Incremental Few-Shot Semantic Segmentation via Multi-Level Switchable Visual Prompts

Anonymous ICCV submission

Paper ID 7277

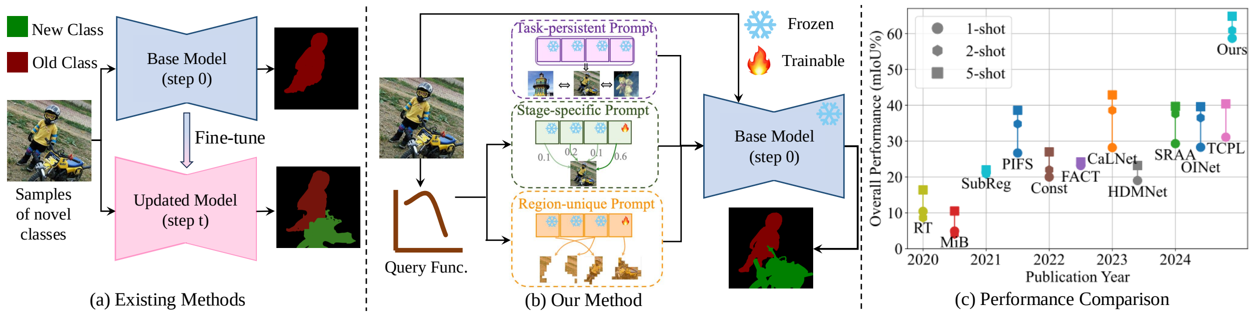

Figure 1. **Comparisons between existing methods and ours.** (a) Existing methods generally fine-tune parameters inherited from old stages and avoid catastrophic forgetting by keeping distribution of old class features. (b) Our method learns novel classes by tailoring multi-granular knowledge to input images achieved by adaptive switching prompts. (c) We compare the overall performance evaluated by Harmonic Mean of mIoU on base and novel classes.

Abstract

Existing incremental few-shot semantic segmentation (IFSS) methods often learn novel classes by fine-tuning parameters from previous stages. This inevitably reduces the distinguishability of old class features, leading to catastrophic forgetting and overfitting to limited new samples. In this paper, we propose a novel prompt-based IFSS method with a visual prompt pool to store and switch multi-granular knowledge across stages, boosting new class learning capability. Specifically, we introduce three levels of prompts: 1) Task-persistent prompts: capturing generalizable knowledge shared across stages, such as foreground-background distributions, to ensure consistent recognition guidance; 2) Stage-specific prompts: adapting to unique requirements of each stage by integrating its discriminative knowledge (e.g., shape difference) with common knowledge from previous stages; and 3) Region-unique prompts: encoding category-specific structures (e.g., edges) to accurately guide the model to retain local details. In particular, we introduce a prompt switching mechanism that adaptively allocates the knowledge required for base and new classes, avoiding interference between prompts and preventing catastrophic forgetting and reducing increasing computation. Our method achieves new state-of-the-art performance, outperforming previous SoTA methods by 30.28% mIoU-N on VOC and 13.90% mIoU-N on COCO under 1-shot.

1. Introduction

Incremental few-shot semantic segmentation (IFSS) [1–5] aims to extend segmentation models to novel classes continuously using limited new data without accessing old data. As models expand with few-shot samples, IFSS faces two critical challenges: 1) catastrophic forgetting of old classes and 2) overfitting to limited samples of novel classes. To address the first, existing methods [1, 6–9] reduce old knowledge forgetting by preserving the distribution consistency of old class features, but fine-tuning still harms the ability to distinguish old classes when learning novel classes and cannot solve the dilemma of competition between new and old abilities (see Fig. 1 (a)). Regarding the second issue, current approaches facilitate the rapid learning of novel classes by tailoring the prototypes of these classes [1–3] or using surrogate modalities to construct inter-class relationships [4, 5]. Unfortunately, due to insufficient model generalization ability and inadequate modal alignment, these methods struggle to effectively differentiate novel classes, resulting in confusion between novel and old classes.

Recently, prompt-based incremental learning [10–14] has garnered attention in image classification. This paradigm maintains an updatable prompt pool for frozen pre-trained vision transformers, adding new prompts to accommodate novel classes and prevent catastrophic forgetting in succeeding stages. However, its application in semantic segmentation remains challenging. On the one hand, embedding multiple background classes into a single continuous feature space leads to a decrease in feature distinguishability and confusion between unseen classes and background features. On the other hand, semantic segmentation requires collaborative prompts with multiple granularities to capture both global context and local details, enabling more meticulous segmentation.

In this paper, we innovatively propose an IFSS method based on multi-level switchable prompts (see Fig. 1 (b)). Specifically, we first introduce a prompt-based dense prediction framework that leverages well-structured text semantics to achieve seamless integration of novel classes with existing ones. First, the framework predefines multiple background semantics to help the model construct clearer background feature boundaries in the feature space, reducing overlap between background and foreground features. Secondly, it introduces a focus decomposition decoder consisting of two separators to align text embeddings with pixel features of foreground and background, respectively. Meanwhile, it further enhances alignment by updating visual prompts inserted into the frozen visual encoder, and expands the model by adding new prompts. This design enables the model to process the salient features of novel classes separately and adapt to background changes during incremental learning, effectively alleviating the confusion between novel and background classes.

Vanilla visual prompts exhibit several limitations in semantic segmentation, including a lack of fine-grained contextual information, growing computational costs as prompts accumulate, and interference from new prompts causing catastrophic forgetting. To address these challenges, we propose a method of multi-level switchable prompts (MSVP), a prompting strategy to balance knowledge retention and adaptability across incremental learning stages. MSVP consists of three levels of prompts: 1) **Task-persistent prompts (TP):** preserving general knowledge shared across stages (e.g., foreground-background distributions); 2) **Stage-specific prompts (SP):** adapting to the unique requirements of each stage by integrating its discriminative knowledge (e.g., shape difference) with common knowledge from previous stages; and 3) **Region-unique prompts (RP):** encoding category-specific structures (e.g., edges) to enhance local detail recovery. TP is frozen after base-stage training, preventing general knowledge from being disrupted by new tasks. SP, initialized during base training and continuously expanded, transfers generalizable experience to novel tasks, balancing rigidity and plasticity. RP is generated for fine-grained information aggregation of new classes. These three prompts enable multi-granularity knowledge transfer across stages, enhancing learning ability. To further mitigate interference and control computational costs, we introduce a flexible prompt-switching mechanism that dynamically tailors prompts for input images. Specifically, a pretrained image encoder (e.g., DINOv2 [15]) serves as a query function, generating global and local query features. Global features aggregate stage-specific prompts through an attention-based mechanism, while local features select region-unique prompts via nearest neighbor matching. This mechanism alleviates interference between prompts from different stages and keeps a constant number of active prompts, thus avoiding catastrophic forgetting and increasing computation.

Our method outperforms previous methods by a large margin (see Fig. 1 (c)) proving the effectiveness of the prompt-based dense prediction framework and multi-level switchable visual prompts. In conclusion, our contributions are summarized as:

1. We propose the first prompt-based IFSS framework, which introduces textual semantics and visual prompts to encode foreground and background classes separately, enabling incremental semantic segmentation.
2. We propose multi-level switchable visual prompts that customizes multi-granular knowledge tailored to input images, enhancing the model's ability to learn novel classes while maintaining knowledge of old classes.
3. Extensive experiments demonstrate the effectiveness of the proposed method. Under the 1-shot condition, it achieves 49.1% mIoU-N on VOC and 25.6% mIoU-N on COCO, setting a new SOTA performance.

2. Related Work

2.1. Incremental Few-Shot Semantic Segmentation

Semantic segmentation [16–21] is a basic computer vision task that involves partitioning an image into meaningful segments. Incremental few-shot semantic segmentation aims at continuously learning to segment novel categories via a few given samples, without forgetting knowledge of old categories. To achieve this, PIFS [1] and OINet [3] adopt a distillation training paradigm to avoid forgetting old knowledge and an effective prototype updating strategy of novel categories to learn novel classes. EHNet [2] maintains the old knowledge using the hyperclass representation bank and adaptively updates it to combine novel classes. Instead of distillation or storing old knowledge, CaLNet [4] employs a class-agnostic mask proposal to generate masks for both base and novel categories and integrates language embedding into visual features to enrich the representation of a few novel categories. However, the mask proposal module is prone to overfit base data, causing low recall rates for novel categories. Different from these methods, we propose an innovative prompt-based incremental few-shot learning method, which learns multi-level visual prompts and filters appropriate prompts for input images, facilitating keeping old knowledge and learning novel classes.

2.2. Prompt-based Incremental Learning

Prompt-based incremental learning has been studied in image classification [10–14]. Inspired by VPT [22], these methods generally freeze the pre-trained parameters and fine-tune only a set of novelly-added learnable prompts at the incremental stage without a rehearsal buffer to store past pristine examples for experience replay, which achieves remarkable performance. L2P [10] is the first method that introduces this training paradigm, which selects the most relevant prompts from a prompt pool in a key-value mechanism. Instead of merely leveraging task-specific prompts, Dual-Prompt [11] proposes to introduce general prompts shared by all tasks, achieving novel SOTA performance. Unlike the above two methods which learn a pool of key-value pairs to select learnable prompts, CODA-Prompt [12] introduces a decomposed prompt that consists of learnable prompt components that assemble to produce attention-conditioned prompts and optimizes the model in an end-to-end fashion. As far as we know, this is the first work that introduces prompt-based incremental learning methods in IFSS. In particular, we propose multi-level prompts to meet the needs of dense prediction tasks for multi-granular contexts.

3. Methods

In this section, we propose a prompt-based IFSS method that expands and updates multi-granularity switchable prompts to learn novel classes. It involves a prompt-based IFSS framework (see Fig. 2) to generate robust class representation and enable the model to expand by simply adding prompts, and an enhanced multi-level prompt generation method (see Fig. 3) to provide fine-grained knowledge and switch prompts to expand the model dynamically.

3.1. Prompt-based IFSS Framework

To avoid the interference of background classes on new class learning, we design an IFSS framework based on frozen visual-language models, which leverages textual semantics and visual prompts to encode foreground and background classes separately, thus enabling incremental learning using prompts. The framework encompasses four key components: image encoding, text encoding, pixel decoding, and targeted optimization objectives.

Image encoding. An image is encoded into a sequence of tokens by a Patch Embedding block as in [23]. Visual prompts concatenated with image tokens are input into each

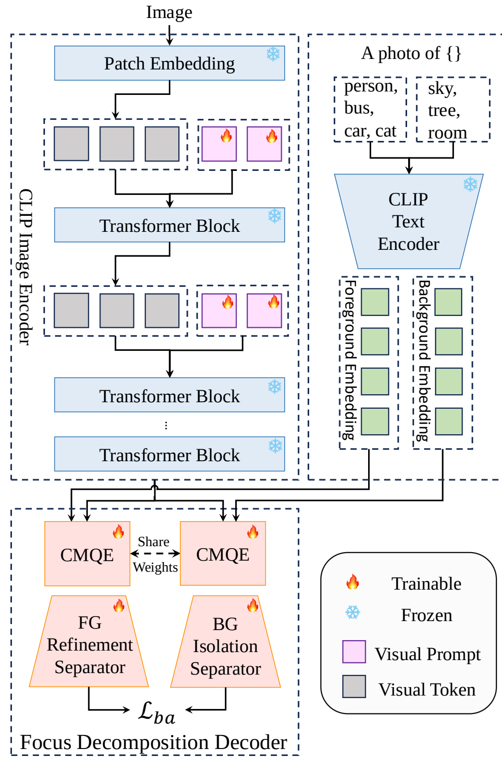

Figure 2. **The proposed IFSS framework.** Image tokens, concatenated with visual prompts, are encoded through successive Transformer blocks. Foreground and background text embeddings, along with image features, are fed into the Focus Decomposition Decoder for final predictions. Two pixel separators are used to distinguish foreground from background and identify specific foreground classes, respectively. The cross-modal query engine (CMQE) generates queries with robust fused modal information.

Transformer block as:

$$[x_{i+1}; p_o] = f_i(x_i, p^{(i)}) \tag{1}$$

where x_i denotes the input tokens output by layer $i-1$, x_{i+1} denotes the output of layer i, p_o denotes the encoded $p^{(i)}$ which will not be input into next layer. These visual prompts are composed of learnable vectors and are distinct between layers, providing an effective way to inject knowledge into pre-trained models. In the incremental stage, the expanded visual prompts are concatenated with the original ones as supplementary to recognize novel classes, as:

$$[x_{i+1}; p'_o] = f_i(x_i, [p^{(i),t-1}; p_e^{(i)}]) \tag{2}$$

where $p_e^{(i)}$ denotes the expanded prompts of layer i of stage $t-1$, $p^{(i),t-1}$ denotes the prompts of stage $t-1$, and $[\cdot;\cdot]$ denotes the concatenation operation. We take this vanilla prompt extension approach as our baseline.

Text encoding. To discriminate background pixels, instead of adding a class of "background", we pre-define a series of background classes, such as ”sky”, ”room”, etc, to pre-isociate backgroud into possible classes. In line with previous works, class names are formatted as ”*a photo of CLASS_NAME*” and encoded using the frozen CLIP Text Encoder to obtain embeddings for both background and foreground classes, denoting as $T_{bg} \in \mathbb{R}^{C_{bg} \times D}$ and $T_{fg} \in \mathbb{R}^{C_{fg} \times D}$, where C_{bg},C_{fg} represent the number of introduced background classes and foreground classes.

Pixel Decoding. We propose a focus decomposition decoder (FDD) to predict pixel semantics by aligning pixel features with class embeddings derived from the language modality. This decoder is composed of a cross-modal query engine (CMQE), inspired by [24], which generates class queries with robust generalization capabilities, along with two separators: one for identifying foreground and background pixels and the other for classifying the semantics of foreground pixels. The dual-decoder architecture enables the model to separately process salient features of new classes and adapt to background variations during incremental learning, making it easier to seamlessly embed features of novel classes into existing class representations.

CMQE integrates visual and linguistic information to generate robust class-specific queries, enhancing segmentation performance on both base and novel classes. Denoting $T_{fg} = \{t_0, t_1, ..., t_{C_{fg}}\}$ and $g \in \mathbb{R}^D$, the class-specific query can be denoted as $Q_{fg} = \{\hat{q}_0, \hat{q}_1, ..., \hat{q}_i, ..., \hat{q}_{C_{fg}}\}$, where $\hat{q}_i$ is calculated as:

$$\hat{q}_i = \texttt{MLP}([t_i \odot g; t_i]) \tag{3}$$

where g denotes the global feature, t_i denotes the class embedding of the i-th class, $\odot$ denotes the Hadamard product and `MLP` is used to align the dimension with pixel features. The background queries Q_{bg} is calculated the same as Q_{fg}.

Subsequently, class-specific queries Q_{bg}, Q_{fg} are forward to BG isolation separator and FG refinement separator to align with pixel features, respectively. We employ a cascaded cross-attention structure to facilitate this alignment, where Q_{bg} and Q_{fg} are the query, and pixel features $\mathbf{P} \in \mathbb{R}^{HW \times d}$ are the keys and values. Finally, we take the scaled dot-product attention in the separator's last block as the final semantic masks, as:

$$\mathbf{M}_{bg} = \frac{\phi_q^{bg}(\hat{\mathbf{Q}}_{bg})\phi_k^{bg}(\mathbf{P})^T}{\sqrt{d_k}}, \mathbf{M}_{fg} = \frac{\phi_q^{fg}(\hat{\mathbf{Q}}_{fg})\phi_k^{fg}(\mathbf{P})^T}{\sqrt{d_k}}, \tag{4}$$

where $\hat{\mathbf{Q}}_{bg}, \hat{\mathbf{Q}}_{fg}$ denote queries aligned by cross-attention blocks before the last block, ϕ denotes the linear projection, d_k is the dimension of the keys, and $\mathbf{M}_{bg} \in \mathbb{R}^{HW \times C_{bg}}$,$\mathbf{M}_{fg} \in \mathbb{R}^{HW \times C_{fg}}$ are the scores of each class.

Optimization objectives. Since we pre-define multiple background classes while the ground truth includes only a generic 'background' label, we propose a background aggregation loss $\mathcal{L}_{ba}$, to address this discrepancy. We regard the maximum scores of all background classes at each pixel as final background scores $\hat{\mathbf{M}}_{\mathbf{bg}} \in \mathbb{R}^{HW \times 1}$, and we regard the maximum logits of all foreground classes at each pixel to represent the likelihood of being a foreground pixel $\hat{\mathbf{M}}_{\mathbf{fg}} \in \mathbb{R}^{HW \times 1}$, as:

$$\begin{aligned} \hat{\mathbf{M}}_{\mathbf{bg}} &= \max_i \mathbf{M}_{\mathbf{bg}}{}^{j,i}, j = 1, 2, ..., HW \\ \hat{\mathbf{M}}_{\mathbf{fg}} &= \max_i \mathbf{M}_{\mathbf{fg}}{}^{j,i}, j = 1, 2, ..., HW. \end{aligned} \tag{5}$$

An intuitive and vanilla way is to optimize two separators independently, as:

$$\begin{aligned} \mathcal{L}_{van} =& \mathcal{L}_{seg}(y, [1 - \hat{\mathbf{M}}_{fg}; \mathbf{M}_{fg}]) \\ &+ \alpha_1 \mathcal{L}_{seg}(\overline{y}, [\hat{\mathbf{M}}_{bg}; 1 - \hat{\mathbf{M}}_{bg}]), \end{aligned} \tag{6}$$

where $y \in \mathbb{R}^{HW \times (C_{fg}+1)}$ denotes the one-hot labels of all classes including background and $\overline{y} \in \mathbb{R}^{HW \times 2}$ denotes the one-hot labels of foreground and background, α_1 is the weight to balance representation of FG and BG, and $\mathcal{L}_{seg}$ denotes the widely used pixel-wise classification loss as in [24] [18], the combination of focal loss [25] and dice loss [26]. To mitigate the potential issue of misaligned optimization directions, we introduce a more flexible loss function that prevents feature confusion among novel, old, and background classes during incremental learning, as:

$$\mathcal{L}_{ba} = \mathcal{L}_{seg}(y, [\hat{\mathbf{M}}_{bg}; \mathbf{M}_{fg}]) + \alpha_2 \mathcal{L}_{seg}(\overline{y}, [\hat{\mathbf{M}}_{bg}; \hat{\mathbf{M}}_{fg}]), \tag{7}$$

which jointly constrains the masks output by the two heads, on the one hand to distinguish the specific categories of foreground pixels as the first term of $\mathcal{L}_{ba}$, and on the other hand to separate foreground and background pixels to alleviate the problem of class imbalance as the second term. This dual constraint encourages accurate pixel classification and facilitates model expansion.

3.2. Multi-level Switchable Visual Prompts

Simply adding visual prompts during incremental learning has shown improvements (see Sec. 4.4). However, it presents three challenges: 1) computation increase: an increase in visual prompts raises computational complexity, 2) information dilution: stage-wise incremental prompts dilute the information of each stage, leading to knowledge forgetting and a diminished capacity to learn novel classes, and 3) insufficient granularity: a single level of prompts fails to meet the multi-granularity contextual needs essential for semantic segmentation. To address these, we propose to switch appropriate multi-level visual prompts tailored for input images, as shown in Fig. 3. It includes task-persistent prompts, stage-specific prompts, and region-unique prompts, alongside a flexible prompt switching mechanism.

Task-persistent Prompts (TP). According to the theory of Complementary Learning Systems (CLS) [27, 28],

ICCV #7277

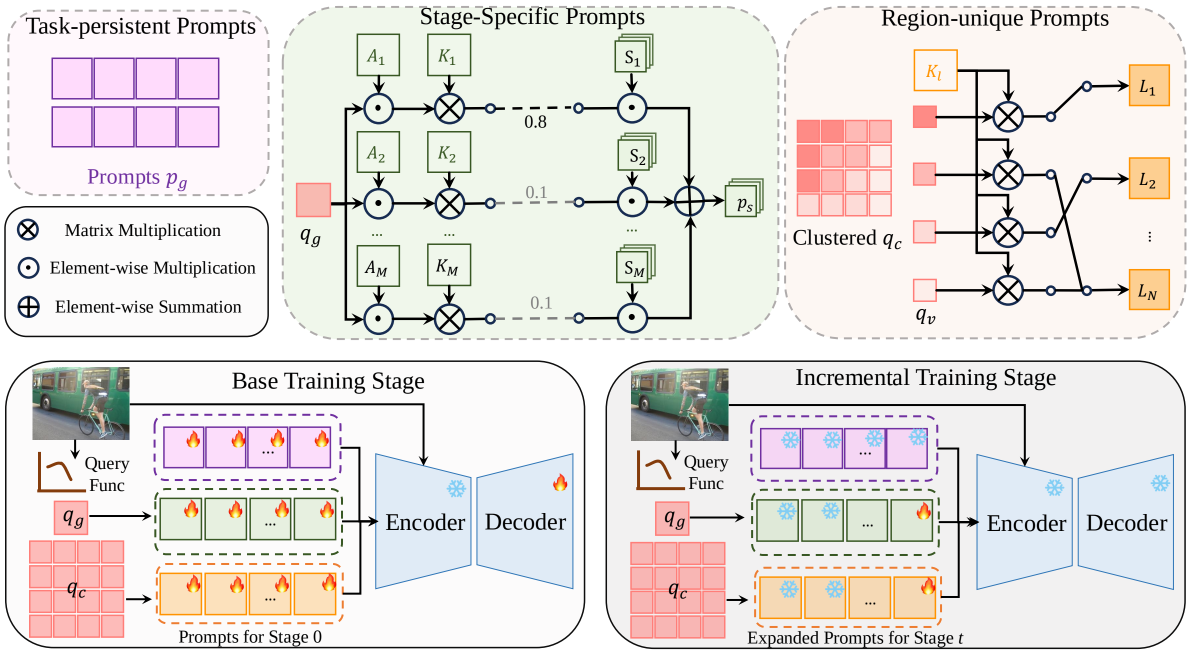

Figure 3. **The pipeline of the proposed multi-level switchable visual prompts.** Images are inputs into the query function to obtain global query features q_g and pixel-wise query features q_c. Stage-specific prompts p_s are generated by an attention-like integration way through q_g. Region-unique prompts p_l are generated by nearest neighbor matching through clustered q_c. Image tokens, concatenated with these selected prompts, are input into pre-trained models to produce predictions. At incremental training stage, the model extends by fine-tuning novelly-added stage-specific prompts and region-unique prompts.

humans learn continually via the synergy between the hippocampus and the neocortex. The former learns pattern-separated representation on specific experiences while the later focuses on learning more general and transferable representation to enhance the capability of learning future stages. Inspired by this, we propose task-persistent prompts to mitigate catastrophic forgetting and capture generalizable representation to facilitate encoding shared semantic structures or relationships for novel classes, including foreground-background distribution, shared contours and edges between categories, and so on. Specifically, $p_g^{(i)} \in \mathbb{R}^{L_g \times D}$ are learnable vectors with pre-defined sequence length L_g and embedding dimension D, which is trained during base training stage to learn generalizable knowledge and frozen during incremental stages.

Stage-specific Prompts (SP). Although TP maintains invariant core knowledge, they lack flexibility for rapid adaptation in incremental stages. To address this, we propose stage-specific prompts that dynamically guide the model to adapt fine-grained distinctions to learn new stages. However, limited training samples hinder learning robust stage-specific representations. To overcome this, we design a knowledge inheritance mechanism that integrates discriminative information from current stages with preserved prior knowledge. This is achieved through an attention-like mechanism that aggregates stage-specific prompts based on the correlation between knowledge required for inputting images and learned knowledge across stages.

Specifically, we take a pre-trained image encoder (e.g., DINOv2 [15]) as query function to obtain the global features $q_g \in \mathbb{R}^{1\times D}$ to aggregate relevant knowledge stored in stage-specific prompts $\mathbf{S}^{(i)} \in \mathbb{R}^{M\times L_s \times D}$, where M denotes the number of current training stage and L_s denotes the number of learnable vectors for each stage. We then calculate the correlation γ between the knowledge required by the input image and knowledge learned from all stages, as:

$$\gamma = \mathtt{Softmax}(< q_g \odot \mathbf{A}^{(i)}, \mathbf{K}^{(i)} > / \tau) \in \mathbb{R}^M \tag{8}$$

where $< \cdot >$ denotes cosine similarity, $\odot$ denotes element-wise multiplication, τ denotes the temperature coefficient, $\mathbf{K}^{(i)} \in \mathbb{R}^{M\times D}$ are keys corresponding to each stage-specific prompt. To allow the query to focus on specific patterns, an attention vector $\mathbf{A}^{(i)} \in \mathbb{R}^{M\times D}$ corresponding to each stage-specific prompt is added. For example, a prompt designed for recognizing car textures can focus on details like headlights while ignoring unrelated features. Additionally, in contrast to [12], where the weight vector is derived directly from the cosine similarity, we employ a normalized similarity computed with softmax as the weight vector, which ensures that stage-specific prompts unrelated to the input image are not aggregated into the final prompts. Note that all the vectors $\mathbf{S}^{(i)}, \mathbf{A}^{(i)}, \mathbf{K}^{(i)}$ are learnable vectors.

After obtaining γ, stage-specific prompts $p_s^{(i)}$ for the input image are aggregated as:

$$p_s^{(i)} = \gamma \cdot \mathbf{S}^{(i)} = \gamma_1 \mathbf{S}_1^{(i)} + \gamma_2 \mathbf{S}_2^{(i)} + ... + \gamma_M \mathbf{S}_M^{(i)}, \quad (9)$$

where $p_s^{(i)} \in \mathbb{R}^{L_s \times D}$. Thus, knowledge required by each image is integrated according the similarity between q_g and $\mathbf{K}^{(i)}$. For example, assuming that the input image contains class 'COW', the knowledge required to segment this class is similar to that of class 'SHEEP' learned from stage s. γ between the query and the key corresponding to stage s and current stage can be calculated as 0.8, 0.2. It means most of knowledge required by the input image can be inherited from stage s and combined with discriminative information learned from current stage.

Region-unique Prompts (RP). As a dense prediction task, semantic segmentation requires more fine-grained knowledge of typical structures for specific categories used for storing local details. To this end, we propose to query the best region-unique prompts conditioned on the distance between the keys and local features of the input image.

Firstly, we obtain the pixel-wise features q_c from the query function. Secondly, querying a prompt for pixel-wise features undoubtedly increases the computational complexity and generates redundant noise. Instead, we apply the `KMeans` algorithm to cluster these features into h categories and let the clustered centroids $q_v \in \mathbb{R}^{h \times D}$ represent the local information of the image. Thirdly, q_v is utilized to query local prompts corresponding to the key $\mathbf{K}_l^{(i)} \in \mathbb{R}^{N \times D}$ closest to q_v, where $N = C_{bg} + C_{fg}$. We have a set of local prompts $\mathbf{L}^{(i)} \in \mathbb{R}^{N \times D}$ for layer i. This process can be formalized as:

$$q_v = \texttt{KMeans}(q_c, h) \quad (10)$$

$$idx = \texttt{argmax}_{1 \le j \le N} q_v \mathbf{K}_{l,j}^{(i)} \quad (11)$$

$$p_l^{(i)} = \mathbf{L}^{(i)}[idx, :] \quad (12)$$

where $p_l^{(i)} \in \mathbb{R}^{h \times D}$ denotes the selected local prompts when given the query q_c. Thereby, a prompt is assigned to each background and foreground class and is queried using the image's local information via nearest matching.

However, the `argmax` operator prevents gradient backpropagation, necessitating additional supervision. To enable end-to-end training, we employ a `Gumbel_Softmax` operation to replace the above procedure, similar to [29], which can be simply formulated as

$$p_l^{(i)} = \texttt{Gumbel\_Softmax}(q_v \mathbf{K}_l^{(i)}) \mathbf{L}^{(i)}. \quad (13)$$

The detailed process is described in Appendix. A.1.

Furthermore, we incorporate attention masks into these visual prompts to prevent prompts specific to individual pixel segments from influencing other segments. When given pixel-wise query features q_c, the feature similarity between pixels can be calculated as $\mathbf{S}_c \in \mathbb{R}^{HW \times HW}$. The pixels with similarity higher than the threshold ζ are the pixels prompted by region-unique prompts, while the rest are the masked pixels, which can be formulated as:

$$\hat{\mathbf{S}}_c = \begin{cases} 0, & \mathbf{S}_c > \zeta \\ -\infty, & \mathbf{S}_c <= \zeta, \end{cases} \quad (14)$$

where a higher ζ means that the area prompted by the region is smaller, while the opposite means it is larger. Finally, extracting the mask corresponding to centroids q_v from $\hat{\mathbf{S}}_c$ and inserting it into the attention mask of the self-attention structure can limit the scope of the region-unique prompts.

Incremental training. During incremental training, task-persistent prompts, stage-specific prompts of previous stages and region-unique prompts of old classes remain frozen. We expand $\mathbf{S}^{(i)}, \mathbf{A}^{(i)}, \mathbf{K}^{(i)}$ to learn knowledge of new stages and expand $\mathbf{K}_l^{(i)}, \mathbf{L}^{(i)}$ to learn local details of new classes, and exclusively train the newly expanded components. The detail is formulated in Appendix A.5.

The multi-level switchable prompts generate multi-granular contextual information essential for semantic segmentation,enhancing new-class adaptability. And it addresses two critical limitations of the conventional method by simply adding prompts: 1) preventing information interference and dilution through adaptive prompts selection thus reducing catastrophic forgetting, (See Tab. 3), and 2) alleviating increasing computation by maintaining fixed input sequence length (See Appendix B.6).

4. Experiments

4.1. Datasets

We conduct experiments on Pascal VOC 2012 [40] and COCO [41, 42] as in previous works [1, 2, 4]. VOC contains 20 classes and one background class. In COCO, we use the 80 classes and the residual classes are labeled as background. We consider 15 and 60 of the classes as base and 5 and 20 classes as novel, for VOC and COCO respectively. The protocols start with pretraining on base classes and multiple steps on novel classes in line with [1, 4], i.e., 5 steps of 1 novel class on VOC and 4 steps of 5 novel classes on COCO. We divide the VOC dataset into 4 folds of 5 classes each and the COCO dataset into 4 folds of 20 classes each. We run experiments 5 times, with each experiment considering one fold at a time as the set of novel classes. In each setting, we explore incremental steps using 1, 2, or 5 images. Following the previous methods [1, 2, 4], we evaluate the performance via three metrics based on the mean Intersection over Union (mIoU): mIoU on base classes (mIoU-B), mIoU on novel classes (mIoU-N) and the Harmonic Mean (HM) of the two.

Table 1. Comparison with SOTA methods on VOC. Bold/Underline indicate SoTA/The Second Best.

	Method	1-shot Base	1-shot Novel	1-shot HM	2-shot Base	2-shot Novel	2-shot HM	5-shot Base	5-shot Novel	5-shot HM
FSC	WI [30]	66.6	16.1	25.9	66.6	19.8	30.5	66.6	21.9	33.0
	DWI [31]	67.2	16.3	26.2	67.5	21.6	32.7	67.6	25.4	36.9
	RT [32]	49.2	5.8	10.4	36.0	4.9	8.6	45.1	10.0	16.4
FSS	AMP [33]	58.6	14.5	23.2	58.4	16.3	25.5	57.1	17.2	26.4
	SPN [34]	49.8	8.1	13.9	56.4	10.4	17.6	61.6	16.3	25.8
IL	LwF [35]	42.1	3.3	6.2	51.6	3.9	7.3	59.8	7.5	13.4
	ILT [36]	43.7	3.3	6.1	52.2	4.4	8.1	59.0	7.9	13.9
	MiB [6]	43.9	2.6	4.9	51.9	2.1	4.0	60.9	5.8	10.5
IFL	SubReg [37]	55.4	13.2	21.3	56.7	12.7	20.8	59.7	13.5	22.0
	Const [38]	58.4	12.1	20.0	61.3	13.4	22.0	62.2	17.2	27.0
	FACT [39]	57.0	14.6	23.2	57.4	15.1	23.9	58.8	15.2	24.2
	PIFS [1]	64.1	16.9	26.7	65.2	23.7	34.8	64.5	27.5	38.6
	OINet [3]	66.1	18.0	28.3	66.3	25.2	36.5	66.4	28.2	39.6
	CaLNet [4]	**74.2**	17.4	28.2	**74.4**	26.1	38.6	**74.7**	30.1	42.9
	SRAA [5]	66.4	18.8	29.3	65.1	26.4	37.6	64.3	28.7	39.7
	Ours	73.04	**49.08**	**58.71**	73.21	**52.19**	**60.93**	73.36	**58.13**	**64.86**

Table 2. Comparison with SOTA methods on COCO. Bold/Underline indicate SoTA/The Second Best.

	Method	1-shot Base	1-shot Novel	1-shot HM	2-shot Base	2-shot Novel	2-shot HM	5-shot Base	5-shot Novel	5-shot HM
FSC	WI [30]	46.3	8.3	14.0	46.5	9.3	15.4	46.3	10.3	16.8
	DWI [31]	46.2	9.2	15.3	46.5	11.4	18.3	46.6	14.5	22.1
	RT [32]	38.4	5.2	9.1	43.8	10.1	16.4	44.1	16.0	23.5
FSS	AMP [33]	36.6	7.9	13.1	36.0	9.2	14.6	33.2	11.0	16.5
	SPN [34]	40.3	8.7	14.3	41.7	12.5	19.2	41.4	18.2	25.3
IL	LwF [35]	41.0	4.1	7.4	42.7	6.5	11.3	42.3	12.6	19.4
	ILT [36]	43.7	6.2	10.8	47.1	10.0	16.5	45.3	15.3	22.8
	MiB [6]	40.4	3.1	5.8	42.7	5.2	9.3	43.8	11.5	18.2
IFL	SubReg [37]	38.4	8.0	13.2	39.5	10.1	16.0	40.0	10.3	16.4
	Const [38]	39.0	8.2	13.6	40.6	11.4	17.8	41.1	11.3	17.7
	FACT [39]	37.9	8.6	14.0	38.9	11.7	18.0	39.4	12.3	18.7
	PIFS [1]	40.4	10.4	16.6	40.1	13.1	19.8	41.1	18.3	25.3
	OINet [3]	41.4	11.7	18.2	41.5	14.4	21.4	41.5	19.7	26.7
	CaLNet [4]	48.4	10.6	17.4	48.5	13.4	21.0	48.6	18.6	26.9
	SRAA [5]	40.7	11.3	17.7	40.5	15.2	22.1	41.0	19.7	26.6
	Ours	**48.85**	**25.60**	**33.59**	**48.52**	**28.05**	**35.54**	**48.61**	**32.38**	**38.86**

4.2. Implementation Details

In this work, we choose ViT-B [23] as the image encoder, and pretrained ViT-B of DINOv2 [15] as our query function, which can output accurate q_g and q_c to query appropriate prompts. Meanwhile, to make training stable and provide meaningful initial keys for region-unique prompts, we take text embeddings encoded by CLIP text encoder in a manner of CoOP [43] as the keys for region-unique prompts, which is detailed in Appendix. A.2. Stage by stage, the backbone and the query function are frozen, all prompts as well as the decoder are trainable. During incremental training, we freeze all parameters but expand and update the stage-specific prompts and region-unique prompts, which is described in Appendix. A.5. We add orthogonality constraints to parameters of stage-specific prompts to avoid interference between existing and new knowledge and reduce catastrophic forgetting. During base training, the model is trained for 20k iterations on VOC and 80k iterations on COCO. During incremental training, for both VOC and COCO, the model is trained for 400 iterations per step.

4.3. Compare with State-of-the-art methods

We mainly conduct comparison between few-shot classification methods (FSC) [30–32], few-shot semantic segmentation methods (FSS) [33, 34], incremental learning methods (IL) [37–39], and incremental few-shot semantic segmentation methods (IFSS).

Evaluation on VOC. The results of 1-shot, 2-shot and 5-shot experiments are presented in Tab. 1. In general, our method achieves novel SOTA performance on novel classes of 49.08% , 52.19% and 58.13% mIoU for 1-shot, 2-shot, and 5-shot scenarios respectively. Additionally, our method also attains the best overall performance with HM scores of 58.71% and 60.93% and 64.86%. Comparing our methods with other methods, it is evident that FSC methods excel in retaining knowledge of base classes, achieving competitive performance on these classes, since FSC methods, such as WI [30] and DWI [31], expand classifiers using class prototypes, thereby preventing the corruption of learned knowledge. In contrast, FSS and IL methods perform poorly on both base and novel classes. While meta-learning helps FSS methods adjust to novel classes, they struggle to retain old knowledge. And IL methods require many novel samples, resulting in low performance on few-shot tasks. Our method remarkably outperforms prior SOTA IFSS method SRAA [5] on novel classes by 30.28%, 25.79%, and 29.43% mIoU in 1-shot, 2-shot and 5-shot scenarios respectively.

Evaluation on COCO. The results are presented in Tab. 2. In general, our method achieves a new SOTA performance on novel classes of 25.60% , 28.05% and 32.38% mIoU for 1-shot, 2-shot, and 5-shot scenarios respectively. Although methods like PIFS [1], OINet [3], and CaLNet [4] enhance novel class learning with refined prototypes or textual knowledge, they remain limited in representing pixel features for novel classes. In contrast, our method customizes contextual information per image, effectively improving novel class representation.

4.4. Ablation Studies

Table 3. Ablation on Multi-level Prompts.

SA	TP	SP	RP	1-shot Base	1-shot Novel	1-shot HM
✓				65.80	47.32	55.05
	✓	✓		72.82	48.45	58.18
	✓		✓	63.26	47.82	54.46
		✓	✓	71.08	46.70	56.36
	✓	✓	✓	**73.04**	**49.08**	**58.71**

Table 4. Ablation on the Framework.

method	1-shot Base	1-shot Novel	1-shot HM
Ours	**73.04**	**49.08**	**58.71**
- FDD	72.86	44.89	55.55
- $\mathcal{L}_{ba}$	70.53	48.73	57.63
- CMQE	70.50	22.20	33.76

ICCV
#7277

Ablation on different prompts. We simply add visual prompts (SA) of each layer of the proposed framework to learn novel classes, which is regarded as our baseline as in Eq. (2). We perform ablations on SA and the proposed prompts. To ensure fairness, each experiment uses the same total number of prompts. Tab. 3 reports results on the 1-shot VOC benchmark. The baseline achieves 65.80% mIoU-B and 47.32% mIoU-N by expanding with a fixed number of prompts, yielding competitive results. Adding TP and SP improves performance of both base and novel classes. TP conveys universal knowledge across stages, while SP provides discriminative representations of the current stage and inherits relevant knowledge from previous ones, enhancing learning ability. With finer-grained RP, novel class performance increases to 49.08% mIoU-N, highlighting the benefits of multi-granular contextual information. Additionally, using only SP and RP leads to a performance decline, as SP may weaken generalizable knowledge from previous stages when adapting to new ones. Introducing TP preserves this knowledge from the data-rich base stage, improving overall performance. We further prove MSVP shows stronger learning capability for novel classes with more training samples in Appendix B.1.

Ablation on the framework. We validate the importance of each component of the framework by removing them one at a time, as in Tab. 4. It can be observed that replacing FDD with a single head markedly reduces performance, particularly on novel classes. That's because FDD decompose semantic segmentation into foreground refinement and background isolation, making the model focus on the former and reducing background interference. $\mathcal{L}_{ba}$ outperforms $\mathcal{L}_{van}$ by 2.51% mIoU on base classes, for the reason that $\mathcal{L}_{van}$ optimizes the two separators independently causing misaligned optimization directions. Besides, as CMQE generates queries with high generalization ability, it enhances the model's capacity to expand, with an improvement of 26.88% mIoU on novel classes.

Ablation on the number of task-persistent prompts L_g. We conduct an experiment on the number of task-persistent prompts as shown in Fig. 4 (a). The graph shows that the performance for both base and novel classes initially increases, peaking at 24 prompts, and then declines as the prompt count increases further. This trend occurs because few prompts provides insufficient shared knowledge, making it harder to prompt the model effectively. Conversely, using too many prompts causes the model to overfit on the base classes, which restricts its flexibility and reduces its capacity to generalize to novel classes.

Ablation on the number of vectors of per stage-specific prompt L_s. We conduct an experiment on the number of vectors of per stage-specific prompt as shown in Fig. 4 (b). Our method achieves its best performance when L_s is 8. This is because when the number of prompts is insufficient, stage-specific knowledge is lacking. Conversely, when the number of prompts is excessive, the model is prone to overfitting a small number of novel class samples.

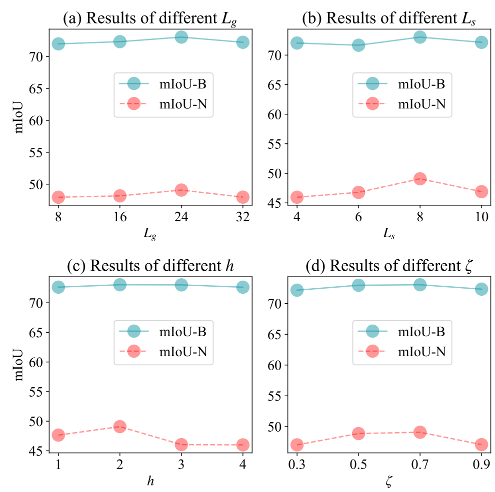

Figure 4. Ablation studies on hyperparameters.

Ablation on the cluster h and the threshold ζ. We conduct two experiments on the hyper-parameters of region-unique prompts, as shown in Fig. 4 (c) and (d). Fig. 4 (c) indicates that the novel class achieves the highest performance when h is 2. Both too few and too many clusters result in decreased performance. Specifically, a low h leads to insufficient granularity in region-unique level prompts, while a high h results in overly fine granularity, which can introduce redundant information. Fig. 4 (d) indicates that when ζ is 0.7, performance on novel classes gets best. A low ζ increases the area of the region-unique prompt, augmenting incorrect pixels, while a high ζ reduces the prompt area, resulting in too few pixels being augmented.

More ablation studies and visualizations are shown in Appendix B and C, including ablation on orthogonality constraints, query function and computation efficiency.

5. Conclusion

In this paper, we propose a novel multi-level prompt-based IFSS method that incorporates a visual prompt pool to store and switch multi-granular knowledge across different stages to enhance the incremental learning. We first design a prompt-based IFSS framework, which leverages textual semantics and visual prompts to encode foreground and background classes separately, enabling incremental semantic segmentation using prompts. Further, we introduce multi-level visual prompts with a switching mechanism to provide the model with multi-granularity contextual information tailored to the image content, thus instructing the model to learn novel classes effectively without forgetting old classes. Extensive experiments on various datasets demonstrate the effectiveness of our method.

References

[1] 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. 1, 2, 6, 7

[2] Guangchen Shi, Yirui Wu, Jun Liu, Shaohua Wan, Wenhai Wang, and Tong Lu. Incremental few-shot semantic segmentation via embedding adaptive-update and hyper-class representation. In *Proceedings of the 30th ACM international conference on multimedia*, pages 5547–5556, 2022. 2, 6

[3] Lianlei Shan, Wenzhang Zhou, Wei Li, and Xingyu Ding. Organizing background to explore latent classes for incremental few-shot semantic segmentation. *arXiv preprint arXiv:2405.19568*, 2024. 1, 2, 7

[4] Leo Shan, Wenzhang Zhou, and Grace Zhao. Incremental few shot semantic segmentation via class-agnostic mask proposal and language-driven classifier. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8561–8570, 2023. 1, 2, 6, 7

[5] Yuan Zhou, Xin Chen, Yanrong Guo, Jun Yu, Richang Hong, and Qi Tian. Advancing incremental few-shot semantic segmentation via semantic-guided relation alignment and adaptation. In *International Conference on Multimedia Modeling*, pages 244–257. Springer, 2024. 1, 7

[6] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Bulo, Elisa Ricci, and Barbara Caputo. Modeling the background for incremental learning in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9233–9242, 2020. 1, 7

[7] Fabio Cermelli, Matthieu Cord, and Arthur Douillard. Comformer: Continual learning in semantic and panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3010–3020, 2023.

[8] Chang-Bin Zhang, Jia-Wen Xiao, Xialei Liu, Ying-Cong Chen, and Ming-Ming Cheng. Representation compensation networks for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7053–7064, 2022.

[9] Umberto Michieli and Pietro Zanuttigh. Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1114–1124, 2021. 1

[10] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 139–149, 2022. 2, 3

[11] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European Conference on Computer Vision*, pages 631–648. Springer, 2022. 3

[12] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11909–11919, 2023. 3, 5

[13] Chenxi Liu, Zhenyi Wang, Tianyi Xiong, Ruibo Chen, Yihan Wu, Junfeng Guo, and Heng Huang. Few-shot class incremental learning with attention-aware self-adaptive prompt. *arXiv preprint arXiv:2403.09857*, 2024.

[14] Martin Menabue, Emanuele Frascaroli, Matteo Boschini, Enver Sangineto, Lorenzo Bonicelli, Angelo Porrello, and Simone Calderara. Semantic residual prompts for continual learning. *arXiv preprint arXiv:2403.06870*, 2024. 2, 3

[15] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2, 5, 7, 3

[16] Kaige Li, Qichuan Geng, Maoxian Wan, Xiaochun Cao, and Zhong Zhou. Context and spatial feature calibration for real-time semantic segmentation. *IEEE Transactions on Image Processing*, 32:5465–5477, 2023. 2

[17] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.

[18] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022. 4

[19] Kaige Li, Qichuan Geng, Maoxian Wan, Xiaochun Cao, and Zhong Zhou. Context and spatial feature calibration for real-time semantic segmentation. *IEEE Transactions on Image Processing*, 32:5465–5477, 2023.

[20] Kaige Li, Qichuan Geng, and Zhong Zhou. Exploring scale-aware features for real-time semantic segmentation of street scenes. *IEEE Transactions on Intelligent Transportation Systems*, 25(5):3575–3587, 2024.

[21] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[22] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning. In *European Conference on Computer Vision*, pages 709–727. Springer, 2022. 3

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 7

[24] Ziqin Zhou, Yinjie Lei, Bowen Zhang, Lingqiao Liu, and Yifan Liu. Zegclip: Towards adapting clip for zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF Con-*

ference on Computer Vision and Pattern Recognition, pages 11175–11185, 2023. 4

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[26] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 4

[27] James L McClelland, Bruce L McNaughton, and Randall C O'Reilly. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419, 1995. 4

[28] Dharshan Kumaran, Demis Hassabis, and James L McClelland. What learning systems do intelligent agents need? complementary learning systems theory updated. *Trends in cognitive sciences*, 20(7):512–534, 2016. 4

[29] Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. Groupvit: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18134–18144, 2022. 6

[30] Hang Qi, Matthew Brown, and David G Lowe. Low-shot learning with imprinted weights. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5822–5830, 2018. 7

[31] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4367–4375, 2018. 7

[32] Yonglong Tian, Yue Wang, Dilip Krishnan, Joshua B Tenenbaum, and Phillip Isola. Rethinking few-shot image classification: a good embedding is all you need? In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 266–282. Springer, 2020. 7

[33] Mennatullah Siam, Boris N Oreshkin, and Martin Jagersand. Amp: Adaptive masked proxies for few-shot segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5249–5258, 2019. 7

[34] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8256–8265, 2019. 7

[35] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017. 7

[36] Umberto Michieli and Pietro Zanuttigh. Incremental learning techniques for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, pages 0–0, 2019. 7

[37] 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. 7

[38] Michael Hersche, Geethan Karunaratne, Giovanni Cherubini, Luca Benini, Abu Sebastian, and Abbas Rahimi. Constrained few-shot class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9057–9067, 2022. 7

[39] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9046–9056, 2022. 7

[40] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Anal. Stat. Model. Comput. Learn., Tech. Rep*, 2007(1-45):5, 2012. 6

[41] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 6

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 6

[43] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 7, 1

[44] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 3

[45] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 3

Incremental Few-Shot Semantic Segmentation via Multi-Level Switchable Visual Prompts

Supplementary Material

A. Implementation Details

In this section, we first elaborate on the process of region-unique prompt switching using `Gumbel_Softmax` and the design of the keys $\mathbf{K}_l$ of region-unique prompts. Secondly, we outline the settings of the proposed method, including the optimizer, learning rate, hyper-parameters, to ensure the reproducibility of the method. Thirdly, we describe the overall data flow and model structure. Finally, we explain the ways to perform incremental learning to expand the model. Due to the double-blind principle, **we will open source the code at our github after the review.**

A.1. Detailed Region-unique Prompt Switching Process

To enable end-to-end training, we employ a `Gumbel_Softmax` operation to switch the appropriate region-unique prompts. We first calculate the similarity between q_v and the keys of region-unique prompts, as

$$d_{k,j} = \frac{exp(q_{v,k}\mathbf{K}_{l,j} + \epsilon_k)}{\sum_{m=1}^{N} exp(q_{v,k}\mathbf{K}_{l,m} + \epsilon_m)} \tag{A1}$$

where $\{\epsilon_k\}$ are i.i.d random samples drawn from the `Gumbel(0,1)` distribution. We compute the region-unique prompt to assign a centroid to by taking the one-hot operation of it `argmax` over all the keys. Since the one-hot assignment operation via `argmax` is not differentiable, we instead use the straight through trick in to compute the assignment matrix as

$$\hat{d} = \texttt{one-hot}(d_{\texttt{argmax}}) + d - \texttt{sg}(d) \tag{A2}$$

where `sg` is the stop gradient operator. With the straight through trick, $\hat{d}$ has the one-hot value of assignment to a single region-unique prompt, but its gradient is equal to the gradient of d, which makes the whole procedure differentiable and end-to-end trainable. After assigning q_v to keys of region-unique prompts, we can easily get the region-unique prompts response to q_v by merging all prompts, as:

$$p_{l,k}^{(i)} = \frac{\sum_{j=1}^{N} \hat{d}_{k,j}\mathbf{L}_j^{(i)}}{\sum_{j=1}^{N} \hat{d}_{k,j}}. \tag{A3}$$

This approach effectively solves the problem of gradient backpropagation by transforming the argmax process into a discrete variable sampling process.

Table A1. Ablation on the design of keys of region-unique prompts.

Method	1-shot		
	Base	Novel	HM
Rand	72.14	47.53	57.30
CoOP	73.04	49.08	58.71

A.2. The Design of the Keys of Region-unique prompts

As described in Sec. 4.2, to ensure stable training and provide meaningful initial keys for region-unique prompts, we leverage text embeddings generated by the CLIP text encoder following the CoOP[43] approach. Specifically, each region-unique prompt, which corresponds to a particular class, is assigned a key that aids in aligning the local features of input images. For each key, we use the CLIP text encoder to produce stable and meaningful embeddings by feeding it a prompt $\sigma = [V]_1[V]_2...[V]_n[CLASS]$, where $[V]_i$ represents vectors with the same dimension as word embeddings, n is a hyperparameter specifying the number of context tokens, and $[CLASS]$ denotes the class name's word embedding. By processing the prompt σ through the text encoder, we obtain a key tailored to each region-unique prompt. This method provides an effective initial value for the keys, mitigating the convergence issues often caused by random initialization, especially in few-shot scenarios. Moreover, it incorporates text modality knowledge, reducing the risk of overfitting to the limited samples in few-shot novel classes.

We further carry out an experiment to compare the performance with and without the keys generated by CoOP on VOC, as shown in Tab. A1. The table shows that with this key generation method, the model's ability to learn new classes is significantly enhanced.

A.3. Settings of the Proposed Method

Table A2. Settings of different datasets.

Dataset	L_g	L_s	h	ζ
VOC	24	8	2	0.7
COCO	40	16	4	0.7

During base training, the backbone and the query function are frozen, all prompts as well as the decoder are trainable. We use AdamW as optimizer with $\beta_1 = 0.9$,

$\beta_2 = 0.9$, weight decay 0.01, and a polynomial learning rate policy with a linear learning rate warmup. The model is trained for 20k iterations on VOC and 80k iterations on COCO with a learning rate of 2×10^{-4} and batch size of 8. During incremental training, we freeze all parameters but update the expanded stage-specific prompts and region-unique prompts. For both VOC and COCO, the model is trained for 400 iterations per step with a learning rate of 2×10^{-4} without the learning rate warmup. We compute the results via single-scale full-resolution images without any post-processing. The settings of model hyper-parameters on different datasets are shown in Tab. A2.

Additionally, the pre-defined background classes are "sky", "wall", "tree", "wood", "grass", "road", "sea", "river", "mountain", "sands", "desk", "bed", "building", "cloud", "lamp", "door", "window", "wardrobe", "ceiling", "shelf", "curtain", "stair", "floor", "hill", "rail", "fence".

A.4. Detail Data Flow and Model Structure

(1) Prompts Generation: The images are input into pre-trained query function (DINOv2) to get global and local query features q_g and q_c which are used to match optimal prompts, formulated by Eq. (8), Eq. (9) and Eq. (13). (2) Image and Text Encoding: Image tokens concatenated with TP, SP and RP are input into each block of CLIP image encoder (ViT-B), which outputs the class token g and pixel features P. The names of each BG and FG classes are input into CLIP text encoder to get text embeddings T_{bg} and T_{fg}. (3) Pixel Decoding: CMQE integrates the image global feature g with text embeddings T_{fg} and T_{bg} to generate class-specific queries Q_{fg} and Q_{bg} as formulated by Eq. (3). FG isolation separator and BG refinement separator share the same structures, composed of three cross-attention blocks. For each block, pixel-wise features P are input as the keys and values. Q_{fg} and Q_{bg} are input as the queries of the first block and the output of each block is input as the queries of the next block. The last block outputs the final masks (Eq. (4)).

A.5. Incremental Learning

We elaborate on the ways in which the model is extended during the incremental stage as follow. During incremental training stage, we need to expand three components: 1) text embeddings of novel classes, 2) slots of stage-specific prompts, 3) slots of region-unique prompts.

For expanding text embeddings, we just add novel classes to foreground text embeddings, which can be denotes as $T^t_{fg} \in \mathbb{R}^{(C^{t-1}_{fg}+N_e)\times D}$, where N_e denotes the number of novel classes, C^{t-1}_{fg} denotes the number of foreground classes of stage $t-1$.

For expanding slots of stage-specific prompts, we concatenate the expanded $\mathbf{A}^{(i)}_e \in \mathbb{R}^{1\times D}, \mathbf{K}^{(i)}_e \in \mathbb{R}^{1\times D}, \mathbf{S}^{(i)}_e \in \mathbb{R}^{1\times L_s \times D}$ to matrix of stage $t-1$. Thereby, the stage-specific prompts integration of stage t can be formulated as:

$$
\begin{aligned}
\gamma =& \texttt{Softmax}(< q_g \odot [\mathbf{A}^{(i),t-1}; \mathbf{A}^{(i)}_e], \\
& [\mathbf{K}^{(i),t-1}; \mathbf{K}^{(i)}_e] > /\tau) \\
p^{(i),t}_s =& \gamma[\mathbf{S}^{(i),t-1}; \mathbf{S}^{(i)}_e] \in \mathbb{R}^{L_s\times D} \qquad \text{(A4)}
\end{aligned}
$$

where $\mathbf{A}^{(i),t-1}$,$\mathbf{K}^{(i),t-1}$,$\mathbf{S}^{(i),t-1}$ denotes the attention matrix, key matrix and stage-specific prompts of stage $t-1$ for layer i, $p^{(i),t}_s$ denotes the integrated stage-specific prompts.

For expanding slots of region-unique prompts, we concatenate the expanded $\mathbf{K}^{(i)}_{le} \in \mathbb{R}^{N_e\times D}, \mathbf{L}^{(i)}_e \in \mathbb{R}^{N_e\times D}$ with matrix of stage $t-1$. Thereby, the region-unique prompts integration of stage t can be formulated as:

$$
\begin{aligned}
p^{(i),t}_l =& \texttt{Gumbel\_Softmax}(q_v[\mathbf{K}^{(i),t-1}_l; \mathbf{K}^{(i)}_{le}]) \\
& [\mathbf{L}^{(i),t-1}; \mathbf{L}^{(i)}_e], \qquad \text{(A5)}
\end{aligned}
$$

where N_e denotes the number of novel classes, $\mathbf{K}^{(i),t-1}_l, \mathbf{L}^{(i),t-1}$ denote the all keys and prompts of stage $t-1$, and $p^{(i),t}_l$ denotes the integrated region-unique prompts.

Note that due to the adoption of a switching mechanism, the number of visual prompts input to the model remain consistent at different stages, effectively preventing a significant increase in computational complexity.

B. More Ablation Studies

In this section, we first conduct experiments to demonstrate the effectiveness of MSVP compared to the baseline, and prove the effectiveness of orthogonality constraints. Secondly, we perform an ablation study on the scale and types of query functions. Thirdly, we prove the computation efficiency of MSVP. Finally, we carry out ablation experiments on the COCO dataset to further validate the proposed method.

B.1. Effectiveness of MSVP

To further demonstrate the effectiveness of the proposed MSVP, we compare the performance under different novel shots between the baseline and MSVP. As shown in Fig. A1, the performance of the baseline model does not increase as much as the proposed model with the growth of the shot. This demonstrates that MSVP effectively boosts the model's capacity to learn novel classes by efficiently switching prompts. In contrast, the baseline model suffers from diluted information as more visual prompts are added, leading to inadequate learning of novel classes. Meanwhile, the baseline's performance on base classes significantly outperforms that of the model with MSVP, and

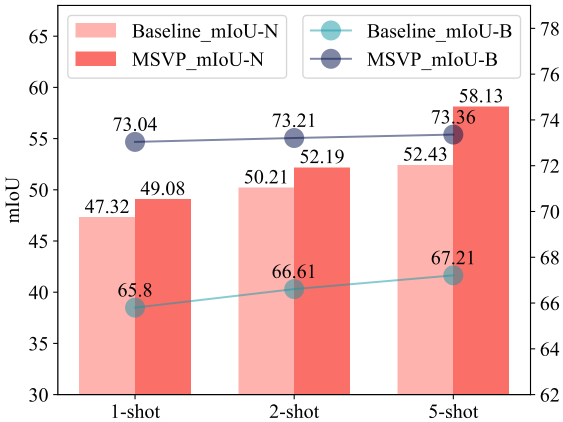

Figure A1. Performance comparison between the baseline and the proposed method on VOC under different shots.

Table A3. Ablation on $\mathcal{L}_{orth}$.

$\mathcal{L}_{orth}$	1-shot		
	Base	Novel	HM
wo	67.35	48.62	56.47
w	73.04	49.08	58.71

improves with more novel class samples. The proposed MSVP stores knowledge of different stages in independent visual prompts and switches them dynamically, thus avoiding knowledge of old stages corrupted.

B.2. Effectiveness of Orthogonality Constraints

We add orthogonality constraints to parameters of stage-specific prompts to avoid interference between existing and new knowledge and reduce catastrophic forgetting. In this section, we conduct an experiment on this loss to prove its effectiveness as in Tab. A3. The table shows that adding this loss function significantly enhances the performance of the base class by 5.69% mIoU, suggesting that it effectively minimizes interference from new knowledge on existing knowledge and helps prevent forgetting of previously learned classes.

B.3. Ablation on Query Function

We choose pretrained ViT-B of DINOv2 as the query function to produce high quality global query features and local query features. In this section, we conduct an ablation study on the scale of the query function, as shown in Tab. A4. The model's performance improves significantly when the query function transitions from ViT-S to ViT-B. However, further scaling from ViT-B to ViT-L results in minimal performance gains, indicating that the larger model size does not substantially enhance effectiveness in our method.

We also perform experiments to evaluate different types of query functions, including MAE[44], BEiT[45], and DINOv2[15], as summarized in Tab. A5. For these experiments, we use ViT-B with various pre-training methods. The results show that DINOv2 achieves the best performance, likely due to its ability to provide both global and local features with high generalizability, thereby switching appropriate prompts accurately. In contrast, MAE and BEiT yield relatively inferior results, which may stem from their limitations in effectively capturing image-specific differences across different stages. Consequently, they struggle to generate tailored prompts that align with the unique characteristics of images at various stages.

Table A4. Ablation on the scale of DINOv2.

Model	1-shot		
	Base	Novel	HM
ViT-S	71.62	48.57	57.88
ViT-B	73.04	49.08	58.71
ViT-L	72.68	48.76	58.36

Table A5. Ablation on different query functions.

Pre-train method	1-shot		
	Base	Novel	HM
MAE	50.44	39.00	43.98
BEiT	42.95	32.69	37.12
DINOv2	73.04	49.08	58.71

B.4. Ablation of Different Prompts on COCO.

We also carry out an experiment of different prompts on COCO to prove the effectiveness of our method, as shown in Tab. A6. When simply adding prompts as Eq. (2), the baseline achieves 44.98% mIoU-B and 23.72% mIoU-N, which has outperformed previous SOTA methods by a large margin. Replacing the vanilla prompt expanding strategy with the proposed task-persistent prompts and stage-specific prompts, the performance increases by 3.28% mIoU-B and 0.93% mIoU-N. That's because task-persistent prompts provide transferable knowledge across stages and stage-specific prompts extract relevant knowledge from other stages and enhance it with discriminative knowledge of the current stage, which offers a flexible way to switch knowledge of different stages, thereby achieving better abilities to keep old knowledge and learn new classes. Furthermore, with finer-grained region-unique prompts, performance on novel classes further rises to 25.60%, for the reason that region-unique prompts provide the model with knowledge of local details of specific classes. The table also shows that excluding task-persistent prompts leads to a performance decrease on both base and novel classes. It proves that general knowledge can not only help models maintain old abilities but also assist models in learning new abilities.

Table A6. Ablation of Different Prompts on COCO.

SA	TP	SP	RP	1-shot SA	1-shot Novel	1-shot HM
✓				44.98	23.72	31.06
	✓	✓		48.26	24.65	32.63
	✓		✓	41.85	24.78	31.13
		✓	✓	44.24	20.54	28.05
	✓	✓	✓	**48.85**	**25.60**	**33.59**

Table A7. Ablation of the Framework on COCO.

Method	1-shot Base	1-shot Novel	1-shot HM
Ours	**48.85**	**25.60**	**33.59**
- FDD	48.18	23.13	31.25
- $\mathcal{L}_{ba}$	45.81	21.92	29.65
- CMQE	44.72	16.57	24.18

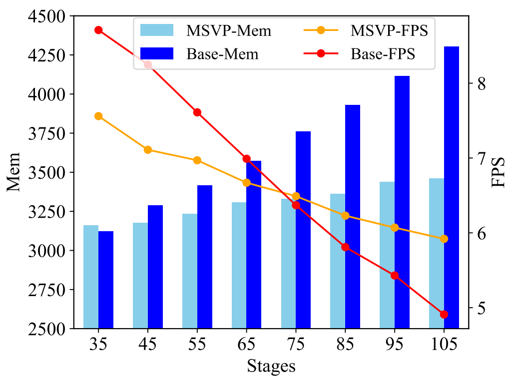

Figure A2. Comparison of MSVP and Base on FPS and Memory.

B.5. Ablation of the Framework on COCO.

We also carry out an experiment of the framework on COCO to prove the effectiveness of our method, as shown in Tab. A6. Replacing FDD with a single decoder results in a decrease in performance on novel classes. That's because FDD enables the model to process the salient features of novel classes and adapt to background changes separately, alleviating the confusion between novel classes and the background. Additionally, jointly optimizing the two separators with $\mathcal{L}_{ba}$ leads to a better performance, which mitigates the misaligned optimization directions caused by $\mathcal{L}_{van}$. It can also be concluded from the table that CMQE plays an important role in maintaining base knowledge and learning novel classes by generating generalizable class embeddings for base classes and novel classes.

B.6. Computation Efficiency.

We conduct an experiment to compare FPS and memory usage between MSVP and Baseline as the incremental phase increases, on an NVIDIA A40 GPU. MSVP exhibits significant advantages in memory efficiency compared to the baseline. This improvement stems from our prompt-switching mechanism, which effectively maintains a constant input sequence length to the transformer model throughout progressive training stages, thereby avoiding memory accumulation. While the baseline shows marginally higher FPS during early incremental stages (<75 stages) due to the introduced query operation of MSVP, our method demonstrates superior computational sustainability as training progresses. Notably, the baseline suffers a sharp FPS degradation as its growing prompt inventory quadratically increases transformer's computational complexity ($O(n^2)$).

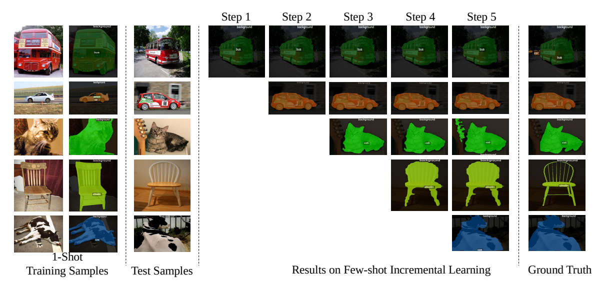

Figure A3. Step-by-step segmentation results of our method. Zoom in for better visualization.

C. Visualization

In Fig. A3, we visualize our step-by-step segmentation results for novel classes. The figure shows that our method effectively retains old class knowledge, enabling the model, even after multiple training rounds, to correctly predict old class samples. Additionally, our approach demonstrates strong novel class learning capability, as it can generalize to other test samples by learning from just one novel class sample. The visualization results effectively demonstrate the validity of the proposed prompt-based IFSS method.