Know Your Attention Maps: Class-specific Token Masking for Weakly Supervised Semantic Segmentation

Joëlle Hanna Damian Borth
University of St.Gallen, Switzerland
{joelle.hanna, damian.borth}@unisg.ch

Abstract

Weakly Supervised Semantic Segmentation (WSSS) is a challenging problem that has been extensively studied in recent years. Traditional approaches often rely on external modules like Class Activation Maps [36] to highlight regions of interest and generate pseudo segmentation masks. In this work, we propose an end-to-end method that directly utilizes the attention maps learned by a Vision Transformer (ViT) [7] for WSSS. We propose training a sparse ViT with multiple [CLS] tokens (one for each class), using a random masking strategy to promote [CLS] token - class assignment. At inference time, we aggregate the different self-attention maps of each [CLS] token corresponding to the predicted labels to generate pseudo segmentation masks[1]*. Our proposed approach enhances the interpretability of self-attention maps and ensures accurate class assignments. Extensive experiments on two standard benchmarks and three specialized datasets demonstrate that our method generates accurate pseudo-masks, outperforming related works. Those pseudo-masks can be used to train a segmentation model which achieves results comparable to fully-supervised models, significantly reducing the need for fine-grained labeled data.*

1. Introduction

Semantic segmentation is a fundamental task in computer vision, aiming to assign a class label to each pixel in an image. This fine-grained understanding of visual data is essential for various domain-specific applications, such as medical imaging, autonomous driving, and remote sensing. Traditionally, achieving high performance in semantic segmentation has relied heavily on the availability of large annotated datasets. However, acquiring such fine-grained labels is often labor-intensive, expensive or even impractical, especially in specialized domains where expert knowledge is required.

[1]Code is available at github.com/HSG-AIML/TokenMasking-WSSS

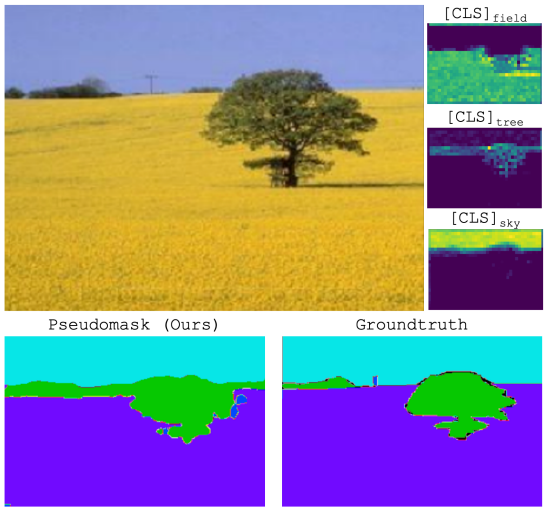


Figure 1. Our method uses multiple [CLS] tokens in ViTs to encourage class-specific self-attention maps. By aggregating these maps, we generate a pseudo-mask (bottom left) comparable to the ground truth (bottom right), without ever using the fine-grained segmentation labels.

To address these challenges, weakly supervised semantic segmentation (WSSS) has emerged as a promising alternative. WSSS leverages weaker forms of supervision, such as image-level labels, to reduce the burden of detailed annotations. A common approach in WSSS involves the use of Class Activation Maps [35], which highlight the most discriminative regions of an image for a particular class. Despite their effectiveness, CAMs suffer from limitations such as coarse localization and an inability to capture fine details, as they only emphasize the most prominent features. Moreover, generating CAMs requires an additional module on top of the existing model, adding complexity to process.
Given these limitations, an alternative approach is to leverage the self-attention maps of Vision Transformers (ViTs,[7]). Unlike CAMs, which rely on additional modules, the attention mechanism in ViTs inherently provides interpretability by capturing spatial dependencies across the

entire image. Vision Transformers, with their self-attention mechanisms, have shown remarkable performance in various vision tasks. However, a significant challenge in using ViTs for WSSS is the inability to assign specific classes to individual attention heads, making it difficult to interpret and utilize their outputs for segmentation tasks effectively [9].

In this work, we propose a novel approach for WSSS that leverages the multi-head self-attention mechanism of ViTs in a more structured manner. Instead of relying on a single `[CLS]` token that aggregates global information, we introduce multiple `[CLS]` tokens, each designed to correspond to a specific class in the segmentation task. This allows the model to decompose the scene into distinct class-specific representations. Ideally, each `[CLS]` token would correspond to one class. However, there is no inherent guarantee that each `[CLS]` token will only capture information relevant to its intended class. To address this, we implement a novel masking mechanism on the `[CLS]` token output embeddings to help ensure that each `[CLS]` token is properly assigned to a single class (Figure 1).

Furthermore, since not all attention heads contribute equally to class-specific feature extraction, we introduce a pruning mechanism to remove redundant or noisy heads. By selectively removing less important heads, we aim to sharpen the attention mechanism, making the remaining heads more interpretable. The ultimate goal is to leverage these refined attention heads to create accurate class-specific segmentation masks, thus enabling the ViT to perform weakly supervised semantic segmentation effectively.

This work aims to bridge the gap between the powerful capabilities of ViTs and the need for class-specific feature embeddings in complex vision tasks. The contributions of this work can be summarized as follows:

- We propose to use multiple `[CLS]` tokens instead of one, and develop a masking mechanism for the `[CLS]` tokens that reliably assigns each `[CLS]` token to its respective class.
- We introduce attention head pruning during training, significantly reducing noise in self-attention maps and improving pseudo-mask quality.
- We validate the application of this method in WSSS on two standard datasets and across three distinct domains: remote sensing, medical imaging, and general scene understanding. Each domain presents unique challenges: remote sensing requires accurate parsing of complex landscapes, medical imaging demands precise object-centric segmentation for diagnostic purposes, and general scene understanding involves diverse and sparse labels.
- We demonstrate that our work can effectively reduce the need for fine-grained labeled data, benefiting both small-scale and large-scale datasets.

2. Related Works

2.1. Weakly Supervised Semantic Segmentation

Weakly supervised semantic segmentation (WSSS) offers an alternative to traditional supervised methods by relying on less precise forms of supervision, such as image-level labels [5, 30], bounding boxes [12, 25], or scribbles [15, 27]. WSSS approaches can be broadly categorized into single-stage and multi-stage methods, each with distinct workflows and objectives.

Single-stage WSSS methods aim to directly generate segmentation maps from weak labels without intermediate processing. These approaches often leverage end-to-end models that integrate weak supervision signals with advanced architectures. Recent advancements include Vision Transformers (ViTs) specifically tailored for WSSS. Hanna *et al*. [9] introduced a multimodal ViT enforcing sparsity in attention heads during training, enabling the direct generation of high-quality segmentation maps. Similarly, Xu *et al*. [33] proposed a Multi-Class Token Transformer that employs class tokens to enhance localization for multiple objects in a single forward pass. These single-stage approaches simplify the WSSS pipeline by bypassing the need for pseudo-mask refinement stages, making them computationally efficient and suitable for real-time applications.

Multi-stage WSSS methods, on the other hand, involve an iterative process where intermediate representations, such as pseudo segmentation masks (pseudo-masks), are generated and refined before final segmentation. This process typically begins with generating coarse localization maps from weak labels, followed by refinement and re-training in a fully supervised manner. One common approach is based on Class Activation Maps (CAMs), initially proposed by Zhou *et al*. [36], which provide rough localization cues by highlighting discriminative image regions. Building on CAMs, Ahn *et al*. [1] introduced AffinityNet, which refines these maps by considering pixel affinities to propagate segmentation information. Wei *et al*. [31] extended this idea with Adversarial Erasing, iteratively suppressing confident CAM regions to uncover complete object extents.

Further advancements in multi-stage WSSS have integrated various forms of weak supervision and iterative refinement strategies. For instance, Chen *et al*. [4] proposed a dual-network approach, where a segmentation network and a pseudo-label generation network are alternately trained to progressively improve pseudo-mask quality. Additionally, recent works [3, 14] have leveraged the Segment Anything Model (SAM) [13] during inference, employing post-processing techniques to refine segmentation results.

2.2. Attention Maps of Vision Transformers

Many studies have recently demonstrated that attention maps generated by Vision Transformers are not only effective for downstream tasks but also offer insights into the decision-making processes of these models. Caron *et al*. [2] introduced a self-supervised learning method called DINO, which significantly enhances the interpretability of ViTs. They observed that self-supervised ViTs can automatically learn class-specific features and produce unsupervised object segmentations. Furthermore, Darcet *et al*. [6] identified artifacts in the attention maps of both supervised and self-supervised ViT networks. By introducing register tokens, they provided an effective solution to eliminate these artifacts, leading to more interpretable attention maps.

3. Approach

In this section, we describe our approach and illustrate it in Figure 2.

3.1. Vision Transformer Basics

Given an image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, it is divided into N patches, each of size $P \times P$. These patches are flattened and embedded into a higher-dimensional space:

$$\mathbf{z}_0 = [\mathbf{x}_p^1\mathbf{E}; \mathbf{x}_p^2\mathbf{E}; \ldots; \mathbf{x}_p^N\mathbf{E}] + \mathbf{E}_{\text{pos}}$$

where $\mathbf{E}$ is the embedding matrix and $\mathbf{E}_{\text{pos}}$ represents positional encodings. A special `[CLS]` token is prepended to the sequence of embedded patches:

$$\mathbf{z}_0 = [\texttt{[CLS]}; \mathbf{z}_0]$$

The sequence is then processed through Transformer encoder layers, leveraging multi-head self-attention (MSA) and feed-forward neural networks to generate feature representations [28]:

$$\mathbf{z}'_\ell = \text{MSA}(\text{LN}(\mathbf{z}_{\ell-1})) + \mathbf{z}_{\ell-1}$$
$$\mathbf{z}_\ell = \text{MLP}(\text{LN}(\mathbf{z}'_\ell)) + \mathbf{z}'_\ell$$

where $\mathbf{z}_\ell$ is the output of the ℓ-th layer, and LN denotes layer normalization. The final output representation corresponding to the `[CLS]` token is used for classification tasks:

$$\mathbf{y} = \text{MLP}(\mathbf{z}_L^{\texttt{[CLS]}})$$

where $\mathbf{z}_L^{\texttt{[CLS]}}$ is the output of the final Transformer layer corresponding to the `[CLS]` token.

3.2. ViT with Multiple `[CLS]` Tokens

We extend the standard ViT architecture, which typically uses a single `[CLS]` token, to incorporate multiple `[CLS]` tokens. Each `[CLS]` token is intended to represent a different class in a multi-label classification setup. This enables the model to handle images with multiple objects and infer segmentation masks for each class separately. The extended embedding sequence is given by:

$$\mathbf{z}_0 = [\texttt{[CLS]}_1; \texttt{[CLS]}_2; \ldots; \texttt{[CLS]}_C; \mathbf{z}_0]$$

where C is the number of classes. Each `[CLS]` token's output $\mathbf{z}_L^{\texttt{[CLS]}_c}$ for class c is processed to generate class-specific predictions.

Additional Register Token Following the work from Darcet *et al*. [6], we append an additional learnable token that the model can use as a register, `[REG]`, to the end of the token sequence. The register token captures general context to prevent it from 'polluting' the `[CLS]` self-attention.

$$\mathbf{z}_0 = [\texttt{[CLS]}_1; \texttt{[CLS]}_2; \ldots; \texttt{[CLS]}_C; \mathbf{z}_0; \texttt{[REG]}]$$

3.3. Random Masking of `[CLS]` Tokens

To address the issue of the lack of hard assignment between each `[CLS]` token and its corresponding class, we implement a `[CLS]` masking strategy during training. This approach aims to reduce interference between classes, as the model would learn to rely on the available tokens to make class-specific decisions, promoting more robust class representations. The process is as follows:

1. **Random Selection**: During each training iteration, a subset of `[CLS]` tokens that do not correspond to the current image labels are randomly selected to be masked. The masking ratio is set to 50%, with a sensitivity analysis of this parameter presented in Section 5.6. Let $\mathcal{Y}$ be the set of true labels for the image $\mathbf{x}$. We define the masking function $m(i)$ for `[CLS]` token CLS_i as follows:

$$m(i) = \begin{cases} 0 & \text{if } i \in \mathcal{Y} \\ 1 & \text{if } i \notin \mathcal{Y} \text{ and randomly selected to be masked} \end{cases}$$

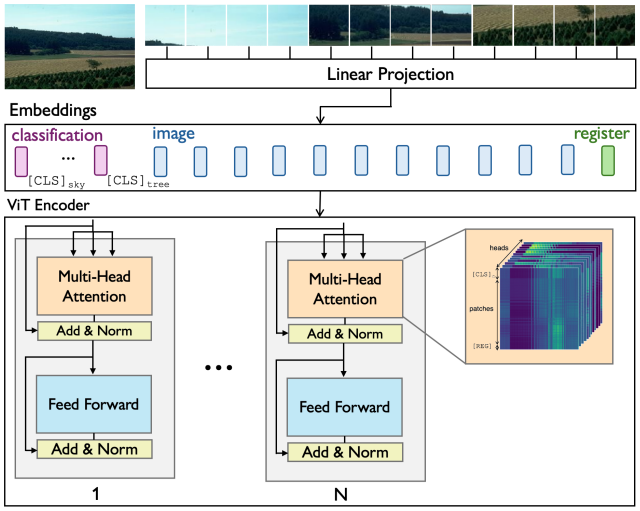
2. **Output Embedding Masking**: To ensure masked `[CLS]` tokens do not contribute to the final predictions, we mask their output embeddings:

$$\mathbf{z}_L^{\texttt{[CLS]}_i} = \mathbf{z}_L^{\texttt{[CLS]}_i} \cdot (1 - m(i))$$

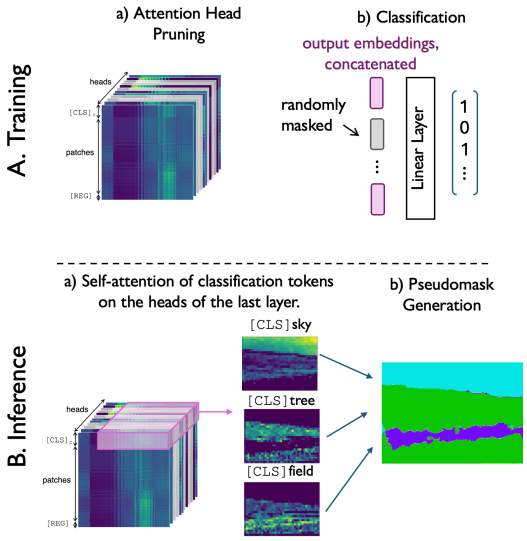
where $\mathbf{z}_L^{\texttt{[CLS]}_i}$ is the final output embedding for the $\texttt{[CLS]}_i$ token.

3.4. Sparse Training of the Vision Transformer

Aside from the challenge of assigning each `[CLS]` token to a specific class, another essential task is making the self-attention of these `[CLS]` tokens interpretable, providing clear and detailed shapes of different classes. Following common findings that many heads can be pruned in a transformer [9, 29], our solution is to prune redundant heads that might introduce noise, leading to clearer and more accurate class-specific representations.



(a) Overview of the method's architecture.



(b) Training and inference pipelines

Figure 2. Our proposed approach divides the input image into patches, which are projected into a sequence of image tokens. To this sequence, we append C classification tokens (`[CLS]`) and a register token `[REG]`. During training, attention heads are pruned to enhance the interpretability of the self-attention maps, and the output embeddings of the `[CLS]` tokens are randomly masked before classification to enforce class assignments. During inference, the self-attention of the `[CLS]` tokens corresponding to the predicted labels are reshaped into the input image dimensions and used to generate pseudo-masks.

To achieve this, we modify the ViT architecture by incorporating gating units g_i into each attention head. Those scalar gates are multiplied by the output of each head i. This modifies the multi-head self-attention (MSA) mechanism as follows:

$$\mathrm{MSA}(Q, K, V) = \mathrm{concat}_i(g_i \cdot \mathrm{head}_i)W^O$$

Here, W^O is a parameter matrix. Our objective is to entirely disable the less significant heads. While L_0 regularization would be ideal for this purpose [17], it is non-differentiable. To address this, we employ a stochastic relaxation method that uses Hard Concrete distributions [18]. By summing the probabilities of heads being non-zero, we can approximate the L_0 norm:

$$\mathcal{L}_{\text{reg}} = \sum_i (1 - P(g_i = 0 \mid \phi_i))$$

Here, ϕ_i are the parameters of the Hard Concrete distribution. The training objective becomes:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{reg}} \tag{1}$$

$\mathcal{L}_{\text{cls}}$ is the loss used for training the classifier, e.g. binary cross-entropy. The parameter λ is used to control the number of heads to be pruned. It is set to 0.01, resulting in approximately two-thirds of the heads being pruned. A sensitivity analysis on that parameter is presented in the Supplementary Material.

3.5. Segmentation Pseudo-mask Generation

The final step in our approach is to generate pseudo segmentation masks by combining the self-attention maps for each predicted class from the corresponding `[CLS]` tokens. For each image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, we extract the representations from the `[CLS]` tokens corresponding to the predicted classes, each denoted as $Z \in \mathbb{R}^{1 \times \frac{H \times W}{P^2}}$. These representations are then reshaped to align with the spatial dimensions of the image, resulting in attention scores of dimensions $\mathbb{R}^{\frac{H}{P} \times \frac{W}{P}}$. Then, we threshold each self-attention map of the `[CLS]` tokens for the predicted classes to binarize them. We then combine all binarized maps, starting with the class of the lowest probability (logit) and ending with the highest. Any pixel not assigned to a class is replaced by the most common value of neighboring pixels.

4. Datasets

We evaluate our approach on two standard datasets: MS COCO 2014 [16] and Pascal VOC 2012 [8] augmented by the SBD dataset [10], following common practices. We also use three domain-specific datasets: the DFC2020 remote sensing dataset [24], the EndoTect Polyp Segmentation Medical Imaging dataset [11], and the ADE20K general scene parsing dataset [37].

The DFC2020 dataset includes ∼5K paired Sentinel 1 (SAR) and Sentinel 2 (multispectral) satellite observations for training, and ∼1K for testing, each image sized 256 ×

256. It provides pixel-level land cover annotations for eight imbalanced classes. Image-level labels are generated for training, keeping classes covering $\geq 10\%$ of the image.

The EndoTect Polyp Segmentation dataset consists of around 110K high-resolution endoscopic images with detailed polyp region annotations. Polyp segmentation is crucial for the early diagnosis of cancer, making this dataset important for evaluating medical imaging tasks with high variability in polyp appearance.

The ADE20K dataset contains images annotated with 150 object categories from various scenes, including urban, indoor, and natural environments, making it suitable for evaluating generalization to real-world scenarios.

Across all datasets, we use classification labels during training and pixel-level annotations to evaluate the weak segmentation accuracy of our method.

5. Experiments

5.1. Models Architecture

We use a ViT-B [7] as our base model with a fully connected layer for classification, to generate pseudomasks as initial segmentation predictions. These pseudomasks are then used to train a UNet model [20], which serves as our segmentation model, to produce the final results.

5.2. Pseudo-masks Evaluation

We evaluate the quality of the pseudo-masks generated by our method by comparing their accuracy against those from existing state-of-the-art WSSS approaches and report the results in terms of mean Intersection over Union (mIoU). Table 1 shows the quantitative performance of our approach on the Pascal VOC 2012 `train` and `val` sets. We note that single-stage methods, which generate pseudomasks in a single step, achieve competitive results with simpler pipelines than multi-stage approaches. Among these, our method achieves the best validation mIoU of 73.7%, slightly outperforming the previous best single-stage method, DuPL [32] (73.5%). Moreover, visually, as shown in Figure 3, the pseudo-masks generated by our method closely match the ground truth segmentation masks on the Pascal VOC and MS COCO datasets.

5.3. Final Segmentation Evaluation

Table 2 compares semantic segmentation results of different Weakly Supervised Semantic Segmentation (WSSS) methods on the VOC validation, VOC test, and COCO validation datasets, measured in mIoU. Multistage methods, such as I/C-CTI [34], achieve strong results, with the best VOC validation mIoU (74.1%). Our method achieves the highest VOC test mIoU (73.5%) and performs strongly on VOC validation (72.7%) and COCO validation (43.2%), highlighting its efficiency and generalization capabilities with a simplified pipeline. Additionally, as shown in Figure 4, the model performs well on specialized datasets like ADE20K and DFC2020, demonstrating its ability to handle diverse tasks. On ADE20K, it produces clear object boundaries in complex scenes, while on DFC2020, it segments detailed land-cover features effectively, even with imbalanced classes. These results further validate the robustness and adaptability of our method across various datasets and tasks.

Table 1. Evaluation of pseudo-masks. The results are evaluated on the VOC train and val sets, and reported in mIoU (%). †denotes using ImageNet-21k pretraining.

Method	Backbone	`train`	`val`
Multi-stage WSSS methods			
ViT-PCM + CRF [21] ECCV 2022	ViT-B†	71.4	69.3
ReCAM [5] CVPR 2022	ResNet50	70.5	-
I/C-CTI [34] CVPR 2024	DeiT-S	73.7	-
Single-stage WSSS methods			
ViT-PCM [21] ECCV 2022	ViT-B†	67.7	66.0
AFA [22] CVPR 2022	MiT-B1	68.7	66.5
ToCo [23] CVPR 2023	ViT-B	72.2	70.5
DuPL [32] CVPR 2024	ViT-B	**75.1**	73.5
Ours	ViT-B	74.5	**73.7**

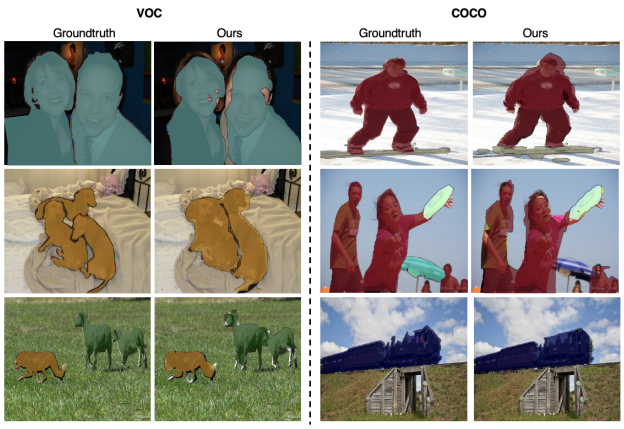


Figure 3. Pseudo-mask results on the Pascal VOC and MS COCO datasets.

5.4. Specialized Datasets

Table 3 shows the performance of our method on specialized datasets (DFC2020, Endotect, and ADE20K). Unlike the standard benchmarks, specialized datasets often contain unique characteristics such as imbalanced classes, fine-grained details, or domain-specific variations, making segmentation more difficult. Our method achieves state-of-the-art results among weakly supervised approaches on all specialized datasets, even surpassing fully supervised methods

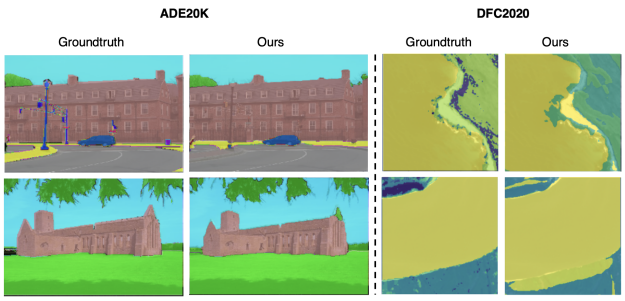


Figure 4. Segmentation results on the ADE20K and the DFC2020 datasets.

on DFC2020. This demonstrates its robustness and adaptability to real-world scenarios where labeled data is scarce. A qualitative comparison for the three datasets is available in the Supplementary Material.

5.5. Ablation Studies

To better understand the contribution of the different components in our method, we perform a series of ablation studies evaluating the effects of the random masking strategy, the additional `[REG]` token, and attention head pruning. The ablation studies are conducted across multiple datasets, with results summarized in both visualizations (Figure 5) and quantitative tables (Tables 4, 5, and 6).

Effect of the Random Masking Strategy We investigate the influence of the random masking strategy on the performance of our method (Section 3.3). The results are reported in Table 4. Randomly masking a subset of the `[CLS]` tokens is designed to ensure that each token is assigned one class to focus on. This strategy appears to be effective as indicated by the significant performance gap between the two setups, across both datasets. Without random masking, class assignments tend to be incorrect, leading to situations where the shapes are accurate, but the assigned classes are wrong, as observed in Figure 5. In the column "Ours w/o Masking" the self-attention matrices fail to capture correct class-specific features, especially for "building", "sky", or "wall". A sensitivity analysis of the masking ratio is performed in Section 5.6 and Figure 6.

Effect of the additional register `[REG]` token We investigate the effect of incorporating an additional learnable

Table 2. Semantic Segmentation Results. The results are evaluated on the VOC val and test sets, and COCO val set, and reported in mIoU (%). †denotes using ImageNet-21k pretraining.

Method	Backbone	VOC		COCO
		val	test	val
Multi-stage WSSS methods				
ReCAM [5] CVPR 2022	DL-V2	68.4	68.2	45.0
MCTformer [33] CVPR 2022	WR-38	71.9	71.6	42.0
I/C-CTI [34] CVPR 2024	ResNet38	**74.1**	73.2	45.4
Single-stage WSSS methods				
AFA [22] CVPR 2022	MiT-B1	66.0	66.3	38.9
ToCo [23] CVPR 2023	ViT-B	69.8	70.5	41.3
DuPL [32] CVPR 2024	ViT-B	72.2	71.6	43.6
Ours	ViT-B	72.7	**73.5**	43.2

Table 3. Performance on specialized datasets, showing fully supervised semantic segmentation methods and weakly supervised semantic segmentation methods. For each dataset we report *Pixel Accuracy* and *mIoU*. The top model overall is underlined, and the best WSSS model is bolded.

		Pixel Accuracy	mIoU
DFC2020	**Supervised Methods**		
	UNet	57.2	53.1
	ViT + seg.	46.4	43.4
	Robinson *et al.* [19] 2021	70	-
	Weakly Supervised Methods		
	ViT + GradCAM	36.6	25.1
	MCTformer [33] 2022	58.8	50.5
	Sparse ViT [9] 2023	57.3	52.2
	Ours	**74.1**	**67.2**
Endotect	**Supervised Methods**		
	UNet	80.3	73.0
	ViT + seg.	67.8	60.0
	Tomar *et al.* [26] 2020	-	78
	Weakly Supervised Methods		
	ViT + GradCAM	63.2	55.9
	MCTformer [33] 2022	71.7	63.3
	Sparse ViT [9] 2023	76.0	68.0
	Ours	**78.4**	**69.8**
ADE20K	**Supervised Methods**		
	UNet	64.3	55.0
	ViT + seg.	43.5	32.2
	Zhou *et al.* [37] 2017	74	61
	Weakly Supervised Methods		
	ViT + GradCAM	32.7	21.3
	MCTformer [33] 2022	45.8	33.3
	Sparse ViT [9] 2023	44.2	34.5
	Ours	**51.8**	**38.2**

Table 4. Effect of random (rnd) masking of a subset of the `[CLS]` tokens (50%) on the pseudo-masks, with considerable improvement on all datasets.

	w/o Rnd Masking	w/ Rnd Masking
MS COCO	41.9	**43.2**
VOC	71.6	**72.7**

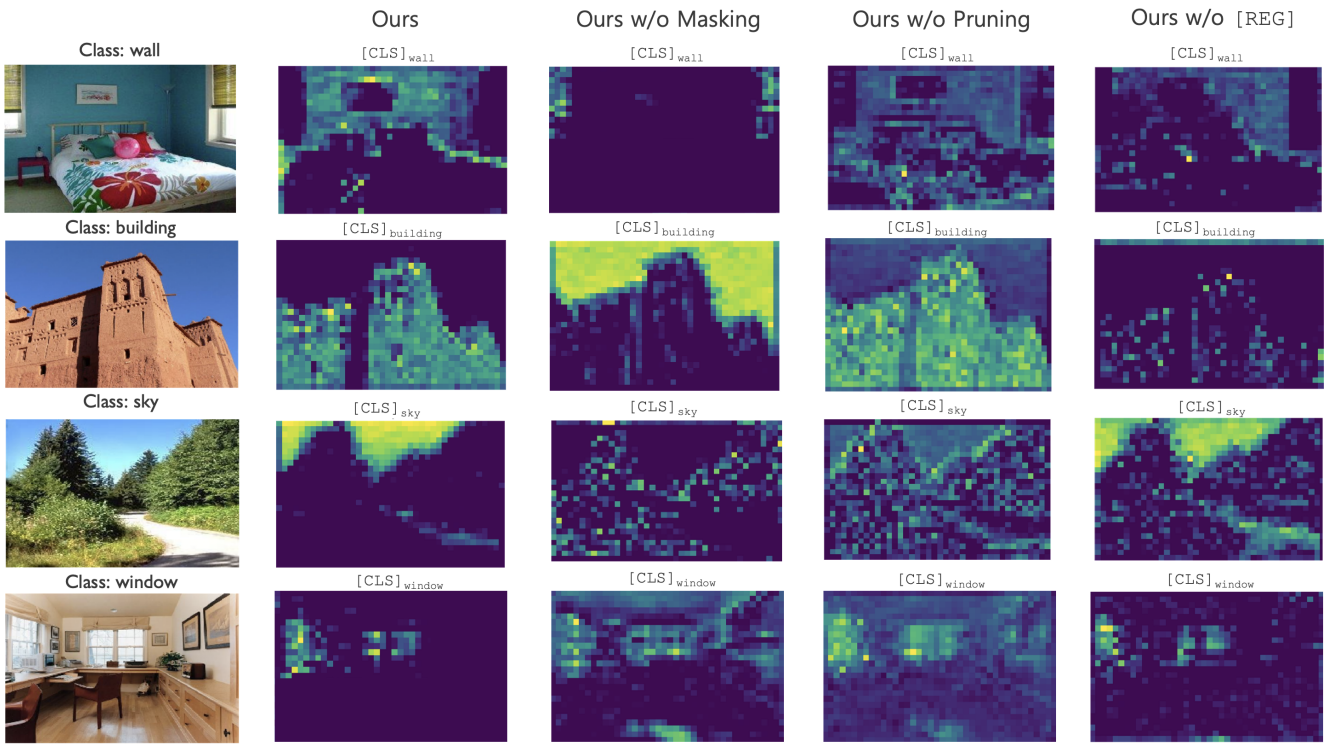


Figure 5. Visual comparison of the impact of different components on the self-attention of [CLS] tokens for the ADE20K dataset. For each image, one class is selected and the reshaped self-attention of the corresponding [CLS] token is displayed. Variations include (a) without random masking, (b) without head pruning, and (c) without appending the additional [REG] token.

token on our model's performance (Section 3.2). Specifically, we compare two configurations: using C [CLS] tokens (one for each class) versus using the same C [CLS] tokens along with an additional [REG] token intended to capture general context. Our results (Table 5) show that incorporating an additional [REG] consistently improves results across datasets. This indicates that the [REG] token provides complementary information that aids in refining class-specific assignments. In Figure 5, the "Ours w/o [REG]" column shows that the absence of the [REG] token results in less precise delineation of features in certain areas. For instance, the boundaries between "sky" and "trees" or "window" and "wall" appear fuzzier and less distinct.

Effect of Attention Head Pruning We explore the impact of pruning attention heads on the performance and interpretability of our Vision Transformer model (Section 3.4). Attention head pruning aims to discard redundant and noisy heads, potentially enhancing the focus of the attention mechanisms. Results are reported in Table 6 and prove that pruning is indeed beneficial. Beyond accuracy, attention head pruning contributes to smoother pseudo-masks, as shown in Figure 5. The "Ours w/o Pruning" column reveals the consequences of retaining redundant attention heads. For example, in the "building" and "sky" classes, the masks appear noisy and fragmented, reflecting the influence of irrelevant attention patterns.

Table 5. Effect of the additional register token ([REG]), with visible improvement on both datasets

	w/o [REG] token	w/ [REG] token
MS COCO	42.8	**43.2**
VOC	72.3	**72.7**

Table 6. Effect of attention head (AH) pruning on the generated pseudo-masks

	w/o AH Pruning	w/ AH Pruning
MS COCO	41.7	**43.2**
VOC	72.0	**72.7**

5.6. Sensitivity Analysis

In this section, we conduct a sensitivity analysis to understand the impact of different masking ratios on the performance of our Vision Transformer model as a classifier and on the accuracy of the generated pseudo-masks. The masking ratio determines the proportion of [CLS] tokens (corresponding to classes that are not in the label) that are ran-

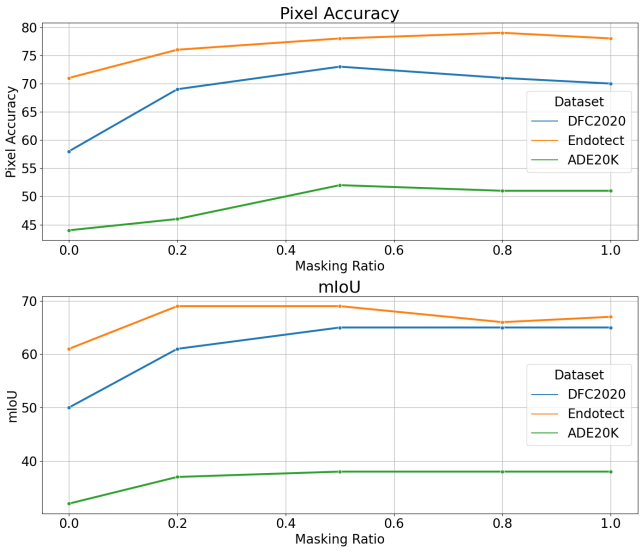


Figure 6. Masking ratios sensitivity analysis for the three domain-specific datasets. Results are reported for the generated pseudo-masks.

domly masked during each training step. We experiment with the following masking ratios: 0% (no masking), 20%, 50%, 80% and 100%. Results are reported in Figure 6. As shown in the figure, increasing the masking ratio generally improves both pixel accuracy and mIoU up to around 50%, after which performance plateaus or slightly declines, indicating an optimal balance between masking and preserving useful class information.

6. General Discussion

The main challenges of using attention maps directly for weakly supervised semantic segmentation without any additional tools are twofold: (i) ensuring that each self-attention map captures distinct object classes in an image and (ii) correctly assigning each of these maps to its corresponding class for pseudo-mask generation. Traditional WSSS methods often struggle with these aspects, leading to coarse or overlapping segmentation masks. Our approach tackles these challenges by leveraging a Vision Transformer architecture with multiple `[CLS]` tokens, where each token corresponds to a specific class. This structured approach encourages the model to develop class-specific representations and enhances interpretability.

A common observation across diverse datasets is that our proposed method consistently produces accurate and well-defined segmentation shapes. This improvement is largely attributable to three key architectural choices: the introduction of an additional `[REG]` token, which captures background and global context; the pruning of redundant attention heads, which sharpens class-specific feature extraction; and the masking strategy, which ensures better class-token assignments by reducing interference between different class representations. These refinements contribute to the generation of high-quality pseudo-masks, which serve as strong initial segmentation candidates.

Another key strength of our approach is its effectiveness in multilabel segmentation tasks, such as ADE20K, where multiple objects coexist within a single image. Unlike single-label segmentation, which relies on coarse feature localization, multilabel segmentation demands precise class-token assignments. Our approach performs very well in these complex scenarios, demonstrating that the combination of multiple `[CLS]` tokens, class-aware masking, and structured attention regularization can lead to robust and interpretable segmentation results.

However, an inherent trade-off emerges: as the number of classes (and consequently, the number of `[CLS]` tokens) grows, so does the model's parameter count. This results in increased computational complexity and memory requirements, posing scalability concerns. While our attention head pruning strategy mitigates some of these burdens by reducing redundancy, optimizing parameter efficiency remains an open challenge. Future research could explore dynamic token assignment strategies, reducing unnecessary computations while maintaining segmentation accuracy. Despite these computational challenges, our results demonstrate that this approach is highly effective when adequate resources are available.

7. Conclusion

To conclude, this works introduces a novel single-stage weakly supervised segmentation method that effectively leverages the self-attention mechanisms of Vision Transformers. By incorporating multiple `[CLS]` tokens, applying random masking, and pruning redundant attention heads, we generate class-specific attention maps that translate into highly accurate pseudo-masks. Our extensive experiments across multiple datasets - including remote sensing, medical imaging, and general scene understanding - demonstrate the robustness and generalization of our approach. Our results show that our method significantly reduces the dependency on fine-grained labels while achieving performance comparable to fully supervised models. As the demand for efficient and scalable segmentation solutions grows, our approach offers a step toward reducing annotation overhead while maintaining high segmentation quality. We believe that further advancements in weak supervision and attention-based learning will continue to push the boundaries of semantic segmentation research.

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4981–4990, 2018. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Herv'e J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9630–9640, 2021. 3

[3] Tianle Chen, Zheda Mai, Ruiwen Li, and Wei-Lun Chao. Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. *ArXiv*, abs/2305.05803, 2023. 2

[4] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2613–2622, 2021. 2

[5] Zhaozheng Chen, Tan Wang, Xiongwei Wu, Xiansheng Hua, Hanwang Zhang, and Qianru Sun. Class re-activation maps for weakly-supervised semantic segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 959–968, 2022. 2, 5, 6

[6] Timoth'ee Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *ArXiv*, abs/2309.16588, 2023. 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. 1, 5

[8] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010. 4

[9] Joelle Hanna, Michael Mommert, and Damian Borth. Sparse multimodal vision transformer for weakly supervised semantic segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2145–2154, 2023. 2, 3, 6

[10] Bharath Hariharan, Pablo Arbeláez, Lubomir D. Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. *2011 International Conference on Computer Vision*, pages 991–998, 2011. 4

[11] Steven Hicks, Debesh Jha, Vajira Lasantha Thambawita, P. Halvorsen, Hugo Lewi Hammer, and M. Riegler. The endotect 2020 challenge: Evaluation and comparison of classification, segmentation and inference time for endoscopy. In *ICPR Workshops*, 2020. 4

[12] Anna Khoreva, Rodrigo Benenson, Jan Hendrik Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1665–1674, 2017. 2

[13] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3992–4003, 2023. 2

[14] Hyeokjun Kweon and Kuk-Jin Yoon. From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19499–19509, 2024. 2

[15] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3159–3167, 2016. 2

[16] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 4

[17] Christos Louizos, Max Welling, and Diederik P. Kingma. Learning sparse neural networks through l0 regularization. *ArXiv*, abs/1712.01312, 2017. 4

[18] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *ArXiv*, abs/1611.00712, 2016. 4

[19] Caleb Robinson, Kolya Malkin, Nebojsa Jojic, Huijun Chen, Rongjun Qin, Changlin Xiao, Michael Schmitt, Pedram Ghamisi, Ronny Hänsch, and Naoto Yokoya. Global land-cover mapping with weak supervision: Outcome of the 2020 ieee grss data fusion contest. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 3185–3199, 2021. 6

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. 5

[21] Simone Rossetti, Damiano Zappia, Marta Sanzari, Marco Schaerf, and F. Pirri. Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation. *ArXiv*, abs/2210.17400, 2022. 5

[22] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16825–16834, 2022. 5, 6

[23] Lixiang Ru, Heliang Zheng, Yibing Zhan, and Bo Du. Token contrast for weakly-supervised semantic segmentation. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3093–3102, 2023. 5, 6

[24] Michael Schmitt, Lloyd Hughes, Pedram Ghamisi, Naoto Yokoya, and Ronny Hänsch. 2020 ieee grss data fusion contest, 2019. 4

[25] Chunfeng Song, Yan Huang, Wanli Ouyang, and Liang Wang. Box-driven class-wise region masking and filling rate guided loss for weakly supervised semantic segmentation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3131–3140, 2019. 2

[26] Nikhil Kumar Tomar, Debesh Jha, Sharib Ali, Haavard D. Johansen, Dag Johansen, M. Riegler, and Paal Halvorsen. Ddanet: Dual decoder attention network for automatic polyp segmentation. In *ICPR Workshops*, 2020. 6

[27] Ozan Unal, Dengxin Dai, and Luc Van Gool. Scribble-supervised lidar semantic segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2687–2697, 2022. 2

[28] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017. 3

[29] Elena Voita, David Talbot, F. Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *57th Conference of the Association for Computational Linguistics*, pages 5797–5808, 2019. 3

[30] Yude Wang, Jie Zhang, Meina Kan, S. Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12272–12281, 2020. 2

[31] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6488–6496, 2017. 2

[32] Yuanchen Wu, Xichen Ye, Kequan Yang, Jide Li, and Xiaoqiang Li. Dupl: Dual student with trustworthy progressive learning for robust weakly supervised semantic segmentation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3534–3543, 2024. 5, 6

[33] Lian Xu, Wanli Ouyang, Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4300–4309, 2022. 2, 6

[34] Sung-Hoon Yoon, Hoyong Kwon, Hyeonseong Kim, and Kuk-Jin Yoon. Class tokens infusion for weakly supervised semantic segmentation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3595–3605, 2024. 5, 6

[35] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2015. 1

[36] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016. 1, 2

[37] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017. 4, 6