

SegViTv2: Exploring Efficient and Continual Semantic Segmentation with Plain Vision Transformers

Bowen Zhang*, Liyang Liu*, Minh Hieu Phan*, Zhi Tian, Chunhua Shen,
and Yifan Liu

June 13, 2023

Abstract We explore the capability of plain Vision Transformers (ViTs) for semantic segmentation using the encoder-decoder framework and introduce **SegViTv2**. In our work, we implement the decoder with the global attention mechanism inherent in ViT backbones and propose the light-weight Attention-to-Mask (ATM) module that effectively converts the global attention map into semantic masks for high-quality segmentation results. Our decoder can outperform the most commonly-used decoder UpperNet in various ViT backbones while consuming only about 5% of the computational cost. For the encoder, we address the concern of the relatively high computational cost in the ViT-based encoders and propose a *Shrunk++* structure that incorporates edge-aware query-based down-sampling (EQD) and query-based up-sampling (QU) modules. The Shrunk++ structure reduces the computational cost of the encoder by up to 50% while maintaining competitive performance. Furthermore, due to the flexibility of our ViT-based architecture, SegVit can be easily extended to semantic segmentation under the setting of continual learning, achieving nearly zero forgetting. Experiments show that our proposed SegViT outperforms recent segmentation methods on three popular benchmarks including ADE20k, COCO-Stuff-10k and PASCAL-Context

datasets. The code is available through the following link: <https://github.com/zbwpx/SegVit>.

Keywords Vision Transformer · Incremental Learning · Semantic Segmentation · Continual Learning

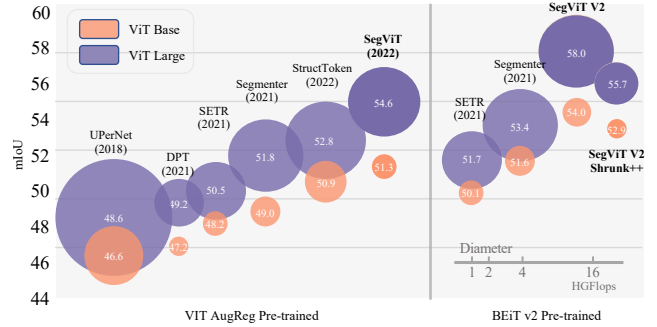


Fig. 1 Comparison with previous methods in terms of performance and efficiency on ADE20K dataset. The orange and purple bubbles in the accompanying graph represent the ViT Base and ViT Large models, respectively, with the size of each bubble corresponding to the FLOPs of the variant segmentation methods. SegViT-BEiT v2 Large achieves state-of-the-art performance with a **58.0%** mIoU on the ADE20K validation set. Additionally, our efficient, optimized version, SegViT-Shrunk-BEiT v2 Large, saves half of the GFLOPs compared to UPerNet, significantly reducing computational overhead while maintaining a competitive performance of **55.7%**.

1 Introduction

Semantic segmentation, a pivotal task in computer vision, demands precise pixel-level classification of input images. Traditional methods, such as Fully Convolutional Networks (FCN) [1], which are widely used in state-of-the-art techniques, employ deep convolutional neural networks (ConvNet) as encoders or base models

Bowen Zhang, Liyang Liu, Minh Hieu Phan, Yifan Liu
The University of Adelaide
E-mail: {b.zhang, akide.liu, vuminhhieuphan, yifan.liu04}
@adelaide.edu.au

Chunhua Shen
Zhejiang University
E-mail: chunhuashen@zju.edu.cn

Zhi Tian
Meituan Inc.
E-mail: zhi.tian@outlook.com

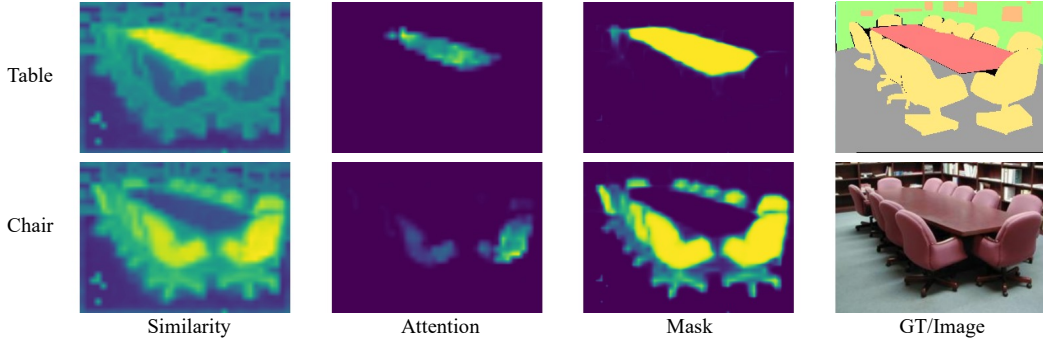


Fig. 2 The overall concept of our Attention-to-Mask decoder. ATM learns the similarity map for each category by capturing the cross-attention between the class tokens and the spatial feature map (Left). Sigmoid is applied to produce category-specific masks, highlighting the area with high similarity to the corresponding class (Middle). ATM enhances the semantic representations by encouraging the feature to be similar to the target class token and dissimilar to other tokens.

and a segmentation decoder to generate dense predictions. Prior works [2–4] have aimed to enhance performance by augmenting context information or incorporating multi-scale information, leveraging the inherent multi-scale and *hierarchical* attributes of the ConvNet architectures.

The advent of the Vision Transformer (ViT) [5] has offered a paradigm shift, serving as a robust backbone for numerous computer vision tasks. ViT, distinct from ConvNet base models, retains a plain and *non-hierarchical* architecture while preserving the resolution of the feature maps. To conveniently leverage existing segmentation decoders for dense prediction, such as U-Net [6] or DeepLab [4], recent Transformer-based approaches, including Swin Transformer [7] and PVT [8], have developed *hierarchical* ViT to extract hierarchical feature representations. However, modifying the original ViT structures requires training the networks from scratch rather than using off-the-shelf plain ViT checkpoints due to the discrepancy between the hierarchical and plain architectures, such as spatial down-sampling [9]. Changing the plain ViT structure eliminates the potential for leveraging rich representations acquired from vision-language pre-training methods such as CLIP [10], BEiT [11], BEiT-v2 [12], MVP [13], and COTS [14]. Hence, there is a clear advantage to developing effective decoders for the original ViT structures in order to leverage those powerful representations. Previous works, such as UPerNet [15] and DPT [16], have primarily focused on hierarchical feature maps and neglected the distinctive characteristics of the plain Vision Transformer. Consequently, these approaches result in computationally demanding operations with limited performance improvement, as illustrated in Figure 1. A recent trend in several works, such as SETR [17] or Segmenter [18], aims to develop decoders specifically tailored for the Plain ViT architec-

ture. However, these designs often represent a simplistic extension of per-pixel classification techniques derived from traditional convolution-based decoders. For example, SETR’s decoder [17] uses a sequence of convolutions and bilinear up-sampling to increase the ViT’s extracted feature maps gradually. It then applies a naive MLP to the extracted features to perform pixel-wise classification, which isolates the neighboring contexts surrounding the pixel. Current pixel-wise classification decoder designs overlook the importance of contextual learning when assigning labels to each pixel.

Another prevalent issue in deep networks, including Transformer, is ‘catastrophic forgetting’ [19, 20], where the model’s performance on previously learned tasks deteriorates as it learns new ones [21–24]. This limitation poses significant challenges for the application of deep segmentation models in dynamic real-world environments. Recently, the rapid development of the foundation model pre-trained on large-scale data has sparked interest among researchers in studying its transferability across various downstream tasks [25]. These models are capable of extracting powerful and generalized representations, which has led to a growing interest in exploring their extensibility to new classes and tasks while retaining the previously learned knowledge representations [26, 27].

Motivated by these challenges, this paper aims to explore how a plain vision transformer can perform semantic segmentation tasks more effectively without the need for a hierarchical backbone redesign. As self-supervision and multi-modality pre-training continue to evolve, we anticipate that the plain vision transformer will learn enhanced visual representations. Consequently, decoders for dense tasks are expected to adapt more flexibly and efficiently to these representations.

In light of these research gaps, we propose **SegViTv2** — a novel, efficient segmentation network that features a plain Vision Transformer and exhibits robustness against forgetting. We introduce a novel Attention-to-Mask (ATM) module that operates as a lightweight component for the SegViT decoder. Leveraging the non-linearity of cross-attention learning, our proposed ATM employs learnable class tokens as queries to pinpoint spatial locations that exhibit high compatibility with each class. We advocate for regions affiliated with a particular class to possess substantial similarity values that correspond to the respective class token.

As depicted in Fig. 2, the ATM generates a meaningful similarity map that accentuates regions with a strong affinity towards the 'Table' and 'Chair' categories. By simply implementing a Sigmoid operation, we can transform these similarity maps into mask-level predictions. The computation of the mask scales linearly with the number of pixels, a negligible cost that can be integrated into any backbone to bolster segmentation accuracy. Building upon this efficient ATM module, we present a novel semantic segmentation paradigm that utilizes the cost-effective structure of plain ViT, referred to as SegViT. Within this paradigm, multiple ATM modules are deployed at various layers to extract segmentation masks at different scales. The final prediction is the summation of the outputs derived from these layers.

To alleviate the computational burdens of plain Vision Transformers (ViTs), we introduce the "Shrunk" and "Shrunk++" structures, which incorporate query-based downsampling (QD) and query-based upsampling (QU). The proposed QD employs a 2x2 nearest neighbor downsampling technique to obtain a sparser token mesh, reducing the number of tokens involved in attention computations. Moreover, we extend QD to edge-aware query-based downsampling (EQD). EQD selectively preserves tokens situated at object edges, as they possess more discriminative information. Consequently, QU recovers the discarded tokens within the object's homogeneous body, reconstructing high-resolution features crucial for accurately dense prediction. By integrating the "Shrunk" structure with the ATM module as the decoder, we achieve computational reductions of up to 50% while maintaining competitive performance levels.

We further extend the application of our SegViT framework to continual learning. With the powerful and generalized representation that the foundation model has learned, this paper aims to study the ability to extend the foundation model to new classes and new tasks without forgetting the knowledge it has learned.

Recent techniques in continual semantic segmentation (CSS) aim to replay old data [28, 29] or distill knowledge from the previous model to mitigate model divergence [24, 30, 31]. These methods necessitate fine-tuning parameters responsible for old tasks, which can disrupt the previously learned solutions, leading to forgetting. In contrast, our proposed SegViT supports learning new classes without encroaching on previously acquired knowledge. We strive to establish a forget-free SegViT framework, which incorporates a new ATM module dedicated to new tasks while keeping all old parameters in a frozen state. Consequently, the proposed SegViT has the potential to virtually eliminate the issue of forgetting.

Our key contributions can be summarized as follows:

- We introduce the Attention-to-Mask (ATM) decoder module, a potent and efficient tool for semantic segmentation. For the first time, we exploit spatial information present in attention maps to generate mask predictions for each category, proposing a new paradigm for semantic segmentation.
- We present the *Shrunk* structure, applicable to any plain ViT backbone, which alleviates the intrinsically high computational expense of the *non-hierarchical* ViT while maintaining competitive performance, as illustrated in Fig. 1. We are the inaugural work capitalizing on edge information to decrease and restore tokens for efficient computation. Our *Shrunk++* version of **SegViTv2**, tested on the ADE20K dataset, achieves an mIoU of 55.7%, with a computational cost of 308.8 GFLOPs, marking a reduction of approximately 50% compared to the original SegViT (637.9 GFLOPs).
- We propose a new SegViT architecture capable of continual learning devoid of forgetting. To our knowledge, we are the first work to seek to completely freeze all parameters for old classes, thereby nearly obliterating the issue of catastrophic forgetting.

2 Related Work

Semantic Segmentation. Semantic segmentation aims to partition an image into regions with meaningful categories. Fully Convolutional Networks (FCNs) used to be the dominant approach to this task. To enlarge the receptive field, several approaches [4, 32] propose dilated convolutions or apply spatial pyramid pooling to capture contextual information at multiple scales. Most semantic segmentation methods aim to classify each pixel directly using a classification loss. This paradigm naturally partitions images into different classes.

Various methods have achieved significant advancements by integrating Transformers into the semantic segmentation task. Early works [7, 33] directly adapt the transformer encoder, designed for classification, into semantic segmentation by fine-tuning it together with segmentation decoders such as UPerNet [15]. Recent approaches [18, 34, 35] have focused on designing the overall segmentation framework to achieve better adaptation. For instance, SETR [17] views semantic segmentation as a sequence-to-sequence task and proposes a pure Transformer encoder combined with a standard convolution-based decoder. SegFormer [34] employs a hierarchical encoder design to extract features from fine-to-coarse levels and a lightweight decoder design for efficient prediction. However, the SegFormer decoder adopts the pyramid structure by fusing multi-scale features, which is specialized for hierarchical ViTs such as Swin Transformer [7]. The above-mentioned methods aim to design either a naive convolution-based decoder or a pyramid-structure decoder for hierarchical base models. Nonetheless, designing an effective decoder specialized for plain ViTs remains an open research question.

Recently, several segmentation methods propose a universal framework that unifies multiple tasks, including instance segmentation, semantic segmentation, and object detection. For example, Mask DINO [36] extends DINO with a mask prediction branch, achieving promising results in the instance, panoptic, and semantic segmentation tasks. Mask2Former [37] enhances MaskFormer [35] by introducing deformable multi-scale attention in the decoder and a masked cross-attention mechanism. OneFormer [38] represents a universal image segmentation framework with a multi-task train-once design, outperforming specialized models in various tasks.

Recent methods [18, 35, 39] propose decoupling the per-pixel classification into image partitioning and region classification. For image partitioning, they use learnable tokens as mask embeddings and associate them with the extracted feature map to generate object masks. For region classification, the learnable tokens are fed to a classifier to predict the class corresponding to each mask. This paradigm allows for global segmentation and alleviates the burden on the decoder to perform per-pixel classification, resulting in state-of-the-art performance [35]. While previous works use generic tokens for mask generation, this work explicitly utilizes class-specific tokens to enhance the semantics of mask embeddings, thereby improving segmentation accuracy.

Mask-oriented Segmentation. Compared to previous Mask-oriented Segmentation methods such as MaskFormer [40] and Mask2Former [37], our method presents several novel conceptual differences and advantages. Specifically, our approach is tailored to address semantic segmentation problems by assigning each class to a fixed token and generating the corresponding mask directly. In contrast, MaskFormer relies on Hungarian matching, with each learnable query corresponding to spatial information instead of category information. Our Attention-to-Mask (ATM) approach eliminates the need for positional embedding, as we utilize the attention map between the class token and the feature map. Our overarching goal is to adapt Plain Vision Transformers for dense prediction, as recent studies have demonstrated that self-supervised learning [12, 41–43] and multimodal learning [10] are enhanced by hierarchical ViT structures. Our approach enhances the representation ability of class tokens by applying transformer blocks.

Previous CNN-based decoders, such as OCRNet [44] and K-Net [45], have demonstrated the effectiveness of the attention mechanism in modeling contextual information. For example, K-Net utilizes semantic kernels (one kernel for each class) and performs convolution operations to generate the semantic mask. In contrast, our proposed ATM module integrates cross-attention mechanisms, allowing for more effective contextual learning. While OCRNet [44] applies cross-attention from the class token to the feature map to enhance feature representations, it still employs a standard linear predictor in the decoder to produce the segmentation map. On the other hand, our proposed ATM module is specifically designed for generating segmentation outputs, paving the way for future research on effective decoders for plain ViT. Additionally, existing convolution-based attention networks such as OCRNet [44], K-Net [45], and DANet [46] adopt the traditional per-pixel classification framework for segmentation generation. In contrast, our proposed SegViT decouples segmentation into mask prediction and classification, which proves advantageous for establishing connections between the class proxy and language representations [47], as well as facilitating continual learning.

Transformers for Vision. Attention-based transformer models have emerged as powerful alternatives to standard convolution-based networks in the realm of image classification tasks. The original ViT [5] represents a plain, non-hierarchical architecture. However, there have been several advancements in the field of hierarchical transformers, such as PVT [8], Swin Transformer [7], TWINS [48], SegFormer [34], and

P2T [49]. These hierarchical transformer models inherit certain design elements from convolution-based networks, including hierarchical structures, pooling, and downsampling with convolutions. Consequently, they can be seamlessly employed as direct replacements for convolutional-based networks and can be coupled with existing decoder heads for tasks such as semantic segmentation.

Self-Supervised Vision Transformers. Self-supervised learning has emerged as a powerful technique for pretraining visual models, eliminating the need for labeled data. One notable self-supervised method is MAE [41] (Masked Autoencoder), which trains a vision transformer to reconstruct masked regions of input images. This approach results in a high generalization capacity. Another significant method is CLIP [10] (Contrastive Language-Image Pre-Training), which involves joint training of a vision transformer and a language model on a large corpus of text and images, leading to the creation of a comprehensive knowledge store. CAE [42] aims to learn image representations that are invariant to context changes and effectively capture underlying semantic content. Furthermore, iBot [50] performs masked visual learning using an online tokenizer and self-distillation mechanism, facilitating semantic representation learning. In our approach, we leverage attention to masks to optimize the extraction of dense hidden representations, thereby enhancing the segmentation capability of our model.

Plain-backbone decoders. In dense prediction tasks like semantic segmentation, high-resolution feature maps generated by backbone play a crucial role. In typical hierarchical transformer models, techniques such as FPN [51] or dilated backbone are employed to generate high-resolution feature maps by merging features from different levels. However, when it comes to plain, non-hierarchical transformer backbone, the resolution remains the same across all layers. SETR [17] proposed a straightforward approach to address segmentation tasks by treating transformer outputs from the base model in a sequence-to-sequence perspective. Segmenter [18] combines class embeddings and transformer patch embeddings and applies several self-attention layers on the combined tokens to learn discriminative embeddings. In their approach, the class tokens are used as input to the ViT backbone, resulting in increased computational complexity. In contrast, our SegViT introduces the class tokens as input to the ATM, the Attention-to-Mask module, thereby reducing computational costs while still benefiting from the integration of class tokens.

Continual Learning. Continual learning (CL) aims to address the issue of forgetting and maintain the performance on previously learned classes while continuously learning new classes [52]. Most CL methods propose regularization techniques for convolution-based networks [53–56] or expand the network architectures to accommodate new tasks [57], thereby avoiding the need to store and replay old data. In recent years, efforts have also emerged to prevent forgetting in Transformer models. Dytox [58] dynamically learns new task tokens, which are then utilized to make the learned embeddings more relevant to the specific task. Life-long ViT [59] and contrastive ViT [60] introduce cross-attention mechanisms between tasks through external key vectors, and they slow down the changes to these keys to mitigate forgetting. Despite the use of complex mechanisms to prevent forgetting, these methods still require fine-tuning of the network for new classes, which can result in interference with previously learned knowledge.

In the field of semantic segmentation, recent research has been devoted to addressing the forgetting issue in continual learning. However, in addition to forgetting, continual semantic segmentation (CSS) also encounters the problem of "background shift." This refers to the situation where foreground object classes from previous tasks are mistakenly classified as background in the current task [30]. REMINDER [24] tackles forgetting in CSS by utilizing class similarity to identify the classes that are more likely to be forgotten. It then focuses on revising those specific classes to mitigate the forgetting problem. RCIL [31] introduces a two-branch convolutional network, with one branch frozen and the other trained to prevent forgetting. At the end of each learning step, the trainable branch is merged with the frozen branch, which can introduce model interference. However, it is worth noting that existing CSS and CL techniques typically involve fine-tuning certain parts of the network dedicated to the old tasks. Unfortunately, this fine-tuning process can lead to forgetting as the model diverges from the previously learned solution.

3 Method

In this section, we will first introduce the overall architecture of our proposed SegViT model. Then, we will proceed to discuss the "Shrunk" architecture, which is designed to reduce the overall computational cost of the model. Additionally, we will delve into the continuous semantic segmentation setting and adapt our SegViT model framework to seamlessly align with this setting.

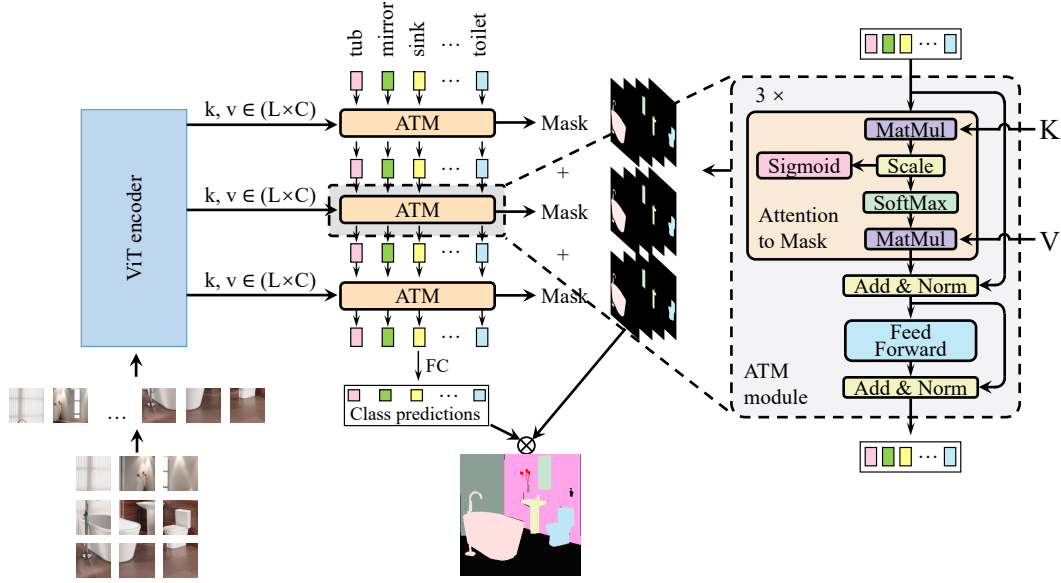


Fig. 3 The overall SegViT structure with the ATM module. The Attention-to-Mask (ATM) module inherits the typical transformer decoder structure. It takes in randomly initialized class embeddings as queries and the feature maps from the ViT backbone to generate keys and values. The outputs of the ATM module are used as the input queries for the next layer. The ATM module is carried out sequentially with inputs from different layers of the backbone as keys and values in a cascade manner. A linear transform is then applied to the output of the ATM module to produce the class predictions for each token. The mask for the corresponding class is transferred from the similarities between queries and keys in the ATM module. We have removed the self-attention mechanism in ATM decoder layers further improve the efficiency while maintaining the performance.

3.1 Overall SegViT architecture

SegViT comprises a ViT-based encoder responsible for feature extraction and a decoder used to learn the segmentation map. In terms of the encoder, we have designed the "Shrunk" structure to decrease the computational costs associated with the plain ViT. As for the decoder, we introduce a novel lightweight module called Attention-to-Mask (ATM). This module generates class-specific masks denoted as M and class predictions denoted as P , which determine the presence of a particular class in the image. The mask outputs from a stack of ATM modules are combined and then multiplied with the class predictions to obtain the final segmentation output. Fig. 3 illustrates the overall architecture of our proposed SegViT.

3.1.1 Encoder

Given an input image $I \in \mathbb{R}^{H \times W \times 3}$, the plain vision transformer backbone reshapes it into a sequence of tokens $\mathcal{F}_0 \in \mathbb{R}^{L \times C}$, where $L = \frac{HW}{P^2}$, P is the patch size, and C is the number of channels. To capture positional information, learnable position embeddings of the same size as \mathcal{F}_0 are added. Subsequently, the token sequence \mathcal{F}_0 undergoes m transformer layers to produce the output. The output tokens for each layer are defined

as $[\mathcal{F}_1, \mathcal{F}_2, \dots, \mathcal{F}_m] \in \mathbb{R}^{L \times C}$. In the case of a plain vision transformer such as ViT, there are no additional modules involved, and the number of tokens remains unchanged for each layer. However, the computational costs of plain ViT can be prohibitively expensive. To address this issue, we introduce the *Shrunk* structure, which enables the development of an efficient ViT-based encoder. Further details regarding the *Shrunk* structure can be found in Section 3.2.

3.1.2 Decoder

Attention-to-Mask (ATM). Cross-attention can be described as the mapping between two sequences of tokens, denoted as $\{\mathbf{v}_1, \mathbf{v}_2\}$. In our case, we define two token sequences: $\mathcal{G} \in \mathbb{R}^{N \times C}$ with a length N equal to the number of classes, and $\mathcal{F}_i \in \mathbb{R}^{L \times C}$. To enable cross-attention, linear transformations are applied to each token sequence, resulting in query (Q), key (K), and value (V) representations. This process is described by Equation (1).

$$\begin{aligned} Q &= \phi_q(\mathcal{G}) \in \mathbb{R}^{N \times C}, \\ K &= \phi_k(\mathcal{F}_i) \in \mathbb{R}^{L \times C}, \\ V &= \phi_v(\mathcal{F}_i) \in \mathbb{R}^{L \times C}. \end{aligned} \quad (1)$$

The similarity map is calculated by computing the dot product between the query and key representations.

Following the scaled dot-product attention mechanism, the similarity map and attention map are calculated as follows:

$$\begin{aligned} S(Q, K) &= \frac{QK^T}{\sqrt{d_k}} \in \mathbb{R}^{N \times L}, \\ \text{Attention}(\mathcal{G}, \mathcal{F}_i) &= \text{Softmax}(S(Q, K))V \in \mathbb{R}^{N \times C}, \end{aligned} \quad (2)$$

where $\sqrt{d_k}$ is a scaling factor with d_k equals to the dimension of the keys.

The shape of the similarity map $S(Q, K)$ is determined by the lengths of the two token sequences, N and L . The attention mechanism updates \mathcal{G} by performing a weighted sum of V , where the weights are derived from the similarity map after applying the softmax function along the L dimension.

In dot-product attention, the softmax function is used to concentrate attention exclusively on the token with the highest similarity. However, we believe that tokens other than those with maximum similarity also carry meaningful information. Based on this intuition, we have designed a lightweight module that generates semantic predictions more directly. To achieve this, we assign \mathcal{G} as the class embeddings for the segmentation task and \mathcal{F}_i as the output of layer i of the ViT backbone. A semantic mask is paired with each token in \mathcal{G} to represent the semantic prediction for each class. The binary mask M is defined as follows:

$$\text{Mask}(\mathcal{G}, \mathcal{F}_i) = \text{Sigmoid}(S(Q, K)) \in \mathbb{R}^{N \times L}. \quad (3)$$

The masks have a shape of $N \times L$, which can be reshaped to $N \times \frac{H}{P} \times \frac{W}{P}$ and bilinearly upsampled to the original image size $N \times H \times W$. The ATM mechanism, as illustrated in the right part of Figure 3, produces masks as its intermediate output during the cross-attention process.

The final output tokens $Z \in \mathbb{R}^{L \times C}$ from the ATM module are utilized for classification. A fully connected layer (FC) parameterized by $W \in \mathbb{R}^{C \times 2}$ followed by the Softmax function is used to predict whether the object class is present in the image or not. The class predictions $\mathcal{P} \in \mathbb{R}^{N \times 2}$ are formally defined as:

$$\mathcal{P} = \text{Softmax}(WZ). \quad (4)$$

Here, $P_{c,1}$ indicates the likelihood of class c appearing in the image. For simplicity, we refer to P_c as the probability score for class c .

The output segmentation map for class $O_s \in \mathbb{R}^{H \times W}$ is obtained by element-wise multiplication of the reshaped class-specific mask M_c and its corresponding prediction score P_c : $O_c = P_c \odot M_c$. During inference,

the label is assigned to each pixel i by selecting the class with the highest score using $\text{argmax}_c O_{i,c}$.

Indeed, plain base models like ViT do not inherently possess multiple stages with features of different scales. Consequently, structures such as Feature Pyramid Network (FPN) that merge features from multiple scales are not applicable to them.

Nevertheless, features from layers other than the last one in ViT contain valuable low-level semantic information, which can contribute to improving performance. In SegViT, we have developed a structure that leverages feature maps from different layers of ViT to enrich the feature representations. This allows us to incorporate and benefit from the rich low-level semantic information present in those feature maps.

SegViT is trained via the classification loss and the binary mask loss. The classification loss (\mathcal{L}_{cls}) minimizes cross-entropy between the class prediction and the actual target. The mask loss ($\mathcal{L}_{\text{mask}}$) consists of a focal loss [61] and a dice loss [62] for optimizing the segmentation accuracy and addressing sample imbalance issues in mask prediction. The dice loss and focal loss respectively minimize the dice and focal scores between the predicted masks and the ground-truth segmentation. The final loss is the combination of each loss, formally defined as:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{focal}} \mathcal{L}_{\text{focal}} + \lambda_{\text{dice}} \mathcal{L}_{\text{dice}} \quad (5)$$

where λ_{focal} and λ_{dice} are hyperparameters that control the strength of each loss function. Previous mask transformer methods such as MaskFormer [35] and DETR [63] have adopted the binary mask loss and fine-tuned their hyperparameters through empirical experiments. Hence, for consistency, we directly use the same values as MaskFormer and DETR for the loss hyperparameters: $\lambda_{\text{focal}} = 20.0$ and $\lambda_{\text{dice}} = 1.0$.

3.2 Shrunk Structure for Efficient Plain ViT Encoder

Recent efforts, such as DynamicViT [64], TokenLearner [65], and SPViT [66], propose token pruning techniques to accelerate vision transformers. However, most of these approaches are specifically designed for image classification tasks and, as a result, discard valuable information. When these techniques are applied to semantic segmentation, they may fail to preserve high-resolution features that are necessary for accurate dense prediction tasks.

In this paper, we propose the *Shrunk* structure, which utilizes query-based down-sampling (QD) to prune the input token sequence \mathcal{F}_i , and query up-sampling (QU) to recover the discarded tokens, thereby

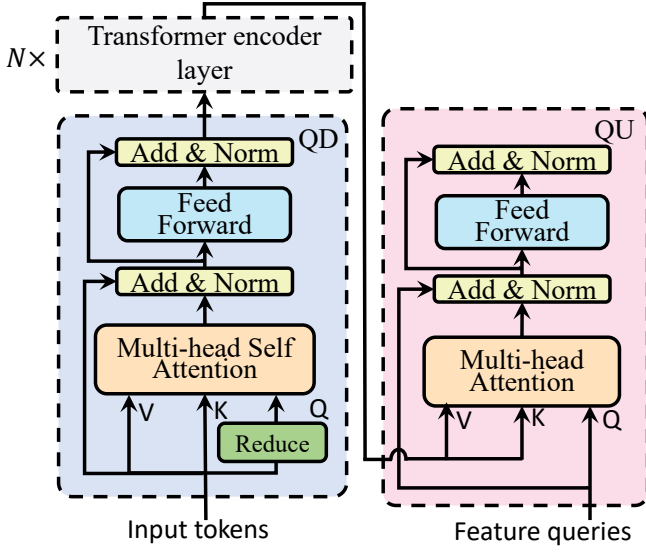


Fig. 4 Architecture of the proposed query-downsampling (QD) layer (blue block) and the query-upsampling (QU) layer (pink block). The QD layer uses an efficient downsampling technique (green block) and removes less informative input tokens used for the query. The QU layer takes a set of trainable query tokens and learns to recover the discarded tokens using multi-head attention.

preserving the fine-detailed features that are crucial for semantic segmentation. The overall architecture of QD and QU is illustrated in Figure 4.

For QD, we have re-designed the Transformer encoder block [67] and incorporated efficient downsampling operations to specifically reduce the number of query tokens. In a Transformer encoder layer, the computational cost is directly influenced by the number of query tokens, and the output size is determined by the query token size. To mitigate the computational burden while maintaining information integrity, a feasible strategy is to selectively reduce the number of query tokens while preserving the key and value tokens. This approach allows for an effective reduction in the output size of the current layer, leading to reduced computational costs for subsequent layers.

For QU, we achieve upsampling by employing either a pre-defined or inherited token sequence with a higher resolution as the query tokens. The key and value tokens are taken from the token sequence obtained from the backbone, which typically has a lower resolution. The output size is dictated by the query tokens with higher resolution. Through the cross-attention mechanism, information from the key and value tokens is integrated into the output. This process facilitates a non-linear merging of information and demonstrates an upsampling behavior, effectively increasing the resolution of the output.

As illustrated in Figure 5, our proposed *Shrunk* structure incorporates the QD and QU modules. Specifically, we integrate a QD operation at the middle depth of the ViT backbone, precisely at the 8th layer of a 24-layer backbone. The QD operation downsamples the query tokens using a 2×2 nearest neighbor downsampling operation, resulting in a feature map size reduction to $1/32$. However, such downsampling can potentially cause information loss and performance degradation. To mitigate this issue, prior to applying the QD operation, we employ a QU operation to the feature map. This involves initializing a set of query tokens with a resolution of $1/16$ to store the information. Subsequently, as the downsampled feature map progresses through the remaining backbone layers, it is merged and upsampled using another QU operation alongside the previously stored $1/16$ high-resolution feature map. This iterative process ultimately generates a $1/16$ high-resolution feature map enriched with semantic information processed by the backbone.

Despite the effectiveness of the proposed *Shrunk* approach in maintaining performance, it requires the integration of the QD operation within the intermediate layers of the backbone. This necessity arises due to the fact that shallow layers primarily capture low-level features, and applying downsampling to these layers would result in significant information loss. Consequently, these low-level layers continue to be computed at a higher resolution, limiting the potential reduction in computational cost. To address this limitation and further optimize the backbone, we introduce enhancements and present a novel architecture called *Shrunk++*. In this enhanced architecture, we incorporate an edge detection module in the QD section and introduce an Edged Query Downsampling (EQD) technique to update the QD process. In addition to the 2×2 nearest downsampling operation that eliminates every 4 consecutive tokens, our approach aims to retain tokens that contain multiple categories, specifically tokens that contain an edge. By preserving the 2×2 sparse tokens, we retain important semantic information, while also preserving the edge tokens to retain detailed spatial information. By retaining both types of information, we minimize the loss of valuable information and overcome the limitations associated with low-level layers. To extract edges, we add a separate branch using a lightweight multilayer perceptron (MLP) edge detection head that learns to detect edges from the input image. The edge detection head operates as an auxiliary branch, trained simultaneously with the main ATM decoder. This head processes the input image, which has the same dimensions as the backbone. Let the input image have C channels,

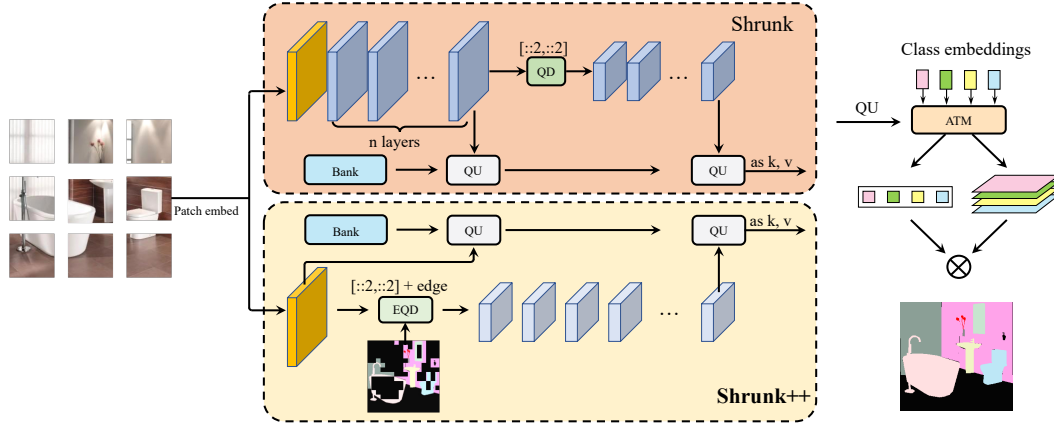


Fig. 5 Illustrations of the *Shrink* and *Shrink++*. In the diagram, the blue and orange boxes respectively refer to the transformer encoder block and the patch embedding block. In SegViT [68], the proposed *Shrink* structure employs query downsampling (QD) on the middle-level features to preserve the information. In the new *Shrink++* architecture, we introduce the Edged Query Downsampling (EQD) technique which consolidates every four adjacent tokens into one token and additionally includes the tokens that contain edges. This enhancement enables downsampling operations to take place before the first layer without significant performance degradation, offering computational savings for the initial layers of the *Shrink* model. The edge information is extracted using a lightweight parallel edge detection head.

aligned with the backbone. The Multi-Layer Perceptron (MLP) in this head consists of three layers, with dimensions C , $C/2$, and 2, respectively. Let I represent the input image, and the output of the MLP can be defined as $E = \text{MLP}(I; W_1, W_2, W_3)$, where W_1, W_2, W_3 are the weights for the three layers. The output E is then passed through a softmax activation function, resulting in $S = \text{Softmax}(E)$. To determine the confidence level of a token belonging to an edge, we apply a threshold τ . In our implementation, we set τ to 0.7. To obtain the ground-truth (GT) edge, we perform post-processing on the GT segmentation map Y . Since the input has been tokenized with a patch size of P , we tokenize the GT and reshape it into a sequence of tokens denoted as $Y \in \mathbb{R}^{(HW/P^2) \times P \times P}$, where the last two dimensions correspond to the patch dimensions. We consider a patch to contain an edge if there exists any edge pixel within the patch. We define the edge mask $Mask_i$ as follows:

$$Mask_i = \begin{cases} 1 & \text{if } \sum_{j,k} Y_{i,j,k} > 0, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

For each element s_i in S , we create a binary edge mask M_i : $M_i = 1$, if $s_i \geq \tau$. The cross-entropy loss is computed between the generated edge mask M_i and the ground-truth edge mask Y_i : $\mathcal{L}_{\text{edge}} = -\sum i Y_i \log(M_i) + (1 - Y_i) \log(1 - M_i)$. By incorporating the Edge Detection head as an auxiliary branch, the *Shrink++* architecture effectively retains detailed spatial contexts throughout the query downsampling process, forming an Edge Query Downsampling (EQD) structure. This EQD structure allows the model to capture and pre-

serve edge information during sparse downsampling, resulting in a significant reduction in computational overhead without compromising performance. The integration of EQD enables the *Shrink++* architecture to strike a remarkable balance between computational efficiency and maintaining high-performance levels.

3.3 Exploration on Continual Semantic Segmentation

Continual semantic segmentation aims to train a segmentation model in T steps without forgetting. At step t , we are given a dataset \mathcal{D}^t which comprises a set of pairs (X^t, Y^t) , where X^t is an image of size $H \times W$ and Y^t is the ground-truth segmentation map. Here, Y^t only consists of labels in current classes \mathcal{C}^t , while all other classes (i.e., old classes $\mathcal{C}^{1:t-1}$ or future classes $\mathcal{C}^{t+1:T}$) are assigned to the background. In continual learning, the model at step t should be able to predict all classes $\mathcal{C}^{1:t}$ in history.

SegViT for Continual Learning. Existing continual semantic segmentation methods [24, 31] propose regularization algorithms to preserve the past knowledge of a specific architecture, DeepLabV3. These methods focus on continual semantic segmentation for DeepLabV3 with a ResNet backbone, which has a less robust visual representation for distinguishing between different categories. Consequently, these methods require fine-tuning model parameters to learn new classes while attempting to retain knowledge of old classes. Unfortunately, adapting the old parameters dedicated to the previous task inevitably interferes with the past knowledge, leading to catastrophic forgetting. In con-

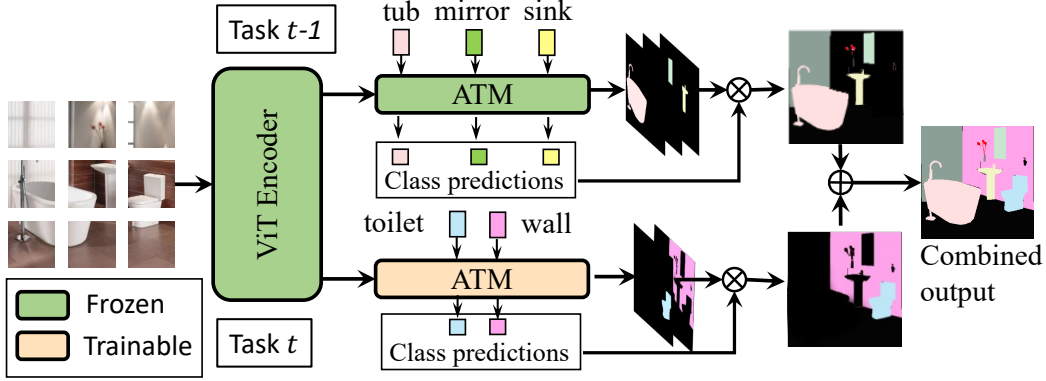


Fig. 6 Overview of SegViT adapted for continual semantic segmentation. When learning a new task t , we grow and train a separate ATM and fully-connected layer to produce mask and class prediction. All the parameters dedicated to the old task $t - 1$, including ATM, FC layer, and the ViT encoder, are frozen. This prevents interfering with the old knowledge, which guarantees no forgetting.

trast, our method, SegViT, decouples class prediction from mask segmentation, making it inherently suitable for a continual learning setting. By leveraging the powerful representation capability of the plain vision transformer, we can learn new classes by solely fine-tuning the class proxy (i.e., the class token) while keeping the old parameters frozen. This approach eliminates the need for fine-tuning old parameters when learning new tasks, effectively addressing the issue of catastrophic forgetting.

When training on a current task t , we add a new sequence of learnable tokens $\mathcal{G}^t \in \mathbb{R}^{|\mathcal{C}^t| \times C}$ with the length equals to the number of classes $|\mathcal{C}^t|$ in the current task. To learn new classes, we grow and train new ATM modules and a fully-connected layer for mask prediction and mask classification. For simplicity, we ignore the parallel structure of ATM modules. A single ATM module refers to multiple ATM modules. Let A^t and W^t denote the ATM module and the weights of the fully connected (FC) layer for task t . All parameters for a previous task, including the ViT encoder, the ATM module, and the FC layer, are completely frozen. Fig. 6 illustrates the overview of our SegViT architecture adapted for continual semantic segmentation.

Given the encoder extracted features \mathcal{F}_T and the class tokens \mathcal{G}^t , the ATM produces the mask predictions M^t and the output tokens Z^t corresponding to the mask:

$$M^t, Z^t = \text{ATM}(\mathcal{G}^t, \mathcal{F}^T). \quad (7)$$

Based on Eq. 4, the class prediction \mathcal{P} is obtained by applying FC on the class token Z^t . The prediction score for each class S_c^t is multiplied with the mask M_c^t to get the segmentation map O_c^t for class c :

$$O_c^t = S_c^t \odot M_c^t, \quad (8)$$

where \odot denotes the element-wise multiplication. The segmentation \hat{O}^t is obtained by taking the class c having the highest score in every pixel, defined as

$$\hat{O}^t = \underset{c \in \mathcal{C}^t}{\operatorname{argmax}} O_{i,c}^t \quad (9)$$

Based on the ground truth Y^t for task t , SegViT is trained using the loss function defined in Eq. 5. To produce the final segmentation for all tasks, we concatenate the outputs O^t from all tasks.

4 Experiments

4.1 Datasets

ADE20K [69] is a challenging scene parsing dataset which contains 20,210 images as the training set and 2,000 images as the validation set with 150 semantic classes.

COCO-Stuff-10K [70] is a scene parsing benchmark with 9,000 training images and 1,000 test images. Even though the dataset contains 182 categories, not all categories exist in the test split. We follow the implementation of mmsegmentation [71] with 171 categories to conduct the experiments.

PASCAL-Context [72] is a dataset with 4,996 images in the training set and 5,104 images in the validation set. There are 60 semantic classes in total, including a class representing ‘background’.

4.2 Implementation details

Transformer backbone. We employ the naive ViT [5] as the backbone for our method. Specifically, we utilize the ‘Base’ variation of ViT for most of our ablation studies and provide results based on the ‘Large’

Table 1 Experiment results on the ADE20K val. split. ‘ms’ means that mIoU is calculated using multi-scale inference. ‘†’ means the models use the backbone weights pre-trained by AugReg [73]. ‘*’ represents the model reproduced under the same settings as the official repo. The GFLOPs are measured at single-scale inference with the given crop size. We report inference speed for our SegViT and reproduce previous methods in terms of Frame Per Second (FPS) on a single A100 device.

Method	Backbone	Crop Size	GFLOPs	mIoU (ss)	mIoU (ms)	Inf time (fps)
UPerNet [15]	ViT-Base	512×512	443.9	46.6	47.5	16.07
DPT* [16]	ViT-Base	512×512	219.8	47.2	47.9	23.63
SETR-MLA* [17]	ViT-Base	512×512	113.5	48.2	49.3	-
Segmenter* [18]	ViT-Base	512×512	129.6	49.0	50.0	20.46
StructToken [74]	ViT-Base	512×512	171.5	50.9	51.8	14.22
MaskFormer [40]	Swin-B(21K)	640×640	198.3	52.7	53.9	-
Mask2Former [37]	Swin-B(21K)	640×640	223.4	53.9	55.1	12.43
SegViT (Ours)	ViT-Base	512×512	120.9	51.3	53.0	31.52
SegViT (<i>Shrunk++</i> , Ours)	BEiTv2-Base	512×512	74.4	52.9	53.3	25.03
SegViT (Ours)	BEiTv2-Base	512×512	120.9	54.0	54.9	23.59
DPT* [16]	ViT-Large†	640×640	800.0	49.2	49.5	9.38
UPerNet [15]	ViT-Large†	640×640	1993.9	48.6	50.0	3.88
SETR-MLA [17]	ViT-Large	512×512	368.6	48.6	50.3	5.17
MCIBI [75]	ViT-Large	512×512	>400	-	50.8	-
Segmenter [18]	ViT-Large†	640×640	671.8	51.8	53.6	4.73
StructToken [74]	ViT-Large†	640×640	774.6	52.8	54.2	4.1
KNet+UPerNet [39]	Swin-L(21K)	640×640	659.3	52.2	53.3	11.28
MaskFormer [40]	Swin-L(21K)	640×640	378.1	54.1	55.6	10.21
Mask2Former [37]	Swin-L(21K)	640×640	402.7	56.1	57.3	8.81
SegViT (ours)	ViT-Large†	640×640	637.9	54.6	55.2	9.37
SegViT(<i>Shrunk</i> , ours)	ViT-Large†	640×640	373.5	53.9	55.1	10.18
SegViT(<i>Shrunk++</i> , ours)	ViT-Large†	640×640	209.1	53.0	54.9	10.26
SegViT (<i>Shrunk++</i> , ours)	BEiTv2-Large†	512×512	210.3	55.1	56.1	9.82
SegViT (ours)	BEiTv2-Large†	512×512	374.0	56.5	58.0	9.39
SegViT (<i>Shrunk++</i> , ours)	BEiTv2-Large†	640×640	308.8	55.7	57.0	9.38
SegViT (ours)	BEiTv2-Large†	640×640	637.9	58.0	58.2	6.25

variation as well. It is worth noting that different pre-trained weights can lead to significant variations in performance, as suggested by Segmenter [18]. Therefore, to ensure a fair comparison, we adopt the pre-trained weights provided by Augreg [73], following the practices of counterparts such as Strudel [18] and StructToken [74]. These weights are obtained through training on ImageNet-21k with strong data augmentation and regularization techniques [73]. To explore the maximum capacity and assess the upper bound of our method, we also conduct experiments using stronger base models such as DEiT v3 [43] and BEiT v2 [12].

Training settings. We use MMSegmentation [71] and follow the commonly used training settings. During training, we applied data augmentation sequentially via random horizontal flipping, random resize with the ratio between 0.5 and 2.0, and random cropping (512×512 for all except that we use 480×480 for PASCAL-Context and 640×640 for ViT-large on ADE20K). The batch size is 16 for all datasets with a total iteration of 160k, 80k, and 80k for ADE20k, COCO-Stuff-10k, and PASCAL-Context respectively.

Evaluation metric. We use the mean Intersection over Union (mIoU) as the metric to evaluate the performance. ‘ss’ means single-scale testing and ‘ms’ test time augmentation with multi-scaled (0.5, 0.75, 1.0, 1.25, 1.5, 1.75) inputs. All reported mIoU scores are in a percentage format. All reported computational costs in GFLOPs are measured using the fvcare¹ library.

4.3 Comparisons with the State-of-the-art Methods

Results on ADE20K. Table 1 reports the comparison with the state-of-the-art methods on ADE20K validation set using ViT backbone. The SegViT uses the ATM module with multi-layer inputs from the original ViT backbone, while the *Shrunk* is the one that conducts QD to the ViT backbone and saves 50% of the computational cost without sacrificing too much performance. Our method achieves State-of-the-art 58.2% (MS) in terms of mIoU with the BEiTv2 Large backbone. To ensure a fair comparison, we evaluate our

¹ <https://github.com/facebookresearch/fvcare>

Table 2 Experiment results on the COCO-Stuff-10K *test*. split. Following published methods, we report the results with multi-scale inference (denoted by ‘ms’). The GFLOPs is measured at single scale inference with a crop size of 512×512 .

Method	Backbone	GFLOPs	mIoU (ms)
DANet [46]	Dilated-ResNet-101	289.3	39.7
MaskFormer [35]	ResNet-101-fpn	81.7	39.8
EMANet [76]	Dilated-ResNet-101	247.4	39.9
SpyGR [77]	ResNet-101-fpn	>80	39.9
OCRNet [3]	HRNetV2-W48	167.9	40.5
GINet [78]	JPU-ResNet-101	>200	40.6
RecoNet [79]	Dilated-ResNet-101	>200	41.5
ISNet [80]	Dilated-ResNeSt-101	228.3	42.1
MCIBI [75]	ViT-Large	>380	44.9
StructToken [74]	ViT-Large	>400	49.1
SenFormer [81]	Swin-Large	>400	50.1
SegViT (<i>Shrunk</i> , ours)	ViT-Large	224.8	49.4
SegViT (ours)	ViT-Large	383.9	50.3
SegViT (<i>Shrunk++</i> , ours)	BEiTv2-Large	213.3	50.54
SegViT (ours)	BEiTv2-Large	388.2	53.46

Table 3 Experimental results on the PASCAL-Context *val*. split. Following published methods, we report the results with multi-scale inference (denoted by ‘ms’). mIoU₅₉: mIoU averaged over 59 classes (without background). mIoU₆₀: mIoU averaged over 60 classes (59 classes plus background). Both metrics were used in the literature, and we report for the 60 classes. The GFLOPs are measured at single scale inference with a crop size of 480×480 .

Method	Backbone	GFLOPs	mIoU ₅₉ (ms)	mIoU ₆₀ (ms)
RefineNet [82]	ResNet-152	-	-	47.3
UNet++ [83]	ResNet-101	-	47.7	-
PSPNet [32]	Dilated-ResNet-101	157.0	47.8	-
Ding <i>et al.</i> [84]	ResNet-101	-	51.6	-
EncNet [85]	Dilated-ResNet-101	192.1	52.6	-
HRNet [86]	HRNetV2-W48	82.7	54.0	48.3
NRD [87]	ResNet-101	42.9	54.1	49.0
GFFNet [88]	Dilated-ResNet-101	-	54.3	-
EfficientFCN [89]	ResNet-101	52.8	55.3	-
OCRNet [3]	HRNetV2-W48	143.9	56.2	-
SETR-MLA [17]	ViT-Large	318.5	-	55.8
Segmenter [18]	ViT-Large	346.2	-	59.0
SenFormer [81]	Swin-Large	-	64.0	-
SegViT (<i>Shrunk</i> , ours)	ViT-Large	186.9	62.3	57.4
SegViT (ours)	ViT-Large	321.6	65.3	59.3
SegViT (<i>Shrunk++</i> , ours)	BEiTv2-Large	179.3	64.91	59.92
SegViT (ours)	BEiTv2-Large	329.7	67.14	61.63

SegViT module with the BEiT-v2 large backbone on a crop size of 512×512 , which consumes 374.0 GFLOPs. Our approach achieves a slightly better performance of 56.5% mIoU compared to Mask2former-Swin-L, which achieves 56.1% with 402.7 GFlops on a crop size of 640×640 . Additionally, our *Shrunk* version with a computational cost reduction of around 50% (308.8 GFlops) achieves a competitive performance of 57.0% (MS) in terms of mIoU. Optimizing SegViT with ViT-Large using the proposed *Shrunk++* reduces the computational cost of *Shrunk* by 3.05 times, while preserving the mIoU. Fig. 7 shows the visual results of different segmentation methods. In contrast to other methods that often confuse similar classes and misclassify related

concepts, our SegViT stands out by more precise object boundary delineation and achieving accurate segmentation of complete objects, even in cluttered scenes.

Results on COCO-Stuff-10K. Table 2 shows the result on the COCO-Stuff-10K dataset. Our method achieves 50.3% which is higher than the previous state-to-the-art StrucToken by 1.2% with less computational cost. Our *Shrunk* version achieves 49.4% with 224.8 GFLOPs, which is similar to the computational cost of a dilated ResNet-101 backbone but with much higher performance. By extending SegViT with the effective *Shrunk++*, we significantly decrease its GFLOPs

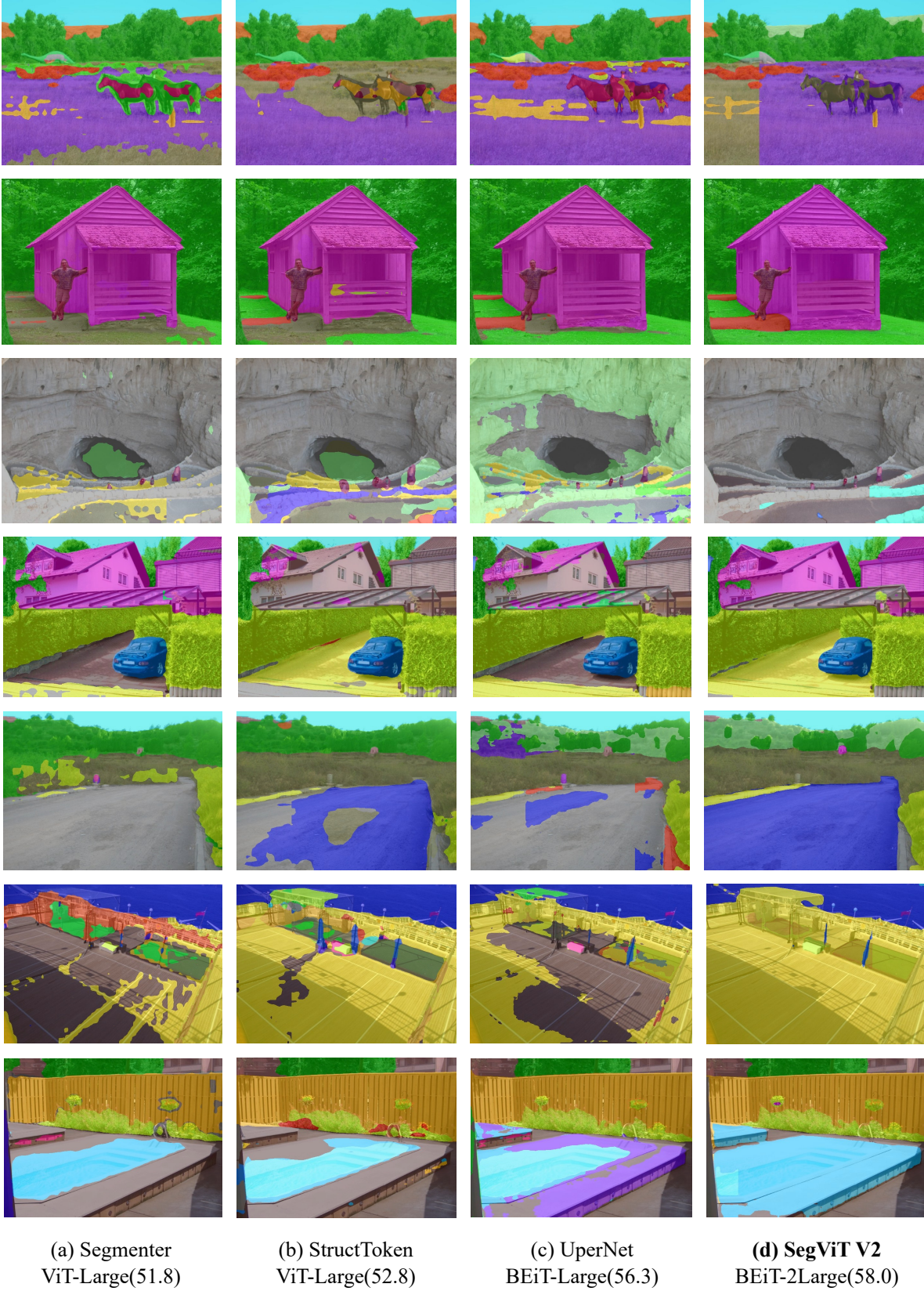


Fig. 7 Visuals results of different segmentation networks and plain ViT backbones on the ADE20K validation set [69]. It includes the following models: (a) Segmenter [18] with ViT large, (b) StructToken [74] with ViT large, (c) UPerNet [15] with BEiT large, and (d) SegViT V2 with BEiTv2 large. The results demonstrate that our methods effectively generate accurate segmentation masks and unlock the potential of plain ViT. Zoom in for better view.

by 1.82 times, while still maintaining the competitive mIoU.

Results on PASCAL-Context. Table 3 shows the results on the PASCAL-Context dataset. We follow HR-Net [86] to evaluate our method and report the results under 59 classes (without background) and 60 classes (with background). Using full SegViT structure without adopting *Shrunk* or *Shrunk++*, we reach mIoU of 67.14% and 61.63% respectively for those two metrics, outperforming the state-of-the-art methods using the ViT backbones with less computational cost. By applying *Shrunk* and *Shrunk++* architecture, the computational cost in terms of GLOPs is reduced by 42% and 45%, respectively. SegViT with *Shrunk++* achieves the best trade-off between accuracy and speed among all methods on the PASCAL-Context dataset.

4.4 Ablation Study

In this section, we conduct extensive ablation studies to show the effectiveness of our proposed methods.

Effect of the ATM module. We conducted an analysis to evaluate the impact of using the proposed ATM module as an encoder. The results are summarized in Table 4. To establish a baseline for comparison, we introduced SETR-naive, which utilizes two 1×1 convolutions to directly derive per-pixel classifications from the final layer of the ViT-Base transformer output. From the results, it is evident that applying the ATM module under the supervision of a conventional cross-entropy loss leads to a performance improvement of 0.5%. However, the performance gains become much more substantial when we decouple the classification and mask prediction processes, supervising each separately. This approach results in a significant performance boost of 3.1%, highlighting the efficacy of the ATM module in enhancing semantic segmentation performance.

Table 4 Comparisons between our proposed ATM module with SETR [17]. ‘CE loss’ indicates the cross-entropy loss commonly used in semantic segmentation. The experiments on the ADE20k dataset are carried out using the ViT-Base backbone.

Decoder	Loss	mIoU (ss)
SETR	CE loss	46.5
ATM	CE loss	47.0 (+0.5)
ATM	\mathcal{L}_{mask} loss	49.6 (+3.1)

Ablation of the feature levels. The effects of using multiple-layer inputs from the backbone to the ATM modules are presented in Table 5. The incorporation of feature maps from lower layers leads to a notable performance improvement of 1.3%. We further investigated the impact of including more layers of features and observed additional gains in performance. After empirical testing, we determined that utilizing three layers yielded optimal results, resulting in an overall mIoU boost of 1.7%. These ablation studies confirm the effectiveness of our proposed ATM decoder and highlight the advantage of incorporating multi-layer features into the segmentation structure. This integration significantly enhances the performance of semantic segmentation tasks.

Table 5 Ablation results of using different layer inputs to the SegViT structure on ADE20K dataset using ViT-Base as the backbone. Involving multi-layer features can bring obvious performance gains.

	Used layers	mIoU (ss)
Single	[12]	49.6
Cascade	[6, 12]	50.9 (+1.3)
Cascade	[6, 8, 12]	51.3 (+1.7)
Cascade	[3, 6, 9, 12]	51.2 (+1.6)

SegViT on hierarchical base models. We conducted an analysis to evaluate the performance of SegViT on hierarchical base models. For comparison, we selected two competitive methods, Maskformer [35] and Mask2former [37]. The results presented in Table 6 indicate that, even though our method was not specifically designed for hierarchical base models, we are still able to achieve competitive performance while maintaining computational efficiency. This demonstrates the applicability of our SegViT approach to various types of ViT-Base models.

Table 6 The experiments use the Swin-Tiny [7] backbone and are carried out on the ADE20K dataset. The GFLOPs are measured at single scale inference with a crop size of 512×512 .

Method	mIoU (ss)	GFLOPs
Maskformer [35]	46.7	57.3
Mask2former [37]	47.7	73.7
SegViT (Ours)	47.1	48.0

Ablation of *Shrunk* and *Shrunk++* strategies. In this section, we analyze the effectiveness of the differ-

Table 7 Ablation results of *Shrunk* and *Shrunk++* version on the ADE20K dataset. We explored various shrink strategies. The GFLOPs are measured at single-scale inference with a crop size of 512×512 on the ViT-Base backbone. QD: query-based downsampling. QU: query-based upsampling. QD_{layer} indicates which layer to apply the QD. QD_{method} indicates the downsampling method for QD.

Structure	QD	QU	QD_{layer}	QD_{method}	Head	mIoU (ss)	GFLOPs
Single	-	-	-	-	SETR	46.5	107.3
Single	-	-	-	-	ATM	49.6 (+3.1)	115.8
Naive <i>Shrunk</i>	✓	-	6	2x2	ATM	46.9 (+0.4)	74.1
<i>Shrunk</i>	✓	✓	6	2x2	ATM	50.0 (+3.5)	97.1
Nearest - TS	✓	✓	0	3x3	ATM	38.9(-7.6)	32.8
Nearest - TS	✓	✓	0	2x2	ATM	43.3(-3.2)	46.1
Shrunk++	✓	✓	0	3x3-Edge	ATM	47.9(+1.4)	69.3
Shrunk++	✓	✓	0	2x2-Edge	ATM	49.9(+3.4)	74.6

ent SegViT structures through an ablation study. Table 7 presents the effects of various techniques employed in each SegViT structure, including query upsampling (QU), query downsampling (QD), token-squeezing (TS) techniques, and segmentation heads. When the ATM head is applied to the ‘Single’ structure with the SETR head, there is a significant performance improvement of 6.67%. This demonstrates the effectiveness of the ATM head in enhancing the performance of the baseline structure. However, when QD is applied to the ‘Single’ structure with the ATM head, there is a performance drop of 2.7%. This indicates that there is information loss during the downsampling process. However, by incorporating QU, the performance is recovered. QU helps recover the discarded information from QD and reconstructs the high-resolution feature map, which is crucial for dense prediction tasks. The *Shrunk* architecture, which utilizes both QU and QD, achieves optimal performance while reducing the computational cost compared to the ‘Single’ structure by 16.15%.

In the proposed *Shrunk++* structure, we analyze the performance of two main token-squeezing techniques: nearest downsampling and edge-aware downsampling. It is important to note that token squeezing is directly applied to the first layer of the network for optimal computational efficiency. Applying naive nearest downsampling with a 3x3 kernel reduces the GFLOPs of the *Shrunk* structure without token-

squeezing by a factor of 2.97. However, reducing the computational cost with 3x3 and 2x2 nearest downsampling leads to a performance drop of 13%. In contrast, by incorporating an additional edge extractor into our *Shrunk++* architecture, we significantly improve the mIoU, achieving performance on par with *Shrunk*, i.e., 49.9%, while only slightly increasing the computational cost to 74.6 GFLOPs. The edge-aware downsampling technique preserves the edge details, thereby preserving discriminative features for dense predictions. Among the different settings, the 2x2 + Naive MLP Edge setting achieves an optimal balance between performance and efficiency.

Ablation of the components in *Shrunk* structure. Table 7 shows the effectiveness of each component (QD and QU) in the *Shrunk* structure. The results presented match the structures illustrated in Fig. 5. Applying the ATM head improves the performance of the ‘Single’ structure with SETR head by 6.67%. When QD is applied to the ‘Single’ structure with ATM head, the performance drops by 2.7%, which indicates the information loss. By applying QU, the performance is recovered. QU helps recover the information discarded by QD and reconstructs the high-resolution feature map, which is crucial for dense prediction tasks. The *Shrunk* architecture uses QU and QD jointly and obtains opti-

Table 8 Ablation results of different decoder methods with their corresponding feature merge types and loss types. ViT-Base is employed as the backbone for all the variants.

Decoder	Multi-level Features		Loss Types			mIoU (ss)
	FPN	Token Merge	Pixel level	Dot product	Attention Mask	
SETR-MLA [17]	✓		✓			48.2
Segmenter [18]			✓			49.0
MaskFormer [35]	✓			✓		46.7
Ours-Variant 1					✓	49.6
Ours-Variant 2		✓		✓		50.6
Ours		✓			✓	51.2

Table 9 Ablation of the QD module in terms of the targets and methods to down-sample. The experiments are carried out on the ViT-Large backbone of ADE20K dataset.

Applied to	Methods	mIoU (ss)
Q	Conv	44.5
Q, K, V	Nearest	52.6
Q	Nearest	53.9

Table 10 Comparisons for various ViT pre-training schedules on the validation set of ADE20K. All results are reported in single-scale inference. The default configuration for these base models is pre-trained on ImageNet-1K with 224 * 224 resolutions. ‘*’ means the models use the backbone weights pre-trained with 384 * 384 resolutions. ‘†’ means the base models pre-trained on imagenet-21K. The proposed SegVit head has a less computational cost and performs better than UPerNet among all pre-training variants.

Backbone	SegViT mIoU	Head FLOPs	UPerNet mIoU	Head FLOPs	ImageNet Acc
MAE Base [41]	49.22 (▲1.12)	6.89(▼329.73)	48.1	336.62	83.66
CLIP Base [10]	50.76 (▲1.16)	6.89(▼329.73)	49.6	336.62	80.20
CAE Base [42]	50.42 (▲0.22)	6.89(▼329.73)	50.2	336.62	83.90
iBot Base [50]	50.58 (▲0.58)	6.89(▼329.73)	50.0	336.62	84.00
Augreg Base*† [73]	51.30 (▲2.66)	6.89(▼329.73)	48.6	336.62	85.49
DEiT v3 Base† [43]	52.40 (▲0.60)	6.89(▼329.73)	51.8	336.62	85.70
BEiT v2 Base† [12]	53.97 (▲0.47)	6.89(▼329.73)	53.5	336.62	86.50
Augreg Large*† [73]	54.60 (▲2.50)	16.36(▼1,366.33)	52.1	1382.69	85.59
DEiT v3 Large*† [43]	55.81 (▲1.21)	16.36(▼1,366.33)	54.6	1382.69	87.70
BEiT v2 Large [43]	58.00 (▲1.3)	16.36(▼868.28)	56.7	884.64	87.30

mal performance while saving the cost compared with the ‘Single’ structure by 16.15%.

Ablation studies on decoder variances. Different decoder methods are associated with specific feature merge types and loss types. In Table 8, we compare the designs of various decoders on a plain ViT backbone. For hierarchical base models like Swin, the resolution of the feature maps in each stage is reduced. Consequently, the adoption of Feature Pyramid Network (FPN) is necessary to obtain feature maps with larger resolutions and rich semantic information. However, in Table 8, we observe that the FPN structure does not perform well with plain vision transformers. In the case of plain ViT base models, the resolution is maintained, and the feature map of the last layer contains the most comprehensive semantic information. Hence, our proposed method, which utilizes tokens to merge features from different levels, achieves better performance. By simply replacing the FPN structure with the ATM-based token merge, we improve the performance from 46.7% to 50.6%. Regarding the loss type, the pixel-level loss refers to the conventional cross-entropy loss applied to the feature map. The dot product loss corresponds to the loss utilized in [63] and [35]. Attention mask loss indicates that mask supervision is directly applied to the similarity map generated by the ATM during attention calculation. By adding loss supervision on the attention mask, as in our proposed method, the performance improves by 0.6%.

Ablation for the QD module. The motivation behind using QD is to leverage the pre-trained weights of the backbone. As shown in Table 9, if we employ a stride-2 convolution with learnable parameters to downsample the query, it will disrupt the pre-trained weights and result in a significant performance decrease.

Applying down-sampling to both the query and the key-value pairs would inevitably lead to information loss during the down-sampling process, which is evident in the weaker performance observed. Through our investigations, we have found that applying 2×2 nearest down-sampling exclusively to the query in the QD module yields better results. This approach allows us to preserve the pre-trained weights of the backbone while achieving the desired down-sampling effect.

4.5 Application 1: A Better Indicator for Feature Representation Learning

Background. Semantic segmentation serves as a fundamental vision task that has been extensively employed in previous research to assess the representation learning capabilities of weakly, fully, and self-supervised base models [12, 41–43]. In prior work, the UPerNet decoder structure has been commonly used for semantic segmentation. However, the UPerNet decoder may not be a suitable indicator for evaluating the feature representation ability of the base model. This is primarily due to its heavier computational requirements and slower convergence rate. Additionally, the feature representation obtained by the base model can vary significantly due to different training strategies employed during the fine-tuning process on semantic segmentation datasets. Consequently, the task of semantic segmentation may not effectively evaluate the feature representation ability of pre-trained models.

Experiment settings. This section presents a comprehensive evaluation of our proposed SegVit on various weakly/fully/self-supervised vision transformers, including those proposed by He et al. [41], Chen et al. [42], Touvron et al. [43], and the BEiT model [12].

Table 11 CSS results on ADE20k in mIoU (%) on 100-50 and 100-10 settings. The relative mIoU reduction compared with the joint training for each method is reported.

Method	100-50 (2 tasks)				100-10 (6 tasks)			
	0-100	101-150	<i>all</i>	<i>avg</i>	0-100	101-150	<i>all</i>	<i>avg</i>
ILT [90]	18.29 (\blacktriangledown 26.1)	14.40 (\blacktriangledown 13.8)	17.00 (\blacktriangledown 22.0)	29.42	0.11 (\blacktriangledown 44.2)	3.06 (\blacktriangledown 25.1)	1.09 (\blacktriangledown 37.9)	12.56
MiB [30]	40.52 (\blacktriangledown 3.9)	17.17 (\blacktriangledown 11.0)	32.79 (\blacktriangledown 6.2)	37.31	38.21 (\blacktriangledown 6.1)	11.12 (\blacktriangledown 17.1)	29.24 (\blacktriangledown 9.8)	35.12
SDR [91]	40.52 (\blacktriangledown 3.8)	17.17 (\blacktriangledown 11.0)	32.79 (\blacktriangledown 6.2)	37.31	37.26 (\blacktriangledown 7.1)	12.13 (\blacktriangledown 16.1)	28.94 (\blacktriangledown 10.1)	34.48
PLOP [92]	41.76 (\blacktriangledown 2.6)	14.52 (\blacktriangledown 13.7)	32.74 (\blacktriangledown 6.3)	37.73	38.59 (\blacktriangledown 5.8)	14.21 (\blacktriangledown 14.0)	30.52 (\blacktriangledown 8.5)	34.48
REMINDER [24]	41.55 (\blacktriangledown 2.8)	19.16 (\blacktriangledown 9.0)	34.14 (\blacktriangledown 4.9)	38.43	38.96 (\blacktriangledown 5.4)	21.28 (\blacktriangledown 6.9)	33.11 (\blacktriangledown 5.9)	37.47
RCIL [31]	42.35 (\blacktriangledown 2.0)	18.47 (\blacktriangledown 9.7)	34.45 (\blacktriangledown 4.6)	38.48	29.42 (\blacktriangledown 15.0)	13.49 (\blacktriangledown 14.0)	28.36 (\blacktriangledown 10.0)	29.93
Oracle - ResNet backbone	44.34	28.21	39.00	-	44.34	28.21	39.00	-
MiB [30]	43.43 (\blacktriangledown 3.2)	30.63 (\blacktriangledown 4.3)	39.19 (\blacktriangledown 3.6)	38.66	39.15 (\blacktriangledown 7.5)	20.37 (\blacktriangledown 14.5)	34.17 (\blacktriangledown 8.6)	39.53
PLOP [92]	43.82 (\blacktriangledown 2.8)	26.23 (\blacktriangledown 8.7)	37.99 (\blacktriangledown 4.8)	38.06	43.25 (\blacktriangledown 3.4)	24.13 (\blacktriangledown 10.8)	36.25 (\blacktriangledown 6.5)	40.28
REMINDER [24]	44.66 (\blacktriangledown 2.0)	26.76 (\blacktriangledown 8.1)	38.73 (\blacktriangledown 4.0)	38.43	43.28 (\blacktriangledown 3.4)	24.33 (\blacktriangledown 10.6)	37.10 (\blacktriangledown 5.6)	41.76
Oracle - ViT backbone	46.63	34.90	42.75	-	46.63	34.90	42.75	-
SegViT-CL (ours)	53.64 (\blacktriangledown0.5)	40.00 (\blacktriangledown5.6)	49.09 (\blacktriangledown2.2)	46.82	53.77 (\blacktriangledown0.3)	35.54 (\blacktriangledown10.0)	47.70 (\blacktriangledown3.6)	50.59
Oracle	54.11	45.60	51.28	-	54.11	45.60	51.28	-

We demonstrate that our method outperforms UPerNet [15] in both self-supervised and multiple modality base models, achieving state-of-the-art performance. Notably, our approach achieves superior performance to UPerNet while utilizing only 5% of the computational cost in terms of the decoder head. Table 10 illustrates that our proposed SegViT head consistently outperforms UPerNet on all base models. For the ViT-Base, our method improves the performance of UPerNet on the CLIP model by 1.16% while significantly reducing the computational cost. Similar observations can be made for ViT-Large base models. Furthermore, compared to UPerNet, our proposed SegViT head exhibits a better alignment between the growth trend of segmentation accuracy and the classification accuracy on ImageNet. This clearly demonstrates the superior efficiency of our SegViT head compared to UPerNet, making it a more suitable indicator for feature representation learning in base models.

4.6 Application2: Continual Semantic Segmentation

Due to the decoupling of class prediction and mask segmentation in our proposed SegVit decoder, we are inherently suitable for a continuous learning setting. This characteristic allows us to learn new classes by solely fine-tuning the class proxy (the class token), leveraging the powerful representation ability of the plain vision transformer while keeping the old parameters frozen. To validate the effectiveness of this new approach to continual learning, we conducted quick fine-tuning experiments following previous continuous learning settings.

Experiment settings. Continual Semantic Segmentation (CSS) has two settings [30]: disjoint and overlapped. In the disjoint setup, all pixels in the images at each step belong to either the previous classes or the

current class. In the overlapped setting, the dataset of each step contains all the images that have pixels of at least one current class, and all pixels from previous and future tasks are labeled as background. The overlapped setting is more realistic and challenging, thus we evaluate the performance of the overlapped setup on the ADE20k dataset.

Following prior published works [24, 30, 92], we perform three experiments: adding 50 classes after training with 100 classes (100-50 setting with 2 steps), adding 50 classes each time after training with 50 classes (50-50 setting with 3 steps), adding 10 classes each time sequentially after training with 100 classes (100-10 setting with 6 steps).

Baselines We conducted a comprehensive comparison of our proposed method against state-of-the-art Continual Semantic Segmentation (CSS) techniques, including RCIL [31], PLOP [92], REMINDER [24], SDR [91], and MiB [30]. To ensure fair comparisons, existing methods were evaluated using DeepLabV3 [93] with ResNet101 and ViT-Base backbones that were pre-trained on ImageNet-21k. The reported results for PLOP, RCIL, and REMINDER were obtained based on the codebases provided by the respective authors. Furthermore, we included the performance of the Oracle model, which represents the upper bound achieved by jointly training on all available data, serving as a benchmark for each method.

Metrics. We evaluate the model performance by five mIoU metrics. First, we compute mIoU for the base classes \mathcal{C}^0 , which reflects model rigidity: the model’s resilience to catastrophic forgetting. Second, we compute mIoU for all incremented classes $\mathcal{C}^{1:T}$, which measures plasticity: the model capacity in learning new tasks. Third, we compute the mIoU of all classes in $\mathcal{C}^{0:T}$ (*all*), which shows the overall performance of models. Fourth, we report the average of mIoU (*avg*) measured step af-

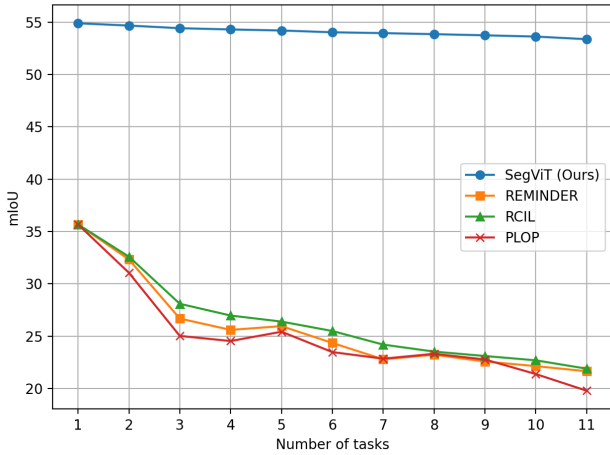


Fig. 8 mIoU of recent CSS methods on the first 100 base classes after incrementally learning new tasks on 100-5 settings with 11 tasks.

Table 12 Performance drop (degree of forgetting) of all classes grouped by tasks in the 100-10 setting. We report the class mIoU when the model first learns the task, and the mIoU when the model last learns it.

Tasks	101-110	111-120	121-130	131-140	141-150	avg
First Time	34.93	39.78	41.10	36.22	27.95	35.99
Last Time	34.51	39.30	40.86	35.09	27.95	35.54
Forgetting	▼0.42	▼0.48	▼0.24	▼1.12	▼0	▼0.45

ter step as proposed by [92], which evaluates performance over the entire continual learning process. To ensure fair comparisons, we evaluate the relative performance of each CSS method in terms of relative mIoU reduction compared with its Oracle model jointly trained on all data.

Results and Discussion. Table 11 shows the results of different CSS methods on ADE20k. Our SegViT-CL consistently outperforms existing methods in *all* mIoU on both settings. In terms of mIoU reduction, the proposed SegViT-CL only decreases the mIoU of the Oracle model by 2.2% on the 100-50 setting, which is two times better than the second-best method, RCIL with ResNet backbone with 4.6% reduction. This substantial enhancement over existing methods underlines the effectiveness of our proposed method in the continual semantic segmentation paradigm. On a long CL setting 100-10 with 6 tasks, ours is almost forgetting-free with a marginal mIoU reduction of 0.3%, while recent CSS methods significantly suffer from forgetting with at least 5.4% mIoU reduction. Using the ViT backbone, existing methods including MiB, REMINDER and PLOP still suffer from high mIoU reductions. Compared with the Oracle, MiB [30], PLOP [92], and REMINDER [24] decrease the mIoU by 8.6%, 6.5% and

5.6% respectively in the 100-10 setting, demonstrating the sub-optimal performance of current CSS methods for ViT architecture. This highlights the need for developing a specialized ViT architecture that is robust to forgetting.

To evaluate the forgetting of every task in the 100-10 setting, we compute the performance drop at the last step compared with its initial mIoU when the model first learns the task. For example, the initial mIoU of task 2 is the mIoU of class 101-110 evaluated at step 2. Similarly, that of task 3 is the mIoU of class 111-120 reported at step 3. Table 12 shows the performance drop at the last step compared with the initial mIoU of each task. Averaged across 5 tasks, the mIoU only drops by 0.45%, which shows that SegViT is robust to forgetting across all tasks in the 100-10 setting. Fig. 8 shows the mIoU on the base classes after incrementally training on many tasks in 100-5, which is a long continual learning setting with 11 tasks. Overall, our SegViT achieves nearly zero forgetting for almost all tasks at the last step. In contrast to previous CSS methods which require partial fine-tuning, the proposed SegViT supports completely freezing old parameters to eliminate interference with past knowledge.

5 Conclusion

This paper presents SegViT, a novel approach for semantic segmentation using plain ViT transformer base models. The proposed method introduces a lightweight decoder that incorporates the Attention-to-mask (ATM) module. Additionally, a *Shrunk++* structure is proposed to reduce the computational cost of the ViT encoder by 50% while maintaining competitive segmentation accuracy. Moreover, this work extends the SegViT framework to address the challenge of continual semantic segmentation, aiming to achieve nearly zero forgetting. By protecting the parameters of old tasks, SegViT effectively mitigates the impact of catastrophic forgetting. Extensive experimental evaluations conducted on various benchmarks demonstrate the superiority of SegViT over UPerNet, while significantly reducing computational costs. The introduced decoder head provides a robust and cost-effective solution for future research in the field of ViT-based semantic segmentation.

Acknowledgments

This work was in part supported by National Key R&D Program of China (No. 2022ZD0118700). Y. Liu’s participation was in part supported by the start-up funding

of The University of Adelaide. B. Zhang's participation was in part supported by Meituan.

References

1. J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2015, pp. 3431–3440.
2. J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, 2020.
3. Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. Eur. Conf. Comp. Vis.*. Springer, 2020, pp. 173–190.
4. L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comp. Vis.*, 2018, pp. 801–818.
5. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *Proc. Int. Conf. Learn. Repren.*, 2021.
6. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.
7. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 10 012–10 022.
8. W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 568–578.
9. Y. Xu, J. Zhang, Q. Zhang, and D. Tao, "Rethinking hierarchies in pre-trained plain vision transformer," *arXiv preprint arXiv:2211.01785*, 2022.
10. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.
11. H. Bao, L. Dong, S. Piao, and F. Wei, "BEiT: BERT pre-training of image transformers," in *International Conference on Learning Representations*, 2022. [Online]. Available: <https://openreview.net/forum?id=p-BhZSz59o4>
12. Z. Peng, L. Dong, H. Bao, Q. Ye, and F. Wei, "BEiT v2: Masked image modeling with vector-quantized visual tokenizers," 2022.
13. L. Wei, L. Xie, W. Zhou, H. Li, and Q. Tian, "Mvp: Multimodality-guided visual pre-training," in *Proc. Eur. Conf. Comp. Vis.*. Springer, 2022, pp. 337–353.
14. H. Lu, N. Fei, Y. Huo, Y. Gao, Z. Lu, and J.-R. Wen, "Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2022, pp. 15 692–15 701.
15. T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comp. Vis.*, 2018, pp. 418–434.
16. R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 12 179–12 188.
17. S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. Torr *et al.*, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021, pp. 6881–6890.
18. R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 7262–7272.
19. R. M. French, "Catastrophic forgetting in connectionist networks," *Trends in cognitive sciences*, vol. 3, no. 4, pp. 128–135, 1999.
20. J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.
21. C. Shao and Y. Feng, "Overcoming catastrophic forgetting beyond continual learning: Balanced training for neural machine translation," *arXiv preprint arXiv:2203.03910*, 2022.
22. Z. Wang, Z. Zhang, C.-Y. Lee, H. Zhang, R. Sun, X. Ren, G. Su, V. Perot, J. Dy, and T. Pfister, "Learning to prompt for continual learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 139–149.
23. Z. Wang, Z. Zhang, S. Ebrahimi, R. Sun, H. Zhang, C.-Y. Lee, X. Ren, G. Su, V. Perot, J. Dy *et al.*, "Dualprompt: Complementary prompting for rehearsal-free continual learning," in *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*. Springer, 2022, pp. 631–648.
24. M. H. Phan, S. L. Phung, L. Tran-Thanh, A. Bouzerdoum *et al.*, "Class similarity weighted knowledge distillation for continual semantic segmentation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2022, pp. 16 866–16 875.
25. O. Ostapenko, T. Lesort, P. Rodríguez, M. R. Arefin, A. Douillard, I. Rish, and L. Charlin, "Continual learning with foundation models: An empirical study of latent replay," in *Conference on Lifelong Learning Agents*. PMLR, 2022, pp. 60–91.
26. V. V. Ramasesh, A. Lewkowycz, and E. Dyer, "Effect of scale on catastrophic forgetting in neural networks," in *Proc. Int. Conf. Learn. Repren.*, 2022.
27. T. Wu, M. Caccia, Z. Li, Y.-F. Li, G. Qi, and G. Hafari, "Pretrained language model in continual learning: A comparative study," in *Proc. Int. Conf. Learn. Repren.*, 2022.
28. A. Maracani, U. Michieli, M. Toldo, and P. Zanuttigh, "Recall: Replay-based continual learning in semantic segmentation," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021.
29. S. Cha, Y. Yoo, T. Moon *et al.*, "Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 10 919–10 930.
30. F. Cermelli, M. Mancini, S. R. Bulò, E. Ricci, and B. Caputo, "Modeling the background for incremental learning in semantic segmentation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2020, pp. 9230–9239.
31. C.-B. Zhang, J.-W. Xiao, X. Liu, Y.-C. Chen, and M.-M. Cheng, "Representation compensation networks for continual semantic segmentation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2022, pp. 7053–7064.

32. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017.
33. X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2022, pp. 12 124–12 134.
34. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.
35. B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.
36. F. Li, H. Zhang, S. Liu, L. Zhang, L. M. Ni, H.-Y. Shum *et al.*, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," *arXiv preprint arXiv:2206.02777*, 2022.
37. B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," 2022.
38. J. Jain, J. Li, M. Chiu, A. Hassani, N. Orlov, and H. Shi, "Oneformer: One transformer to rule universal image segmentation," *arXiv preprint arXiv:2211.06220*, 2022.
39. W. Zhang, J. Pang, K. Chen, and C. C. Loy, "K-net: Towards unified image segmentation," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.
40. B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," 2021.
41. K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.
42. X. Chen, M. Ding, X. Wang, Y. Xin, S. Mo, Y. Wang, S. Han, P. Luo, G. Zeng, and J. Wang, "Context autoencoder for self-supervised representation learning," *arXiv preprint arXiv:2202.03026*, 2022.
43. H. Touvron, M. Cord, and H. Jégou, "Deit iii: Revenge of the vit," in *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*. Springer, 2022, pp. 516–533.
44. Y. Yuan, X. Chen, X. Chen, and J. Wang, "Segmentation transformer: Object-contextual representations for semantic segmentation," *arXiv preprint arXiv:1909.11065*, 2019.
45. W. Zhang, J. Pang, K. Chen, and C. C. Loy, "K-net: Towards unified image segmentation," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 326–10 338, 2021.
46. J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019, pp. 3146–3154.
47. Z. Zhou, B. Zhang, Y. Lei, L. Liu, and Y. Liu, "Zegclip: Towards adapting clip for zero-shot semantic segmentation," *arXiv preprint arXiv:2212.03588*, 2022.
48. X. Chu, Z. Tian, Y. Wang, B. Zhang, H. Ren, X. Wei, H. Xia, and C. Shen, "Twins: Revisiting the design of spatial attention in vision transformers," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021.
49. Y.-H. Wu, Y. Liu, X. Zhan, and M.-M. Cheng, "P2t: Pyramid pooling transformer for scene understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022.
50. J. Zhou, C. Wei, H. Wang, W. Shen, C. Xie, A. Yuille, and T. Kong, "ibot: Image bert pre-training with online tokenizer," *Proc. Int. Conf. Learn. Repren.*, 2022.
51. T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017, pp. 2117–2125.
52. Z. Chen and B. Liu, *Lifelong Machine Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning, 2016.
53. Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, pp. 2935–2947, 2018.
54. A. Douillard, M. Cord, C. Ollion, T. Robert, and E. Valle, "Podnet: Pooled outputs distillation for small-tasks incremental learning," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2020, pp. 86–102.
55. M. Kang, J. Park, and B. Han, "Class-incremental learning by knowledge distillation with adaptive feature consolidation," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2022, pp. 16 071–16 080.
56. Y. Peng, J. Qi, Z. Ye, and Y. Zhuo, "Hierarchical visual-textual knowledge distillation for life-long correlation learning," *Int. J. Comp. Vis.*, vol. 129, pp. 921–941, 2021.
57. S. Yan, J. Xie, and X. He, "Der: Dynamically expandable representation for class incremental learning," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021, pp. 3014–3023.
58. A. Douillard, A. Ramé, G. Couairon, and M. Cord, "Dytox: Transformers for continual learning with dynamic token expansion," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2022, pp. 9285–9295.
59. Z. Wang, L. Liu, Y. Duan, Y. Kong, and D. Tao, "Continual learning with lifelong vision transformer," in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2022, pp. 171–181.
60. Z. Wang, L. Liu, Y. Kong, J. Guo, and D. Tao, "Online continual learning with contrastive vision transformer," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2022, pp. 631–650.
61. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2017, pp. 2980–2988.
62. F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *3DV*. IEEE, 2016, pp. 565–571.
63. N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2020, pp. 213–229.
64. Y. Rao, W. Zhao, B. Liu, J. Lu, J. Zhou, and C.-J. Hsieh, "Dynamicvit: Efficient vision transformers with dynamic token sparsification," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 13 937–13 949.
65. M. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova, "Tokenlearner: Adaptive space-time tokenization for videos," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, pp. 12 786–12 797, 2021.
66. Z. Kong, P. Dong, X. Ma, X. Meng, W. Niu, M. Sun, X. Shen, G. Yuan, B. Ren, H. Tang *et al.*, "Spvit: Enabling faster vision transformers via latency-aware soft token pruning," in *Proc. Eur. Conf. Comp. Vis.* Springer, 2022, pp. 620–640.
67. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
68. B. Zhang, Z. Tian, Q. Tang, X. Chu, X. Wei, C. Shen, and Y. Liu, "Segvit: Semantic segmentation with plain

- vision transformers,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2022.
69. B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, “Scene parsing through ade20k dataset,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017, pp. 633–641.
 70. H. Caesar, J. Uijlings, and V. Ferrari, “Coco-stuff: Thing and stuff classes in context,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018, pp. 1209–1218.
 71. MMSegmentation, “MMSegmentation: OpenMMLab semantic segmentation toolbox and benchmark,” <https://github.com/open-mmlab/mms Segmentation>, 2020.
 72. R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Fidler, R. Urtasun, and A. Yuille, “The role of context for object detection and semantic segmentation in the wild,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2014, pp. 891–898.
 73. A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, “How to train your vit? data, augmentation, and regularization in vision transformers,” 2021.
 74. F. Lin, Z. Liang, J. He, M. Zheng, S. Tian, and K. Chen, “Structtoken: Rethinking semantic segmentation with structural prior,” 2022.
 75. Z. Jin, T. Gong, D. Yu, Q. Chu, J. Wang, C. Wang, and J. Shao, “Mining contextual information beyond image for semantic segmentation,” in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 7231–7241.
 76. X. Li, Z. Zhong, J. Wu, Y. Yang, Z. Lin, and H. Liu, “Expectation-maximization attention networks for semantic segmentation,” in *Proc. IEEE Int. Conf. Comp. Vis.*, 2019, pp. 9167–9176.
 77. X. Li, Y. Yang, Q. Zhao, T. Shen, Z. Lin, and H. Liu, “Spatial pyramid based graph reasoning for semantic segmentation,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2020, pp. 8950–8959.
 78. T. Wu, Y. Lu, Y. Zhu, C. Zhang, M. Wu, Z. Ma, and G. Guo, “Ginet: Graph interaction network for scene parsing,” in *Proc. Eur. Conf. Comp. Vis.* Springer, 2020, pp. 34–51.
 79. W. Chen, X. Zhu, R. Sun, J. He, R. Li, X. Shen, and B. Yu, “Tensor low-rank reconstruction for semantic segmentation,” in *Proc. Eur. Conf. Comp. Vis.* Springer, 2020, pp. 52–69.
 80. Z. Jin, B. Liu, Q. Chu, and N. Yu, “Isnet: Integrate image-level and semantic-level context for semantic segmentation,” in *Proc. IEEE Int. Conf. Comp. Vis.*, 2021, pp. 7189–7198.
 81. W. Bousselham, G. Thibault, L. Pagano, A. Machireddy, J. Gray, Y. H. Chang, and X. Song, “Efficient self-ensemble framework for semantic segmentation,” *arXiv preprint arXiv:2111.13280*, 2021.
 82. G. Lin, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-path refinement networks for high-resolution semantic segmentation,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017, pp. 1925–1934.
 83. Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested U-net architecture for medical image segmentation,” in *Proc. Deep Learning in Medical Image Analysis Workshop*, 2018, pp. 3–11.
 84. H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, “Context contrasted feature and gated multi-scale aggregation for scene segmentation,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018, pp. 2393–2402.
 85. H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, “Context encoding for semantic segmentation,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2018, pp. 7151–7160.
 86. K. Sun, Y. Zhao, B. Jiang, T. Cheng, B. Xiao, D. Liu, Y. Mu, X. Wang, W. Liu, and J. Wang, “High-resolution representations for labeling pixels and regions,” 2019.
 87. B. Zhang, Z. Tian, C. Shen *et al.*, “Dynamic neural representational decoders for high-resolution semantic segmentation,” vol. 34, 2021.
 88. X. Li, H. Zhao, L. Han, Y. Tong, S. Tan, and K. Yang, “Gated fully fusion for semantic segmentation,” in *Proc. AAAI Conf. on Arti. Intel.*, vol. 34, no. 07, 2020, pp. 11 418–11 425.
 89. J. Liu, J. He, J. Zhang, J. Ren, and H. Li, “EfficientFCN: Holistically-guided decoding for semantic segmentation,” in *Proc. Eur. Conf. Comp. Vis.*, 2020.
 90. U. Michieli and P. Zanuttigh, “Incremental learning techniques for semantic segmentation,” in *Proc. IEEE Int. Conf. Comp. Vis. Workshops*, 2019, pp. 3205–3212.
 91. —, “Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021, pp. 1114–1124.
 92. A. Douillard, Y. Chen, A. Dapogny, and M. Cord, “Plop: Learning without forgetting for continual semantic segmentation,” in *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2021.
 93. L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” *arXiv preprint arXiv:1706.05587*, 2017.