GenFRC: Generative Feature Replay and Calibration for Non-Exemplar Class-Incremental Learning

Anonymous Author(s)

Abstract. Image classification is a fundamental computer vision task with broad practical applications. However, real-world visual systems inherently learn in an incremental manner as category distributions shift temporally and spatially. This long-term learning brings threats such as storage constraints and data privacy concerns, which prevent retaining previous data. Consequently, models suffer from catastrophic forgetting. Recent advances in Non-Exemplar Class-Incremental Learning (NECIL) aim to address this issue by enabling continual learning without access to prior class samples. However, continuous model updates lead to an inevitable drift between the old and new networks, disrupting classification boundaries. To address these challenges,, we propose a Generative Feature Replay and Calibration (GenFRC) method for NECIL. GenFRC consists of two key components: a Hybrid-Conditional Variational Replay (HCVR) module for local feature replay, and a Fourier Kernel Alignment Module (FKAM) for global feature calibration. Extensive experiments on two benchmarks under varying protocols demonstrate that GenFRC consistently outperforms state-of-the-art NECIL approaches.

Keywords: Incremental Learning · Lifelong Learning · Catastrophic Forgetting.

1 Introduction

Image classification is a fundamental computer vision task with broad practical applications [2, 14, 41]. However, real-world visual systems must learn incrementally, constantly adapting to new categories and changing visual contexts [37]. This incremental learning scenario challenges the traditional "train once, deploy forever" paradigm, which struggles to efficiently accommodate new classes [32]. Such limitations necessitate repeated collection and retraining with all historical data, resulting in significant computational and storage burdens. Moreover, the dependency on past data also raises privacy concerns under increasingly strict data regulations [17]. Consequently, the development of Incremental Learning (IL) frameworks becomes crucial for both theoretical advancements and real-world deployment.

IL is challenging due to catastrophic forgetting, where models forget previously learned classes when learning new knowledge [32]. This phenomenon stems from the absence of explicit long-term memory mechanisms in deep neural networks [11]. As models adapt to new class distributions, parameter updates often

overwrite discriminative pathways for old classes, gradually weakening gradient signals tied to earlier tasks [7]. The resulting imbalance between acquiring new knowledge and retaining old representations leads to performance instability. Mainstream methods alleviate forgetting by replaying a subset of old samples, such exemplar-based strategies are impractical under strict privacy constraints or limited resources. In contrast, Non-Exemplar Class-Incremental Learning (NECIL) offers a more viable solution by expanding class knowledge without relying on prior samples, making it particularly suited for privacy-sensitive and resource-constrained settings.

Recent NECIL methods primarily develop along three technical directions: generative replay, parameter regularization, and architecture optimization. Generative replay methods [25] preserve old knowledge by generating synthetic samples or leveraging knowledge distillation, but remains dependent on data storage and generator quality. Parameter regularization methods [15, 34, 34] protect old knowledge without accessing historical samples, but impose constraints that limit learning capability when faced with long task sequences or significant distributional differences. Architecture optimization methods [20, 33] dynamically expand or prune network structures to isolate tasks, effectively reducing forgetting but incurring linearly increasing complexity with more tasks.

To mitigate catastrophic forgetting, we propose synthesizing representative pseudo-features of previously learned classes and adaptively calibrating them using the feature representations of newly encountered classes. Building on this idea, we propose GenFRC method to regenerate high-fidelity pseudo-features for old classes while concurrently aligning them with new tasks. GenFRC integrates feature replay and distribution alignment through two key modules. The Hybrid-Conditional Variational Replay (HCVR) synthesizes diverse and discriminative historical features by conditioning on semantic, geometric, and synthetic attributes. The Fourier Kernel Alignment Module (FKAM) ensures global consistency between past and current feature distributions through a Random Fourier Feature approximation of Maximum Mean Discrepancy. The synergy between local feature reconstruction and global distribution alignment facilitates stable NECIL. Contributions can be summarized as:

1. We propose Hybrid-Conditional Variational Replay (HCVR) to generate rich and diverse pseudo-features that effectively mitigate forgetting at the representation level.
2. We propose Fourier Kernel Alignment Module (FKAM) to preserve distributional coherence across tasks, reducing feature drift during incremental updates.
3. We conduct comprehensive experiments across multiple class-incremental learning benchmarks, demonstrating that GenFRC consistently outperforms state-of-the-art approaches.

2 Related Work

2.1 Class-Incremental Learning (CIL)

CIL aims to incorporate new categories sequentially while retaining prior knowledge. To tackle the core challenge of catastrophic forgetting, existing methods are typically classified into three categories: replay-based, parameter regularization, and model expansion. Replay-based methods [23, 25, 28] retain a subset of past samples for rehearsal, but raise concerns about memory usage and privacy. Regularization methods [5, 15, 34] constrain updates to important weights, enabling non-exemplar learning, though performance often degrades over time or under distribution shifts. Model expansion [20, 33] mitigates forgetting by allocating task-specific capacity, but incurs scalability and efficiency costs. In real-world applications with tight memory, privacy, or computational budgets, these strategies face practical limitations.

2.2 Non-Exemplar CIL (NECIL)

In privacy-sensitive scenarios where storing historical samples is infeasible, NECIL addresses the CIL problem through alternative mechanisms such as knowledge distillation, generative replay and parameter isolation. Distillation-based approaches [18, 19, 27, 35] transfer knowledge by using the output logits or intermediate representations from previous models to supervise the learning of new tasks. Generative replay methods [9, 21, 31] utilize generative models to simulate data distributions of past tasks. Parameter isolation techniques [22, 26] freeze the backbone and learn translation layers, reducing interference. While distillation avoids storing real data, it is sensitive to distribution drift. Generative approaches mitigate forgetting but suffer from high computational costs and instability. Parameter isolation methods offer effective forgetting control but may increase inference overhead. Therefore, NECIL remains an evolving area with growing emphasis on hybrid strategies that balance forgetting, efficiency, and representation quality.

3 Methodology

3.1 Overview

Fig. 1 illustrates the GenFRC method proposed for NECIL. It comprises two complementary components: (i) Hybrid-Conditional Variational Replay (HCVR), designed to synthesize representative pseudo-features of past classes in the absence of stored exemplars; and (ii) Fourier Kernel Alignment Module (FKAM), which alleviates feature drift through efficient Maximum Mean Discrepancy (MMD) approximations via Random Fourier Features (RFF). These modules collaboratively enable effective incremental learning without storing historical exemplars.

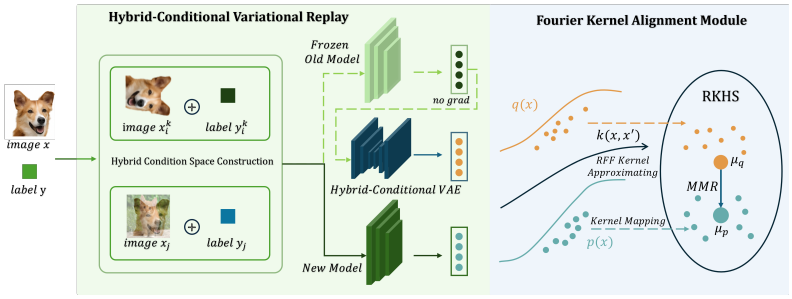

Fig. 1: Overview of the proposed GenFRC. Hybrid-Conditional Variational Replay (HCVR) generates old-class features locally, while the Fourier Kernel Alignment Module (FKAM) aligns feature distributions across tasks globally.

3.2 Hybrid-Conditional Variational Replay (HCVR) for Local Feature Replay

In non-exemplar setups, catastrophic forgetting arises from the loss of class discriminability over time. HCVR addresses this issue within a probabilistic generative method that synthesises representative old-class features conditioned on three complementary sources of information: semantic class identity, geometric transformations, and synthetic mix augmentations.

Probabilistic Feature Replay with Hybrid Conditioning. Let $p_\theta(f_{\text{old}} \mid c)$ be the decoder likelihood of an old-class feature f_{old} given the hybrid condition c. A variational encoder $q_\phi(z \mid f_{\text{old}}, c)$ approximates the true posterior $p(z \mid f_{\text{old}}, c)$ over latent variable $z \in \mathbb{R}^d$. The ELBO is

$$\log p_\theta(f_{\text{old}} \mid c) \;\geq\; \mathbb{E}_{q_\phi}[\log p_\theta(f_{\text{old}} \mid z, c)] \;-\; \beta\, D_{\text{KL}}(q_\phi(z \mid f_{\text{old}}, c) \,\|\, p(z)), \qquad (1)$$

where $p(z) = \mathcal{N}(0, I)$ and β is the β-VAE coefficient that regulates disentanglement [13].

The hybrid condition vector c integrates three complementary sources of information to enhance the generative model's capability. Semantic class identity encodes the original class identity: for each class label $y \in \{1, \ldots, K\}$, the corresponding vector is defined as $c_{\text{cls}} \in \{0, 1\}^K$, where K denotes the number of semantic classes. Geometric transformation captures rotation transformations applied to input samples. For a rotation $r \in \{0°, 90°, 180°, 270°\}$, an expanded label space is constructed as $y_{\text{geo}} = y \times 4 + k$, where $k \in \{0, 1, 2, 3\}$ corresponds to the rotation angle r. The associated condition vector is $c_{\text{geo}} \in \{0, 1\}^{4K}$. Synthetic mix augmentation models intermediate classes generated by mix augmentation. Given two samples (x_i, x_j) and a mixing coefficient $\lambda \sim \text{Beta}(\alpha, \alpha)$, the synthetic condition vector is constructed as a convex combination of the respective class encodings that $c_{\text{syn}} = \lambda c_{\text{cls}}^{(i)} + (1 - \lambda) c_{\text{cls}}^{(j)} \in [0, 1]^K$. The final hybrid condition is constructed through learnable fusion:

$$c = W_{\text{cls}} \cdot c_{\text{cls}} \oplus W_{\text{geo}} \cdot c_{\text{geo}} \oplus W_{\text{syn}} \cdot c_{\text{syn}}, \qquad (2)$$

where $W_{\{\cdot\}}$ are learnable projection matrices and $\oplus$ denotes concatenation. This design preserves distinct information channels while enabling adaptive weighting of different condition types.

Variational Inference and Feature Reconstruction. The encoder $q_\phi(z|f,c)$ maps hybrid-conditioned features to latent parameters:

$$\mu, \log\sigma^2 = \text{MLP}([f_{\text{old}}; c]), \quad z = \mu + \epsilon \odot \exp\left(\frac{\log\sigma^2}{2}\right). \tag{3}$$

The decoder $p_\theta(f_{\text{old}}|z,c)$ reconstructs features through:

$$\hat{f}_{\text{old}} = \text{MLP}([z;c]) + \epsilon_{\text{obs}}, \quad \epsilon_{\text{obs}} \sim \mathcal{N}(0, \sigma_{\text{obs}}^2 I), \tag{4}$$

where σ_{obs} models reconstruction uncertainty. The reconstruction loss in Eq. (1) thus becomes:

$$\log p_\theta(f_{\text{old}}|z,c) = -\frac{1}{2\sigma_{\text{obs}}^2}\|f_{\text{old}} - \hat{f}_{\text{old}}\|_2^2 + \text{const}. \tag{5}$$

Dynamic Replay via Feature Manifold Interpolation. During incremental stages t, synthetic features are generated through latent space traversal conditioned on historical class manifolds:

$$\hat{f}_{\text{old}}^{(i)} = \text{Decoder}\big(z^{(i)}, c^{(i)}\big), \quad z^{(i)} \sim \mathcal{N}(0, I), \tag{6}$$

where $c^{(i)} = c_{\text{cls}}^{(i)} \oplus c_{\text{geo}}^{(i)} \oplus c_{\text{syn}}^{(i)}$. The mix components $c_{\text{syn}}^{(i)}$ are dynamically adjusted based on the current task's class distribution:

$$c_{\text{syn}}^{(i)} = \sum_{j=1}^{K_{\text{old}}} \pi_j c_{\text{cls}}^{(j)}, \quad \pi \sim \text{Dirichlet}(\alpha_{\text{syn}}), \tag{7}$$

where K_{old} is the number of old classes and α_{syn} controls interpolation diversity.

Let $\mathcal{F}_{\text{old}}$ and $\mathcal{F}_{\text{syn}}$ denote the true and synthetic feature distributions. Under Lipschitz continuity assumptions of the decoder, the Wasserstein distance between distributions is bounded by:

$$W_2(\mathcal{F}_{\text{old}}, \mathcal{F}_{\text{syn}}) \leq L_{\text{dec}}\mathbb{E}[\|z - z'\|_2] + \sigma_{\text{obs}}, \tag{8}$$

where L_{dec} is the decoder's Lipschitz constant and z' is the latent variable for real features, following [1, 30]. Minimizing Eq. (1) directly reduces this bound through the reconstruction term and latent regularization.

The final HCVR loss function is:

$$\mathcal{L}_{\text{CVAE}} = \|f_{\text{old}} - \hat{f}_{\text{old}}\|_2^2 + \beta_{\text{CVAE}} \cdot D_{\text{KL}}\left(q(z \mid f,c) \,\|\, \mathcal{N}(0,I)\right). \tag{9}$$

3.3 Fourier Kernel Alignment Module (FKAM) for Global Feature Calibration

FKAM addresses feature drift by aligning feature distributions between incremental tasks using MMD metric, approximated efficiently with RFF. An adaptive regularization schedule balancing stability and plasticity ensures smooth transitions in representation spaces across tasks.

Kernel-Induced Feature Space Alignment. The MMD metric quantifies distributional discrepancy within a Reproducing Kernel Hilbert Space (RKHS). Let P and Q denote feature distributions of old and new tasks, respectively. The squared MMD is defined as:

$$\mathrm{MMD}^2(P,Q) = \mathbb{E}_{x,x'\sim P}[k(x,x')] + \mathbb{E}_{y,y'\sim Q}[k(y,y')] - 2\mathbb{E}_{\substack{x\sim P\\ y\sim Q}}[k(x,y)], \quad (10)$$

where $k(\cdot,\cdot)$ is typically the Radial Basis Function (RBF) kernel $k(x,y) = \exp(-\gamma\|x-y\|^2)$. Direct computation requires $\mathcal{O}((N+M)^2)$ operations for N new and M old features, which becomes prohibitive as tasks accumulate.

RFF Approximation with Error Bounds. By Bochner's theorem, any stationary kernel admits the following expression:

$$k(x,y) = \mathbb{E}_\omega[\sqrt{2}\cos(\omega^\top x + b)\sqrt{2}\cos(\omega^\top y + b)], \quad (11)$$

with $\omega \sim \mathcal{N}(0, 2\gamma I)$ and $b \sim U[0, 2\pi]$. Using Monte Carlo approximation with D_{rff} samples, we obtain:

$$k(x,y) \approx \varphi(x)^\top \varphi(y) = \frac{2}{D_{\mathrm{rff}}} \sum_{i=1}^{D_{\mathrm{rff}}} \cos(\omega_i^\top x + b_i)\cos(\omega_i^\top y + b_i). \quad (12)$$

The approximation error decays as $O(1/\sqrt{D_{\mathrm{rff}}})$ [24], ensuring theoretical consistency. Substituting Eq. (12) into Eq. (10) yields the RFF-MMD loss:

$$\mathcal{L}_{\mathrm{FKAM}} = \left\| \frac{1}{N}\sum_{i=1}^{N} \varphi(f_{\mathrm{new}}^{(i)}) - \frac{1}{M}\sum_{j=1}^{M} \varphi(f_{\mathrm{old}}^{(j)}) \right\|_2^2. \quad (13)$$

This reduces computational complexity to $\mathcal{O}((N+M)D_{\mathrm{rff}})$, linear in the number of samples.

Adaptive Regularization with Stability-Plasticity Tradeoff. The regularization strength $\lambda(t)$ is dynamically adjusted during training according to:

$$\lambda(t) = \lambda_0 \cdot \eta(t), \quad \eta(t) = \max\left(1 - \frac{t}{T}, \kappa\right), \quad (14)$$

where t is the current epoch, T the total number of epochs, $\lambda_0 = 50$ is the initial weight, and $\kappa = 0.1$ is the minimum retention ratio. This scheduling design emphasizes stability during early training stages, where $\eta \approx 1$ ensures strong feature alignment to suppress abrupt feature drift. As training progresses, the linear decay of $\eta(t)$ gradually increases plasticity, allowing the model to flexibly adapt to new task features. Meanwhile, the lower bound κ guarantees that a minimum level of knowledge retention is maintained throughout training, preventing complete forgetting of old class distributions.

3.4 Integrated Loss Function

The overall training objective combines classification, feature replay, and feature calibration terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{HCVR}} + \lambda_f \mathcal{L}_{\text{FKAM}}, \tag{15}$$

where $\mathcal{L}_{\text{cls}}$ denotes the cross-entropy classification loss, $\mathcal{L}_{\text{HCVR}}$ includes both reconstruction and regularization terms derived from the ELBO, $\mathcal{L}_{\text{FKAM}}$ serves as a global regularizer to minimize distributional shifts in the feature space across tasks, and λ_f is hyperparameter controlling the contributions of FKAM losses.

4 Experiments

4.1 Datasets

We follow previous works by evaluating our proposed GenFRC model on two widely-used public datasets: CIFAR-100 and TinyImageNet. CIFAR-100 consists of 100 classes, with 500 training images and 100 test images per class. TinyImageNet contains 200 classes, with 500 training images and 50 test images per class. We adopt the conventional NECIL setting to construct incremental learning (IL) scenarios. Specifically, for CIFAR-100, the model is initially trained on 50, 50, and 40 classes, followed by training across 5, 10, and 20 IL stages. For TinyImageNet, the model is first trained on 100 classes, followed by training across 5, 10, and 20 IL stages.

4.2 Evaluation Criteria

We evaluate performance using three metrics: Last-stage Accuracy, Average Accuracy, and Average Forgetting. Last-stage Accuracy (A_K) measures model performance after the final task, indicating its ability to retain and generalize knowledge across all K tasks. Average Accuracy ($\bar{A}$) reflects overall performance throughout the learning process, computed as $\bar{A} = \frac{1}{K+1}\sum_{k=0}^{K} A_k$, where A_k is the accuracy at stage k. Average Forgetting ($\bar{F}$) captures performance degradation on earlier tasks, defined as $\bar{F} = \frac{1}{K-1}\sum_{k=1}^{K-1}(\max_{l\leq k} A_{k,l} - A_{K,l})$, where $A_{k,l}$ is the accuracy on task l after stage k, and $A_{K,l}$ is the final accuracy on task l.

Table 1: Last-stage accuracy comparison with state-of-the-art methods on CIFAR-100 and TinyImageNet (%).

Method	**CIFAR100**			**TinyImageNet**		
	5 stages	10 stages	20 stages	5 stages	10 stages	20 stages
Joint-Train	77.3	77.3	77.3	54.2	54.2	54.2
Fine-Tune	9.0	4.8	3.3	7.1	3.7	2.0
LwF [18]	24.0	16.5	14.7	14.7	7.6	3.1
iCaRL-CNN [25]	47.8	42.2	40.1	24.8	20.0	15.2
iCaRL-NME	55.0	48.5	46.1	30.5	25.6	18.5
EEIL [4]	50.2	47.6	42.2	35.0	33.7	27.6
PASS [39]	56.4	50.7	46.9	42.5	40.3	34.8
IL2A [38]	53.9	45.8	44.2	39.5	36.6	30.0
R-DFCIL [8]	54.8	50.0	37.0	40.8	37.9	32.0
SSRE [40]	57.0	56.6	51.9	41.5	41.2	41.0
EDG [10]	56.0	54.3	49.3	38.1	38.0	34.9
FeTrIL [22]	58.1	57.6	52.5	42.9	42.4	41.3
FGKSR [35]	59.0	57.9	54.3	45.0	43.4	41.9
TASS [19]	59.3	57.9	53.8	44.1	43.9	43.6
FCS [16]	62.1	60.4	58.4	46.0	45.0	42.6
GenFRC	**64.2**	**62.4**	**59.2**	**47.9**	**46.0**	**44.0**

Table 2: Average accuracy comparison with state-of-the-art methods on CIFAR-100 dataset (%).

Method	5 stages	10 stages	20 stages
LwF [18]	45.9	27.4	20.1
DeeSIL [3]	60.0	50.6	38.1
ABD [29]	63.8	62.5	57.4
PASS [39]	63.5	61.8	58.1
IL2A [38]	66.0	60.3	57.9
SSRE [40]	65.9	65.0	61.7
FeTrIL [22]	66.3	65.2	61.5
SEED [26]	70.9	69.3	62.9
PRL [27]	71.3	70.2	68.4
FGKSR [35]	68.2	70.1	66.9
TASS [19]	68.8	67.4	62.8
FCS [16]	70.4	69.0	68.3
GenFRC	**72.6**	**70.8**	**69.2**

Table 3: Average Forgetting comparison with state-of-the-art methods on CIFAR-100 dataset (%).

Method	5 stages	10 stages	20 stages
LwF [18]	50.8	55.0	58.0
iCaRL-CNN [25]	42.8	46.5	51.0
iCaRL-NME [25]	27.8	31.9	28.8
EEIL [4]	23.4	26.7	32.4
PASS [39]	19.6	26.6	27.8
IL2A [38]	28.5	39.3	41.3
SSRE [40]	18.4	19.5	19.0
EDG [10]	21.9	23.8	24.7
FeTrIL [22]	17.2	18.8	23.4
TASS [19]	16.4	17.8	17.8
FGKSR [35]	16.4	17.2	16.6
FCS [16]	12.2	16.7	15.9
GenFRC	**9.9**	**14.4**	**14.2**

4.3 Implementation Details

We use the widely adopted ResNet-18 architecture [12] as the backbone network, initialized from scratch. The model is optimized using the Adam optimizer with an initial learning rate of 1×10^{-3} and a weight decay of 2×10^{-4}. Training is conducted over 100 epochs, with the learning rate decayed by a factor of 0.1 every 45 epochs. A batch size of 64 is employed for all experiments. Image augmentation follows the protocols in [6, 36]. For reproducibility, the random

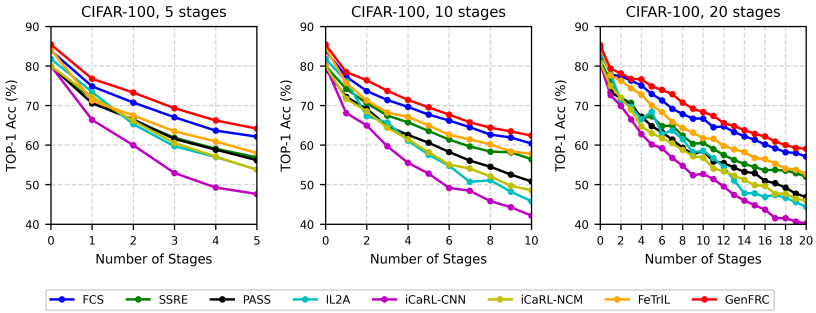

Fig. 2: Complete classification accuracy of each stage on CIFAR-100.

seed was set to 2025. All experiments are executed using PyTorch on a single NVIDIA A6000 GPU.

Table 4: Ablation study of different components.

Method	CIFAR-100		
	5 stages	**10 stages**	**20 stages**
Base	62.1	60.4	58.4
w/ HCVR	63.3	61.2	58.6
w/ FKAM	63.2	61.4	58.8
GenFRC	**64.2**	**62.4**	**59.1**

4.4 Comparison with state-of-the-arts

Main Results We compare GenFRC method with various NECIL approaches, including LwF [18], DssSIL [3], ABD [29], PASS [39], IL2A [38], SSRE [40], R-DFCIL [8], EDG [10], FeTrIL [22], FCS [16], SEED [26], FGKSR [35], TASS [19] and PRL [27], as well as two exemplar-based CIL methods, iCaRL [25] and EEIL [4]. Across multiple scenarios, GenFRC consistently outperforms previous CIL methods. Additionally, we include two special experimental settings: Joint-Train and Fine-Tune. Joint-Train involves training with all data at once, serving as an upper bound, while Fine-Tune refers to directly fine-tuning the model without any anti-forgetting algorithms.

Table 1 presents the Last-stage accuracy (A_K) results. GenFRC outperforms all compared methods across various settings. Specifically, on CIFAR-100, we achieve 64.2%, 62.4%, and 59.2% under the 5-, 10-, and 20-stage settings, respectively, and 47.9%, 46.0%, and 44.0% on TinyImageNet. These results establish a new state-of-the-art. Compared to the strongest prior competitor FCS, GenFRC yields absolute gains of +2.1%, +2.0%, +0.8% on CIFAR-100 and +1.9%, +1.0%, +1.4% on TinyImageNet. Notably, the margin to the joint-training upper bound is reduced to just 13.1% in the 5-stage CIFAR-100 setting.

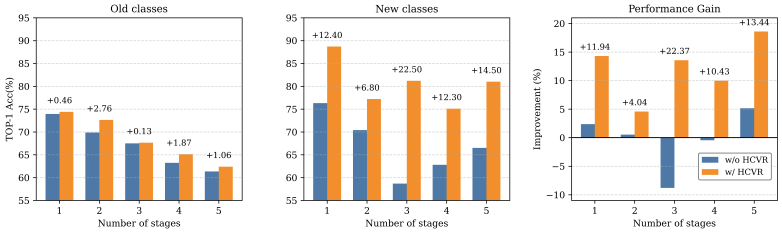

Fig. 3: Top-1 accuracy of the old classes, new classes, and the performance gain caused by the integration of HCVR from different stages.

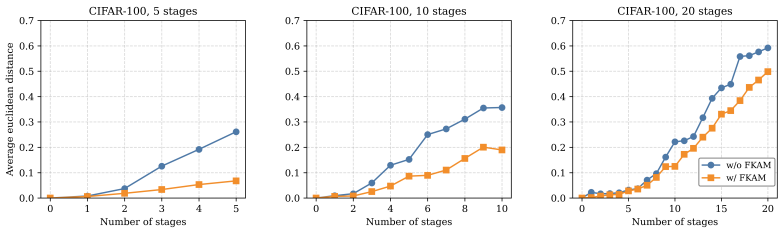

Fig. 4: Average Euclidean distance between the old network and the new network at each stage on CIFAR-100.

Tables 2 and 3 report Average Accuracy ($\bar{A}$) and Average Forgetting (AF) for CIFAR-100. GenFRC attains the highest $\bar{A}$ across all settings, with values of 72.6%, 70.8%, and 69.2% for the 5, 10, and 20-stage settings respectively, surpassing FCS and SEED. Simultaneously, GenFRC achieves the lowest AF scores of **9.9, 14.4, 14.2** percentage points. These consistent improvements in both average accuracy and forgetting demonstrate that the synergy between HCVR and FKAM not only enhances knowledge acquisition but also mitigates catastrophic forgetting, enabling superior long-term retention without relying on stored exemplars.

Average Curve Fig. 2 illustrates the stage-wise top-1 accuracy on CIFAR-100 for the 5-, 10-, and 20-stage settings. Throughout the entire trajectory, GenFRC consistently outperforms the other methods. GenFRC maintains an accuracy of over 64% after five incremental updates. In the more challenging 10- and 20-stage settings, GenFRC achieves last-stage accuracy of around 62% and 59%, respectively. The performance of GenFRC decays notably more slowly.

4.5 Ablation Studies

Table 4 reports the quantitative ablation results on CIFAR-100, while Figs. 3 and4 provide a stage-wise breakdown of the effects of each component.

Effectiveness of HCVR: We present a visualization of classification accuracy on both old and new classes in Fig. 3, along with the corresponding performance gain across incremental stages. Notably, the integration of HCVR consistently improves accuracy for both class types at all stages. These gains can be attributed to HCVR's ability to reconstruct local features and replay old knowledge, thereby mitigating forgetting. To further examine this effect, we compare the performance gains across stages between models with and without HCVR. As shown, HCVR leads to significant improvements, especially in later stages where the accumulation of tasks typically amplifies forgetting. This highlights HCVR's effectiveness in preserving learned representations and facilitating continual learning.

Effectiveness of FKAM: We visualize the average Euclidean distance between the feature centers extracted by the old network and the new network across different incremental stages on CIFAR-100, with 5, 10, and 20-stage configurations in Fig. 4. A smaller distance indicates better alignment between the evolving feature spaces. Across all settings, FKAM consistently yields lower feature drift. This result suggests that FKAM effectively preserves feature continuity by transferring the class-wise feature centers from the old network to the new one. By mitigating feature space distortion and suppressing representation drift, FKAM plays a critical role in alleviating forgetting caused by semantic shift, thereby improving overall model stability during continual learning.

Table 5: Impact of kernel choice inside the FKAM evaluated on CIFAR-100. "K_n" employs a fixed $n \times n$ Gaussian kernel, and *Fourier* uses Random Fourier Features.

Kernel Variant	A_K	$\bar{A}$	$\overline{Mem}$ **(GB)**
K_1	62.72	70.45	28.27
K_3	62.77	70.46	
K_5	62.82	70.46	
Fourier	63.18	72.45	6.69

Table 6: Impact of λ_f and D_{rff} evaluated on CIFAR-100.

$\boldsymbol{\lambda_f}$	$\boldsymbol{D}_{\text{rff}}$	A_K	$\bar{A}$
50	256	62.39	72.27
	512	63.18	72.45
100	256	62.27	72.11
	512	62.46	72.15
150	256	61.38	71.67
	512	62.38	72.01

4.6 Effect of Hyperparameter Settings

To investigate the influence of FKAM and its associated hyperparameter configurations, we analyse the kernel variant, the MMD weight λ_f, and the RFF dimension D_{rff}. The corresponding quantitative results are summarized in Tables 5 and 6.

Kernel Variant in FKAM: Compared to fixed Gaussian kernels, the RFF-based kernel consistently achieves better results, improving A_K from 62.82 to 63.18 and $\bar{A}$ from 70.46 to 72.45. This improvement stems from two factors: (i) RFF approximates shift-invariant kernels using explicit mappings, enabling faster and scalable computation compared to implicit kernel evaluations; (ii) its randomized projection captures richer frequency patterns, offering a more expressive and data-adaptive alignment space without requiring kernel size tuning.

MMD Regularization Weight λ_f: Table 6 shows that increasing λ_f from 50 to 150 consistently degrades performance. A large weight likely over-constrains the feature alignment objective, hindering the model's ability to adapt to new tasks. A moderate setting ($\lambda = 50$) offers a better trade-off by preserving old knowledge while allowing sufficient plasticity for learning new classes.

RFF Dimension $D_{\mathbf{rff}}$: 512 RFF dimensions yield better kernel approximations and tighter alignment. Increasing D_{rff} from 256 to 512 improves both A_K and $\bar{A}$, indicating enhanced capacity to represent complex feature distributions.

5 Conclusion

In this paper, we propose a Generative Feature Replay and Calibration (GenFRC) method to address the challenging problem of catastrophic forgetting in non-exemplar class-incremental learning (NECIL). GenFRC consists of two key components: Hybrid-Conditional Variational Replay (HCVR) and Fourier Kernel Alignment Module (FKAM). In detail, HCVR synthesizes representative pseudo-features without exemplars, effectively alleviating catastrophic forgetting while preserving the discriminative representations of previously learned classes. FKAM leverages random Fourier features to approximate the RBF-based MMD to further mitigate representation drift, enabling a computational reduction from $\mathcal{O}(N+M)^2)$ to $\mathcal{O}(N+M)D)$. Comprehensive evaluations on multiple class-incremental benchmarks demonstrate the superior performance and stability of GenFRC compared to state-of-the-art NECIL approaches.

References

1. Arjovsky, M., Chintala, S., Bottou, L.: Wasserstein generative adversarial networks. In: International conference on machine learning. pp. 214–223. PMLR (2017)
2. Azizi, S., Mustafa, B., Ryan, F., Beaver, Z., Freyberg, J., Deaton, J., Loh, A., Karthikesalingam, A., Kornblith, S., Chen, T., et al.: Big self-supervised models advance medical image classification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3478–3488 (2021)
3. Belouadah, E., Popescu, A.: Deesil: Deep-shallow incremental learning. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
4. Castro, F.M., Marín-Jiménez, M.J., Guil, N., Schmid, C., Alahari, K.: End-to-end incremental learning. In: Proceedings of the European conference on computer vision (ECCV). pp. 233–248 (2018)
5. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: ECCV (2018), https://arxiv.org/abs/1801.10112
6. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation strategies from data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 113–123 (2019)
7. Elsayed, M., Mahmood, A.R.: Addressing loss of plasticity and catastrophic forgetting in continual learning. In: The Twelfth International Conference on Learning Representations

8. Gao, Q., Zhao, C., Ghanem, B., Zhang, J.: R-dfcil: Relation-guided representation learning for data-free class incremental learning. In: European Conference on Computer Vision. pp. 423–439. Springer (2022)
9. Gao, R., Liu, W.: Ddgr: Continual learning with deep diffusion-based generative replay. In: International Conference on Machine Learning. pp. 10744–10763. PMLR (2023)
10. Gao, Z., Xu, C., Li, F., Jia, Y., Harandi, M., Wu, Y.: Exploring data geometry for continual learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24325–24334 (2023)
11. Hadsell, R., Rao, D., Rusu, A.A., Pascanu, R.: Embracing change: Continual learning in deep neural networks. Trends in cognitive sciences **24**(12), 1028–1040 (2020)
12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
13. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. In: International conference on learning representations (2017)
14. Jeong, J., Zou, Y., Kim, T., Zhang, D., Ravichandran, A., Dabeer, O.: Winclip: Zero-/few-shot anomaly classification and segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19606–19616 (2023)
15. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
16. Li, Q., Peng, Y., Zhou, J.: Fcs: Feature calibration and separation for non-exemplar class incremental learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 28495–28504 (2024)
17. Li, W., Gao, H.a., Gao, M., Tian, B., Zhi, R., Zhao, H.: Training-free model merging for multi-target domain adaptation. In: European Conference on Computer Vision. pp. 419–438. Springer (2024)
18. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)
19. Liu, X., Zhai, J.T., Bagdanov, A.D., Li, K., Cheng, M.M.: Task-adaptive saliency guidance for exemplar-free class incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23954–23963 (2024)
20. Mallya, A., Lazebnik, S.: Packnet: Adding multiple tasks to a single network by iterative pruning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7765–7773 (2018)
21. Meng, Z., Zhang, J., Yang, C., Zhan, Z., Zhao, P., Wang, Y.: Diffclass: Diffusion-based class incremental learning. In: European Conference on Computer Vision. pp. 142–159 (2024)
22. Petit, G., Popescu, A., Schindler, H., Picard, D., Delezoide, B.: Fetril: Feature translation for exemplar-free class-incremental learning. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3911–3920 (2023)
23. Prabhu, A., Torr, P.H., Dokania, P.K.: Gdumb: A simple approach that questions our progress in continual learning. In: European Conference on Computer Vision. pp. 524–540. Springer (2020)

24. Rahimi, A., Recht, B.: Random features for large-scale kernel machines. Advances in neural information processing systems **20** (2007)
25. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
26. Rypeść, G., Cygert, S., Khan, V., Trzcinski, T., Zieliński, B.M., Twardowski, B.: Divide and not forget: Ensemble of selectively trained experts in continual learning. In: The Twelfth International Conference on Learning Representations
27. Shi, W., Ye, M.: Prospective representation learning for non-exemplar class-incremental learning. Advances in Neural Information Processing Systems **37**, 995–1018 (2024)
28. Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. In: Advances in Neural Information Processing Systems. pp. 2990–2999 (2017), https://arxiv.org/abs/1705.08690
29. Smith, J., Hsu, Y.C., Balloch, J., Shen, Y., Jin, H., Kira, Z.: Always be dreaming: A new approach for data-free class-incremental learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9374–9384 (2021)
30. Tolstikhin, I., Bousquet, O., Gelly, S., Schoelkopf, B.: Wasserstein auto-encoders. arXiv preprint arXiv:1711.01558 (2017)
31. van de Ven, G.M., Siegelmann, H.T., Tolias, A.S.: Brain-inspired replay for continual learning with artificial neural networks. Nature communications **11**(1), 1–14 (2020)
32. Wang, L., Zhang, X., Su, H., Zhu, J.: A comprehensive survey of continual learning: Theory, method and application. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
33. Yoon, J., Yang, E., Lee, J., Hwang, S.J.: Lifelong learning with dynamically expandable networks. In: 6th International Conference on Learning Representations, ICLR 2018. International Conference on Learning Representations, ICLR (2018)
34. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: International conference on machine learning. pp. 3987–3995. PMLR (2017)
35. Zhai, J.T., Liu, X., Yu, L., Cheng, M.M.: Fine-grained knowledge selection and restoration for non-exemplar class incremental learning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6971–6978 (2024)
36. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412 (2017)
37. Zhao, B., Mac Aodha, O.: Incremental generalized category discovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19137–19147 (2023)
38. Zhu, F., Cheng, Z., Zhang, X.y., Liu, C.l.: Class-incremental learning via dual augmentation. In: Thirty-Fifth Conference on Neural Information Processing Systems (2021)
39. Zhu, F., Zhang, X.Y., Wang, C., Yin, F., Liu, C.L.: Prototype augmentation and self-supervision for incremental learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5871–5880 (2021)
40. Zhu, K., Zhai, W., Cao, Y., Luo, J., Zha, Z.J.: Self-sustaining representation expansion for non-exemplar class-incremental learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9296–9305 (2022)
41. Zhu, X., Bilal, T., Mårtensson, P., Hanson, L., Björkman, M., Maki, A.: Towards sim-to-real industrial parts classification with synthetic dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4454–4463 (2023)