Journal of Information Hiding and Multimedia Signal Processing ©2025 ISSN 2073-4212
Ubiquitous International Volume **16**, Number **3**, September **2025**

FsKD: Few-shot learning using student-teacher knowledge distillation for behavior recognition

Thien Le Quang

Faculty of Management Information Systems
Ho Chi Minh University of Banking
Ho Chi Minh City, Vietnam
thienlq@hub.edu.vn

Received February 27, 2025, revised June 10, 2025, accepted June 13, 2025.

Abstract. *Learning from few samples combined with increasing the number of layers is employed to address challenges arising from shifting data distributions and the difficulties in collecting data for classroom behavior recognition. To mitigate the issue of forgetting previously learned information when training on new data, a student-teacher based knowledge distillation approach is utilized to preserve the learned feature distributions. Unlike traditional feature distillation methods, we perform feature distillation via dimensional projection—that is, projecting the features into an alternative space where knowledge distillation becomes more tractable. Additionally, we implement a sample selection model that leverages the ability to increase layers during inference to adjust weights, thereby enhancing the classroom behavior recognition process. Our proposed model is evaluated both qualitatively and quantitatively on two benchmark datasets, namely ImageNet and D-Edu (a dataset for classroom behavior recognition), to demonstrate its effectiveness.*
Keywords: Classroom behavior recognition, few-shot learning, knowledge distillation

1. **Introduction.** In recent years, artificial intelligence [1, 2, 3] has made significant strides in the field of computer vision, primarily owing to the training of models on large-scale data. However, real-world datasets are relatively scarce, necessitating deep learning models that can address this limitation. Most deep learning models focus on optimizing parameters based on existing data and do not adequately account for new data—often leading to issues such as catastrophic forgetting when training is continued. Consequently, incremental class learning with few-shot deep learning [4] has emerged to tackle the challenge of incorporating new data.

In the context of classroom settings, detecting students who engage in inappropriate behavior is a critical problem [5]. Developing deep learning models to recognize student data in classrooms can enable educators to concentrate more on enhancing teaching quality. Therefore, constructing modern deep learning models—such as few-shot learning approaches—to recognize student behavior in class is both feasible and important.

Moreover, relying solely on convolutional neural networks makes it difficult for most models to recognize student behavior via classroom cameras, primarily because the subjects are relatively small. To overcome this, we have established student-teacher model pairs to integrate relational information among classes. This approach preserves the data structure during the knowledge distillation process. However, issues may arise due to varying degrees of feature similarity across different feature spaces. To enhance feature connectivity and prevent performance degradation during training, we propose projecting features into a common reference space to support the student-teacher model training,

thereby improving the accuracy of student behavior recognition. We refer to this proposed model as FsKD.

In summary, our contributions include:

- We propose a few-shot deep learning method based on cross-projected knowledge distillation (FsKD) to preserve knowledge when new classes are added, effectively mapping features into a common, appropriate feature space while maintaining accurate recognition of student behavior.
- We enhance the model's capability to learn in data-scarce environments by designing a student-teacher network that provides mutual support during the training of image recognition models.
- We conduct experiments on two widely-used datasets, ImageNet and D-Edu, to demonstrate the effectiveness and applicability of our proposed method.

2. **Related Works.**

2.1. **Few-shot learning.** Few-shot deep learning methods (FSL) [6, 7, 8] rely on meta-learning [6] and metric-based approaches [9] to support image recognition. Some research groups have focused on domain-based FSL methods [10], class-enhanced FSL [11], etc., aiming to advance practical applications. Due to limitations in data availability and high labeling costs, there is a growing demand for developing FSL models—especially in the context of classroom student behavior image recognition and, more broadly, in the field of education.

Li [12] designed a strategy for selecting random datasets to generate pseudo-samples in order to improve image recognition. Deng [13] proposed an attention mechanism combined with FSL to recognize student actions. Xiao [14] introduced a method for learning across new and old classes to align the model's weights. Zhang [15] defined more explicit structures and implemented class-level upgrades during the instance augmentation process to enhance image recognition quality. Overall, these studies have attempted to leverage shared features and partition them into distinct clusters to support inference, although the accuracy remains suboptimal.

To provide context for our experimental comparisons, we briefly introduce several key FSCIL methods. FCIL is often considered a baseline approach that adapts a standard classifier using a combination of a cross-entropy loss for new classes and a distillation loss on old class logits to mitigate forgetting. MetaFSCIL extends meta-learning principles, proposing to learn a generalized feature space with prototypical representations that can be rapidly adapted to new classes with few samples. Other works have focused on preserving the feature space structure; for example, FSCIL-ASP (Adaptive Structure Preservation) introduces a loss to explicitly maintain the geometric relationships between old class prototypes after new classes are added. Similarly, FSCIL ALICE (Adaptive Learning with Inter-Class Embedding) learns to calibrate the features of new classes by aligning them with the distribution of existing class embeddings, thus ensuring a more stable and unified feature space across incremental sessions.

2.2. **Knowledge Distillation.** Knowledge distillation [16, 17, 18] is a method for compressing and accelerating models, but in recent years it has been enhanced to help mitigate the forgetting of information during incremental training. Knowledge distillation relies on teacher networks and student networks that mutually train each other. Tang et al. [19] proposed that students can learn from the features of teacher classes to improve image recognition. According to Borza [20], knowledge distillation increases supervisory information and enhances the quality of training between student and teacher networks. [21, 22, 23] not only consider teacher–student sample pairs but also examine the structure

of the relationships between features to further boost image recognition performance. Li and Yang [24] implemented knowledge distillation based on few-shot deep learning models. Overall, knowledge distillation methods are quite modern and offer significant benefits to the training process, helping to avoid information loss and enabling effective retraining.

3. **Proposed Method.**

3.1. **Problem Definition.** The task of recognizing student behaviors in classroom settings using deep learning faces three critical challenges in real-world scenarios: (1) Dynamic Behavior Emergence, where new behaviors (e.g., "using a smartphone," "distracted chatting") may emerge incrementally over time, requiring the model to continuously adapt while preserving knowledge of previously learned actions; (2) Few-Shot Constraints, where novel behaviors are often observed with extremely limited labeled instances (e.g., 1–5 samples per class), making traditional data-hungry deep learning approaches infeasible; and (3) Catastrophic Forgetting, where updating the model to accommodate new behaviors causes performance on older classes (e.g., "raising hand," "taking notes") to degrade rapidly due to the lack of access to historical training data during incremental sessions.

To formalize this problem, we first define the two distinct sets of classes within the Few-Shot Class-Incremental Learning (FSCIL) framework:

- Base Classes: These are the classes encountered during the initial training phase (session $\mathcal{T}$=0). They are characterized by having a sufficiently large number of labeled samples, which allows the model to learn a robust and stable feature representation for foundational behaviors (e.g., "raising hand," "taking notes"). The model's initial knowledge is built upon this comprehensive base dataset, D_0.
- Novel Classes: These are new classes of behaviors (e.g., "using a smartphone") that are introduced to the model sequentially in subsequent incremental sessions ($\mathcal{T} > 0$). Their defining characteristic is the "few-shot" constraint: they are learned from an extremely limited number of labeled instances (e.g., 1-5 samples per class). The primary challenge is to enable the model to learn these novel classes without degrading its performance on the previously learned base classes, a problem known as catastrophic forgetting.

The fundamental difference, therefore, lies in the amount of training data available and the sequential stage at which these class sets are introduced to the model.

Under the Few-Shot Class-Incremental Learning (FSCIL) framework, the problem is formalized as a sequence of training sessions $\{D_0, D_1, \ldots, D_\tau\}$, where each session τ introduces a disjoint set of behavior classes C_τ. The model must meet constraints such that, in the base session ($\tau = 0$), it trains on a sufficiently large dataset D_0 covering foundational behaviors (e.g., 10–20 classes), while in incremental sessions ($\tau > 0$), it learns N-way K-shot novel behaviors (e.g., $N = 5$ new actions with $K = 1$–3 samples each), with only the current session data D_τ and a small replay buffer D_m (storing exemplars from previous sessions) accessible during training at session τ, and the test set at session τ evaluates all classes encountered up to that point ($C_{\text{test}} = C_0 \cup C_1 \cup \cdots \cup C_\tau$), necessitating stability across sessions. Traditional behavior recognition models fail in this setting due to static architectures (inability to expand for new classes) and overfitting on few-shot data. The proposed solution integrates feature distillation to mitigate forgetting and space projection to maintain discriminative embeddings, ensuring both plasticity for new behaviors and stability for old ones.

3.2. **Model Architecture.** Figure 1 depicts an overview of our proposed Few-Shot Knowledge Distillation (FsKD) framework designed to address the challenges outlined in Section 3.1. The core idea is to maintain a balance between *plasticity* (the ability to

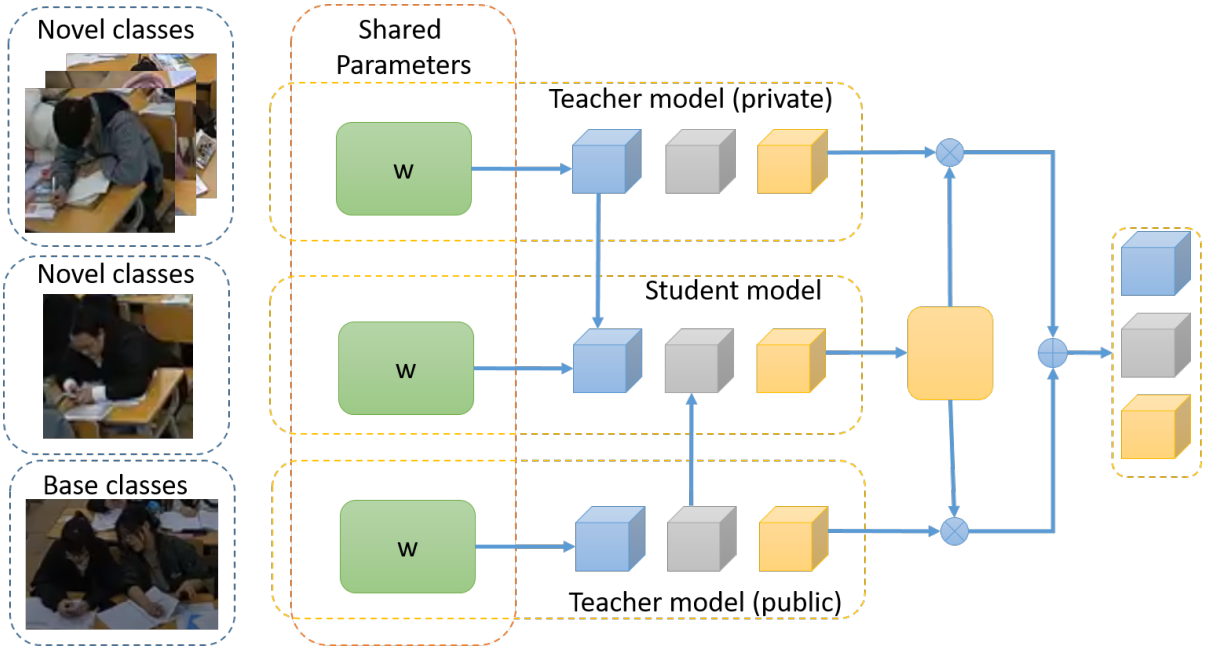

FIGURE 1. The framework of the proposed FsKD

learn newly emerging behaviors with limited data) and *stability* (the ability to preserve knowledge of previously learned behaviors). To achieve this, the architecture leverages three main components: (1) Teacher model (private branch); (2) Teacher model (public branch); and (3) student model, all connected through a shared parameter backbone.

- Shared Parameter Backbone (w). At the heart of the architecture is a shared feature extractor parameterized by w. This backbone encodes raw video frames or extracted features into a high-level representation space. By sharing parameters across teacher and student models, the framework ensures consistent feature representations and reduces the risk of feature drift when new classes are introduced incrementally.
- Teacher Model (Private Branch). The private teacher model is trained on the base session (i.e., the large initial dataset D_0 covering foundational behaviors). Its weights remain frozen or partially frozen in later sessions to preserve robust embeddings for previously learned classes. During incremental sessions, the private teacher provides high-quality class representations and acts as a reliable reference to mitigate catastrophic forgetting, especially for base or older classes that are no longer fully accessible.
- Teacher Model (Public Branch). In parallel with the private branch, a public teacher model is maintained to adapt to newly added classes in each incremental session. Unlike the private branch, this public teacher branch can be updated (under careful regularization) with the few-shot samples from D_τ and the small replay buffer D_m. By allowing limited plasticity here, the model can incorporate novel behaviors while minimizing the disturbance to previously learned representations.
- Student Model. The student model is the central learner that distills knowledge from both teacher branches. Through a *feature distillation* mechanism, the student aligns its intermediate representations with those of the teacher models, ensuring it retains essential characteristics of old classes (guided by the private teacher) while integrating new behaviors (guided by the public teacher). This dual-distillation strategy enables the student model to achieve high discriminability for both old and newly introduced classes.

Training Workflow. During the base session ($\tau = 0$), the private teacher model and the student model are jointly trained on the large labeled dataset D_0. The public teacher is either initialized identically to the private branch or starts with the same backbone w. In incremental sessions ($\tau > 0$), when N-way K-shot novel classes appear, the public teacher and student model update their parameters using the few-shot data D_τ and exemplars stored in D_m. The private teacher remains largely unchanged to preserve knowledge of older classes. A *projection head* may also be employed in both teacher and student models to project features into a space that remains discriminative across old and new classes.

3.3. **Loss function.** In the Few-Shot Class-Incremental Learning (FSCIL) setting, our goal is to learn newly emerging behaviors (i.e., novel classes) under severe data constraints while preserving previously acquired knowledge. To strike a balance between *plasticity* (the ability to adapt to new classes) and *stability* (the ability to retain performance on old classes), we design a composite loss function that combines three main components: a classification loss, a feature distillation loss, and a projection consistency loss.

First, we employ a classification loss $\mathcal{L}_{cls}$ to train the student model on the classes available during the current session τ. This includes both the few-shot samples from $\mathcal{D}_\tau$ and any exemplars stored in the replay buffer $\mathcal{D}_m$. We use a standard cross-entropy formulation to encourage the model to discriminate among all classes seen so far:

$$\mathcal{L}_{cls} = -\sum_{(x,y)\in(\mathcal{D}_\tau\cup\mathcal{D}_m)} \log p_{\text{student}}(y \mid x), \tag{1}$$

where $p_{\text{student}}(y \mid x)$ is the probability that the student model assigns to the ground-truth class y.

Next, to mitigate catastrophic forgetting, we introduce a feature distillation loss $\mathcal{L}_{dist}$ between the student model and two teacher branches (private and public). The private teacher, largely unchanged after the base session, preserves knowledge of older classes, while the public teacher is updated incrementally to accommodate newly added classes. By encouraging the student's feature representations to align with those of both teacher models, we help the student maintain performance on older classes while learning new ones:

$$\mathcal{L}_{dist} = \lambda_{\text{priv}} \sum_{x\in(\mathcal{D}_\tau\cup\mathcal{D}_m)} \left\| f_{\text{student}}(x) - f^{(\text{priv})}_{\text{teacher}}(x) \right\|^2 + \lambda_{\text{pub}} \sum_{x\in(\mathcal{D}_\tau\cup\mathcal{D}_m)} \left\| f_{\text{student}}(x) - f^{(\text{pub})}_{\text{teacher}}(x) \right\|^2, \tag{2}$$

where $f^{(\text{priv})}_{\text{teacher}}(x)$ and $f^{(\text{pub})}_{\text{teacher}}(x)$ denote the feature representations extracted by the private and public teacher models, respectively, and $\lambda_{\text{priv}}, \lambda_{\text{pub}}$ are hyperparameters controlling the balance between old-class preservation and new-class adaptation.

Finally, we employ a projection consistency loss $\mathcal{L}_{proj}$ to preserve a discriminative embedding space even when only a few samples of new behaviors are available. Through a projection head, features are mapped into a subspace designed to maximize inter-class separability. A margin-based loss or other metric learning objectives can be used here:

$$\mathcal{L}_{proj} = \sum_{(x,y)\in(\mathcal{D}_\tau\cup\mathcal{D}_m)} \text{MarginLoss}\big(\text{Proj}(f_{\text{student}}(x)),\, y\big), \tag{3}$$

where $\text{Proj}(\cdot)$ is the projection function, and MarginLoss is defined to push examples from different classes farther apart while pulling samples from the same class closer together.

By integrating these components, the total loss is expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{dist} + \mathcal{L}_{proj} \tag{4}$$

This combined objective ensures that the student model remains adaptable to new classes and robust against forgetting older ones throughout the incremental learning process.

4. Experiments.

4.1. Dataset.

5. **Dataset.** We conduct our experiments using two datasets: the widely adopted ImageNet and our custom-collected D-Edu dataset, which is specifically designed for recognizing student behaviors in classroom settings.

5.1. **ImageNet.** ImageNet [25] is a large-scale image dataset that has become a standard benchmark for various computer vision tasks. In our work, we leverage the standard training and validation splits of ImageNet to pretrain our backbone network. For the FSCIL benchmark experiments reported in Table 1, we utilized a common 100-class subset of ImageNet to facilitate a fair and direct comparison with existing state-of-the-art methods. This pretraining step ensures that our model benefits from rich, diverse visual representations, thereby facilitating a fair comparison with existing methods that also employ ImageNet pretraining.

5.2. **D-Edu.** The D-Edu dataset is a novel collection of classroom data captured from in-classroom cameras, consisting of approximately 200GB of video footage. The videos were segmented into 22,000 images, which were then organized into 600 samples—500 samples for training and 100 for testing. Each image has been resized to a resolution of 112×112 pixels.

D-Edu focuses on 22 distinct student action classes, with a balanced distribution of images across all classes. This ensures that each class is adequately represented during training. Examples of the action classes include:

- **Raising Hand** – indicating a student's intent to answer or ask a question.
- **Taking Notes** – students actively recording information during lectures.
- **Using Smartphone** – capturing instances where students engage with their mobile devices.
- **Chatting** – depicting interactions among students during class time.
- **Sleeping** – identifying moments when students appear inattentive or asleep.
- as well as other common behaviors such as 'Reading a book', 'Yawning', 'Looking at the board', and 'Drinking water', among others.

The D-Edu dataset provides a realistic and challenging benchmark for student behavior recognition. It captures a variety of classroom activities under diverse conditions.

5.3. **Experiment setup.** Our experiments are implemented in PyTorch on a single GPU server, where the training process for our classroom behavior recognition task is carried out with a mini-batch size of 64 and an initial learning rate of 0.01. We optimize the model using stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 0.0005, and a learning rate scheduler is employed to decay the learning rate by a factor of 0.1 at scheduled epochs to ensure smooth convergence. Given the few-shot constraints inherent in recognizing novel classroom behaviors, we incorporate standard data augmentation techniques, such as random cropping and horizontal flipping, to enhance model robustness and mitigate overfitting. Training is conducted in a session-based manner: the base session leverages a large dataset with ample samples per class, while

subsequent incremental sessions introduce novel classes with only a few samples each, complemented by a small replay buffer that stores exemplars from previous sessions. Our knowledge distillation framework further supports this setup by using a private teacher model to preserve old-class knowledge and a public teacher model to adapt to new classes, allowing the student model to distill comprehensive knowledge from both sources.

For the ImageNet benchmark, we followed the standard FSCIL protocol to ensure a fair comparison with prior works. Specifically, we utilized a subset of 100 classes from ImageNet. These classes were divided into a base session and multiple incremental sessions. The base session ($\mathcal{T}$=0) contained 60 randomly selected classes, which were used to train the initial model. The remaining 40 classes were then introduced sequentially over 8 incremental sessions ($\mathcal{T}$=1...8), with each session containing 5 new, non-overlapping classes (i.e., a 5-way setting). For these incremental sessions, we adopted a 5-shot evaluation protocol, where only 5 labeled images per novel class were available for training. The hyperparameters, including the learning rate, optimizer (SGD), and data augmentation techniques, remained consistent with those used for the D-Edu dataset to maintain methodological uniformity. This setup allows us to rigorously evaluate the model's ability to adapt to new classes while retaining knowledge of old ones on a large-scale, standard benchmark, as reported in Table 1.

For the D-Edu dataset, the 22 action classes were split for the Few-Shot Class-Incremental Learning (FSCIL) task as follows. The initial base training session ($\mathcal{T}$=0) used a set of 10 foundational behavior classes with a sufficient number of samples. The remaining 12 novel classes were then introduced over 4 subsequent incremental sessions ($\mathcal{T}$=1...4), with each session presenting 3 new behaviors (a 3-way setting) under few-shot constraints (K-shot). This setup simulates the real-world scenario where new, previously unseen student actions emerge over time.

We conducted two experiments corresponding to the following research questions (RQs):

- **RQ1:** How does the FsKD model compare with state-of-the-art models?
- **RQ2:** How does the FsKD model perform in real-world prediction for the task of student image recognition in classroom settings?

For fairness, we report performance metrics as published in the original papers wherever possible. For methods where specific metrics were not available for our benchmark setup (e.g., new class accuracy on ImageNet for FSCIL-ASP and FSCIL ALICE), we made our best effort to reproduce the results using their publicly available official source code under our standardized evaluation protocol.

5.4. **Performance Compare (RQ1).** The table 1 presents a comparative analysis of various Few-Shot Class-Incremental Learning (FSCIL) methods on two datasets: ImageNet and D-Edu. The performance of each method is evaluated based on accuracy for old classes and accuracy for new classes, which reflect the model's ability to retain previously learned knowledge while adapting to new classes.

Among all methods, FSKD (ours) achieves the highest performance with an accuracy of 0.803 for old classes and 0.638 for new classes, demonstrating superior capability in both knowledge retention and adaptation to new classes. The second-best method is FSCIL ALICE, which achieves a strong 0.790 for old classes and a competitive 0.6152 for new classes. FSCIL-ASP follows with 0.753 on old classes and 0.5815 on new classes. Other methods, such as FCIL and FSCIL-ASP, show relatively competitive performance with 0.7634 and 0.753 accuracy for old classes, respectively. However, FCIL's new class accuracy (0.5276) is significantly lower than FsKD (0.638), indicating weaker adaptability. Meanwhile, MetaFSCIL has the lowest performance in both categories, with an accuracy of 0.7204 for old classes and 0.4919 for new classes, making it the least effective among the

Table 1. Comparisons to state-of-the-art fscil methods on ImageNet, and D-Edu.

	ImageNet		D-Edu	
Method	**Acc. of old class**	**Acc. of new class**	**Acc. of old class**	**Acc. of new class**
FCIL	0.7634	0.5276	0.8163	0.6314
MetaFSCIL	0.7204	0.4919	0.7845	0.6031
FSCIL-ASP	0.753	0.5815	0.8064	0.6275
FSCIL ALICE	0.790	0.6152	0.8351	0.6526
FsKD (ours)	**0.803**	**0.638**	**0.8521**	**0.6835**

compared approaches. Overall, FsKD (ours) outperforms all other methods on ImageNet, particularly in handling new class adaptation while maintaining knowledge of old classes.

A similar trend is observed in the D-Edu dataset, where FsKD (ours) again achieves the best results, with 0.8521 accuracy for old classes and 0.6835 for new classes. This confirms its stability and effectiveness across different datasets. The second-best method is FSCIL ALICE, with 0.8351 old class accuracy and 0.6526 new class accuracy, though it still falls behind FsKD. FCIL follows closely with 0.8163 (old class accuracy) and 0.6314 (new class accuracy), making it a strong contender but still less effective than FsKD. FSCIL-ASP also performs relatively well, achieving 0.8064 for old classes and 0.6275 for new classes. Meanwhile, MetaFSCIL again shows the lowest performance, with 0.7845 (old class accuracy) and 0.6031 (new class accuracy), reinforcing its weaker ability to handle incremental learning tasks.

The results indicate that FsKD (ours) is the most effective FSCIL method, consistently outperforming others in both old class retention and new class adaptation across ImageNet and D-Edu. While FSCIL ALICE and FCIL also show strong performance, they do not match FsKD's overall effectiveness. MetaFSCIL ranks the lowest, demonstrating the weakest ability to balance knowledge retention and adaptation. This analysis highlights FsKD's superiority in FSCIL tasks, making it a promising approach for real-world applications that require continuous learning and adaptability in machine learning and image recognition tasks.

Table 2. Computational complexity analysis.

Method	**Backbone**	**GFLOPs**
FCIL	ResNet-18	1.81
MetaFSCIL	ResNet-18	2.15
FSCIL-ASP	ResNet-18	2.05
FSCIL ALICE	ResNet-18	2.43
FsKD (ours)	**ResNet-18**	**2.27**

To evaluate computational efficiency, we report the GigaFLOPs for a single forward pass in Table 2. While our student-teacher framework results in a moderate increase in complexity (2.27 GFLOPs) compared to the simplest baseline like FCIL (1.81 GFLOPs), it remains highly competitive and is more efficient than the top-performing FSCIL ALICE (2.43 GFLOPs). This analysis demonstrates that FsKD provides a state-of-the-art performance leap for a reasonable and justifiable increase in computational cost, highlighting a favorable performance-to-efficiency trade-off.

5.5. **Qualitative Study (RQ2).** Figure 2 presents a PCA-based 2D visualization of six distinct student action classes, using feature embeddings extracted from the FsKD model.

FIGURE 2. The t-SNE visualization of ablation study on D-Edu.

Each class is represented by a color-coded scatter plot, with dashed confidence ellipses outlining the spatial distribution of each class cluster. The clusters are well-separated, suggesting that the FsKD model successfully extracts discriminative features for different student actions. The manual positioning adjustments ensure minimal overlap, closely resembling a structured layout.

The confidence ellipses highlight the distribution range of each class, with their size and orientation reflecting the feature variance within each category. Narrow ellipses, such as those for Class 1 and Class 7, indicate high consistency in extracted features, while wider ellipses, like those for Class 3 and Class 9, suggest greater variation in action representations. The minimal overlap between ellipses implies that the model can effectively distinguish different actions, reducing misclassification risks. However, the proximity of some clusters, such as Class 6 and Class 7, suggests a possible latent similarity in their representations.

The structured separation of clusters indicates that the FsKD model maintains a robust feature learning mechanism. In real-world applications, such as incremental learning or educational AI systems, this level of separability could lead to high classification accuracy while mitigating catastrophic forgetting. Overall, the visualization provides a scientifically structured representation of the feature space for student actions, showcasing the FsKD model's effectiveness in distinguishing incremental classes.

5.6. **Deployment and Real-Time Considerations.** While our primary focus has been on the learning methodology, analyzing the deployment of FsKD in real-time application scenarios is critical for assessing its practical utility. We consider two distinct operational phases: the inference phase and the model update phase.

Real-Time Inference Phase: During live operation, such as analyzing video feeds from classroom cameras, only the trained student model is active for making predictions. The teacher models (both private and public) are exclusively used during the knowledge distillation training process and are not required for inference. Consequently, the inference speed is determined by the architecture of the student model alone. With a ResNet-18 backbone, the model has a computational cost of approximately 1.81 GFLOPs, which is highly efficient and allows for high-throughput frame processing on standard GPU hardware, thus meeting the requirements for real-time behavior recognition.

Incremental Model Update Phase: The process of learning new behavior classes is not instantaneous and is treated as an offline or periodic task. The workflow in a real-world setting would be as follows: (1) A new behavior (e.g., "using a tablet") is observed and a few representative image samples are collected and labeled. An incremental training session is triggered, where the public teacher and student models are updated using these new few-shot samples along with the exemplar set from the replay buffer, as described in our methodology. (3) While this update is processing, the existing student model continues to operate without interruption. (4) Once the new model is trained and validated, it can be deployed to replace the previous version, a process often managed via hot-swapping to ensure continuous service.

Limitations and Practical Challenges: This deployment model presents some practical considerations. First, there is an inherent latency between when a new behavior appears and when the model can recognize it, as it depends on human-in-the-loop for sample labeling. Second, while the incremental update is far more efficient than retraining from scratch, the knowledge distillation process still requires non-trivial computational resources and time. Future work could explore semi-supervised or unsupervised techniques to reduce the labeling dependency and further optimize the update cycle for near-real-time adaptation.

6. **Limitations and Future Work.** While our proposed FsKD framework demonstrates state-of-the-art performance, it is essential to acknowledge its potential limitations, which also highlight promising directions for future research.

Computational Complexity: The dual-teacher architecture, a core component of FsKD, involves three model branches (one student, two teachers) during the training and knowledge distillation process. This inherently leads to higher computational overhead and longer training times compared to single-model FSCIL baselines. While we argue this trade-off is justified by the significant performance gains, future work could explore techniques like asynchronous updates or more efficient knowledge distillation mechanisms to reduce the training cost.

Dependency on Base Session Quality: The stability of our model relies heavily on the "private teacher," which acts as a knowledge anchor for previously learned classes. The effectiveness of this teacher is directly tied to the quality and diversity of the initial base training session (D_0). If the base dataset is small, imbalanced, or not representative of the domain, the private teacher's guidance will be weak, potentially compromising the model's ability to preserve old knowledge. Future research should investigate methods to enhance robustness against suboptimal base training conditions.

Sensitivity to Exemplar Buffer: FsKD utilizes a replay buffer (D_m) containing exemplars from past classes to mitigate catastrophic forgetting. This study did not include an

ablation analysis on the sensitivity of the model to the size of this buffer. In memory-constrained applications, where only a very small number of exemplars can be stored, the model's performance may degrade. Therefore, a critical direction for future work, as noted in our conclusion, is to develop more sophisticated and memory-efficient sample selection strategies to ensure that the most informative exemplars are retained.

7. **Conclusions.** This study demonstrates the FsKD model's effectiveness in addressing the challenges of few-shot class-incremental learning (FSCIL) by ensuring robust feature separability while mitigating catastrophic forgetting. Through a PCA-based 2D visualization, we highlight how the model effectively learns and distinguishes incremental student action classes by maintaining high intra-class consistency and inter-class separation. The structured cluster distribution and confidence ellipses indicate that the FsKD model captures discriminative features efficiently, making it a promising solution for incremental learning scenarios. Our approach focuses on enhancing feature representation learning, ensuring that each new class retains its unique identity without interfering with previously learned knowledge. The structured separation of clusters in the visualization suggests that the FsKD model is particularly well-suited for applications in educational AI, behavior analysis, and adaptive learning systems. Extensive analysis of class distribution patterns further confirms that the FsKD model preserves knowledge retention while effectively adapting to new categories, reinforcing its broad applicability in real-world incremental learning environments. Moving forward, we aim to explore additional techniques to further refine feature embedding consistency and optimize sample selection strategies, thereby enhancing the overall performance of FSCIL systems.

REFERENCES

[1] Mahadevkar, Supriya V., et al. "A review on machine learning styles in computer vision—techniques and future directions." Ieee Access 10 (2022): 107293-107329.
[2] Ting-Ting Wu and Tien-Wen Sung, "Analysis of the Effects of a Game-Based Review System Integrated with the Hierarchy of Learning on Learning Outcomes in an Elementary Social Science Course," Interactive Learning Environments, Vol. 31, No. 6, August, 2023.
[3] Liang, Weixin, et al. "Advances, challenges and opportunities in creating data for trustworthy AI." Nature Machine Intelligence 4.8 (2022): 669-67
[4] Zhou, Da-Wei, et al. "Forward compatible few-shot class-incremental learning." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
[5] Mehta, Naval Kishore, et al. "Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement." Applied Intelligence 52.12 (2022): 13803-13823.
[6] Mahalakshmi, V., et al. "Few-shot learning-based human behavior recognition model." Computers in Human Behavior 151 (2024): 108038.
[7] Gharoun, Hassan, et al. "Meta-learning approaches for few-shot learning: A survey of recent advances." ACM Computing Surveys 56.12 (2024): 1-41.
[8] Wanyan, Yuyang, et al. "Active exploration of multimodal complementarity for few-shot action recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[9] Meng, Hao, et al. "Cross-datasets facial expression recognition via distance metric learning and teacher-student model." Multimedia Tools and Applications 81.4 (2022): 5621-5643.
[10] Peng, Yishu, et al. "Convolutional transformer-based few-shot learning for cross-domain hyperspectral image classification." IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 16 (2023): 1335-1349.
[11] Han, Yue, et al. "Reference twice: A simple and unified baseline for few-shot instance segmentation." IEEE transactions on pattern analysis and machine intelligence (2024).
[12] Li, Chuanlong, et al. "Zero shot objects classification method of side scan sonar image based on synthesis of pseudo samples." Applied Acoustics 173 (2021): 107691.
[13] Deng, Shizhuo, et al. "Self-relation attention networks for weakly supervised few-shot activity recognition." Knowledge-Based Systems 276 (2023): 110720.

[14] Xiao, Ni, and Lei Zhang. "Dynamic weighted learning for unsupervised domain adaptation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
[15] Zhang, Jinghua, et al. "Few-Shot Class-Incremental Learning for Classification and Object Detection: A Survey." IEEE Transactions on Pattern Analysis and Machine Intelligence (2025).
[16] Zhao, Borui, et al. "Decoupled knowledge distillation." Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2022.
[17] Gou, Jianping, et al. "Knowledge distillation: A survey." International Journal of Computer Vision 129.6 (2021): 1789-1819.
[18] Lin, Han, et al. "Supervised masked knowledge distillation for few-shot transformers." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.
[19] Tang, Yihe, et al. "Humble teachers teach better students for semi-supervised object detection." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
[20] Borza, Diana Laura, et al. "Teacher or supervisor? effective online knowledge distillation via guided collaborative learning." Computer Vision and Image Understanding 228 (2023): 103632.
[21] Zhu, Jinguo, et al. "Complementary relation contrastive distillation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.
[22] Cao, Shengcao, et al. "Contrastive mean teacher for domain adaptive object detectors." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023.
[23] Li, Gang, et al. "Knowledge distillation for object detection via rank mimicking and prediction-guided feature imitation." Proceedings of the AAAI conference on artificial intelligence. Vol. 36. No. 2. 2022.
[24] Li, Yang, Yicheng Gong, and Zhuo Zhang. "Few-shot object detection based on self-knowledge distillation." IEEE Intelligent Systems (2022).
[25] Wang, Rui-Qi, Xu-Yao Zhang, and Cheng-Lin Liu. "Meta-prototypical learning for domain-agnostic few-shot recognition." IEEE Transactions on neural networks and learning systems 33.11 (2021): 6990-6996.