Supplementary Materials: SIA-OVD: Shape-Invariant Adapter for Bridging the Image-Region Gap in Open-Vocabulary Detection

Anonymous Authors

1 MORE ABLATION STUDIES

We provide additional experimental results of various configurations and architectures within the shape-invariant adapter architecture. For all experiments, we conduct the ground-truth bounding boxes as input and utilize AP50 of base and target categories as evaluation metrics.

Learnable residual factor. SIA keeps a set of N independent adapters $\{Adapter_j, j = 1, ..., N\}$. Each adapter consists of two fully connected layers $\mathbf{W}_1^j$ and $\mathbf{W}_2^j$, a ReLU activation layer, and a residual factor λ. In this paper, we initially set the residual factor to 0.2. To better evaluate the effectiveness of this factor, we parameterize the residual factor as a learnable parameter and employ softmax to confine it within the range of 0 to 1. The results are shown in Table 1. While introducing additional learnable residual factors led to significant performance gains on base categories, it resulted in a notable 2.38 AP50 decrease in performance on novel categories.

Table 1: Training with learnable residual factor. The results of the proposed SIA is highlighted in bold format.

Adapter Number	Learnable	AP50	
(N)	Residual Factor	Novel	Base
2		68.56	81.34
2	✓	67.10	82.29
4		68.38	81.56
4	✓	66.06	82.63
10		**68.65**	**81.37**
10	✓	68.02	82.42
20		67.83	80.99
20	✓	66.83	81.99

Different architectures of SIA. To study the effect of different architectures of SIA, we conduct an ablation study by replacing "Linear-Relu-Linear" with only a Linear layer. The results are shown in Table 2, where "L-R-L" represents a Linear-ReLU-Linear architecture, and "L" represents a Linear layer. The utilized architecture "L-R-L" leads in AP50 both on base and novel categories.

2 MORE VISUALIZATIONS

Classification on Ground-Truth Bounding Boxes. In Figure 1, we show more examples of classification results calculated by CLIP [2], CORA [3], and SIA (ours) on ground-truth bounding boxes from the COCO-OVD validation set.

Distribution of Adapted Region Features. Figure 2 shows the visualization results of the region features adapted by SIA for 17 novel categories in the COCO-OVD validation set using t-SNE [1]. We report the results of CLIP [2], CORA [3], and our SIA with both RN50 and RN50x4 backbone, which can be demonstrated that our SIA achieves a more obvious separation of embeddings belonging to different categories than CLIP and CORA.

Table 2: Different architectures of SIA, where "L-R-L" represents a Linear-ReLU-Linear architecture, and "L" represents a Linear layer. The results of the proposed SIA is highlighted in bold format.

Adapter Number	Architecture		AP50	
(N)	L-R-L	L	Novel	Base
2	✓		68.56	81.34
2		✓	67.82	81.23
4	✓		68.38	81.56
4		✓	67.43	81.06
10	✓		**68.65**	**81.37**
10		✓	66.99	81.26
20	✓		67.83	80.99
20		✓	67.17	80.99

REFERENCES

[1] Pavlin G Poličar, Martin Stražar, and Blaž Zupan. 2019. openTSNE: a modular Python library for t-SNE dimensionality reduction and embedding. *BioRxiv* (2019), 731–877.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. 8748–8763.

[3] Xiaoshi Wu, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. CORA: Adapting CLIP for Open-Vocabulary Detection With Region Prompting and Anchor Pre-Matching. In *CVPR*. 7031–7040.

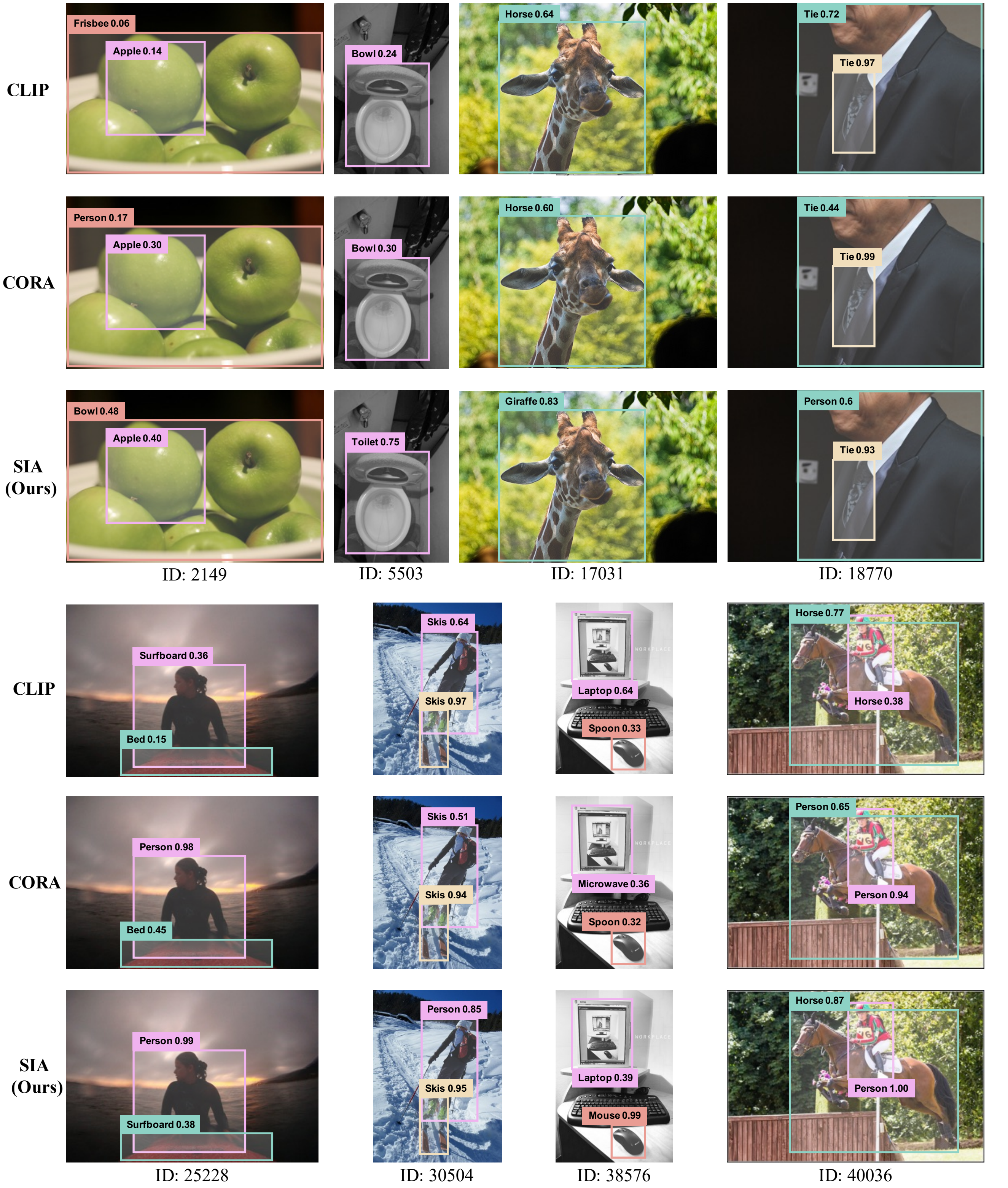

Figure 1: Visualization of region classification results and confidence scores for ground-truth bounding boxes from the COCO-OVD validation set.

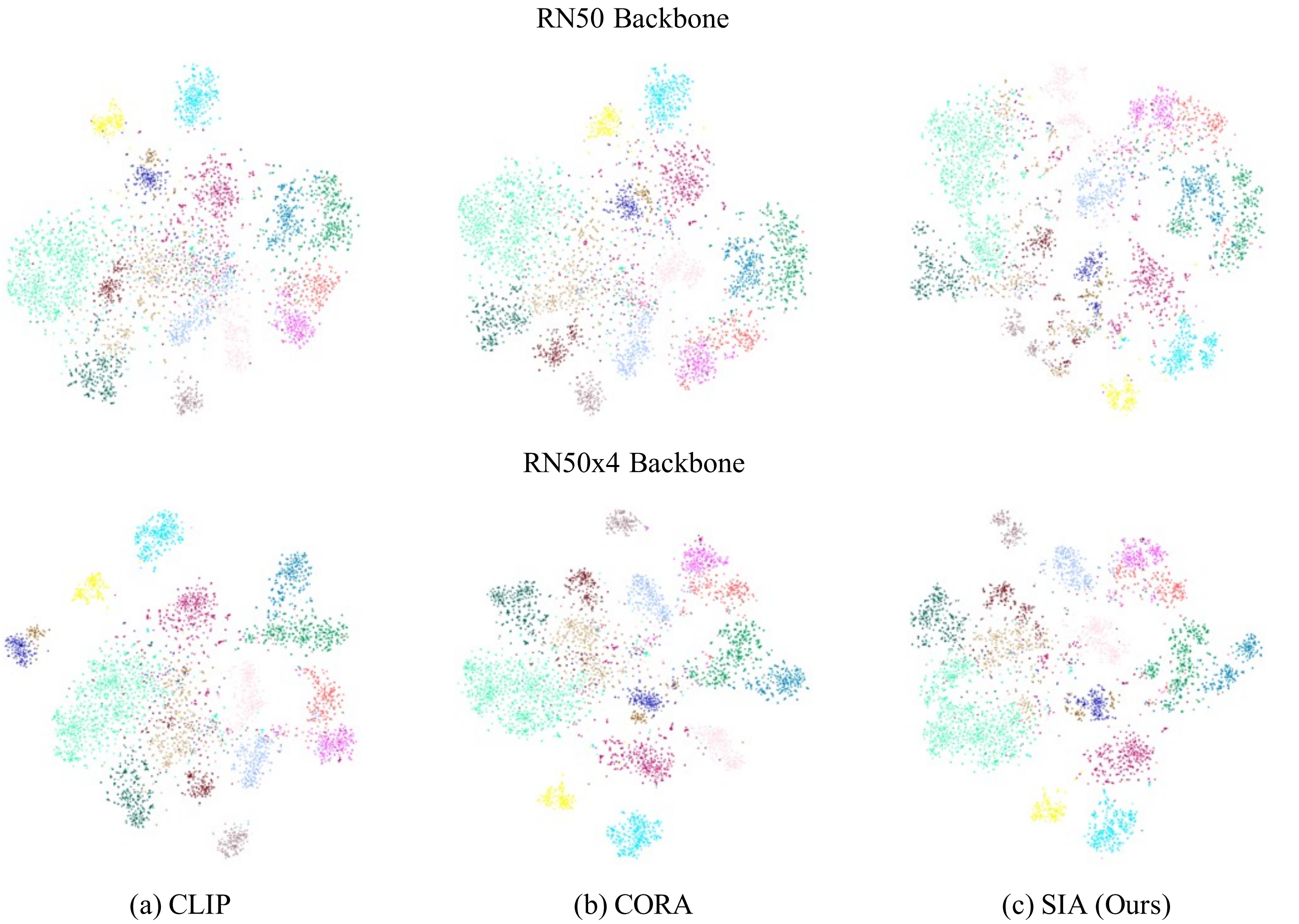

Figure 2: Visualization of the region features belong to 17 novel categories in the COCO-OVD validation set encoded by CLIP, CORA, and our SIA with RN50 backbone via t-SNE.