This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

Naomi Kombol[1] Ivan Martinović[1] Siniša Šegvić[1] Giorgos Tolias[2]

[1]Faculty of Electrical Engineering and Computing
University of Zagreb

[2]VRG, Faculty of Electrical Engineering
Czech Technical University in Prague

Abstract

Foundational Vision Transformers (ViTs) have limited effectiveness in tasks requiring fine-grained spatial understanding, due to their fixed pre-training resolution and inherently coarse patch-level representations. These challenges are especially pronounced in dense prediction scenarios, such as open-vocabulary segmentation with ViT-based vision-language models, where high-resolution inputs are essential for accurate pixel-level reasoning. Existing approaches typically process large-resolution images using a sliding-window strategy at the pre-training resolution. While this improves accuracy through finer strides, it comes at a significant computational cost. We introduce ***SPAR****: **S**ingle-**P**ass **A**ny-**R**esolution ViT, a resolution-agnostic dense feature extractor designed for efficient high-resolution inference. We distill the spatial reasoning capabilities of a finely-strided, sliding-window teacher into a single-pass student using a feature regression loss, without requiring architectural changes or pixel-level supervision. Applied to open-vocabulary segmentation, SPAR improves single-pass baselines by up to 10.5 mIoU and even surpasses the teacher, demonstrating effectiveness in efficient, high-resolution reasoning. Code: https://github.com/naomikombol/SPAR*

1. Introduction

Vision Transformers (ViTs) [11] have become the backbone of modern computer vision, powering foundation models like CLIP [32], DINO [4, 16, 35], and SigLIP [41, 50] through contrastive and self-supervised learning. ViTs process images as sequences of patches using self-attention [42], but due to its quadratic complexity, are typically pre-trained at a single, low resolution. While this suffices for image-level, it limits performance on dense prediction tasks like segmentation, which require fine-grained, per-pixel understanding.

This limitation is particularly severe in training-free open-vocabulary segmentation (OVS), where models must segment categories specified by text without access to pixel-level supervision. OVS approaches leverage ViTs as image feature extractors within Vision-Language models (VLMs), like CLIP [32], for their strong pre-training, but struggle with high-resolution test inputs due to discrepancies with the training resolution. Two common strategies attempt to address this issue: (i) interpolating positional encodings and (ii) sliding-window inference.

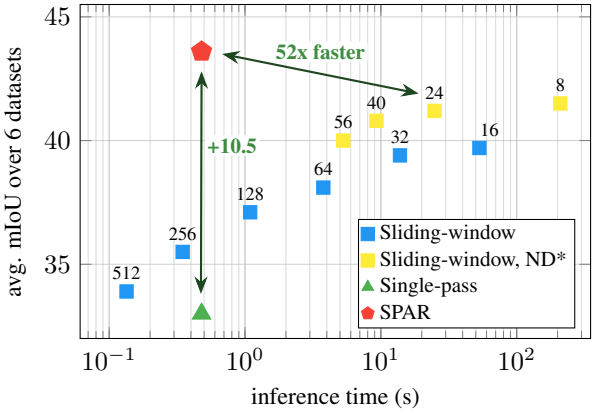


Figure 1. **Performance *vs*. inference time trade-off.** Comparison between the pre-trained SigLIP2 – `ViT-B-16` with single-pass or sliding-window (stride value reported in text labels) inference, and our single-pass SPAR-distilled model. The teacher has sliding windows of size 512 × 512 with stride 24. We report average performance across six datasets along with average inference time for an 1024×2048 image. *ND: stride not divisible by patch size.

The first approach supports single-pass processing and is efficient, but yields poor accuracy. In contrast, we show that the second approach, when using small strides and highly overlapping windows, significantly improves performance. Sliding-window inference enables a patch to appear in multiple contexts and enhances prediction accuracy through aggregation in overlapping regions. Interestingly, strides that are not divisible by the patch size lead to even better results, as sub-patch areas are exposed to diverse contexts.

However, sliding-window inference with small strides increases computational cost, making it impractical for real-world use (see Fig. 1). Despite growing architectural advances [9, 22, 35, 41] designed to handle variable resolutions, they are either not widely adopted or, as indicated in our experiments, have unexplored room for improvement on standard OVS benchmarks.

To address these challenges, we propose SPAR: a method for enabling efficient resolution-agnostic feature extraction in ViTs via teacher-student distillation. SPAR transfers the spatial reasoning of a frozen VLM teacher that operates in a finely-strided sliding-window manner, to a fast, single-pass student. The student learns to align its image features with the teacher's using a regression loss, without requiring architectural changes or additional supervision. As shown in Fig. 1, SPAR surpasses the teacher's performance while maintaining single-pass efficiency.

SPAR uses the unchanged ViT architecture of the VLM and is compatible with a wide range of ViT backbones. During training, the student is exposed to diverse resolutions and aspect ratios, promoting robust generalization. We find that unfreezing only a small subset of layers suffices for strong resolution tolerance. To reduce compute, teacher features are precomputed and reused across training iterations. Our OVS experiments with vanilla zero-shot predictors show that SPAR is effective with ViTs of OpenCLIP [5], SigLIP2 [41], but also DINOv3.txt [35], which already includes components making it resolution resilient.

In summary, we introduce SPAR, a framework for producing resolution-agnostic ViTs that deliver strong accuracy in a single forward pass. SPAR is applied to the vision encoder of common VLMs and requires no architectural modifications or additional labels. Rather than introducing new components, SPAR focuses on a principled training strategy that ingrains resolution flexibility without sacrificing efficiency or feature-space alignment. The resulting model excels in training-free open-vocabulary segmentation across resolutions, matching or surpassing finely-strided sliding-window teachers while being up to $52\times$ faster.

2. Related Work

Open-vocabulary semantic segmentation (OVSS). Unlike standard semantic segmentation [10, 52], open-vocabulary semantic segmentation (OVSS) allows specifying and recognizing arbitrary categories using text. Progress on this task has closely followed advances in Vision–Language Models (VLMs). CLIP [32] introduces contrastive pretraining on web-scale image–text pairs, demonstrating remarkable zero-shot classification capability. Subsequent models, such as ALIGN [15] and the SigLIP family [41, 50], further scale datasets and refine training objectives – the latter transitioning from contrastive formulations to more efficient, sigmoid-based matching losses – thereby improving both semantic alignment and localization. LSeg [23] is the first to adapt CLIP for OVSS by aligning per-pixel visual features with frozen text embeddings. MaskCLIP [54] exposes spatial cues within CLIP's vision tower by modifying its final attention layer to recover fine-grained localization.

Current methods are broadly categorized as *training-based* [6, 10, 14, 16, 23] and *training-free* [21, 43, 51, 54]. Training-based approaches include: (i) methods that correlate text embeddings with regions produced by decoupled mask proposal generators [10, 14, 24, 47, 48]; (ii) models that couple CLIP features with proposal generation [49]; and (iii) variants that directly adapt CLIP for dense prediction [20, 49, 54]. Training-free methods, on the other hand, aim to preserve the pre-trained weights of CLIP without any fine-tuning. Some rectify localization weaknesses within CLIP itself by refining attention calculation and token intermixing [1, 20, 43, 54], others incorporate spatial priors from vision foundation models [21, 34, 51], or aid localization with generative priors [17, 38, 44]. Another line of work performs propagation of features or predictions to improve their spatial consistency [36, 46]. Across all these works, Vision Transformers (ViTs) remain the dominant backbone, and their ingrained resolution sensitivity persists as a shared limitation; one that our method addresses.

Resolution-agnostic transformers. While Vision Transformers (ViTs) are powerful, fixed-resolution pretraining limits their robustness to varying input sizes. The naive remedy of interpolating positional encodings to match new image resolutions causes notable performance degradation [7, 25, 39, 40]. Pix2Struct [22] is among the first to address this issue by fixing token sequence length and resizing images with aspect ratios intact, breaking from traditional sliding-window inference [25, 27]. Building on this idea, several works introduce multi-resolution training to aid robustness. ResFormer [39] trains on square crops of varying sizes and expands positional encodings during attention with neighborhood-focused components. FlexiViT [3] removes fixed patch sizes, randomly sampling them to expose the model to diverse sequence lengths. NaViT [9] factorizes positional encodings along spatial axes, enabling processing of arbitrary-resolution images with native aspect ratios. SigLIP2's *NaFlex* [41] unifies these ideas, combining flexible patching with variable sequence lengths for improved resolution generalization. However, NaFlex remains optimized for image–text alignment at the image- and not patch-level. This limits its use for dense prediction tasks like OVSS, as demonstrated in Sec. 4. Overall, these methods enhance robustness through diverse pretraining resolutions rather than post hoc adaptation.

Recent works continue this trend. ViTAR [13] expands training resolutions via fuzzy positional encoding, using lossy compression to reduce attention costs, while UniViTAR [31] merges image and video modeling with aspect ratio preservation, introducing non-standard ViT changes. OryxMLLM [26] trains on native resolutions using dynamic feature compression. In contrast, our approach adapts ViTs to varying resolutions *post hoc*, without requiring additional labeled data or architectural modifications.

3. Method

3.1. Task Definition

Open Vocabulary Segmentation (OVS) aims to assign a semantic class to every pixel in an image $X \in \mathbb{R}^{3\times H\times W}$. The set of classes $\mathcal{C}$, containing C class names, is arbitrary and can be specified at inference time.

3.2. Preliminary

Vision Transformers (ViTs) process images by dividing them into square patches, projecting each patch into a vector representation, and adding positional encodings based on patch location. These vectors are concatenated into a sequence and passed through transformer blocks. A special CLS token is typically included, but we omit it here.

Foundational ViTs are traditionally pre-trained on square images of fixed $K \times K$ image resolution and consist of an encoder $f : \mathbb{R}^{3\times K\times K} \to \mathbb{R}^{d\times n}$ that produces a feature map $V = f(X) \in \mathbb{R}^{d\times n}$ per input image, where $n = k \cdot k$ is the number of patches, d is the feature dimension, and (k, k) are the spatial dimensions after reshaping into a tensor of size $d \times k \times k$. Due to patching, $k = K/P$, where P is the patch size. Positional encodings are learned for that resolution only, *i.e.* on a $k \times k$ grid of possible positions.

Vision-Language Models (VLMs) combine a vision encoder with a text encoder to map images and text into a shared feature space. Let $F \in \mathbb{R}^{d\times C}$ denote the concatenated text features for all classes in $\mathcal{C}$.

OVS is performed by computing dot-product similarities between normalized image features and text features. The 2D map of class similarities for all classes given by

$$Y(X) = \mathrm{norm}(V)^\top \mathrm{norm}(F) \in [0, 1]^{C\times n},$$

where $\mathrm{norm}(\cdot)$ denotes ℓ_2 normalization applied along the feature dimension, is reshaped and upsampled to the resolution of X using bilinear interpolation. Upsampling is necessary as the feature map V has low resolution due to the patchification of input during inference, which limits segmentation quality, especially for small objects. Operating at higher resolutions, along with arbitrary aspect ratios, is critical for accurate segmentation. Therefore, for image $X \in \mathbb{R}^{3\times H\times W}$ of arbitrary size, we require an encoder capable of $f : \mathbb{R}^{3\times H\times W} \to \mathbb{R}^{d\times n}$, with $n = h \cdot w$, that produces feature maps of spatial resolution proportional to the input size, *i.e.* $h = H/P$, $w = W/P$.

3.3. Pre-trained Baselines

Single-pass inference at arbitrary resolutions is feasible with ViT-based VLMs since self-attention can handle sequences of varying length. The required adjustment is interpolating positional encodings to match the new resolution. However, this approach suffers from performance degradation due to the training-inference resolution mismatch, and due to interpolation, which disrupts the absolute understanding of positional information learned during training.

Sliding-window is a common strategy for processing an arbitrary-resolution image $X \in \mathbb{R}^{3\times H\times W}$. The image is divided into m overlapping, or at least adjacent, windows of size $K \times K$, each processed independently. We denote the image windows by $\{X_{w_i}\}_{i=1}^m$, where $X_{w_i} \in \mathbb{R}^{3\times K\times K}$. The stride s denotes how much we *stride* between neighboring windows horizontally and/or vertically, controlling their overlap.

Feature maps are computed as $V_{w_i} = f(X_{w_i})$, and class similarities as $Y(X_{w_i}) = V_{w_i}^\top F$. The final prediction is obtained by stitching:

$$Y_{\text{stitch}}(X) = \text{stitch}(\{Y(X_{w_i})\}_{i=1}^m) \in [0, 1]^{C\times h\times w}. \quad (1)$$

Stitching refers to merging the individual window predictions into a single coherent output map, typically by averaging overlapping regions and aligning them to their original spatial positions.

This method preserves the native resolution of the vision encoder, avoiding single-pass issues, but incurs higher computational cost proportional to the total number of windows m. Even with batch processing, sliding-window inference is slower than single-pass for small strides. Nevertheless, small strides yield better performance by increasing window overlap, allowing each pixel to be seen in multiple contexts. This redundancy improves robustness through prediction averaging, akin to test-time augmentation.

3.4. Single-Pass Any-Resolution (SPAR)

Performance *vs*. inference time trade-off. Observing the trade-off between speed and accuracy in single-pass versus sliding-window inference (see Fig. 1), we propose a feature distillation approach that combines their advantages. A student model is trained to mimic the feature embeddings of a sliding-window teacher while maintaining the efficiency of single-pass inference. Both models have the same architecture, allowing the student to be initialized with the same weights and keep the same feature space. A method overview is presented in Figure 2.

Sliding-window teacher. Assume $X \in \mathbb{R}^{3\times H\times W}$ is a training image of arbitrary resolution and aspect ratio. Instead of stitching predictions, we stitch features over windows of size $K \times K$ producing the teacher feature map by

$$V_{\text{teacher}}(X) = \text{stitch}(\{f(X_{w_i})\}_{i=1}^m),$$

where stitching refers to merging feature maps $f(X_{w_i})$ from m windows into a single coherent representation aligned to the feature map layout that the single-pass model would produce, by averaging overlapping regions.

Distillation loss. The student model $g : \mathbb{R}^{3\times H\times W} \to \mathbb{R}^{d\times n}$ processes the entire image in a single pass to produce

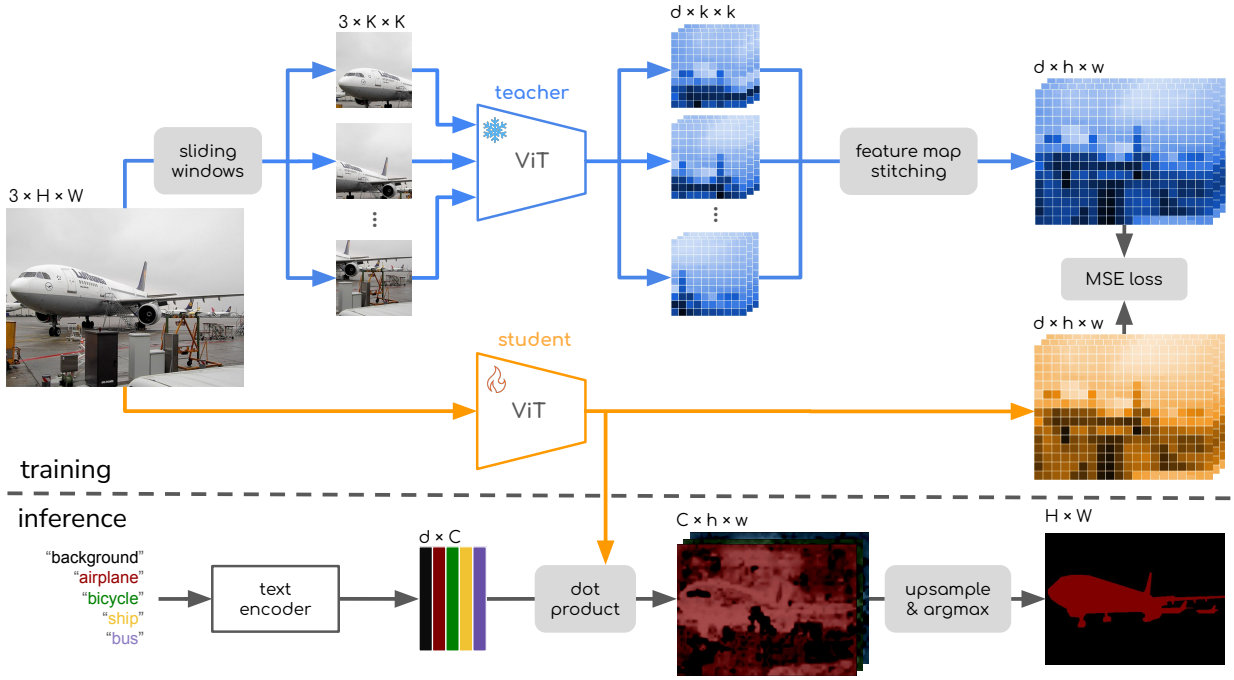


Figure 2. **Overview of SPAR.** During training, the teacher branch uses a frozen foundational vision encoder to generate feature maps via a sliding-window process followed by stitching. Stitching refers to merging the feature maps of overlapping windows into a unified representation aligned with the original image layout. The student branch, initialized from the same pre-trained weights, trains to match the teacher's output using efficient single-pass inference. At inference time, the student model enables fast and accurate segmentation at diverse resolutions and aspect ratios using a single forward pass.

$V_{\text{student}}(X) = g(X)$. The distillation loss minimizes the mean squared error between teacher and student features:

$$\mathcal{L}_{\text{distill}} = \|V_{\text{teacher}}(X) - V_{\text{student}}(X)\|_2^2.$$

Training. The teacher remains frozen during training. We empirically observe that training only the last blocks of the student is a good choice for the standard OVS settings, but training all parameters excels at very large resolution inference and other tasks. Optimization is performed over a dataset of images with diverse resolutions and aspect ratios. No annotations are required since the loss operates solely on feature maps; a generic image dataset suffices.

Training runs for multiple epochs, and teacher feature maps are precomputed and stored to save time. To enable easier feature-level stitching, all training images are resampled bilinearly to have dimensions divisible by the patch size P, and window size K is also divisible by P. If stride s is divisible by P, stitching occurs at the encoder's native feature resolution due to full patch alignment. However, experiments show that using a stride not divisible by P improves performance by exposing pixels to more diverse contexts. In such cases, if s is divisible by P/r, stitching is performed by upsampling all feature maps by a factor of r before merging, while then downsampling by a factor of r after merging to restore the original feature resolution.

4. Experiments

4.1. Implementation Details

Models. We employ 3 pre-trained VLMs. Most experiments use SigLIP2 [41] with the `ViT-B-16` vision encoder, pre-trained on 512×512 images. SigLIP2 employs attention pooling over patch tokens with a learned query to obtain the final image-level representation. We skip the pooling step and directly project the value representation of patch tokens through the output linear layer of the attention-pooling and then the rest of the network. In this way, the patch representation we obtain is compatible with that of the text encoder. For the CLIP [32] experiments, we use OpenCLIP [5] with `ViT-B-16` and native image size 224×224. We follow MaskCLIP [54] and set the last encoder's attention matrix to identity. Lastly, we use DINOv.3txt [35] with `ViT-L-16` and native image size 384×384. Inference is performed in accordance with the original instructions for semantic segmentation [35]. As the image-level representation is a concatenation of a CLS token and the average of patch representations, we keep only the second half of the textual feature dimensions for later similarity comparison.

Methods. For SPAR, we evaluate single-pass inference, and for pre-trained models, we use two variants. The *single-pass baseline*, which has exactly the same runtime complexity as ours, and the *sliding-window baseline* for different values of stride s. In particular, for $s = 24$ this corresponds

Models	Inference	Upsampler	Voc21	Voc20	CS	ADE	C60	C59	Mean_6		
SigLIP2 [41] – ViT-B-16											
NaFlex [41]	single-pass	bilinear	35.8	66.6	22.8	16.2	23.7	25.2	31.7		
Pre-trained	sliding-window	bilinear	45.0	77.5	**38.4**	21.8	30.7	34.0	41.2		↴
Pre-trained	single-pass	bilinear	36.1	71.3	23.5	16.8	24.5	26.1	33.1	↴	
SPAR	single-pass	bilinear	**47.3**	**81.5**	**38.4**	**23.4**	**33.8**	**37.2**	**43.6**	**+10.5**	**+2.4**
LPOSS [36]+Pre-tr.	single-pass	bilinear	46.1	89.6	34.5	19.9	30.9	35.2	42.7	↴	
LPOSS [36]+SPAR	single-pass	bilinear	**51.2**	**89.7**	**39.2**	**25.8**	**34.5**	**39.8**	**46.7**	**+4.0**	
Pre-trained	single-pass	AnyUp [45]	42.4	82.3	33.2	23.1	30.1	34.5	40.9	↴	
SPAR	single-pass	AnyUp [45]	**51.0**	**86.2**	**38.6**	**26.1**	**37.1**	**41.5**	**46.8**	**+5.9**	
OpenCLIP [5] – ViT-B-16											
Pre-trained	sliding-window	bilinear	**48.5**	55.1	**30.8**	15.4	25.2	26.6	33.6		↴
Pre-trained	single-pass	bilinear	43.1	52.3	17.7	11.5	20.3	21.4	27.7	↴	
SPAR	single-pass	bilinear	**48.5**	**57.6**	25.7	**16.2**	**28.2**	**30.3**	**34.4**	**+6.7**	**+0.8**
Pre-trained	single-pass	AnyUp [45]	48.5	**61.0**	20.9	14.8	25.3	26.9	32.9	↴	
SPAR	single-pass	AnyUp [45]	**50.3**	59.2	**26.1**	**17.0**	**29.2**	**32.0**	**35.6**	**+2.7**	
DINOv3.txt [35] – ViT-L-16											
Pre-trained	sliding-window	bilinear	42.6	90.8	39.7	24.9	**31.8**	34.4	44.0		↴
Pre-trained	single-pass	bilinear	**46.0**	89.5	35.9	24.4	32.1	34.8	43.8	↴	
SPAR	single-pass	bilinear	43.1	**91.3**	**40.1**	**25.4**	31.6	**35.0**	**44.4**	**+0.6**	**+0.4**
Pre-trained	single-pass	AnyUp [45]	**46.2**	89.9	32.4	24.9	**32.4**	**35.1**	43.5	↴	
SPAR	single-pass	AnyUp [45]	42.8	**91.5**	**36.1**	**25.6**	31.7	35.0	**43.8**	**+0.3**	

Table 1. **Performance evaluation via mIoU on six datasets.** We report results for 3 backbones comparing SPAR to the pre-trained model and NaFlex [41]. Sliding-window inference uses stride $s = 24$, which corresponds to the teacher model. We combine SPAR with LPOSS [36] to show complementarity with current open-vocabulary semantic segmentation methods. Prediction upsampling is either with bilinear interpolation or AnyUp [45]. Labels as follows: CS: Cityscapes, C60: Context60, C59: Context59. Best results per section are **bolded**.

to the settings of the *teacher model* during distillation. Note that, when evaluating sliding-window processing, stitching happens for class similarities at the pixel-level, while during distillation we stitch together upsampled vision features and resample back down to patch-level as discussed in 3. Additionally, the $s = 256$ case for SigLIP2, corresponds to the commonly used setting in OVS literature [36, 43, 46] of taking half the window size due to its low runtime cost. The window size K is set to the native image size of each model, *i.e.* $K = 512$ for SigLIP2, $K = 224$ for MaskCLIP and $K = 384$ for DINOv3.txt. We further show complementarity by combining SPAR with learnable upsampling by AnyUp [45], and label propagation performed by LPOSS [36]. The former is a feature-agnostic model that operates independently per-feature dimension. Even though it is not its original use, we apply it to upsample class similarities and not the features themselves in all benchmarks, which has reduced complexity, *i.e.* $C \ll D$. LPOSS is a training-free OVSS approach that refines initial patch-based predictions from a vision-language model by propagating VLM-labels (CLIP) across patches in accordance with spatial and DINOv2-semantic similarity. For LPOSS, we replace both models with SPAR-SigLIP2 to highlight its improved spatial coherence. We additionally report NaFlex performance: SigLIP2's native variant to handle images of any resolution. Unless otherwise stated, bilinear interpolation is used to upsample predictions.

Datasets and evaluation. We evaluate on six standard open-vocabulary semantic segmentation benchmarks derived from four datasets: Pascal VOC [12], Pascal Context [30], ADE20K [53] and Cityscapes [8]. VOC21 and Context60 variants include an additional background class over VOC20 and Context59. To allow evaluation for models that have large native resolution, *e.g.* SigLIP2, we bilinearly resize images of all datasets to have their shorter side equal to 672, with the exception of Cityscapes, which is used at its original resolution. We use mean Intersection over Union (mIoU) as the metric of choice for semantic segmentation. By Mean_6 we denote the average mIoU over all six datasets. We train on a subset of 25k images from SA-1B [19], an 11-million-image dataset for class-agnostic semantic segmentation. Please note that we do not use any ground-truth annotations from SA-1B.

Training of SPAR. For training, we chose the teacher with sliding windows and stride $s = 24$ for its good balance of performance and inference time (see Fig. 1). We apply image augmentations once per image, as the stitched feature map from the teacher network is computed once and stored. An image is augmented with random resizing of the shorter side to $[512 - 2048]$ in the default case, and $[512 - 2560]$ in the experiments with larger resolution. We apply axis-independent random cropping, with side lengths from 512 to the maximum possible, and horizontal flipping, each with a 0.5 probability.

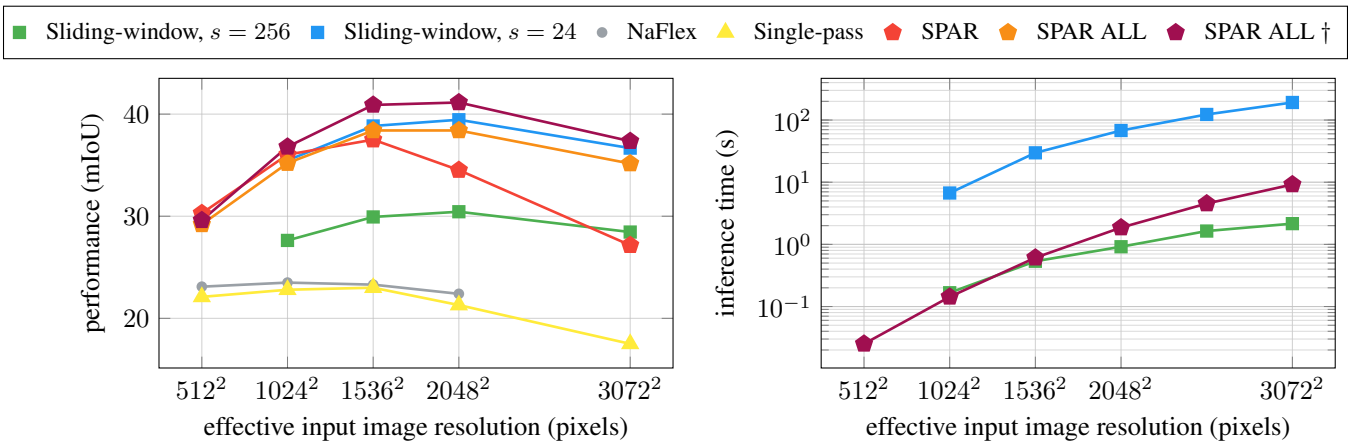


Figure 3. **Performance (left) and inference time (right) *vs*. image resolution.** We compare SPAR to the pre-trained single-pass and sliding-window models, and single-pass NaFlex. SigLIP2 – `ViT-B-16` with native resolution 512×512, and $K = 512$ used. Results on Cityscapes with resolution corresponding to the denoted area. Resolution 512 does not qualify for $K = 512$. Single-pass models, including SPAR, have the same inference time. †: larger resolution used in training (short side up to 2560 pixels instead of 2048). ALL: training the whole network, otherwise only the last 2 blocks.

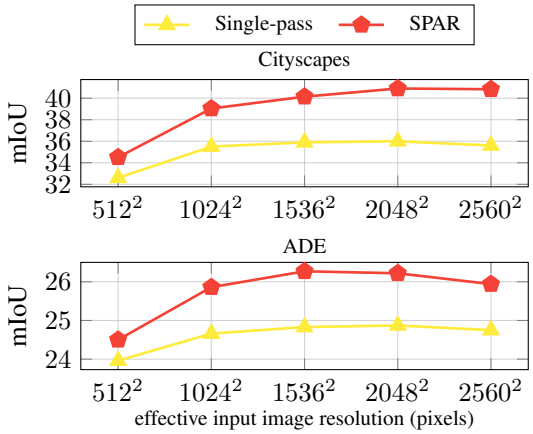


Figure 4. **Performance vs. resolution for `DINOv3.txt`.**

We train for 10 epochs with a constant learning rate $2 \cdot 10^{-5}$ with the AdamW [28] optimizer and weight decay 10^{-4} while tuning only the ViT's last two blocks, unless otherwise mentioned. The default choice for the training set consists of 25k images from SA-1B [19]. More training details in the supplement.

4.2. Main Results

SPAR improvements and compatibility. We report the quantitative results across three different backbones in Tab. 1. Compared to pre-trained single-pass inference, SPAR yields major gains of +10.5 and +6.7 mIoU for SigLIP2 and OpenCLIP, respectively. Moreover, SPAR consistently surpasses the teacher across all models in average performance as well as individually on most datasets.

For DINOv3.txt, SPAR produces only slight gains over the teacher. DINOv3.txt already performs well in single-pass inference, leaving limited room for distillation from sliding-windows. This is likely due to the model's RoPE encodings [37] and an additional high-res fine-tuning stage. Nevertheless, SPAR on Cityscapes, which has larger resolution test images, achieves a substantial gain from 35.9 to 40.1 mIoU, effectively recovering performance on images that deviate most from the training resolution.

Moreover, we see that the SigLIP2 NaFlex variant does not perform competitively compared to either SPAR or the sliding-window baseline, demonstrating its limitations for dense prediction tasks despite its additional training for resolution robustness. AnyUp [45], combined with the single-pass baseline, overall does not manage to exceed SPAR. However, in tandem, it proves compatible and boosts the performance by an appreciable margin, notably by 3.2 on average for SPAR-SigLIP2. Interestingly, AnyUp is not effective when combined with DINOv3.txt. Finally, the combination of SPAR and LPOSS [36] further boosts performance (another +3.1 mIoU), demonstrating how SPAR-trained models are complementary to other OVS methods.

Performance on large resolution inference. Pre-trained single-pass inference and NaFlex do not benefit from larger resolutions, which is not the case for pre-trained sliding-window inference. The default variant of SPAR benefits from increasing effective resolution up to 1.5k^2, and surpasses the sliding-window approach when $s = 256$ but not the more costly $s = 24$. Nevertheless, we observe that training the full network improves SPAR's ability to perform large resolution inference, which is further enhanced by increasing the maximum shorter side length in training from 2048 to 2560. These two choices help with large, but not with small, resolution inference. Overall, SPAR achieves the best performance across all inference-time resolutions.

Training Configuration	#params (M)	CS	ADE	Mean_6
Pre-trained model				
Single-pass		23.5	16.8	33.1
Sliding-windows $s = 24$		38.4	21.8	41.2
SPAR model				
ALL params	93.5	<u>38.5</u>	22.8	42.5
ALL params †	93.5	**40.5**	22.7	42.0
Last block	7.1	36.9	<u>23.3</u>	<u>43.3</u>
Last 2 blocks (default)	14.2	38.4	**23.4**	**43.6**
Last 2 blocks †	14.2	38.5	23.0	42.9
Last 3 blocks	21.3	37.5	23.1	42.9
Patch projection	0.6	25.6	21.4	38.0
Positional encoding	0.8	31.4	21.4	39.6
Last 2 blocks - `MLP`	9.4	37.5	22.6	42.7
Last 2 blocks - `QKV`	3.5	34.9	22.2	41.7

Table 2. **Effect of fine-tuning different network elements**. We show mIoU on Cityscapes (CS), ADE20K, and the mean over six datasets for SPAR-SigLIP2 – `ViT-B-16`. Fine-tuning only the last two transformer blocks, **bolded** for emphasis, yields the best overall performance. #params: number of fine-tuned parameters. †: larger resolution used in training. Best result is **bolded**, second best is <u>underlined</u>.

Training Set	CS	ADE	Mean_6
Pre-trained model			
Sliding-window $s = 24$	<u>38.4</u>	21.8	41.2
SPAR model			
ADE20k+CS+VOC	**39.3**	**23.6**	<u>43.4</u>
ADE20k	37.4	<u>23.5</u>	43.0
SA-1B 1.25k (5%)	35.4	21.9	41.1
SA-1B 2.5k (10%)	36.4	22.4	42.1
SA-1B 12.5k (50%)	38.0	23.2	43.2
SA-1B 25k (100%)	<u>38.4</u>	23.4	**43.6**
SA-1B 25k (diff. subset)	37.7	23.2	**43.6**
SA-1B 50k (200%)	38.1	<u>23.5</u>	**43.6**

Table 3. **Impact of distillation data composition and scale**. We show mIoU on Cityscapes (CS), ADE20K, and the mean over six datasets for SPAR-SigLIP2 – `ViT-B-16` when training on different unlabeled sources. Comparable results across all settings show that SPAR does not rely on in-domain data, and performance saturates with 25k distillation images. Best result is **bolded**, second best is <u>underlined</u>.

Due to the observation in Tab. 1 that with SPAR, DINOv3.txt benefits at larger resolution, we replicate the same experiment for DINOv3.txt in Fig. 4. We observe that SPAR consistently improves DINOv3.txt performance across all resolutions, despite the design and training choices of DINOv3.txt that already promote resolution robustness.

Performance *vs*. inference time. In Fig. 1, we analyze the trade-off between inference time and performance of SPAR SigLIP2 and the pre-trained models for different variants. SPAR achieves higher results than the best sliding-window variant, while being $52\times$ faster than the teacher, and having a $+10.5$ point mIoU performance gain over the pre-trained single-pass model, while keeping the same inference time. For the sliding-window variants, we observe smaller stride values leading to much better performance than the setting of $s = 256$, *i.e*. half the window size which is commonly adopted in the OVS literature due to its low runtime cost. Furthermore, without compromising inference time, we see that a stride value not divisible by patch size gives a good performance boost, justifying its use for the SPAR teacher.

We further analyze the performance and inference time trade-off for increasing image resolution in Fig. 3, with images scaled to the same area as denoted on the axis. Compared to single-pass inference, the sliding-window approach with $s = 24$ incurs a substantial runtime cost, being two orders of magnitude slower. All SPAR variants are noticeably cheaper, providing a great performance-cost trade-off.

4.3. Additional Analyses

Fine-tuning different network layers. In Tab. 2 we summarize the results of training assorted configurations of unfrozen parameters for SPAR built on SigLIP2, reporting Cityscapes, ADE20K and mean mIoU over six datasets. When varying the number of trainable blocks, fine-tuning only the final block already recovers, and slightly surpasses, the teacher's performance, tuning the last two blocks yields a minor extra gain, while tuning more layers slightly degrades performance. To isolate the contribution of specific components, we fine-tune either only the `MLP` or the `QKV` projections. They both perform worse than full block training, with `MLP` being slightly better than `QKV` by using far more parameters. Tuning only the patch projection yields only a minor improvement compared to the single-pass pre-trained model, while training positional encodings achieves noticeably better performance, yet still not enough to surpass the sliding-window teacher. Training all parameters or with larger resolutions does not immediately surpass the normal regime: their value is better displayed in the experiment with larger resolution inference (see Fig. 3).

Influence of the distillation dataset. In Tab. 3, we validate that SPAR is not dependent on in-domain data. When trained on the union of ADE20K, Cityscapes, and VOC training splits, instead of an SA-1B subset, the performance on corresponding validation images drops slightly (-0.2 mIoU). Restricting the data to ADE20K images preserves performance, except on Cityscapes, where larger, unseen image resolutions lead to a $\sim$1.9 mIoU drop. We initially sampled 25k SA-1B (out of 11M) images to roughly match the combined size of ADE20K, Cityscapes, and VOC training splits, and then examined whether smaller subsets suffice. Using 5%, 10%, and 50% of this data, we observe consistent performance gains with larger subsets up to 25k images, beyond which additional samples yield no further improvement. When the 25k SA-1B subset is randomly resampled, results remain consistent, indicating that SPAR is robust to the specific choice of distillation data.

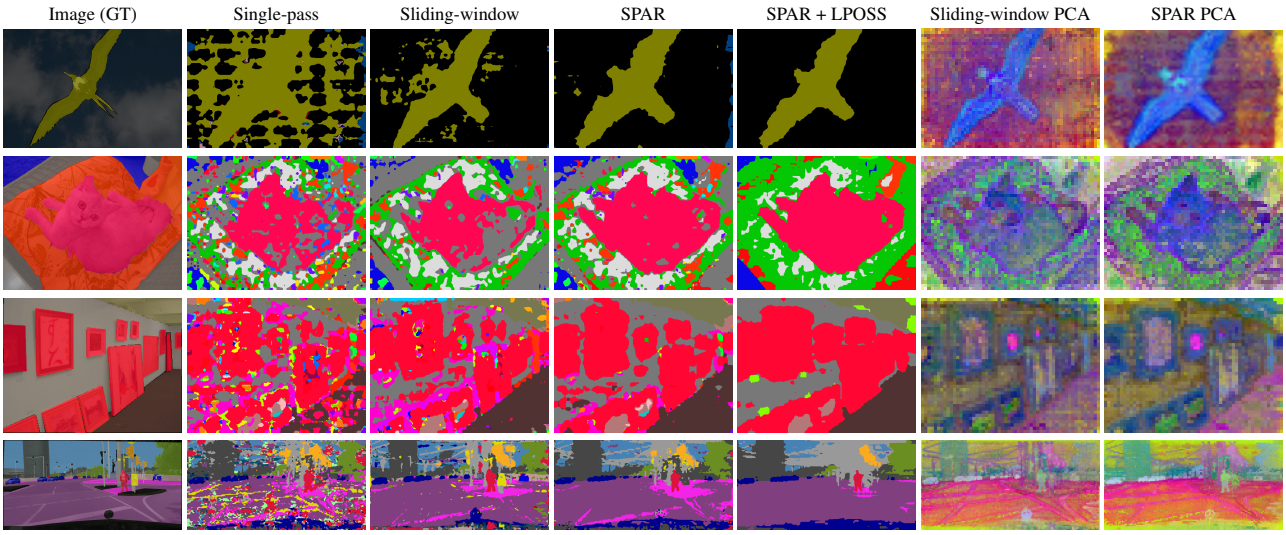


Figure 5. **Qualitative segmentation and PCA analysis results.** We show that SPAR yields less noisy and spatially smoother predictions than the teacher, further improved by LPOSS [36]. Images are from VOC21 [12], Context60 [30], ADE20K [53] and Cityscapes [8].

SigLIP2 – `ViT-B-16`	VOC21	CS	ADE
Linear Probe			
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
Hummingbird			
Pre-trained single-pass	31.9	26.7	13.5
SPAR (Last 2 blocks)	41.0	30.0	16.0
SPAR (ALL)	**54.4**	**39.9**	**22.7**
Mask-Oracle Panoptic Segmentation			
Pre-trained single-pass	N/A	31.5	33.6
SPAR (Last 2 blocks)	N/A	28.6	34.0
SPAR (ALL)	N/A	**52.4**	**41.1**

Table 4. **SPAR on vision-only dense prediction tasks**. We show mIoU for linear probing and Hummingbird [2] segmentation on VOC21, Cityscapes (CS) and ADE20K. Oracle panoptic segmentation [29, 33] shows PQ on the latter two datasets. Improvement across all benchmarks attests to SPAR's benefit to visual representation quality. Best result is **bolded**.

SPAR on vision-only tasks. Our contribution with SPAR has applications beyond OVS, as it operates on the vision encoder only. To demonstrate this, we evaluate on vision-only dense prediction tasks (see Tab. 4), including linear probing, kNN segmentation (Hummingbird [2]), and oracle-guided panoptic segmentation via mask pooling [29, 33]. SPAR consistently improves performance in all settings, with the largest gains on Cityscapes, whose resolution deviates most from pretraining. Full network finetuning (ALL) further enhances performance. Additional details are provided in the supplementary material.

4.4. Qualitative Analysis

Figure 5 presents qualitative results on VOC21, Context60, ADE20K, and Cityscapes for SPAR with SigLIP2. SPAR not only matches, but often surpasses the teacher in spatial smoothness and semantic coherence, producing cleaner boundaries and reduced noise (e.g., the paintings example). PCA visualizations further highlight improved feature structure, with increased inter-object separability and smoother intra-object consistency without loss of fine details. For instance, the cat is more clearly separated from the cushion, while street scenes exhibit sharper transitions that preserve structure. As expected from distillation, errors largely mirror those of the teacher, including misclassifications and noisy regions, though these can be mitigated with complementary methods such as LPOSS. Additional examples are provided in the supplementary material.

5. Conclusion

We improve ViTs for dense feature extraction beyond their fixed pre-training resolution. Our method, SPAR, distills from a costly sliding-window teacher into a single-pass model that tolerates varying image resolutions while being significantly faster. With no additional labeled data or architectural changes, SPAR achieves performance comparable to, or even surpassing, the teacher on open-vocabulary semantic segmentation; gaining up to 10.5 mIoU over the single-pass baseline and reducing inference time by up to $52\times$. We demonstrate the effectiveness of our lightweight post-training technique with three different ViTs: SigLIP2, OpenCLIP and DINOv3.txt. SPAR is additionally shown to be effective in vision-only dense predictions tasks, with the trained models likely to be found useful for applications beyond the ones examined in this work.

Acknowledgements This work was supported by the Junior Star GACR GM 21-28830M, Croatian Recovery and Resilience Fund – NextGenerationEU (grant C1.4 R5-I2.01.0001) and Croatian Science Foundation (grant IP-2020-02-5851). The access to the computational infrastructure of the OP VVV funded project CZ.02.1.01/0.0/0.0/16_019/0000765 "Research Center for Informatics" is also gratefully acknowledged. We thank Tomáš Vojíř, Ivan Grubišić and Noa Garcia, for valuable feedback that helped improve the manuscript. In memory of Siniša Segvić, who passed away before this work was published. He will be missed as a dear colleague, mentor, friend, deep thinker, and moral compass.

References

[1] M Arda Aydın, Efe Mert Cırpar, Elvin Abdinli, Gozde Unal, and Yusuf H Sahin. Itaclip: Boosting training-free semantic segmentation with image, text, and architectural enhancements. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4142–4152, 2025. 2

[2] Ivana Balazevic, David Steiner, Nikhil Parthasarathy, Relja Arandjelović, and Olivier Henaff. Towards in-context scene understanding. *Advances in Neural Information Processing Systems*, 36:63758–63778, 2023. 8, 2

[3] Lucas Beyer, Pavel Izmailov, Alexander Kolesnikov, Mathilde Caron, Simon Kornblith, Xiaohua Zhai, Matthias Minderer, Michael Tschannen, Ibrahim Alabdulmohsin, and Filip Pavetic. Flexivit: One model for all patch sizes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14496–14506, 2023. 2

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 1

[5] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 2818–2829. IEEE, 2023. 2, 4, 5, 3

[6] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 2

[7] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, and Chunhua Shen. Conditional positional encodings for vision transformers. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016. 5, 8, 3, 4

[9] Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36:2252–2274, 2023. 1, 2

[10] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11583–11592, 2022. 2

[11] 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. In *International Conference on Learning Representations*, 2021. 1

[12] Mark Everingham, Luc Van Gool, Christopher Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010. 5, 8

[13] Qihang Fan, Quanzeng You, Xiaotian Han, Yongfei Liu, Yunzhe Tao, Huaibo Huang, Ran He, and Hongxia Yang. Vitar: Vision transformer with any resolution, 2024. 2

[14] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *European conference on computer vision*, pages 540–557. Springer, 2022. 2

[15] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *CoRR*, abs/2102.05918, 2021. 2

[16] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24905–24916, 2025. 1, 2

[17] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion models for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 299–317. Springer, 2024. 2

[18] Tommie Kerssies, Daan de Geus, and Gijs Dubbelman. How to benchmark vision foundation models for semantic segmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 2

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 5, 6, 1

[20] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pages 143–160. Springer, 2024. 2

[21] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pages 70–88. Springer, 2024. 2

[22] Kenton Lee, Mandar Joshi, Iulia Raluca Turc, Hexiang Hu, Fangyu Liu, Julian Martin Eisenschlos, Urvashi Khandelwal, Peter Shaw, Ming-Wei Chang, and Kristina Toutanova. Pix2struct: Screenshot parsing as pretraining for visual language understanding. In *International Conference on Machine Learning*, pages 18893–18912. PMLR, 2023. 1, 2

[23] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 2

[24] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070, 2023. 2

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[26] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx MLLM: On-demand spatial-temporal understanding at arbitrary resolution. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[27] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 6

[29] Ivan Martinović, Josip Šarić, Marin Oršić, Matej Kristan, and Siniša Šegvić. Dearli: Decoupled enhancement of recognition and localization for semi-supervised panoptic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6209–6220, 2025. 8, 2

[30] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2014. 5, 8

[31] Limeng Qiao, Yiyang Gan, Bairui Wang, Jie Qin, Shuang Xu, Siqi Yang, and Lin Ma. Univitar: Unified vision transformer with native resolution, 2025. 2

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 2, 4

[33] Josip Šarić, Ivan Martinović, Matej Kristan, and Siniša Šegvić. What holds back open-vocabulary segmentation? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4256–4266, 2025. 8, 2

[34] Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23487–23497, 2025. 2

[35] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 1, 2, 4, 5

[36] Vladan Stojnić, Yannis Kalantidis, Jiří Matas, and Giorgos Tolias. Lposs: Label propagation over patches and pixels for open-vocabulary semantic segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9794–9803, 2025. 2, 5, 6, 8

[37] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 6

[38] Lin Sun, Jiale Cao, Jin Xie, Xiaoheng Jiang, and Yanwei Pang. Cliper: Hierarchically improving spatial representation of clip for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23199–23209, 2025. 2

[39] Rui Tian, Zuxuan Wu, Qi Dai, Han Hu, Yu Qiao, and Yu-Gang Jiang. Resformer: Scaling vits with multi-resolution training, 2023. 2

[40] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021. 2

[41] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. 1, 2, 4, 5, 3

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1

[43] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European Conference on Computer Vision*, pages 315–332. Springer, 2024. 2, 5, 1

[44] Jinglong Wang, Xiawei Li, Jing Zhang, Qingyuan Xu, Qin Zhou, Qian Yu, Lu Sheng, and Dong Xu. Diffusion model is secretly a training-free open vocabulary semantic segmenter. *IEEE Transactions on Image Processing*, 2025. 2

[45] Thomas Wimmer, Prune Truong, Marie-Julie Rakotosaona, Michael Oechsle, Federico Tombari, Bernt Schiele, and Jan Eric Lenssen. Anyup: Universal feature upsampling. *arXiv preprint arXiv:2510.12764*, 2025. 5, 6

[46] Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, pages 320–337. Springer, 2024. 2, 5

[47] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2955–2966, 2023. 2

[48] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 2

[49] Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023. 2

[50] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1, 2

[51] Dengke Zhang, Fagui Liu, and Quan Tang. Corrclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24677–24687, 2025. 2

[52] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 2

[53] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 5, 8, 3, 4

[54] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European conference on computer vision*, pages 696–712. Springer, 2022. 2, 4, 3