FoSSIL: A Unified Framework for Continual Semantic Segmentation in 2D and 3D Domains

Anonymous authors
Paper under double-blind review

Abstract

Evolving visual environments challenge continual semantic segmentation by introducing the complexities of *class-incremental learning*, *domain-incremental learning*, *limiting available annotations*, and necessitating the use of unlabeled data. In this work, we present the framework FoSSIL (**F**ew-sh**o**t **S**emantic **S**egmentation for **I**ncremental **L**earning), which extensively benchmarks continual semantic segmentation, spanning both 2D natural scenes and 3D medical volumes. Our evaluation encompasses diverse and realistic settings, leveraging both labeled (few-shot) and unlabeled data. Building on this benchmark, we introduce *guided noise injection* to mitigate overfitting due to novel few-shot classes from various domains. Furthermore, we leverage semi-supervised learning for unlabeled data to augment few-shot novel classes. We propose a *filtering* mechanism to remove highly confident but incorrectly predicted pseudo-labels, further improving performance. Results across class-incremental, few-shot, and domain-incremental scenarios with unlabeled data validate our strategies for robust semantic segmentation in complex, evolving settings, highlighting both the effectiveness and generality of our approach. Our findings illustrate that the proposed framework forms a simple yet powerful recipe for continual semantic segmentation in dynamic real-world environments. Our large-scale benchmarking across natural 2D and medical 3D domains exposes key failure modes of existing methods and offers a roadmap for building robust continual segmentation models.

1 Introduction

"I hear and I forget. I see and I remember. I do and I understand." – Confucius

Figure 1: A major challenge in continual learning where the models must segment the objects (e.g., a car or organ) as they appear from different domains over time with few-shot data.

The pursuit of truly intelligent systems necessitates continuous learning and adaptation in open-world environments. While continual learning (CL) (Wang et al. (2024); Yuan & Zhao (2024)) has advanced across various tasks, a critical gap remains in addressing the significant complexities of real-world dense prediction tasks like semantic segmentation. In demanding applications such as autonomous driving and medical image analysis, semantic segmentation models are confronted with a continuous influx of data characterized by both novel semantic categories or *class-incremental learning* (CIL Zhou et al. (2024)) and evolving data distributions or *domain-incremental learning* (DIL Mirza et al. (2022)), posing a formidable challenge to their adaptability and robustness (Figure 1).

This discrepancy between idealized CL scenarios and real-world semantic understanding poses significant challenges. In the realistic setting of continual learning, prevalent CL methods struggle with catastrophic forgetting, significantly exacerbated by shifts in the semantic label space and input data characteristics. Furthermore, data scarcity necessitates effective *few-shot learning* (Tao et al.

(2020); Qiu et al. (2023); Tian et al. (2024)) within these continuous learning streams. The confluence of these factors creates a challenging landscape where models must rapidly adapt to new concepts with limited supervision while preserving previously acquired knowledge. Specifically, the need to balance plasticity (acquiring new knowledge) and stability (retaining old knowledge) becomes paramount, yet exceedingly difficult, under these conditions. Figure 2 systematically evaluates different combinations of constraints, such as continual learning with a varying set of classes (CIL) and changing data distributions (DIL), and scarcely labeled data (few-shot learning), highlighting the individual and combined impact of these challenges on model performance, further emphasising the importance of addressing this problem comprehensively.

Leveraging unlabeled data through *semi-supervised learning* (Kang et al. (2023b); Cui et al. (2024)) holds immense potential. However, the dynamic introduction of novel classes complicates the reliable utilization of pseudo-labels, as initial model biases can lead to the propagation of incorrect information. A key challenge here is that this initial bias can compound over time, progressively degrading the model's ability to learn effectively.

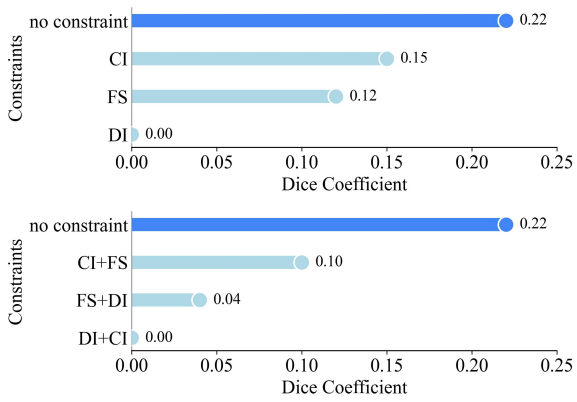

Figure 2: Class-incremental (CI), few-shot (FS), and domain-incremental (DI) constraints all lead to significantly reduced Dice scores compared to the unconstrained baseline ("no constraint") on a 3D U-Net model.

Despite its practical significance, the realistic setting of few-shot learning and semi-supervised learning for complex tasks like semantic segmentation remains largely under underexplored in the context of continual learning. Existing CL methods, often evaluated on simpler tasks, are not designed to handle these complexities. The core challenge lies in developing a learning framework that can effectively handle the data distributions that change over time and the need for rapid adaptation to new classes with limited labeled data, while simultaneously mitigating catastrophic forgetting and the propagation of errors from noisy pseudo-labels.

This work directly confronts these critical, largely unaddressed challenges in continual learning for semantic segmentation. Our proposed framework **FoSSIL** (**F**ew-sh**o**t **S**emantic **S**egmentation for **I**ncremental **L**earning) investigates continual learning in realistic and demanding scenarios where semantic classes and data domains evolve and may *reappear* over time, constrained by few-shot data within each learning session. Crucially, FoSSIL models the real-world occurrence where a previously seen class or domain can reappear, with the constraint that any given incremental session introduces novelty in either the class set or the data domain, but not concurrently. In contrast to prevailing approaches, FoSSIL addresses the intertwined challenges of evolving classes and domains under data scarcity by employing an exemplar-free prototype replay (Chen et al. (2023)) with a novel *guided noise injection* scheme and *refinement* of pseudo-labels in a semi-supervised setting that takes advantage of widely accessible unlabeled data. This enables effective learning, robustness, and knowledge retention, achieving strong generalization across diverse semantic segmentation architectures.

To ground our contributions, we conduct extensive benchmarking that spans both 2D natural and 3D medical domains, systematically evaluating class-incremental, domain-incremental, few-shot, and semi-supervised continual learning settings. FoSSIL benchmarks a wide spectrum (around *twenty-five*) of state-of-the-art methods with detailed ablations on *nine* datasets with different *backbones* like U-Net (Ronneberger et al. (2015)), DeepLabv3+ (Chen et al. (2018)) and Transformers-based (Kirillov et al. (2023)). It highlights key failure modes in current approaches, such as overfitting in few-shot regimes, difficulties in adapting across domains, and error amplification from pseudo-labels, and proposes novel strategies to overcome them. We find that fine-tuning popular backbones (e.g., U-Net, DeepLabv3+, MedFormer Gao et al. (2022)) on novel few-shot classes from varied domains, whether with partially frozen or fully unfrozen weights, leads to severe performance drops in incremental sessions, highlighting the severity of the problem (Figure 3).

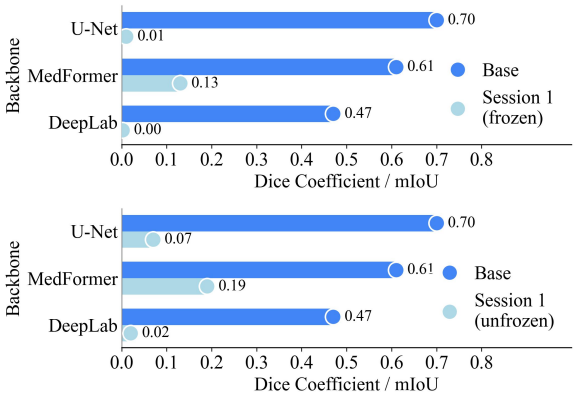

Figure 3: Performance of various backbones on the FoSSIL benchmark with partially frozen or fully unfrozen weights. Session 1 denotes the first incremental session following the common base session.

Our key contributions to multi-constrained continual learning for semantic segmentation are fourfold: **(i) Firstly,** we provide an extensive benchmark of continual learning for semantic segmentation through the FoSSIL framework, incorporating multiple realistic constraints across nine datasets from both 2D natural and 3D medical domains. We *re-implemented* and *adapted* around twenty-five closely related methods with open-source implementation to run on our proposed benchmark. This evaluation exposes limitations in current approaches. **(ii) Secondly,** we employ an exemplar-free prototype replay strategy for continual learning in both class- and domain-incremental settings with few-shot data, improving memory efficiency and privacy by avoiding the storage of raw samples. FoSSIL's novelty further lies in integrating a guided noise injection strategy, which regularizes the model and enhances generalization across novel classes from multiple domains with few-shot data. **(ii) Thirdly,** We leverage *semi-supervised* learning to augment few-shot classes using readily available unlabeled data. Here, the novelty lies in employing learned prototypes to filter highly confident but incorrectly predicted regions within pseudo-labels, thereby enhancing the quality of supervision obtained from unlabeled data. **(iii) Finally,** our innovative approaches demonstrate strong generalization across diverse semantic segmentation architectures (3D U-Net, DeepLabv3+, Transformers), even outperforming models pre-trained on large-scale datasets like SAM (Kirillov et al. (2023); Kerssies et al. (2024)) on the FoSSIL benchmark.

2 Related Work

Class-Incremental/Domain-Incremental Learning in Semantic Segmentation: MiB Cermelli et al. (2020) pioneered incremental semantic segmentation by addressing background shift with distillation losses and classifier initialization. CLIP-CT Zhang et al. (2023) builds on this by using pseudo-labeling and CLIP-guided Radford et al. (2021) organ-specific heads for efficient adaptation to new classes. MDIL Garg et al. (2022) addresses semantic segmentation across domains using a dynamic architecture that factorizes parameters into domain-invariant and domain-specific components. It employs domain-aware residual units, domain-specific normalization, and adaptive distillation to balance stability and plasticity.

Few-shot Class-Incremental Learning: Cermelli et al. (2021) proposed Prototype-based Incremental Few-Shot Segmentation (PIFS), which integrates prototype learning with knowledge distillation to learn new classes from few samples without access to old training data. Subspace regularization (Subspace Akyürek et al. (2021)) mitigates catastrophic forgetting and overfitting by constraining novel class weights to the subspace spanned by base class weights, optionally incorporating semantic information from class names. Hersche et al. (2022) introduced C-FSCIL (Constrained Few-shot Class-Incremental Learning), which leverages hyperdimensional computing with a frozen meta-learned feature extractor, a trainable fully connected layer, and a dynamically growing memory of quasi-orthogonal prototypes. FACT (Zhou et al. (2022)) reserves embedding space for future classes via virtual prototypes and uses manifold mixup to forecast novel classes. Liu et al. (2022) presented an entropy-regularized data-free replay (Gen-Replay) method that synthesizes uncertain samples from previous classes without storing real data. Yang et al. (2023) proposed NC-FSCIL, a neural collapse–inspired framework that fixes classifier prototypes as a simplex equiangular tight frame. Qiu et al. (2023) introduced GAPS, a model-agnostic framework for few-shot incremental semantic segmentation that addresses partial annotations by generating fully labeled data via copy-paste synthesis. SoftNet (Kang et al. (2023a)) is inspired by the Regularized Lottery Ticket Hypothesis, which uses adaptive soft masks to decompose a network into major and minor subnetworks. The major subnetwork mitigates forgetting, while the minor subnetwork adapts

to novel classes with limited overfitting through joint weight–mask optimization. Jiang et al. (2023) presented FSCIL-SS that combines pseudo-labeling with knowledge distillation to learn novel classes from limited examples while retaining existing knowledge. FeCAM (Goswami et al. (2023)), an exemplar-free method leverages class-specific covariance matrices and the Mahalanobis distance to improve prototype-based classification.

Semi-Supervised Learning based approaches: Killamsetty et al. (2021) proposed RETRIEVE, a coreset selection framework for efficient and *robust* semi-supervised learning. NNCSL (Kang et al. (2023b)) is a soft nearest-neighbor framework for continual semi-supervised learning, tackling catastrophic forgetting on unlabeled representations and overfitting on limited labeled samples. UaD-CE (Cui et al. (2024)), an uncertainty-aware distillation framework with class equilibrium for semi-supervised learning that balances pseudo-label generation, while the uncertainty-aware distillation module selects reliable exemplars for adaptive knowledge transfer, mitigating both overfitting and forgetting.

Miscellaneous: Khosla et al. (2020) introduced a supervised *contrastive learning* (SupCL) framework that extends *self-supervised* contrastive approaches to leverage label information for *improved representation learning*. SimCLR (Chen et al. (2020)) is a contrastive learning framework that learns visual representations by maximizing agreement between differently augmented views of the same image, without requiring labels (UnSupCL). Robinson et al. (2021) proposed a hard negative sampling framework (UnSupCL-HNM) for contrastive learning that addresses the challenge of selecting informative negative samples without supervision. Wang et al. (2021) bridged multi-task learning (MTL) and gradient-based meta-learning by showing that both share the same optimization formulation via joint training and regularized bi-level optimization. Bouniot et al. (2022) analyzed *few-shot learning* through multi-task representation theory, highlighting differences between gradient-based (e.g., MAML) and metric-based meta-learning methods in satisfying optimal predictor assumptions. Liu et al. (2023) proposed a CLIP-driven *universal* model for *multi-organ* segmentation and tumor detection, addressing *partial label* problems by incorporating CLIP Radford et al. (2021) text embeddings to capture semantic relationships. Franco et al. (2024) proposed HALO, a hyperbolic neural network approach for pixel-level active learning in semantic segmentation under *domain shift*.

To the best of our knowledge, no existing method jointly addresses CIL, DIL, and few-shot learning for semantic segmentation. The proposed FoSSIL framework addresses this limitation and leverages unlabeled data to augment scarce few-shot classes.

3 FoSSIL Framework

We formalize the multi-constrained continual learning problem for semantic segmentation as a sequence of sessions $\mathcal{S} = \{\mathcal{S}_0, \mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_T\}$ where each session $\mathcal{S}_t$ is characterized by both a semantic class space $\mathcal{C}_t$ and a domain distribution $\mathcal{D}_t$. At each session $\mathcal{S}_t$, a learning model encounters a dataset $\mathbb{D}_t = \{(x_i, y_i)\}_{i=1}^{N_t}$ where $x_i \in \mathcal{X}$ represents input images drawn from the domain $\mathcal{D}_t$ and $y_i \in \mathcal{Y}_t \subseteq \mathcal{C}_t$ denotes the pixel-wise semantic labels, with $|\mathbb{D}_t| = N_t$ being the number of available samples in session t. $\mathcal{S}_0$ denotes the base session, which contains domain(s) with abundant labeled data (N_0), while the remaining sessions are few-shot sessions with sparsely labeled domains.

3.1 Classes, Domains, and Data Across Sessions

The continual learning sequence accommodates three realistic scenarios that distinguish our framework from idealized settings. **Same Classes, Different Domains:** Different sessions may share semantic classes while those classes belong to different domains: $\mathcal{C}_t = \mathcal{C}_{t'}$ and $\mathcal{D}_t \neq \mathcal{D}_{t'}$. **Different Classes, Same Domain:** New classes may be introduced across sessions belonging to the same domain: $\mathcal{C}_t \neq \mathcal{C}_{t'}$ and $\mathcal{D}_t = \mathcal{D}_{t'}$. **Different Classes, Different Domain:** The most challenging scenario with both semantic and domain shift: $\mathcal{C}_t \neq \mathcal{C}_{t'}$ and $\mathcal{D}_t \neq \mathcal{D}_{t'}$. Importantly, we explicitly exclude the case where same classes of the same domain are repeated across sessions.

Each incremental session ($t \neq 0$) is trained on few-shot labeled data, where $N_t \ll N_0$ and $N_t = K \cdot |\mathcal{C}_t|$. Here, K denotes the number of labeled examples per class (typically $K \in \{5, 10, 20, 30\}$). In addition, each session may also access unlabeled data, defined as $\mathcal{U}_t = \{x_j^{(u)}\}_{j=1}^{M_t}$, where $M_t \gg N_t$ indicates the number of unlabeled samples drawn from the same domain $\mathcal{D}_t$.

3.2 EXEMPLAR FREE PROTOTYPE REPLAY

For each class $c \in \mathcal{C}_t$, we extract a compact prototype from the feature space. Given a trained model with intermediate feature extractor ϕ. For each sample (x_i, y_i), embedding $E_i = \phi(x_i) \in \mathbb{R}^{D \times H \times W}$.

In session t, for class c, we extract features corresponding to class pixels as $\mathcal{F}_c^{(i)} = \{E_i : y_i = c\}$.

The class prototype is given as,

$$p_c^{(i)} = \frac{1}{|\mathcal{F}_c^{(i)}|} \sum_{\mathbf{f} \in \mathcal{F}_c^{(i)}} \mathbf{f} \tag{1}$$

Across all samples in session t containing class c:

$$P_c = \frac{1}{NS_c} \sum_{j \in NS_c} \frac{p_c^{(j)}}{||p_c^{(j)}||_2} \tag{2}$$

where $NS_c \subseteq N_t$ is the number of samples containing class c.

In session $t+1$, we replay all prototypes learned from previous session, $\mathcal{P}_t = \{P_c\}$ for all $c \in \mathcal{C}_t$ as,

$$\mathcal{L}_{\text{proto}} = \sum_{P_c \in \mathcal{P}_t} \mathcal{L}_{\text{CE}}(F(P_c), c) \tag{3}$$

where F is the final classifier layer and $\mathcal{L}_{\text{CE}}$ is the cross-entropy loss. $\mathcal{L}_{\text{proto}}$ optimizes the model parameters in session $t+1$.

3.3 GUIDED NOISE INJECTION

The guided noise injection mechanism regulates noise using parameter gradients, which serve as a proxy for determining the appropriate magnitude of noise to add to each parameter.

The method maintains a gradient buffer G that accumulates squared gradients $(\nabla_{w_{ij}} \mathcal{L})^2$ where $\mathcal{L}$ is the loss function and w_{ij} are the weight parameters of the classifier layer F with weight matrix $\mathbb{W}$.

For any $G_{ij} \in G$ the inverse is computed as $G_{ij}^{-1} = \frac{1}{G_{ij}+\epsilon}$ where $\epsilon = 10^{-8}$ ensures numerical stability. To control the noise magnitude, the inverse gradients are normalized to a bounded range:

$$\tilde{G}_{ij}^{-1} = \frac{1 + G_{ij}^{-1} - \min(G^{-1})}{1 + \max(G^{-1}) - \min(G^{-1})} \tag{4}$$

The weights $\mathbb{W}$ are perturbed as:

$$\tilde{\mathbb{W}} = \mathbb{W} + \tilde{G}^{-1} \odot \mathcal{N}(0, I) \tag{5}$$

Hence the noise $\mathcal{N}(0, I)$ added to $\mathbb{W}$ is guided by $\tilde{G}^{-1}$.

Large gradients correspond to low noise injection, whereas small gradients allow for higher noise injection. Critical weight parameters with large gradients, which are actively contributing to learning, receive minimal noise injection, whereas parameters that have begun overfitting and no longer contribute significantly are injected with higher noise for regularization.

3.4 PROTOTYPE-GUIDED PSEUDO-LABEL REFINEMENT

To leverage abundant unlabeled data while mitigating the risk of noise and error propagation, we introduce a prototype-guided pseudo-label refinement (or filtering) strategy within a mean-teacher (Tarvainen & Valpola (2017)) based framework. To mitigate confirmation bias in standard pseudo-labeling, we introduce a mechanism that validates pseudo-labels through both predictive confidence and feature-space consistency with prototypes.

For an unlabeled input $x_j^{(u)}$, the student network M_s and teacher network M_t generate pseudo-label predictions, denoted by $\hat{y}_s$ and $\hat{y}_t$, respectively:

$$\hat{y}_s, \mathcal{F}_s = M_s(x_j^{(u)}), \quad \hat{y}_t, \mathcal{F}_t = M_t(x_j^{(u)}) \tag{6}$$

where $\mathcal{F}_s$ and $\mathcal{F}_t$ are features representations learned by student and teacher, respectively.

We compute the confidence of the predictions ($\hat{y}_s$ and $\hat{y}_t$) as,

$$c'(p,q) = \arg\max_c(\text{softmax}(\hat{y}(p,q))), \quad \text{conf}(p,q) = \max(\text{softmax}(\hat{y}(p,q))) \tag{7}$$

where $c'(p,q)$ is class for pixel (p,q), $\text{conf}(p,q)$ is confidence of class $c'(p,q)$ and $\hat{y}(p,q)$ is pseudo-label at pixel (p,q).

To validate predictions, the cosine similarity between features and prototypes is computed as,

$$\text{sim}(p,q) = \frac{\mathcal{F}(p,q) \cdot P_{c'(p,q)}}{||\mathcal{F}(p,q)||_2 ||P_{c'(p,q)}||_2} \tag{8}$$

where $\mathcal{F}(p,q)$ is feature of the pixel (p,q) and $P_{c'(p,q)}$ is prototype corresponding to the predicted class $c'(p,q)$.

A pseudo-label at (p,q) is retained only if it satisfies the following conditions:

$$\text{valid}(p,q) = (\text{conf}(p,q) > \tau_{\text{conf}}) \quad \text{and} \quad (\text{sim}(p,q) > \tau_{\text{sim}}) \tag{9}$$

where τ_{conf} and τ_{sim} are empirically determined thresholds.

The consistency loss in mean-teacher operates only on validated pseudo-labels:

$$\mathcal{L}_{\text{consistency}} = \frac{1}{|\mathcal{V}|} \sum_{(p,q)\in\mathcal{V}} ||\hat{y}_s(p,q) - \hat{y}_t(p,q)||_2^2 \tag{10}$$

where $\mathcal{V} = \{(p,q) : \text{valid}_s(p,q) \quad \text{and} \quad \text{valid}_t(p,q)\}$ represents pixels validated by both models.

This verification mechanism reduces pseudo-label noise by requiring both high prediction confidence and feature-space similarity to class prototypes. This ensures the pseudo-labels are robust to domain shifts, preventing the amplification of errors in the student-teacher feedback loop.

Table 1: Summary of FoSSIL Benchmarks. $|\mathcal{C}_t|$ denotes the number of classes in session i. 'SS' denotes Semi-Supervised.

Benchmark	Session 0 (Base)	Session 1	Session 2	Session 3	Session 4	Session 5
Med FoSSIL-Disjoint	$\|\mathcal{C}_0\| = 15$ (TS)	$\|\mathcal{C}_1\| = 5$ (AMOS)	$\|\mathcal{C}_2\| = 6$ (BCV)	$\|\mathcal{C}_3\| = 4$ (MOTS)	$\|\mathcal{C}_4\| = 3$ (BraTS)	$\|\mathcal{C}_5\| = 4$ (VerSe)
Med FoSSIL-Mixed	$\|\mathcal{C}_0\| = 10$ (AMOS)	$\|\mathcal{C}_1\| = 8$ (BCV, MOTS)	$\|\mathcal{C}_2\| = 6$ (TS, AMOS)	$\|\mathcal{C}_3\| = 4$ (MOTS, TS)	$\|\mathcal{C}_4\| = 7$ (Brats, VerSe)	–
Med SS-FoSSIL	$\|\mathcal{C}_0\| = 15$ (TS)	$\|\mathcal{C}_1\| = 5$ (AMOS)	$\|\mathcal{C}_2\| = 6$ (BCV)	$\|\mathcal{C}_3\| = 4$ (MOTS)	$\|\mathcal{C}_4\| = 3$ (BraTS)	$\|\mathcal{C}_5\| = 4$ (VerSe)
Natural-FoSSIL	$\|\mathcal{C}_0\| = 10$ (BDD)	$\|\mathcal{C}_1\| = 5$ (IDD)	$\|\mathcal{C}_2\| = 5$ (BDD, IDD)	–	–	–
SS-Natural-FoSSIL	$\|\mathcal{C}_0\| = 10$ (BDD)	$\|\mathcal{C}_1\| = 2$ (Cityscapes)	$\|\mathcal{C}_2\| = 2$ (IDD)	$\|\mathcal{C}_3\| = 3$ (IDD)	–	–

4 FoSSIL Benchmarks

We construct five challenging benchmarks for 3D medical and 2D natural scene segmentation, designed to simulate realistic clinical and autonomous driving scenarios with *multiple sessions*, *diverse domains*, and a *large number of novel classes*. Each benchmark features a base learning session on a large dataset followed by incremental sessions with limited labeled data (few-shot) and with significant domain shifts.

3D Medical FoSSIL Benchmarks: We develop three 3D medical benchmarks using data from **TotalSegmentator** (TS) (Wasserthal et al. (2023)), **AMOS** (Ji et al. (2022)), **BCV** (Landman et al. (2015), **MOTS** (Zhang et al. (2021)), **BraTS** (Menze et al. (2014)), and **VerSe** (Sekuboyina et al. (2021)). All three benchmarks adopt a few-shot learning setup, using 5 training samples per class for incremental sessions, progressing from normal to tumor segmentation.

Table 2: Performance of baselines on Med FoSSIL-Disjoint benchmark (3-sessions). Results reported as Dice coefficients (0-1).

Method	Session 0	Session 1	Session 2
PIFS Cermelli et al. (2021)	0.700	0.129	0.078
NC-FSCIL Yang et al. (2023)	0.394	0.077	0.081
CLIP-CT Zhang et al. (2023)	0.475	0.186	0.141
MiB Cermelli et al. (2020)	0.700	0.271	0.096
MDIL Garg et al. (2022)	0.779	0.115	0.097
C-FSCIL Hersche et al. (2022)	0.787	0.334	0.297
SoftNet Kang et al. (2023a)	0.820	0.305	0.146
GAPS Qiu et al. (2023)	0.700	0.334	0.253
FSCIL - SS Jiang et al. (2023)	0.700	0.115	0.089
Subspace Akyürek et al. (2021)	0.257	0.054	0.040
Gen-Replay Liu et al. (2022)	0.700	0.076	0.102
FeCAM Goswami et al. (2023)	0.700	0.048	0.042
FACT Zhou et al. (2022)	0.357	0.071	0.0278
MAML Bouniot et al. (2022)	0.700	0.0006	0.059
MAML + regularizer Bouniot et al. (2022)	0.700	0.001	0.062
MTL Wang et al. (2021)	0.700	0.079	0.0880
UnSupCL Chen et al. (2020)	0.700	0.039	0.0882
SupCL Khosla et al. (2020)	0.700	0.058	0.0421
UnSupCL-HNM Robinson et al. (2021)	0.700	0.035	0.0676
FoSSIL (U-Net)	0.736	**0.460**	**0.398**

Table 3: Performance on Med FoSSIL-Disjoint benchmark (6-sessions). All values are reported as Dice coefficients (0-1).

Method	Session 0	Session 1	Session 2	Session 3	Session 4	Session 5
U-Net Vanilla	0.700	0.076	0.057	0.047	**0.030**	0.042
FoSSIL (U-Net)	0.736	**0.460**	**0.398**	**0.329**	0.025	**0.324**

The three medical benchmarks: (i) **Med FoSSIL-Disjoint**, a 6-session, 37-class protocol with disjoint classes and domains; (ii) **Med FoSSIL-Mixed**, a 5-session, 35-class setup allowing recurrence of either classes or domains (but not both) and mixing datasets per session; and (iii) **Med Semi-Supervised-FoSSIL**, a semi-supervised variant of Med FoSSIL-Disjoint augmented with 8–30 unlabeled samples per session. Please refer to Table 1 for various classes and domains.

2D Natural Scene FoSSIL Benchmarks: We introduce two benchmarks for autonomous driving scenarios using data from **BDD100K** (Yu et al. (2020)), **Cityscapes** (Cordts et al. (2016)), and **IDD** (Varma et al. (2019)). These benchmarks feature a few-shot learning with 10 training samples per class. The two natural scene benchmarks for autonomous driving: (i) **Natural-FoSSIL**, a 3-session setup using BDD100K, Cityscapes, and IDD, designed to test representation adaptation under domain shifts and class recurrence; and (ii) **Semi-Supervised Natural-FoSSIL**, a 4-session variant that augments new classes with 400 unlabeled images per class to reflect realistic scenarios with limited annotations but abundant raw data. Our code is available at https://github.com/anony34/FoSSIL. Please refer to the Appendix for details.

Table 4: Performance on Natural-FoSSIL benchmark. All values are reported as mIoU (0-100).

Method	Session 0	Session 1	Session 2
DeepLab Vanilla	47.76	2.18	3.86
GAPS Qiu et al. (2023)	47.76	23.42	16.68
MiB Cermelli et al. (2020)	47.76	2.50	2.37
MDIL Garg et al. (2022)	48.54	1.59	3.02
SAM Vanilla Kerssies et al. (2024)	66.0	32.6	30.81
FoSSIL (SAM)	66.0	**33.2**	**31.22**

Table 5: Performance on Med FoSSIL-Mixed benchmark. All values are reported as Dice coefficients (0-1).

Method	Session 0	Session 1	Session 2	Session 3	Session 4
U-Net Vanilla	0.571	0.216	0.133	0.074	0.045
CLIP-driven Liu et al. (2023)	0.717	**0.417**	0.227	0.196	0.089
MedFormer Vanilla	0.613	0.198	0.134	0.052	0.067
SwinUNetr Vanilla	0.605	0.197	0.133	0.082	0.082
FoSSIL (SwinUNetr)	0.605	0.318	0.275	0.254	0.210
FoSSIL (MedFormer)	0.622	0.367	**0.287**	**0.288**	**0.228**

Table 6: Performance on Semi-Supervised Natural-FoSSIL benchmark. All values are reported as mIoU (0-100).

Method	Session 0	Session 1	Session 2	Session 3
DeepLab Vanilla	47.76	1.04	1.51	0.43
MDIL Garg et al. (2022)	47.76	1.87	1.43	0.39
MiB Cermelli et al. (2020)	47.76	5.97	1.59	0.42
UaD-CE Cui et al. (2024)	47.76	1.88	1.74	0.69
NNCSL Kang et al. (2023b)	47.76	0.79	1.27	0.46
HALO Franco et al. (2024)	47.76	1.78	2.02	1.27
RETRIEVE Killamsetty et al. (2021)	47.76	1.57	1.89	0.39
GAPS Qiu et al. (2023)	47.76	19.73	18.76	14.45
FoSSIL + GAPS	47.76	**27.84**	**27.69**	**25.47**

5 Results and Analysis

We use mean Intersection over Union (mIoU), ranging from 0 to 100, for 2D natural scene benchmarks, and the Dice coefficient (Dice score), ranging from 0 to 1, for 3D medical benchmarks, as evaluation metrics. In each *incremental session*, we evaluate the classes introduced in the current session along with all classes encountered in previous sessions. The goal is to retain previously learned knowledge while effectively acquiring new information, handling data scarcity, and adapting to domain shifts. A *Vanilla* baseline consists of a plain backbone with no mechanisms to handle any constraints. Gen-Replay Liu et al. (2022) is implemented with a diffusion model adapted from Dorjsembe et al. (2024).

Across the medical benchmarks, baselines collapse after just two sessions in *Med FoSSIL-Disjoint* (Table 2), while FoSSIL with a U-Net backbone sustains strong performance across all five, demonstrating robustness to multiple constraints as shown in Table 2 and Table 3. In *Med FoSSIL-Mixed*, transformer backbones such as MedFormer, SwinUNetr, and a CLIP-driven U-Net all degrade over sessions, with the latter dropping despite pretraining on 21 of 35 classes (Table 5), highlighting the benchmark's difficulty. In *Med Semi-Supervised-FoSSIL*, adding unlabeled data significantly boosts FoSSIL (Table 7, Figure 4b), unlike existing semi-supervised methods that fail to exploit it. **This demonstrates that leveraging readily available unlabeled data can substantially improve multi-constraint continual learning for semantic segmentation**. In natural scene benchmark (*Natural-FoSSIL*), even large-scale models like SAM Kirillov et al. (2023), pretrained on

Table 7: Performance on Med Semi-Supervised-FoSSIL benchmark. Results reported as Dice coefficients (0-1).

Method	Session 0	Session 1	Session 2	Session 3	Session 4	Session 5
U-Net Vanilla	0.700	0.076	0.0578	0.0472	0.0302	0.0429
NNCSL (U-Net) Kang et al. (2023b)	0.700	0.048	0.0477	0.030	0.011	0.0404
UaD-CE (U-Net) Cui et al. (2024)	0.700	0.082	0.0750	0.0670	0.0313	0.0487
FoSSIL (U-Net)	0.736	**0.554**	**0.4449**	**0.414**	**0.0576**	**0.368**
MedFormer Vanilla	0.659	0.065	0.062	0.059	0.051	0.040
UaD-CE (MedFormer) Cui et al. (2024)	0.659	0.052	0.0479	0.0646	0.0369	0.0323
NNCSL (MedFormer) Kang et al. (2023b)	0.659	0.142	0.0955	0.144	0.010	0.048
FoSSIL (MedFormer)	0.640	**0.431**	**0.368**	**0.335**	**0.323**	**0.293**

a billion masks, exhibit forgetting (Table 4), yet FoSSIL consistently improves SAM, as well as U-Net and transformer backbones, showing broad applicability. Finally, in *Semi-Supervised Natural-FoSSIL*, FoSSIL serves as a plug-and-play module that leverages unlabeled data to enhance GAPS results (Table 6), further underscoring its effectiveness across diverse baselines.

Ablations: In the Med FoSSIL-Mixed benchmark, where FoSSIL improves the performance of MedFormer, we removed guided noise injection ($\tilde{G}^{-1}$) from Equation 5, and the results are plotted in Figure 5a. As shown, there is a significant drop in performance, highlighting the importance of the proposed guided noise injection strategy, which helps regularize the model under few-shot data and domain shifts. In the Semi-Supervised Natural-FoSSIL benchmark, where FoSSIL improves GAPS using unlabeled data, we removed the pseudo-label refinement strategy and evaluated FoSSIL's performance, as shown in Figure 5b. It is evident that the refinement strategy contributes to the improved performance of FoSSIL.

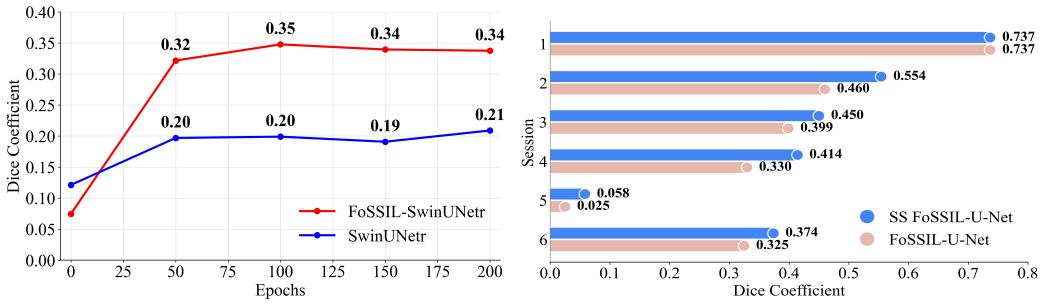

Figure 4: a) Variation of performance of SwinUnetr Vanilla and with FoSSIL over the epochs. This illustrates how FoSSIL sustains performance across epochs. b) Performance of FoSSIL without unlabeled data (Med FoSSIL-Disjoint) and with unlabeled data (Med Semi-Supervised-FoSSIL). 'SS' is Semi-Supervised.

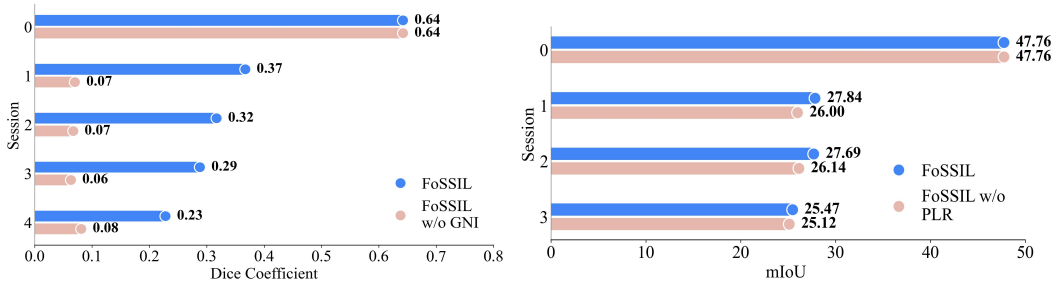

Figure 5: a) FoSSIL without Guided Noise Injection (GNI) evaluated with MedFormer (Med FoSSIL-Mixed). b) FoSSIL without Pseudo-Label Refinement (PLR) on Semi-Supervised Natural-FoSSIL benchmark.

6 Conclusion

We evaluated existing continual learning methods—including class-incremental, domain-incremental, few-shot, and semi-supervised approaches—against the proposed benchmarks, which reveal a substantial performance gap that current methods have yet to close. This underscores the urgent need for robust methods capable of handling multiple constraints in continual semantic segmentation, as even large pre-trained and foundational models exhibit performance degradation. Our proposed framework, FoSSIL, mitigates performance drop across sessions, demonstrating the effectiveness of guided noise injection and pseudo-label refinement strategies. It clearly demonstrates that using readily available unlabeled data can significantly improve multi-constraint continual learning for semantic segmentation. In the future, FoSSIL will be extended to more challenging settings, such as continual learning with open-vocabulary, detection, and other dense prediction tasks that remain largely unexplored. We will also evaluate and refine other large foundational and vision–language models to assess their performance on the FoSSIL benchmark.

REFERENCES

Afra Feyza Akyürek, Ekin Akyürek, Derry Tanti Wijaya, and Jacob Andreas. Subspace regularizers for few-shot class incremental learning. *arXiv preprint arXiv:2110.07059*, 2021.

Quentin Bouniot, Ievgen Redko, Romaric Audigier, Angélique Loesch, and Amaury Habrard. Improving few-shot learning through multi-task representation learning theory. In *European Conference on Computer Vision*, 2022.

Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulò, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Fabio Cermelli, Massimiliano Mancini, Yongqin Xian, Zeynep Akata, and Barbara Caputo. Prototype-based incremental few-shot semantic segmentation. In *Proceedings of the 32nd British Machine Vision Conference*, November 2021.

Jinpeng Chen, Runmin Cong, Yuxuan Luo, Horace Ip, and Sam Kwong. Saving 100x storage: Prototype replay for reconstructing training sample distribution in class-incremental semantic segmentation. *Advances in Neural Information Processing Systems*, 36:35988–35999, 2023.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (eds.), *Computer Vision – ECCV 2018*, pp. 833–851, Cham, 2018. Springer International Publishing. ISBN 978-3-030-01234-2.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Yawen Cui, Wanxia Deng, Haoyu Chen, and Li Liu. Uncertainty-aware distillation for semi-supervised few-shot class-incremental learning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(10):14259–14272, 2024. doi: 10.1109/TNNLS.2023.3277018.

Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4084–4093, 2024. doi: 10.1109/JBHI.2024.3385504.

Luca Franco, Paolo Mandica, Konstantinos Kallidromitis, Devin Guillory, Yu-Teng Li, Trevor Darrell, and Fabio Galasso. Hyperbolic active learning for semantic segmentation under domain shift. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=hKdJPMQvew.

Yunhe Gao, Mu Zhou, Di Liu, Zhennan Yan, Shaoting Zhang, and Dimitris N. Metaxas. A data-scalable transformer for medical image segmentation: Architecture, model efficiency, and benchmark, 2022.

Prachi Garg, Rohit Saluja, Vineeth N Balasubramanian, Chetan Arora, Anbumani Subramanian, and C.V. Jawahar. Multi-domain incremental learning for semantic segmentation. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2080–2090, 2022. doi: 10.1109/WACV51458.2022.00214.

Dipam Goswami, Yuyang Liu, Bartłomiej Twardowski, and Joost van de Weijer. FeCAM: Exploiting the heterogeneity of class distributions in exemplar-free continual learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Asx5eDqFZl.

Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9047–9057, 2022. doi: 10.1109/CVPR52688.2022.00885.

Yuanfeng Ji, Haotian Bai, Chongjian GE, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, and Ping Luo. AMOS: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. URL https://openreview.net/forum?id=Vk4-HUnkEak.

Chengjia Jiang, Tao Wang, Sien Li, Jinyang Wang, Shirui Wang, and Antonios Antoniou. Few-shot class-incremental semantic segmentation via pseudo-labeling and knowledge distillation. In *2023 4th International Conference on Information Science, Parallel and Distributed Systems (ISPDS)*, pp. 192–197, 2023. doi: 10.1109/ISPDS58840.2023.10235731.

Haeyong Kang, Jaehong Yoon, Sultan Rizky Hikmawan Madjid, Sung Ju Hwang, and Chang D. Yoo. On the soft-subnetwork for few-shot class incremental learning. In *The Eleventh International Conference on Learning Representations*, 2023a. URL https://openreview.net/forum?id=z57WK5lGeHd.

Zhiqi Kang, Enrico Fini, Moin Nabi, Elisa Ricci, and Karteek Alahari. A soft nearest-neighbor framework for continual semi-supervised learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11868–11877, October 2023b.

Tommie Kerssies, Daan De Geus, and Gijs Dubbelman. How to benchmark vision foundation models for semantic segmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1162–1171, 2024.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *arXiv preprint arXiv:2004.11362*, 2020.

Krishnateja Killamsetty, Xujiang Zhao, F. Chen, and Rishabh K. Iyer. Retrieve: Coreset selection for efficient and robust semi-supervised learning. *ArXiv*, abs/2106.07760, 2021. URL https://api.semanticscholar.org/CorpusID:235436029.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Bennett Landman, Zhoubing Xu, Juan Igelsias, Martin Styner, Thomas Langerak, and Arno Klein. Miccai multi-atlas labeling beyond the cranial vault–workshop and challenge. In *Proc. MICCAI multi-atlas labeling beyond cranial vault—workshop challenge*, volume 5, pp. 12. Munich, Germany, 2015.

Huan Liu, Li Gu, Zhixiang Chi, Yang Wang, Yuanhao Yu, Jun Chen, and Jin Tang. Few-shot class-incremental learning via entropy-regularized data-free replay. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pp. 146–162, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20052-6. doi: 10.1007/978-3-031-20053-3_9. URL https://doi.org/10.1007/978-3-031-20053-3_9.

Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21152–21164, 2023.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10): 1993–2024, 2014.

M Jehanzeb Mirza, Marc Masana, Horst Possegger, and Horst Bischof. An efficient domain-incremental learning approach to drive in all weather conditions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3001–3011, 2022.

Ri-Zhao Qiu, Peiyi Chen, Wangzhe Sun, Yu-Xiong Wang, and Kris Hauser. GAPS: Few-shot incremental semantic segmentation via guided copy-paste synthesis, 2023. URL https://openreview.net/forum?id=cDVL245jZa.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=CR1XOQ0UTh-.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Anjany Sekuboyina, Malek E Husseini, Amirhossein Bayat, Maximilian Löffler, Hans Liebl, Hongwei Li, Giles Tetteh, Jan Kukačka, Christian Payer, Darko Štern, et al. Verse: a vertebrae labelling and segmentation benchmark for multi-detector ct images. *Medical image analysis*, 73:102166, 2021.

Xiaoyu Tao, Xiaopeng Hong, Xinyuan Chang, Songlin Dong, Xing Wei, and Yihong Gong. Few-shot class-incremental learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12183–12192, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/68053af2923e00204c3ca7c6a3150cf7-Paper.pdf.

Songsong Tian, Lusi Li, Weijun Li, Hang Ran, Xin Ning, and Prayag Tiwari. A survey on few-shot class-incremental learning. *Neural Networks*, 169:307–324, 2024. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2023.10.039. URL https://www.sciencedirect.com/science/article/pii/S0893608023006019.

Girish Varma, Anbumani Subramanian, Anoop Namboodiri, Manmohan Chandraker, and CV Jawahar. Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pp. 1743–1751. IEEE, 2019.

Haoxiang Wang, Han Zhao, and Bo Li. Bridging multi-task learning and meta-learning: Towards efficient training and effective adaptation. In *International Conference on Machine Learning*. PMLR, 2021.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5362–5383, 2024.

Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. Totalsegmentator: robust segmentation of 104 anatomic structures in ct images. *Radiology: Artificial Intelligence*, 5 (5):e230024, 2023.

Yibo Yang, Haobo Yuan, Xiangtai Li, Zhouchen Lin, Philip Torr, and Dacheng Tao. Neural collapse inspired feature-classifier alignment for few-shot class-incremental learning. In *ICLR*, 2023.

Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2636–2645, 2020.

Bo Yuan and Danpei Zhao. A survey on continual semantic segmentation: Theory, challenge, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Jianpeng Zhang, Yutong Xie, Yong Xia, and Chunhua Shen. Dodnet: Learning to segment multi-organ and tumors from multiple partially labeled datasets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1195–1204, 2021.

Yixiao Zhang, Xinyi Li, Huimiao Chen, Alan L. Yuille, Yaoyao Liu, and Zongwei Zhou. Continual learning for abdominal multi-organ and tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023: 26th International Conference, Vancouver, BC, Canada, October 8–12, 2023, Proceedings, Part II*, pp. 35–45, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-43894-3. doi: 10.1007/978-3-031-43895-0_4. URL https://doi.org/10.1007/978-3-031-43895-0_4.

Da-Wei Zhou, Fu-Yun Wang, Han-Jia Ye, Liang Ma, Shiliang Pu, and De-Chuan Zhan. Forward compatible few-shot class-incremental learning. In *CVPR*, 2022.

Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.