SPAR: Single-Pass Any-Resolution ViT for Open-vocabulary Segmentation

Supplementary Material

6. Training Details

We conduct all training on two `NVIDIA RTX A6000` GPUs using precomputed image features generated by the corresponding teacher model in sliding-window mode with a stride of 24. For feature generation, we use the smallest upsampling factor r such that the upsampled feature maps of cropped windows will have their corresponding image sub-patches fully overlapping. Stitching is then done by simply aligning and averaging feature maps. For $s = 24$, $P = 16$, we choose $r = 2$ which is the smallest value of r for which s is divisible by P/r. Both the up- and down-sampling is bilinear. To accommodate variable input sizes despite using precomputed features, we train with a batch size of 1. Larger batches would require generating teacher features in grouped batches so that all samples share the same sequence length. This would either reduce the effective diversity of the training set, as image tuples would always co-occur in the same context, or introduce additional complexity to gradient accumulation, along with masking in attention to prevent cross-sample interaction. Attention does not usually support unpadded sequences of different lengths in a batch, which would necessitate flattening together all sequences into a single long one. Training is performed in mixed `float16` precision.

By default, for the initial rescaling in image augmentation, we use MMCV's `RandomResize` with `scale=(2048,1024)`, `ratio_range=(0.5,1)` and `keep_ratio=True`. For the extended image range experiments, denoted by †, we adjust `scale=(2560,2560)` and `ratio_range=(0.2,1)`, while still keeping aspect ratio intact with `keep_ratio=True`. The dataset class names and their feature extraction are as in [43] and its official Github implementation. The costs for SPAR SigLIP2 – `ViT-B-16` time-wise include 9 hours for feature extraction on a single A6000 GPU, and 1.5 hours of training on 2 A6000. Feature storage takes about 170GB.

For experiments involving SA-1B [19], we use the first 25k images from archive files `sa_000000.tar`, `sa_000001.tar`, and `sa_000002.tar`, as named in the master file for downloading SA-1B. For the alternate seed used in Tab. 3, we instead sample images sequentially from the randomly chosen `sa_000165.tar`, `sa_000205.tar`, and `sa_000569.tar`.

7. Performance Across Seeds

We report repeated experimental trials to assess the repeatability of SPAR. For each setting, we compute the mean and standard deviation across three independent runs. The aggregated results for individual tuning configurations experiments are presented in Tabs. 5 and 6, and mirror Tabs. 2 and 3 in the main paper. By Mean_6 we denote the average mIoU over six datasets: Voc21, Voc20, Cityscapes, ADE20K, Context60, and Context59. $\overline{\text{Mean}_6}$ and σ_{Mean_6} indicate the mean and standard deviation of Mean_6 over three independent runs.

Training Configuration	Reported Mean_6	$\overline{\text{Mean}_6} \pm \sigma_{\text{Mean}_6}$
SPAR model		
All params	42.5	42.3 ± 0.27
Last block	43.3	43.2 ± 0.14
Last 2 blocks (default)	**43.6**	$\mathbf{43.6 \pm 0.04}$
Last 3 blocks	42.9	43.1 ± 0.15
Patch projection	38.0	38.1 ± 0.11
Positional encoding	39.6	39.6 ± 0.03
Last 2 blocks - `MLP`	42.7	42.8 ± 0.07
Last 2 blocks - `QKV`	41.7	41.7 ± 0.04

Table 5. **Effect of fine-tuning different network elements (mean and std performance across three seeds)**. We show mean mIoU over six datasets, along with the mean and standard deviation across three independent runs, for various training configurations of SPAR-SigLIP2 – `ViT-B-16`. Fine-tuning only the last two transformer blocks, **bolded** for emphasis, achieves the best overall performance and maintains the second smallest standard deviation.

Training Set	Reported Mean_6	$\overline{\text{Mean}_6} \pm \sigma_{\text{Mean}_6}$
SPAR model		
ADE20k+CS+VOC	43.4	43.6 ± 0.27
ADE20k	43.0	43.1 ± 0.12
SA-1B 1.25k (5%)	41.1	41.0 ± 0.09
SA-1B 2.5k (10%)	42.1	41.9 ± 0.31
SA-1B 12.5k (50%)	43.2	43.0 ± 0.15
SA-1B 25k (100%)	43.6	43.6 ± 0.04
SA-1B 25k (diff. subset)	43.6	43.3 ± 0.30
SA-1B 50k (200%)	43.6	43.5 ± 0.10

Table 6. **Impact of distillation data composition and scale (mean and std performance across three seeds)**. We show mean mIoU over six datasets, along with the mean and standard deviation across three independent runs, for SPAR-SigLIP2 – `ViT-B-16` when training on different unlabeled sources. Comparable results across all settings show SPAR does not rely on in-domain data, and performance saturates at 25k distillation images.

8. Measuring Inference Time

In this section, we provide details on how we measure inference time for the experiments reported in Fig. 1. To ensure a fair comparison, we only accumulate the time required for the forward passes needed to process each image. In

single-pass mode, this corresponds to timing the underlying Vision Transformer (ViT) for a single image. For sliding-window inference, we sum the time required to process each sub-batch of window crops. The sub-batch size is set to 60 and is kept constant across experiments; if an image's total number of windows is smaller, they are processed in a single batch. All experiments are conducted on a single `NVIDIA RTX A6000` GPU, and inference time is measured by accumulating the differences between start and end timestamps using the native `time` package. Before timing each forward pass, we perform 10 warm-up passes with the data. Measurement is done on the 11th pass.

9. SPAR with LPOSS

For experiments combining SPAR-SigLIP2 with LPOSS [36], we tune the γ hyperparameter, which controls the Laplacian computation during label propagation. For single-pass inference, we set $\gamma = 10$ to better align with the label distribution produced by SPAR, while for pretrained single-pass and sliding-window inference we retain the default $\gamma = 1$, as we found these settings to perform best for each approach.

10. Vision-only dense prediction tasks details

We utilize the code of [18] with only minor adaptations to enable training and evaluation in a single pass using native image resolutions. We disable scale jittering and cropping augmentations in the training source code: scale jittering would distort images in a way that does not reflect reality, and cropping to a fixed size is inconsistent with our goal of training a transformer capable of processing images at their native aspect ratio. Horizontal flipping remains enabled during training. In Tab. 7, we additionally report linear probing results when using the code's default augmentations and resizing of images: 512^2 for VOC21 and ADE20K, and 1024^2 for Cityscapes. Training the last two blocks and all parameters becomes more similar for VOC21 and ADE20K, while the gap on Cityscapes closes. SPAR still provides a noticeable performance benefit.

We use the official implementation of Hummingbird [2] for KNN segmentation, leaving images at their native resolution and aspect ratio. Due to A6000 memory limitations, we utilize only 50% of VOC21 and 30% of ADE20K training images to construct the index used to classify patches from the evaluation images. Cityscapes uses the full training set, while the other two datasets use the largest subset that avoids out-of-memory errors. We report the mean over three trials, as the training images are randomly sampled, and observe minimal fluctuations: the standard deviation never exceeds 0.4 mIoU points.

The panoptic segmentation experiments report ADE20K results for images resized to 800^2, while Cityscapes is kept at its native resolution, following standard practice. We additionally evaluated other resolutions and observed the same trend: training all parameters yields the highest-quality representations and emerges as a promising approach for future research in resolution-agnostic panoptic segmentation.

SigLIP2 – `ViT-B-16`	VOC21	CS	ADE
Linear Probe - native resolution			
Pre-trained single-pass	67.1	54.1	36.0
SPAR (Last 2 blocks)	**70.2**	56.2	**38.1**
SPAR (All)	68.9	**66.7**	36.5
Linear Probe - repository resolution			
Pre-trained single-pass	71.2	57.0	37.7
SPAR (Last 2 blocks)	74.9	60.9	**40.0**
SPAR (All)	**75.0**	**67.1**	39.1

Table 7. **SPAR linear probing performance using different resolutions**. We show mIoU for linear probing on VOC21, Cityscapes (CS) and ADE20K when utilizing native image resolution vs default resizing of [18] for training and evaluation. Improved results across all benchmarks show SPAR's benefit to visual representation quality. Best result is **bolded**.

GT Panoptic	ADE				CS
	512^2	640^2	800^2	1024^2	1024×2048 (native)
Pre-trained single pass	33.9	34.3	33.6	31.0	31.5
SPAR Last 2 blocks	34.6	35.2	34.0	30.7	28.6
SPAR All parameters	**37.9**	**39.7**	**41.1**	**39.0**	**52.4**

Table 8. **SPAR mask-oracle panoptic segmentation results for different resolutions**. We show PQ for panoptic segmentation on Cityscapes (CS) and ADE20K when using ground truth masks for pooling features [29, 33]. Improved results across resolution show SPAR-All's benefit to visual representation quality. Best result is **bolded**.

11. SPAR and other distillation schemes

In Tab. 9, we explore additional distillation targets. We isolate the effect of multi-resolution training by distilling from a single-pass teacher, for which we precompute the image features and maintain the same setup as described in Sec. 4. During training, the student sees the image additionally bilinearly interpolated by a random factor and attempts to align its features, which are bilinearly up- or downsampled to match the teacher's as needed. As observed, this yields only a +2.9 mIoU average improvement, highlighting the importance of the teacher's multi-context supervision. We also explore using a teacher with a finer stride of $s = 16$, which underperforms by 3 mIoU, emphasizing the benefit of observing pixels in the context of different patches.

To quantify potential benefits that might be missed by not aligning class similarities, we experiment with using the class lists of ADE20K and Cityscapes, respectively. This approach overfits to the domain of the class set used

SigLIP2 – ViT-B-16	Voc21	Voc20	CS	ADE	C60	C59	Mean_6
Pre-trained single-pass	36.1	71.3	23.5	16.8	24.5	26.1	33.1
Multi-resolution distill.	38.0	76.8	26.3	19.0	26.8	29.0	36.0
SPAR w/ Teach. $s = 16$	43.8	77.6	36.2	21.3	31.0	34.0	40.6
SPAR w/ Teach. $s = 24$	47.3	**81.5**	**38.4**	**23.4**	**33.8**	**37.2**	**43.6**
Class sim. distill. ADE	**48.0**	75.4	37.1	23.0	32.8	36.4	42.1
Class sim. distill. CS	44.0	64.8	38.3	19.8	26.0	31.8	37.4

Table 9. **SPAR compared to other distillation schemes**. We show mIoU over six datasets, along with the mean. The results underscore the importance of SPAR using a sliding-window teacher with a stride allowing for non-overlapping patches between windows as well as aligning at the visual feature level. Distilling class similarities overfits to the utilized classes. Best results are **bolded**.

(e.g., using Cityscapes classes yields high performance on Cityscapes but not on other datasets), reaffirming that SPAR does not require any knowledge of the target domain to be effective and in fact benefits from the classless approach.

12. Qualitative Analysis

We provide additional semantic segmentation results in Fig. 6 and PCA visualizations in Fig. 7 for ADE20K [53] and Cityscapes [8]. Utilized backbones, MaskCLIP [54] with OpenCLIP [5] and SigLIP2 [41], are indicated per-row in the figures. The segmentation maps demonstrate how SPAR denoises teacher predictions while preserving semantic alignment. SPAR-OpenCLIP improves delineation of bathroom elements and buildings (1st and 3rd columns in Fig. 6) while suppressing noisy classes on the roads (4th and 5th columns). SPAR-SigLIP2 behaves similarly, yet slightly more robustly. The PCA visualizations show further how SPAR improves inter-object separability and smooths intra-object consistency without losing finer details. The former is visible in the PCA from both backbones, e.g., in the clearer separation of people from the wall or bedroom elements (1st and 2nd columns in Fig. 7), while improved intra-object consistency is most apparent in the interior of the camper (3rd column).

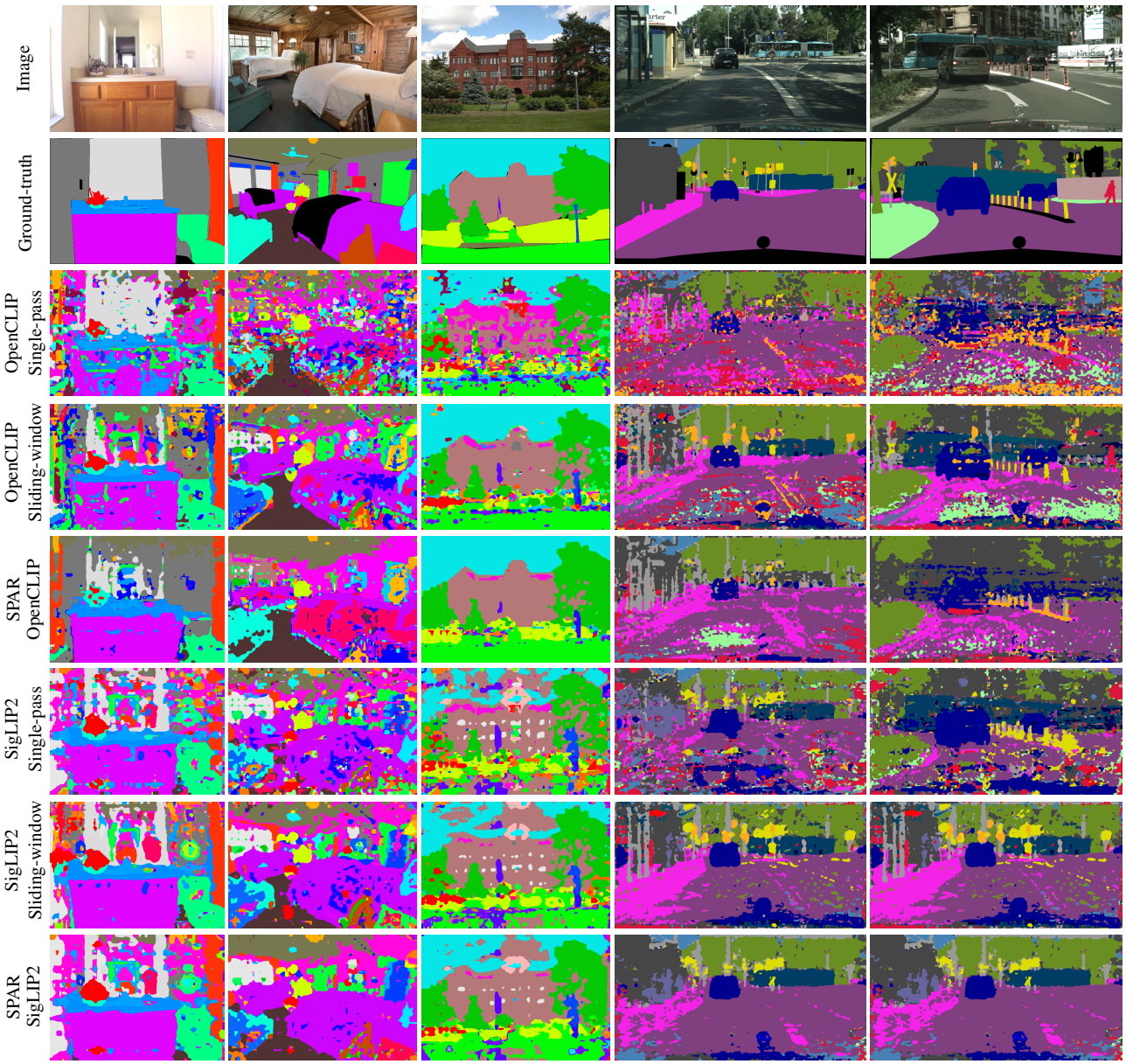

Figure 6. **Qualitative segmentation results.** We show SPAR yields more coherent and smoother predictions than the teacher. Images are from ADE20K [53] (first three columns) and Cityscapes [8] (last two columns).

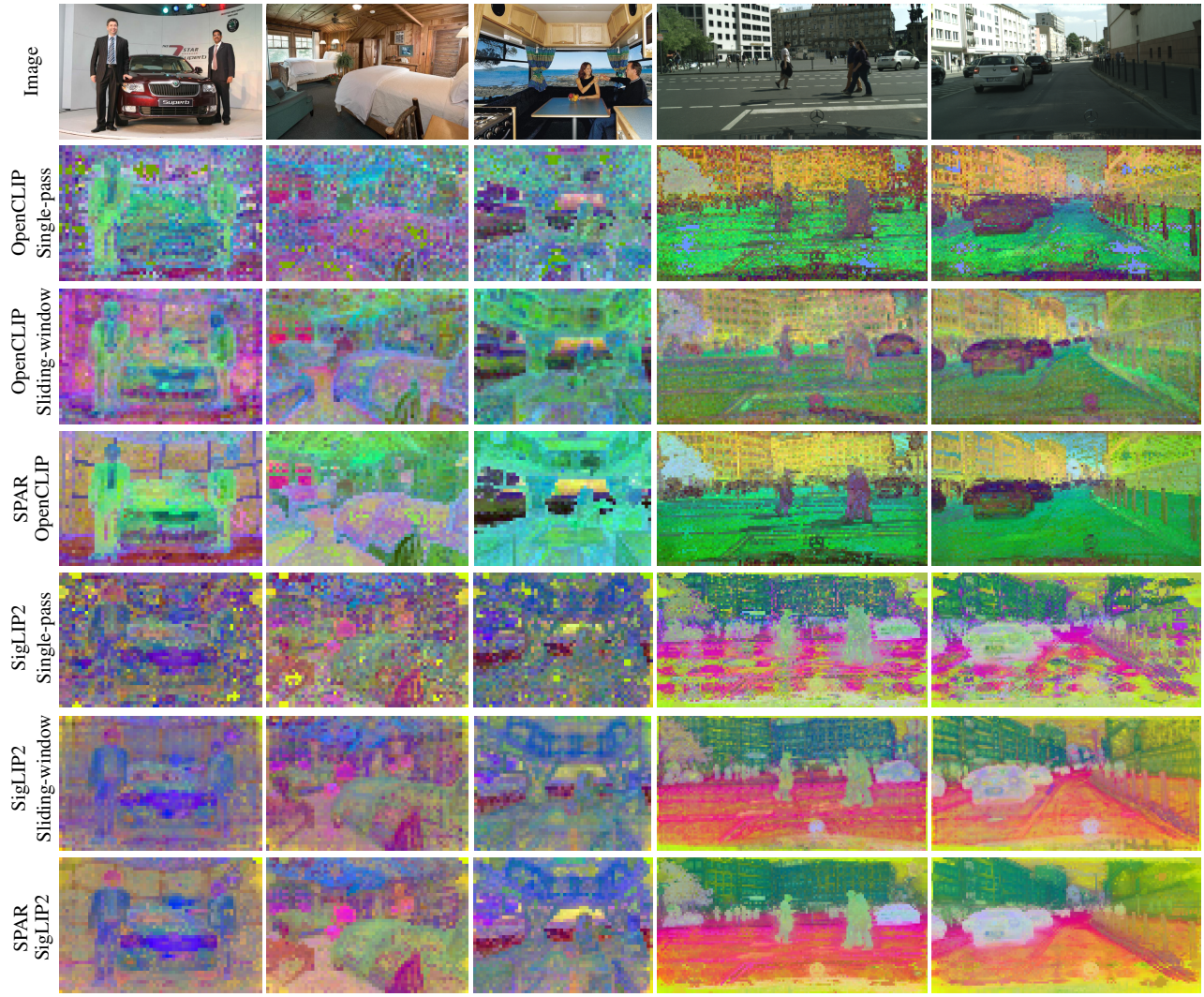

Figure 7. **Feature quality visualization via PCA.** We project features onto the same PCA basis, computed from the teacher's sliding-window representations, to enable a consistent and semantically aligned comparison across models. SPAR improves inter-object separability and smooths intra-object consistency without losing fine details. The former is visible in the clearer distinction between people and the wall behind them (1st column) or pedestrians on the street (4th column), while improved intra-object consistency is most apparent in the interior of the camper (3rd column). Images are from ADE20K [53] (first three columns) and Cityscapes [8] (last two columns).