Compressing model with few class-imbalance samples: An out-of-distribution expedition

Tian-Shuang Wu TIANSHUANGWU@HHU.EDU.CN
Key Laboratory of Water Big Data Technology of Ministry of Water Resources, College of Computer Science and Software Engineering, Hohai University, Nanjing, China

Shen-Huan Lyu ✉ LVSH@HHU.EDU.CN
Key Laboratory of Water Big Data Technology of Ministry of Water Resources, College of Computer Science and Software Engineering, Hohai University, Nanjing, China
Department of Computer Science, City University of Hong Kong, Hong Kong, China
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

Yanyan Wang YANYAN.WANG@HHU.EDU.CN
Key Laboratory of Water Big Data Technology of Ministry of Water Resources, College of Computer Science and Software Engineering, Hohai University, Nanjing, China
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

Ning Chen CHE-N-ING@HHU.EDU.CN
Key Laboratory of Water Big Data Technology of Ministry of Water Resources, College of Computer Science and Software Engineering, Hohai University, Nanjing, China

Zhihao Qu QUZHIHAO@HHU.EDU.CN
Key Laboratory of Water Big Data Technology of Ministry of Water Resources, College of Computer Science and Software Engineering, Hohai University, Nanjing, China

Baoliu Ye YEBL@NJU.EDU.CN
State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

Abstract

Few-sample model compression aims to compress a large pre-trained model into a compact one using only a few samples. However, previous methods typically assume a balanced class distribution, which is costly under severe data scarcity. In the presence of imbalance, the compressed model exhibits significant performance degradation. We propose a novel framework named OOD-Enhanced Few-Sample Model Compression (OE-FSMC), introducing out-of-distribution (OOD) samples with dynamically assigned labels to prevent bias during the compression process. To avoid overfitting the OOD samples, we incorporate a joint distillation loss and a class-dependent regularization term. Extensive experiments on multiple benchmark datasets show that our framework can be seamlessly incorporated into existing few-sample model compression methods, effectively mitigating the accuracy degradation caused by class imbalance.

Keywords: Few-shot learning, Network compression, Class imbalance, Image classification

✉. Corresponding author

This paper was accepted by *Pattern Recognition Letters* on 10 January 2026 (received 25 May 2025; revised 12 December 2025).

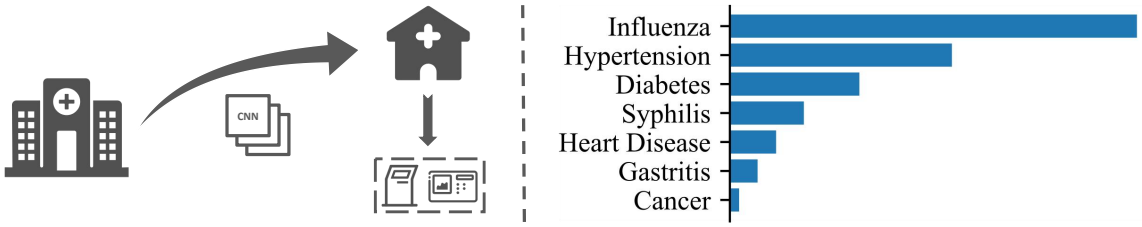


Figure 1: Overview of introduction. **Left:** Illustration of few-sample model compression workflow for deployment in small hospitals. **Right:** Class imbalance in disease cases.

1 Introduction

As deep learning models have grown in size and complexity, they require increasing computational and storage resources, which limits their deployment on edge devices such as cameras or smartwatches. To compress the model, network pruning methods [8, 13] try to remove less significant weights or channels, while knowledge distillation methods [10] let the compact model learn from soft labels of the pre-trained model, and quantization methods [18, 21] try to reduce the precision of model weights and activations. However, they typically rely on large datasets to maintain performance, which is often impractical in real-world scenarios constrained by privacy, security, or data acquisition limitations.

In domains such as healthcare and finance, where sensitive data are strictly limited, few-sample model compression methods [2, 14, 23, 24, 27] have gained prominence as a compromise between privacy and performance. For example, a small hospital that aims to deploy an intelligent diagnostic system may lack sufficient local data to train a high-quality model from scratch. With few-sample compression techniques, it can adopt a pre-trained model from a larger institution and use its limited data to compress and fine-tune that model (*cf.* Fig. 1). This paradigm enables lightweight deployment on edge devices while avoiding direct data usage.

Although previous few-sample model compression strategies have shown promising results in optimizing model performance with limited data, they have not considered the high likelihood of class imbalance occurring under the few-sample setting. These methods assume balanced class distributions under N-way K-shot settings (*i.e.*, K samples per class), which rarely reflect real-world conditions. For example, in medical diagnosis (*cf.* Fig. 1), common diseases like flu dominate, while rare conditions such as cancer lack sufficient samples. This imbalance introduces training bias and distorts the compression process, often reducing the model's ability to retain information from the minority class. Moreover, the complexity of the compression process causes the impact of class imbalance to accumulate across stages. As shown in Table 1, these challenges make most traditional class imbalance mitigation strategies ineffective under few-sample compression scenarios.

To resolve these challenges, we propose a novel framework, OOD-Enhanced Few-Sample Model Compression (OE-FSMC), which incorporates *out-of-distribution* (OOD) samples during the compression process to achieve dynamic balance. Inspired by Open-sampling [25],

for each OOD instance, we sample the label from a predefined complementary distribution to rebalance class priors. Differently, we dynamically adjust this distribution strategy at every stage to accommodate the complexity of compression. To handle extreme scenarios with few or even zero samples, we incorporate Laplace smoothing. Furthermore, we introduce a joint distillation loss and a class-dependent regularization term to prevent the model from overfitting OOD samples. Our main contributions can be summarized as follows:

- To the best of our knowledge, we are the first to point out and address the class imbalance problem in the few-sample model compression.
- We propose a novel framework, OOD-Enhanced Few-Sample Model Compression (OE-FSMC), that leverages out-of-distribution data to mitigate the class imbalance problem during the compression process.
- Our method readily integrates with mainstream few-sample model compression approaches, enhancing their robustness against class imbalance. Moreover, it is agnostic to the specific model architecture.

2 Related Work

2.1 Few-sample model compression

Few-sample model compression aims to derive compact models from pre-trained overparameterized networks using few samples. Bai et al. [2] developed cross distillation (CD), which interleaves the teacher and student hidden layers to suppress inter-layer error propagation. FSKD [14] introduces learnable 1×1 convolutions on student network blocks, optimizing auxiliary parameters to bridge block-level representation gaps with teacher model. MiR [24] aligns outputs at the penultimate layer of teacher and pruned student models, then substituted all layers except the head before the penultimate layer in the teacher model with the trained student model. Block dropping methods [23, 27] replace traditional filter pruning with block dropping. However, these methods overlook the risk of class imbalance in few-sample scenarios.

2.2 Class imbalance

In classification tasks, class imbalance occurs when the sample sizes of different classes vary significantly [19], causing the model to focus on majority classes and neglect critical minority-class information. Current solutions for addressing class imbalance problems can be classified into three main categories: data-level, algorithm-level, and hybrid approaches. Data-level methods achieve class balance by either removing majority-class samples [15–17] or generating additional minority-class samples [1, 4, 22]. Algorithm-level methods [6, 9, 28] attempt to mitigate the preference for majority classes by modifying existing machine learning algorithms. Hybrid methods typically combine data-level or algorithm-level strategies with ensemble learning [5, 7]. Although the class imbalance problem has been extensively studied across various tasks, research on the generalization performance of these methods in the context of few-sample model compression remains limited.

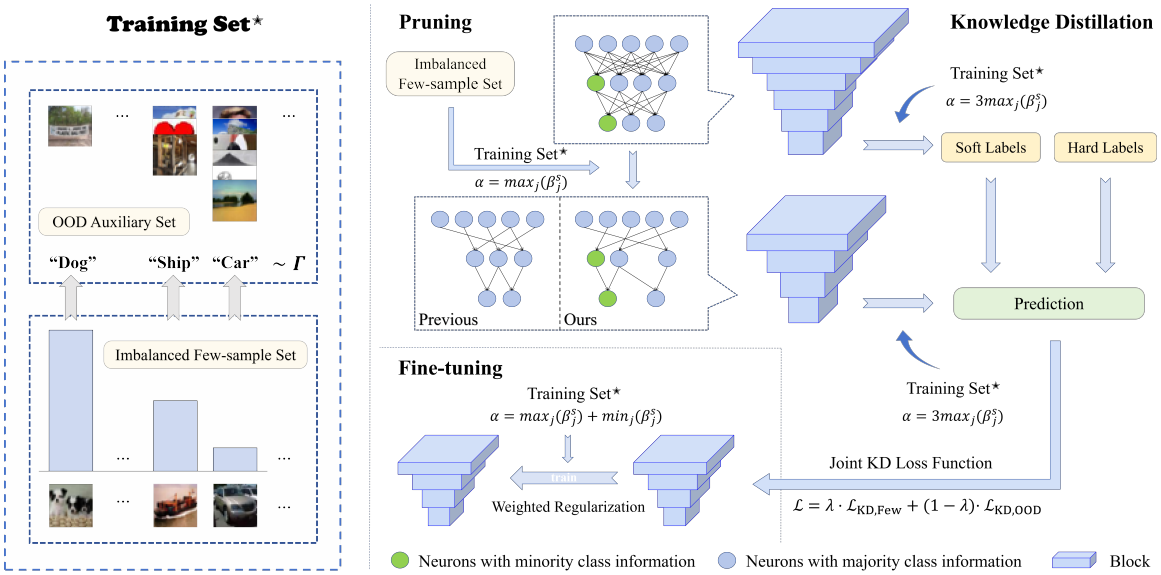


Figure 2: Illustration of OE-FSMC. **Left:** Label assignment strategy for OOD samples, where labels are dynamically assigned based on the complementary distribution. **Right:** Framework of OE-FSMC. The framework consists of three stages: (1) **Pruning:** The pre-trained teacher model is pruned using OOD samples to better retain channels associated with minority classes; (2) **Knowledge Distillation:** The student model learns from the teacher with a joint loss function balancing original and OOD data; (3) **Fine-tuning:** The model is fine-tuned with a class-dependent regularization term to mitigate overfitting.

3 Method

3.1 Problem definition

Few-sample Model Compression aims to get a compact model from the pre-trained redundant model with few samples for multi-class classification tasks, where the input space is represented by $\mathcal{X} \subset \mathbb{R}^d$, and the label space $\mathcal{Y}$ is $\{1, \ldots, K\}$. The pre-trained model is trained on the full dataset $\mathcal{D}_{\text{full}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$, $\boldsymbol{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, consisting of N samples. Under the few-sample setting, the dataset used for compression is denoted as $\mathcal{D}_{\text{few}} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{M}$, $\boldsymbol{x}_i \in \mathcal{X}$, $y_i \in \mathcal{Y}$, consisting of M samples, where $M \ll N$. Let m_j denote the number of samples of class j in $\mathcal{D}_{\text{few}}$, such that $M = \sum_{j=1}^{K} m_j$. We consider the **Class Imbalance** problem in $\mathcal{D}_{\text{few}}$. Let $P_{\text{full}}(X, Y)$ and $P_{\text{few}}(X, Y)$ denote the data distributions corresponding to the full dataset and the few-sample dataset, respectively. When suffering from class imbalance, the class-conditional distribution on the few-sample dataset remains the same as that of the full dataset (*i.e.*, $P_{\text{few}}(X|Y) = P_{\text{full}}(X|Y)$), while the class prior differs (*i.e.*, $P_{\text{few}}(Y) \neq P_{\text{full}}(Y)$). In addition, we assume that both the class-conditional distribution and the class prior on the test set are identical to those of the full dataset (*i.e.*, $P_{\text{test}}(X|Y) = P_{\text{full}}(X|Y)$, $P_{\text{test}}(Y) = P_{\text{full}}(Y)$).

Table 1: Top-1 accuracy of rebalancing methods when compressing VGG-16 on CIFAR-10 using cross-distillation [2] with a few-sample dataset of 10 samples.

Method	Accuracy (%)
Baseline (no rebalance)	70.34
SMOTE [4]	67.28
Undersampling [16]	63.73
ROS [11]	69.79
ROS+ [11]	69.82
SSP [26]	**71.05**
LDAM-RAW [3]	65.54
LDAM-DRW [3]	67.39
Balanced Softmax [20]	66.07
Open-sampling [25]	**70.80**

3.2 The urgent need to address class imbalance in few-sample model compression

In few-sample scenarios, it is inherently difficult to maintain class balance due to the scarcity of training data. Consequently, class imbalance becomes a common and often unavoidable issue. As research on few-sample model compression advances, it is imperative to recognize and address this issue. Unlike in the general case, the impact of class imbalance in few-sample compression amplifies as the compression process proceeds. Therefore, we consider this to be a more severe and complex problem. Specifically, pruning methods are more likely to remove channels associated with minority classes, and in knowledge distillation, the student model tends to absorb more information from majority classes. This leads to severely degraded performance on minority classes and poor generalization of the compressed model. Furthermore, fine-tuning exacerbates this bias, pushing the model even further toward the majority classes. The results in Table 2 have confirmed the existence of this problem and demonstrate its severe impact, thus urgently requiring effective solutions.

3.3 Feasibility of using OOD data

Ochal et al. [19] have demonstrated that some traditional imbalance mitigation methods can generalize well to few-shot learning. However, our experimental results in Table 1 show that these methods do not yield comparable effectiveness in the context of few-sample model compression. Traditional undersampling techniques inevitably discard already scarce data, which is especially detrimental in low-data regimes. Oversampling methods [11], on the other hand, are prone to overfitting on the minority class. Methods that rely on high-dimensional and continuous feature spaces [3, 20] also tend to fail, as the compression process reduces feature representation complexity. Approaches that incorporate in-distribution unlabeled data [26] perform relatively better, but they often incur high costs in few-sample model compression scenarios. Therefore, it remains a significant challenge to mitigate the

impact of class imbalance when both data availability and model capacity are severely constrained.

Under such circumstances, although the Open-Sampling method [25] does not achieve the highest accuracy, it has drawn our attention for its low data requirements and high computational efficiency. By leveraging out-of-distribution (OOD) samples to rebalance class priors, this method offers a lightweight yet effective solution. Furthermore, Theorem 1 provides a rigorous theoretical guarantee that incorporating OOD samples into the training process is provably non-toxic:

Theorem 1 (Wei et al. [25])**.** *When labels are uniformly sampled from the label space within the distribution, augmenting the training set* $\mathcal{D}_{\text{mix}} = \mathcal{D}_{\text{few}} \cup \mathcal{D}_{\text{OOD}}$ *does not affect the prediction of the Bayesian classifier:*

$$\arg\max_{y \in \mathcal{Y}} P_{\text{mix}}(\boldsymbol{x}|y) P_{\text{mix}}(y) = \arg\max_{y \in \mathcal{Y}} P_{\text{full}}(\boldsymbol{x}|y) P_{\text{full}}(y) \ , \tag{1}$$

where $P_{\text{mix}}(X, Y)$ *represents the underlying data distribution of* $\mathcal{D}_{\text{mix}}$*.*

Theorem 1 indicates that when OOD samples are assigned uniformly random labels (i.e., each class has an equal probability of being selected), these labels do not introduce systematic bias to the prior distribution $P_{\text{full}}(Y)$ of the original task. Even if the OOD data samples are unrelated to the target task, the uniformity of their labels ensures that their influence on the Bayesian decision boundary is effectively canceled out. There is no inherent limitation preventing this theorem from being generalized to the few-sample model compression scenario. This implies that the feature space of the student model remains aligned with that of the teacher model after incorporating OOD instances during the compression process. These theoretical results provide a solid foundation for us to design a framework based on OOD data to tackle the class imbalance problem during the few-sample model compression process.

3.4 Our method: A framework using OOD data

In this section, we propose a novel framework called OOD-Enhanced Few-Sample Model Compression (OE-FSMC), as illustrated in Fig. 2. It leverages the Open-sampling [25] technique to address the class imbalance problem in the few-sample model compression.

3.4.1 Label assignment strategy for OOD data

Open-sampling [25] employs a complementary distribution as the label assignment strategy for OOD samples. This strategy allocates more OOD instances to the minority class while ensuring the stability of the Bayesian classifier's prediction. Specifically, the complementary sampling rate Γ_j for class j is defined as:

$$\Gamma_j = \frac{\alpha - \beta_j}{K \cdot \alpha - 1} \ , \tag{2}$$

where $\beta_j = m_j / M$ represents the original distribution weight for class j, $\sum_{i=1}^{K} \Gamma_i = 1$ ensures that the sampling rates are normalized. The hyperparameter α controls the label assignment strategy. When $\alpha = \max_j(\beta_j)$, the assignment is biased toward minority classes, effectively

allocating more labels to minority classes, thus achieving the strongest rebalancing effect. In contrast, as $\alpha \to \infty$, the label distribution approaches uniformity, with each class receiving a probability of approximately $1/K$, thereby minimizing the potential toxicity of OOD data.

However, in few-sample settings, certain classes may have extremely limited or even zero samples. This situation often introduces bias into the original class prior estimation $\beta_j = m_j/M$, which in turn causes Γ_j in Eq. (2) to assign all OOD samples to that class, injecting significant noise. To address this, we apply Laplace smoothing to adjust the class prior estimate:

$$\beta_j^{\mathrm{s}} = \frac{m_j + \delta}{M + K \cdot \delta} \ , \tag{3}$$

where $\delta = 1$ serves as the smoothing factor. This correction ensures that even when $m_j = 0$, the smoothed prior $\beta_j^{\mathrm{s}} > 0$, which avoids instability during complementary distribution computation. The final complementary sampling rate for class j is defined as:

$$\Gamma_j^{\mathrm{s}} = \frac{\alpha - \beta_j^{\mathrm{s}}}{K \cdot \alpha - 1}. \tag{4}$$

3.4.2 OOD-enhanced few-sample model compression

OOD-enhanced pruning: We first prune the pre-trained teacher model. In this stage, class imbalance weakens the significance of minority classes, causing the pruning algorithm to incorrectly remove channels associated with them. Once pruned, these channels are difficult to recover in later stages, leading to irreversible loss.

To address this, we introduce OOD samples during pruning to rebalance the class prior. The label assignment strategy for OOD samples follows Eq. (4), where we set $\alpha = \max_j(\beta_j^{\mathrm{s}})$ to retain as many minority-related channels as possible. Although this leads to more aggressive pruning of majority-class channels, they are relatively easier to recover in subsequent stages.

OOD-enhanced knowledge distillation: We distill knowledge from the teacher into the pruned student model. In this stage, class imbalance causes the student's logits for minority classes to update less aggressively.

To force greater emphasis on minority classes, we inject OOD data during distillation and apply the label assignment strategy from Eq. (4). We observe that OOD data exhibits heightened toxicity during the distillation stage, likely due to a mismatch between complementary distribution labels and the teacher's logits. Interestingly, this "toxic" signal enhances model robustness. Empirically, we find the balance point $\alpha = 3\max_j(\beta_j^{\mathrm{s}})$ reduces the toxicity of OOD samples while maximizing the robustness of the model.

To further prevent overfitting to OOD samples during distillation, we compute a joint distillation loss over the original few-sample dataset and the auxiliary OOD samples:

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{\mathrm{KD,Few}} + (1 - \lambda) \cdot \mathcal{L}_{\mathrm{KD,OOD}} \ , \tag{5}$$

where λ is a balancing coefficient that adjusts the relative contributions of the original dataset and the OOD instances to the knowledge distillation process.

Algorithm 1 OE-FSMC.

Input: Pre-trained teacher model T, Imbalanced few-sample dataset $\mathcal{D}_{\text{few}} = \{(x_i, y_i)\}_{i=1}^{M}$, OOD dataset $\mathcal{D}_{\text{OOD}}$.

Output: Compact student model S.

Initialization: Calculate class prior distribution β_j, smoothed prior β_j^{s} (Eq. (3)), and smoothed complementary sampling rate Γ_j^s (Eq. (4)).

// Stage 1: OOD-enhanced pruning

Set $\alpha \leftarrow \max_j(\beta_j^{\text{s}})$.

Sample OOD batch and assign labels using $\Gamma^s(\alpha)$.

Obtain the initial student S by pruning a copy of T using mixed minibatches from $\mathcal{D}_{\text{few}}$ and $\mathcal{D}_{\text{OOD}}$.

// Stage 2: OOD-enhanced knowledge distillation

Set $\alpha \leftarrow 3\max_j(\beta_j^{\text{s}})$.

for epoch $= 1$ to E_{KD} **do**

  Sample (x, y) from $\mathcal{D}_{\text{few}}$ and $\tilde{x}$ from $\mathcal{D}_{\text{OOD}}$.

  Assign complementary labels $\tilde{y} \sim \Gamma^s(\alpha)$ to $\tilde{x}$.

  Compute Joint Distillation Loss $\mathcal{L}$ (Eq. (5)).

  Update S by back-propagation.

end for

// Stage 3: OOD-enhanced fine-tuning

Set $\alpha \leftarrow \max_j(\beta_j^{\text{s}}) + \min_j(\beta_j^{\text{s}})$.

for epoch $= 1$ to E_{FT} **do**

  Sample (x, y) from $\mathcal{D}_{\text{few}}$ and $\tilde{x}$ from $\mathcal{D}_{\text{OOD}}$.

  Assign complementary labels $\tilde{y} \sim \Gamma^s(\alpha)$ to $\tilde{x}$.

  Compute Regularized Loss $\mathcal{L}_{\text{total}}$ (Eq. (8)).

  Update weights of S.

end for

return Student model S.

The distillation losses for the original few-sample dataset and the OOD samples can be expressed as:

$$\begin{aligned} \mathcal{L}_{\text{KD,Few}} &= \frac{1}{M}\sum_{i=1}^{M}\sum_{y} P_T(y|\boldsymbol{x}_i)\log\frac{P_T(y|\boldsymbol{x}_i)}{P_S(y|\boldsymbol{x}_i)}\ , \\ \mathcal{L}_{\text{KD,OOD}} &= \frac{1}{|\mathcal{D}_{\text{OOD}}|}\sum_{i=1}^{|\mathcal{D}_{\text{OOD}}|}\sum_{\tilde{y}} P_T(\tilde{y}\mid\tilde{\boldsymbol{x}}_i)\log\frac{P_T(\tilde{y}\mid\tilde{\boldsymbol{x}}_i)}{P_S(\tilde{y}\mid\tilde{\boldsymbol{x}}_i)}\ , \end{aligned} \tag{6}$$

where $|\mathcal{D}_{\text{OOD}}|$ is the size of OOD samples, $\tilde{\boldsymbol{x}}_i$ is the ith OOD sample, $\tilde{y}$ is sampled from the complementary distribution Γ_j^{s}, $P_T(y|\boldsymbol{x})$ is the teacher model's predicted probability of label y for input x and $P_S(y|\boldsymbol{x})$ is the student model's predicted probability of the same.

OOD-enhanced fine-tuning: To mitigate class imbalance during fine-tuning, we introduce OOD samples and apply the label assignment strategy from Eq. (4). We set $\alpha = \max_j(\beta_j^{\text{s}}) + \min_j(\beta_j^{\text{s}})$, which represents the "sweet spot" between OOD toxicity and

imbalance correction [25] and we find that this choice remains optimal under the current condition.

Directly applying standard cross-entropy loss in this setting may result in overfitting to the OOD samples, leading to convergence issues. To address this, we introduce a regularization term for the OOD instances:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\tilde{\boldsymbol{x}} \sim P_{\text{ood}}(X)} \left[\gamma_{\tilde{y}} \cdot \ell \left(f(\tilde{\boldsymbol{x}}; \boldsymbol{\theta}), \tilde{y} \right) \right] \ , \tag{7}$$

where $\ell(\cdot)$ is the cross-entropy loss and $\gamma_{\tilde{y}} = \Gamma_j^{\text{s}} \cdot K$ is a class-dependent weight factor. The overall loss function is:

$$\begin{aligned} \mathcal{L}_{\text{total}} = & \mathbb{E}_{((\boldsymbol{x},y) \sim P_{\text{few}}(X,Y))} \left[\ell \left(f(\boldsymbol{x}; \boldsymbol{\theta}), y \right) \right] \\ & + \eta \cdot \mathbb{E}_{(\tilde{\boldsymbol{x}}) \sim P_{\text{ood}}(X)} \left[\gamma_{\tilde{y}} \cdot \ell \left(f(\tilde{\boldsymbol{x}}; \boldsymbol{\theta}), \tilde{y} \right) \right]. \end{aligned} \tag{8}$$

Here, η controls the strength of the regularization term.

Relation to open-sampling: In this work, we extend the Open-sampling [25] algorithm to the few-sample model compression setting by incorporating Laplace smoothing for class-prior correction, replacing the fixed "sweet-spot" α with a dynamic, stage-wise OOD label assignment strategy, and drastically reducing the required OOD sample count by adopting $|\mathcal{D}_{\text{OOD}}| = M$ instead of the original $|\mathcal{D}_{\text{OOD}}| \gg M$. Our framework unifies pruning, distillation, and fine-tuning under strict data constraints, yielding a generic, extensible solution for few-sample model compression.

To clarify the multi-stage workflow of OE-FSMC, we provide the pseudocode in Algorithm 1.

4 Experiment

To address the following research questions, we conduct extensive experiments on three publicly available datasets:

- **RQ1:** Is class imbalance a significant issue for few-sample model compression methods?
- **RQ2:** Can our framework effectively mitigate the issue of class imbalance in few-sample model compression, and is it compatible with current state-of-the-art few-sample model compression methods?
- **RQ3:** How effective is each component of our framework?

4.1 Experiment setting

Data description: To validate that our framework effectively alleviates the class imbalance issue in few-sample model compression, we select image classification datasets CIFAR-10 ($K = 10$), CIFAR-100 ($K = 100$), and ILSVRC-2012 ($K = 1000$), where K denotes the number of classes. In the few-sample setting, to facilitate comparison with balanced scenarios, we sample a total of $C \times K$ examples, where $C \in \{1, 2, 3, 5, 10\}$, corresponding to balanced configurations with C samples per class. For each class, the number of samples is set according to the imbalance standard in the long-tailed CIFAR-10 / 100 [12], adopting a

Table 2: Test accuracy (%) of VGG-16 on long-tailed CIFAR-10 under balanced and imbalanced conditions. ↓ indicates the accuracy loss in the imbalanced case.

Num	10		20		50		100	
Method	Balance	Imbalance	Balance	Imbalance	Balance	Imbalance	Balance	Imbalance
CD	73.89	70.34(**3.55**↓)	81.65	76.39(**5.26**↓)	82.86	80.47(**2.39**↓)	83.56	82.24(**1.32**↓)
FSKD	76.18	73.51(**2.67**↓)	84.20	81.66(**2.54**↓)	88.63	86.31(**2.32**↓)	89.18	87.41(**1.77**↓)
MiR	75.42	72.84(**2.58**↓)	79.18	76.57(**2.61**↓)	84.27	82.49(**1.78**↓)	85.71	83.85(**1.86**↓)
PRACTISE	76.91	75.78(**1.13**↓)	82.07	81.26(**0.81**↓)	86.35	85.98(**0.37**↓)	89.71	89.35(**0.35**↓)
DC-ViT	70.47	67.62(**2.85**↓)	77.58	74.39(**3.19**↓)	83.77	81.85(**1.92**↓)	85.46	84.12(**1.34**↓)

long-tailed distribution with an imbalance ratio of 100. OOD samples are randomly sampled from the Tiny Images dataset and filtered to remove those that overlap with the training data.

Compared methods: To demonstrate the effectiveness of the proposed OE-FSMC in mitigating the class imbalance, it is integrated with the most authoritative few-sample model compression methods at present. These include the cross-distillation approach CD [2], the block-level alignment method FSKD [14], the layer-replacement strategy MIR [24], and block dropping approaches PRACTISE [23] and DC-ViT [27].

Evaluation metric: As in previous work, we evaluate the effectiveness of our framework based on top-1 accuracy. For each method, we conduct five independent experiments and report the mean and standard deviation.

Parameter setting: In our experimental setup, the number of incorporated OOD samples is configured to match the size of the few-sample training set, with a batch size of 128 and a learning rate of 0.0005. It is important to emphasize that we retain the pruning rate, learning rate, and other hyperparameter settings as specified in the original source code and corresponding papers [2, 14, 23, 24, 27]. Our goal is to analyze the shared challenges across different methods, rather than to compare their relative performance. This setup enables a fair evaluation of model behavior under class imbalance and highlights both the effectiveness and generalizability of the proposed OE-FSMC approach. All experiments are conducted on an RTX-4090 24GB GPU and an I9-13900KF CPU.

4.2 Validation and analysis (RQ1)

The results in Table 2 show that when the dataset suffers from class imbalance, the performance of existing few-sample model compression methods significantly decreases. Among them, the PRACTISE method is less affected, likely because the block-dropping strategy is less sensitive to imbalance. However, under imbalance conditions, its latency increases significantly, and since "latency-accuracy" is the key evaluation metric emphasized by PRACTISE, its performance is also affected by the imbalance issue in some ways.

4.3 Effectiveness and generalizability (RQ2)

In this experiment, we evaluated the effectiveness of the OE-FSMC method on different model architectures and datasets, specifically including VGG-16 on long-tail CIFAR-10, ResNet-32 on long-tail CIFAR-100, and ResNet-34 on long-tail ILSVRC-2012. Table 3

Table 3: Test accuracy (%) of VGG-16 on long-tailed CIFAR-10, ResNet-32 on CIFAR-100 and ResNet-34 on long-tailed ILSVRC-2012 with different training set sizes.

Dataset	Long-tailed CIFAR-10			Long-tailed CIFAR-100			Long-tailed ILSVRC-2012		
Num	10	20	50	100	200	500	1000	2000	3000
CD	70.34 ± 1.31	76.39 ± 0.84	80.47 ± 0.45	53.38 ± 0.95	62.71 ± 0.40	67.40 ± 0.27	69.49 ± 0.34	70.07 ± 0.19	70.14 ± 0.22
+ Ours	$\mathbf{78.90 \pm 0.33}$	$\mathbf{81.02 \pm 0.28}$	$\mathbf{82.80 \pm 0.24}$	$\mathbf{56.33 \pm 0.47}$	$\mathbf{65.18 \pm 0.35}$	$\mathbf{68.84 \pm 0.16}$	$\mathbf{69.84 \pm 0.22}$	$\mathbf{70.56 \pm 0.20}$	$\mathbf{70.58 \pm 0.18}$
FSKD	73.51 ± 0.85	81.66 ± 0.72	86.31 ± 0.54	57.78 ± 0.39	63.29 ± 0.26	68.15 ± 0.05	68.54 ± 0.15	69.95 ± 0.17	70.27 ± 0.19
+ Ours	$\mathbf{75.93 \pm 0.46}$	$\mathbf{84.67 \pm 0.32}$	$\mathbf{88.24 \pm 0.25}$	$\mathbf{59.46 \pm 0.21}$	$\mathbf{65.88 \pm 0.19}$	$\mathbf{69.17 \pm 0.07}$	$\mathbf{69.06 \pm 0.16}$	$\mathbf{70.18 \pm 0.10}$	$\mathbf{70.60 \pm 0.09}$
MiR	72.84 ± 0.35	76.57 ± 0.41	82.49 ± 0.22	57.69 ± 0.18	64.92 ± 0.30	68.70 ± 0.13	66.36 ± 0.17	67.23 ± 0.15	67.48 ± 0.07
+ Ours	$\mathbf{75.25 \pm 0.42}$	$\mathbf{77.93 \pm 0.35}$	$\mathbf{83.60 \pm 0.27}$	$\mathbf{60.37 \pm 0.33}$	$\mathbf{66.53 \pm 0.24}$	$\mathbf{69.83 \pm 0.17}$	$\mathbf{66.85 \pm 0.20}$	$\mathbf{67.97 \pm 0.11}$	$\mathbf{68.12 \pm 0.05}$
PRACTISE	75.78 ± 0.24	81.26 ± 0.21	85.98 ± 0.19	58.42 ± 0.13	65.14 ± 0.18	69.02 ± 0.05	70.90 ± 0.12	71.89 ± 0.08	72.41 ± 0.03
+ Ours	$\mathbf{76.59 \pm 0.15}$	$\mathbf{81.77 \pm 0.08}$	$\mathbf{86.25 \pm 0.12}$	$\mathbf{59.57 \pm 0.09}$	$\mathbf{66.08 \pm 0.05}$	$\mathbf{69.62 \pm 0.10}$	$\mathbf{71.52 \pm 0.11}$	$\mathbf{72.17 \pm 0.06}$	$\mathbf{72.85 \pm 0.09}$
DC-ViT	67.62 ± 0.76	74.39 ± 0.35	81.85 ± 0.41	57.83 ± 0.30	64.40 ± 0.28	68.38 ± 0.26	72.04 ± 0.19	73.35 ± 0.02	74.71 ± 0.10
+ Ours	$\mathbf{69.02 \pm 0.31}$	$\mathbf{75.94 \pm 0.25}$	$\mathbf{83.27 \pm 0.20}$	$\mathbf{58.33 \pm 0.45}$	$\mathbf{64.96 \pm 0.29}$	$\mathbf{69.01 \pm 0.16}$	$\mathbf{72.97 \pm 0.03}$	$\mathbf{74.09 \pm 0.05}$	$\mathbf{75.25 \pm 0.02}$

Table 4: Results of ablation study.

PR	KD	FT	CD	FSKD	MiR	PRACTISE	DC-ViT
			70.34	73.51	72.84	75.78	67.62
✓			73.18	73.97	73.39	76.03	68.13
✓	✓		75.54	74.61	73.98	76.26	68.45
✓	✓	✓	**78.90**	**75.93**	**75.25**	**76.59**	**69.02**

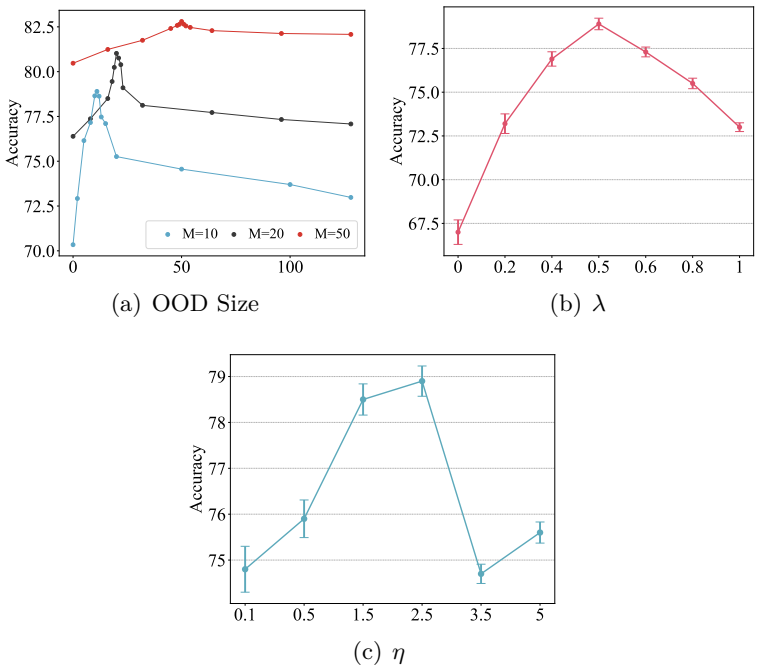
presents the performance of several mainstream few-sample model compression strategies before and after combining with OE-FSMC, with a focus on addressing the class imbalance problem. The results clearly show that after incorporating the OE-FSMC method, the accuracy of each few-sample model compression strategy improved significantly, sometimes even outperforming the results obtained under balanced conditions. Notably, the performance improvement brought by OE-FSMC is more pronounced when the number of samples is small, suggesting that OE-FSMC is particularly effective in mitigating class imbalance in scenarios with limited data.

4.4 Ablation study (RQ3)

As detailed in the Methods section, OE-FSMC can be divided into three components, PR, KD, and FT, corresponding to our enhancement strategies in the pruning, knowledge distillation, and fine-tuning stages, respectively. To evaluate the effectiveness of each component, we conduct ablation experiments on VGG-16 with the CIFAR-10 dataset and adopt five popular few-sample compression methods as baseline compression architectures. Table 4 presents the quantitative results for different combinations of the components. As shown, each component independently improves performance under class imbalance, and the results are optimal when all three components are combined, demonstrating the strongest robustness to class imbalance.

4.5 Hyperparameter sensitivity analysis

To investigate the impact of the size of OOD samples on model performance, we conducted experiments under three different training set sizes. The results in Fig. 3(a) demonstrate



(a) OOD Size (b) λ

(c) η

Figure 3: Results of hyperparameter experiments. In this figure: a) Influence of the number of OOD samples. b) Sensitivity analysis of $\lambda \in [0, 1]$. c) Sensitivity analysis of $\eta \in [0.1, 5]$.

that the model achieves the best performance when the size of the OOD samples is comparable to the original training set. This differs from the conclusion in [25], which suggests that the number of OOD samples should significantly exceed that of the training set. This discrepancy may arise because OOD samples exhibit stronger toxicity under few-sample conditions.

Fig. 3(b) illustrates the effect of different values of λ in Eq. (5) on model accuracy. We vary its value in $\{0, 0.2, 0.4, 0.5, 0.6, 0.8, 1.0\}$, and the result indicates that the model achieves the best performance when $\lambda = 0.5$. This indicates that assigning equal weight to the original training data and the OOD samples during knowledge distillation achieves a balanced outcome, effectively preserving the features of minority classes while avoiding over-reliance on OOD samples.

To further explore the effect of η in Eq. (8), we search η in $\{0.1, 0.5, 1.5, 2.5, 3.5, 5\}$. From Fig. 3(c), we observe that the model performance improves significantly when $\eta = 2.5$. This finding suggests that moderate regularization effectively constrains the model's parameter search space, thereby enhancing generalization. When η is too large, the regularization term dominates, suppressing the model's ability to learn features from both the training and OOD samples, leading to a decline in performance. Overall, our ablation experiments provide additional evidence for the effectiveness of our proposed method, further demonstrating that it can achieve strong performance under reasonable parameter configurations.

Table 5: Test accuracy (%) of VGG-16 on long-tailed CIFAR-10, ResNet-32 on long-tailed CIFAR-100, and ResNet-34 on long-tailed ILSVRC-2012 under different class imbalance ratios.

Dataset	Long-tailed CIFAR-10			Long-tailed CIFAR-100			Long-tailed ILSVRC-2012		
Imbalance Ratio	IR=10	IR=50	IR=100	IR=10	IR=50	IR=100	IR=10	IR=50	IR=100
CD	82.87 ± 0.38	82.37 ± 0.27	82.24 ± 0.30	69.63 ± 0.22	68.52 ± 0.13	68.44 ± 0.15	70.98 ± 0.14	70.40 ± 0.11	70.35 ± 0.07
+ Ours	$\mathbf{83.62 \pm 0.31}$	$\mathbf{83.53 \pm 0.30}$	$\mathbf{83.50 \pm 0.25}$	$\mathbf{70.37 \pm 0.16}$	$\mathbf{70.01 \pm 0.10}$	$\mathbf{69.97 \pm 0.12}$	$\mathbf{71.35 \pm 0.12}$	$\mathbf{70.84 \pm 0.08}$	$\mathbf{70.81 \pm 0.05}$
FSKD	88.12 ± 0.27	87.59 ± 0.31	87.41 ± 0.24	70.89 ± 0.13	69.47 ± 0.06	69.25 ± 0.08	71.12 ± 0.19	70.71 ± 0.16	70.60 ± 0.15
+ Ours	$\mathbf{89.01 \pm 0.26}$	$\mathbf{88.92 \pm 0.18}$	$\mathbf{88.84 \pm 0.16}$	$\mathbf{71.94 \pm 0.10}$	$\mathbf{71.23 \pm 0.05}$	$\mathbf{71.05 \pm 0.02}$	$\mathbf{71.53 \pm 0.17}$	$\mathbf{71.41 \pm 0.09}$	$\mathbf{71.27 \pm 0.12}$
MiR	84.59 ± 0.21	84.04 ± 0.19	83.85 ± 0.14	71.03 ± 0.15	69.91 ± 0.17	69.79 ± 0.14	69.67 ± 0.05	68.82 ± 0.04	68.68 ± 0.07
+ Ours	$\mathbf{85.50 \pm 0.18}$	$\mathbf{85.32 \pm 0.20}$	$\mathbf{85.25 \pm 0.16}$	$\mathbf{72.18 \pm 0.13}$	$\mathbf{71.77 \pm 0.08}$	$\mathbf{71.70 \pm 0.08}$	$\mathbf{69.85 \pm 0.08}$	$\mathbf{69.16 \pm 0.06}$	$\mathbf{69.09 \pm 0.04}$
PRACTISE	89.49 ± 0.13	89.39 ± 0.05	89.35 ± 0.02	71.28 ± 0.08	70.29 ± 0.05	70.14 ± 0.03	73.91 ± 0.14	73.37 ± 0.12	73.29 ± 0.10
+ Ours	$\mathbf{89.71 \pm 0.09}$	$\mathbf{89.65 \pm 0.02}$	$\mathbf{89.61 \pm 0.04}$	$\mathbf{72.35 \pm 0.04}$	$\mathbf{70.86 \pm 0.06}$	$\mathbf{70.78 \pm 0.04}$	$\mathbf{74.12 \pm 0.11}$	$\mathbf{73.84 \pm 0.07}$	$\mathbf{73.72 \pm 0.08}$
DC-ViT	84.66 ± 0.26	84.25 ± 0.28	84.12 ± 0.27	70.42 ± 0.29	69.60 ± 0.18	69.47 ± 0.23	75.81 ± 0.15	75.32 ± 0.14	75.25 ± 0.11
+ Ours	$\mathbf{85.90 \pm 0.22}$	$\mathbf{85.73 \pm 0.23}$	$\mathbf{85.66 \pm 0.22}$	$\mathbf{71.28 \pm 0.16}$	$\mathbf{70.21 \pm 0.09}$	$\mathbf{70.13 \pm 0.11}$	$\mathbf{76.17 \pm 0.12}$	$\mathbf{75.76 \pm 0.10}$	$\mathbf{75.73 \pm 0.07}$

Table 6: Training-time overhead.

Method	Inference FLOPs	Training Time (min)	Overhead
FSKD	1×	102	–
FSKD+OE-FSMC	1×	115	+12.7%

4.6 Robustness under different imbalance ratios

We further investigated the robustness of our method by varying the imbalance ratio and summarizing the results in Table 5. Overall, the change in accuracy induced by increasing the imbalance ratio is modest under a fixed extreme few-sample budget, because the very limited number of labeled samples constrains the granularity of the data distribution, so that the per-class sample counts under different imbalance ratio settings remain statistically similar. Nevertheless, across all datasets and baseline methods, our approach consistently yields performance gains at all considered imbalance ratios, indicating stable generalization across imbalance levels. Moreover, compared with the original baselines, OE-FSMC mitigates the performance degradation caused by higher imbalance ratios, providing further evidence of its robustness to varying degrees of class imbalance.

4.7 Computational cost analysis

We compare the end-to-end training time of OE-FSMC with a strong few-sample compression baseline (FSKD) under identical settings, and summarize the results in Table 6. Since OE-FSMC only introduces additional OOD samples during training, the inference-time FLOPs and memory footprint of the students remain identical to their baselines. On CIFAR-100 with IR = 100, OE-FSMC takes 115 minutes compared to 102 minutes for FSKD, representing a 12.7% increase. This additional cost mainly stems from OOD sampling and complementary label generation, but remains modest because the OOD set is reused across all three stages and has the same size as the few-sample training set. Given the consistent accuracy improvements reported in Tables 3 and 5, this overhead appears acceptable for practical deployment.

5 Conclusion

In this paper, we proposed OE-FSMC, a novel few-sample model compression framework that enhances robustness against class imbalance during the compression process by leveraging OOD samples. To the best of our knowledge, we are the first to identify and resolve the class imbalance problem in the context of few-sample model compression. We followed the label assignment idea of Open-sampling [25], but we further integrated Laplace smoothing for few-shot scenarios and dynamically adjusted the assignment strategy at each stage. In addition, we employed a joint distillation loss and a class-dependent regularization to prevent overfitting.

CRediT authorship contribution statement

Tian-Shuang Wu: Writing – Original draft, Methodology, Formal analysis, Conceptualization. **Shen-Huan Lyu:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Yanyan Wang:** Writing – review & editing, Validation, Resources. **Ning Chen:** Writing – review & editing, Visualization, Software. **Zihao Qu:** Writing – review & editing, Validation, Funding acquisition. **Baolu Ye:** Writing – review & editing, Validation, Funding acquisition.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments and Disclosure of Funding

This work was supported by the National Natural Science Foundation of China (No. 62306104, 62441225 and 62572171), Basic Research Program of Jiangsu (No. BK20253011), Hong Kong Scholars Program (No. XJ2024010), Research Grants Council of the Hong Kong Special Administrative Region, China (GRF Project No. CityU11212524), Natural Science Foundation of Jiangsu Province (No. BK20230949), Jiangsu Association for Science and Technology (No. JSTJ2024285), China Postdoctoral Science Foundation (No. 2023TQ0104).

References

[1] L. Abdi and S. Hashemi. To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE Transactions on Knowledge and Data Engineering*, 28 (1):238–251, 2015.

[2] H. Bai, J. Wu, I. King, and M. Lyu. Few shot network compression via cross distillation. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 3203–3210, 2020.

[3] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in Neural Information Processing Systems*, 32, 2019.

[4] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[5] N. V. Chawla, A. Lazarevic, L. O. Hall, and K. W. Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *Proceedings of 7th European Conference on Principles and Practice of Knowledge Discovery in Databases*, pages 107–119, 2003.

[6] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, F. Herrera, A. Fernández, S. García, M. Galar, R. C. Prati, et al. Cost-sensitive learning. In *Learning from Imbalanced Data Sets*, pages 63–78. Springer, Cham, 2018.

[7] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 42(4): 463–484, 2011.

[8] Y. He, X. Zhang, and J. Sun. Channel pruning for accelerating very deep neural networks. In *Proceedings of the 16th the IEEE International Conference on Computer Vision*, pages 1398–1406, 2017.

[9] Y.-X. He, D.-X. Liu, S.-H. Lyu, C. Qian, and Z.-H. Zhou. Multi-class imbalance problem: A multi-objective solution. *Information Sciences*, 680:121156, 2024.

[10] G. Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[11] N. Japkowicz and S. Stephen. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5):429–449, 2002.

[12] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical Report UT-ML-TR-2009-001, University of Toronto, 2009.

[13] H. Li, A. Kadav, I. Durdanovic, H. Samet, and H. P. Graf. Pruning filters for efficient convnets. In *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[14] T. Li, J. Li, Z. Liu, and C. Zhang. Few sample knowledge distillation for efficient network compression. In *Proceedings of the 33rd IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14639–14647, 2020.

[15] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang. Clustering-based undersampling in class-imbalanced data. *Information Sciences*, 409:17–26, 2017.

[16] X.-Y. Liu, J. Wu, and Z.-H. Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 39(2):539–550, 2008.

[17] R. Mohammed, J. Rawashdeh, and M. Abdullah. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In *Proceedings of the 11th International Conference on Information and Communication Systems*, pages 243–248, 2020.

[18] M. Nagel, M. v. Baalen, T. Blankevoort, and M. Welling. Data-free quantization through weight equalization and bias correction. In *Proceedings of the 17th IEEE/CVF International Conference on Computer Vision*, pages 1325–1334, 2019.

[19] M. Ochal, M. Patacchiola, J. Vazquez, A. Storkey, and S. Wang. Few-shot learning with class imbalance. *IEEE Transactions on Artificial Intelligence*, 2023.

[20] J. Ren, C. Yu, X. Ma, H. Zhao, S. Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems*, 33:4175–4186, 2020.

[21] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio. Fitnets: Hints for thin deep nets. In *Proceedings of 3rd International Conference on Learning Representations*, pages 1–13, 2015.

[22] S. Sharma, A. Gosain, and S. Jain. A review of the oversampling techniques in class imbalance problem. In *Proceedings of the 5th International Conference on Innovative Computing and Communications*, pages 459–472, 2022.

[23] G.-H. Wang and J. Wu. Practical network acceleration with tiny sets. In *Proceedings of the 36th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20331–20340, 2023.

[24] H. Wang, J. Liu, X. Ma, Y. Yong, Z. Chai, and J. Wu. Compressing models with few samples: Mimicking then replacing. In *Proceedings of the 35th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 701–710, 2022.

[25] H. Wei, L. Tao, R. Xie, L. Feng, and B. An. Open-sampling: Exploring out-of-distribution data for re-balancing long-tailed datasets. In *Proceedings of the 39th International Conference on Machine Learning*, pages 23615–23630, 2022.

[26] Y. Yang and Z. Xu. Rethinking the value of labels for improving class-imbalanced learning. In *Advances in Neural Information Processing Systems 33*, pages 19290–19301, 2020.

[27] H. Zhang, Y. Zhou, and G.-H. Wang. Dense vision transformer compression with few samples. In *Proceedings of the 37th IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15825–15834, 2024.

[28] Z.-H. Zhou and X.-Y. Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on Knowledge and Data Engineering*, 18(1):63–77, 2005.