Correspondence as Video: Test-Time Adaption on SAM2 for Reference Segmentation in the Wild

Supplementary Material

Table 5. **Parameter sizes and inference speeds of SAM2 at different scales.** As the scale of increases, the parameter size also rises, while the inference speed maintaining real-time level.

	tiny	small	base_plus	large
Size (M)	38.9	46.0	80.8	224.4
Speed (FPS)	47.2	43.3	34.8	24.2

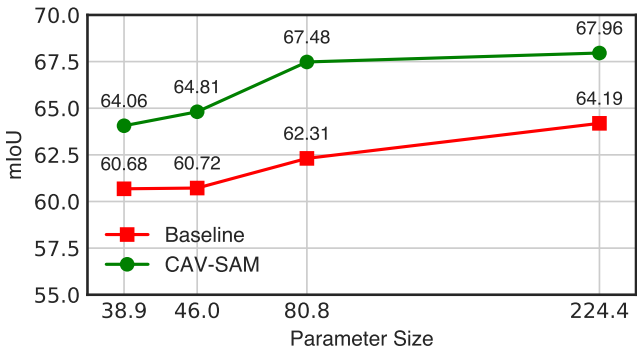

Figure 11. **Performance variation with model size in SAM2 encoder configurations.** The result illustrates the performance of both the baseline method and CAV-SAM across different scales of the SAM2 encoder, demonstrating that CAV-SAM maintains robust performance even with the smallest model configuration, while performance consistently improves with larger model scales.

A. Performance across Different Scales

SAM2 [31], like its predecessor, offers multiple encoder scales: `tiny`, `small`, `base_plus`, and `large`, each with distinct parameter sizes and inference speeds, as detailed in Tab. 5. Our experiments focused on the smallest configuration, `tiny`, to minimize the possible influence of SAM2's built-in video segmentation capabilities on our method's effectiveness. However, in practical applications, larger models generally adhere to scaling laws, often resulting in enhanced performance. We evaluated both the baseline discussed in Sec. 4.4 and CAV-SAM across all model scales, with results depicted in Fig. 11. These findings indicate that our method achieves robust performance even with the smallest model, while performance improves with larger model scales as well. It is worth mentioning that since our TTGA module only fine-tunes the FPN `neck` of the SAM2 encoder, the increase in the parameter size of SAM2's encoder does not result in significant computational overhead.

Table 6. **Comparative performance of CAV-SAM and VRP-SAM across different label types.** Evaluations are performed on the CD-FSS benchmark using *points* and *boxes* labels.

	Point	Box	Mask
VRP-SAM	39.11	48.92	57.50
CAV-SAM	55.72	61.28	64.06

B. Other Types of Reference Labels

In the main body of the paper, we define reference segmentation as using reference images and their pixel-level *masks*. However, as prompt segmentation becomes a new trend, it is now possible to use other types of sparse labels (such as *points* or *boxes*) for reference segmentation. Our CAV-SAM can seamlessly utilize these alternative label types by first generating pixel-level pseudo-labels from them, and then proceeding with the original steps of CAV-SAM using these pseudo-labels. We conducted additional evaluations of our CAV-SAM using point and box label types and compared it to VRP-SAM, which also supports multiple label types. All the datasets in CD-FSS benchmark contain only mask as labels. Therefore we follow SEEM[1] to generate point and box labels by randomly simulating user inputs based on the ground truth mask of the reference image. The results presented in Tab. 6 demonstrate that our method significantly outperforms existing approaches across various label types. We attribute this success to SAM2's robust video segmentation capability and TTGA module, both are not sensitive to the accuracy of the sparse labels of the reference image.

C. Additional Qualitative Results

Due to space constraints, we can only include one visualization example for each dataset in the main text. Additional visualization results will be provided in Fig. 12 in this supplementary material. Our CAV-SAM demonstrates excellent performance across all four datasets in the CD-FSS benchmark. The DBST module effectively generates pseudo-video sequences with semantic transitions, while the TTGA module derives pseudo-labels for the first half of the pseudo-video sequence using only a single reference image and its mask. These pseudo-labels serve as additional prompts for SAM2, further aligning the model with the geometric variance present in the pseudo-video sequences.

[1]Zou, Xueyan, et al. "Segment everything everywhere all at once." Advances in Neural Information Processing Systems 36 (2024).

ICCV
#13298

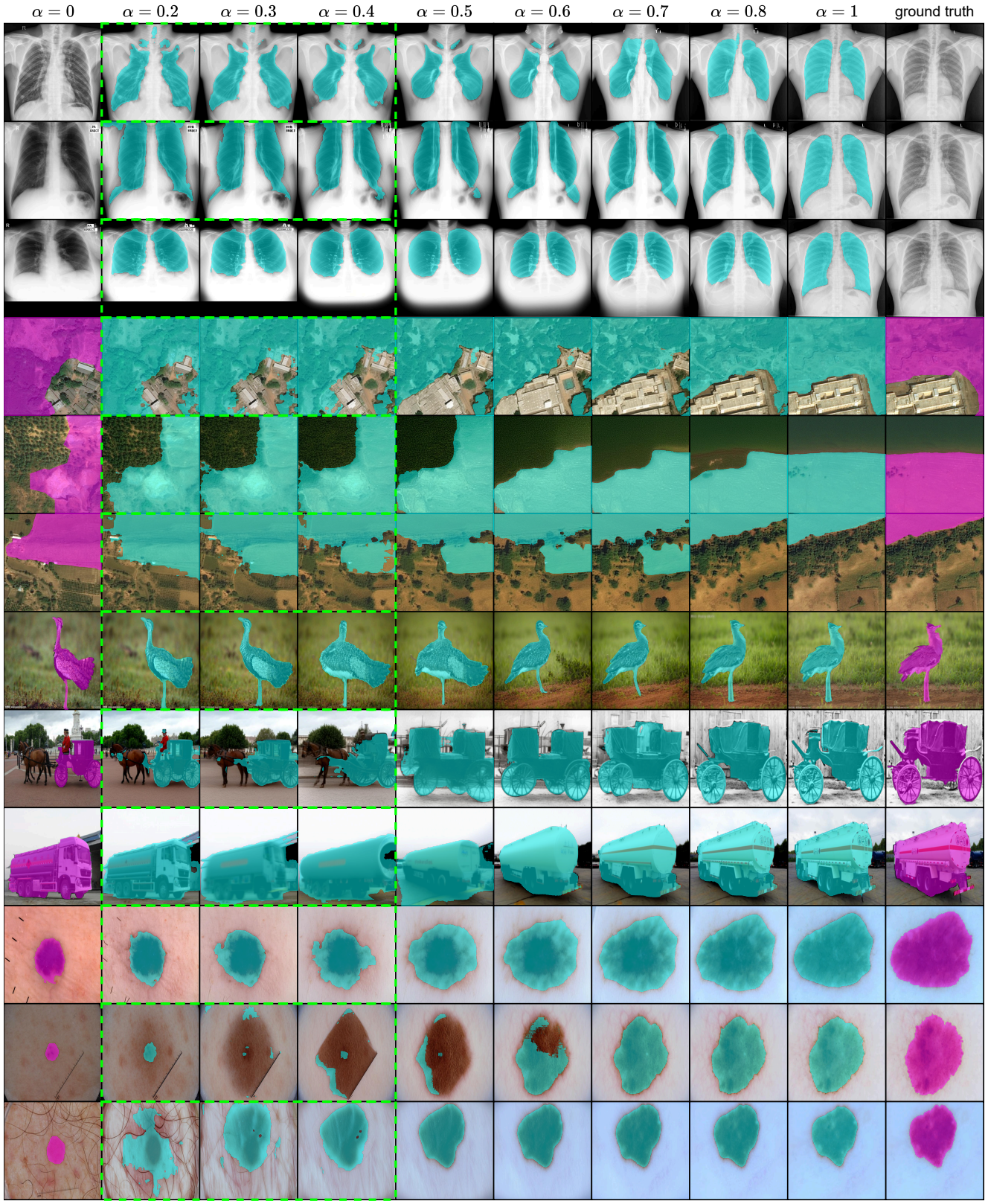

Figure 12. **Additional visualization of 1-shot segmentation results across four datasets.** The **magenta** mask denotes ground truth, while the **cyan** mask shows predictions. **Green** dashed lines outline frames where the TTGA module prompts SAM2.