# WP-FSCIL: A Well-Prepared Few-shot Class-incremental Learning Framework for Pill Recognition

Jinghua Zhang, Chen Li, Marco Cristani, Hongzan Sun, Marcin Grzegorzek, and Huiling Chen

***Abstract*— Few-shot Class-incremental Pill Recognition (FSCIPR) aims to develop an automatic pill recognition system that requires only a few training data and can continuously adapt to new classes, providing technical support for applications in hospitals, portable apps, and assistance for visually impaired individuals. This task faces three core challenges: overfitting, fine-grained classification problems, and catastrophic forgetting. We propose the Well-Prepared Few-shot Class-incremental Learning (WP-FSCIL) framework, which addresses overfitting through a parameter-freezing strategy, enhances the robustness and discriminative power of backbone features with Center-Triplet (CT) loss and supervised contrastive loss for fine-grained classification, and alleviates catastrophic forgetting using a multi-dimensional Knowledge Distillation (KD) strategy based on flexible Pseudo-feature Synthesis (PFS). By flexibly synthesizing any number of old-class features, the PFS strategy resolves the issue of insufficient samples in the KD process, enabling Response-based KD (KD1) and Relation-based KD (KD2) to comprehensively preserve old knowledge. The effectiveness of WP-FSCIL has been validated through experiments conducted on two publicly available pill datasets. These experiments show that WP-FSCIL outperforms existing state-of-the-art methods, demonstrating its superior performance.**

***Index Terms*— Pill recognition, Class-incremental learning, Few-shot learning, Metric learning, Knowledge distillation**

## I. Introduction

The “Medication Without Har” initiative by the World Health Organization emphasizes that unsafe and incorrect medication practices are a major source of preventable harm in healthcare systems worldwide, resulting in an estimated annual cost of $42 billion [1]. Medication errors can occur at any stage, such as prescribing or dispensing, and are often influenced by factors like fatigue, which can lead to severe harm or even death. To address these challenges, the World Health Organization’s initiative aims to reduce medication errors and enhance patient safety. With the progresses made by machine learning in many fields [2]–[11], automatic pill recognition technology has emerged as a promising solution [12]–[16], which primarily involves using computer vision and machine learning techniques to accurately identify pills based on their visual characteristics, thereby reducing medication errors and preventing potential adverse pill events.

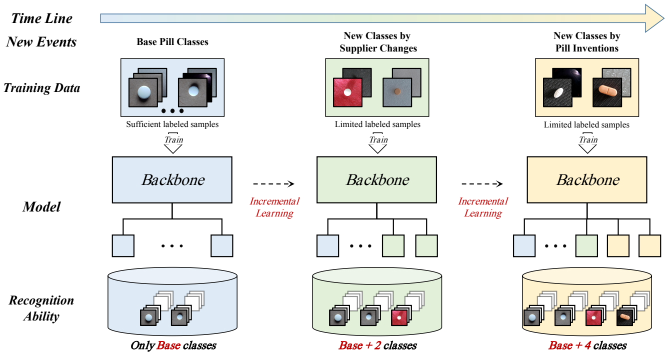


Fig. 1: An example of FSCIPR under the hospital scenario. In the hospital, as time progresses, the range of pill classes requiring identification may expand due to changes in suppliers or the invention of new pills. Whenever new pills are introduced, automatic pill recognition systems should effectively learn to identify these new classes using limited samples while retaining the ability to recognize previously known pills.

The automatic pill recognition system holds significant and wide-ranging value across healthcare, home medication, and pharmaceutical management [17]–[19]. In healthcare institutions, this system provides pharmacists and nursing staff with an additional verification layer to ensure the accuracy of medication type and dosage, significantly reducing medication errors and enhancing patient safety. This is particularly crucial in scenarios involving multi-pill dispensing in hospital wards. For chronic patients, the elderly, and those with visual impairments, pill recognition technology offers reliable medication

This work was supported in part by the National Natural Science Foundation of China under Grant 82220108007 and Grant 62306323, and in part by the Outstanding Youth Project from the Hunan Provincial Department of Education under Grant 24B0865.

Jinghua Zhang (zhangjingh@foxmail.com) is with the College of Intelligence Science and Technology, National University of Defense Technology, Changsha, China. Chen Li (lichen@bmie.neu.edu.cn) is with the College of Medicine and Biological Information Engineering, Northeastern University, Shenyang, China. Marco Cristani (marco.cristani@univr.it) is with the Department of Computer Science, University of Verona, Verona, Italy. Hongzan Sun (sunhz@sj-hospital.org) is with Shengjing Hospital, China Medical University, Shenyang, China. Marcin Grzegorzek (marcin.grzegorzek@uni-luebeck.de) is with the Institute of Medical Informatics, University of Luebeck, Luebeck, Germany. Huiling Chen (chenhuiling.jlu@gmail.com) is with the College of Computer Science and Artificial Intelligence, Wenzhou University, Wenzhou, China.

Huiling Chen and Chen Li are the corresponding authors.

verification in home settings, helping to prevent confusion or repeated medication, thereby supporting safe adherence to medical instructions. Automatic pill recognition facilitates automated inventory control within pharmaceutical companies, warehouses, and pharmacies in pharmaceutical management. This ensures the accuracy of pill types, quantities, and batch information, stabilizing the pill supply chain. Furthermore, recent advancements in mobile applications incorporate pill recognition features that leverage camera-based image recognition to provide real-time medication information, offering convenient, self-service medication support for users in their daily lives. These application scenarios underscore the profound significance of automatic pill recognition systems in reducing medication errors, enhancing patient safety, and improving the efficiency of healthcare system management.

Currently, research in the automatic pill recognition field is relatively limited, with most focusing on traditional machine learning and computer vision methods. For example, *Kim et al.* [15] proposed a Multi Combination Pattern Labeling method that improves pill classification accuracy and reliability through feature point extraction and edge recognition. *Wang et al.* [14] introduced the MCIR-YOLO system, which enhances the YOLOv5s model through multimodal fusion techniques to address the challenge of distinguishing white pills. *Heo et al.* [16] developed a deep learning-based automatic pill recognition system that combines image classification, text detection, and a language model for label correction, significantly improving recognition accuracy.

Despite the growing body of research, most of these systems are static, making it difficult to adapt to new pills and constantly changing classes. Additionally, minimizing the reliance on labeled samples is essential, given the high cost of data annotation and the requirement for specialized expertise. To address these issues, *Ling et al.* [20] developed the few-shot pill recognition method, and *Nguyen et al.* [21] were the first to consider a continual learning scenario for pill recognition. However, in practice, the addition of new pills is often accompanied by a lack of labeled samples. For example, in hospital scenarios, as illustrated in Fig. 1, the range of pill classes requiring identification may expand over time due to changes in suppliers or the invention of new pills. Whenever new pills are introduced, automatic pill recognition systems must effectively learn to identify these new classes using few-shot samples while retaining the ability to recognize previously known pills. Similarly, on mobile devices, personal users may need to customize new pills, creating a demand for automatic pill recognition systems that handle both few-shot and class-incremental challenges. These scenarios underscore the urgent need for automatic pill recognition systems that can address these dual challenges effectively, though research in this area remains in its early stages.

To address the Few-shot Class-incremental Learning (FSCIL) challenges in pill recognition, we propose a novel framework, Well-Prepared Few-shot Class-incremental Learning (WP-FSCIL), which tackles key issues in Few-shot Class-incremental Pill Recognition (FSCIPR). These challenges include overfitting, where limited training samples during incremental learning lead to poor generalization; catastrophic forgetting, where new class introduction can result in the loss of knowledge for old classes; and the fine-grained challenge, where similarities in pill color and shape, as well as varying angles and lighting, increase both inter- and intra-class confusion, complicating model performance.

To overcome these challenges, we introduce WP-FSCIL, a framework designed with one training strategy and two core designs. Following the training strategy used in most FSCIL methods, we freeze most of the backbone parameters after the base session training to mitigate the overfitting problem. This first core design: in the base session training, we leverage Center-Triplet (CT) loss to enhance the model's ability to distinguish fine-grained features effectively. To enhance the backbone's ability to generalize to future incremental classes, we incorporate supervised contrastive loss during the base session. Additionally, we design a set of pseudo-classes based on pill visual characteristics to simulate potential future scenarios, thereby augmenting the model's forward compatibility. This second core design: during the subsequent incremental sessions, the focus shifts to retaining previously acquired knowledge while learning new classes, thus mitigating catastrophic forgetting. To this end, we propose a novel multi-dimensional Knowledge Distillation (KD) strategy grounded in flexible Pseudo-feature Synthesis (PFS). This strategy enables the flexible synthesis of an arbitrary number of pseudo-old class features, combined with real incremental data features, to facilitate the KD process. By generating sufficient pseudo-features, this approach addresses the limitations of few-shot scenarios. During distillation, we employ Response-based KD (KD1) and Relation-based KD (KD2) [22] to comprehensively retain knowledge of previously learned classes. This helps the model handle new and old classes effectively while preserving strong recognition performance.

To evaluate the effectiveness of WP-FSCIL, we performed experiments on the FCPill and *m*CURE datasets. The results show that our framework surpasses existing advanced methods, achieving outstanding performance across various evaluation metrics. To highlight the contributions of our approach, we summarize the key points below:

- **Proposed a well prepared FSCIPR framework:** We designed the WP-FSCIL framework for pill recognition, systematically addressing overfitting, fine-grained issues, and catastrophic forgetting with targeted strategies in both base and incremental sessions.
- **Introduced a comprehensive base session training strategy:** In the base session, we introduced CT loss for fine-grained discriminability, supervised contrastive loss for generalization, and combined virtual and real classes to enhance forward compatibility for future incremental classes.
- **Proposed a multi-dimensional knowledge distillation strategy for incremental learning:** In incremental sessions, we developed a multi-dimensional KD strategy with PFS to transfer classification and feature relationships, preserving previous knowledge.
- **Superior experimental performance:** Experiments on two public datasets show that WP-FSCIL outperforms existing methods, demonstrating its effectiveness in ad-

dressing FSCIPR challenges.

The structure of this paper is as follows: Section II covers the foundational knowledge and related studies; Section III provides an in-depth description of our proposed methods; Section IV presents the experimental setup and performance analysis; and Section V offers the conclusion.

## II. Related Work

### A. Automatic Pill Recognition

Currently, research on pill recognition predominantly focuses on two tasks: pill classification [15], [16], [20], [21], [23], [24] and pill detection [14], [18], [19], [25], [26]. Our paper focuses on classification tasks. Existing approaches can be categorized into traditional image processing techniques and deep learning methods. Traditional techniques often rely on expert-designed feature extraction frameworks rooted in domain knowledge, subsequently paired with machine learning algorithms for pill classification. For example, the early pill recognition system proposed by *Lee et al.* [23] relied heavily on feature engineering, showcasing the importance of attributes such as shape, color, size, and imprint in achieving effective pill recognition. With the rise of deep learning, models can independently learn complex features from images, removing the need for manual feature extraction. Many recent studies have used deep learning for pill recognition. For instance, *Zeng et al.* [24] proposed the MobileDeepPill system for automatic pill recognition via smartphones in complex environments. This system improves recognition robustness, captures pill features, and reduces model size through a triplet loss function, a multi-CNN model, and a KD framework.

Deep learning-based methods are more flexible and perform better than traditional methods but require large amounts of data and struggle to adapt to changing environments. To address these challenges, *Ling et al.* [20] developed a few-shot classification framework, and *Nguyen et al.* [21] proposed an incremental multi-stream fusion framework. However, incremental scenarios in pill recognition often come with limited samples, introducing the FSCIL challenge, which remains in its early research stages, with only our early work [13] addressing this area.

### B. Few-shot Learning

Few-shot learning has emerged as a crucial area of research to overcome the limitations posed by insufficient training samples, as conventional deep learning models typically rely on extensive labeled datasets to achieve robust generalization [27]; however, acquiring sufficient labeled samples is time-consuming and costly in many real-world applications, such as pill recognition. Few-shot learning aims to effectively learn new classes from limited training samples, achieving satisfactory classification performance [28]–[31]. Current few-shot learning methods primarily include meta-learning, metric learning, and data augmentation approaches. Meta-learning approaches focus on helping models adapt rapidly to new tasks. For example, Prototype Bayesian Meta-Learning [32] constructs task-specific initialization in a Bayesian framework, combining variational inference and prototype-conditioned priors to enable rapid adaptation and improved generalization on new tasks. Metric learning methods address few-shot classification by defining similarity metrics in the feature space. For instance, *Zhou et al.* [33] proposed an automatic metric search to reduce the need for manual effort and domain knowledge in designing metrics. Data augmentation techniques enrich training datasets by generating additional samples. Recently, *Meng et al.* [34] fine-tuned an autoregressive PLM on limited samples, using it as a generator to create novel training examples, effectively expanding the dataset and addressing few-shot learning challenges.

Although existing few-shot learning methods have established a systematic theoretical framework and made progress in pill recognition, they primarily focus on improving the performance of few-shot classes while neglecting the preservation of initial class performance and adaptability to dynamically added new classes, leading to a disconnect with real-world scenarios. In contrast, FSCIPR is more aligned with practical applications, effectively addressing the dual challenges of limited samples and dynamic data streams.

### C. Class-incremental learning

Class-incremental learning has become an important research direction for scenarios that require dynamically expanding classes. In traditional deep learning, models are usually trained once on a static dataset; however, in pill recognition tasks, new pill classes continuously emerge with the development and release of new medications, requiring models to learn new classes to adapt to evolving data continuously. Class-incremental learning aims to incorporate new classes while minimizing catastrophic forgetting of previously learned knowledge [35]–[37]. Representative class-incremental learning methods include Data Replay (DR), Dynamic Networks, and KD approaches. DR methods save or generate data from old classes, enabling joint training with new class data to prevent forgetting previous knowledge. For example, *Jodelet et al.* [38] introduced SDDR, a method that uses a diffusion model to generate data for previously learned classes, aiding in preserving past knowledge. Dynamic network approaches adapt the model's representational capacity to changing data streams. For instance, *Hu et al.* [39] proposed DNE, a method that better balances accuracy and model complexity. KD methods preserve old-class knowledge through distillation. For example, *Wen et al.* [40] introduced MTD, which identifies multiple diverse teacher models for effective knowledge retention.

While class-incremental learning research has made substantial progress, most methods assume ample training data in incremental scenarios, disregarding the reality of data scarcity and the high costs of annotation. In dynamic pill recognition scenarios, new pill classes often emerge unexpectedly, with limited sample availability. Compared to addressing class-incremental problems alone, FSCIPR techniques are more attuned to real-world demands, as they adeptly handle the twin challenges of data scarcity and dynamic class evolution.

### D. Few-shot Class-incremental Learning

FSCIL is a specialized subset of class-incremental learning that addresses scenarios with limited sample availability [41]. Its goal is to allow models to incrementally learn new classes from limited labeled samples while preserving previous knowledge of old classes [41]–[44]. The core challenges in FSCIL lie in addressing overfitting and catastrophic forgetting, especially as the scarcity of training data in incremental learning scenarios exacerbates these issues.

In dynamic domains such as pill recognition, FSCIL addresses the necessity for models to adapt flexibly to new pill classes without undergoing extensive retraining on continuously expanding datasets. In tasks such as pill recognition, where new pill classes are regularly introduced, FSCIL effectively enables models to incorporate these new classes. Current FSCIL methods mainly fall into the "Feature Extractor + Softmax Classifier" framework and the "Feature Embedding + Nearest Mean Classifier" framework. The "Feature Extractor + Softmax Classifier" approach typically trains the entire network during incremental learning, often incorporating mechanisms such as DR or KD to mitigate forgetting. For instance, the CEC algorithm [45] employs a Graph Attention Network (GAT) to optimize the relations between base and new class prototypes. This refinement allows the classifier to establish better decision boundaries in complex category environments. In contrast, the "Feature Embedding + Nearest Mean Classifier" approach focuses on learning class prototypes in a feature space, mapping samples into an embedding space where semantic differences are more apparent, and classifying based on the nearest mean. The FACT framework [46] introduces the concept of forward compatibility, enhancing the model's capacity to adapt to future new classes.

Although FSCIL methods have achieved notable progress, there is still room for improvement. In FSCIPR, the challenges extend beyond overfitting and catastrophic forgetting, as fine-grained classification also need to be addressed. To tackle these challenges, we propose a novel framework, WP-FSCIL, which employs the following strategies: mitigating overfitting through a parameter-freezing strategy; enhancing the model's robustness and discriminative capability in fine-grained classification tasks by incorporating CT loss and supervised contrastive loss; and alleviating catastrophic forgetting with a multi-dimensional KD strategy based on flexible PFS.

## III. Method

This section initially outlines the problem setting for FSCIL, followed by a discussion of the principal challenges associated with FSCIPR. Subsequently, we present an overview of the proposed framework WP-FSCIL and offer a detailed exposition of its components.

### A. Problem Setting

FSCIL typically consists of a base session and multiple incremental sessions. The base session aims to provide sufficient data during the initial training to ensure a good initialization for the model, while the incremental sessions focus on learning new classes from limited training samples without forgetting previously acquired knowledge. In FSCIL, the training and testing datasets can be mathematically represented as $\{D^0_{train}, \ldots, D^n_{train}\}$ and $\{D^0_{test}, \ldots, D^n_{test}\}$, respectively, with $n$ indicating the total number of incremental sessions in the task. $D^0_{train}$ represents the base session training dataset, containing a substantial amount of labeled data. For each integer $i$ ranging from 1 to $n$, $D^i_{train}$ adopts an $N$-way $K$-shot format, indicating that in session $i$, the training dataset comprises $N$ classes, each containing $K$ labeled samples. $D^i_{test}$ represents the testing dataset for session $i$. For any integers $i$ and $j$ between 0 and $n$ where $i \neq j$, the corresponding label spaces of $D^i_{train}$ and $D^j_{train}$, denoted as $C^i$ and $C^j$, are disjoint, *i.e.*, $C^i \cap C^j = \emptyset$. When the learning process reaches session $i$, only $D^i_{train}$ is accessible, while the complete and original training datasets from previous sessions are unavailable (some methods may store some samples for reply). The evaluation for session $i$ is performed on a combined set of testing datasets from session 0 to session $i$, represented as $D^0_{test} \cup \cdots \cup D^i_{test}$.

### B. Challenge Analysis

In this section, we summarize the main challenges faced in FSCIPR, including overfitting due to limited training data, catastrophic forgetting in the incremental learning process, and fine-grained classification difficulty among similar pill classes.

*1) Overfitting:* In FSCIPR, limited training data often leads the model to focus on minimizing prediction errors within the training set. This approach is prone to significant discrepancies between empirical and expected risks, particularly when the training dataset adopts an $N-$way $K-$shot format. As a result, the model tends to overfit, excelling on training data but underperforming on test data. Furthermore, as new classes are added incrementally, continual reliance on this unreliable empirical risk minimization can hinder the model from reaching an optimal state, challenging its stability and reliability in the following sessions.

*2) Catastrophic Forgetting:* FSCIPR requires a balance between maintaining stability for previous knowledge and exhibiting plasticity for new classes. When new classes are introduced, unrestricted adjustment of existing model parameters can shift decision boundaries towards these new classes, leading to catastrophic forgetting. Conversely, placing too much emphasis on retaining old knowledge can hinder the model's ability to acquire new classes effectively.

*3) Fine-grained Challenges:* In FSCIPR, the subtle visual differences between many pill classes demand a model with robust discriminative capabilities. Introducing new classes that closely resemble existing ones increases classification difficulty and may lead to confusion between previously learned and newly added classes. Additionally, the scarcity of data for new classes, combined with their high similarity to known classes, further intensifies the challenge for the model to differentiate among them accurately. This problem can be found in the pill examples in Fig. 1 and Fig. 4.

### C. Overall Framework

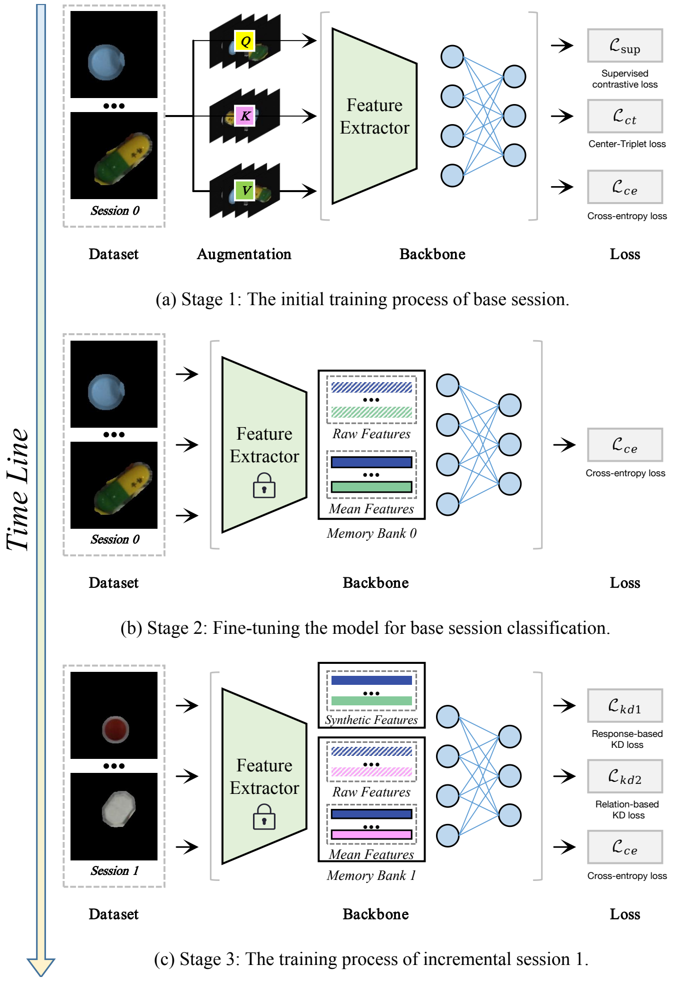
To comprehensively address the challenges in FSCIPR, we propose a framework for pill recognition, termed WP-FSCIL. Deep neural networks typically require substantial training data; however, in FSCIPR, only limited samples are available for incremental sessions. Training the entire network with such sparse data directly leads to significant overfitting issues. Furthermore, FSCIPR involves multiple incremental sessions, demanding that the model preserve prior knowledge while integrating newly introduced classes. Directly training the entire network with new data could cause the decision boundaries to shift towards the new classes, leading to catastrophic forgetting. Thus, following most other FSCIL methods, we freeze most of the backbone parameters after the base session training to mitigate both overfitting and forgetting.

The WP-FSCIL training process consists of two main phases, as illustrated in Fig. 2: the base session training phase and the incremental session training phase. The base session training phase consists of two stages: Stage 1 represents the foundational training, designed to ensure the model attains a robust initialization from sufficient base data; Stage 2 involves fine-tuning the model to enhance its adaptability for classification tasks within the base session; Stage 3 is the incremental session training phase, focusing on incremental learning which updates specific parameters to accommodate new classes.

In Stage 1, our primary objective is to endow the backbone with strong generalization capabilities to adapt to future incremental classes and enhance its discriminative ability for fine-grained pill classes. To achieve this, we employ a forward-compatible strategy and supervised contrastive learning. The forward-compatible strategy constructs pseudo-classes based on existing classes to simulate potential future incremental classes. It lets the model get the power to adapt to future incremental classes during base session training. Additionally, to further improve generalization, we introduce a modified supervised contrastive loss based on self-supervised contrastive loss, enhanced by class-supervision information. This approach enables the model to capture inter-class differences, thereby supporting effective feature extraction for future incremental learning. To strengthen the model's ability to distinguish fine-grained classes, the CT loss is introduced, merging the advantages of triplet and center losses to encourage greater inter-class separation and tighter intra-class clustering, thereby enhancing the backbone's discriminative capabilities.

In Stage 2, the model undergoes further fine-tuning with cross-entropy loss to perform the classification in the base session. During this stage, a few original features extracted by the backbone are stored along with the class prototypes (mean feature vectors for each class) in a memory bank to support knowledge retention in future sessions. This selective feature storage strategy significantly reduces storage requirements compared to conventional raw DR methods.

In Stage 3, the model learns new classes while preserving old knowledge. To overcome catastrophic forgetting, Stage 3 introduces the multi-dimensional KD strategy, which combines KD1 to preserve classification abilities and KD2 to maintain relationships among samples, thereby facilitating the effective retention of previous knowledge. However, conventional KD typically requires sufficient samples, which is incompatible with the few-shot setting in FSCIPR. Therefore, we synthesize pseudo features using features and class prototypes stored in the memory bank to aid in KD. The memory bank is also updated continuously across the incremental sessions.



(a) Stage 1: The initial training process of base session.

(b) Stage 2: Fine-tuning the model for base session classification.

(c) Stage 3: The training process of incremental session 1.

Fig. 2: We propose the WP-FSCIL framework for FSCIPR: In the first stage, virtual classes are synthesized to enable the model to adapt to future classes in advance. Additionally, CT loss is introduced to enhance fine-grained discriminative ability, while supervised contrastive loss improves the model's generalization capability. The second stage focuses on fine-tuning the model, optimizing classification performance in the base session using cross-entropy loss, and storing relevant features from the base session. The third stage employs a multi-dimensional knowledge distillation strategy based on flexible PFS (combining response knowledge distillation loss and relational knowledge distillation loss) to acquire new knowledge while effectively preventing catastrophic forgetting of old classes.

### D. Base Session Learning

The base session training aims to establish a strong backbone with robust generalization and fine-grained discrimination by introducing forward-compatible learning, supervised contrastive loss, and CT loss.

*1) Forward-compatible Learning:* Since FSCIPR involves multiple incremental sessions, enough training samples are available during the base session, but only limited supervised information is available in the incremental sessions. Therefore, building a strong generalization capability for the backbone during the base session training is critical for our method's performance in subsequent incremental sessions. One of the most straightforward strategies is to train the model during the base session in a way that anticipates potential future scenarios, thus constructing a backbone that is forward-compatible with future incremental classes.

We observed that pills differ from other natural images, as their shapes and appearances are typically fixed during manufacturing. Leveraging this characteristic, we construct virtual pill classes to simulate potential future incremental classes. Specifically, we use existing pill class information to create a series of pseudo-classes through transformations in size and color. This process is illustrated in Fig. 3. Using data augmentation techniques (e.g., color transformation, cropping, rotation, and flipping), we generated virtual classes and merged them with the original classes to form a new augmented dataset. The generation process is defined as $(x_{rv},\ y_{rv}) = \mathcal{F}(x,\ y)$, where $(x,\ y)$ corresponds to a sample and its label from the original base training data, $\mathcal{F}$ is the applied transformation function, and $(x_{rv},\ y_{rv})$ denotes a transformed sample-label pair included in the augmented dataset comprising both real and virtual data. These virtual classes are combined with existing base classes for backbone initialization. This approach offers two main advantages: first, introducing virtual classes enriches the semantic information available during base session training; second, virtual classes act as placeholders in the feature space, paving the way for the integration of future incremental classes. Following the integration of virtual classes, the classification loss function for the backbone can be articulated as follows:

$$\mathcal{L}_{cls}\left(\phi; x_{rv}, y_{rv}\right) = \mathcal{L}_{ce}\left(\phi\left(x_{rv}\right), y_{rv}\right), \tag{1}$$

where $\phi$ represents the trained model, and $\mathcal{L}_{ce}$ denotes the cross-entropy loss.

*2) Supervised MoCo-based Contrast Learning:* In addition to equipping the backbone with forward compatibility to address potential future scenarios, we are also dedicated to enhancing the model's generalization capability and ability to discriminate fine-grained pill classes. In recent years, many foundational model techniques in computer vision have gained attention due to their outstanding performance across various tasks. Among them, self-supervised contrastive learning techniques, exemplified by MoCo, leverage dynamic dictionaries to autonomously extract information from images without requiring labeled data and have proven effective in foundational model training. We introduce a supervised contrastive loss function built upon MoCo to enhance model performance further. Compared to self-supervised contrastive loss, supervised contrastive loss incorporates class supervision, bringing feature representations of the same class closer together while pushing those of different classes further apart. This enables the model to capture inter-class differences better and extract more discriminative and efficient features, thereby supporting FSCIPR tasks.

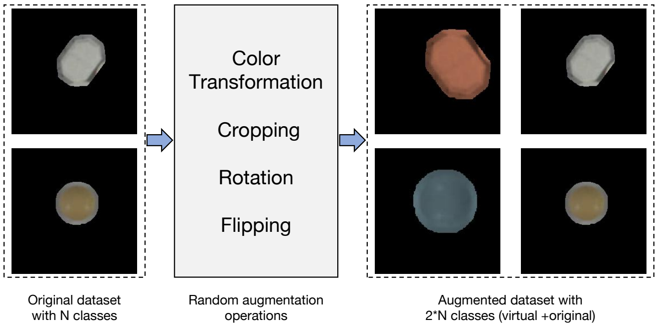

Color Transformation
Cropping
Rotation
Flipping
Original dataset with N classes
Random augmentation operations
Augmented dataset with 2*N classes (virtual +original)

Fig. 3: Virtual class generation process. The process of generating virtual classes through random augmentation operations, such as color transformation, cropping, rotation, and flipping. The original dataset with $M$ classes is expanded to form an augmented dataset with $2M$ classes, combining virtual and original classes.

Specifically, for a batch of image-label pairs $\{(x_i, y_i)\}_{i=0}^{b}$, random augmentations are applied to each image, resulting in the creation of a query view $x_q = Aug_q\left(x\right)$ and a key view $x_k = Aug_k\left(x\right)$. These augmented views are then transformed into $L_2$-normalized representations $q$ and $k$ through $\zeta\left(\cdot\right)$, where $\zeta = h \circ g$ represents a composition of the full image encoder $g$ and the projector $h$. The supervised MoCo-based contrastive loss $\mathcal{L}_{sup}$ is calculated using these representations to optimize the model's feature discrimination capabilities, formulated as:

$$\mathcal{L}_{sup}(\zeta; x_i, y_i, \mathcal{T}) = -\frac{1}{|\mathbf{k}_+|}\sum_{k_+\in\mathbf{k}_+}\log\frac{\exp(q_i^\top k_+/\mathcal{T})}{\sum_{k\in\mathbf{k}}\exp(q_i^\top k/\mathcal{T})}, \tag{2}$$

where $\mathbf{k}$ is the set of all the key representations, $\mathbf{k}_+$ indicates the positive set, meaning the elements in $\mathbf{k}$ that belong to the same class as $x_i$, and the $\mathcal{T}$ is a temperature parameter. After the introduction of the supervised contrastive loss, the joint training loss can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \alpha\mathcal{L}_{sup}, \tag{3}$$

where $\mathcal{L}_{ce}$ is the cross-entropy loss, $\mathcal{L}_{sup}$ represents the supervised contrast loss, and $\alpha$ is a hyper-parameter that weights the importance of $\mathcal{L}_{sup}$.

*3) Center-Triplet Loss:* Although we have previously introduced virtual class-based forward compatibility strategies and supervised contrastive loss to improve the model's forward compatibility, generalization, and, to some extent, discriminative capability, there remains a need to enhance its discriminative power further when dealing with fine-grained pill classes and high inter-class similarity. Fine-grained classes are characterized by minimal differences between classes and significant variations within the same class, posing a significant challenge to the model's feature extraction and classification capabilities. To address this issue, we employ the CT loss [13] to enhance intra-class compactness and inter-class separation. This ensures that samples belonging to the same class are more tightly grouped within the feature space, and samples from different classes are pushed further apart, improving the model's discriminative capability and classification accuracy.

The CT loss combines the advantages of Center loss [47] and Triplet loss [48]. It anchors a class by minimizing the distance between its samples and the class center while ensuring it is smaller than the distance to the nearest different class center. The core objective of CT loss is to enhance intra-class compactness and inter-class separability by optimizing the feature representations, thereby improving the model's overall performance in distinguishing between classes. The CT loss is mathematically expressed as:

$$\mathcal{L}_{ct}\left(\phi; x\right) = \max(0, m + \left\|\phi\left(x\right) - c_y\right\| - \min_{j \neq y} \left\|c_y - c_j\right\|), \tag{4}$$

where the triplet $(x, c_y, c_j)$ includes a sample $x$, its class center $c_y$, and the nearest different class center $c_j$. The loss updates the backbone to ensure the distance between the sample's feature and its center $c_y$ is less than the distance to $c_y$ by a margin $m$. After the introduction of the CT loss, the joint training loss can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \alpha\mathcal{L}_{sup} + \beta\mathcal{L}_{ct}, \tag{5}$$

where $\lambda$ balances the CT and classification losses.

### *E. Incremental Session Learning*

In incremental session learning, the core goal is to ensure that the model can retain previously learned knowledge while learning new classes, avoiding catastrophic forgetting. However, conventional KD methods often rely on large amounts of samples to maintain the stability of old knowledge. However, the issue of limited samples in incremental sessions of FSCIPR makes conventional KD difficult. To address this, we propose a PFS strategy, which generates pseudo-features of old classes to mitigate the impact of sample scarcity, thereby reducing storage costs and facilitating the transfer of old knowledge. We introduce a multi-dimensional KD approach based on this synthesis strategy, including response-based and relation-based distillation (KD1 and KD2). These methods comprehensively leverage the relationships between old and new classes, preserving the stability of old knowledge while strengthening the model's capability to acquire new classes. This enables a smooth transition in the incremental learning process, prevents catastrophic forgetting, and maintains model performance.

*1) Pseudo Feature Synthesis:* In incremental learning, DR and KD are commonly employed to maintain the stability of previously learned knowledge. However, these methods typically require additional memory to store samples from prior learning sessions, leading to increasing storage demands as new knowledge is introduced. FSCIPR faces more complex challenges than other incremental learning tasks, particularly due to the limited sample availability during incremental sessions, making traditional DR and KD methods difficult to apply directly. Additionally, storing data can pose potential privacy risks for patients. To tackle these challenges, a strategy integrating PFS is introduced, allowing the flexible generation of reliable features without storing extensive datasets, thereby reducing storage requirements and addressing privacy concerns.

Our approach maintains a frozen feature extractor during incremental sessions while training only the fully connected layers to adapt to new classes. This allows us to utilize the features extracted by the frozen feature extractor for DR and KD, avoiding storing raw image data, significantly reducing storage requirements, and minimizing privacy concerns. However, due to the few-shot samples, relying solely on these features may be insufficient for a comprehensive knowledge review. Therefore, we introduce a strategy combining existing feature vectors with model predictions to synthesize and select reliable pseudo-features, effectively overcoming the limitations of the few-shot setting. This strategy supports the transfer of old knowledge during the DR and KD processes, ensuring the stability of previously learned knowledge while strengthening the model's ability to acquire incremental classes.

Our PFS method is detailed in Alg. 1, which operates during session $i$. It takes the training dataset, trained model, and specified numbers of stored and synthesized features per class as input. The algorithm extracts and computes the mean features for each class, randomly selects and stores some features, and then synthesizes pseudo-features using a combination of randomly generated scalars and stored features. By leveraging model predictions, reliable pseudo-features are selected to form the synthesized pseudo-feature set, effectively utilizing existing class features and overcoming the limitations of knowledge review under few-shot conditions.

*2) Data Replay and Multi-dimensional Knowledge Distillation:* After generating reliable pseudo-features, our framework ensures a balanced integration of new and old knowledge through DR and Multi-dimensional KD. The flexibility in adjusting the number of synthesized pseudo-features allows it to meet the requirements of both DR and KD. Specifically, during session $t$, we first perform DR using pseudo-features representing old classes. This ensures that the fully connected layers of the model learn both the new class features and the pseudo-features of the old classes, thus achieving an equilibrium between integrating new knowledge and retaining previously learned information. The mathematical representation is as follows:

$$\mathcal{L}_{total}\left(\psi_t; f_{rv}, y_{rv}\right) = \mathcal{L}_{ce}\left(\psi_t\left(f_{rv}\right), y_{rv}\right), \tag{6}$$

where $\psi_t$ denotes the fully connected layers, $f_{rv}$ represents the feature from the union of features derived from the ongoing session and the synthesized pseudo-features of earlier sessions, and $y_{rv}$ is the corresponding label.

In addition to DR, our framework introduces multi-dimensional KD to facilitate knowledge transfer from the old model to the new one. The distillation process consists of two main components: KD1, which utilizes the Kullback-Leibler divergence to measure differences in probability distributions, and KD2, which incorporates both Euclidean distance and cosine similarity to capture relational information. The Kullback-Leibler divergence evaluates the variation across the softened probability outputs produced by the teacher model and the

**Algorithm 1:** PFS for session $i$

```
Input  : Training dataset $D^i_{train}$, trained model $\phi_i$
         (with feature extractor $\varphi_i$ and fully connected
         layers $\psi_i$), number of features $P$ to be stored
         per class, number of pseudo-features $Q$ to
         synthesize per class.
Output: Set of synthesized pseudo-features $S$.
Initialize an empty set $S$ for storing pseudo-features.;
for each class $c \in D^i_{train}$ do
    Extract features $\{f_{c_i}\}_{i=1}^{N_c}$ for all $N_c$ samples in
      class $c$ using $\varphi_i$.;
    Compute mean feature $\mu_c = \frac{1}{N_c}\sum_{i=1}^{N_c} f_{c_i}$.;
    Store $P$ randomly selected features and mean
      feature $\mu_c$ in a memory bank $M_c$.;
end
for each class $c \in D^i_{train}$ do
    Set $count_c = 0$;
    while $count_c < Q$ do
        Select a random feature vector $f$ from $M_c$;
        Generate a random scalar $\alpha \in (0, 1)$;
        Synthesize pseudo-feature
          $f_v = \alpha f + (1 - \alpha)\mu_c$;
        Predict class label of $f_v$ using $\psi_i(f_v)$;
        if $\psi_i(f_v)$ predicts class $c$ then
            Append $f_v$ to $S$;
            Increment $count_c$ by 1;
        end
    end
end
```

student model, mathematically expressed as:

$$\mathcal{L}_{kd1}\left(\psi_t, \psi_{t-1}; f_{rv}\right) = KL\left(\frac{\psi\left(f_{rv}\right)}{T}, \frac{\psi_{t-1}\left(f_{rv}\right)}{T}\right), \tag{7}$$

where $T$ denotes the temperature parameter.

To further preserve consistency in the relationships between samples, we introduce KD2, which enhances the transfer of old knowledge by aligning both the pairwise distance and angular relationships between samples in the feature space of the new and old models. This distillation approach consists of two main components. First, the distance relationship loss ensures that the pairwise distances between samples in the new and old models are aligned. This loss is expressed as:

$$\mathcal{L}_{dist}\left(\varphi_t, \varphi_{t-1}; f_{rv}\right) = \frac{1}{N}\sum_{i,j}\left|\frac{\text{dist}\left(\varphi_{t-1}(f_{rv_i}), \varphi_{t-1}(f_{rv_j})\right)}{\overline{\text{dist}(\varphi_{t-1})}} - \frac{\text{dist}\left(\varphi_t(f_{rv_i}), \varphi_t(f_{rv_j})\right)}{\overline{\text{dist}(\varphi_t)}}\right|, \tag{8}$$

where $\text{dist}(\cdot,\cdot)$ denotes the Euclidean distance between features, and $\overline{\text{dist}(\varphi)}$ represents the normalized mean distance for the features of the respective model.

In addition, the angular relationship loss ensures consistency of angular relationships between feature vectors in the student and teacher models. This is formulated as:

$$\mathcal{L}_{\text{angle}}\left(\varphi_t, \varphi_{t-1}; f_{rv}\right) = \frac{1}{N}\sum_{i,j}\left|\cos\left(\mathbf{v}_t(f_{rv_i}, f_{rv_j})\right) - \cos\left(\mathbf{v}_{t-1}(f_{rv_i}, f_{rv_j})\right)\right|, \tag{9}$$

where $\mathbf{v}_t(f_{rv_i}, f_{rv_j})$ and $\mathbf{v}_{t-1}(f_{rv_i}, f_{rv_j})$ denote the normalized pairwise feature differences between samples in the new and old models, respectively, and $\cos(\cdot)$ represents cosine similarity.

The total KD2 loss is defined as a weighted combination of the two components:

$$\mathcal{L}_{kd2} = \lambda \cdot \mathcal{L}_{\text{distance}} + (1 - \lambda) \cdot \mathcal{L}_{\text{angle}}, \tag{10}$$

where $\lambda_{\text{distance}}$ and $\lambda_{\text{angle}}$ are weighting factors for the distance and angular relationship losses, respectively. This KD2 approach effectively preserves the consistency of feature relationships between the student and teacher models, thereby facilitating the smooth transfer of knowledge and maintaining stability during incremental learning sessions.

The joint training loss in incremental sessions combines classification loss, KD1 and KD2, formulated as:

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \delta\mathcal{L}_{kd1} + \varepsilon\mathcal{L}_{kd2}, \tag{11}$$

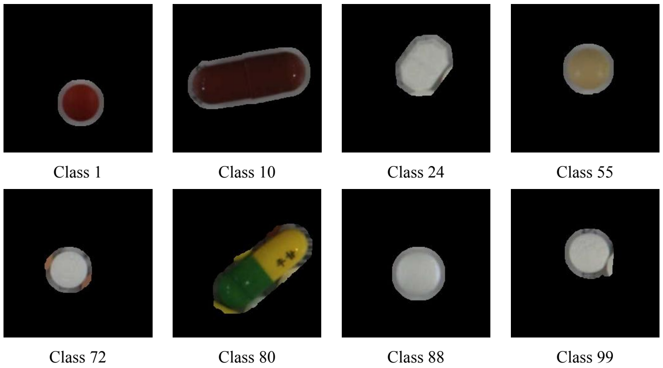
where $\delta$ and $\varepsilon$ balance the KD1 loss and KD2 loss.

## IV. Experiments

### A. Dataset and Evaluation Metric

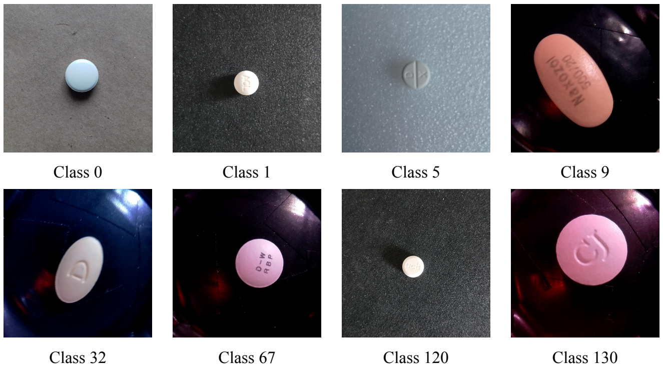
**FCPill:** The FCPill dataset was introduced to evaluate pill recognition within the FSCIL setting [13]. It comprises 100 classes, each with 600 high-resolution images, offering a robust base for experiments in few-shot learning and class-incremental learning. To follow the FSCIL setup, the dataset is organized with 60 classes forming the base session. Additionally, 40 classes are distributed across eight incremental sessions, with each session introducing 5 classes. In each incremental session, the training dataset follows the 5-way 5-shot format, where 5 samples per class are randomly selected for training. Fig. 4 illustrates representative images from FCPill, showcasing pills' varied appearances and packaging across different classes.

***m*CURE:** In addition to the FCPill dataset, we evaluated our method using another publicly available pill image dataset, CURE [20]. Originally designed for few-shot learning tasks, the CURE dataset comprises 1,873 images distributed across 196 classes, each containing approximately 45 samples. Following the setup established by *Zhang et al.* [13], we adapted it to evaluate pill recognition within the FSCIL setting. Specifically, we used a subset of CURE, referred to as *m*CURE, which consists of 171 classes for experimentation. The *m*CURE dataset comprised 91 base classes and 80 incremental classes. The incremental classes were further divided across eight incremental sessions, with each session's training data organized in a 10-way 5-shot format. Fig. 5 presents representative examples.


Class 1 Class 10 Class 24 Class 55

Class 72 Class 80 Class 88 Class 99

Fig. 4: Examples in the FCPill dataset. It can be found that all pills are structured, with most being capsules or round tablets. However, unlike other image datasets, the pills in this dataset present fine-grained challenges, as categories 72, 88, and 99 show high similarity, adding extra difficulty to FSCIPR.


Class 0 Class 1 Class 5 Class 9

Class 32 Class 67 Class 120 Class 130

Fig. 5: Examples in the *m*CURE dataset. Similar to FCPill, the pills are structured in shape, but there is a high degree of similarity between different categories, such as categories 0, 1, and 120. These fine-grained challenges further increase the difficulty of FSCIPR.

**Evaluation Metric:** To comprehensively evaluate our framework, we utilized three metrics: 1) Accuracy for each session; 2) *Performance Drop* (PD), which measures the absolute decline in accuracy from the first session to the last, calculated as $PD = A_0 - A_N$, where $A_0$ represents the classification accuracy in the base session, and $A_N$ represents the accuracy in the final session [41]; and 3) The *Average Accuracy* (AA) across all sessions. Higher values indicate better performance for both accuracy and AA, whereas a lower value is preferable for PD.

### B. Implementation Details

In most FSCIL research, ResNet18 serves as the standard backbone network, and we adopted it for the pill datasets in our study as well. An additional fully connected layer is attached to the network's output. The training process uses the SGD optimizer with a learning rate of 0.1, momentum set to 0.9, and a weight decay of 0.0005. Upon completing the base session training, all parameters, except those in the fully connected layers, are frozen; only these layers are trained during the incremental sessions. Classification leverages the softmax function. Our implementation utilizes PyTorch 2.1 and Python 3.9, with training on an Nvidia Tesla V100 GPU.

### C. Comparison with State of The Arts

Our research evaluates the proposed method against several advanced approaches using the FCPill and *m*CURE datasets. The evaluated methods include CEC [45], LIMIT [49], FACT [46], ALICE [50], SSFE-Net [51], BiDistFSCIL [52], and SAVC [53], chosen for their strong performance and relevance to FSCIL. Comprehensive results for FCPill and *m*CURE are presented in Fig. 6, Tab. I and Tab. II, providing a thorough comparison of our approach with these methods across different sessions, thereby highlighting the strengths and effectiveness of our framework.

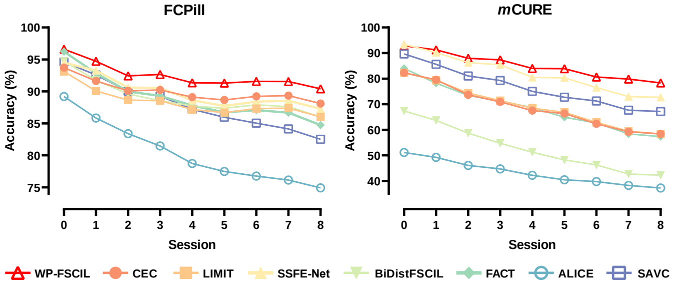


Fig. 6: Comparison with SOTA methods on FCPill and *m*CURE. Our method, WP-FSCIL, comprehensively surpasses other methods.

We compare WP-FSCIL with several advanced methods on the FCPill dataset, and the results are provided in Fig.6 and Tab.I, highlighting WP-FSCIL's remarkable advantages in various dimensions. From the perspective of individual session performance, WP-FSCIL establishes a robust baseline with an impressive accuracy of 96.6% in the initial session. It maintains high stability throughout the incremental sessions, with accuracies of 91.35% in Session 4 and 90.40% in Session 8. This consistent performance across sessions underscores its ability to handle incremental learning scenarios effectively. In contrast, other methods, such as SAVC and FACT, exhibit more pronounced declines. For instance, SAVC's accuracy drops from 94.62% in Session 0 to 82.50% in Session 8, and FACT decreases from 96.22% to 84.73%. Regarding AA, WP-FSCIL achieves 92.51%, outperforming all competing methods, including SSFE-Net (89.94%) and CEC (90.01%). This highlights WP-FSCIL's superior capacity for balancing learning new and old classes. Furthermore, regarding PD, WP-FSCIL achieves a competitive value of 6.20%, slightly higher than CEC's 5.59% but substantially lower than FACT's 11.49% and SAVC's 12.12%. It is important to note that while PD provides a straightforward measure of accuracy decline from the first to the last session, it does not fully capture the resistance to forgetting across all sessions. Therefore, WP-FSCIL's exceptional AA further validates its robustness and effectiveness. Overall, WP-FSCIL demonstrates unparalleled performance in individual session accuracy, average accuracy,

TABLE I: Comparison results of our WP-FSCIL against other SOTA methods on FCPill. (In %).

| Dataset | Method | Venue | Accuracy in each session | | | | | | | | | AA↑ | PD↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | | |
| FCPill | CEC [45] | CVPR 21 | 93.71 | 91.63 | 90.08 | 90.22 | 89.10 | 88.67 | 89.22 | 89.34 | 88.11 | 90.01 | **5.59** |
| | LIMIT [49] | TPAMI 22 | 93.11 | 90.07 | 88.65 | 88.54 | 87.18 | 86.68 | 87.33 | 87.39 | 86.01 | 88.33 | 7.10 |
| | FACT [46] | CVPR 22 | 96.22 | 92.84 | 89.98 | 89.31 | 87.80 | 86.72 | 87.09 | 86.67 | 84.73 | 89.04 | 11.49 |
| | ALICE [50] | ECCV 22 | 89.20 | 85.84 | 83.40 | 81.46 | 78.73 | 77.48 | 76.76 | 76.13 | 74.91 | 80.43 | 14.29 |
| | SSFE-Net [51] | WACV 23 | 94.49 | 93.26 | 90.61 | 90.53 | 88.63 | 87.72 | 88.40 | 88.54 | 87.29 | 89.94 | 7.20 |
| | BiDistFSCIL [52] | CVPR 23 | 94.71 | 91.74 | 89.61 | 88.54 | 87.73 | 87.36 | 87.90 | 87.62 | 86.00 | 89.02 | 8.71 |
| | SAVC [53] | CVPR 23 | 94.62 | 92.57 | 90.02 | 89.28 | 87.20 | 85.95 | 85.04 | 84.14 | 82.50 | 87.92 | 12.12 |
| | **WP-FSCIL** | - | **96.60** | **94.72** | **92.44** | **92.65** | **91.35** | **91.31** | **91.58** | **91.55** | **90.40** | **92.51** | 6.20 |

and stability, confirming its significant advantage in the challenging FSCIPR task. This consistent superiority across key metrics establishes WP-FSCIL as a reliable FSCIL framework.

Fig. 6 and Tab. II compare WP-FSCIL with other advanced methods on the *m*CURE dataset, highlighting WP-FSCIL's significant advantages from multiple perspectives. First, in the base session, WP-FSCIL achieves a classification accuracy of 92.82%, which is slightly lower than that of SSFE-Net, potentially due to SSFE-Net leveraging a deeper ResNet-50 for training assistance [51]. In the incremental sessions, WP-FSCIL exhibits relatively small performance declines. For instance, it maintains accuracies of 84.01% and 78.35% in Session 4 and Session 8, respectively, showcasing high stability. In contrast, other methods exhibit more substantial performance drops; for example, SAVC's accuracy decreases from 89.63% in Session 0 to 67.14% in Session 8, and SSFE-Net drops from 93.41% to 72.71%. In terms of AA, WP-FSCIL achieves 85.12%, significantly surpassing SSFE-Net (82.03%) and SAVC (76.59%), thereby confirming its superior overall capability to learn new and old classes simultaneously. WP-FSCIL's performance PD of 14.47% is slightly higher than ALICE's 13.88%. Overall, WP-FSCIL demonstrates outstanding performance in single-session accuracy, average accuracy, and stability, establishing its superiority in FSCIPR tasks.

### D. Ablation Study

Ablation studies were conducted to validate the significance of the proposed components, with a focus on the key aspects of our method. These components include virtual class generation (VCG), CT loss and supervised contrastive learning (Loss), PFS, KD1, and KD2. The results for FCPill and *m*CURE are presented in Tab. III and Tab. IV, demonstrating the impact of each component on overall performance.

Tab. III showcases the ablation study results on the FCPill dataset, highlighting the impact of each module on overall performance through the evaluation of different component combinations. The results clearly illustrate the importance of each component. Using only the baseline model (without any additional components), the accuracy in the initial session (Session 0) is 95.02%, but it drops significantly to 51.38% in Session 8, with an AA of only 70.41% and a PD of 43.64%, indicating the baseline model's insufficient ability to maintain performance in incremental learning. Introducing virtual class generation (VCG) significantly enhances forward compatibility, improving AA to 76.48% and reducing PD to 32.99%. This improvement is reflected in the incremental sessions, where accuracy rises to 63.22% in Session 8, demonstrating how VCG helps the model adapt to new classes. Adding the joint loss (CT loss and supervised contrastive learning) on top of VCG further improves performance. The model achieves an AA of 77.57%, with accuracy in Session 8 increasing to 61.20%. This demonstrates the effectiveness of the joint loss in enhancing fine-grained feature discrimination and generalization ability. However, PD slightly increases to 35.40%, indicating room for improvement in cross-session stability. Incorporating PFS results in significant performance gains. AA improves to 92.41%, and PD dramatically decreases to 6.55%. This improvement demonstrates that PFS significantly stabilizes performance across sessions, as evidenced by the Session 8 accuracy rising to 90.05%. Adding KD1 on top of PFS maintains the AA at 92.41%, while further reducing PD to 6.33%, validating KD1's effectiveness in mitigating forgetting through distillation. Finally, incorporating KD2 achieves optimal overall performance, increasing AA to 92.51% and reducing PD to 6.20%. KD2 further stabilizes model performance by preserving the consistency of sample relationships in the feature space, as reflected in the high accuracy across all sessions, such as 91.35% in Session 4 and 90.40% in Session 8.

Tab. IV presents the ablation study results conducted on the *m*CURE dataset, illustrating the contributions of various components in our WP-FSCIL framework. The basic model, without incorporating any additional components, achieves an initial session accuracy of 84.84%, which drops sharply to 39.15% by Session 8. This results in a low AA of 56.81% and a high PD of 45.69%, highlighting severe forgetting across incremental sessions. Introducing virtual class generation (VCG) improves AA to 57.08% but results in a higher PD of 49.53%, indicating its role in forward compatibility yet has a limited effect on forgetting mitigation. Adding the combined CT loss and supervised contrastive learning (Loss) significantly enhances the AA to 66.28% while reducing the PD to 44.09%, showcasing its effectiveness in fine-grained feature learning and generalization. The inclusion of PFS results in a remarkable improvement, achieving an AA of 85.20% and substantially lowering the PD to 15.29%. Adding

TABLE II: Comparison results of our WP-FSCIL against other SOTA methods on *m*CURE. (In %).

| Dataset | Method | Venue | Accuracy in each session | | | | | | | | | AA↑ | PD↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | | |
| *m*CURE | CEC [45] | CVPR 21 | 82.26 | 79.52 | 73.65 | 70.92 | 67.52 | 66.35 | 62.36 | 59.24 | 58.40 | 68.91 | 23.86 |
| | LIMIT [49] | TPAMI 22 | 82.26 | 79.33 | 74.35 | 71.40 | 68.44 | 66.79 | 62.87 | 59.42 | 58.31 | 69.24 | 23.95 |
| | FACT [46] | CVPR 22 | 84.00 | 78.24 | 73.39 | 71.20 | 68.67 | 64.90 | 62.68 | 58.39 | 57.31 | 68.75 | 26.69 |
| | ALICE [50] | ECCV 22 | 51.10 | 49.21 | 46.04 | 44.71 | 42.21 | 40.46 | 39.74 | 38.23 | 37.22 | 43.21 | **13.88** |
| | SSFE-Net [51] | WACV 23 | **93.41** | 90.22 | 86.24 | 85.46 | 80.48 | 80.27 | 76.53 | 72.95 | 72.71 | 82.03 | 20.70 |
| | BiDistFSCIL [52] | CVPR 23 | 67.36 | 63.66 | 58.69 | 54.71 | 51.22 | 48.19 | 46.23 | 42.70 | 42.22 | 52.78 | 25.14 |
| | SAVC [53] | CVPR 23 | 89.63 | 85.57 | 81.04 | 79.28 | 75.04 | 72.72 | 71.26 | 67.64 | 67.14 | 76.59 | 22.49 |
| | **WP-FSCIL** | - | 92.82 | **91.26** | **87.95** | **87.33** | **84.01** | **83.90** | **80.63** | **79.83** | **78.35** | **85.12** | 14.47 |

TABLE III: Ablation study on FCPill. For short, **VCG** indicates virtual class generation; **Loss** stands for the combination of our proposed CT loss and supervised contrastive learning. (In %)

| VCG | Loss | PFS | KD1 | KD2 | Accuracy in each session | | | | | | | | | AA↑ | PD↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | | |
| | | | | | 95.02 | 86.50 | 79.59 | 74.24 | 69.39 | 65.25 | 58.16 | 54.19 | 51.38 | 70.41 | 43.64 |
| ✓ | | | | | 96.21 | 89.35 | 81.40 | 77.97 | 74.63 | 71.45 | 68.11 | 65.98 | 63.22 | 76.48 | 32.99 |
| ✓ | ✓ | | | | 96.60 | 91.43 | 84.90 | 81.99 | 76.86 | 71.29 | 69.24 | 64.65 | 61.20 | 77.57 | 35.40 |
| ✓ | ✓ | ✓ | | | 96.60 | 95.13 | 92.71 | 92.80 | 91.23 | 90.83 | 91.17 | 91.19 | 90.05 | 92.41 | 6.55 |
| ✓ | ✓ | ✓ | ✓ | | 96.60 | 94.73 | 92.40 | 92.58 | 91.18 | 91.09 | 91.43 | 91.46 | 90.27 | 92.41 | 6.33 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 96.60 | 94.72 | 92.44 | 92.65 | 91.35 | 91.31 | 91.58 | 91.55 | 90.40 | 92.51 | 6.20 |

TABLE IV: Ablation study on *m*CURE. The abbreviation is consistent with Table III. (In %)

| VCG | Loss | PFS | KD1 | KD2 | Accuracy in each session | | | | | | | | | AA↑ | PD↓ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | | |
| | | | | | 84.84 | 74.41 | 64.24 | 60.28 | 53.76 | 48.03 | 45.41 | 41.20 | 39.15 | 56.81 | 45.69 |
| ✓ | | | | | 89.63 | 73.98 | 63.24 | 58.52 | 52.59 | 48.03 | 45.54 | 42.10 | 40.10 | 57.08 | 49.53 |
| ✓ | ✓ | | | | 92.82 | 82.72 | 72.03 | 68.90 | 63.65 | 61.18 | 55.98 | 50.51 | 48.73 | 66.28 | 44.09 |
| ✓ | ✓ | ✓ | | | 92.82 | 91.79 | 89.01 | 87.40 | 84.53 | 83.93 | 80.49 | 79.26 | 77.53 | 85.20 | 15.29 |
| ✓ | ✓ | ✓ | ✓ | | 92.82 | 91.21 | 88.08 | 87.42 | 84.05 | 83.66 | 80.52 | 79.26 | 77.90 | 84.99 | 14.92 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 92.82 | 91.26 | 87.95 | 87.33 | 84.01 | 83.90 | 80.63 | 79.83 | 78.35 | 85.12 | 14.47 |

KD1 maintains a similar AA of 84.99% and slightly lowers the PD to 14.92%, validating KD1's contribution to retaining old knowledge. Finally, the addition of KD2 further optimizes the framework's performance, achieving the highest AA of 85.12% and reducing PD to 14.47%. This demonstrates KD2's critical role in preserving sample relationships and mitigating forgetting. Overall, these components work synergistically, enabling our WP-FSCIL to excel in the FSCIPR task.

### *E. Impact of Hyper-parameter*

A comprehensive analysis was performed to examine how various hyper-parameters influence the performance of the proposed WP-FSCIL framework on the FCPill and *m*CURE datasets, as depicted in Fig. 7 and Fig. 8. Due to the unique characteristics of each dataset, specific hyper-parameter tuning is required for each dataset.

For the FCPill dataset, the influence of $\alpha$ from Eq. 5 is shown in Fig. 7(a). When $\alpha$ is set to 0.04, the method reaches its peak performance across all sessions. Regarding the CT loss, the analysis focuses on $\beta$ from Eq. 5 and the margin $m$ from Eq. 4. As illustrated in Fig. 8(a), the peak performance in the final session is achieved when $\beta$ is 0.06 and $m$ is 4. For PFS, the key hyper-parameters include $P$ (the number of real features stored per class) and $Q$ (the number of pseudo-features synthesized per class) from Alg. 1. Fig. 8(b) illustrates that the optimal performance in the final session is achieved with $P = 5$ and $Q = 8$. It is worth noting that while $P = 5$ might give the impression of complete replay, the stored data are not the raw image data. Furthermore, our method can achieve comparable performance with other values of $P$; the parameter tuning here is solely aimed at achieving optimal performance. For KD1, the primary hyper-parameters are the temperature parameter $T$ and the weighting coefficient $\delta$ in the overall loss function (Eq. 11). As depicted in Fig. 8(c), the optimal performance on the FCPill dataset is observed when $\delta = 0.8$ and $T = 3$. We also observed that KD1 sensitivity on the FCPill dataset is lower than on *m*CURE. This can be attributed to two factors: first, as the temperature increases, the teacher model already provides sufficient soft information, and further increases do not enhance the effective information conveyed; second, the characteristics or distribution of the FCPill dataset make class distinctions less dependent on temperature adjustments, reducing the impact of temperature and, consequently, the sensitivity during the KD1 process. For KD2, the analysis focuses on the ratio of distance relationships and cosine relationships ($\lambda$ in Eq. 10) and the weighting coefficient $\varepsilon$ in Eq. 11. As illustrated in Fig. 8(d), setting $\varepsilon = 1.0$ and $\lambda = 0.1$ yields the highest performance in the final session.

Overall, on the FCPill dataset, the best performance is achieved with the hyper-parameters $\{\alpha, \beta, m, P, Q, \delta, T, \varepsilon, \lambda\}$ = $\{0.04, 0.06, 4, 5, 8, 0.8, 3, 1.0, 0.1\}$. Similarly, for the *m*CURE dataset, the best performance is observed with the hyper-parameters $\{\alpha, \beta, m, P, Q, \delta, T, \varepsilon, \lambda\} = \{0.04, 0.1, 2, 3, 11, 0.7, 5, 0.5, 0.3\}$.

## V. Conclusion

To conclude, our paper introduces WP-FSCIL, a novel framework specifically developed to tackle the key challenges of the FSCIPR task, including overfitting, fine-grained classification issues, and catastrophic forgetting. To mitigate the overfitting risk associated with the limited samples in incremental sessions, WP-FSCIL employs a parameter-freezing strategy after base session training, allowing only the fully connected layers to be updated. This approach preserves the model's generalization capabilities, effectively preventing overfitting caused by the few-shot setting. For fine-grained classification, the framework integrates CT loss and supervised contrastive learning during the base session to enable the backbone network to learn robust and discriminative feature representations. To address catastrophic forgetting, WP-FSCIL incorporates multiple innovative strategies. During the base session, virtual class generation is introduced, equipping the model with forward compatibility to anticipate and adapt to future incremental learning scenarios in advance. In incremental sessions, a multi-dimensional KD strategy based on flexible PFS is employed. This strategy can synthesize any number of pseudo-features for old classes, addressing the limitations of KD caused by insufficient samples. Furthermore, the framework utilizes both KD1 and KD2 to comprehensively retain knowledge of previously learned classes, enhancing resistance to forgetting. Extensive experiments on two publicly available pill datasets were performed to assess the performance of WP-FSCIL. In comparison with advanced FSCIL methods, WP-FSCIL achieved superior results, demonstrating its effectiveness and robustness. Additionally, comprehensive ablation studies validated the contributions of each component within the framework, while a detailed hyperparameter analysis provided insights into the impact of key parameters on performance and identified optimal configurations. These findings collectively highlight the significant potential of WP-FSCIL as a robust solution for FSCIPR tasks.

Despite the significant advantages demonstrated by WP-FSCIL in FSCIPR tasks, there are still several limitations that warrant further investigation. First, the current FSCIL setting typically relies on the N-Way K-Shot assumption, which is overly idealized and may not fully align with real-world applications. Future research should focus on exploring more realistic settings, such as managing imbalanced data or dynamic variations in sample numbers encountered in real-world scenarios. Second, the current study focuses primarily on single-object classification tasks, while pill targets in real-world applications often appear in multi-object, multi-class contexts. Therefore, future work could further investigate methods for incremental learning in multi-object, multi-class scenarios, enhancing the system's applicability in complex environments. Lastly, with the growing adoption of smart devices, multi-terminal collaboration has become a critical trend. Future research could explore developing few-shot class-incremental pill recognition systems within a federated learning framework, enabling effective data privacy protection while improving the model's generalization and robustness across multiple devices. These directions not only broaden the applicability of WP-FSCIL but also provide new insights and challenges for research in the FSCIPR field.

[1] WHO, https://www.who.int/initiatives/medication-without-harm.

[2] X. Ma, B. Zhang, C. Ma, and Z. Ma, "Co-regularized nonnegative matrix factorization for evolving community detection in dynamic networks," *Information Sciences*, vol. 528, pp. 265–279, 2020.

[3] X. Ma, L. Yu, P. Wang, and X. Yang, "Discovering dna methylation patterns for long non-coding rnas associated with cancer subtypes," *Computational biology and chemistry*, vol. 69, pp. 164–170, 2017.

[4] X. Ma and L. Gao, "Discovering protein complexes in protein interaction networks via exploring the weak ties effect," *BMC systems biology*, vol. 6, pp. 1–15, 2012.

[5] X. Ma, W. Tang, P. Wang, X. Guo, and L. Gao, "Extracting stage-specific and dynamic modules through analyzing multiple networks associated with cancer progression," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 15, no. 2, pp. 647–658, 2016.

[6] D. Hu, Z. Dong, K. Liang, H. Yu, S. Wang, and X. Liu, "High-order topology for deep single-cell multi-view fuzzy clustering," *IEEE Transactions on Fuzzy Systems*, 2024.

[7] D. Hu, K. Liang, S. Zhou, W. Tu, M. Liu, and X. Liu, "scdfc: a deep fusion clustering method for single-cell rna-seq data," *Briefings in Bioinformatics*, vol. 24, no. 4, p. bbad216, 2023.

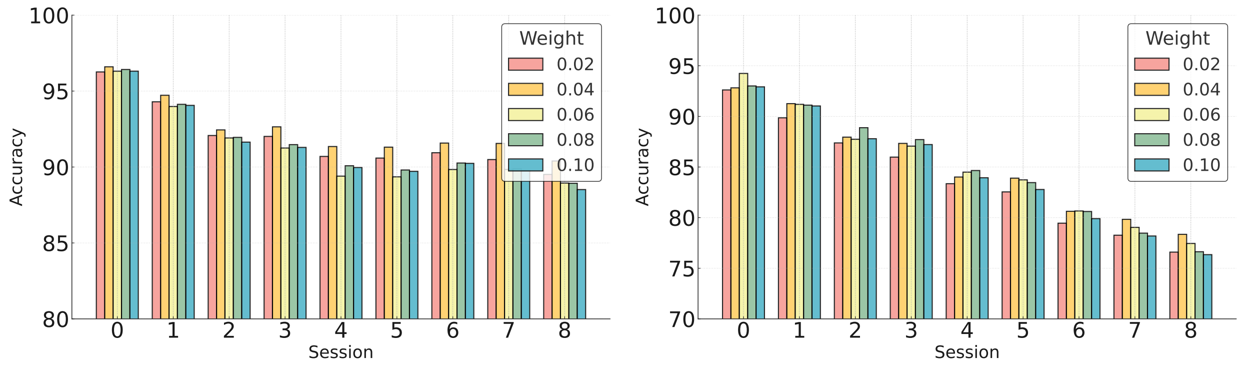
[8] D. Hu, R. Guan, K. Liang, H. Yu, H. Quan, Y. Zhao, X. Liu, and K. He, "scegg: an exogenous gene-guided clustering method for single-cell transcriptomic data," *Briefings in Bioinformatics*, vol. 25, no. 6, p. bbae483, 2024.

[9] H. Jiang, Y. Yin, J. Zhang, W. Deng, and C. Li, "Deep learning for liver cancer histopathology image analysis: A comprehensive survey," *Engineering Applications of Artificial Intelligence*, vol. 133, p. 108436, 2024.

[10] J. Zhang, C. Li, Y. Yin, J. Zhang, and M. Grzegorzek, "Applications of artificial neural networks in microorganism image analysis: a comprehensive review from conventional multilayer perceptron to popular convolutional neural network and potential visual transformer," *Artificial Intelligence Review*, vol. 56, no. 2, pp. 1013–1070, 2023.

[11] J. Zhang, C. Li, S. Kosov, M. Grzegorzek, K. Shirahama, T. Jiang, C. Sun, Z. Li, and H. Li, "Lcu-net: A novel low-cost u-net for environmental microorganism image segmentation," *Pattern Recognition*, vol. 115, p. 107885, 2021.

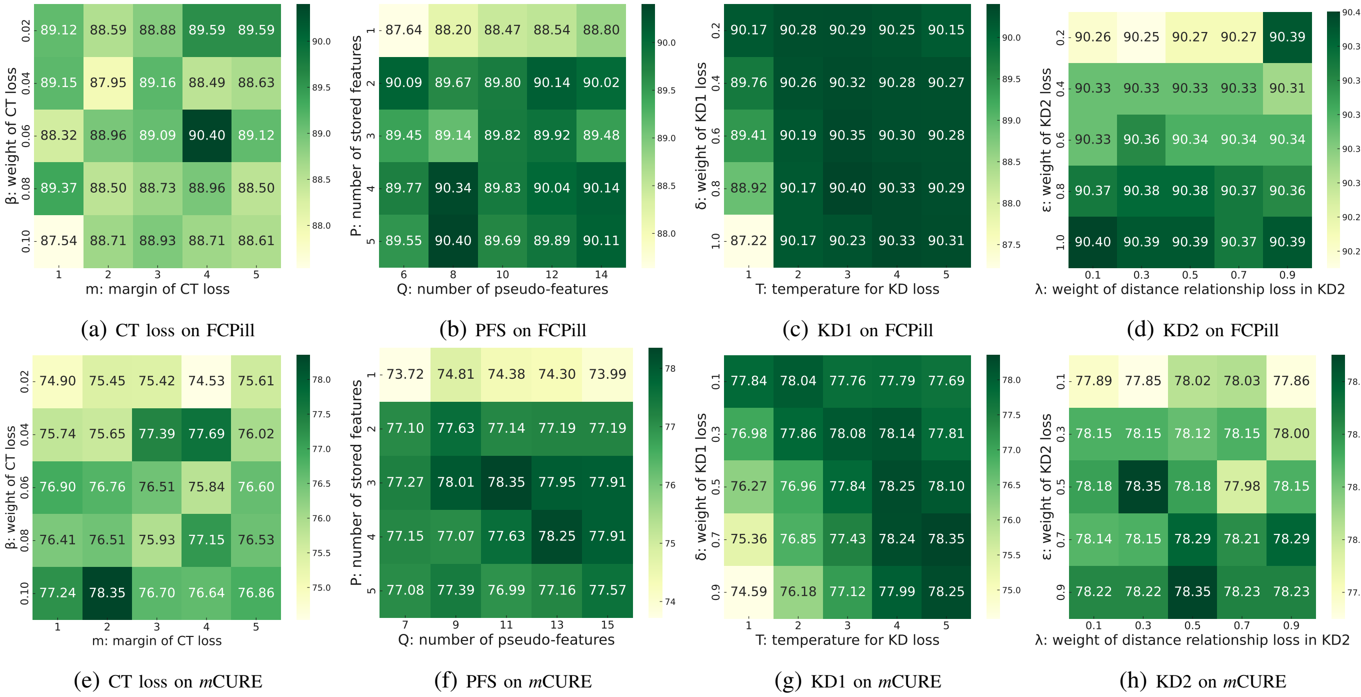
[12] J. Yu, Z. Chen, S.-i. Kamata, and et al., "Accurate system for automatic pill recognition using imprint information," *IET Image Processing*, vol. 9, no. 12, pp. 1039–1047, 2015.



(a) FCPill

(b) *m*CURE

Fig. 7: The influence of the weight of supervised contrastive loss on the performance of our method across sessions on the FCPill and *m*CURE datasets. It can be observed that for both the FCPill and *m*CURE datasets, the best performance is achieved when the weight is set to 0.04.



(a) CT loss on FCPill

(b) PFS on FCPill

(c) KD1 on FCPill

(d) KD2 on FCPill

(e) CT loss on *m*CURE

(f) PFS on *m*CURE

(g) KD1 on *m*CURE

(h) KD2 on *m*CURE

Fig. 8: The influence of hyper-parameters of each component on the performance of our method in the final session on the FCPill and *m*CURE datasets. It can be observed that on the FCPill dataset, the best performance is achieved with hyper-parameters $\{\beta, m, P, Q, \delta, T, \varepsilon, \lambda\} = \{0.06, 4, 5, 8, 0.8, 3, 1.0, 0.1\}$. On the *m*CURE dataset, the optimal performance is obtained with $\{\beta, m, P, Q, \delta, T, \varepsilon, \lambda\} = \{0.1, 2, 3, 11, 0.7, 5, 0.5, 0.3\}$.

[13] J. Zhang, L. Liu, K. Gao, and et al., "A forward and backward compatible framework for few-shot class-incremental pill recognition," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[14] M. Wang, Y. Jiang, B. Xu, and et al., "Mcir-yolo: White medication pill classification using multi-band infrared images," *IEEE Photonics Journal*, 2024.

[15] S. Kim, E.-Y. Park, J.-S. Kim, and et al., "Combination pattern method using deep learning for pill classification," *Applied Sciences*, vol. 14, no. 19, p. 9065, 2024.

[16] J. Heo, Y. Kang, S. Lee, and et al., "An accurate deep learning–based system for automatic pill identification: Model development and validation," *Journal of medical Internet research*, vol. 25, p. e41043, 2023.

[17] A. Cunha, T. Adão, and P. Trigueiros, "Helpmepills: A mobile pill recognition tool for elderly persons," *Procedia Technology*, vol. 16, pp. 1523–1532, 2014.

[18] Y.-Y. Ou, A.-C. Tsai, X.-P. Zhou, and et al., "Automatic drug pills detection based on enhanced feature pyramid network and convolution neural networks," *IET Computer Vision*, vol. 14, no. 1, pp. 9–17, 2020.

[19] H.-J. Kwon, H.-G. Kim, S.-W. Jung, and et al., "Deep learning and detection technique with least image-capturing for multiple pill dispensing inspection," *Journal of Sensors*, vol. 2022, 2022.

[20] S. Ling, A. Pastor, J. Li, and et al., "Few-shot pill recognition," in *Proceedings of CVPR*, 2020, pp. 9789–9798.

[21] T.-T. Nguyen, H. H. Pham, P. Le Nguyen, and et al., "Multi-stream fusion for class incremental learning in pill image classification," in *Proceedings of ACCV*, 2022, pp. 4565–4580.

[22] J. Gou, B. Yu, S. J. Maybank, and et al., "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021.

[23] Y.-B. Lee, U. Park, and A. K. Jain, "Pill-id: Matching and retrieval of drug pill imprint images," in *Proceedings of ICPR*. IEEE, 2010, pp.

2632–2635.

[24] X. Zeng, K. Cao, and M. Zhang, "Mobiledeeppill: A small-footprint mobile deep learning system for recognizing unconstrained pill images," in *Proceedings of MobiSys*, 2017, pp. 56–67.

[25] N. Pornbunruang, V. Tanjantuk, and T. Titijaroonroj, "Drugtionary: Drug pill image detection and recognition based on deep learning," in *Proceedings of ICCIT*. Springer, 2022, pp. 43–52.

[26] W.-J. Chang, L.-B. Chen, C.-H. Hsu, and et al., "A deep learning-based intelligent medicine recognition system for chronic patients," *IEEE Access*, vol. 7, pp. 44 441–44 458, 2019.

[27] M. M. Rahaman, E. K. Millar, and E. Meijering, "Breast cancer histopathology image-based gene expression prediction using spatial transcriptomics data and deep learning," *Scientific Reports*, vol. 13, no. 1, p. 13604, 2023.

[28] Y. Wang, Q. Yao, J. T. Kwok, and et al., "Generalizing from a few examples: A survey on few-shot learning," *ACM Computing Surveys*, vol. 53, no. 3, pp. 1–34, 2020.

[29] Y. He, T. Li, R. Ge, and et al., "Few-shot learning for deformable medical image registration with perception-correspondence decoupling and reverse teaching," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 3, pp. 1177–1187, 2021.

[30] H. Gharoun, F. Momenifar, F. Chen, and et al., "Meta-learning approaches for few-shot learning: A survey of recent advances," *ACM Computing Surveys*, vol. 56, no. 12, pp. 1–41, 2024.

[31] H. Quan, X. Li, D. Hu, and et al., "Dual-channel prototype network for few-shot pathology image classification," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[32] M. Fu, X. Wang, J. Wang, and et al., "Prototype bayesian meta-learning for few-shot image classification," *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[33] Y. Zhou, J. Hao, S. Huo, and et al., "Automatic metric search for few-shot learning," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[34] Y. Meng, M. Michalski, J. Huang, and et al., "Tuning language models as training data generators for augmentation-enhanced few-shot learning," in *Proceedings of ICML*. PMLR, 2023, pp. 24 457–24 477.

[35] D.-W. Zhou, Q.-W. Wang, Z.-H. Qi, and et al., "Class-incremental learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[36] H. Liu, Y. Zhou, B. Liu, and et al., "Incremental learning with neural networks for computer vision: a survey," *Artificial Intelligence Review*, vol. 56, no. 5, pp. 4557–4589, 2023.

[37] M. Masana, X. Liu, B. Twardowski, and et al., "Class-incremental learning: survey and performance evaluation on image classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 5, pp. 5513–5533, 2022.

[38] Q. Jodelet, X. Liu, Y. J. Phua, and et al., "Class-incremental learning using diffusion model for distillation and replay," in *Proceedings of ICCV*, 2023, pp. 3425–3433.

[39] Z. Hu, Y. Li, J. Lyu, and et al., "Dense network expansion for class incremental learning," in *Proceedings of CVPR*, 2023, pp. 11 858–11 867.

[40] H. Wen, L. Pan, Y. Dai, and et al., "Class incremental learning with multi-teacher distillation," in *Proceedings of CVPR*, 2024, pp. 28 443–28 452.

[41] J. Zhang, L. Liu, O. Silvén, M. Pietikäinen, and D. Hu, "Few-shot class-incremental learning for classification and object detection: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[42] J. Zhang, P. Zhao, Y. Zhao, and et al., "Few-shot class-incremental learning for retinal disease recognition," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[43] X. Tao, X. Hong, X. Chang, and et al., "Few-shot class-incremental learning," in *Proceedings of CVPR*, 2020, pp. 12 183–12 192.

[44] L. Sun, M. Zhang, B. Wang, and et al., "Few-shot class-incremental learning for medical time series classification," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 4, pp. 1872–1882, 2023.

[45] C. Zhang, N. Song, G. Lin, and et al., "Few-shot incremental learning with continually evolved classifiers," in *Proceedings of CVPR*, 2021, pp. 12 455–12 464.

[46] D.-W. Zhou, F.-Y. Wang, H.-J. Ye, and et al., "Forward compatible few-shot class-incremental learning," in *Proceedings of CVPR*, 2022, pp. 9046–9056.

[47] Y. Wen, K. Zhang, Z. Li, and et al., "A discriminative feature learning approach for deep face recognition," in *Proceedings of ECCV*. Springer, 2016, pp. 499–515.

[48] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of CVPR*, 2015, pp. 815–823.

[49] D.-W. Zhou, H.-J. Ye, L. Ma, and et al., "Few-shot class-incremental learning by sampling multi-phase tasks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[50] C. Peng, K. Zhao, T. Wang, and et al., "Few-shot class-incremental learning from an open-set perspective," in *Proceedings of ECCV*, 2022, pp. 382–397.

[51] Z. Pan, X. Yu, M. Zhang, and et al., "Ssfe-net: Self-supervised feature enhancement for ultra-fine-grained few-shot class incremental learning," in *Proceedings of WACV*, 2023, pp. 6275–6284.

[52] L. Zhao, J. Lu, Y. Xu, and et al., "Few-shot class-incremental learning via class-aware bilateral distillation," in *Proceedings of CVPR*, 2023, pp. 11 838–11 847.

[53] Z. Song, Y. Zhao, Y. Shi, and et al., "Learning with fantasy: Semantic-aware virtual contrastive constraint for few-shot class-incremental learning," in *Proceedings of CVPR*, 2023, pp. 24 183–24 192.