**Academic Journal of Science and Technology**
ISSN: 2771-3032 | Vol. 23, No. 2, 2026

# Prototype Compensation Method for Small-Scale Incremental Learning Based on Attention-Guided Sample Augmentation and Causal Constraints

**Jiamin Zhi, Kexin Zhang, Meiqi Zhao**

SIAS UNIVERSITY FACULTY OF ENGINEERING Zhengzhou, Henan 451100, China

**Abstract:** In small-shot, replay-free incremental learning scenarios, the scarcity of training samples can lead to weak model feature representation and insufficient causal K-nearest neighbor candidate feature pools, resulting in weight estimation bias and prototype drift, thereby exacerbating catastrophic forgetting. Existing causality-constrained methods have good stability, but they lack plasticity under small-shot conditions; ordinary data augmentation only expands the training samples and lacks causal representation constraints. To address this, this paper proposes a prototype compensation method that combines attention-guided sample augmentation with causal constraints. Attention-guided data augmentation (AGDA) is used to generate semantically invariant synthetic samples, expanding the causal K-nearest neighbor candidate pool to optimize weight estimation; combined with causal regularization to protect old class knowledge, supervised contrastive loss is used to optimize the feature space, and prototype drift is corrected through adversarial shift compensation. Experiments were conducted on 4-task and 8-task small-shot settings of Caltech 101 and Oxford Flowers 102. The results show that, compared to the contrastive baseline, this method improves average accuracy by 1.2%–3.3% and average incremental accuracy by 1.0%–3.0%, effectively reducing the forgetting rate, providing a feasible solution path for replay-free incremental learning with scarce samples.

**Keywords:** Small-start non-replay incremental learning; catastrophic forgetting; attention-guided sample augmentation; causal constraints; prototype drift compensation; stability-plasticity.

## 1. Introduction

Deep learning has achieved excellent results in tasks such as image classification and object recognition, but traditional deep learning models are built on the assumption that all training data are available in advance and independently and identically distributed (i.i.d.)[1]. In real-world applications, data often arrive in batches as a sequential data stream, with new classes constantly emerging, giving rise to class-incremental learning. Under the class-incremental learning setting, during the training phase, only the current task's data can be accessed, and all historical samples are unavailable; during the testing phase, the model needs to recognize all learned classes within a unified global class space. This setting is more aligned with practical applications in open environments but also brings significant challenges, namely the catastrophic forgetting problem[2]. When the model learns new classes, the parameters are updated under the influence of new task gradients, which disrupts the feature space and class prototypes formed during training on old tasks, causing a sharp decline in recognition performance on old classes.

The fundamental contradiction in class-incremental learning lies in the trade-off between stability and plasticity[3]. Stability represents the model's ability to retain knowledge of old classes and resist forgetting; plasticity represents the model's ability to fit new tasks and capture features of new classes. There is a natural trade-off between stability and plasticity: overemphasizing stability imposes strict constraints on model parameter updates, inhibiting the model from learning new knowledge and decreasing plasticity; fully releasing plasticity, allowing the model to significantly update parameters to adapt to new data, can destroy the feature representations of old knowledge, aggravating catastrophic forgetting[4]. Existing class-incremental learning techniques can be categorized into five major types: memory replay, parameter and optimization constraints, model prediction correction, model structure design, and pretrained model transfer.

1) Memory replay methods[5]: Use a sample buffer to save some historical old samples and train with a mixture of old and new samples to alleviate forgetting. However, there is the problem of storage overhead increasing with the number of tasks, and storing original samples also poses privacy risks; generative replay avoids storing samples, but the quality of generated samples is unstable and can easily introduce distribution bias.

2) Parameter and optimization constraint methods[6]: Represented by regularization, knowledge distillation, and gradient constraints, these are rehearsal-free approaches that do not require saving historical samples. However, most algorithms are designed for large bootstrapped scenarios; when transferred to small-scale scenarios, initial feature representations are incomplete, the association between old and new tasks is weak, and prototype drift worsens, causing a significant drop in algorithm performance.

3) Model prediction correction methods[7]: Focus on the classification output layer to correct prediction biases of new and old classes through logit recalibration, weight alignment, etc., with low deployment overhead. However, this only remedies the output side and cannot suppress feature representation drift at the root, and the correction effect noticeably degrades when feature space distortion is severe.

4) Model architecture design methods[8]: Rely on task masks or dynamically expand network modules to isolate new and old tasks, reducing parameter interference. The drawback is that the number of model parameters continuously grows with the task sequence, resulting in high inference overhead

in long-sequence incremental scenarios, which is unfavorable for deployment on resource-constrained devices.

5) Pre-trained model transfer methods[9]: Use pre-trained general prior knowledge combined with prompt learning and adapter fine-tuning to achieve incremental learning, showing remarkable performance under large data conditions. However, they heavily depend on pre-trained weights and are less applicable in scenarios with small datasets or lacking pre-trained models.

In recent years, causal reasoning has gradually been introduced into class-incremental learning research[10]. In small-start, no-replay class-incremental learning, the number of samples in the first task is limited, there is no historical sample buffer, and subsequent tasks arrive in batches, so the model cannot store the original samples of previous tasks[11]. Datasets have relatively few samples per class, such as the object recognition dataset Caltech 101 and the flower dataset Oxford Flowers 102. These real-world datasets naturally suffer from sample scarcity. From the perspective of causal graphs, the fundamental reason for forgetting in no-replay class-incremental learning is the collision causal structure formed by the composition of new and old data. Conditional independence occurs in the collision structure, blocking the information transmission path from old data to predictions, preventing old task knowledge from being passed to subsequent training stages. Causal effect extraction methods use causal interventions to break the collision structure conditioned on old features, utilize a K-nearest neighbor mechanism to calculate similarity weights, reproduce the causal contribution of old data to predictions, and achieve implicit replay without storing historical samples, significantly improving model stability. However, this method heavily depends on the richness of feature samples. In small-start scenarios, the number of samples available for neighbor searching is limited, coverage of the feature distribution is insufficient, and bias occurs in the estimation of causal effect weights, limiting the model's plasticity in learning new classes, and prototype drift remains significant.

Data augmentation is an important means to improve model plasticity[12]. Ordinary random data augmentation, such as random cropping, flipping, and color jitter, can easily destroy key discriminative semantics of images and introduce irrelevant noise; generative augmentation requires high training cost, and generated samples are at risk of semantic drift. Semantic-preserving sample generation for incremental tasks is a feasible approach to alleviating sample scarcity. The existing OCAGA method uses Grad-CAM to identify key visual areas of important categories in images, avoids altering the main object, and only modifies the background and non-critical regions to generate synthetic samples with unchanged category labels. However, this method is based on the premise of having a sample replay buffer, and the augmented samples are trained together with old samples from the buffer. In this paper, targeting no-replay, small-start constraints, no storage of old samples in a buffer exists; directly applying this augmentation strategy to merely expand the training set does not solve the issue of biased weight estimation caused by insufficient candidate features for causal KNN sampling in no-replay scenarios[13]. Simply relying on sample expansion only improves the model's ability to fit new classes but cannot block the intrinsic mechanism of forgetting from the causal representation level, and the model's stability still has shortcomings.

Observing the two technical approaches reveals that they have strong complementarity: causal effect extraction excels at ensuring stability from the perspective of causal-level representation but is limited by the sample size; attention-guided data augmentation excels at enhancing plasticity at the sample level but lacks causal representation constraints. Most existing studies use only one of these techniques, with few works combining them.

To address the stability-plasticity dilemma in small-start incremental learning without replay, this paper establishes a single core research focus: integrating causal effect extraction with attention-guided data augmentation to collaboratively balance stability and plasticity in class incremental learning from the dual dimensions of representation constraint and sample augmentation. The overall design is divided into four levels:

Sample level: Expand the K-nearest neighbor candidate sample pool to mitigate the small-start sampling deficiencies, and include the semantically consistent synthetic samples generated by attention-guided data augmentation (AGDA) in the causal effect extraction computation process. By using semantically consistent synthetic samples to expand the feature candidate set, feature distribution diversity is increased, addressing the issues of poor K-nearest neighbor sampling representativeness and inaccurate causal weight estimation caused by the insufficient number of original samples in small-start scenarios. Semantically consistent synthetic samples are only involved in model training, are not stored, and incur no additional memory overhead.

Representation constraint level: Causal regularization ensures model stability. Introduce a causal effect regularization loss to impose constraints on the parameter updates of the new feature extractor. By causal intervention, the colliding causal structure in the no-replay scenario is resolved, reproducing the causal contribution of old data to model predictions. In a no-replay framework without storing historical original samples, this suppresses the impact of new task gradients on the representation of old knowledge and guarantees the model's stability in remembering old classes.

Feature learning level: Enhance sample diversity to improve model plasticity. Semantically consistent synthetic samples generated by AGDA, together with original samples, participate in classification training, enriching the diversity of the training set, alleviating overfitting risks due to limited samples in small-start conditions, and improving the model's plasticity in fitting new classes. Supervised contrastive loss is also introduced, constructing positive and negative sample pairs from original samples and semantically consistent synthetic samples of the same class, optimizing the geometric structure of the feature embedding space, achieving intra-class feature aggregation and inter-class feature separation, further enhancing feature discrimination.

Prototype correction level: Adversarial shift compensation suppresses prototype drift. Adversarial samples are generated based on both original and semantically consistent synthetic samples to estimate the prototype drift vector between the new and old feature extractors, compensating and correcting the historical old class prototypes, mitigating prototype drift during incremental training, and further reducing catastrophic forgetting.

In summary, the modules described cooperate to simultaneously preserve old knowledge and learn new classes within a no-replay algorithm framework, achieving collaborative optimization of stability and plasticity. Experiments are conducted on the two publicly available

datasets Caltech 101 and Oxford Flowers 102 with limited sample sizes, using small-start no-replay experiments in 4-task and 8-task equal splits, verifying the effectiveness of the proposed method.

# 2. Related Work

## 2.1. Replay-free Incremental Learning Based on Causal Reasoning

Causal reasoning provides a brand-new theoretical analysis tool for analyzing the internal mechanisms of catastrophic forgetting in quasi-incremental learning. Unlike traditional studies that only focus on parameter updates and feature drift, causal reasoning reveals the underlying causes of forgetting from the perspective of causal relationships between variables, offering new solutions for non-replay-type incremental learning. Hu[14] et al proposed the approach of distilling causal effects from data. By constructing causal maps for quasi-incremental learning, they found that under the constraint of no playback, the training process cannot access historical old samples. Old data $D_{old}$ $D_{new}$ with newly arrived task data form a collision causal structure. The information transmission path of old data is blocked by collision variables, and old task knowledge cannot contribute causally to the predictive outputs of new models. This is the core incentive for catastrophic forgetting in scenarios without replay.

Based on the above theory, we introduce a causal effect extraction mechanism into small-start, replay-free incremental learning scenarios, build a complete algorithm framework, use K-nearest neighbor weighting for implicit replay of old knowledge, and rely on adversarial samples to handle prototype drift compensation. This framework validates the stability gain of causal interventions, but all sampling relies solely on the original task sample set. When samples are scarce, the candidate feature pool for K-neighbor retrieval is limited, weight estimation introduces bias, and plasticity is constrained. To address this limitation, this paper introduces attention-guided generated semantic-preserving synthetic samples to expand the candidate sample pool for KNN.

Although this type of causality-driven replay-free method has achieved decent results, its performance heavily depends on the richness of the training sample features. Under the strict conditions of small-start equal splits, it still has obvious shortcomings[15]. In the small-start setup, the first task has few classes and a limited overall sample size, and each subsequent incremental task also provides too few feature samples. This limits the size of the candidate feature pool available for K-nearest neighbor (K-NN) retrieval, and the representativeness of the retrieved neighbors drops, directly causing bias in estimating the causal effect weights $W_i$. Distorted weight estimates reduce the actual effect of the causal regularization: on one hand, the implicit replay of old knowledge is weakened, and on the other hand, it interferes with fitting features of new classes. This ultimately shows up as insufficient model plasticity and a hindered ability to learn new classes[16]. Meanwhile, the adversarial prototype drift compensation module in this framework also relies on sample quality: when samples are scarce, the quality of generated adversarial samples drops, introducing extra error in estimating prototype drift and weakening prototype correction, which further limits the overall performance of the model in small-start scenarios. Therefore, this paper introduces AGDA to augment samples and make up for the lack of K-NN samples.

## 2.2. Attention-Guided Data Augmentation Techniques

In the class-incremental learning paradigm with a replay sample buffer, the limited capacity of the sample buffer can lead to an insufficient number of reusable old samples. When repeatedly training on a small number of samples, the model is prone to overfitting, which reduces feature discriminability and causes feature entanglement across different classes. To address this issue, To tackle this pain point, some researchers proposed the OCAGA framework and designed a set of attention-driven sample generation strategies. This strategy uses Grad CAM to output class heatmaps to locate the discriminative parts of an image, then applies max pooling to broaden the heatmap response region, and constructs a reverse mask that only modifies the non-discriminative background areas of the image, generating synthetic samples while keeping class labels unchanged. The samples generated this way don't need to be written into the replay buffer; they only participate in the network's forward and backward passes, avoiding storage pressure and enriching the diversity of training samples. It's important to note that the entire OCAGA mechanism is designed for rehearsal-based incremental learning, where augmented samples help the old samples in the buffer for classification training. In contrast, the scenario in this paper is small-start class incremental learning without replay—there's no old-sample buffer, and the semantic-preserved synthetic samples are not simply used for classification training. The key improvement here is that the generated semantic-preserved samples are fed into the candidate feature set for causal K NN retrieval, helping improve causal weight estimation. The augmentation method in that paper lacks causal representation constraints and can't address the representation forgetting problem caused by conflicting causal structures under no-replay conditions, which is why this paper adds a causal constraint module.

## 2.3. Class-incremental learning stability-plasticity balance

The stability-plasticity dilemma is a core underlying issue in the field of continual learning. Existing solutions can be divided into three categories: 1) Stability-biased: Methods like regularization, knowledge distillation, and causal intervention protect old knowledge by constraining parameter updates, but they tend to suppress the model's ability to learn new classes, impairing plasticity[17]; 2) Plasticity-biased: Replay, data augmentation, and model expansion enrich the training data or increase model capacity, enhancing the ability to learn new classes, but they bring storage and computational costs and lack protection for old representations, resulting in insufficient stability[18]; 3) Collaborative balance schemes: These take both into account and are the current research hotspot, but most existing collaborative schemes are replay-distillation or structure expansion-regularization, and fusion schemes integrating causal reasoning with semantic-preserving data augmentation have not yet been thoroughly studied. Extracting causal effects can ensure stability, while attention-guided augmentation can enhance plasticity; the two are naturally complementary. This paper combines them, using semantic-preserving synthetic samples in the nearest-neighbor sampling for causal effect computation, which both expands the training samples and improves the quality of

causal effect estimation, achieving reciprocal benefits. Few existing works combine causal representation constraints with semantic-preserving data augmentation, and this is the starting point of this paper.。

# 3. A Class-Incremental Learning Method Enhanced by Fusion of Causal Effect Extraction and Attention Guidance

This paper proposes a small-shot class incremental learning prototype compensation method based on attention-guided sample augmentation and causal constraints. The overall model is divided into a feature extraction module and a prototype classification module. The core idea is to introduce semantically consistent synthetic samples generated via attention guidance into the causal constraint computation process, expanding the K-nearest neighbor candidate sample pool to compensate for the sampling shortcomings caused by the limited number of original samples in small-shot scenarios. The overall workflow of the model is as follows: 1) Once new task data arrives, attention-guided sample augmentation is performed to generate semantically consistent synthetic samples; 2) Original samples and AGDA-generated semantic-preserving synthetic samples are fed into the frozen old feature extractor for feature extraction; 3) The set of original samples and semantic-preserving synthetic samples together form the candidate sample pool, K-nearest neighbor retrieval is carried out, and similarity weights are calculated to construct a causal constraint loss for parameter regulation of the new feature extractor; 4) Supervised contrastive loss is introduced, with original samples and same-class semantic-preserving synthetic samples forming positive and negative pairs, optimizing the feature embedding space to achieve intra-class cohesion and inter-class separation; 5) The prototype classification module combines original samples and semantic-preserving synthetic samples to generate adversarial samples, estimate the prototype drift vector, and perform prototype compensation, ultimately using the NCM nearest class mean classifier for class prediction.

## 3.1. Attention-Driven Sample Augmentation for Causal Candidate Sets

In the small-start, no-replay scenario discussed in this paper, there is no historical sample buffer, and each incremental task can access only a limited amount of samples. On one hand, training directly on limited samples is prone to overfitting; on the other hand, when the subsequent causal constraint module performs K-nearest neighbor retrieval, if the number of candidate features is too small, the estimated results of similarity weights will show significant bias. To address this, this paper draws on the idea of attention-driven sample generation and designs a sample augmentation process suitable for the no-replay setting.

To prevent the new network parameters from shifting the old class discriminative semantics during iteration, the heatmap calculation uses the frozen feature extractor of the old task throughout, executing Grad CAM to obtain the original class activation heatmap $M_0$. The original heatmap often focuses only on a very small discriminative region of the object, which can result in insufficient mask coverage. Therefore, max pooling is used to expand the heatmap, obtaining the expanded mask $\widetilde{M}_0$. Based on this mask, a reverse mask $1-\widetilde{M}_0$ is constructed: the pixels of the original object's discriminative region are kept unchanged, while random geometric transformations are applied only to the non-discriminative background areas marked by the mask, synthesizing a new perturbed sample $\hat{x}$, with the sample label remaining the same as the original image. The formula for sample synthesis is as follows：

$$\hat{x} = (x_0 \odot M_0) + (x_n \odot (1 - M_0))$$

In the formula: $\odot$ represents the Hadamard element-wise product; $x_0$ is the original input task sample; $x_n$ is the sample after random geometric transformation; $\hat{x}$ is the final semantically preserved synthetic sample, and the semantically preserved synthetic sample maintains the exact same class label as the original sample.

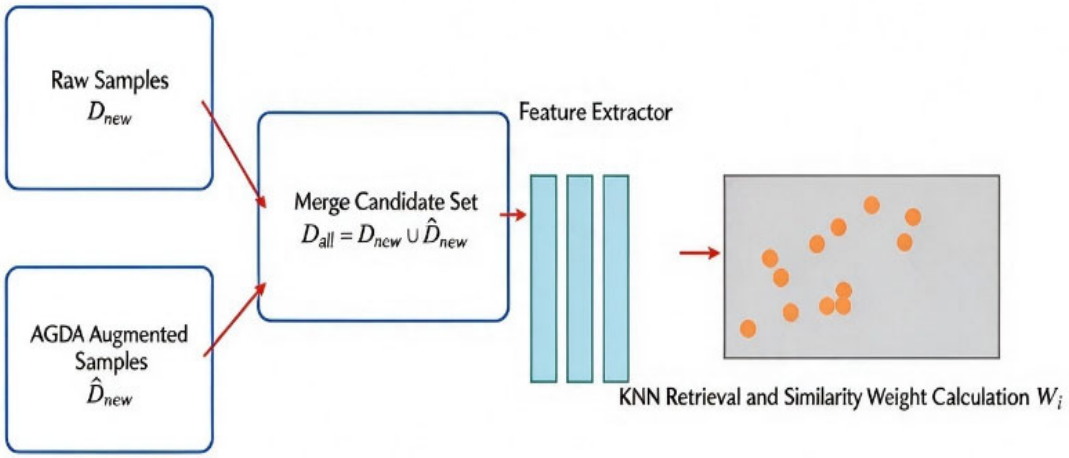



**Figure 2-1.** Schematic of the Attention-Guided Sample Augmentation Module

The existing OCAGA method is a replay framework where semantically preserved synthetic samples are used together with old samples from the sample cache for classification training. The application scenario in this paper is completely different, with no replay sample cache; therefore, the use of semantically preserved synthetic samples is changed: the original task sample set $D_{new}$ and the semantically preserved synthetic sample set $\widehat{D}_{new}$ are merged to form a joint candidate feature set $D_{all} = D_{new} \cup \widehat{D}_{new}$, which is directly supplied to the downstream causal constraint module for K-nearest neighbor feature retrieval. By leveraging the expanded feature set, the sampling bias caused by the scarcity

of candidate features under small-start conditions is alleviated, improving the robustness of the similarity weight $W_i$ estimation.

## 3.2. Causal Effect Extraction of Fused Augmented Samples

Analyzing the forgetting mechanism of replay-free incremental learning from the perspective of causal graph theory: When historical old samples cannot be accessed, the old data $D_{old}$ and the new task data $D_{new}$ form a collision causal structure. The colliding variable blocks the causal path from the old data to the model's prediction $Y_{new}$, preventing knowledge from old tasks from influencing the new model's output, thereby triggering catastrophic forgetting. The core idea of causal intervention is to condition on the old features $X_{old}$ to remove the collision constraint, re-establish the causal contribution of the old data to the model's predictions, achieve sample-independent implicit replay, and ensure model stability. The existing causal effect extraction methods only perform K-nearest neighbor sampling using the raw new task samples $D_{new}$, with the corresponding causal effect calculation expression being:

$$Effect_D = \sum_{i\in\{N_1,\ ,N_K\}} P(Y_{new}|D_{new})\, W_i$$

Where: $N_1, N_2,\ , N_K$ represent the indices of the K samples in the old feature space that are closest to the query sample features; $W_i$ is the similarity weight calculated based on feature distance; $P(Y_{new}|D_{new})$ is the predicted output probability of the sample under the new feature extractor.

This paper incorporates the semantic-preserving synthetic sample set $\widehat{D}_{new}$ generated by the AGDA module into the candidate sample pool, constructing a joint candidate set $D_{all} = D_{new} \cup \widehat{D}_{new}$ that combines the original samples with the semantic-preserving synthetic samples, thereby expanding the feature space for nearest neighbor search. The improved causal effect calculation formula is as follows:

$$Effect_D = \sum_{i\in\{N_1,\ ,N_K\}} P(Y_{new}|D_{all})\, W_i$$

Semantic-preserving synthetic samples provide more differentiated feature distributions, reducing the sampling bias caused by the scarcity of samples in small-scale initialization scenarios and making the estimation of similarity weights $W_i$ more robust. The computed causal effects are further used to construct a causal regularization loss $L_{causal}$, which is incorporated as a constraint into the overall loss function, imposing constraints on the parameter updates of the new feature extractor. This preserves the causal impact of previous tasks on the model's predictions without storing any historical original samples, suppresses the destruction of old knowledge representations by new task gradients, and maintains model stability.

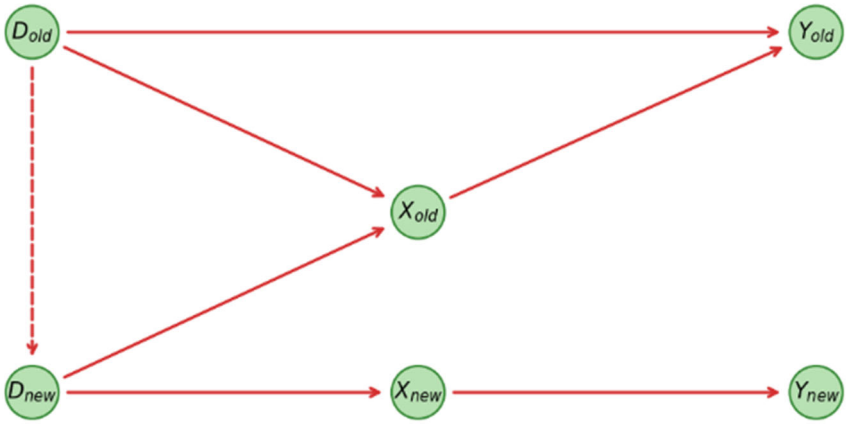



**Figure 2-2.** Schematic Diagram of the Causal Effect Extraction Module for Fusion-Enhanced Samples

At the same time, a supervised contrastive loss (SupCon) is introduced to optimize the geometric structure of the feature space. Original samples and AGDA semantic-preserving synthetic samples of the same category are treated as positive sample pairs, while samples of different categories are treated as negative sample pairs. Through the constraints of the loss function, features of the same class are clustered together in the embedding space, while features of different classes are separated, further enhancing the discriminative ability of the feature representation. The process of optimizing the feature space with supervised contrastive loss is shown in Figure 2-3, achieving intra-class aggregation and inter-class separation by constructing positive and negative sample pairs.

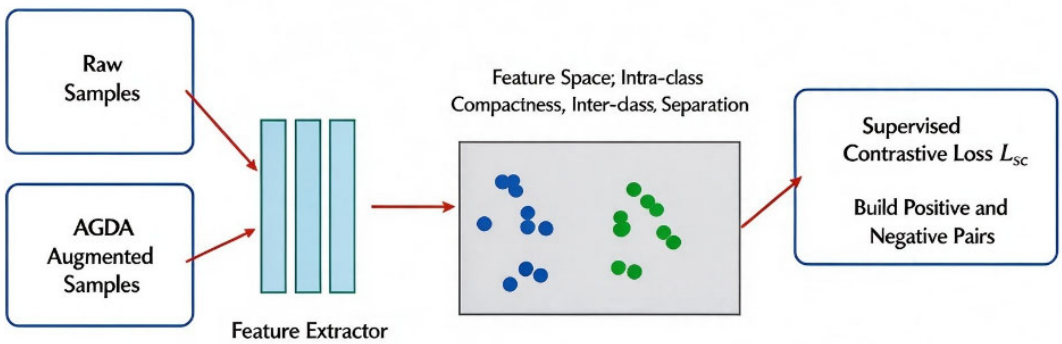



**Figure 2-3.** Schematic diagram of the contrast feature optimization module

## 3.3. Counteracting prototype drift compensation

After the feature extraction network is updated, shifts in the feature space can cause category prototype drift, resulting in performance degradation of the NCM nearest class mean classifier. To address this, an adversarial prototype drift compensation strategy is introduced in the prototype classification branch, using both the original samples and AGDA semantic-preserved synthetic samples to jointly generate an adversarial sample set $X_{adv}^{k}$. The adversarial samples are then fed into the frozen old feature extractor $f_{t-1}$ (with parameters fixed after the previous task training) and the current new feature extractor $f_t$ to be updated. By comparing the differences in the features output by the adversarial samples under the two networks, the prototype drift amount for category k is quantified.

$$\triangle_{t-1\to t}^{k}=\frac{1}{|X_{adv}^{k}|}\sum_{x_{adv}\in X_{adv}^{k}}(f_t(x_{adv})-f_{t-1}(x_{adv}))$$

In the formula: $X_{adv}^{k}$ represents the set of adversarial samples belonging to class k; $|X_{adv}^{k}|$ is the number of samples in this set; $\triangle_{t-1\to t}^{k}$ denotes the overall drift vector of the prototype of class k from task t-1 to task t. Based on the obtained drift vector, a compensatory update is applied to the previously saved old class prototype $P_{t-1}^{k}$, resulting in a class prototype adapted to the current feature space.

$$p_t^k = P_{t-1}^k + \triangle_{t-1\to t}^k$$

The prototype $p_t^k$ after compensation is used for subsequent NCM nearest class mean classification. This strategy estimates the overall shift in the feature space through adversarial samples, corrects the positions of class prototypes, mitigates the prototype drift problem during the incremental iteration process, and improves the recognition accuracy of old and new classes in the classification stage. The calculation process of the adversarial prototype drift compensation module is shown in Figure 2-4, which completes the compensation of historical class prototypes based on the drift vector estimated between the old and new feature extractors using adversarial samples.

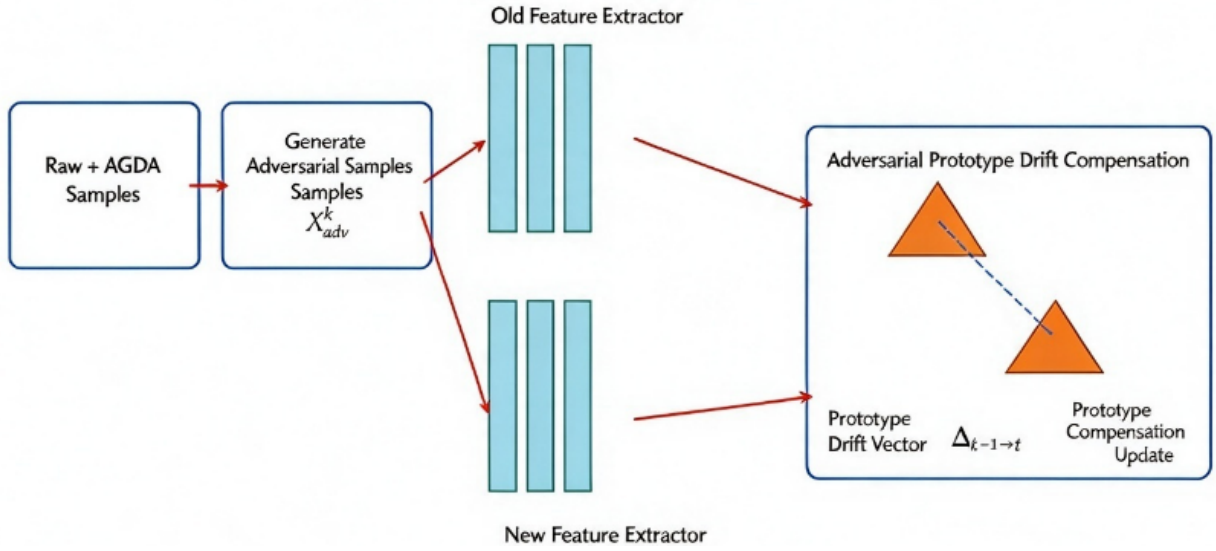



**Figure 2-4.** Schematic diagram of the prototype drift compensation module

## 3.4. Total loss function

To collaboratively achieve joint optimization of stability and plasticity, the total loss in this work is composed of three weighted components: the cross-entropy classification loss $L_{ce}$, the supervised contrastive loss $L_{sc}$, and the causal regularization loss $L_{causal}$. Among them, the cross-entropy classification loss is responsible for driving the model to learn the class labels of the original samples and the AGDA semantically preserved synthesized samples, ensuring the model's ability to fit new classes and maintaining model plasticity; the supervised contrastive loss is used to optimize the geometric distribution of the embedding space, enhancing feature discriminability; the causal regularization loss, as a constraint at the representation level, preserves the causal contribution of old tasks to the model's predictions, suppresses the destruction of old knowledge representations by gradients from new tasks, and guarantees model stability. The expression for the total loss is defined as:

$$L_{total} = L_{ce} + \lambda_1 L_{sc} + \lambda_2 L_{causal}$$

$L_{ce}$ is the cross-entropy classification loss. The original sample set $D_{new}$ and the AGDA-generated semantic-preserving synthetic sample set $\widehat{D}_{new}$ jointly participate in the loss calculation, constraining the model output to remain consistent with the true class labels, fully utilizing augmented samples to enhance learning ability for new classes;

$L_{sc}$ is the supervised contrastive loss (SupCon), where positive sample pairs are constructed using original samples and semantic-preserving synthetic samples of the same class, and negative sample pairs are constructed using samples from different classes, bringing features of the same class closer and pushing features of different classes apart, optimizing the feature embedding space;

$L_{causal}$ is the causal regularization loss, constructed from the improved causal effect $Effect_D$, used to constrain the update of new feature extractor parameters and reproduce the causal effects of old data under no-replay conditions;

$\lambda_1$ and $\lambda_2$ are non-negative hyperparameters that control the weights of the supervised contrastive loss and causal regularization loss in the total loss, balancing the impact of the three losses on model updates.

1) When $\lambda_1$ = 0, the model no longer performs the supervised contrastive constraint, causing feature space geometric optimization to fail;

2) When $\lambda_2 = 0$, the causal regularization constraint is disabled, and the model degenerates into a baseline model enhanced only by AGDA, losing the stability guarantee from causal effects.

During the training phase, all samples $D_{all} = D_{new} \cup \widehat{D}_{new}$ are input into the network, and the parameters of the new feature extractor are updated by minimizing the total loss $L_{total}$. After training, the compensated and corrected prototypes for each class are saved. In the inference phase, only the NCM nearest class mean classifier is used for class prediction; semantic-preserving synthetic samples are not generated during inference, avoiding additional computational overhead.

Integrating all the above modules, the complete algorithm and overall research framework of this paper are shown in Figure 2-5, sequentially undergoing AGDA sample augmentation, causal feature selection, contrastive feature optimization, and adversarial prototype correction, completing small-scale rehearsal-free class incremental learning.

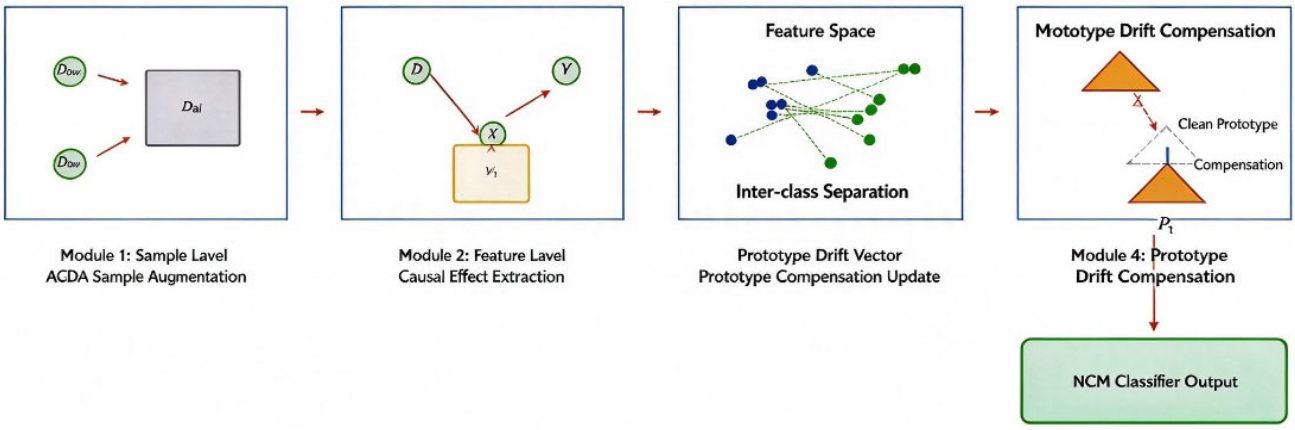



**Figure 2-5.** Overall Research Framework of the Method in This Paper

# 4. Experimental Research

## 4.1. Experimental Setup

### 4.1.1. Dataset and Task Division

This article selects two image datasets with limited sample sizes, Caltech 101 and Oxford Flowers 102, to conduct small-startup, replay-free class-incremental learning experiments [19]. Caltech 101 contains 101 object categories, each with a limited number of samples, with 40–80 images per category; Oxford Flowers 102 contains 102 flower image categories, each with only 40 samples. Both datasets are naturally suitable for small-startup experimental scenarios. All images are uniformly preprocessed and scaled to 224×224 pixels, tasks are divided equally, and the model does not store any historical raw samples at any time, strictly following the replay-free constraint.

1) Caltech 101: In the 4-task division, the first 3 tasks have 25 categories each, and the 4th task has 26 categories; in the 8-task division, the first 7 tasks have 12 categories each, and the 8th task has 17 categories.

2) Oxford Flowers 102: In the 4-task division, the first 3 tasks have 25 categories each, and the 4th task has 27 categories; in the 8-task division, the first 7 tasks have 12 categories each, and the 8th task has 18 categories.

### 4.1.2. Network and Hyperparameter Settings

The backbone network is ResNet 18, and the experimental framework is PyTorch; batch size batch-size = 32; base training epochs = 80; SGD optimizer with momentum 0.9 and weight decay $5 \times 10^{-4}$; initial learning rate 0.01, using a step decay strategy. In the causal effect module, the number of neighbors K = 10; adversarial perturbation step size α = 20, number of attack iterations i = 3; loss balance hyperparameters $\lambda_1 = 0.3$, $\lambda_2 = 0.4$. The method proposed in this paper belongs to the replay-free class of incremental learning and does not store any historical raw samples.

### 4.1.3. Evaluation metric

This paper quantifies the model's performance using three commonly used metrics in the field of class incremental learning: Average Accuracy (AA), Average Incremental Accuracy (AIA), and Average Forgetting (F).

Let the total number of tasks be T. After the model finishes training on the t-th incremental task, the classification accuracy on the test set of the k-th task is denoted as $a_{t,k}$, where $1 \le k \le t \le T$.

1) Average Accuracy (AA)[20]: Average accuracy is the core metric in incremental learning. It represents the mean accuracy of the model on the test sets of all learned classes after finishing the first t tasks, reflecting the overall classification performance of the model.

$$AA_t = \frac{1}{t}\sum_{k=1}^{t} a_{t,k}$$

When all T tasks are finished training, the final average accuracy:

$$AA = AA_T = \frac{1}{T}\sum_{k=1}^{T} a_{T,k}$$

2) Average Incremental Accuracy (AIA)[21]: Average Incremental Accuracy measures the model's average recognition accuracy for all seen classes after learning each new task, focusing on evaluating the model's ability to continually learn new classes (plasticity):

$$AIA = \frac{1}{T}\sum_{t=1}^{T}(\frac{1}{t}\sum_{k=1}^{t} a_{t,k})$$

3) Average Forgetting Rate F[22]: The forgetting rate is used to measure how much the performance on old classes has dropped, reflecting the model's stability. $a_{k,k}$ represents the highest accuracy of the model on the test set of task k right after finishing learning task k; $a_{T,k}$ represents the model's accuracy on the test set of task k after training on all tasks. The forgetting rate of task k:

$$f_k = a_{k,k} - a_{T,k}$$

Overall average forgetting rate:

$$F = \frac{1}{T-1}\sum_{k=1}^{T-1} f_k$$

The higher the forgetting rate F, the more severe the forgetting of old knowledge; F=0 means no catastrophic forgetting.

Here, the average accuracy (AA) is used to measure a model's overall recognition ability across all historical classes after learning all incremental tasks; average incremental accuracy (AIA) evaluates the model's continuous learning ability throughout the incremental learning process, focusing on its plasticity; average forgetting rate (F) quantifies how much the performance on old classes drops during subsequent task training, which reflects the model's stability.

#### 4.1.4. Comparing with baseline methods

Compare multiple representative rehearsal-free incremental learning baselines:

Finetune: directly fine-tune, no anti-forgetting strategy;

EWC[23]: Elastic Weight Consolidation, a parameter regularization method;

LwF[24]: Learning without Forgetting, based on knowledge distillation;

WA[25]: Weight Alignment method;

ADC[26]: Adversarial Drift Compensation prototype incremental method;

Causal only: only uses causal effect extraction method;

AGDA only: only uses attention-guided data augmentation method;

IID[27]: joint training, all data trained simultaneously, theoretical performance upper bound。

## 4.2. Experimental results

#### 4.2.1. Calteh 101 Dataset Experiment Results

**Table 3-1.** Calteh 101 Dataset Experiment Results (Small Initialization Setting, Unit: %)

| Method | 4 Task Division | | | 8 Task Assignments | | |
|---|---|---|---|---|---|---|
| | AA | AIA | F | AA | AIA | F |
| IID | 74.22 | 74.22 | 0.00 | 74.22 | 74.22 | 0.00 |
| Finetune | 19.43 | 35.17 | 72.41 | 11.26 | 26.84 | 79.53 |
| EWC | 20.87 | 36.72 | 69.04 | 13.11 | 29.71 | 75.62 |
| LWF | 31.64 | 50.23 | 54.82 | 24.37 | 41.25 | 62.48 |
| WA | 56.32 | 65.47 | 27.93 | 45.18 | 59.34 | 34.76 |
| ADC | 57.26 | 66.35 | 26.47 | 46.23 | 60.18 | 33.14 |
| Causal-only | 59.41 | 68.42 | 23.86 | 48.35 | 61.92 | 30.42 |
| AGDA-only | 58.14 | 67.24 | 25.41 | 47.16 | 60.77 | 32.08 |
| Ours | 62.14 | 70.95 | 20.12 | 51.47 | 64.36 | 26.65 |

Result Analysis: 1) IID serves as the performance upper bound; Finetune, EWC, and LwF perform poorly in the small-start scenario of Caltech-101, with severe catastrophic forgetting; 2) Causal only (only causal effects) achieves decent stability, but is limited by the number of samples, with a bottleneck in plasticity; AGDA only (only attention enhancement) improves the ability to learn new classes, but its forgetting rate is higher than our method; 3) Our combined method outperforms all baselines in both AA and AIA metrics under the 4-task and 8-task setups, while also achieving the lowest average forgetting rate F. Under the 4-task split, compared to Causal only, the average accuracy increases by 2.73% and the average forgetting rate decreases by 3.74%; under the 8-task split, the average accuracy increases by 3.12%. The accuracy curves of different methods as tasks progress are shown in Figures 3.1 and 3.2 in Section 3.5.

#### 4.2.2. Experimental results of the Oxford Flowers 102 dataset

**Table 3-2.** Experimental results on the Oxford Flowers 102 dataset (small initialization setup, unit: %)

| Method | 4 Task Division | | | 8 Task Division | | |
|---|---|---|---|---|---|---|
| | AA | AIA | F | AA | AIA | F |
| IID | 52.74 | 52.74 | 0.00 | 52.74 | 52.74 | 0.00 |
| Finetune | 12.37 | 21.26 | 75.84 | 9.14 | 18.73 | 81.47 |
| EWC | 17.42 | 32.08 | 66.13 | 11.26 | 24.82 | 73.25 |
| LWF | 21.73 | 33.94 | 62.45 | 18.85 | 27.16 | 68.33 |
| WA | 36.94 | 46.31 | 34.87 | 29.26 | 41.74 | 40.15 |
| ADC | 39.27 | 48.63 | 30.74 | 31.83 | 42.96 | 36.82 |
| Causal-only | 40.46 | 50.37 | 28.53 | 33.24 | 44.27 | 34.47 |
| AGDA-only | 39.52 | 49.12 | 30.06 | 32.37 | 43.41 | 35.73 |
| Ours | 43.18 | 52.84 | 24.21 | 36.42 | 46.93 | 30.64 |

Result Analysis: The Oxford Flowers 102 dataset has fewer samples per category, so the overall model recognition accuracy drops compared to Caltech 101. However, our method still outperforms the baseline across the board. For the 4-task split, the average accuracy improves by 2.72% and the average forgetting rate decreases by 4.32% compared to Causal only; for the 8-task split, the average accuracy goes up by 3.18%. This shows that on a flower dataset with even fewer samples, attention-guided sample augmentation combined with causal constraint strategies is still effective. The accuracy change curves for each task can be seen in Figures 3-3 and 3-4 in Section 3.5.

## 4.3. Ablation experiment

The ablation experiments were carried out on the Caltech 101 dataset under the 4-task split to verify the synergistic benefits of the attention-guided sample augmentation module and the causal constraint module[28], with results shown in

Table 3-3.

**Table 3-3.** Ablation experiment results (Caltech 101, 4 tasks, unit: %)

| Causal Effect Extraction | Attention-Guided Enhancement | AA | AIA | F |
|---|---|---|---|---|
| — | — | 19.43 | 35.17 | 72.41 |
| √ | — | 59.41 | 68.42 | 23.86 |
| — | √ | 58.27 | 67.24 | 25.41 |
| √ | √ | 62.14 | 70.95 | 20.12 |

Ablation analysis: When both modules are disabled, the model suffers from severe catastrophic forgetting; enabling only the causal constraint improves model stability, but it's limited by the original sample size and lacks plasticity; enabling only attention-guided sample augmentation increases sample diversity, but without the causal representation constraint, stability is weaker than with the causal constraint alone. When both modules are enabled, the model achieves the highest average accuracy (AA), the highest average incremental accuracy (AIA), and the lowest average forgetting rate (F). Experiments show that attention-guided sample augmentation and causal constraints have a synergistic effect, confirming the rationale of the core improvements proposed in this paper.

## 4.4. Supplementary parameter sensitivity experiment (K-value ablation)

In causal effect extraction, K represents the number of K-nearest neighbor samples, and the choice of K directly affects the quality of causal similarity weight estimation. In this section, based on Caltech 101 (4 task splits), we compare the average precision changes of Causal only and our method under different K values.

**Table 3-4.** Ablation experiment of K value parameter (Caltech 101, 4 tasks, unit: %)

| K | 5 | 8 | 10 | 12 | 15 | 20 |
|---|---|---|---|---|---|---|
| Causal-only | 57.86 | 58.53 | 59.41 | 58.97 | 58.04 | 56.31 |
| Ours | 60.27 | 61.42 | 62.14 | 61.58 | 60.73 | 58.86 |

Analysis: The best performance for the Causal only baseline appears at K=10. After introducing attention-guided sample augmentation, our method outperforms the Causal only baseline across all K values; the optimal K is still 10. At the same time, the overall performance curve is smoother, which shows that semantic-preserving synthetic sample augmentation enriches the pool of causal candidate samples, improving the model's robustness to neighboring K values and reducing reliance on hyperparameter tuning in low-data scenarios.

## 4.5. Visualization analysis of experimental results

In the previous section, we used tables to complete the quantitative statistics for each dataset, ablation studies, and parameter sensitivity. In this section, we use curves, bar charts, feature dimensionality reduction visualizations, and radar charts to intuitively analyze the effects of our method from the angles of trends, module gains, feature space, and multidimensional overall performance.

For the Caltech 101 dataset under the 4-task and 8-task splits, the accuracy changes of each baseline method as incremental tasks progress are shown in Figures 3-1 and 3-2. As tasks keep increasing, the accuracy of all methods gradually decreases. The red curve representing our method stays at the top throughout and has the smallest decline slope, indicating that when continually learning new classes, the performance decay of old classes is slower, allowing it to balance learning new classes and remembering old ones.

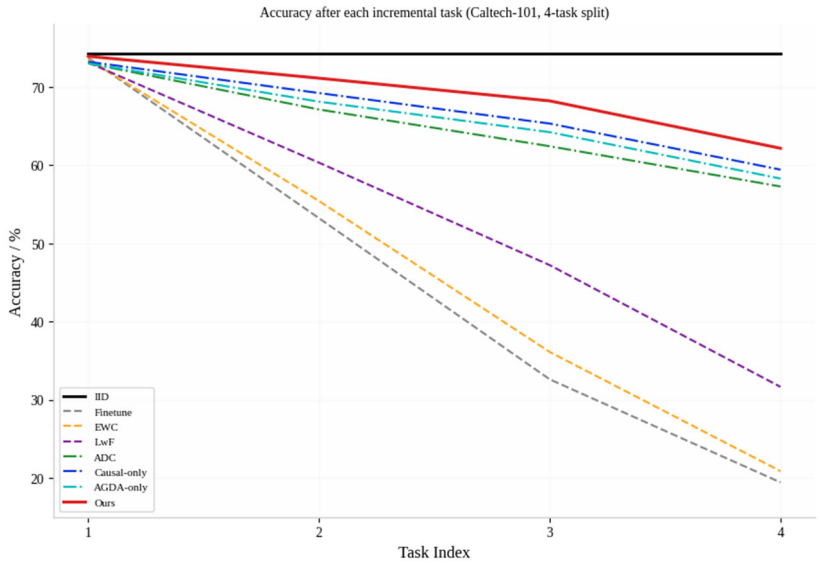



**Figure 3-1.** Accuracy changes at each incremental task stage for Caltech 101 (4-task split)

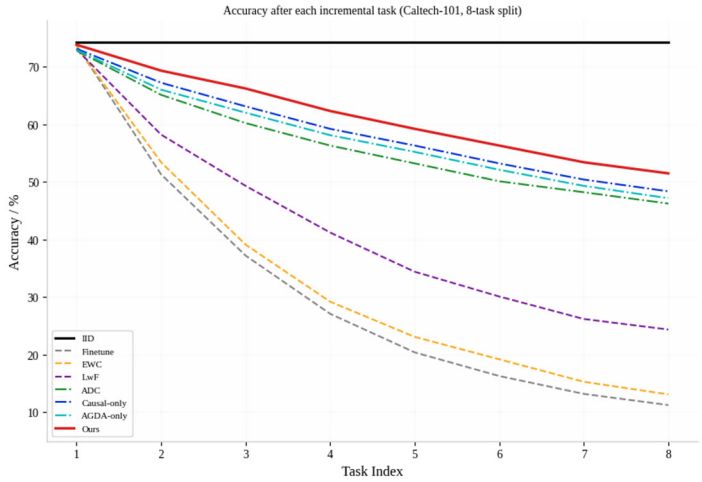



**Figure 3-2.** Accuracy changes at each incremental task stage for Caltech 101 (8-task split)

According to the results in Figures 3-1 and 3-2, we can see that as the number of tasks increases, forgetting gets worse. The method in this paper has higher accuracy across all task nodes compared to the other baselines, and the performance drop is significantly smaller than that of the Causal only and AGDA only baselines.

On the Oxford Flowers 102 dataset, where samples are even more scarce, the experimental curves for the 4-task and 8-task splits are shown in Figures 3.3 and 3.4. Even when the number of samples per class is very low, our method still outperforms the baselines at all task points, demonstrating that the fusion strategy generalizes well on small-sample datasets.

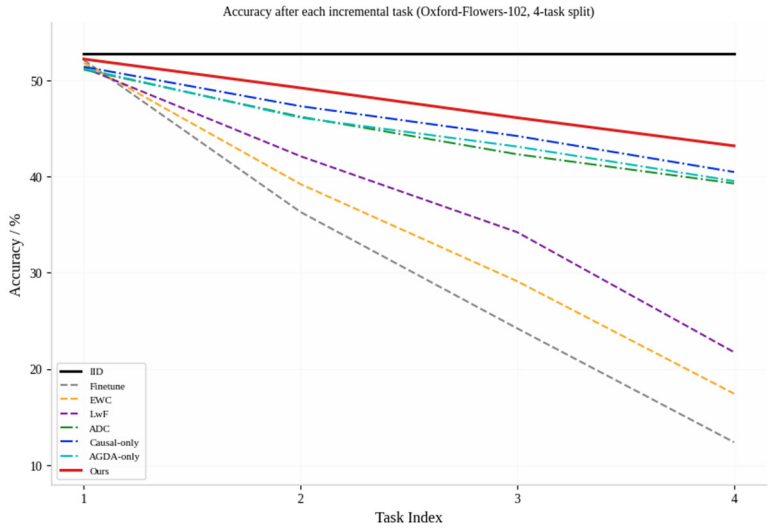



**Figure 3-3.** Accuracy changes at each incremental task stage for Oxford Flowers 102 (4-task split)

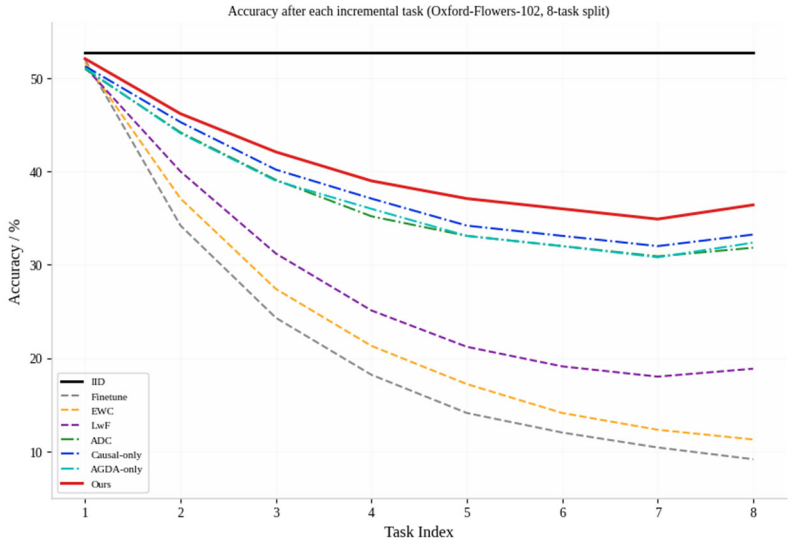



**Figure 3-4.** Accuracy changes at each incremental task stage for Oxford Flowers 102 (8-task split)

The bar chart visualization results of the ablation experiment are shown in Figure 3-5.

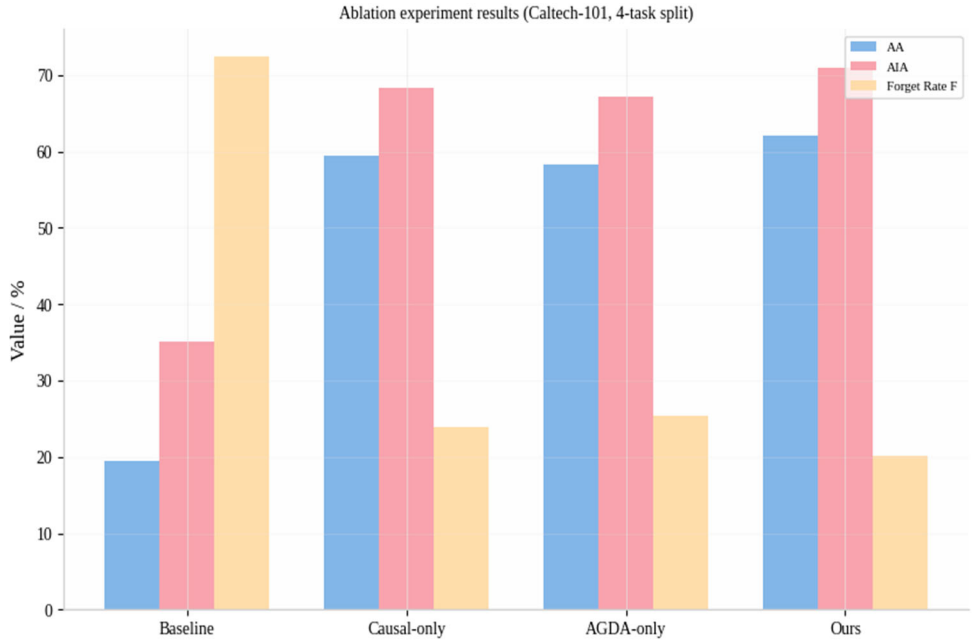



**Figure 3-5.** Comparison of module ablation experiment results on Caltech 101 (4-task split)

The sensitivity results for the number of neighbors K are shown in Figure 3-6. Comparing the curves for Causal only and our method, we can see that after introducing attention-guided sample augmentation to expand the candidate sample pool, our method outperforms the baseline across all K values. At the same time, the peak of the curve is smoother, indicating that semantically preserved synthetic samples improve the model's robustness to the neighbor parameter and alleviate the sampling issues of K neighbors in small-start scenarios.

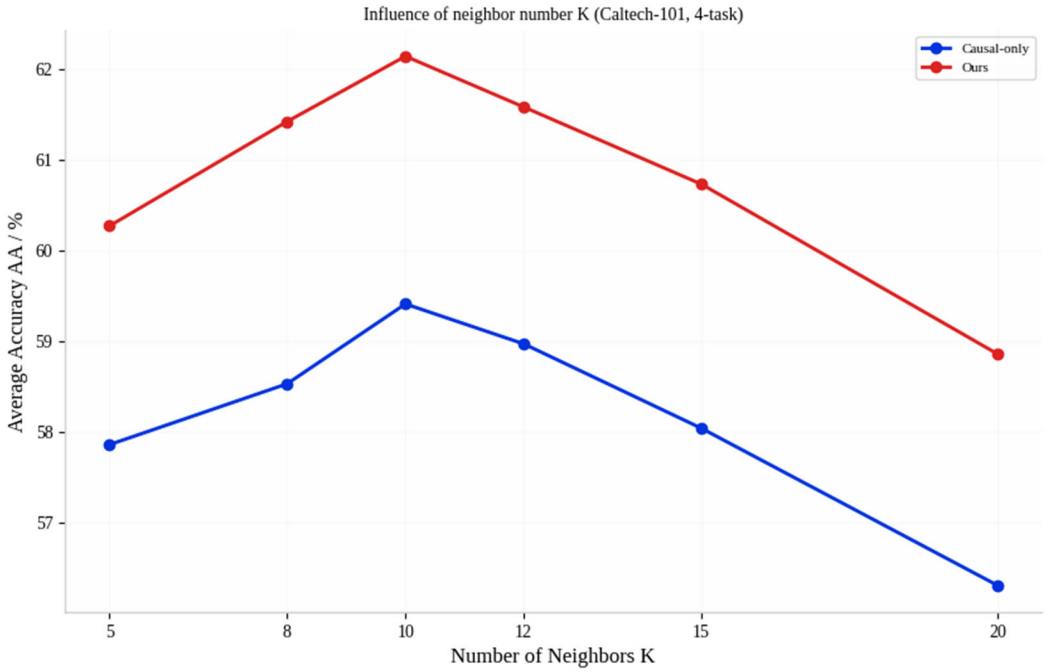



**Figure 3-6.** The effect of the number of neighbors K on the model's average accuracy (Caltech 101, 4 tasks)

To visually observe the optimization effect in the feature space, we used t-SNE to reduce the high-dimensional features for visualization, as shown in Figure 3-7. Here, (a) is the Causal only method, and (b) is our Ours method. With Causal only, the feature space shows quite noticeable overlap between classes. In contrast, our method, with AGDA sample augmentation and supervised contrastive loss working together, results in same-class samples clustering more tightly, clearer decision boundaries between different classes, and effectively enhanced feature discriminability.

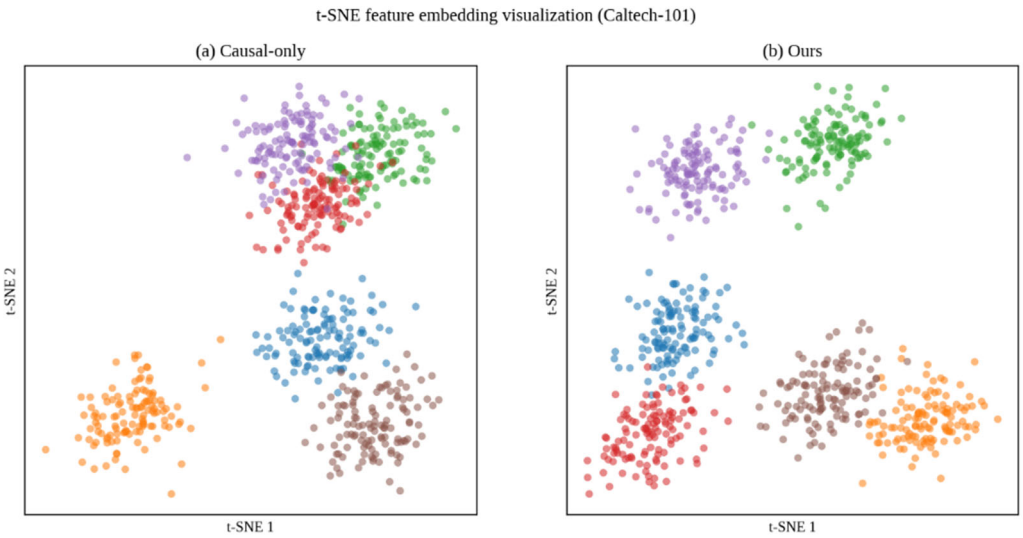



**Figure 3-7.** t-SNE feature embedding visualization comparison (CIFAR 100)

Figure 3-7 is divided into two subplots: (a) Causal only: there is some overlap between categories in the feature space, and the cohesion within classes is average; (b) Our method: samples of the same class are more tightly clustered, boundaries between different classes are clear, and category overlap is significantly reduced. Based on the results in this figure, it shows that AGDA preserves semantics while expanding the sample pool with synthetic samples, and combined with causal effect constraints, it optimizes the geometric structure of the feature space and alleviates prototype drift.

Across multiple evaluation dimensions like mean average precision, resistance to forgetting, accuracy on new classes, and inference efficiency[29], the radar chart of overall performance for each method is shown in Figure 3-8. Our method achieves the best results in mean average precision, resistance to forgetting, and accuracy on new classes, while its inference efficiency is roughly on par with the baseline, achieving a balanced trade-off between stability and plasticity.

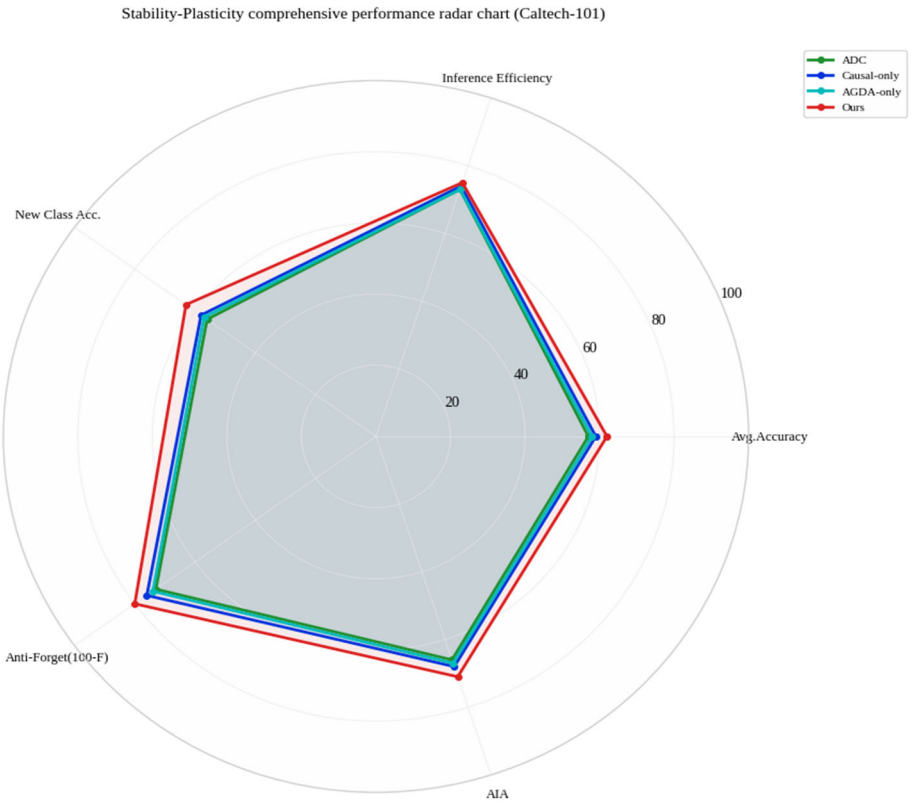



**Figure 3-8.** Radar Chart of Stability and Plasticity Comprehensive Performance

## 4.6. Comprehensive Analysis

1) Source of stability: The extraction of causal effects breaks collision structures through causal intervention, restoring the causal contribution of old data to model predictions and constraining the feature extractor parameter updates through causal regularization loss, preserving old knowledge at the representation level; compensating for prototype drift through adversarial prototypes corrects the prototype shift caused by incremental training and suppresses

the performance degradation of old class recognition; experimentally, the method in this paper shows that the average forgetting rate F is the lowest in all experimental settings, and the performance decay of old classes is minimal.

2) Source of plasticity: AGDA attention-guided augmentation generates semantic-preserving synthetic samples, expanding training samples, alleviating the scarcity of samples in small initial scenarios, and reducing the risk of overfitting; semantic-preserving synthetic samples participate in K-nearest neighbor sampling of causal effects, expanding the candidate feature pool and improving causal weight estimation, allowing the model to better learn new task features; supervised contrastive loss fully utilizes original samples and semantic-preserving synthetic samples, learning more discriminative features and enhancing the fitting ability for new classes; experimentally, the stage accuracy after each task training is higher, and the model demonstrates excellent learning capability for newly added classes.

In summary, this method neither over-constrains parameter updates like strong regularization methods, which suppress plasticity, nor focuses solely on fitting new classes like pure data augmentation methods, ignoring the causal-level protection of old representations. It achieves a synergistic balance between stability and plasticity. However, the method in this paper still relies on the quality of AGDA-generated semantic-preserving synthetic samples. When the dataset sample resolution is extremely low, the effectiveness of augmented samples decreases; this will be further investigated in future research.

# 5. Summary and Outlook

This paper focuses on small-start incremental learning scenarios without replay where samples are scarce. It studies the stability-plasticity dilemma and proposes a small-start incremental learning prototype compensation method based on attention-guided sample augmentation and causal constraints. The main improvement of this work lies in introducing semantic-preserving synthetic samples generated by attention-guided augmentation into the causal constraint computation process, expanding the candidate feature sample pool for K-nearest neighbor retrieval, and reducing bias in causal weight estimation caused by insufficient original samples. At the same time, causal regularization protects old task representations, supervised contrastive loss optimizes the feature embedding space, and an adversarial shift compensation strategy is used to correct prototype drift during incremental training.

On the Caltech 101 and Oxford Flowers 102 small-sample image datasets, we set up no-replay incremental experiments with 4-task and 8-task equal splits. Comparative experiments show that, compared to the baseline, our method improves average accuracy by 1.2%–3.3%, average incremental accuracy by 1.0%–3.0%, and significantly reduces average forgetting rate. Ablation experiments verify the synergistic effects of the attention-guided sample augmentation module and the causal constraint module. Parameter ablation results indicate that expanding the candidate pool with semantic-preserving synthetic samples improves model robustness to the K parameter in nearest neighbor retrieval. t-SNE feature visualizations show that our method achieves more compact intra-class features and better inter-class separation, effectively mitigating prototype drift and balancing stability and plasticity. In the future, dynamically adjusting $\lambda_1$ and $\lambda_2$ through an adaptive loss weight strategy according to the sample size of different tasks can help achieve a dynamic trade-off between stability and plasticity.

# References


[1] Huang, J., Xu, L., Liu, Y., et al. (2026). Reducing catastrophic forgetting via gradient-based task organization in class-incremental learning. Neurocomputing, 706, 134964. https://doi.org/10.1016/j.neucom.2026.134964.

[2] Wang, J., Hu, M., Li, N., Al-Ali, A., & Suganthan, P. N. (2026). Randomized neural network with adaptive forward regularization for online task-free class incremental learning. Neural Networks, 203, 109115. https://doi.org/10.1016/j.neunet.2026.109115

[3] Yuan, Z., Sun, B., Wang, Z., et al. (2026). Causal-enhanced machine learning framework for long-term wind power probabilistic forecasting. Reliability Engineering and System Safety, 272, 112532. https://doi.org/10.1016/j.ress.2026.112532

[4] Gudapati, M., Borah, A., & S, K. R. (2025). Incremental approach: A workflow model for semi-supervised learning environment. Engineering Applications of Artificial Intelligence, 160, 111815. https://doi.org/10.1016/j.engappai.2025.111815

[5] Jiang, M., Fan, J., & Li, F. (2025). Enhancing few-shot class-incremental learning through prototype optimization. Applied Intelligence, 55(10), 745. https://doi.org/10.1007/s10489-025-06605-z

[6] Li, D., Nair, C. K. N., & Wang, K. I. K. (2025). ADDAEIL: Anomaly detection with drift-aware ensemble-based incremental learning. Algorithms, 18(6), 359. https://doi.org/10.3390/a18060359

[7] Yuan, J., Chen, H., Tian, S., et al. (2025). Prompt-based learning for few-shot class-incremental learning. Alexandria Engineering Journal, 120, 287–295. https://doi.org/10.1016/j.aej.2025.02.008

[8] Ciarelli, M. P., Oliveira, E., & Salles, T. O. E. (2013). Impact of the characteristics of data sets on incremental learning. Artificial Intelligence Research, 2(4), 63–74. https://doi.org/10.5430/air.v2n4p63

[9] Hua, X., & Ding, S. (2011). Incremental learning algorithm for support vector data description. Journal of Software, 6(7), 1166–1173. https://doi.org/10.4304/jsw.6.7.1166-1173

[10] Zhang, H. S., Qu, L. X., & Wang, Y. X. (2011). Incremental regressive learning algorithm of support vector machine and its application. Advanced Materials Research, 216, 301–306. https://doi.org/10.4028/www.scientific.net/amr.216.301

[11] Devi, P., & Robi, P. (2007). An ensemble-based incremental learning approach to data fusion. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 37(2), 437–450.

[12] Hayes, T. L., Kafle, K., Shrestha, R., et al. (2020). REMIND your neural network to prevent catastrophic forgetting. In Computer Vision – ECCV 2020 (pp. 466–483). Springer.

[13] Al-Daoud, E. (2006). Incremental learning of auto-association multilayer perceptrons network. International Arab Journal of Information Technology, 3(1), 16–19.

[14] Hu, X. T., Tang, K. H., Miao, C. Y., et al. (2021). Distilling causal effect of data in class-incremental learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3957–3966). IEEE. https://doi.org/10.1109/CVPR46437.2021.00395

[15] Liu, M., Quan, T., & Yao, T. (2006). Tracking maneuvering target based on neural fuzzy network with incremental neural learning. Journal of Systems Engineering and Electronics,

17(2), 343–349. https://doi.org/10.1016/S1004-4132(06)60060-1

[16] Sun, J., Wang, J., & Xu, Y. (2004). Incremental learning with balanced update on receptive fields for multi-sensor data fusion. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 34(1), 659–665.

[17] Lee, J., & Cho, S. (2003). An learning algorithm to find the optimized network structure in an incremental model. Journal of Internet Computing and Services, 4(5), 69–76.

[18] Lange, S., & Grieser, G. (2002). On the power of incremental learning. Theoretical Computer Science, 288(2), 277–307. https://doi.org/10.1016/S0304-3975(01)00404-2

[19] Zhang, J., Liu, L., Silven, O., et al. (2025). Few-shot class-incremental learning for classification and object detection: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2025.3529038

[20] Farah, J., Khaled, J., & Ouajdi, K. (2025). Intrusion detection based on concept drift detection and online incremental learning. International Journal of Pervasive Computing and Communications, 21(1), 81–115. https://doi.org/10.1108/ijpcc-12-2023-0358

[21] Taie, W., ElGeneidy, K., Yacoub, A. A., et al. (2024). Addressing catastrophic forgetting in payload parameter identification using incremental ensemble learning. Frontiers in Robotics and AI, 11, 1470163. https://doi.org/10.3389/frobt.2024.1470163

[22] Zhu, K., Xu, J., Zhou, L., et al. (2024). Dmaf: Data-model anti-forgetting for federated incremental learning. Cluster Computing, 28(1), 30. https://doi.org/10.1007/s10586-024-04697-9

[23] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., et al. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114

[24] Li, Z., & Hoiem, D. (2018). Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(12), 2935–2947. https://doi.org/10.1109/TPAMI.2017.2773087

[25] Zhao, B., Xiao, X., Gan, G., et al. (2020). Maintaining discrimination and fairness in class incremental learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13208–13217). IEEE. https://doi.org/10.1109/CVPR42600.2020.01322

[26] Goswami, D., Soutif-Cormerais, A., Liu, Y., et al. (2024). Resurrecting old classes with new data for exemplar-free continual learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 23456–23465). IEEE. https://doi.org/10.1109/CVPR52733.2024.02474

[27] Douillard, A., Cord, M., Ollion, C., Robert, T., & Valle, E. (2020). PODNet: Pooled outputs distillation for small-tasks incremental learning. In Computer Vision – ECCV 2020 (pp. 86–102). Springer.

[28] Wang, N., Wang, H., Yang, S., et al. (2024). Semi-supervised incremental domain generalization learning based on causal invariance. International Journal of Machine Learning and Cybernetics, 15(10), 4815–4828. https://doi.org/10.1007/s13042-024-02199-z

[29] Sun, W., Li, Q., Zhang, J., et al. (2023). Class incremental learning based on identically distributed parallel one-class classifiers. Neurocomputing, 556. https://doi.org/10.1016/j.neucom.2023.126579.