ROS-SAM: High-Quality Interactive Segmentation for Remote Sensing Moving Object

Supplementary Material

1. The details of our used dataset

Details of the dataset used in this study are presented in Table 1. This dataset poses significant challenges due to its diverse range of image and object sizes, with the average object size being relatively small (MS COCO defines small objects as those measuring smaller than 32×32 pixels, and the majority of objects in this dataset fall below that threshold.). Additionally, Figure 1 illustrates sample images from the training set, which are generated using large-scale jittering (LSJ) and random rotations. For original data, we resize images randomly within a scale range of 0.1 to 4.0. If the resized dimensions exceed 1024×1024 pixels, the images are cropped to fit; if they are smaller, gray borders are added to the bottom-right corner to pad the image. This data augmentation strategy effectively enhances both the diversity and quantity of training samples.

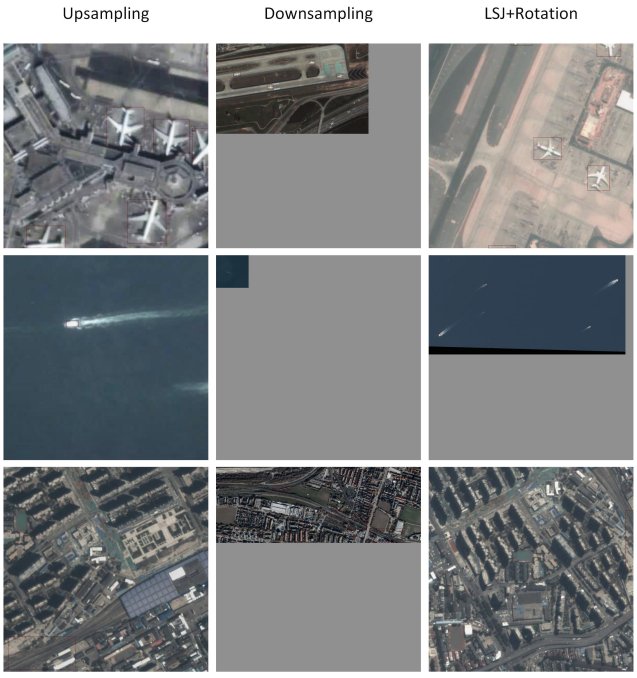

Figure 1. Visualization of partial training samples.

2. Visualization of experimental results

In Figure 2, we visualize the prediction results, highlighting how our methods enhance the model's performance. First, a comparison between the results of the original SAM and our pipeline reveals that SAM's predictions are significantly less consistent with the object's actual shape. This inconsistency arises primarily from bias introduced by multiple resizing operations, particularly evident in the ship prediction, where the result extends beyond the prompt box. Second, the proposed decoder improves the object's shape refinement by leveraging early texture information. However, it introduces some degree of feature misjudgment due to limitations in feature interpretation. Finally, when the proposed fine-tuning method is integrated with the image encoder, the resulting predictions are highly accurate and closely align with the ground truth labels.

Attribute	Number
Number of training sets	8,333
Number of test sets	2,284
Maximum image size	2160×1080
Minimum image size	512×512
Number of objects	52,123
Maximum object size	8422
Minimum object size	12
Average object size	518
Maximum ratio of the object to image	0.023339
Minimum ratio of the object to image	3.111e-6

Table 1. Details of the dataset used.

3. Zero-Shot segmentation on object tracking

To evaluate its generalization ability, we test our model on several remote sensing object tracking datasets, as shown in Figures 3. Notably, the predictions of our model in this zero-shot setting remain of high quality, demonstrating significantly superior zero-shot capabilities compared to SAM in remote sensing. For airplanes, both our model and SAM can identify airplanes. However, SAM's predictions are highly imprecise, often resembling a four-pointed star rather than the actual shape of an airplane. In contrast, our model achieves fine-grained segmentation, accurately identifying details such as the engine position. For trains, SAM fails to produce meaningful predictions, whereas our approach shows a clear advantage. For ships, predictions are generally accurate, as the distinct features of ships contrast significantly with the water's surface. Our results highlight challenging scenarios where ships and waves overlap, and our model delivers significantly more accurate predictions in these cases.

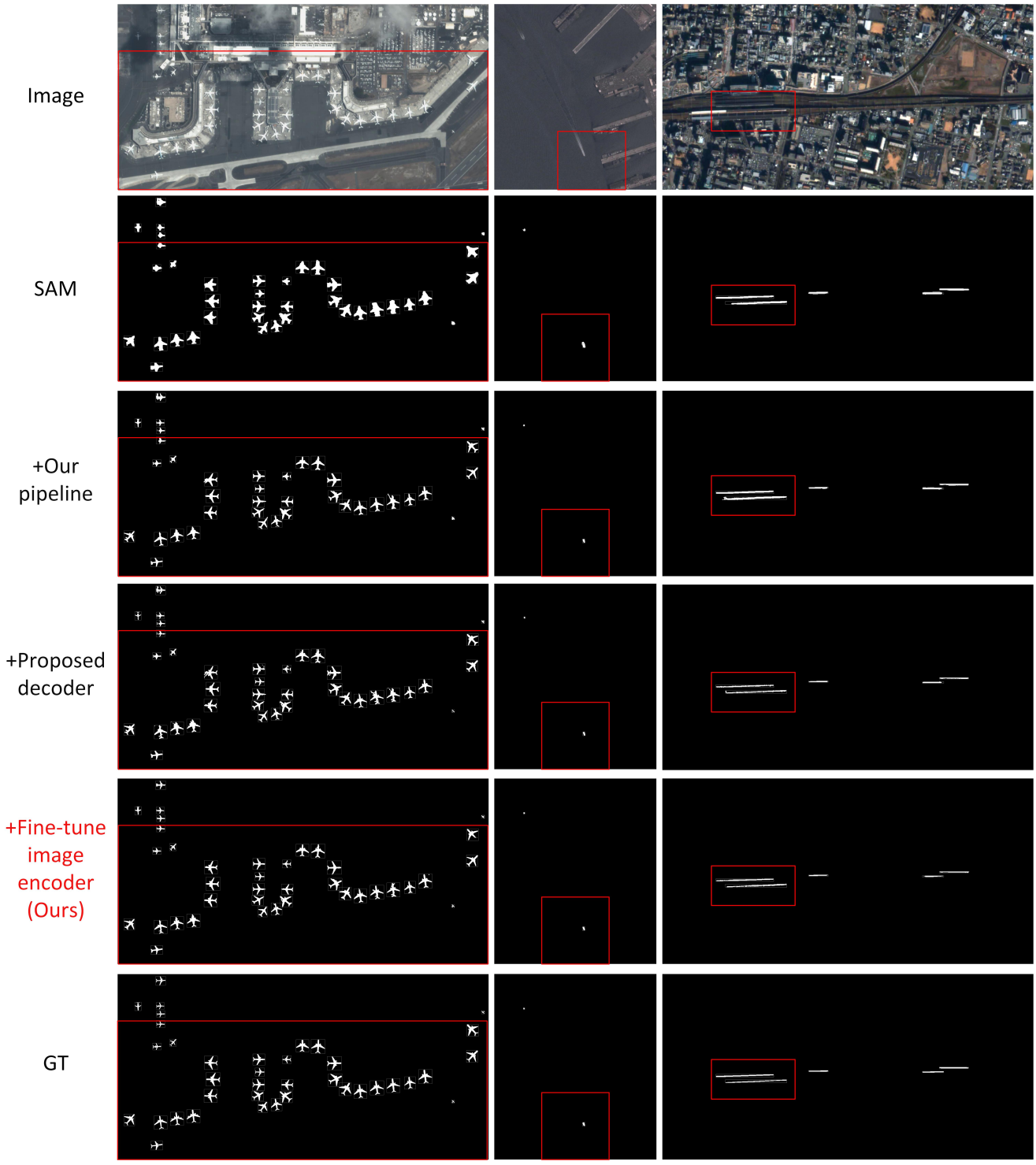

Figure 2. Visualized results of our proposed ROS-SAM. In the second row, the predictions extend beyond the object's actual shape, particularly for the train, where the prediction nearly fills the entire prompt box. In the third row, the predictions lack fine-grained detail and, similar to the second row, fail to distinguish between the foreground and background of the train. In the fourth row, feature discrimination is notably poor, resulting in multiple incorrect predictions; for instance, the boarding bridge and other objects are misclassified as airplanes.

Figure 3. Visualization results of ROS-SAM on the remote sensing object tracking dataset.