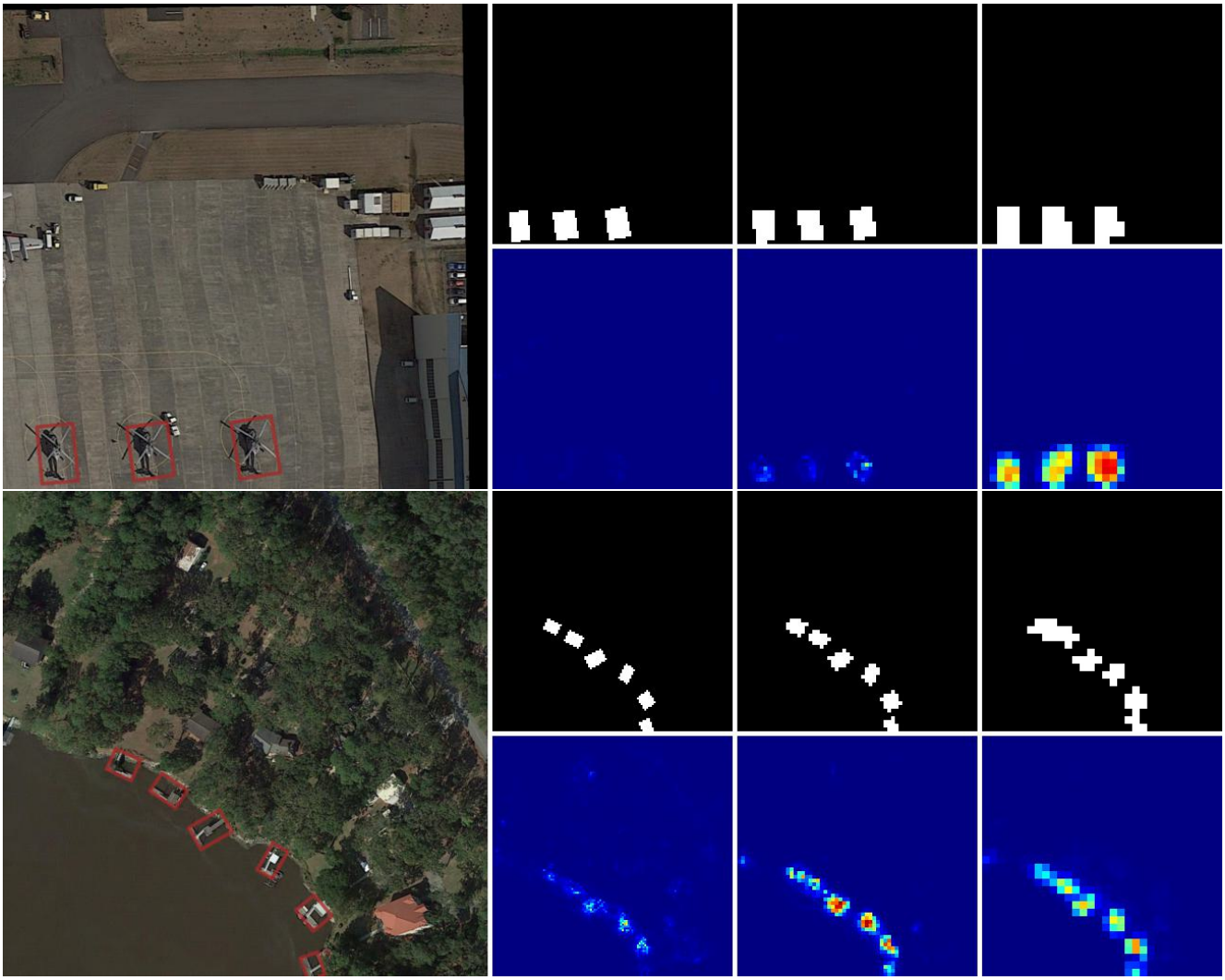

Figure 1. Visualization of single-object detection results on DOTAv2. The figure shows rotated bounding boxes (left), multi-scale enhanced visual features $\{C_3, C_4, C_5\}$ (lower right), and multi-scale learned masks $\{X_3, X_4, X_5\}$ (upper right) generated by the proposed VISO model in response to corresponding text input for detecting one specific object category. The masks demonstrate the high sparsity.

A. Appendix

A.1. Visualization

In this section, we visualize the detection results on DOTAv2, showcasing rotated bounding boxes, multi-scale learned masks $\{X_3, X_4, X_5\}$ and heatmaps of multi-scale enhanced visual features $\{C_3, C_4, C_5\}$ of VISO.

We first consider single-object detection in DOTAv2, i.e., only one object or one category of object is of interest. We take "Detect the helicopter"as the text input of VISO. The visualization is shown in the first row of Figure 1. We take "Detect the harbor"as the text input of VISO. The visualization is shown in the second row of Figure 1.

We further consider multi-object detection in DOTAv2, i.e., multiple objects or multiple categories of objects are of interest. We take "Detect the plane", "Detect the ship", "Detect the storage tank", "Detect the baseball diamond", "Detect the tennis court", "Detect the basketball court", "Detect the ground track field", "Detect the harbor", "Detect the bridge", "Detect the large vehicle", "Detect the small vehicle", "Detect the helicopter", "Detect the roundabout", "Detect the soccer ball field", "Detect the swimming pool", "Detect the container crane", "Detect the airport", "Detect the helipad"as text inputs of VISO. The visualization is shown in Figure 2.

For 1024×1024 input image size, the scale of C_3 is 128 (1/8), the scale of C_4 is 64 (1/16), and the scale of C_5 is 32 (1/32). We resize them to the original size in visualization.

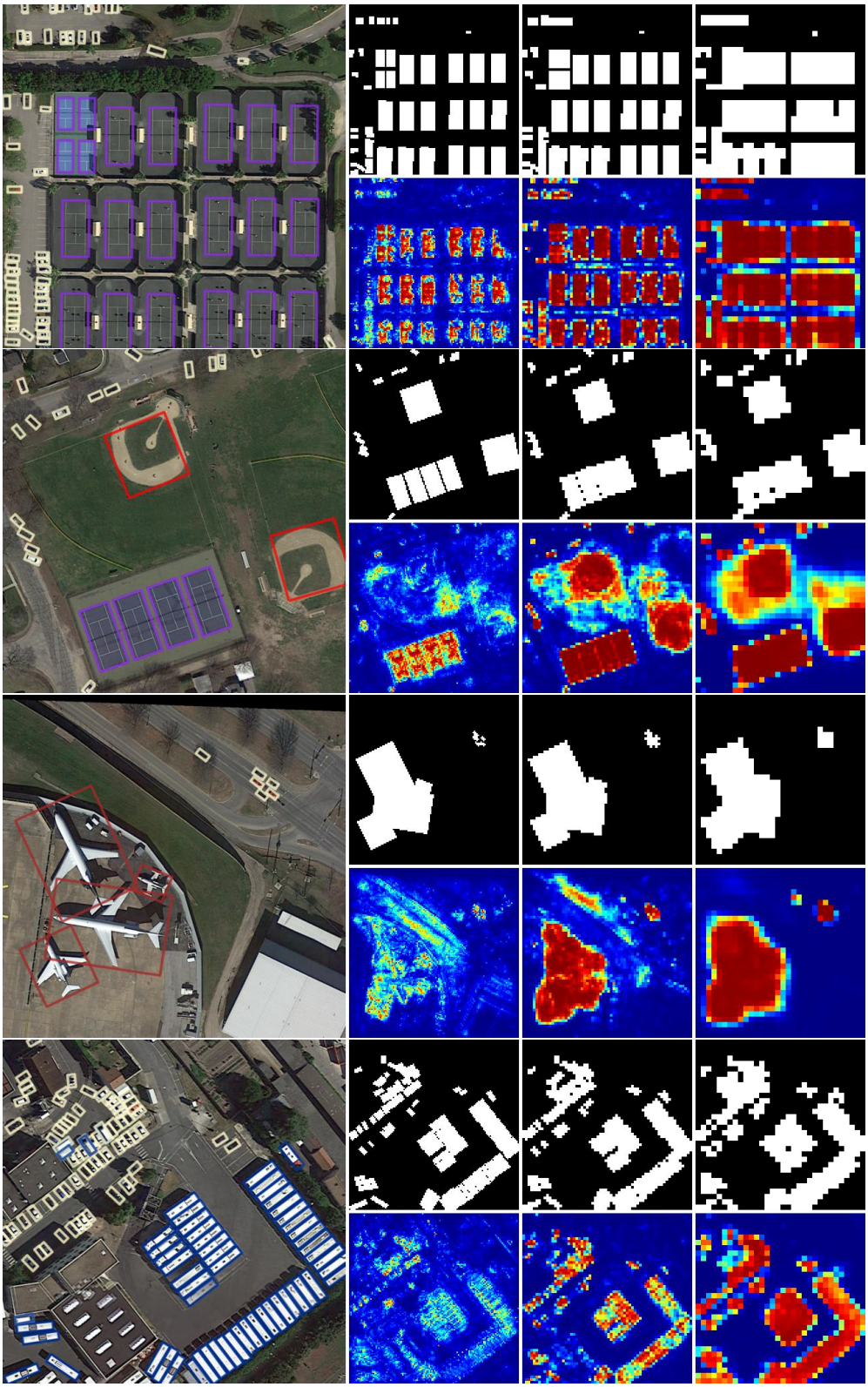

Figure 2. Visualization of multi-object detection results on DOTAv2. The figure shows rotated bounding boxes (left), multi-scale enhanced visual features $\{C_3, C_4, C_5\}$ (lower right), and multi-scale learned masks $\{X_3, X_4, X_5\}$ (upper right) generated by the proposed `VISO` model in response to various text inputs for detecting and distinguishing different object categories.

A.2. Details of fair comparison

Based on the zero-shot results in Tab.3, we further train YOLO-World (YW), which has comparable parameter sizes and structure to VISO, on our 3.4M training set and compare the results in Table 1. We find that VISO increases 3.1% AP on average for 3 test sets on M variant without sparsity and 1.0% AP on L variant without sparsity, highlighting VISO's effectiveness in distinguishing object-centric feature extraction. When applying sparsity to VISO, VISO can still achieve remarkable FLOPs compared with dense models.

Model	MAR20		HRSC2016		VEDAI	
	FLOPs(G) ↓	AP(%) ↑	FLOPs(G) ↓	AP(%) ↑	FLOPs(G) ↓	AP(%) ↑
YW-M	113	89.9	113	69.6	113	43.6
YW-L	229	89.2	229	70.3	229	48.3
	No sparsity → Sparsity					
VISO-M	113→ 59	**90.2**→90.1	113→68	**75.6**→74.6	113→55	**46.6**→46.2
VISO-L	229→130	90.0→90.0	229→134	70.7→70.1	229→118	**50.2**→49.9

Table 1. Training YOLO-World on the same training set with VISO and test on three benchmarks in terms of AP (%) and FLOPs (G).

Additionally, we present Pareto front figures in Figure 3 showing the AP (%) versus FLOPs (G) for three variants of VISO on the MAR20 test set, evaluated under three levels of sparse conversion.

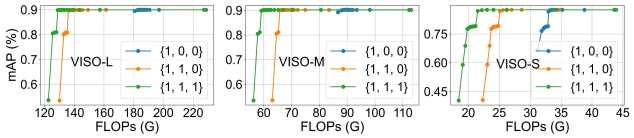

Figure 3. FLOPs (G) V.S. AP (%) of VISO-S/M/L on MAR20 test set under three level-wise sparse conversion.