

VISTA-CLIP: Visual Incremental Self-Tuned Adaptation for Efficient Continual Panoptic Segmentation

Manjunath D^{*1}, Shrikar Madhu^{*2}, Aniruddh Sikdar^{*3}, Suresh Sundaram^{1,3}

¹Department of Aerospace Engineering, Indian Institute of Science, Bengaluru, India

²Kotak IISc AI-ML Centre (KIAC), Indian Institute of Science, Bengaluru, India

³Robert Bosch Centre for Cyber Physical Systems, Indian Institute of Science, Bengaluru, India

{manjunathdl, shrikarm, aniruddhss, vssuresh} @iisc.ac.in

Abstract

Continual learning seeks to mimic human adaptability by progressively acquiring new knowledge without forgetting previous knowledge. Although deep neural networks perform well on static tasks, they are prone to catastrophic forgetting, where learning new classes significantly degrades performance on earlier ones. While most research has focused on classification, detection, and semantic segmentation, comprehensive continual learning frameworks for panoptic segmentation remain largely unexplored. To bridge this gap, we propose VISTA-CLIP (Visual Incremental Self-Tuned Adaptation) that balances plasticity and stability. VISTA-CLIP introduces three key components: (1) Feature Space Exploratory Learnable Perturbations (FLEX), for enhancing feature generalization via learnable noise injection, (2) Visual prompt learning applied at the input level, enabling dynamic adaptation to novel classes without altering core network parameters, and (3) textual feature integration from the CLIP text encoder into class-specific prompt embeddings to improve plasticity. Extensive experiments validate the effectiveness of VISTA-CLIP, delivering an average performance gain of 1.2% across all tasks and a notable 2.5% improvement in the challenging 50-50 split (50 base classes followed by 50 incremental). These results position VISTA-CLIP as a scalable and efficient solution for continual panoptic segmentation, pushing the boundaries of lifelong learning in dynamic, real-world settings

1. Introduction

Image segmentation is a fundamental computer vision task that involves dividing an image into meaningful segments, with various segmentation tasks defined based on segment definitions [24–26]. Panoptic segmentation is one of the most sophisticated approaches to image segmentation, com-

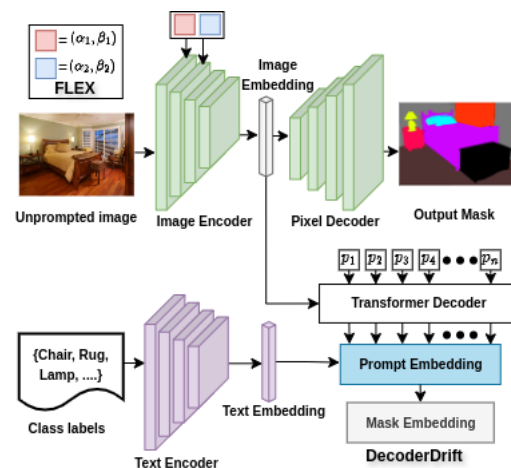


Figure 1. The green blocks represent the encoder-decoder structure that is trainable during base class training. FLEX is introduced to first two stages of the image encoder. It has two learnable parameters α and β for each stage that regulate the perturbation levels of channel statistics. The CLIP text encoder of DecoderDrift that generates text embedding of same shape with that of prompt embedding. The text embeddings are added with the prompt embeddings to enrich the prompt information during novel class training.

binning semantic segmentation [30] with instance segmentation (identifying individual objects within categories) by assigning each pixel a semantic class while distinguishing different instances into separate segments. Recent progress in panoptic segmentation has been driven by convolutional neural networks [7, 17] and transformer-based architectures [2, 9, 10].

Conventional approaches, trained on static datasets with fixed class definitions [14, 27–29], often necessitate complete retraining on both existing and newly introduced classes to accommodate updates. A more computationally efficient alternative involves selectively fine-tuning the pre-

^{*}Equal contribution.

trained model using only the new class data, thereby avoiding redundant training overhead. However, this approach often results in catastrophic forgetting, significantly impairing performance on previously learned classes [19]. Ideally, a model should continuously learn new classes while retaining knowledge of previously learned ones, without requiring full retraining [23]. The dynamic nature of the real world necessitates models that not only grasp the present but also adapt and evolve over time. Continual image segmentation [3, 20, 21] addresses this need by enabling models to learn new classes incrementally over time without forgetting the old classes. However, simultaneously retaining knowledge of previously learned classes (mitigating catastrophic forgetting [10]) while effectively incorporating new class information (ensuring plasticity) remains a significant challenge.

Continual Semantic Segmentation (CSS) addresses the challenge of incrementally updating a model’s knowledge over time, a problem that has garnered increasing attention within the segmentation community [4, 11, 33]. The interest in CSS arises from its practical relevance and the unique challenges it poses, such as background shift [3]. Unlike standard segmentation, CSS defines background relative to current classes only, causing it to include both past and future classes. Semantic ambiguity can lead the model to misclassify old class pixels as background, exacerbating catastrophic forgetting [3]. To address this, most continual segmentation approaches employ techniques like knowledge distillation [3], [11], [33] and pseudo-labeling [6], [11], [22]. These challenges are often rooted in semantic confusion, where visually similar objects are misclassified, an issue intensified in continual learning due to the model’s lack of awareness of future classes. While knowledge distillation helps mitigate catastrophic forgetting by transferring knowledge from the old model to the new one, and pseudo-labeling enables learning of previously seen classes using inferred labels, both approaches come with trade-offs. These methods typically involve extra network passes and demand careful hyperparameter tuning, such as setting appropriate distillation loss weights and pseudo-label thresholds, which adds to the overall training complexity and computational cost [8]. As classes are incrementally introduced, distillation-based methods face growing challenges in scalability and efficiency. Moreover, most state-of-the-art approaches are narrowly designed for semantic segmentation, limiting their adaptability to broader segmentation tasks. ECLIPSE [16], for example, targets continual semantic segmentation but is constrained by its exclusive use of decoder-level prompts for novel classes. It omits prompt integration in the image space and fails to leverage textual encodings, leaving the issues of plasticity and catastrophic forgetting insufficiently addressed.

To overcome these limitations, we propose VISTA-CLIP

(Visual Incremental Self-Tuned Adaptation for Efficient Continual Panoptic Segmentation), a unified framework designed to tackle the core challenges of continual learning, i.e., catastrophic forgetting (rigidity) and limited plasticity, in panoptic segmentation. VISTA-CLIP integrates three synergistic components: Feature Space Exploratory Learnable Perturbations (FLEX), DecoderDrift, and a Visual Prompting strategy [1], collectively addressing both plasticity and forgetting. FLEX is applied solely to the shallow layers of the image encoder during base class training, where it perturbs the channel-wise feature statistics, namely, the mean and standard deviation. This is achieved through two trainable parameters, α and β , which modulate the extent of perturbation, promoting more resilient feature representations, as shown in the Fig. 1. Despite the limited performance of CLIP-based encoders in panoptic segmentation, we propose the DecoderDrift strategy, which introduces text-based perturbations from the CLIP text encoder into class-specific prompt embeddings of the transformer decoder during novel class training. To further enhance model plasticity, visual prompting is also employed during this stage. Our approach incurs only a minimal increase in trainable parameters over the state-of-the-art network [16], yet consistently delivers significantly better performance across all evaluation settings. Our key contributions are as follows:

- We introduce VISTA-CLIP, a framework designed for continual panoptic segmentation that effectively mitigates catastrophic forgetting while enhancing scalability. Our approach integrates three core components: FLEX, DecoderDrift, and visual prompting.
- Feature Space Exploratory Learnable Perturbations (FLEX) introduces two trainable parameters within the image encoder to modulate the perturbation levels of channel statistics during base class training.
- During novel class training, we propose the DecoderDrift strategy alongside visual prompt learning. DecoderDrift leverages the CLIP text encoder to inject class-specific text-based perturbations into the output of the transformer decoder, thereby improving the model’s adaptability to new classes.
- Experiments on the ADE20K dataset across various settings demonstrate that VISTA-CLIP outperforms state-of-the-art models while introducing only a minimal increase in the number of trainable parameters.

2. Related Works

Panoptic Segmentation It is an emerging and effective task that combines both semantic and instance segmentation, offering a holistic understanding of a scene. It analyzes stuff (amorphous regions such as sky or road) and things (countable objects like cars or pedestrians) within a unified framework. Several prior works have employed panoptic seg-

mentation such as [8], [5], [11]. In recent years, there have been notable advancements in panoptic segmentation using transformers, particularly for urban scene parsing, which demands complex scene understanding such as [31], [32], [10], [15]. This is mainly due to the architectural designs that include self and cross attention mechanisms which facilitate processing the multiscale features thereby helping in achieve state-of-the-art performance in semantic, instance and panoptic segmentation. Although significant advancements have been achieved, panoptic segmentation models still lag behind the performance of standalone instance and semantic segmentation models, preventing them from fully realizing their potential as a unified approach.

Continual learning for segmentation Continual learning for panoptic segmentation has emerged as a more challenging and underexplored task. Continual learning for semantic segmentation has predominantly utilized knowledge distillation techniques, such as [5] [8] [11]. To address the growing demands of real-world applications, continual segmentation has evolved into a more sophisticated task. Distillation-based methods commonly leverage pseudo-labeling strategy to alleviate catastrophic forgetting and minimize semantic drift. A prior approach for panoptic segmentation using transformers was introduced by Mask2Former [8]. This method presented masked attention, a modified form of cross-attention that focuses on extracting localized features within the foreground area of each predicted mask. Additionally, it incorporated a multi-scale processing strategy to enhance small object predictions. Recent research work ECLIPSE [16] has emphasized the need for efficient networks that can manage continual learning without substantially increasing the number of trainable parameters. ECLIPSE had successfully integrated VPT with Mask2Former to mitigate catastrophic forgetting by using unique queries and MLPs for each classes. Notably, the effect of logit manipulation can be evidently seen to handle object and no-object prediction challenges to prevent error propagation and semantic drift. Despite the progress made by [16], continual panoptic segmentation still lags behind in performance and the architecture is unexplored to its full potential. Previous works including the state-of-the-art method like [8], [10], [16] heavily focused on catastrophic forgetting / rigidity as observed in the results and there are very few work that tries to balance rigidity and plasticity problems in the same work. Our method demonstrates a more effective balance between plasticity and rigidity, achieving this with only a minor increase in the number of trainable parameters.

Prompt Learning It originates from the prompting strategies used in large language models within Natural Language Processing (NLP). This concept has since been extended to fuse text and image modalities, enabling better feature alignment and improved performance [18]. Vi-

sual Prompting (VP) and Visual Prompt Tuning (VPT) are two key approaches centered on image-based prompting. VP refines prompts directly in the pixel space serving as Parameter Efficient Fine Tuning (PEFT) strategy, whereas VPT incorporates learnable tokens into Vision Transformer (ViT) architectures for prompt tuning. While these methods demonstrate the potential of visual-only prompting, they typically fall short compared to alternative techniques like linear probing. Recent work on visual prompting was proposed for domain adaptation [13] has proposed severity aware visual prompt tuning with image level severity metric to measure the severity differences among each image.

3. Preliminaries

We review ECLIPSE [16] in this section, a framework designed for continual panoptic segmentation. We adopt ECLIPSE [16] as our baseline architecture, an efficient, knowledge-distillation-free method for continual panoptic segmentation.

ECLIPSE [16] is built upon Mask2Former [8], a transformer-based framework designed for universal segmentation tasks encompassing semantic, instance, and panoptic segmentation. Mask2Former performs mask prediction through a process known as mask classification, where each predicted mask is associated with a specific class label. ECLIPSE has three modules: image encoder M_{enc} , pixel decoder and transformer decoder. The image encoder and pixel decoder together generates the masks for all objects. The transformer decoder takes the image embedding from image encoder with N learnable queries, it generates N mask embeddings using self and cross attention with the image embeddings. The initial training has all the parameters trainable on the set of base classes:

$$(s^1, m^1) = M(x, Q^1) \quad (1)$$

where, x represents the input, M is the model, Q denotes task-specific queries, s refers to the class logits, and m indicates the predicted output masks. After base training, ECLIPSE adopts freeze and fine-tuning approach repeatedly. When new classes are introduced, all existing parameters are frozen, and only a new set of learnable prompt embeddings Q_t and separate MLP layers in the transformer decoder are fine-tuned to recognize the new classes. This is denoted by,

$$M(x, Q^t) = MLP^t(M_{trans}(\varepsilon_{img}, Q^t)) \otimes \varepsilon_{pixel} \quad (2)$$

$$\varepsilon_{img} = M_{enc}(x), \varepsilon_{pixel} = M_{pixel}(\varepsilon_{img}) \quad (3)$$

where, ε_{img} represents the image embedding, derived through the image encoder M_{enc} , pixel decoder M_{pixel} , and transformer decoder M_{trans} . Each set of prompts Q_t is treated as a distinct task-specific module, dedicated exclusively to recognizing the corresponding set of classes

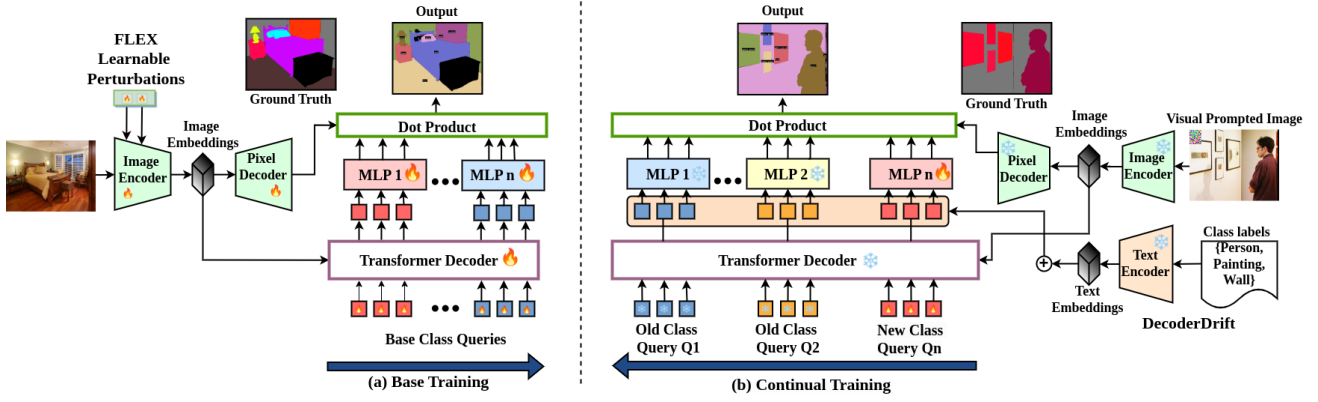


Figure 2. **Overview of VISTA-CLIP** During base class training, VISTA-CLIP applies FLEX to perturb shallow image encoder features, enhancing generalization through learnable feature modulation. For novel class training, it leverages DecoderDrift with CLIP text encoder to introduce class-specific textual perturbations, while visual prompts guide the model’s adaptation.

C_t . During inference, outputs from all prompt sets $Q_{1:t}$ are aggregated, enabling the model to segment all previously learned classes, given by,

$$(s^{1:t}, m^{1:t}) = M(x, Q^{1:t}) \quad (4)$$

To mitigate semantic confusion and representation drift, ECLIPSE introduces a simple yet effective strategy known as logit manipulation. This inference-only technique employs sigmoid activation with independent logits, enabling the model to better exploit the distinct characteristics of each class. For more details, please refer to [16].

4. Method

4.1. Problem Formulation

This work tackles the continual learning setting for the advanced task of panoptic segmentation, which unifies both semantic and instance-level segmentation. We adopt the same experimental setup as in [16][5]. At each discrete time step $t = 1, \dots, T$, a new training dataset D_t is introduced, consisting of image-label pairs (x_t, y_t) , where x_t is an input image and y_t represents its associated segmentation label. At each time step t , the labels y_t are provided only for the currently available classes C_t . Classes from previous tasks ($C_{1:t-1}$) and future tasks ($C_{t+1:T}$) are inaccessible. During inference, the model is required to perform segmentation for all classes in $C_{1:t}$, while addressing two key challenges: avoiding catastrophic forgetting of previously learned classes ($C_{1:t-1}$) and incrementally learning new classes in C_t to maintain adaptability (plasticity).

4.2. VISTA-CLIP

Our work builds upon the ECLIPSE architecture [16] for panoptic segmentation. ECLIPSE adopts a freeze-and-fine-tune strategy, leveraging prompt tuning to address continual learning more efficiently. However, its primary focus

is on mitigating catastrophic forgetting, with limited consideration for model plasticity. To address this limitation, we propose VISTA-CLIP, which introduces only a minimal increase in trainable parameters (from 0.601M to 0.608M), yet achieves a significant performance improvement over ECLIPSE.

4.2.1. Generalization using Feature Space Exploratory Learnable Perturbations (FLEX)

We introduce a novel approach, FLEX, aimed at enhancing model robustness to domain shifts in incremental learning settings. Inspired by NP [12], originally proposed for domain generalization, FLEX uses feature channel statistics, specifically the mean and standard deviation from shallow layers, to address the issue of domain style overfitting. This challenge arises when domain styles show limited variation, yet the contextual information differs greatly, a typical trait in robust segmentation scenarios. Let the image encoder be denoted as M_{enc} . In the FLEX approach, we introduce two trainable parameters, α and β , which regulate the perturbation levels of the channel statistics, specifically the standard deviation σ_c and mean μ_c , in the first two layers of M_{enc} . FLEX mechanism can be mathematically represented as,

$$y = \sigma_s^* \frac{x - \mu_c}{\sigma_c} + \mu_s^* \quad (5)$$

where, $\sigma_s^* = \alpha \sigma_c$ and $\mu_s^* = \beta \mu_c$ represent the perturbed channel-level statistics, where $\{\sigma_c, \mu_c\}$, are the original standard deviation and mean, respectively. The factors $\{\alpha, \beta\}$ are random noise terms, and y denotes the normalized feature map. This can be further simplified as:

$$y = \alpha x + (\beta - \alpha) \mu_c \quad (6)$$

This enables the model to learn domain-invariant representations that enables base classes to generalize better. Notably, FLEX is applied only during the base class training

Method	Trainable Params	KD	100-5 (11 tasks)			100-10 (6 tasks)			100-50 (2 tasks)		
			1-100	101-150	all	1-100	101-150	all	1-100	101-150	all
FT	44.9M	✗	0.0	25.8	8.6	0.0	2.9	1.0	0.0	1.3	0.4
MiB [3]	44.9M	✓	24.0	6.5	18.1	27.1	10.0	21.4	35.1	19.3	29.8
PLOP [11]	44.9M	✓	28.1	15.7	24.0	30.5	17.5	26.1	40.2	22.4	34.3
CoMFormer [5]	44.9M	✓	34.4	15.9	28.2	36.0	17.1	29.7	40.2	23.5	34.6
ECLIPSE [16]	0.60M	✗	39.5	14.7	31.2	41.2	16.5	33	39.3	21.1	33.6
Ours	0.68M	✗	41.5	15.4	32.8	42.1	17.2	33.8	41.9	21.7	35.1
joint			43.2	32.1	39.5	43.2	32.1	39.5	43.2	32.1	39.5

Table 1. **Continual Panoptic Segmentation** results on ADE20K dataset in PQ when the number of base classes $|C^1|$ is (a) 100 and (b) 50. *KD* denotes using distillation strategies, which demands more trainable parameters and computational overhead. All methods use the same network of Mask2Former with ResNet-50 backbone. *joint* means an oracle setting training all classes offline at once.

phase ($t=1$), where all model components, including the image encoder, pixel decoder, transformer decoder with base prompts, and MLPs, are trainable, as illustrated in Fig. 2. FLEX is not used during inference.

4.2.2. Visual Prompting

We adapt Visual Prompts (VP) [1] for continual learning in panoptic segmentation, modifying it to align with the pre-processing pipeline used in ECLIPSE. In VP, the prompt is directly added to the input region, resulting in a prompted image. During training, gradient updates are applied solely to the prompt parameters and model parameters, while the input image itself remains unchanged. ECLIPSE applies a preprocessing strategy that dynamically scales images with upper-left alignment, followed by zero-padding to meet the required input resolution. In our approach, we introduce trainable visual prompts by padding the upper-left corner of input images with prompts of fixed size 30. This ensures the model consistently receives a fixed prompt patch, regardless of image scaling. Mathematically this is represented as modifying the image as a pre-processing step of appending prompts on the image as:

$$x = x + p \quad (7)$$

where $x \in \mathbb{R}^{3 \times H \times W}$ and $p \in \mathbb{R}^{3 \times 30 \times 30}$. This prompting strategy is applied exclusively during training on novel classes. It has proven effective in enhancing the model’s plasticity, enabling it to learn novel classes more efficiently within the limited training iterations available, while introducing only a minimal increase in trainable parameters relative to prior methods.

4.2.3. DecoderDrift

In realistic continual learning scenarios, novel classes generally have a limited number of annotated samples, and the training iterations are deliberately constrained to prevent overfitting. To address this challenge, we introduce text-based perturbations using CLIP-text encoder M_{text} , into the prompt embeddings of the transformer decoder, targeting the class-specific prompts encountered at each stage of

continual learning, as illustrated in Fig. 2. We refer to this strategy as DecoderDrift, which reflects the intentional deviation of the decoder’s prompt-driven representations. In ECLIPSE, each set of prompts Q_t functions as a distinct task-specific module, dedicated exclusively to recognizing the class set C_t . The transformer decoder receives image embeddings alongside the corresponding learnable prompt embeddings Q_t . Text perturbations ε_{text} are applied to the transformer decoder’s output, given by,

$$M(x, Q^t) = MLP^t(M_{trans}(\varepsilon_{img}, Q^t) + \varepsilon_{text}) \otimes \varepsilon_{pixel} \quad (8)$$

These tokens are fed into the MLPs, which compute a dot product with the output of the pixel decoder, as illustrated in Fig. 2.

5. Experiment

5.1. Experimental Setting.

We compare our method against recent approaches for continual panoptic segmentation, including PLOT, MiB, CoMFormer, and ECLIPSE. Experiments are conducted on the ADE20K dataset, which comprises 150 classes—100 ‘thing’ and 50 ‘stuff’ categories. This dataset encompasses a diverse range of indoor and outdoor scenes, with an average of 9.9 classes and 19.5 instances per image. We evaluate all methods on continual panoptic segmentation tasks using the Panoptic Quality (PQ) metric, which is defined as the product of Recognition Quality (RQ) and Segmentation Quality (SQ) as follows:

$$PQ = \frac{\sum_{(p,g) \in TP} \text{IoU}(p,g)}{|TP|} \times \frac{|TP|}{|TP| + \frac{1}{2}|FP| + \frac{1}{2}|FN|}$$

where $\text{IoU}(p,g)$ is the intersection-over-union between the predicted mask(p) and the ground truth(g). And TP, FP, FN represents the true-positive, false-positive, false-negative respectively. We report the accuracy for base class, novel class and all classes after the last continual step as followed by [16].

Method	Trainable Params	KD	50-10 (11 tasks)			50-50 (3 tasks)		
			1-50	51-150	all	1-50	51-150	all
FT	44.9M	✗	0.0	1.7	1.1	0.0	12.0	8.1
MiB [3]	44.9M	✓	34.9	7.7	16.8	42.4	15.5	24.4
PLOP [11]	44.9M	✓	39.9	15.0	23.3	45.8	18.7	27.7
CoMFormer [5]	44.9M	✓	38.5	15.6	23.2	45.0	19.3	27.9
ECLIPSE [16]	0.60M	✗	48.9	15.2	26.5	46.8	16.5	27.3
Ours	0.68M	✗	48.5	17.8	28.0	48.0	20.4	29.8
joint			50.2	34.1	39.5	50.2	34.1	39.5

Table 2. **Continual Panoptic Segmentation** results on ADE20K dataset in PQ when the number of base classes $|C^1|$ is (a) 100 and (b) 50. *KD* denotes using distillation strategies, which demands more trainable parameters and computational overhead. All methods use the same network of Mask2Former with ResNet-50 backbone. *joint* means an oracle setting training all classes offline at once.

5.1.1. Continual Learning Protocols

The continual learning paradigm is characterized by two primary categories: *base classes* and *novel classes*, which are used to manage the balance between plasticity and rigidity during the learning process. Notably, previous studies classify continual learning into two settings: *Overlapped* and *Disjoint*, based on factors such as class distribution, annotation, and difficulty level. In the disjoint setting, the dataset at each step D_t contains samples exclusively from the current classes C_t , with no overlap between steps. Each learning step introduces entirely new and distinct classes. While this setting is straightforward, it is often criticized for being restrictive and less practical in real-world scenarios. In contrast, the overlapped setting allows data samples at any learning step D_t to contain pixels from past classes $C_{1:t-1}$, current classes C_t , and even future classes $C_{t+1:T}$. However, only the current classes C_t are annotated in the labels Y_t , while all other pixels (from past and future classes) are treated as background. This setting better reflects real-world scenarios, where previously encountered information may still appear in new data.

For instance, in Table 1, the *100-5* training setup refers to 100 base classes, typically derived from a large, labeled dataset, and 5 novel classes that are incrementally introduced. The goal is to enable the model to learn the new classes without forgetting the previously learned 100 classes. Since the ADE20K dataset contains 150 classes, the continual learning setup excludes the 100 base classes and introduces novel classes in batches of 5, resulting in a total of 11 tasks, 1 task for the base classes and 10 tasks covering the 50 novel classes (5 classes per task).

5.1.2. Implementation Details

To ensure a fair comparison, we adopt the training configurations used in ECLIPSE. We follow the experimental setup that leverages Mask2Former with a ResNet-50 backbone for the continual panoptic segmentation task. To mitigate rigidity (i.e., catastrophic forgetting) during base class training, we apply the FLEX strategy, introducing it only

FLEX	Vis-Prompt	DecoderDrift	mAP
✗	✗	✗	27.3
✓	✗	✗	29.5
✓	✓	✗	29.6
✓	✓	✓	29.8

Table 3. Ablation study of VISTA-CLIP’s core components on the ADE20K dataset under the 50-50 training split.

to the first two stages of the image encoder. To address the challenge of plasticity, we incorporate text-based feature perturbations using the CLIP text encoder M_{text} , generating embeddings of dimension 256 that are added to each class’s prompt embedding during the novel class training phase. Additionally, we reintroduce Visual Prompting by padding a 30-pixel-wide prompting patch around the input image as a preprocessing step, applied only during the training of novel classes. The CLIP text encoder is used to generate 256-dimensional text feature vectors, which are integrated into the corresponding class prompt embeddings of the transformer decoder, as illustrated in Fig. 2. For training, we use a combination of Dice loss and binary cross-entropy loss for the mask prediction, and binary cross-entropy loss for classification, following bipartite matching between predicted outputs and ground truth annotations.

5.2. Experimental Results

5.2.1. Continual Panoptic Segmentation

We benchmark our approach against recent state-of-the-art methods, including Eclipse (the only non-distillation-based baseline), several knowledge distillation techniques (MiB, PLOP, CoMFormer), and basic fine-tuning (FT). All methods are implemented using the Mask2Former architecture with ResNet-50 backbones and evaluated across six incremental learning settings, as standardized in prior works. While existing distillation-based methods rely heavily on pseudo-labeling and large parameter budgets to combat catastrophic forgetting, our framework achieves superior performance using only 1.4% of trainable parameters.



Figure 3. Visual prompt learning with varying prompt placements on the input image, showcasing visualizations of six different strategies as outlined in Table 6

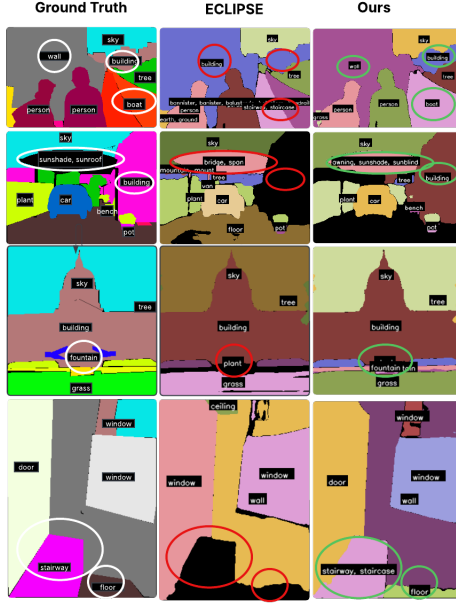


Figure 4. Qualitative examples for identifying correct class labels panoptic segmentation.

Our method leverages feature-space perturbations to enhance the robustness and plasticity of the encoder, while visual prompting—augmented with text-based prompt embedding perturbations—effectively balances the retention of prior knowledge and adaptability to new classes. Quantitative evaluations show consistent improvements across all incremental settings, supported by qualitative visualizations that highlight enhanced segmentation accuracy. These performance gains are attributed to our dual-perturbation strategy: targeted noise injection in the image encoder boosts feature resilience, and trainable text-guided prompts improve cross-modal alignment. Together, they offer a more effective solution to the plasticity–stability trade-off than parameter-intensive distillation techniques or isolated prompting-based models.

6. Ablation Studies

Impact of FLEX To assess the impact of feature-space perturbation, we conduct a qualitative comparison between ECLIPSE enhanced with NP [12], an approach that injects

random noise into the feature space to improve generalization to the target domain, and our proposed FLEX method, as presented in Table 4. It can be seen from the Table 4 while introducing vanilla NP PQ, RQ and PQ base increases but there is a significant drop in SQ score. Whereas, introducing our proposed method FLEX increases the performance across PQ, RQ and PQ base similar to NP with relatively small drop in SQ performance. Notably, all the experiments in Table 4 is carried only for base as FLEX is introduced only during base class training. As shown in Table 3, integrating FLEX with the base ECLIPSE model results in a notable performance gain, enhancing the model’s robustness while preserving its rigidity. Importantly, FLEX introduces only a minimal increase in trainable parameters from 0.6M to 0.8M, maintaining the overall efficiency of the network. This lightweight design contributes to performance improvements ranging from 0.6% to 2.2% across various settings. Figure 4 presents the qualitative results.

	ECLIPSE	with NP	with FLEX
PQ	25.2	26.6	27.1
SQ	53.3	51.6	52.3
RQ	29.6	31.3	32.0
PQ Base	37.7	39.9	40.7

Table 4. Ablation study of feature level perturbations on ECLIPSE for 100-50 setting trained only for base classes.

Effect of position and type of visual prompts Identifying the appropriate visual prompt design and placement of the prompts can influence the performance of the network significantly as discussed in [13]. We conduct an extensive ablation study on the locations of the prompts on the input image as show in Table 6 and Fig. 3, where we observe increase in the prompt size above 30 results in degradation of the models performance. Our study shows that introducing prompts around the image, specifically in the top-left corner padding, yields superior results compared to other prompt placement strategies, as shown in Table 6. In general we see an overall improvement ranging from 0.2% to 0.7% improvement across various settings.

Effect of text features on prompt embeddings While recent works have explored the integration of textual and visual features to enhance overall model performance, our

	100-5			100-10			100-50			50-10			50-50		
	Base	Novel	All	Base	Novel	All	Base	Novel	All	Base	Novel	All	Base	Novel	All
Ours + Clip Text	40.9	15.5	32.5	42.0	17.2	33.7	42.2	20.8	35.1	48.4	18.1	28.2	47.7	19.8	29.1
Ours + Rand Text	41.4	15.4	32.8	42.0	17.5	33.7	41.9	21.6	35.1	48.4	17.7	28.0	48.0	20.4	29.8

Table 5. Ablation study comparing CLIP text weight initialization with random weight initialization on ECLIPSE across various settings.

Position	100-50 (2 tasks)			
	Size	1-100	101-150	all
Four Corners	30	42.2	20.76	35.0
Padding around	30	42	19.36	34.45
Center	30	42.2	19.12	34.51
Left padding	30	42.3	20.16	34.91
Four corners	30	41.8	20.87	34.81
Top Left	30	42.3	21.16	35.28

Table 6. Ablation study on the impact of visual prompt placement at different image locations during novel class training.

method is the first to leverage an ImageNet-pretrained vision encoder in conjunction with text features embedded from CLIP-pretrained text encoder, into class-specific prompts within the transformer decoder. As shown in Table 5, we find that using a randomly initialized text encoder yields comparable, and occasionally slightly better, results than using a CLIP-pretrained text encoder. We hypothesize that this is due to the limited number of training iterations for novel classes, resulting in minimal impact from the pre-training. Our experiments, which compare the effects of random versus CLIP-derived text embeddings added to the prompt embeddings, confirm that the performance difference between the two is marginal.

7. Conclusion

In this work, we present novel strategies to address the dual challenges of catastrophic forgetting and limited plasticity in continual panoptic segmentation. Our approach integrates feature-space perturbation through visual prompting with semantic guidance via textual information incorporated into the prompt embeddings. We propose FLEX, a mechanism for injecting learnable perturbations into the image encoder during base class training, and introduce DecoderDrift, a strategy that applies text-based perturbations using the CLIP text encoder during novel class training. Additionally, visual prompts are incorporated during novel training in a manner aligned with the preprocessing pipeline. Extensive experiments demonstrate that our method not only strengthens base class representation but also enables effective adaptation to novel classes in a continual learning setup. Results on the ADE20K dataset validate the effectiveness of our approach, achieving state-of-the-art performance with a significantly reduced number of trainable parameters. As part of this work, exploring world models like CLIP could benefit the network in class-

incremental learning.

8. Acknowledgments

This work was supported by The Kotak IISc AI-ML Centre (KIAC).

References

- [1] Hyojin Bahng, Ali Jahani, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, 2022. 2, 5
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. 1
- [3] Fabio Cermelli, Massimiliano Mancini, Samuel Rota Buló, 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. 2, 5, 6
- [4] Fabio Cermelli, Antonino Geraci, Dario Fontanel, and Barbara Caputo. Modeling missing annotations for incremental learning in object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3700–3710, 2022. 2
- [5] 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. 3, 4, 5, 6
- [6] Sungmin Cha, YoungJoon Yoo, Taesup Moon, et al. Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. *Advances in neural information processing systems*, 34:10919–10930, 2021. 2
- [7] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12475–12485, 2020. 1
- [8] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 2, 3
- [9] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34:17864–17875, 2021. 1
- [10] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask

- transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 1, 2, 3
- [11] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4040–4050, 2021. 2, 3, 5, 6
- [12] Qi Fan, Mattia Segu, Yu-Wing Tai, Fisher Yu, Chi-Keung Tang, Bernt Schiele, and Dengxin Dai. Normalization perturbation: A simple domain generalization method for real-world domain shifts. *arXiv preprint arXiv:2211.04393*, 2022. 4, 7
- [13] Ziyang Gong, Fuhao Li, Yupeng Deng, Deblina Bhattacharjee, Xianzheng Ma, Xiangwei Zhu, and Zhenming Ji. Coda: Instructive chain-of-domain adaptation with severity-aware visual prompt tuning. In *European Conference on Computer Vision*, pages 130–148. Springer, 2024. 3, 7
- [14] Prajwal Gurunath, Sumanth Udupa, Aditya Gandhamal, Shrikar Madhu, Aniruddh Sikdar, Suresh Sundaram, et al. Indraeye: Infrared electro-optical uav-based perception dataset for robust downstream tasks. *arXiv preprint arXiv:2410.20953*, 2024. 1
- [15] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2989–2998, 2023. 3
- [16] Beomyoung Kim, Joonsang Yu, and Sung Ju Hwang. Eclipse: Efficient continual learning in panoptic segmentation with visual prompt tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3346–3356, 2024. 2, 3, 4, 5, 6
- [17] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6399–6408, 2019. 1
- [18] Nilakshan Kunananthaseelan, Jing Zhang, and Mehrtash Harandi. Lavip: language-grounded visual prompting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2840–2848, 2024. 3
- [19] Xiaoli Li, Savitha Ramasamy, ArulMurugan Ambikapathi, Suresh Sundaram, and Haytham M Fayek. *Towards Human Brain Inspired Lifelong Learning*. World Scientific, 2024. 2
- [20] 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. 2
- [21] Firat Ozdemir and Orcun Goksel. Extending pretrained segmentation networks with additional anatomical structures. *International journal of computer assisted radiology and surgery*, 14:1187–1195, 2019. 2
- [22] Minh Hieu Phan, Son Lam Phung, Long Tran-Thanh, Abdeslam Bouzerdoun, et al. Class similarity weighted knowledge distillation for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16866–16875, 2022. 2
- [23] Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016. 2
- [24] Aniruddh Sikdar, Sumanth Udupa, Suresh Sundaram, and Narasimhan Sundararajan. Fully complex-valued fully convolutional multi-feature fusion network (fc 2 mfn) for building segmentation of insar images. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 581–587. IEEE, 2022. 1
- [25] Aniruddh Sikdar, Jayant Teotia, and Suresh Sundaram. Contrastive learning-based spectral knowledge distillation for multi-modality and missing modality scenarios in semantic segmentation. *arXiv preprint arXiv:2312.02240*, 2023.
- [26] Aniruddh Sikdar, Sumanth Udupa, Prajwal Gurunath, and Suresh Sundaram. Deepmao: Deep multi-scale aware over-complete network for building segmentation in satellite imagery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 487–496, 2023. 1
- [27] Aniruddh Sikdar, Qiranul Saadiyeen, Prahlad Anand, and Suresh Sundaram. Ssl-rgb2ir: Semi-supervised rgb-to-ir image-to-image translation for enhancing visual task training in semantic segmentation and object detection. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5017–5023. IEEE, 2024. 1
- [28] Aniruddh Sikdar, Jayant Teotia, and Suresh Sundaram. Skd-net: Spectral-based knowledge distillation in low-light thermal imagery for robotic perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9041–9047. IEEE, 2024.
- [29] Aniruddh Sikdar, Jayant Teotia, and Suresh Sundaram. Ogp-net: Optical guidance meets pixel-level contrastive distillation for robust multi-modal and missing modality segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6922–6930, 2025. 1
- [30] Sumanth Udupa, Prajwal Gurunath, Aniruddh Sikdar, and Suresh Sundaram. Mrfp: Learning generalizable semantic segmentation from sim-2-real with multi-resolution feature perturbation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5904–5914, 2024. 1
- [31] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *European conference on computer vision*, pages 108–126. Springer, 2020. 3
- [32] Huiyu Wang, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Max-deeplab: End-to-end panoptic segmentation with mask transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5463–5474, 2021. 3
- [33] 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. 2