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

Supplementary Material

8. Implementation details

Our method, as outlined in the Approach Section (Section 3), is implemented using the PyTorch framework [20]. During training, we use the AdamW optimizer [17] with parameters $\beta_1 = 0.9$ and $\beta_2 = 0.95$. A cosine learning rate schedule with a linear warmup was applied. Each experiment was trained for 200 epochs.

For the Transformer's hyperparameters, we use patch sizes of 16×16 for the MS-COCO, PascalVOC, ADE20K and EndoTect datasets, and 14×14 for the DFC2020 dataset. The model was configured with a depth of 12 blocks and 16 heads. The λ value controlling the pruning was set to 0.01 (Equation 1), and we used a masking ratio of 50% for the `[CLS]` tokens.

Table 7. Model parameters for the specialized dataset

	DFC2020	EndoTect	ADE20K
Image size	224×224	512×640	200×320
Nb. of channels	15	3	3
Nb. of classes	8	23	151
Patch size	14×14	16×16	16×16
Depth	12	12	12
Nb. of heads	16	16	16
Nb. of parameters	87.7M	86.6M	122M

Table 8. Effects of λ (Equation 1) on sparsity rate (percentage of pruned heads) and multi-label accuracy.

		$\lambda = 0$	$\lambda = 0.001$	$\lambda = 0.01$	$\lambda = 0.1$
DFC2020	**Sparsity rate**	0	46	69	78
	Classifier Acc.	86.2	86.2	86.1	84.2
	Pixel Acc.	70.0	71.1	74.1	72.9
	mIoU	59.8	64.3	67.2	65.5
EndoTect	**Sparsity rate**	0	64	79	88
	Classifier Acc.	84.5	84.1	84.0	81.8
	Pixel Acc.	79.3	79.2	78.4	76.0
	mIoU	69.9	69.8	69.8	68.6
ADE20K	**Sparsity rate**	0	55	60	77
	Classifier Acc.	93.5	93.9	94.1	93.0
	Pixel Acc.	47.7	49.9	51.8	48.2
	mIoU	37.3	37.4	38.2	37.8

9. Sensitivity Analysis

In this section, we conduct a sensitivity analysis to examine the impact of enforcing sparsity during training, on the performance of our Vision Transformer model as a multi-label classifier and on the accuracy of the generated pseudomasks. We vary the λ parameter in the objective function (shown in Equation 1). We experiment with the following values: $\lambda = 0$ (no pruning), $\lambda = 0.001$, $\lambda = 0.01$, and $\lambda = 0.1$. Results are reported in Table 8. The resulting models have different numbers of heads retained. The larger λ, the sparser the network becomes.

10. Qualitative Results on the Specialized Datasets

A qualitative comparison against various baselines in Figure 7 for the ADE20K, Figure 8 for the DFC2020 and Figure 9 for the EndoTect medical imaging dataset highlight that our method is the closest to the groundtruth. Across all datasets, our approach results in pseudomasks with more accurate shapes and better class assignments compared to other WSSS methods.

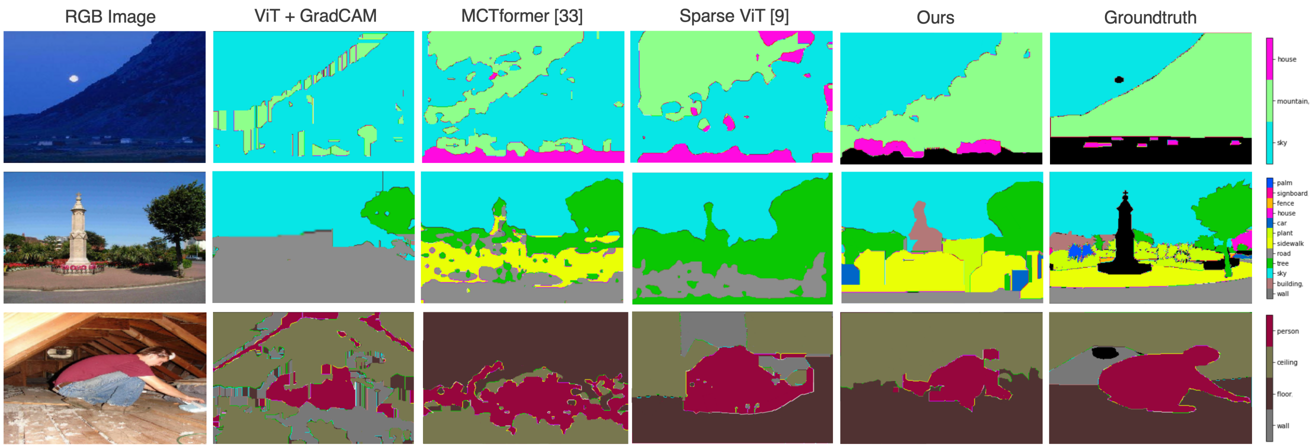

Figure 7. Qualitative comparison of our approach with other weakly supervised methods and the groundtruth, for the ADE20K dataset.

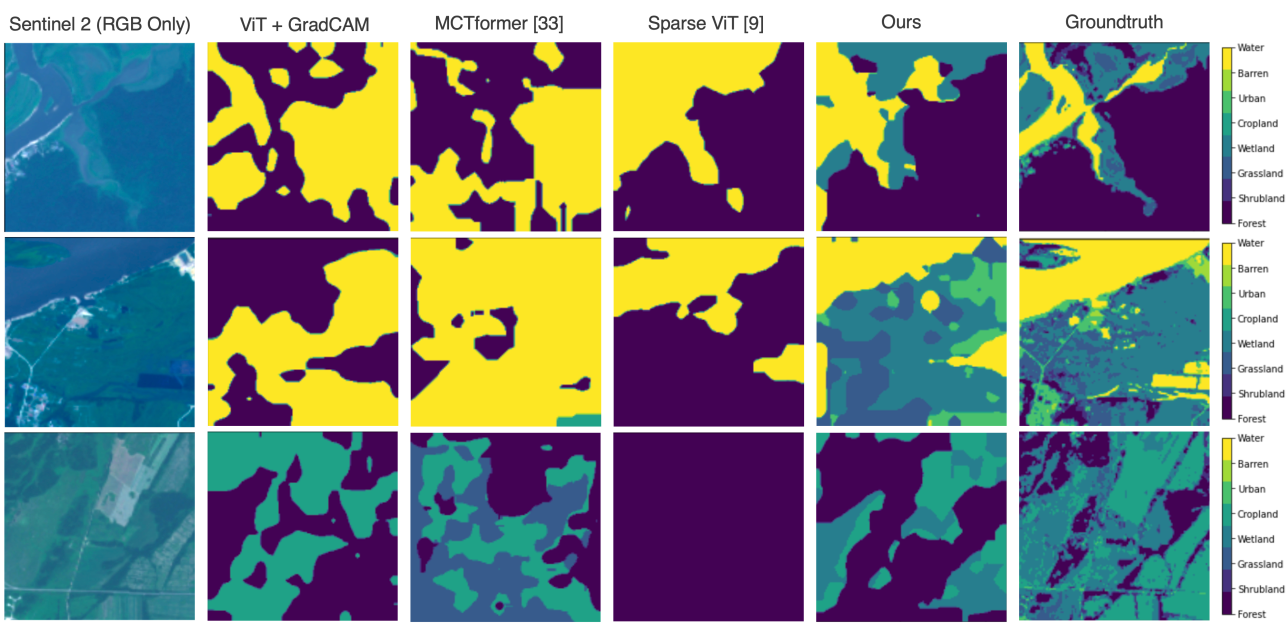

Figure 8. Qualitative comparison of our approach with other weakly supervised methods and the groundtruth, for the DFC2020 dataset.

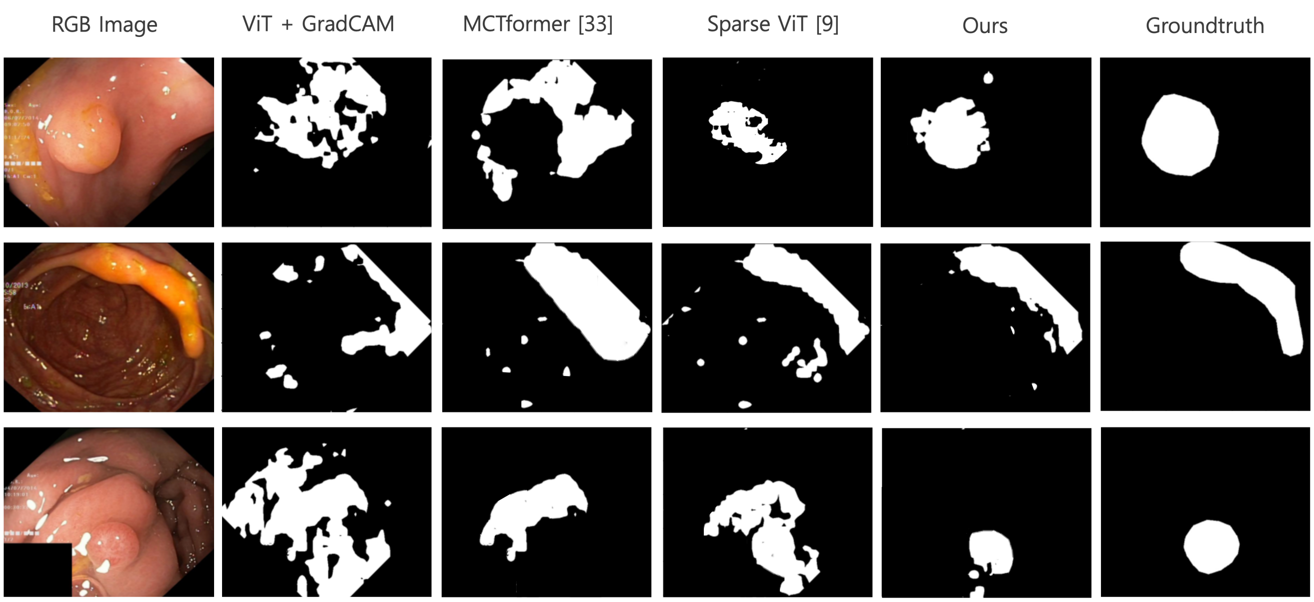

Figure 9. Qualitative comparison of our approach with other weakly supervised methods and the groundtruth, for the EndoTect dataset.