MILE: Mixture of Incremental LoRA Experts for Continual Semantic Segmentation across Domains and Modalities

Shishir Muralidhara[1], Didier Stricker[1,2], and René Schuster[1,2]

[1] German Research Center for Artificial Intelligence (DFKI),
Trippstadter Str. 122, 67663 Kaiserslautern, Germany
[2] University of Kaiserslautern-Landau (RPTU),
Gottlieb-Daimler-Str. 47, 67663 Kaiserslautern, Germany
{firstname.lastname}@dfki.de

Abstract. Continual semantic segmentation requires models to adapt to new domains or modalities without sacrificing performance on previously learned tasks. Expert-based learning, in which task-specific modules specialize in different domains, has proven effective in mitigating forgetting. These methods include dynamic expansion, which suffers from scalability issues, or parameter isolation, which constrains the ability to learn new tasks. We introduce Mixture of Incremental LoRA Experts (MILE), a modular and parameter-efficient framework for continual segmentation across both domains and modalities. MILE leverages Low-Rank Adaptation (LoRA) to instantiate lightweight experts for each new task while keeping the pretrained base network frozen. Each expert is trained exclusively on its task data, thus avoids overwriting previously learned information. A prototype-guided gating mechanism dynamically selects the most appropriate expert at inference. MILE achieves the benefits of expert-based learning while overcoming its scalability limitations. It requires only a marginal parameter increase per task and tens of LoRA adapters are needed before matching the size of a single full model, making it highly efficient in both training and storage. Across domain- and modality-incremental benchmarks, MILE achieves strong performance while ensuring better stability, plasticity, and scalability.

Keywords: Continual Learning, Continual Semantic Segmentation, Domain Incremental Learning, Modality Incremental Learning

1 Introduction

Semantic segmentation is essential for autonomous driving, as it provides precise pixel-level understanding of the environment and supports key perception tasks. Despite significant progress driven by deep learning, most existing semantic segmentation models operate under rigid assumptions. They are trained under closed-world settings with large labeled datasets from a fixed domain, assume a consistent set of input modalities, and are designed to solve a single task. In real-world scenarios, autonomous vehicles are continually exposed to a wide range

of distributional shifts. Domain shifts arise from environmental changes such as transitioning between different geographical locations, varying weather conditions, and lighting, all of which can drastically alter the visual appearance of the same semantic classes. These shifts frequently result in performance degradation [8], as static models fail to generalize beyond their training. Retraining or fine-tuning on new data leads to catastrophic forgetting [16], where performance on previously learned domains or modalities is adversely affected. Continual Learning (CL) has emerged as a promising paradigm to address the limitations of static models by enabling them to incrementally learn new tasks while preserving previously learned knowledge. This introduces the stability-plasticity dilemma [17], a trade-off between learning new information (plasticity) and preserving prior knowledge (stability). This challenge arises because updating model parameters to learn new tasks inevitably leads to overwriting previous weights resulting in forgetting. In safety-critical applications such as autonomous driving, ensuring consistent and reliable performance is paramount, as any degradation in perception capabilities can lead to potentially hazardous outcomes. An effective strategy to mitigate forgetting and the stability-plasticity trade-off is expert-based learning, where separate expert modules are assigned to a task. This approach minimizes interference between tasks by training each expert in isolation, preventing the overwriting of previously learned weights and preserving task-specific knowledge. However, this approach suffers from poor scalability, as the model size increases linearly with the number of tasks or domains.

To address these challenges, we propose MILE: Mixture of Incremental LoRA Experts, a framework for continual semantic segmentation across domains and modalities. MILE leverages Low-Rank Adaptation (LoRA) [9] to enable computationally efficient updates for adapting to new tasks. It leverages the strengths of expert-based learning by using task-specific modules to prevent interference and forgetting, while addressing scalability limitations. Instead of instantiating entire networks for each domain, MILE introduces compact LoRA-based expert modules that represent only a small fraction of the full model parameters. This allows the system to scale to a large number of domains, requiring many such experts before even approaching the size of a single full model. As new domains or modalities are encountered, MILE instantiates and trains a corresponding LoRA expert, while keeping the shared pretrained weights frozen. This significantly reduces the number of trainable parameters, thereby lowering the computational cost of training and storage. During inference, we dynamically infer the domain of the input using a gating network and select the corresponding task expert.

2 Background and Related Works

Incremental learning involves training models on a sequence of tasks, enabling them to acquire new knowledge while retaining previously learned information. Depending on the objective of the task, two main incremental learning settings [10] are commonly considered: Class-incremental learning, where the input distribution remains fixed but new, non-overlapping subsets of classes are

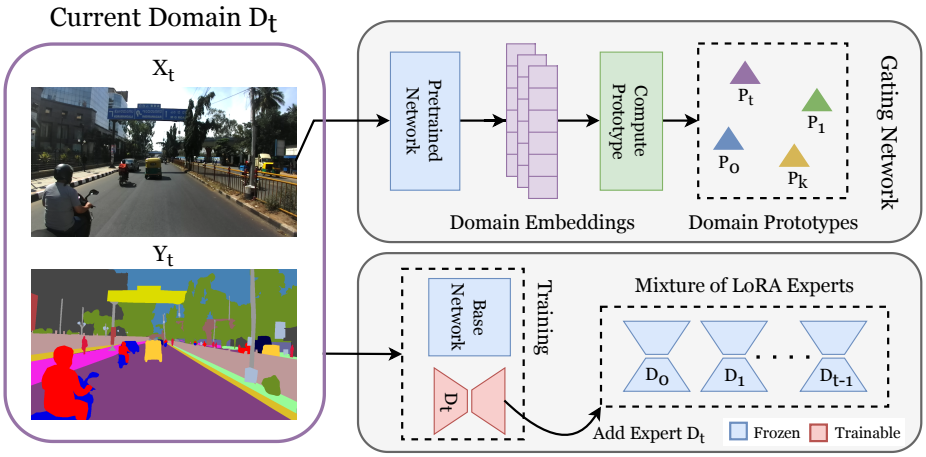

Fig. 1: Overview of the training process in MILE. Using the input images X_t, the domain prototype P_t is computed using a frozen pretrained network. LoRA expert D_t is instantiated and trained while keeping the base network frozen.

introduced sequentially. This setting has been extensively explored in image classification [33], and more recently in semantic segmentation [31]. In domain-incremental learning, the output space remains consistent, and the input distribution changes across tasks. Modality-incremental learning [7] extends domain-incremental learning to scenarios where new modalities are incrementally introduced. Additionally, recent works propose novel incremental settings that address evolving ontologies [12, 19] or simultaneous domain and class shifts [26, 22], reflecting increasingly complex real-world learning challenges.

2.1 Domain-Incremental Learning

The task of adapting to new domains has been widely explored in domain adaptation (DA) and domain generalization (DG) methods [4]. DA methods assume access to both source and target domain data during adaptation to the target domain. In contrast, domain-incremental learning (DIL) involves sequentially learning new domains without access to data from previous domains. Kalb *et al.* [10] evaluate the effectiveness of continual learning approaches for both class and domain incremental semantic segmentation, and observe replay-based methods are more effective for the latter. However, with replay-based approaches, there is an overhead for storing images for rehearsal or generating images and the subsequent pseudo-labeling. MDIL [6] addresses incremental learning across geographic domains by combining shared domain-invariant parameters with domain-specific adapters and decoders, and uses the task-ID at inference to select the appropriate domain-specific components. PSS [20] incrementally adapts to adverse driving conditions through a dynamically growing collection of domain-specific expert networks, which are automatically selected

at inference using a convolutional autoencoder ensemble. Replaying Styles [3] uses low-level style embeddings to replay past domains without storing raw data. RCIL [32] incrementally learns cities from Cityscapes [2] using two parallel branches to decouple old and new knowledge.

2.2 Modality-Incremental Learning

Modality-Incremental Learning (MIL) [7] extends incremental learning to scenarios where data from new sensor modalities such as RGB, depth, infrared, or thermal are introduced sequentially. Unlike domain-incremental learning, where the input distribution shifts within the same modality, MIL introduces drastically different input characteristics, making feature alignment and knowledge transfer more challenging. Similar to DIL, the label space remains consistent across tasks, with the same set of classes learned. Hegde *et al*. [7] propose DRMN, which uses modality-specific relevance maps to activate disjoint subsets of network parameters. Even under joint training where all modalities are available simultaneously, a single model struggles to learn effectively [7], highlighting the inherent difficulty of handling modality differences. MILE simultaneously addresses shifts across both domains and modalities, enabling the model to adapt to new environments and sensor types within a unified framework.

2.3 Expert-based Learning

Expert-based approaches mitigate catastrophic forgetting by assigning dedicated sub-networks and networks *i.e*. experts to individual tasks or domains. Expert learning can be achieved either dynamically, through network expansion, or within a fixed network through parameter isolation. Dynamic network expansion adds new expert networks [24, 1] as tasks arrive, enabling the model to expand its capacity over time. Parameter isolation operates within a fixed network capacity by assigning task-specific parameters. This can be achieved through masking [14] important weights for previous tasks, creating task-specific paths [5], or pruning [15] to free capacity for new tasks. Notably, several recent continual semantic segmentation methods [6, 20, 7] leverage expert-based learning to achieve strong performance. However, a critical limitation of expert-based approaches is scalability. Dynamically growing networks increase model size and storage costs as new tasks are added, which can become prohibitive over long sequences. Conversely, parameter isolation may struggle as the number of tasks increases, risking saturation of learning capacity and limiting adaptability.

3 Mixture of Incremental LoRA Experts

We propose Mixture of Incremental LoRA Experts (MILE), a framework that overcomes the scalability limitations of expert-based learning while retaining their benefits of task-specialization and strong knowledge preservation. MILE builds on Low-Rank Adaptation (LoRA) [9] to enable modular and scalable

adaptation. For each new domain or modality, a lightweight LoRA module is added while the base network and previous LoRA modules remain frozen. MILE does not require task-ID at inference and dynamically identifies the current domain using a lightweight gating network and selects the most appropriate LoRA expert for the given input. MILE offers the following benefits:

- **Parameter Efficiency and Scalability**. MILE incrementally adapts to new domains by adding compact LoRA adapters that constitute only a small portion of the full model size, ensuring scalability to long task sequences.
- **Reduced Forgetting through Expert Isolation**. By confining adaptation to these small parameter subsets, MILE preserves previously learned information by preventing adverse interference, thereby avoiding forgetting.
- **Stability and Plasticity:** MILE overcomes the trade-off by isolating past knowledge in independent LoRA experts while adding new ones for full plasticity, ensuring strong performance on both old and new tasks.
- **Resource-Efficient Adaptation:** By training only a small subset of parameters, MILE reduces both computational and memory overhead, making it well-suited for continual learning in resource-constrained environments.
- **Unified Learning of Domains and Modalities:** Through a modular mixture-of-experts approach, MILE supports both domain and modality-incremental learning within a single architecture.

3.1 Training with Low-Rank Adaptation

Incremental learning involves a model sequentially learning from a series of tasks $T = \{t_1, t_2, \ldots, t_n\}$ while retaining knowledge from previous tasks. Each task t is associated with task-specific data $D_t = (X_t, Y_t)$. In both domain- and modality-incremental settings, the input distribution shifts between tasks ($X_{t-1} \neq X_t$), but the number of classes C in Y remains consistent across all tasks. Adapting to new tasks typically necessitates retraining the entire network, which is computationally prohibitive in resource-constrained environments.

In contrast, our approach leverages Low-Rank Adaptation [9] for parameter-efficient continual learning [21]. LoRA is a parameter-efficient fine-tuning (PEFT) method that adapts large pre-trained models to downstream tasks with minimal computational cost. For a network with weights $W \in \mathbb{R}^{d\times k}$, LoRA introduces a small, low-rank update ΔW, represented as the product of two matrices: $B \in \mathbb{R}^{d\times r}$ and $A \in \mathbb{R}^{r\times k}$, where $r \ll d$ determines the number of trainable parameters. During training, only the LoRA weights are updated, while the base network weights W remain frozen. This significantly reduces both the computational and memory requirements, making LoRA particularly suitable for continual learning in resource-constrained settings.

We use LoRA to incrementally adapt to new domains or modalities. An overview of the training process with MILE is presented in Fig. 1. For each new task t, we train only the corresponding LoRA weights, denoted as ΔW_t. The full model for any task can be reconstructed by combining the frozen base network weights with the appropriate task-specific LoRA weights. This modular design

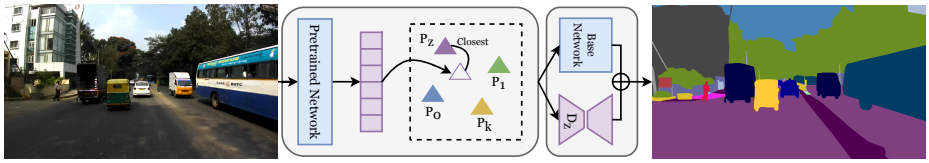

Fig. 2: Overview of inference in MILE. For an image, domain features are extracted using the frozen network, and the gating network infers the domain z by finding the closest prototype and then selects the expert D_z for segmentation.

requires storing only LoRA weights, enabling efficient scaling with a large number of tasks and addressing the scalability concerns of expert-based learning. It would take tens of LoRA modules before their combined size approaches that of a full model, highlighting the storage efficiency of our approach. Despite updating only a subset of parameters through LoRA, MILE achieves performance comparable to full network training, while significantly improving computational efficiency.

3.2 Inference with Prototype-Based Gating

Our approach trains task-specific LoRA experts for each domain or modality, and a key challenge at inference is selecting the appropriate expert without reintroducing catastrophic forgetting at the domain inference stage. In incremental learning, where experts are added incrementally, it is not possible to jointly train a classifier to route inputs to the appropriate expert. Updating such a classifier with new tasks can inadvertently overwrite knowledge from previously learned domains. To overcome this limitation, we propose a prototype-based gating mechanism that avoids the need for a trainable classifier.

For each task t, we compute a domain prototype $\mathbf{p}_t$ as the mean of the features of all training samples $X_t \in D_t$:

$$\mathbf{p}_t = \frac{1}{|D_t|} \sum_{(x,y)\in D_t} f_\theta(x), \tag{1}$$

where $f_\theta(\cdot)$ denotes a frozen feature extractor. $\mathbf{p}_t$ serves as a compact representation of the overall domain characteristics. During inference, given an input sample x, we first extract its feature representation: $\mathbf{z} = f_\theta(x)$. The gating function $g(\mathbf{z})$ selects the expert corresponding to the prototype most similar to $\mathbf{z}$:

$$g(\mathbf{z}) = \underset{t\in\{1,\ldots,n\}}{\text{argmax}}\ \text{cosine\_sim}(\mathbf{z}, \mathbf{p}_t) \tag{2}$$

where n is the total number of tasks.

An overview of inference in MILE with the prototype-based gating network is presented in Fig. 2. This approach offers several advantages: It enables domain identification without requiring a trainable classifier, eliminating the need for additional model updates. It scales efficiently with an increasing number of

tasks, as it only requires storing a single prototype per task, ensuring minimal memory overhead. Additionally, it inherently mitigates forgetting, since the feature extractor remains frozen throughout the process. The fixed feature space ensures consistent performance as new tasks are added without interfering with or altering the representations of previously learned tasks.

4 Experiments and Results

In this section, we describe the datasets and task settings, drawing from prior work on domain- and modality-incremental learning. We evaluate MILE under realistic challenges, including adverse weather, geographic shifts, and modality changes, and report results with a focus on scalability and efficiency.

4.1 Datasets

We evaluate MILE on diverse segmentation datasets capturing domain and modality shifts. Cityscapes (CS) [2] provides urban street scenes under clear weather with 19 classes, while ACDC [25] and Berkeley DeepDrive (BDD) [30] introduce domain variations due to adverse weather, illumination, and geographic diversity using the same label set. Indian Driving Dataset (IDD) [27] includes challenging conditions such as unstructured environments with 26 classes, including domain-specific labels. Freiburg Thermal [28] introduces a modality shift between aligned RGB and thermal infrared image pairs with 13 classes.

4.2 Implementation and Baselines

For semantic segmentation, we use SegFormer-B5 [29] from SATS [23] for all methods. MILE uses LoRA with rank r=16 which corresponds to 5% of the total model parameters, enabling parameter-efficient adaptation. We use batch size 12 and the learning rates are 0.01/0.001 for standard training and 0.05/0.005 for LoRA during initial/incremental steps. The gating network uses a pretrained ConvNeXt-Tiny [13] for feature extraction. We evaluate MILE against standard continual learning baselines: Single-Task (ST) models are trained independently on each task and serve as a reference to measure forgetting. Fine-Tuning (FT) adapts the model sequentially to new tasks, resulting in positive forward transfer but significant forgetting of previous tasks. Joint Training (JT) trains a single model on all task data simultaneously, typically providing an upper-bound performance. Elastic Weight Consolidation (EWC) [11] constrains model updates to preserve parameters important for previous tasks. Incremental Learning Techniques (ILT) [18] uses knowledge distillation to retain knowledge across tasks. We additionally report comparisons with PSS [20], MDIL [6], and DRMN [7] using results from their original papers. All methods are evaluated using performance change $\triangle P$ relative to corresponding single-task models.

Table 1: Results (mIoU) on the sequence $CS_{day} \rightarrow ACDC_{fog} \rightarrow ACDC_{rain} \rightarrow ACDC_{snow} \rightarrow ACDC_{night}$ after learning all tasks.

Method	*CS* $\triangle P$ (%)	*Fog* $\triangle P$ (%)	*Rain* $\triangle P$ (%)	*Snow* $\triangle P$ (%)	*Night* $\triangle P$ (%)	**Average**
ST / PSS†	71.79 -	60.05 -	57.83 -	60.03 -	47.86 -	59.51
Fine-Tune	63.26 **- 11.88**	65.43 **+ 08.96**	58.22 **+ 00.67**	56.41 **- 06.03**	50.93 **+ 06.44**	58.85
EWC	65.48 **- 08.79**	68.13 **+ 13.46**	57.87 **+ 00.07**	55.99 **- 06.73**	48.42 **+ 01.19**	59.18
ILT	66.33 **- 07.61**	65.89 **+ 09.73**	58.04 **+ 00.36**	55.73 **- 07.16**	39.19 **- 18.10**	57.04
MILE†	71.21 **- 00.81**	70.55 **+ 17.49**	64.17 **+ 10.96**	66.20 **+ 10.28**	49.95 **+ 04.39**	64.42
MILE	71.21 **- 00.81**	70.30 **+ 17.07**	64.06 **+ 10.77**	65.16 **+ 08.55**	49.95 **+ 04.39**	64.14
Joint-Train	72.79 **+ 00.70**	75.65 **+ 25.98**	67.05 **+ 15.94**	67.07 **+ 11.73**	52.19 **+ 09.07**	66.85

† denotes domain inference using an oracle.

4.3 Adverse Weather Domains

Using Cityscapes (CS) [2] and ACDC [25] we formulate a fine-grained categorization of adverse weather conditions with the following task sequence: $CS_{day} \rightarrow ACDC_{fog} \rightarrow ACDC_{rain} \rightarrow ACDC_{snow} \rightarrow ACDC_{night}$, consisting of 5 tasks. PSS [20] is evaluated on a broader categorization between normal and adverse weather conditions grouped together. We use the results from the fully trained single-task models to represent PSS evaluated with an oracle during inference. The results after learning all tasks are presented in Table 1. A limitation with single-task models is the lack of positive forward transfer when learning new tasks. When learning tasks with small incremental datasets, such as the individual domains from ACDC, all sequential learning methods demonstrate noticeable improvements. Constraining parameter updates with EWC [11] and ILT [18], hinders learning on new tasks and, although less severe than fine-tuning, still leads to partial forgetting of earlier tasks. MILE† with an oracle to provide domain-ID consistently improves performance across all ACDC domains compared to single-task models, and comparable results on the initial task. When using a gating network, MILE achieves results close to the oracle, with only a slight drop in performance due to misrouted samples.

4.4 Geographical Domains

Using Cityscapes (CS) [2], BDD [30], and IDD [27], we evaluate the task sequence $CS \rightarrow BDD \rightarrow IDD$, where each task represents a domain shift in the geographical location. To enable sequential learning, we use IDD with labels mapped to the 19 CS classes. The task-wise results after learning a new domain and evaluating on all previously seen domains are presented in Table 2. In contrast to the results presented in Table 1, we can observe that the sequentially trained models do not exhibit any positive forward transfer. Even with the fine-tuning, where learning on the new tasks is not regularized or constrained, the results on the new domain *i.e.* BDD in *Step* 2 and IDD in *Step* 3 fall short of the single-task models. With EWC [11] and ILT [18], we observe results

Table 2: Step-wise results (mIoU) for the task sequence $CS \rightarrow BDD \rightarrow IDD$ representing different geographical locations.

Method	*Step* 1 : *CS*	*Step* 2 : $CS \rightarrow BDD$		*Step* 3 : $CS \rightarrow BDD \rightarrow IDD$			**Average**
	CS	*CS* $\triangle P$ (%)	*BDD* $\triangle P$ (%)	*CS* $\triangle P$ (%)	*BDD* $\triangle P$ (%)	*IDD* $\triangle P$ (%)	
ST / PSS†	71.79	71.79 -	60.28 -	71.79 -	60.28 -	74.10 -	68.72
Fine-Tune	71.79	63.35 **- 11.76**	60.22 **- 00.10**	55.53 **- 22.65**	50.22 **- 16.69**	73.54 **- 00.76**	59.76
EWC	71.79	67.81 **- 05.54**	57.64 **- 04.38**	59.76 **- 16.76**	52.30 **- 13.24**	67.52 **- 08.88**	59.86
ILT	71.79	68.94 **- 03.97**	55.58 **- 07.80**	64.22 **- 10.54**	53.79 **- 10.77**	64.28 **- 13.25**	60.76
MILE†	71.21	71.21 **- 00.81**	59.12 **- 01.92**	71.21 **- 00.81**	59.12 **- 01.92**	72.37 **- 02.33**	67.57
MILE	71.21	71.21 **- 00.81**	58.99 **- 02.14**	71.21 **- 00.81**	58.35 **- 03.20**	71.05 **- 04.12**	66.87
Joint-Train	66.35	66.35 **- 07.58**	56.24 **- 06.70**	66.35 **- 07.58**	56.24 **- 06.70**	71.32 **- 03.75**	64.64

† denotes domain inference using an oracle.

marginally better than fine-tuning, highlighting the difficulty of using a single model in incrementally learning new domains. The joint training model, which had previously improved performance across all domains, underperforms compared to the single-task models in this setting, presumably due to substantial differences in domain characteristics across continents. Our proposed approach MILE, achieves results closest to the single-task models.

4.5 Modality-Incremental

The results on the Freiburg Thermal [28] dataset highlight the challenges posed by the significant distribution shift across modalities. We evaluate using the task sequence $RGB \rightarrow IR \rightarrow Gray$ and present the results in Table 3. Unlike DIL, this setting does not require domain inference, as the task-ID that corresponds to the sensor is inherently known. As expected, the large modality shifts present a significant challenge. Even the joint training model, with access to all modalities and no forgetting, underperforms, highlighting that a single model struggles to capture all modality-specific features. Single-task models achieve the highest per-

Table 3: Step-wise results (mIoU) for the task sequence $RGB \rightarrow IR \rightarrow Gray$ representing modality shifts.

Method	*Step* 1 : *RGB*	*Step* 2 : $RGB \rightarrow IR$		*Step* 3 : $RGB \rightarrow IR \rightarrow Gray$			**Average**
	RGB	*RGB* $\triangle P$ (%)	*IR* $\triangle P$ (%)	*RGB* $\triangle P$ (%)	*IR* $\triangle P$ (%)	*Gray* $\triangle P$ (%)	
ST / PSS	80.16	80.16 -	63.03 -	80.16 -	63.03 -	78.51 -	73.90
Fine-Tune	80.16	10.29 **- 87.16**	56.98 **- 09.60**	78.50 **- 02.07**	06.68 **- 89.40**	77.99 **- 00.66**	54.39
EWC	80.16	66.22 **- 17.39**	49.71 **- 21.13**	78.18 **- 02.47**	05.68 **- 90.99**	76.90 **- 02.05**	53.59
ILT	80.16	09.91 **- 87.64**	43.14 **- 31.56**	63.53 **- 20.75**	08.68 **- 86.23**	62.28 **- 20.67**	44.83
MILE^	78.51	78.51 **- 02.06**	58.50 **- 07.19**	78.51 **- 02.06**	58.50 **- 07.19**	76.46 **- 02.61**	71.16
MILE°	78.51	78.51 **- 02.06**	60.76 **- 03.60**	78.51 **- 02.06**	60.76 **- 03.60**	76.94 **- 02.00**	72.07
Joint-Train	78.64	78.64 **- 01.90**	23.62 **- 62.53**	78.64 **- 01.90**	23.62 **- 62.53**	77.16 **- 01.72**	59.81

∧ denotes sequentially trained models, ⋄ denotes models trained independently.

Table 4: Final results for the task $RGB \rightarrow IR \rightarrow Gray$.

Method	RGB $\triangle P$ (%)	IR $\triangle P$ (%)	Gray $\triangle P$ (%)	Average
Single-Task	80.16 -	63.03 -	78.51 -	73.90
MILE	78.51 **- 02.06**	60.76 **- 03.60**	76.94 **- 02.00**	72.07
Single-Task [7]	76.41 -	59.56 -	74.56 -	70.18
RMN [7]	73.13 **- 04.29**	55.01 **- 07.64**	68.29 **- 08.41**	65.48
DRMN [7]	73.21 **- 04.19**	54.95 **- 07.74**	69.38 **- 06.95**	65.85

formance, indicating that isolating modalities allows the model to learn modality-specific characteristics without interference, demonstrating task-specialization, one of the advantages of expert-based learning. The stability-plasticity trade-off in EWC and ILT, becomes even more challenging under large shifts. We study the influence of task sequence order under additional experiments to understand how modality sequences affect forward transfer and forgetting. Within the MILE framework, we observe that sequentially trained $MILE^{\wedge}$ underperforms compared to single-task $MILE^{\diamond}$, which is trained independently on each modality. These results highlight the challenges of modality-incremental learning and the limitations of joint training and sequential adaptation.

Additionally, we compare our method against RMN and DRMN from [7], using the relative performance $\triangle P$ to the corresponding single-task models. The results are summarized in Table 4. With parameter isolation in RMN and DRMN, the network capacity to learn new tasks becomes exhausted, affecting performance on later modalities. In contrast, MILE leverages the full network representation through individual LoRA modules, achieving performance closest to that of the corresponding single-task models across all modalities.

5 Additional Experiments

We present additional experiments, including the influence of task sequence in modality-incremental learning, domain incremental learning with heterogeneous labels, domain inference results, and an analysis of the scalability of MILE.

5.1 Influence of Task Sequence

The order in which tasks are presented plays a key role in balancing stability and plasticity between previously learned and new tasks. Previously in Table 3, we observed the performance on the first modality learned improve after learning the last task due to their visual similarity. Here, we evaluate using the sequence $RGB \rightarrow Gray \rightarrow IR$ and the results are presented in Table 5. Single-task, joint training and MILE trained independently are not affected by the task order. For sequentially trained methods, in Step 2, where $Gray$ is learned after RGB, the performance on the first modality remains relatively stable. This suggests that

Table 5: Step-wise results (mIoU) for the task sequence $RGB \rightarrow Gray \rightarrow IR$ representing modality shifts.

Method	*Step 1 : RGB*	*Step 2 : RGB → Gray*		*Step 3 : RGB → Gray → IR*			**Average**
	RGB	*RGB* $\triangle P$ (%)	*Gray* $\triangle P$ (%)	*RGB* $\triangle P$ (%)	*Gray* $\triangle P$ (%)	*IR* $\triangle P$ (%)	
ST / PSS	80.16	80.16 -	78.51 -	80.16 -	78.51 -	63.03 -	76.47
Fine-Tune	80.16	78.71 **- 01.81**	78.29 **- 00.28**	21.42 **- 73.28**	20.73 **- 73.60**	59.19 **- 06.09**	33.78
EWC	80.16	78.23 **- 02.41**	76.74 **- 02.25**	66.47 **- 07.08**	61.71 **- 21.40**	50.05 **- 20.59**	59.41
ILT	80.16	79.35 **- 01.01**	75.23 **- 04.18**	17.75 **- 77.86**	16.26 **- 79.29**	43.34 **- 31.24**	25.78
MILE°	78.51	78.51 **- 02.06**	76.94 **- 02.00**	78.51 **- 02.06**	76.94 **- 02.00**	60.76 **- 03.60**	72.07
Joint-Train	78.64	78.64 **- 01.90**	77.16 **- 01.72**	78.64 **- 01.90**	77.16 **- 01.72**	23.62 **- 62.53**	59.81

the shared visual characteristics between RGB and $Gray$ mitigate forgetting. This behaviour contrasts with the earlier sequence $RGB \rightarrow IR \rightarrow Gray$ in Table 3, where the second task IR affects the performance on RGB due to larger domain dissimilarity. The addition of IR, which differs significantly from the previous modalities, adversely interferes with the previous tasks and results in substantial forgetting. Consequently, the average performance in Table 5 across all modalities is lower than the previous task sequence $RGB \rightarrow IR \rightarrow Gray$.

5.2 Domain-Incremental Learning with Heterogeneous Labels

In classical domain-incremental learning, the set of classes C between domains remains consistent with $C_{t-1} = C_t$, and the model is only required to adapt to distributional shifts in the input space. However, in many real-world scenarios, the label spaces across domains are not homogeneous. For instance, the IDD dataset [27] contains 26 classes, including domain-specific labels like auto-rickshaw, which have no direct correspondence in BDD [30] or Cityscapes [2]. This illustrates how heterogeneous labels impose additional complexity: The model must expand its semantic space while still retaining representations of overlapping classes (*e.g.* car, person, or traffic light) that are present across domains. One common strategy to mitigate this issue is to map heterogeneous labels back into a common label space as done in Table 2 but this discards domain-specific and fine-grained information and undermines the utility of the model

Table 6: Final results (mIoU) for the task sequence $CS \rightarrow BDD \rightarrow IDD$.

Method	**CS** $\triangle P$ (%)	**BDD** $\triangle P$ (%)	**IDD** $\triangle P$ (%)	**Average**
Single-Task	71.79 -	60.28 -	65.54 -	65.87
MILE†	71.21 **- 00.81**	59.12 **- 01.92**	64.85 **- 01.05**	65.06
Single-Task [6]	72.55 -	54.10 -	61.97 -	62.87
MDIL† [6]	59.19 **- 18.41**	49.66 **- 08.21**	59.16 **- 04.53**	56.00

† denotes domain inference using an oracle.

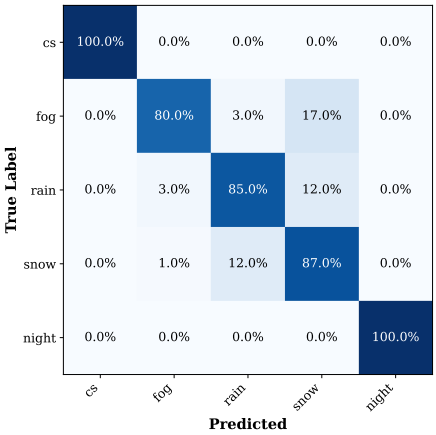

Fig. 3: Normalized confusion matrix for domain inference illustrating the effectiveness of the gating network.

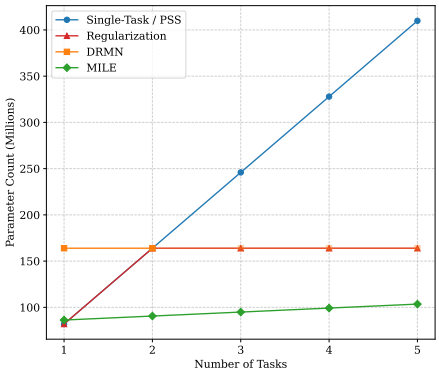

Fig. 4: Parameter growth across tasks for different continual learning approaches. MILE adds only a small number of LoRA parameters per task ensuring scalability.

in that domain. Expert-based learning approaches with dedicated decoders [6] or separate networks remain unaffected by label heterogeneity, as each expert is trained independently and evaluated only on its corresponding label set. The results on the task-sequence $CS \rightarrow BDD \rightarrow IDD$ with 26 classes in IDD is presented in Table 6 and compared with MDIL [6]. Beyond quantitative evaluation, Fig. 5 presents qualitative results. Unlike joint training, which is constrained by a shared label space, MILE preserves domain-specific information, retaining unique classes such as auto-rickshaw in IDD.

5.3 Domain Inference

We present the results of our domain inference stage, where the gating network identifies the domain and routes inputs to the corresponding task expert. The normalized confusion matrices for the classification results on the validation sets are presented in Fig. 3. Even in the challenging multi-domain classification across five weather domains, our approach achieves good performance. The overall performance of MILE using this gating network for domain inference is close to MILE with an oracle for domain inference, highlighting the effectiveness of routing to the correct task expert. The main advantage of our approach is that it does not require any additional training for domain inference, as we only use domain prototypes computed using frozen features.

5.4 Scalability in MILE

Figure 4 illustrates the total parameter count increase with the number of tasks for different continual learning approaches. Single-task and PSS [20] grow lin-

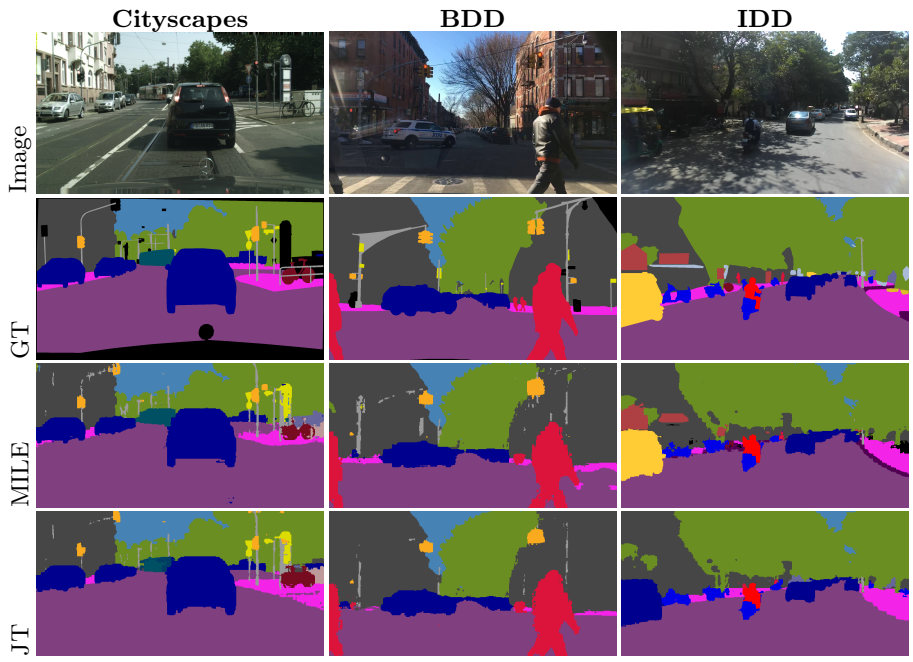

Fig. 5: Qualitative comparison of segmentation results for domain-incremental learning of geographical domains using Cityscapes [2], BDD [30], and IDD [27].

early, as each task has a dedicated model, which limits scalability. Regularization-based methods, maintain a constant parameter count across tasks, and only require the previous task model for regularizing the training of incremental tasks, but this reduces plasticity and can limit performance. DRMN [7] has a higher initial parameter count due to the relevance maps but its fixed network capacity can constrain learning over long task sequences.

MILE achieves a balance between these extremes: Using LoRA, MILE adds only a small number of parameters with each new task. Notably, even with several LoRA modules added for multiple tasks, the total parameter count remains well below that of a full task-specific model. This demonstrates a trade-off in continual learning approaches: Expert-learning based methods are highly effective but suffer from poor scalability, whereas regularization methods are scalable but are less flexible. MILE addresses these competing objectives, retaining the advantages of task-specialization while limiting parameter growth as new tasks are added, making it a practical solution for scalable and efficient continual learning.

6 Conclusion

In this work, we propose MILE (Mixture of Incremental LoRA Experts), a modular and parameter-efficient framework for continual semantic segmentation across

domains and modalities. MILE uses LoRA to add lightweight task-specific experts while keeping the base network frozen, preventing knowledge overwriting. A prototype-based gating mechanism selects the appropriate expert at inference. We extensively evaluate MILE across diverse domain-incremental settings, including adverse weather, geographical domains, and modalities. Across these tasks, MILE achieves performance on par with upper-bound baseline, demonstrating its effectiveness in balancing scalability, efficiency, and performance.

Acknowledgments

This work was partially funded by the German Federal Ministry of Research, Technology, and Space under the project COPPER (16IW24009).

References

1. Aljundi, R., Chakravarty, P., Tuytelaars, T.: Expert gate: Lifelong learning with a network of experts. In: CVPR (2017)
2. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
3. Deng, Y., Xiang, X.: Replaying styles for continual semantic segmentation across domains. In: Asian Conference on Pattern Recognition (2023)
4. Farahani, A., Voghoei, S., Rasheed, K., Arabnia, H.R.: A brief review of domain adaptation. Advances in Data Science and Information Engineering (2021)
5. Fernando, C., Banarse, D., Blundell, C., Zwols, Y., Ha, D., Rusu, A.A., Pritzel, A., Wierstra, D.: Pathnet: Evolution channels gradient descent in super neural networks. arXiv:1701.08734. (2017)
6. Garg, P., Saluja, R., Balasubramanian, V.N., Arora, C., Subramanian, A., Jawahar, C.: Multi-domain incremental learning for semantic segmentation. In: WACV (2022)
7. Hegde, N., Muralidhara, S., Schuster, R., Stricker, D.: Modality-incremental learning with disjoint relevance mapping networks for image-based semantic segmentation. In: WACV (2025)
8. Hell, F., Hinz, G., Liu, F., Goyal, S., Pei, K., Lytvynenko, T., Knoll, A., Yiqiang, C.: Monitoring perception reliability in autonomous driving: Distributional shift detection for estimating the impact of input data on prediction accuracy. In: ACM CSCS (2021)
9. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: ICLR (2022)
10. Kalb, T., Roschani, M., Ruf, M., Beyerer, J.: Continual learning for class-and domain-incremental semantic segmentation. In: IV (2021)
11. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences (2017)
12. Lin, Z., Pathak, D., Wang, Y.X., Ramanan, D., Kong, S.: Continual learning with evolving class ontologies. NeurIPS (2022)

13. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. CVPR (2022)
14. Mallya, A., Davis, D., Lazebnik, S.: Piggyback: Adapting a single network to multiple tasks by learning to mask weights. In: ECCV (2018)
15. Mallya, A., Lazebnik, S.: Packnet: Adding multiple tasks to a single network by iterative pruning. In: CVPR (2018)
16. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of Learning and Motivation. Elsevier (1989)
17. Mermillod, M., Bugaiska, A., Bonin, P.: The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. Frontiers in Psychology (2013)
18. Michieli, U., Zanuttigh, P.: Incremental learning techniques for semantic segmentation. In: ICCVW (2019)
19. Muralidhara, S., Bukhari, S., Schneider, G., Stricker, D., Schuster, R.: Cleo: Continual learning of evolving ontologies. In: ECCV (2024)
20. Muralidhara, S., Schuster, R., Stricker, D.: Domain-incremental semantic segmentation for autonomous driving under adverse driving conditions. In: ICPRAM (2025)
21. Muralidhara, S., Stricker, D., Schuster, R.: Clora: Parameter-efficient continual learning with low-rank adaptation. arXiv:2507.19887 (2025)
22. Park, M.Y., Lee, J.H., Park, G.M.: Versatile incremental learning: Towards class and domain-agnostic incremental learning. In: ECCV (2024)
23. Qiu, Y., Shen, Y., Sun, Z., Zheng, Y., Chang, X., Zheng, W., Wang, R.: Sats: Self-attention transfer for continual semantic segmentation. Pattern Recognition (2023)
24. Rusu, A.A., Rabinowitz, N.C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., Pascanu, R., Hadsell, R.: Progressive neural networks. arXiv:1606.04671 (2016)
25. Sakaridis, C., Dai, D., Van Gool, L.: ACDC: The adverse conditions dataset with correspondences for semantic driving scene understanding. In: ICCV (2021)
26. Toldo, M., Michieli, U., Zanuttigh, P.: Learning with style: Continual semantic segmentation across tasks and domains. TPAMI (2024)
27. Varma, G., Subramanian, A., Namboodiri, A., Chandraker, M., Jawahar, C.: Idd: A dataset for exploring problems of autonomous navigation in unconstrained environments. In: WACV (2019)
28. Vertens, J., Zürn, J., Burgard, W.: Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. In: IROS (2020)
29. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. NeurIPS (2021)
30. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: CVPR (2020)
31. Yuan, B., Zhao, D.: A survey on continual semantic segmentation: Theory, challenge, method and application. TPAMI (2024)
32. Zhang, C.B., Xiao, J.W., Liu, X., Chen, Y.C., Cheng, M.M.: Representation compensation networks for continual semantic segmentation. In: CVPR (2022)
33. Zhou, D.W., Wang, Q.W., Qi, Z.H., Ye, H.J., Zhan, D.C., Liu, Z.: Class-incremental learning: A survey. TPAMI (2024)