

Supplementary Materials: Unleashing the Power of Generic Segmentation Models: A Simple Baseline for Infrared Small Target Detection

Anonymous Authors

A OVERFITTING ISSUE OF TRAINING LARGE VISION MODELS IN THE CONTEXT OF IRSTD

In machine learning theory, a phenomenon known as overfitting exists, wherein a model excels on the training data but struggles to generalize to new, unseen data. This phenomenon is often depicted by a "U-shaped" curve in the model's performance throughout training, showing an initial performance increase and then a decrease as the model converges. As shown in Figure 2, we finetune three large models: SAM, SAM-HQ, and our teacher model Semantic-SAM on the SIRST dataset and visualize their performance curves and training loss. We can see from the figure that, as the training loss decreases, the performance of all three models grows initially and then decreases, exhibiting a serious overfitting issue. Although we mitigate this issue by adopting early stopping to stop the training process when performance starts to degrade, the potential of these models is partially constrained, and the performance is inferior to our proposed model.

B EXCLUDED IMAGE FOR THE IRSTD1K DATASET

IRSTD1k dataset contains 1001 images. We follow [2] to choose 800 images as the train set. We carefully check the labels and images of the whole dataset and exclude 6 images from the test set due to their incorrect or misleading annotations. To ensure fairness, we test all methods under the same settings and provide details of these excluded images in Figure 3.

Specifically, in Figure 3a, there is an unannotated bright spot in the upper right corner of the image. This spot appears similar to the targets in the center of the image to the naked eye, and there is insufficient additional information to determine whether it should be considered a target. In Figure 3b, based on the definition of small targets given by [1, 2], the plane on the right side of the image appears too large to be considered a small target. Additionally, the bright spot framed in red lacks annotation. In figure 3c, the spot framed in red misses annotation. The ground truth mask of the target framed in red in Figure 3d is inaccurately labeled, as the mask appears much larger than the corresponding target itself. Moreover, we delete Figure 3e. Even to the human eye, it is impossible to determine if the location marked by a red box is indeed a target. Finally, we exclude Figure 3f due to missing annotation for the location indicated by the red box.

C SPEED-ACCURACY TRADEOFFS

In Figure 5, we investigate the throughput and accuracy of our model against IRSTD SOTA methods, SAM, and SAM variants. Our model achieves a better trade-off between throughput and accuracy.

Compared to the latest IRSTD methods, our model surpasses UIU-net, DNA-net, and ISNet by approximately 5 IoU while maintaining a comparable inference speed to UIU-net. In addition, our approach outperforms large vision models such as SAM, SAM-HQ, and the teacher model Semantic-SAM in both performance and throughput, demonstrating that our proposed method fully harnesses the potential of generic segmentation models. Despite the faster throughput of efficient SAM variants like Efficient-SAM and MobileSAM, our method still reaches 'real-time' speed and exceeds them by a large margin of 9 IoU.

D VISUAL COMPARISON ON THE NUDT DATASET

In Figure 4, we visually compare outputs of different methods on the same input image from the NUDT dataset. Our model demonstrates significant superiority over 10 other methods across different scenarios. Concretely, for the first and second input images where the targets are bright and clear, although all methods can detect and locate small targets, our model produces outputs with the finest-grained details closest to the ground truth. In complex scenarios when targets are dimmer and obscured by their surroundings, such as the third and fourth input images in Figure 4, other methods either fail to accurately locate targets or produce low-quality masks. In contrast, our method consistently delivers robust results, effectively segmenting highly discriminative objects.

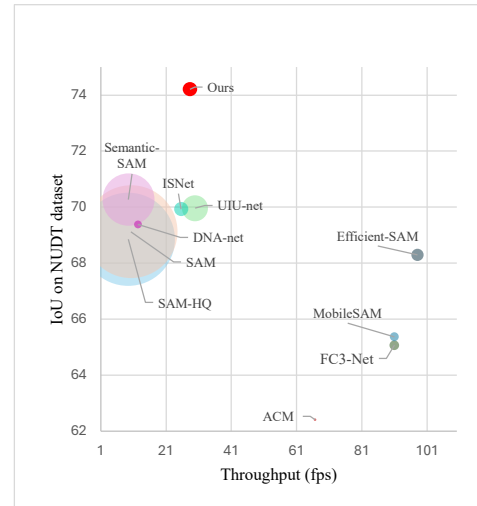


Figure 1: The comparisons of the IoU and throughput on a single Nvidia GeForce 4090 GPU. The circle size refers to the model size. Batch size is set to 1, and the experiments are conducted on the IRSTD1k dataset.

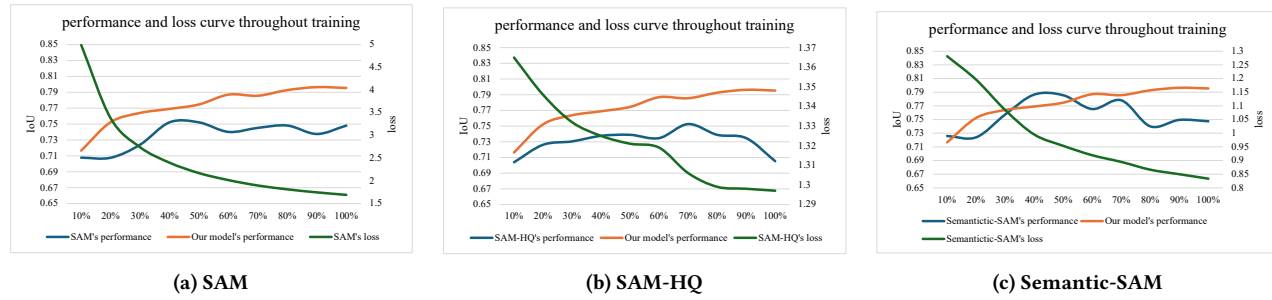


Figure 2: The performance and training loss curve of SAM, SAM-HQ, and Semantic-SAM on SIRST dataset. We can observe a "U-shaped" in performance as the training proceeds. This indicates that the model overfits to the dataset

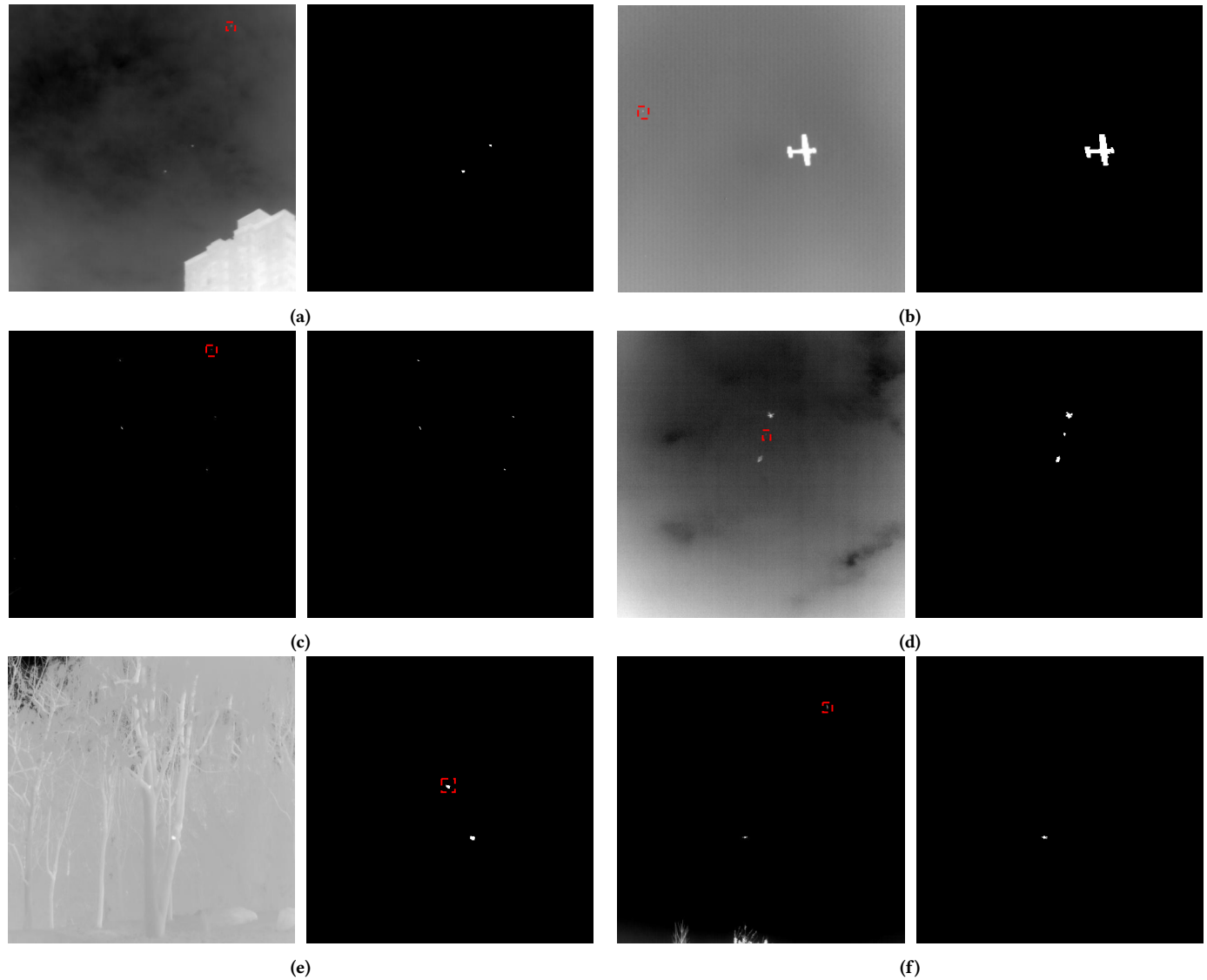


Figure 3: Deleted images from the testing split of the IRSTD1k dataset



Figure 4: Visual comparison among 11 methods on the NUDT dataset.

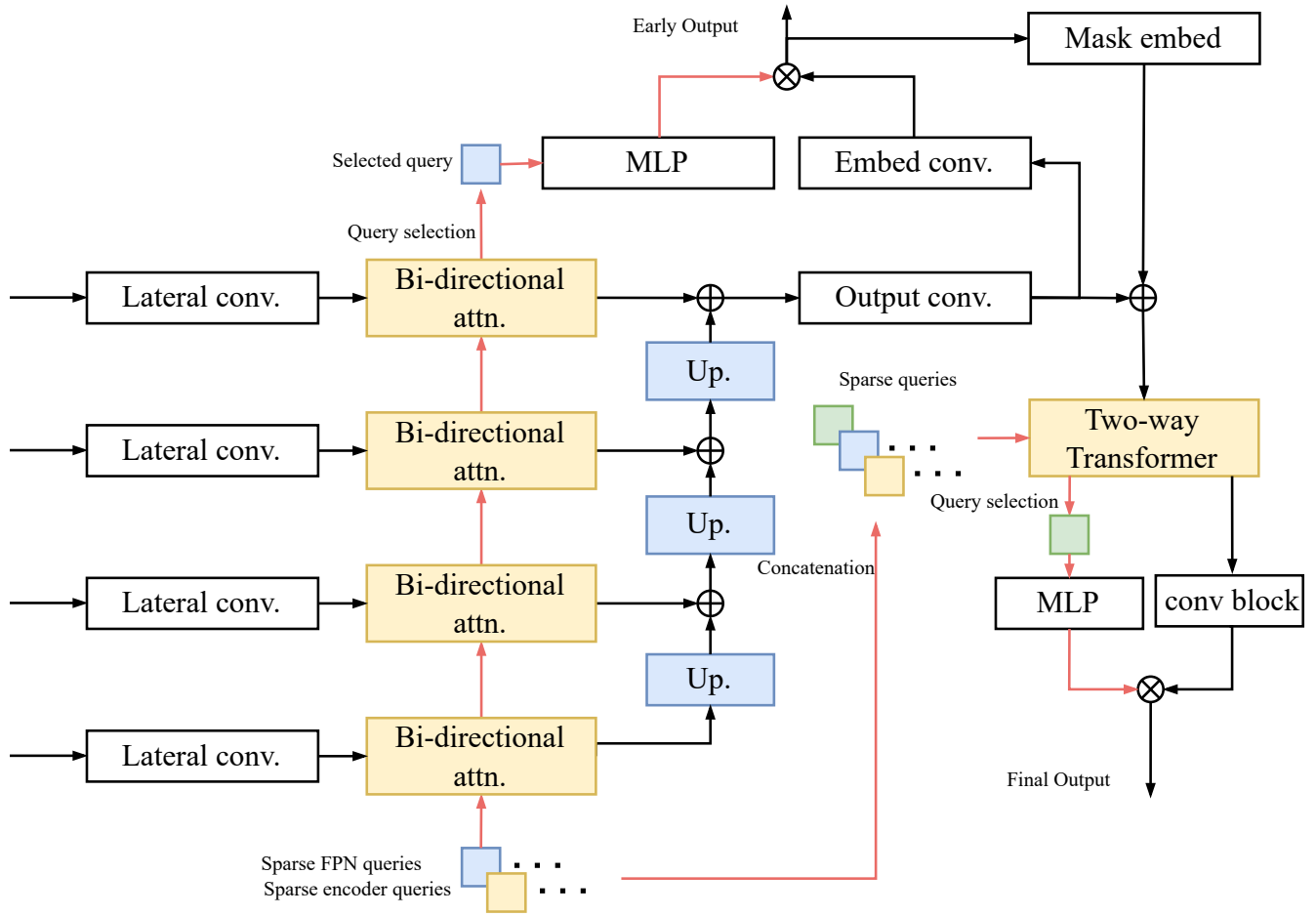


Figure 5: Details in FPN and modified SAM decoder. The proposed query design effectively levitates multi-level information and prompts final decoding progress

REFERENCES

- [1] Xin Wu, Danfeng Hong, and Jocelyn Chanussot. 2022. UIU-Net: U-Net in U-Net for infrared small object detection. *IEEE Transactions on Image Processing* 32 (2022), 364–376.
- [2] Mingjin Zhang, Rui Zhang, Yuxiang Yang, Haichen Bai, Jing Zhang, and Jie Guo. 2022. ISNet: Shape matters for infrared small target detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 877–886.