ATAS: Any-to-Any Self-Distillation for Enhanced Open-Vocabulary Dense Prediction

Supplementary Material

1. Semantic Coherence and Fine-Grained Vision-Language Alignment

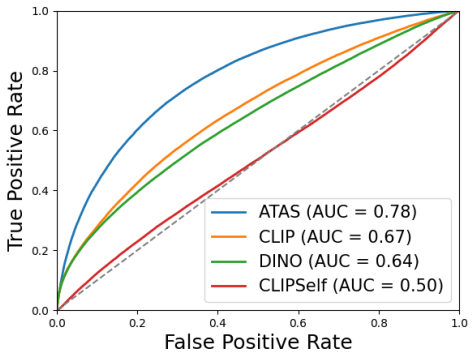

Figure 1. **ROC Curve on Pascal VOC.** illustrating semantic coherence of DINO, CLIP, and fine-tuned CLIP models.

Semantic Coherence There are conflicting perspectives on CLIP's semantic coherence. Studies such as PACL [14] and CLIPtrase [16] emphasize CLIP's inherent ability to effectively align patch-level representations of semantically similar regions. PACL further demonstrates that CLIP outperforms other models, like DINO [3], in maintaining this coherence. In contrast, ProxyCLIP [10] argues that CLIP's coherence is insufficient for aligning with global semantics, limiting its applicability for dense prediction tasks. It suggests that models with stronger semantic coherence, like DINO, should be used to enhance performance. The two methods take different approaches to computing semantic coherence. PACL evaluated semantic coherence based on feature correspondences [6], whereas ProxyCLIP relied on attention scores [1]. Given that our method employs feature correspondences for weighted aggregation, we adopted the PACL approach for measuring semantic coherence.

To explicitly assess CLIP's semantic coherence in Fig. 1, we conducted an experiment measuring its semantic coherence. We followed the experimental settings proposed in PACL [14] with some modifications due to practical constraints. Specifically, we computed patch embeddings using the CLIP vision encoder on the Pascal VOC dataset and predicted whether two patches belong to the same class based on cosine similarity.

Patch labels were assigned using majority voting, based on the pixel-wise class annotations provided by the annotation map. For each class in a batch of 128 images, we sampled 8 patches per image, resulting in comparisons between patches from different images within the same batch. The cosine similarity was computed for all paired samples. For pairs belonging to the same class, the target label was set to 1, whereas for pairs from different classes, the target label was set to 0.

Model	Accuracy
CLIP	69.48
CLIPSelf	70.66
ATAS	82.46

Table 1. **Patch-level classification accuracy.** indicating fine-grained alignment of CLIP and fine-tuned CLIP models.

Top-k	Number of Grids	Image Size	Average Accuracy
10	1	256	33.02
	2	512	17.54

Table 2. **Average Accuracy per Grids from Scene-Centric Dataset.**

Fine-Grained Vision-Language Alignment During the pretraining of CLIP, contrastive learning utilizes the image CLS token and text CLS token, without explicitly using local patches. Consequently, many prior works have aimed to enhance CLIP's dense prediction capabilities by focusing on aligning local patches. To empirically evaluate the degree of alignment in CLIP in Tab. 1, we conducted an alignment experiment as proposed in PACL. We predicted the class of each patch by identifying the class whose text embedding has the highest cosine similarity to the patch embedding. For this experiment, we utilized all patch embeddings for each class from every image in the Pascal VOC dataset and assigned patch labels using the same method employed in the semantic coherence experiments.

2. On the Selection of Data and Augmentation

2.1. Rationale for Choosing an Object-Centric Dataset

Our goal is to enhance the performance of dense prediction tasks by employing a self-distillation strategy for CLIP's vision encoder [15]. In our framework, both the global image embedding and local embeddings from the teacher model are

Datasets	Methods	Resolution	Segmentation	Detection
COCO	CLIPSelf	1024	33.8	42.5
	ATAS	1024	37	44.89
ImageNet	CLIPSelf	960	36.98	45.70
	ATAS	384	37.74	40.85
		576	38.51	43.59
		960	**39.34**	**46.38**

Table 3. **Performance Comparison Across Datasets and Resolutions.**

Mosaic	Segmentation mIoU	Detection AP_{50}
No mosaic	39.41	35.95
4	38.33	45.94
2,4	**39.75**	45.85
2,4,6	39.34	**46.38**

Table 4. **Effect of different mosaic augmentation strategies on dense prediction performance.**

used as pseudo-labels. The teacher's global embedding is transferred to both the student's global and local representations, so ensuring its semantic reliability is critical.

However, dense prediction models are often trained on scene-centric datasets, such as COCO [11], which we find to be an unreliable source of high-quality pseudo-labels, especially for the global embedding. To quantify this issue, we evaluated the region-level Top-10 accuracy on scene images, as shown in Tab. 2. For this experiment, each image is divided into grids, and each grid cell is classified via its CLS token. A prediction is considered correct if any of the top-10 predicted classes for a segment match the ground-truth labels of the entire image. This evaluation revealed a sharp drop in accuracy, from 33.02% for the whole image to just 17.54% when using a 2×2 grid. This result quantifies the poor semantic consistency within scene-centric images and underscores their limitations in generating reliable pseudo-labels for self-distillation.

The benefit of object-centric datasets for self-distillation is evident in Tab. 3. Notably, our best performance occurs when using ImageNet, indicating that object-centric supervision provides more reliable signals for dense prediction. CLIPSelf shows the same trend, performing better with ImageNet than with COCO. This underscores that object-centric data is essential for obtaining a stable global embedding, which in turn guides reliable learning of both global and local features.

2.2. Motivation for Mosaic Augmentation

We investigate the effectiveness of using object-centric datasets for dense prediction tasks. However, as shown in Tab. 4, using them without mosaic augmentation leads to a significant drop in detection accuracy. This performance gap stems from the limited scene complexity of object-centric datasets, which typically contain fewer objects per image

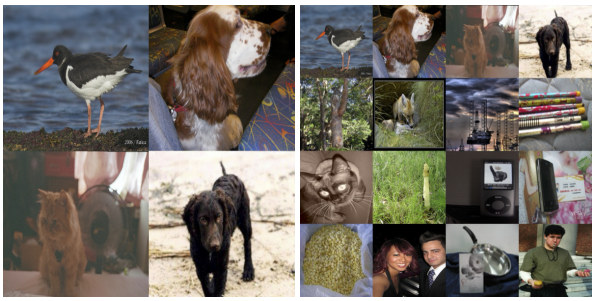

Figure 2. **Mosaic images for training.**

than scene-centric datasets. As a result, models trained solely on such datasets struggle with dense prediction in complex, multi-object scenes. To address this limitation, we apply mosaic augmentation to synthesize training samples that include multiple objects. We experiment with various mosaic configurations (2×2, 4×4, and 6×6), and observe that combining these yields the best performance. These improvements validate our hypothesis that multi-object awareness plays a crucial role in dense prediction.

2.3. Implementation and Visualization of Mosaic Augmentation

In this paper, we use a combination of 2×2, 4×4, and 6×6 mosaics. During training, we utilized 36 images to construct mosaic images $x_{\text{msc}} \in \mathbb{R}^{3\times960\times960}$ for ViT-B, which we randomly organized into either (1) a single 6x6 mosaic image with 36 individual images, each of size $x_i \in \mathbb{R}^{3\times160\times160}$; (2) two 4x4 mosaic images, each containing 16 individual images of size $x_j \in \mathbb{R}^{3\times240\times240}$ and a single 2x2 mosaic image with 4 individual images, each of size $x_k \in \mathbb{R}^{3\times480\times480}$. An example of mosaic augmentation is shown in Fig. 2. The image on the left is an example of 2×2 mosaic image using four 480×480 images and the image on the right is an example of 4×4 mosaic image using 16 240×240 images.

3. Ablation-setting

We conducted ablation experiments on two tasks: zero-shot scenarios in open vocabulary semantic segmentation and open-vocabulary object detection. For both tasks, we utilized the CLIP Vision Transformer(ViT) base model and applied our proposed ATAS framework. The weights of the CLIP image encoder were replaced with our trained models before ablation studies.

In semantic segmentation, we adopted the MaskCLIP [21] approach. Experiments were performed using this approach on three datasets—Pascal-VOC [5], Pascal-Context [13], and COCO-Stuff [2]—using the average mIoU as evaluation metric.

In object detection, we evaluate the detection performance of

Method	Model	VOC20		PC-59		COCO-Stuff		ADE20		CityScapes		Average	
		mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc	mIoU	mAcc
ClearCLIP	CLIP	80.9	90.4	**35.9**	57.4	**23.9**	42.6	16.7	35.5	**30.0**	**49.8**	**37.5**	55.1
	CLIPSelf	74.4	85.2	30.4	53.5	19.9	42.0	15.3	36.3	23.5	40.9	32.7	51.6
	ATAS	**82.4**	**92.5**	33.8	**59.6**	22.8	**47.3**	**17.0**	**40.7**	28.5	49.7	36.9	**58.0**

Table 5. **Results of open-vocabulary semantic segmentation in *zero-shot* scenarios.** The model that achieved the best performance is marked in **bold**, while the model with the second highest performance is marked with underline

F-ViT using OpenAI CLIP ViT-B, trained for 3 epochs with a batch size of 64 on the OV-COCO dataset. The evaluation considers Average Precision (AP) across all categories, including both base and novel classes.

4. Additional results of ClearCLIP

In this paper, we investigate the zero-shot scenario for the open-vocabulary semantic segmentation task, leveraging the intrinsic capabilities of the CLIP Vision Transformer (ViT) for dense prediction tasks. For this experiment, we applied three models—CLIP [15], CLIPSelf [18] and ours—across four different methods: (1) vanilla CLIP, (2) MaskCLIP [21], (3) SCLIP [17] and (4) ClearCLIP [9]. We utilized the original implementation of all methods, except for ClearCLIP.

ClearCLIP provides an in-depth analysis of the CLIP ViT architecture and proposes three architectural modifications to improve segmentation performance: (1) removal of the residual connection, (2) query-query self-attention, and (3) removal of the feed-forward networks. To extract dense features x_{dns} from image $x \in \mathbb{R}^{N\times H\times W\times C}$ in CLIP ViTs,

$$q = \mathbf{Proj}_{\text{q}}(\mathbf{LN}(x)), v = \mathbf{Proj}_{\text{v}}(\mathbf{LN}(x)) \tag{1}$$

$$x_{\text{dns}} = \mathbf{Proj}(\mathbf{Attn}(q, q) \cdot v) \tag{2}$$

where Proj denotes projection layer, LN express layer normalization and Attn represents self-attention.

These methods exhibit some architectural differences compared to our approach. We utilize MaskCLIP [21] method to extract dense features,

$$v = \mathbf{Proj}_{\text{v}}(\mathbf{LN}(x)) \tag{3}$$

$$x_{\text{attn}} = x + \mathbf{Proj}(v) \tag{4}$$

$$x_{\text{dns}} = x_{\text{attn}} + \mathbf{FFN}(\mathbf{LN}(x_{\text{attn}})) \tag{5}$$

Where FFN denotes feed-forward network.

Unlike ClearCLIP, our method performs self-attention and trains the feed-forward network. To minimize performance variations resulting from structural differences, we conducted experiments with two different versions of ClearCLIP.

$$x_{\text{attn}} = x + \mathbf{Proj}(\mathbf{Attn}(q, q) \cdot v) \tag{6}$$

While the results in Tab. 5 are based on the original ClearCLIP method without residual connection. ATAS achieved superior performance in both versions. Notably, when local representations were extracted using the original ClearCLIP method in CLIPSelf, a drop in mIoU was observed across all datasets compared to OpenAI CLIP. In contrast, our method exhibited more robust performance, with some datasets even showing relative improvements. This highlights the robustness of our approach in effectively managing the architectural differences in the methods used for local representation extraction.

5. Exploring Architectural Differences in CAT-Seg v1 and v2

In this paper, we used two versions of CAT-Seg [4] for the open-vocabulary semantic segmentation task. The CATSeg used in CLIPSelf [18] underwent architectural modifications and a refined model is now publicly available. To ensure fairness in our comparison, we conducted experiments using both versions of CAT-Seg.

Both CAT-Seg v1 and CAT-Seg v2 are trained on the COCO-Stuff [2] dataset. Two versions employ a cost-aggregation framework designed to align semantically similar local and global representations. While both versions leverage CLIP vision transformer(ViT) for extracting local features, CAT-Seg v1 also incorporates auxiliary feature extraction models: ResNet-101 [7] and Swin Transformer [12], which are known to capture more localized representations compared to the CLIP ViT. Specifically, ResNet-101 was paired with ViT base, while the Swin Transformer was paired with ViT large. In contrast, CAT-Seg v2 relies solely on CLIP ViT for cost aggregation, achieving superior performance compared to CAT-Seg v1. This suggests that CLIP's intrinsic properties contribute effectively to semantic segmentation tasks. Additionally, CAT-Seg v2 fine-tunes the CLIP text encoder during training, whereas CAT-Seg v1 keeps it frozen.

We executed experiments comparing CLIP [15], CLIPSelf, and ours on CAT-Seg. For CLIP and CLIPSelf, we utilized pre-trained weights, while our method was trained from scratch. All three models—CLIP, CLIPSelf, and ours—were applied to CAT-Seg and trained from scratch. Our approach demonstrates enhanced performance over both methods, while CLIPSelf showed a performance drop in CAT-Seg v2. This indicates that

Method	ADE-150	ADE-847	PC-59	PC-459	PAS-20	PAS-20b	Average
SAN [20]	33.3	13.7	60.2	17.1	95.5	-	-
SED [19]	35.2	13.9	60.6	22.6	96.1	-	-
CAT-Seg (v1)	31.5	10.8	62.0	20.4	96.6	81.8	50.5
CAT-Seg (v1) + CLIPSelf	34.8	12.4	61.7	20.8	96.8	81.0	51.3
CAT-Seg (v1) + ATAS	34.1	12.4	62.5	21.3	96.6	81.5	51.4
CAT-Seg (v2)	37.9	15.9	**63.0**	**24.0**	96.7	81.9	53.2
CAT-Seg (v2) + CLIPSelf	37.6	15.7	**63.0**	23.9	96.7	82.0	53.2
CAT-Seg (v2) + ATAS	**38.0**	**16.2**	62.7	**24.0**	**96.9**	**82.1**	**53.3**

Table 6. **Results of open-vocabulary semantic segmentation (mIoU) with CAT-Seg using the ViT-Large model.**

Method	OV-COCO			OV-LVIS			
	AP_{50}	AP_{50}^{base}	AP_{50}^{novel}	mAP	mAP_c	mAP_f	mAP_r
F-ViT [18]	31.4	36.9	16.0	14.1	11.4	19.7	8.3
F-ViT + CLIPSelf [18]	42.5	46.9	29.8	20.1	16.2	23.8	21.6
F-ViT + ATAS	46.4	50.5	**34.7**	20.7	16.4	24.8	**22.1**

Table 7. **Detailed results of open-vocabulary object detection with OpenAI CLIP on OV-COCO and OV-LVIS.**

CLIPSelf heavily relies on specific training methods. In contrast, our approach emphasizes the importance of promoting CLIP's semantic coherence and fine-grained vision-language alignment, which enhances its performance on dense prediction tasks.

6. Detailed Results for CAT-Seg with ViT-Large model

To evaluate whether our approach performs efficiently on large models, we conducted the experiment on a ViT-Large model using the CAT-Seg. As shown in Tab. 6, we observed consistent performance improvements regardless of the training method. As detailed in Sec. 5, each version of CAT-Seg is trained with a different network architecture. In particular, our model, when combined with CAT-Seg v2, achieved the highest performance on most datasets compared to other models. This demonstrates that our approach is effective regardless of the model size and can be applied to various CLIP-based methodologies for dense prediction tasks.

7. Additional Results of Open-Vocabulary Object Detection

In the main results of open-vocabulary object detection (OVOD), we employed the EVA-CLIP model for its capability. We present detailed results using the OpenAI CLIP model for OVOD in Tab. 7, where our method still achieves superior performance across all metrics.

8. Qualitative Examples

Fig. 3 shows the results of the zero-shot scenario of open-vocabulary semantic segmentation (OVSS) performed on the Pascal VOC [5] dataset using the MaskCLIP [21] framework. Fig. 4 presents the results of using CAT-Seg v2 [4] on the Pascal Context [13] dataset. Finally, Fig. 5 displays the results of OVOD performed on the OV-COCO [11] dataset using the F-ViT [8] model.

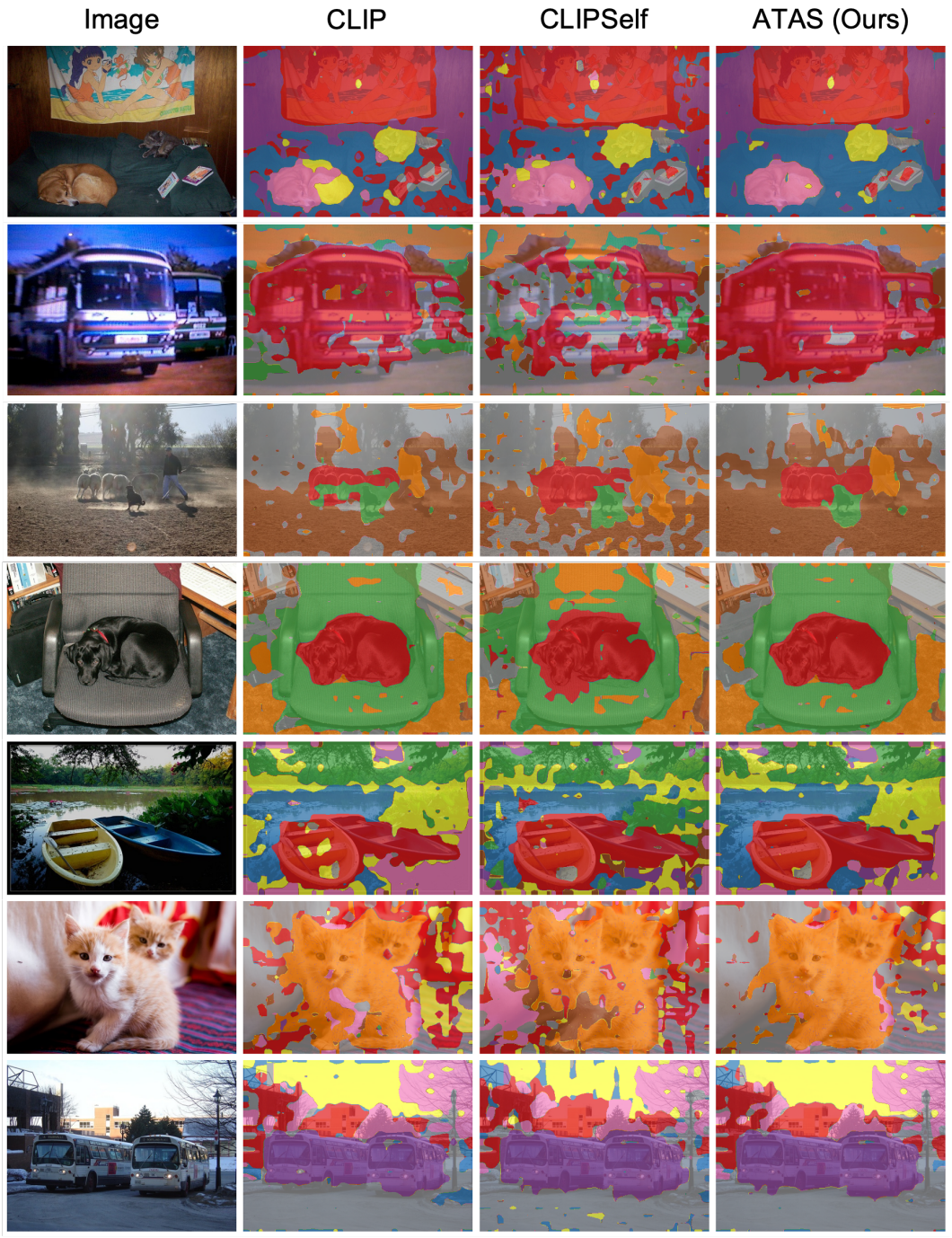

Figure 3. **Visualization of MaskCLIP Segmentation Results.** that compares dense prediction abilities of CLIP, CLIPSelf, and ATAS. The images are from Pascal VOC dataset.

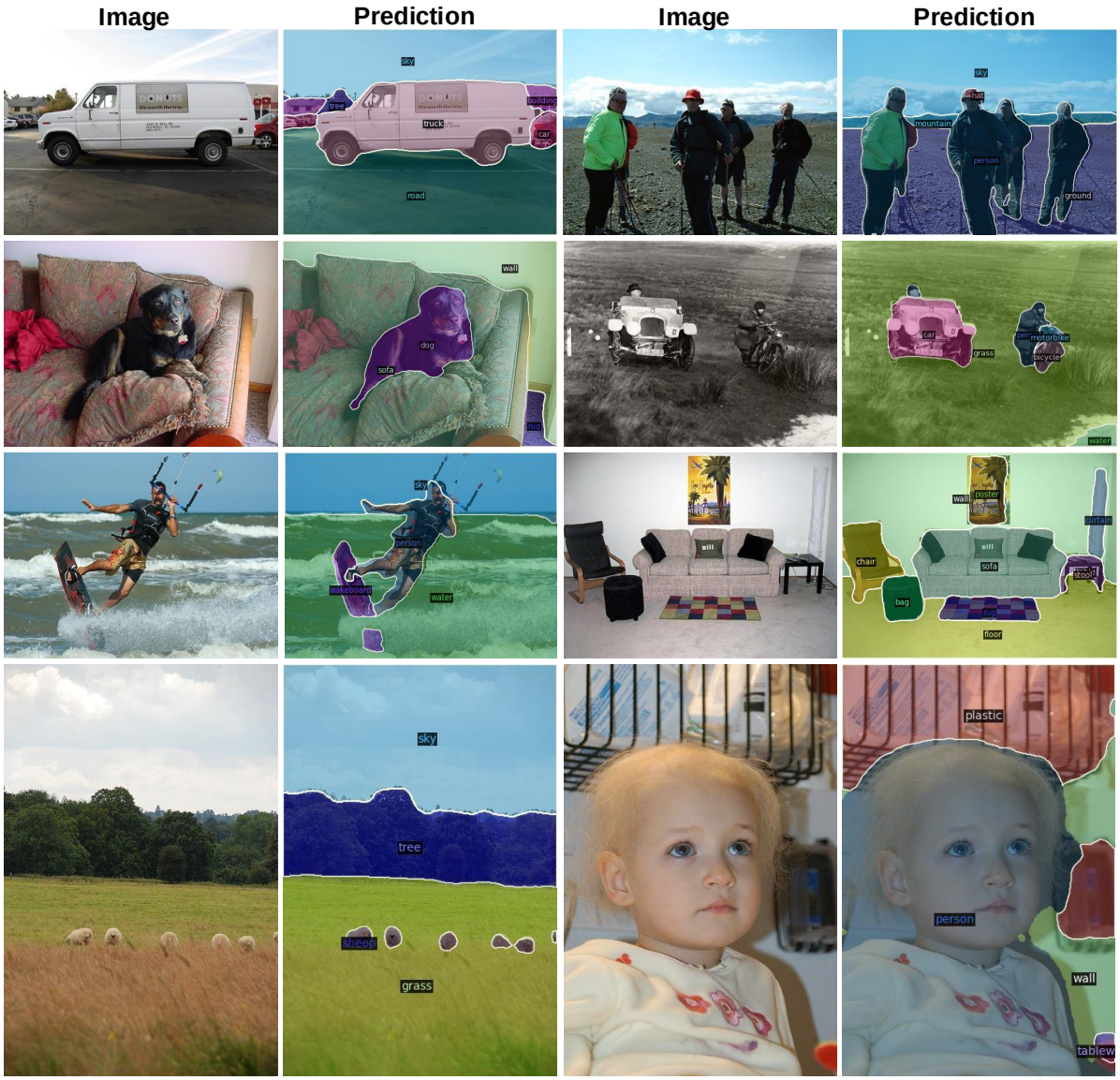

Figure 4. **Visualization result of CAT-Seg v2 with ATAS Results.** The images are from Pascal Context [13] with 459 categories.

Figure 5. **Visualization of Object Detection Results.** The red boxes indicate predictions for novel classes, while the blue boxes represent predictions for base classes.

References

[1] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3828–3837, 2024. 1

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 2, 3

[3] 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

[4] Seokju Cho, Heeseong Shin, Sunghwan Hong, Seungjun An, Seungjun Lee, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2303.11797*, 2023. 3, 4

[5] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111:98–136, 2015. 2, 4

[6] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*, 2022. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[8] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. Open-vocabulary object detection upon frozen vision and language models. In *The Eleventh International Conference on Learning Representations*, 2022. 4

[9] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. *arXiv preprint arXiv:2407.12442*, 2024. 3

[10] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *ECCV*, 2024. 1

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 2, 4

[12] 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. 3

[13] 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 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. 2, 4, 6

[14] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip H.S. Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19413–19423, 2023. 1

[15] 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, 3

[16] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *European Conference on Computer Vision*. Springer, 2024. 1

[17] 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, 2025. 3

[18] Size Wu, Wenwei Zhang, Lumin Xu, Sheng Jin, Xiangtai Li, Wentao Liu, and Chen Change Loy. CLIPSelf: Vision transformer distills itself for open-vocabulary dense prediction. In *The Twelfth International Conference on Learning Representations*, 2024. 3, 4

[19] Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3426–3436, 2024. 4

[20] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 4

[21] 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, 3, 4