A. Supplementary Material

Here we provide more information about the CAT-Seg model and how we have applied it in this work. Additionally, to demonstrate the adaptability of our method to different backbones and to reinforce the choice of CAT-Seg, we provide a small ablation against a similar CLIP-based open-vocabulary model, Side-Adapter Network (SAN) [41].

A.1. CAT-Seg Overview

CAT-Seg is a CLIP-based segmentation model originally intended as a zero-shot open-vocabulary model for natural images. We chose CAT-Seg in this work with the intention of leveraging its zero-shot open-vocabulary capability for domain adaption to remote sensing datasets.

An overview of the CAT-Seg architecture is shown in Figure 7. CAT-Seg is trained in a full-fine-tuning regime, with the exception of the CLIP image encoder, in which parameter-efficient fine-tuning (PEFT) is enforced. Typically, CLIP is used to output a similarity map between an input image and a text prompt. CAT-Seg generates a similarity score between the input image and each semantic class in the vocabulary, creating a similarity score for every class, stacked to form a "cost volume".

Cost volume features are learned spatially with a single Swin Transformer [24] block, then features across class-wise similarity maps are learned with another Swin Transformer block. The output is passed to a transposed convolution-based decoder that upsamples the aggregated cost features. Finally, for each pixel the cost slice with the highest activation is chosen as the predicted value, resulting in a dense prediction.

The same principle that enables its zero-shot open-vocabulary performance is what makes CAT-Seg attractive for domain adaptation; deep CLIP weights can be leveraged to make reasonable predictions for unseen classes without additional fine-tuning (in the case of natural images). Additionally, the PEFT scheme by which the CLIP encoder is trained is attractive for active learning, as fewer model parameters may lead to more rapid adaptation.

A.2. PEFT with CAT-Seg

The PEFT scheme of CAT-Seg is shown in Figure 2. Within the CLIP encoder, every model parameter is frozen except for the Q and V projection matrices. There is not a strong theoretical basis provided by the authors for this scheme, but they ablate many different PEFT schemes and find that so-called "QK PEFT" outperforms other PEFT schemes as well as full fine-tuning of the CLIP encoder, interestingly.

We do not ablate this PEFT scheme in our transfer learning task because it is out of the scope of this work, which is concentrated mainly on pixel acquisition strategies, but we surmise that the reduction in parameters provided by PEFT can lead to faster convergence in active learning due to reduced complexity than full-fine-tuning of CLIP. It is true that full fine-tuning could conceivably lead to better results after many shots, due to increased capacity for learning the domain shift, but because the focus of active learning is maximizing performance in few shots, we leave full fine-tuning unexplored.

Table 1. Compute cost of entropy and EGL.

Method	Runtime (ms)	FLOPs (G)
Entropy	15	0
EGL	467	33.8
PAGs	482	33.8

A.3. Ablation of PAGs with SAN backbone

Finally, to reinforce our choice of CAT-Seg as a backbone, and to demonstrate the flexibility of the framework, we provide some additional results with a SAN [41] backbone. SAN is similar to CAT-Seg in that it is an open-vocabulary segmentation model with a CLIP backbone. There are two basic differences. First, both CLIP encoders are kept completely frozen, and CLIP features are fused with the learnable adaptor network. Additionally, this method does not aggregate CLIP features, but instead uses CLIP embeddings to propose and rank binary masks, which are fused to form a dense prediction, similar to MaskFormer [6]. We follow the same procedure used for CAT-Seg to adapt SAN to the geospatial domain and for active learning on the target datasets. The results of active learning on Vaihingen at a 1% budget are shown in Figure 8. We first note that zero-shot capabilities of SAN are much weaker than CAT-Seg on this task, in spite of undergoing the same domain adaptation steps. Active learning is generally successful, though all methods are generally less stable than their CAT-Seg counterparts, and converge to a lower mIOU. This may be a result of a weaker starting point or the unique training dynamics of SAN, or both.

A.4. Compute Cost of EGL

EGL incurs a significant runtime cost due to the need for a gradient computation pass. We show the runtime and compute cost of computing entropy, EGL, and PAGs on a single image in Table 1. Metrics are computed for a NVIDIA RTX 6000 GPU. Due to the setting of active learning, the runtime of EGL/PAGs is not a major impediment, as images are not being sampled on a large scale.

A.5. Additional Vaihingen Acquisition Maps

We show a full suite of annotation maps across acquisition methods for an image from Vaihingen (Figure 9), in addition to the set shown for Peru in the primary manuscript.

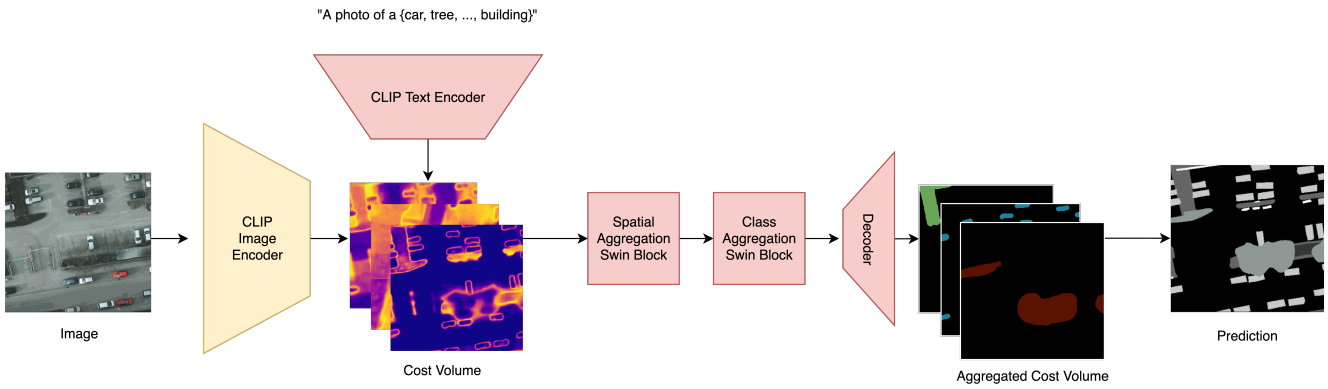

Figure 7. Overview of the CAT-Seg Model. A set of per-class cost maps is aggregated spatially and channel-wise to form a set of probabiliy maps for each class. These probabiliy maps are argmaxed to form the sementation prediction.

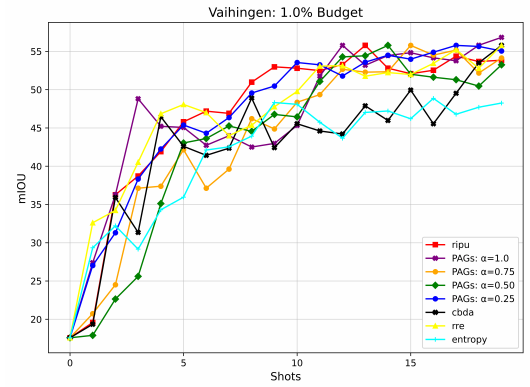

Figure 8. PAGs and baseline results on Vaihingen using SAN as a backbone, 1% labeling budget.

The Vaihingen image is typical of the Vaihingen set, containing large and small buildings, roads, and cars, and is a good choice for evaluating uncertainty due to the representation of classes and abundance of shadows, though Vaihingen images in general do not contain as many semantic regions per image as in Peru.

A.6. Comparison of Linear Combination Variants

Figure 10 compares performance for different linear combinations of PAGs. While $\alpha = 1.0$ is the dominant variant, it is not the best available in some cases. In general, other variants still closely approach $\alpha = 1.0$.

A.7. Tanzania Results

Following Figure 4, we include the results for the Tanzania dataset in Figure 11.

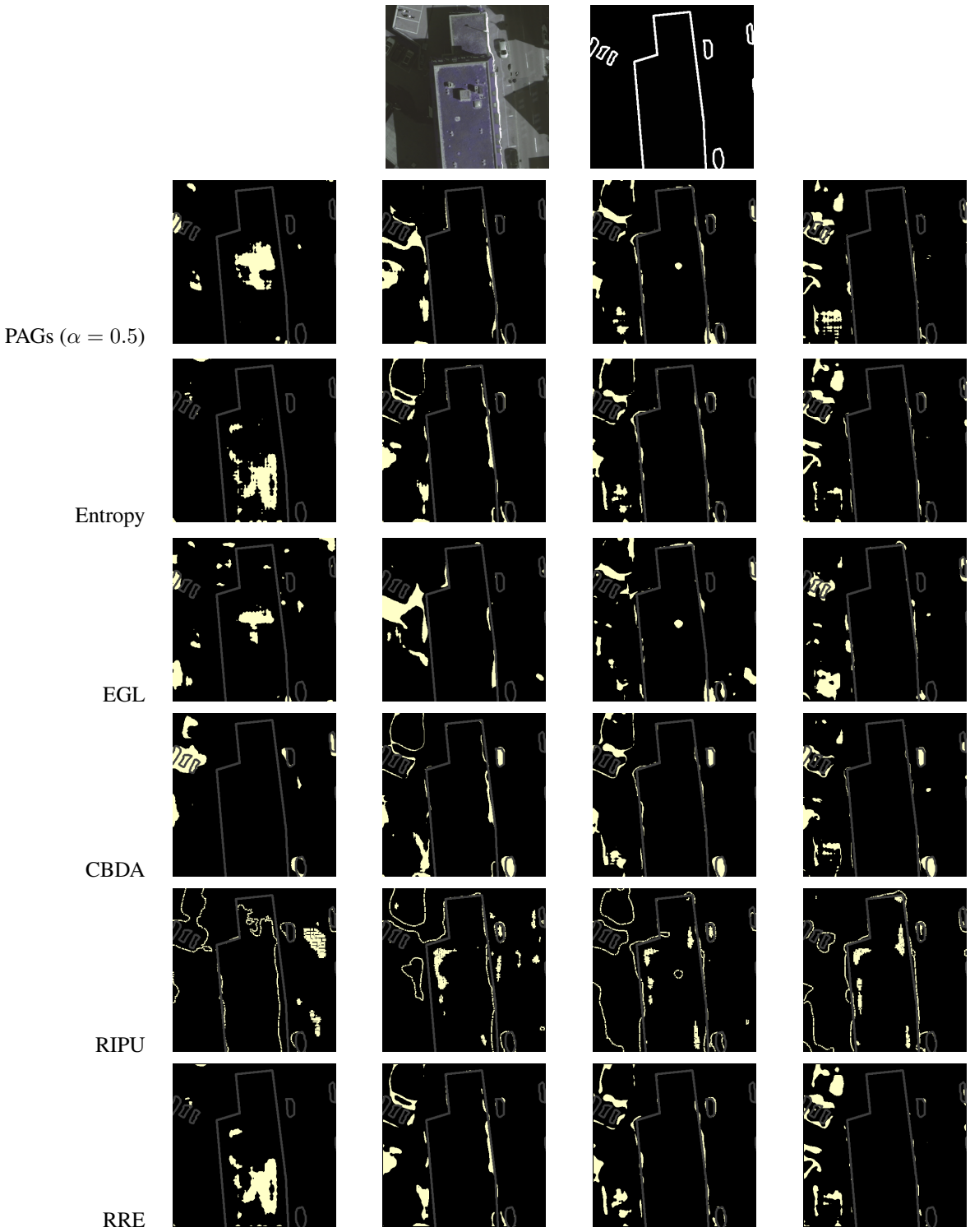

Figure 9. Sample annotation masks from each method at 0, 4, 9, and 19 shots (respectively, from left to right) with a 5% labeling budget. We overlay the ground truth class borders to show how each method approaches class border regions, but this is for illustration only; the methods do not have access to the ground truth labels. The original Vaihingen image and class boundary are shown at the top of the figure for reference

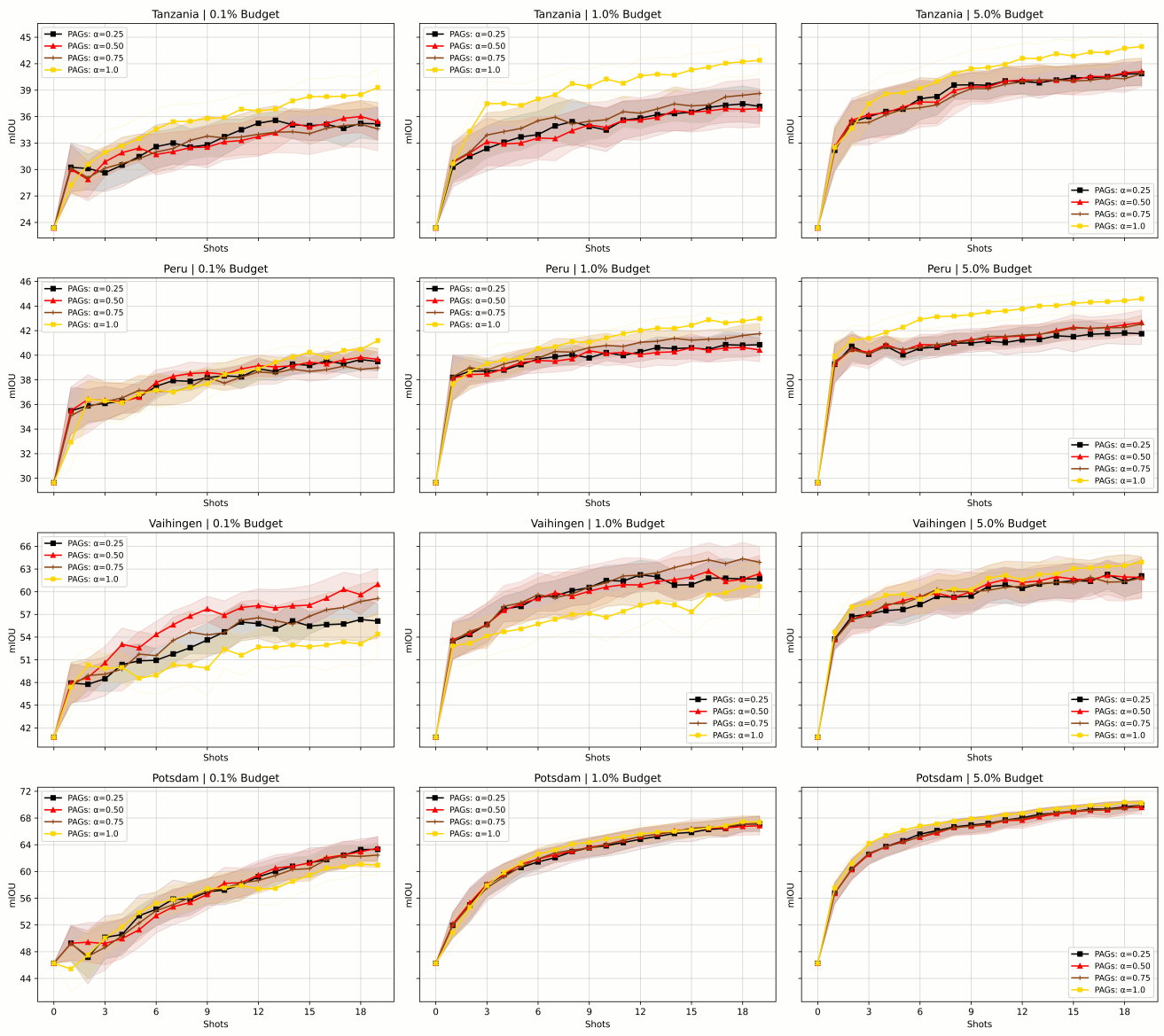

Figure 10. Comparison of different linear combinations of PAGs across datasets and budgets. Each point is the mean of 15 trials, and the surrounding shading is the 95% confidence interval.

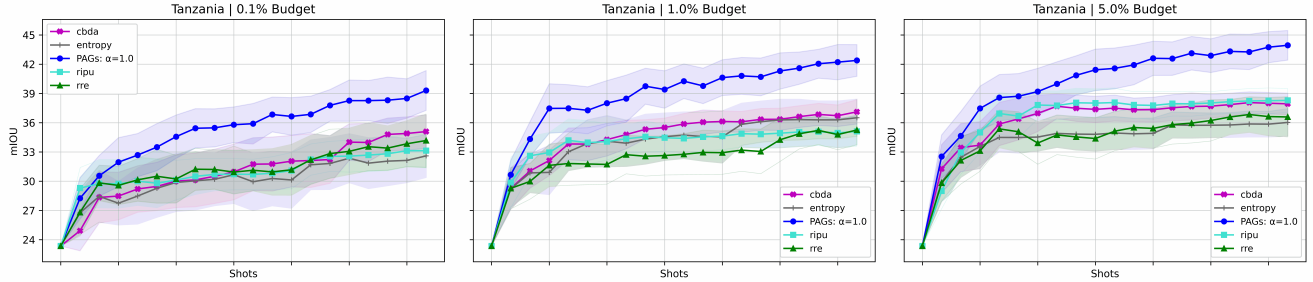

Figure 11. PAGs vs baselines on the Tanzania dataset. Each point is the mean of 15 trials, and the surrounding shading is the 95% confidence interval.