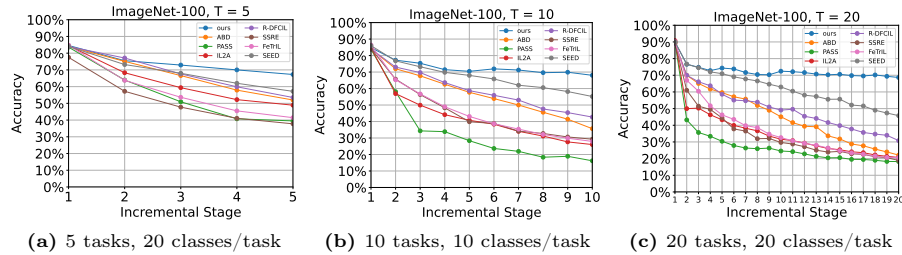
use ResNet18 [10] as the backbone for all methods. For our training, the SGD optimizer is adopted to train 40 epochs. The learning rate is initially set to 0.1 with a decay factor of 0.1 after 30 epochs. The weight decay is set to 0.0001 and batch size of 128. We train and report all methods from scratch with original implementations.

5.2 Results and Analysis

CIFAR100. We report the results of our method and SOTA exemplar-free CIL methods on CIFAR100 in Tab. 2. As observed, our method achieves the highest average and final accuracy among all approaches with non-marginal improvements. Moreover, as CIL becomes more difficult with a larger N (such as 20), the baselines suffer from significant accuracy drop (such as from 51.42% to 42.87% for SEED [29] when increasing N from 10 to 20), while our method still maintains high accuracy close to that of smaller N (such as our final accuracy from 58.4% to 57.11%) with larger improvements over baselines. Notably, compared with SOTA exemplar-free CIL method SEED(ICLR 2024) [29], when

Table 3: Evaluation on ImageNet100 with protocol that equally split 100 classes into N tasks.

Approach	$N = 5$		$N = 10$		$N = 20$	
	Acc_{avg}	Acc_L	Acc_{avg}	Acc_L	Acc_{avg}	Acc_L
Upper Bound		80.41		80.41		80.41
ABD [33] (ICCV 2021)	67.12	52.00	57.06	35.66	45.75	22.10
PASS [54] (CVPR 2021)	55.75	39.50	33.75	16.18	27.30	18.24
IL2A [53] (NeurIPS 2021)	62.66	48.91	43.46	26.04	35.59	20.72
R-DFCIL [7] (ECCV 2022)	68.42	53.50	59.36	42.70	49.99	30.80
SSRE [55] (CVPR 2022)	52.25	37.76	46.00	29.28	34.96	18.90
FeTril [22] (WACV 2023)	58.40	41.44	46.44	27.92	37.64	20.62
SEED [29] (ICLR 2024)	69.08	58.17	67.55	55.17	62.26	45.77
Ours	74.85	67.26	73.87	67.02	72.51	68.68

**Fig. 5: Incremental Accuracy on ImageNet100.** Our method greatly outperforms all baseline methods in all incremental settings. Our method achieves more significant improvements in more incremental task settings (e.g. increase N from 5 to 10 or to 20)

$N = 20$, our method is 9.68 percent more accurate for the average incremental accuracy Acc_{avg} and 14.24 percent more accurate for the final accuracy Acc_L .

We further present the detailed incremental accuracy of various learning phases for $N = 5, 10$, and 20 on CIFAR100 in Fig. 4. We observe that our curve drops significantly slower than all baseline methods with the highest accuracy at various phases, demonstrating our superior performance to mitigate the forgetting of previously learned knowledge over baseline methods.

ImageNet100. In Tab. 3, we present the results of our method and SOTA exemplar-free CIL methods on ImageNet100. Similarly, our method outperforms all baselines in terms of the average accuracy and final accuracy with non-marginal improvements. As CIL becomes more difficult with a larger N , the advantages or improvements of our method become more significant. Compared with SOTA exemplar-free CIL method seed [29], for $N = 20$, our method is 10.25 percent more accurate for Acc_{avg} and 22.91 percent more accurate for Acc_N .

Table 4: Ablation Study results of comparison between our method with all components and without the multi-distribution-matching diffusion model (MDM), without multi-domain adaptation reformation (MDA), and without selective synthetic image augmentation (SSIA). The ablation study is conducted on ImageNet100 with $N = 5$.

MDM	MDA	SSIA	Acc_{avg}	Acc_N
\times	\checkmark	\checkmark	59.71	51.17
\checkmark	\times	\checkmark	65.29	55.22
\checkmark	\checkmark	\times	62.37	52.94
\checkmark	\checkmark	\checkmark	74.85	67.26

The detailed incremental accuracy of various learning phases for $N = 5$, 10, and 20 on ImageNet100 are presented in Fig. 5. As observed, our method keeps the highest accuracy at almost all of the learning phases or stages. As it goes through more learning phases, our method can maintain almost consistent accuracy, outperforming baselines (which suffer from significant accuracy drops) with larger improvements. The results demonstrate that our method performs much better to mitigate the catastrophic forgetting problem in CIL.

5.3 Ablation Studies

We ablate the three major components in our method on ImageNet100 with $N = 5$. In each phase, 5 new classes are learned. We present our ablation results in Tab. 4. The results demonstrate that all proposed components contribute greatly. We further show that all three components are crucial to achieving better plasticity *vs.* stability balance through an ablation study in Fig. 6.

Multi-Distribution Matching (MDM). Without finetuning diffusion models with a multi-distribution matching technique, the average accuracy Acc_{avg} drops by 15.14 percent (74.85% *vs.* 59.71%), and the final classification accuracy Acc_N drops by 16.09 percent (67.26% *vs.* 51.17%). From Fig. 6, we also observe that MDM serves a crucial role in reviewing previous knowledges (*i.e.* stability).

Multi-Domain Adaptation (MDA). Without reforming exemplar-free CIL into a multi-domain adaptation problem, the average accuracy Acc_{avg} drops by 9.56 percent, and the final accuracy Acc_N drops by 12.04 percent. MDA also contributes to building model stability as shown in Fig. 6.

Selective Synthetic Image Augmentation (SSIA). Without further enhancement from selective synthetic image augmentation, the average accuracy Acc_{avg} drops by 12.48 percent, and the final accuracy Acc_L drops by 14.97 percent. Furthermore, Fig. 6 shows that SSIA helps the model not only learn new knowledge (*i.e.* plasticity) but also remember the knowledge from previous tasks.

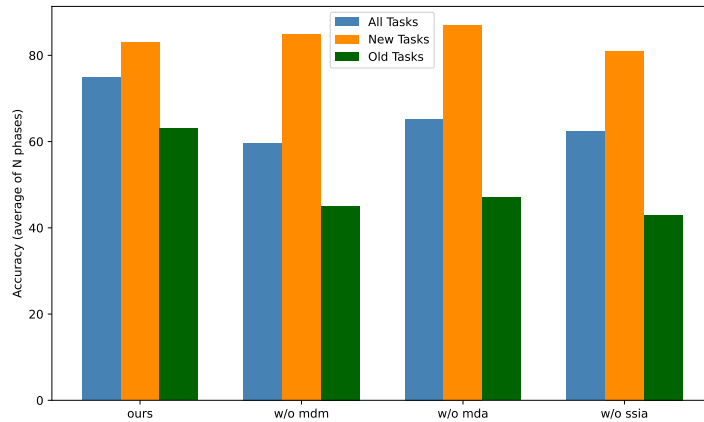


Fig. 6: Ablation Study about Stability-Plasticity Balance. Our method with all three components shows a better balance *vs.* w/o each of the three components.

6 Conclusion

In this paper, we introduce a novel exemplar-free CIL approach to address catastrophic forgetting and stability and plasticity imbalance caused by the domain gap between synthetic and real data. Specifically, our method generates synthetic data using multi-distribution matching (MDM) diffusion models to explicitly bridge the domain gap and unify quality among all training data. Selective synthetic image augmentation (SSIA) is also applied to enlarge training data distribution, enhancing the model’s plasticity and bolstering the efficacy of our method’s final component, multi-domain adaptation (MDA). With the proposed integrations, our method then reforms exemplar-free CIL to a multi-domain adaptation problem to implicitly address the domain gap problem during incremental training. Our method achieves state-of-the-art performance in various exemplar-free CIL settings on CIFAR100 and ImageNet100 benchmarks. In the ablation study, we proved that each component of our method is significant to best perform in exemplar-free CIL.

Limitations and Future Works One potential limitation of our method is the training time for each incremental task. In specific, the time to finetune a generative model using LoRA. This limitation is very common in exemplar-free methods that utilize synthetic data. In our case, we deduce in each incremental phase, the time to finetune an MDM diffusion model is proportional to the number of new classes to learn. In future work, we aim to explore strategies to streamline this process, thereby enhancing a shorter exemplar-free CIL training process.

Acknowledgement

This work is partially supported by the Army Research Office/Army Research Laboratory via grant W911-NF-20-1-0167 to Northeastern University, National Science Foundation CCF-1937500.

References

1. Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023.
2. Yulai Cong, Miaoyun Zhao, Jianqiao Li, Sijia Wang, and Lawrence Carin. Gan memory with no forgetting. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16481–16494. Curran Associates, Inc., 2020.
3. Arthur Douillard, Matthieu Cord, Charles Ollion, Thomas Robert, and Eduardo Valle. Podnet: Pooled outputs distillation for small-tasks incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.
4. Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
5. Victor Gallego. Personalizing text-to-image generation via aesthetic gradients. *arXiv preprint arXiv:2209.12330*, 2022.
6. Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *J. Mach. Learn. Res.*, 17(1):2096–2030, 2016.
7. Qiankun Gao, Chen Zhao, Bernard Ghanem, and Jian Zhang. R-dfcil: Relation-guided representation learning for data-free class incremental learning. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022.
8. Rui Gao and Weiwei Liu. Ddgr: continual learning with deep diffusion-based generative replay. In *Proceedings of the 40th International Conference on Machine Learning, ICML’23*. JMLR.org, 2023.
9. 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.
10. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
11. Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.
12. Saihui Hou, Xinyu Pan, Chen Change Loy, Zilei Wang, and Dahua Lin. Learning a unified classifier incrementally via rebalancing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
13. Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

14. Q. Jodelet, X. Liu, Y. Phua, and T. Murata. Class-incremental learning using diffusion model for distillation and replay. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 3417–3425, Los Alamitos, CA, USA, 2023. IEEE Computer Society.
15. Ronald Kemker and Christopher Kanan. Fearnert: Brain-inspired model for incremental learning. *arXiv preprint arXiv:1711.10563*, 2017.
16. Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report*, 2009.
17. Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.
18. Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in Neural Information Processing Systems*, 35:5775–5787, 2022.
19. Zichong Meng, Changdi Yang, Jun Liu, Hao Tang, Pu Zhao, and Yanzhi Wang. Instructgie: Towards generalizable image editing. *arXiv preprint arXiv:2403.05018*, 2024.
20. Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 18–24 Jul 2021.
21. Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. VAEs meet diffusion models: Efficient and high-fidelity generation. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.
22. Grégoire Petit, Adrian Popescu, Hugo Schindler, David Picard, and Bertrand Delezoide. Fetril: Feature translation for exemplar-free class-incremental learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3911–3920, January 2023.
23. Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.
24. Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
25. Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
26. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.
27. Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.
28. Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 2015.
29. Grzegorz Rypeść, Sebastian Cygert, Valeriya Khan, Tomasz Trzcinski, Bartosz Michał Zieliński, and Bartłomiej Twardowski. Divide and not forget: Ensemble

- of selectively trained experts in continual learning. In *The Twelfth International Conference on Learning Representations*, 2024.
30. Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.
 31. Flavio Schneider. Archisound: Audio generation with diffusion. *arXiv preprint arXiv:2301.13267*, 2023.
 32. Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.
 33. James Smith, Yen-Chang Hsu, Jonathan Balloch, Yilin Shen, Hongxia Jin, and Zsolt Kira. Always be dreaming: A new approach for data-free class-incremental learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
 34. Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.
 35. Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
 36. Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
 37. Brandon Trabucco, Kyle Doherty, Max A Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. In *The Twelfth International Conference on Learning Representations*, 2024.
 38. Fu-Yun Wang, Da-Wei Zhou, Liu Liu, Han-Jia Ye, Yatao Bian, De-Chuan Zhan, and Peilin Zhao. BEEF: Bi-compatible class-incremental learning via energy-based expansion and fusion. In *The Eleventh International Conference on Learning Representations*, 2023.
 39. Fu-Yun Wang, Da-Wei Zhou, Han-Jia Ye, and De-Chuan Zhan. Foster: Feature boosting and compression for class-incremental learning. In *European conference on computer vision*, pages 398–414. Springer, 2022.
 40. Zhendong Wang, Yifan Jiang, Yadong Lu, Pengcheng He, Weizhu Chen, Zhangyang Wang, Mingyuan Zhou, et al. In-context learning unlocked for diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.
 41. Zifeng Wang, Zheng Zhan, Yifan Gong, Yucai Shao, Stratis Ioannidis, Yanzhi Wang, and Jennifer Dy. Dualhsic: Hsic-bottleneck and alignment for continual learning. In *International Conference on Machine Learning*, pages 36578–36592. PMLR, 2023.
 42. Zifeng Wang, Zheng Zhan, Yifan Gong, Geng Yuan, Wei Niu, Tong Jian, Bin Ren, Stratis Ioannidis, Yanzhi Wang, and Jennifer Dy. SparCL: Sparse continual learning on the edge. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
 43. Chenshen Wu, Luis Herranz, Xialei Liu, yaxing wang, Joost van de Weijer, and Bogdan Raducanu. Memory replay gans: Learning to generate new categories without forgetting. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

44. Fei Ye and Adrian G Bors. Learning latent representations across multiple data domains using lifelong vaegan. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX 16*, pages 777–795. Springer, 2020.
45. Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724, 2020.
46. Jianhao Yuan, Jie Zhang, Shuyang Sun, Philip Torr, and Bo Zhao. Real-fake: Effective training data synthesis through distribution matching. In *The Twelfth International Conference on Learning Representations*, 2024.
47. Jie Zhang, Chen Chen, Weiming Zhuang, and Lingjuan Lyu. Target: Federated class-continual learning via exemplar-free distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4782–4793, 2023.
48. Jie Zhang, Xiaohua Qi, and Bo Zhao. Federated generative learning with foundation models. *arXiv preprint arXiv:2306.16064*, 2023.
49. Didi Zhu, Yinchuan Li, Yunfeng Shao, Jianye Hao, Fei Wu, Kun Kuang, Jun Xiao, and Chao Wu. Generalized universal domain adaptation with generative flow networks. In *ACM International Conference on Multimedia (MM) 2023*, 2023.
50. Didi Zhu, Yinchuan Li, Junkun Yuan, Zexi Li, Kun Kuang, and Chao Wu. Universal domain adaptation via compressive attention matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6974–6985, 2023.
51. Didi Zhu, Yinchuan Li, Min Zhang, Junkun Yuan, Jiashuo Liu, Kun Kuang, and Chao Wu. Bridging the gap: neural collapse inspired prompt tuning for generalization under class imbalance. *arXiv preprint arXiv:2306.15955*, 2023.
52. Didi Zhu, Zhongyi Sun, Zexi Li, Tao Shen, Ke Yan, Shouhong Ding, Kun Kuang, and Chao Wu. Model tailor: Mitigating catastrophic forgetting in multi-modal large language models. *arXiv preprint arXiv:2402.12048*, 2024.
53. Fei Zhu, Zhen Cheng, Xu-Yao Zhang, and Cheng-lin Liu. Class-incremental learning via dual augmentation. *Advances in Neural Information Processing Systems*, 34:14306–14318, 2021.
54. Fei Zhu, Xu-Yao Zhang, Chuang Wang, Fei Yin, and Cheng-Lin Liu. Prototype augmentation and self-supervision for incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5871–5880, June 2021.
55. Kai Zhu, Wei Zhai, Yang Cao, Jiebo Luo, and Zheng-Jun Zha. Self-sustaining representation expansion for non-exemplar class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9296–9305, 2022.