Towards Fine-grained Interactive Segmentation in Images and Videos Supplementary Materials

Yuan Yao Qiushi Yang Miaomiao Cui Liefeng Bo
Tongyi Lab, Alibaba Group
{ryan.yy, yangqiushi.yqs, miaomiao.cmm, liefeng.bo}@alibaba-inc.com

1. Implementation Details

1.1. Training scheme

During training, we fix the model parameters of the pre-trained SAM2 model while only make the proposed modules learnable. To support flexible segmentation prompts, we train SAM2Refiner by sampling five types of prompts including bounding boxes only, randomly sampled positive points only, the mixture of bounding boxes and randomly sampled positive points, the mixture of bounding boxes and randomly sampled negative points and degraded masks. We set a probablity list of $[0.3, 0.2, 0.2, 0.2, 0.1]$ for above prompt types and randomly sample these types in each training iterations. We set the number of sampled points as 10, the positive points and negative points are sampled based on the GT masks. We generate the degraded masks by adding random Gaussian noise in the boundary regions of the GT masks. For generalizability to different object scales and orientation, we use large-scale jittering and flipping strategies for data augment.

1.2. Details of RFB

The RFB block in Prompt Retargeting module aims to reduce the channel number as well as enhance the discriminability of features in small patches. Figure 1 demonstrates the architecture of the RFB block. Three parallelized branch are applied to achieve multi-size receptive fields. In each branch, a 1×1 convolutional layer is followed by $1 \times n$ plus a $n \times 1$ convolutional layer, and a 3×3 convolutional layer with a corresponding rate (i.e., padding and dilation) n. The three branches are then concatenated and encoded via a 1×1 convolutional layer. The final augmented feature is obtained by adding the concatenated feature and a shortcut feature encoded from another 1×1 convolutional layer with input feature.

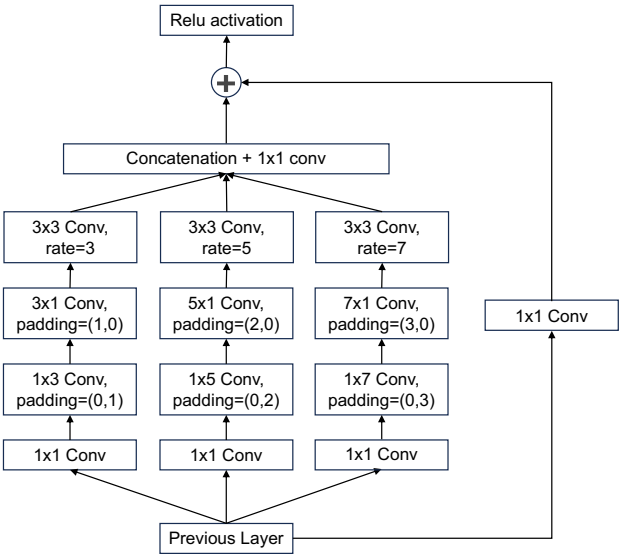

Figure 1. The architecture of the modified RFB block in Prompt Retargeting module.

2. More Results of Our Method

2.1. Comparison to HQ-SAM2 and SAM2 finetuning

We provide two experiments to further enhance the fairness of the experimental setup. Firstly, we compare our method with HQ-SAM2, with the same training-set and same SAM2 backbone. Secondly, we finetune SAM2's decoder with the same training-set. Results in Tab. 1 demonstrates the superiority of our method.

Table 1. Comparison to HQ-SAM2 and SAM2 finetuning.

Model	DIS		COIFT		HRSOD		ThinObject	
	mIoU	mBIoU	mIoU	mBIoU	mIoU	mBIoU	mIoU	mBIoU
HQ-SAM2 (hiera_l)	68.7	60.9	95.2	91.1	88.5	82.5	88.6	77.8
Finetune SAM2's decoder	81.4	74.4	95.0	91.7	92.5	85.6	93.4	87.4
Ours (hiera_l)	**85.2**	**80.6**	**96.6**	**93.6**	**94.7**	**90.2**	**95.9**	**91.0**

2.2. Efficiency evaluation

We provide efficiency evaluation of the proposed method and compare it with SAM and SAM2 in Tab. 2. The proposed method demonstrates substantial improvements over SAM in both accuracy and efficiency. Although our model architecture is built upon SAM2 with three supplementary modules, it results in a minor decline in fps and a modest increase in memory consumption as opposed to SAM2. In the LA module, sub-images and the original image are concatenated along the channel dimension, forming a batch before feeding it into the encoder, and the additional attention operation is conducted within low-dimensional features. These design thus brings only negligible extra overhead. Compared to SAM2, the proposed method introduces only a few trainable parameters (5%), which is lightweight for training. Therefore, the efficiency trade-off can be considered acceptable given the substantial improvement in model accuracy.

Table 2. Efficiency evaluation of the proposed method. Results are tested on a single RTX-3090 GPU.

Method	FPS	Inference Mem(G)	Params(M)	Learnable(M)
SAM(vit_h)	2.24	10.3	2446	2446
SAM2(hiera_l)	5.76	4.6	216	216
Ours(hiera_l)	4.92	5.6	228	12

2.3. More qulitative results

In Figure 2, we present qualitative results of the straightforward predictions using bounding box prompt from the proposed SAM2Refiner, SAM and HQ-SAM in some challenging samples. As illustrated in the figure, SAM and HQ-SAM are susceptible to interference from objects surrounding the main subject, often resulting in noisy masks. In contrast, our proposed SAM2Refiner exhibits an outstanding ability to capture exquisitely fine details and accurately perceive complex topological structures.

2.4. Interaction Evaluation

To verify the flexible prompting ablity of the proposed method, Figure 3 provides segmentation results before and after single point click with user intention. It can be observed that our method outperforms SAM and HQ-SAM in the first round segmentation while strictly follows the second round prompting with correct and accurate results.

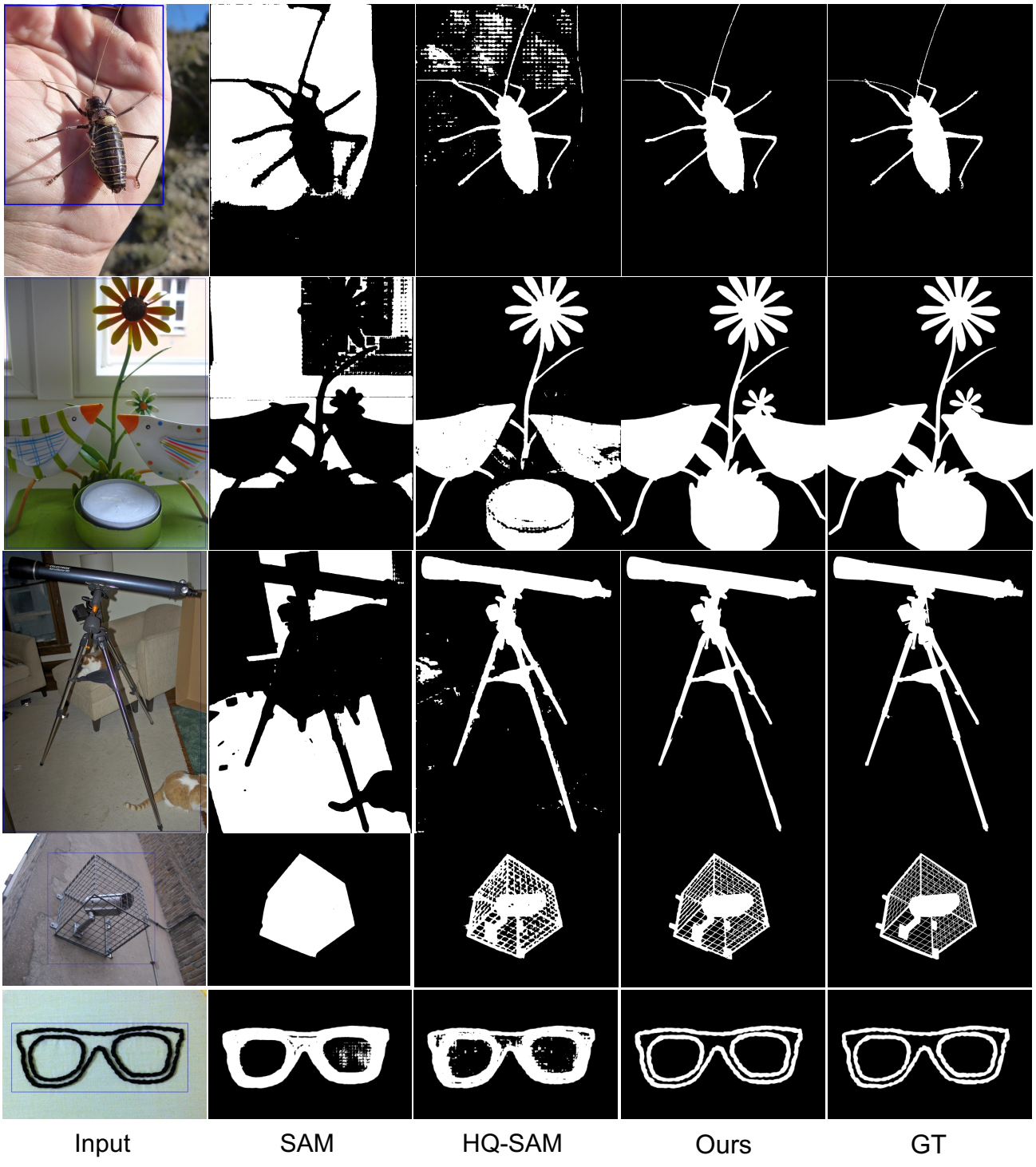

Figure 2. Qulitative comparisions between the proposed SAM2Refiner with SAM and HQ-SAM. Our SAM2Refiner demonstrates a remarkable capability to capture fine details in challenge samples. The blue box represent the bounding box prompt for each case.

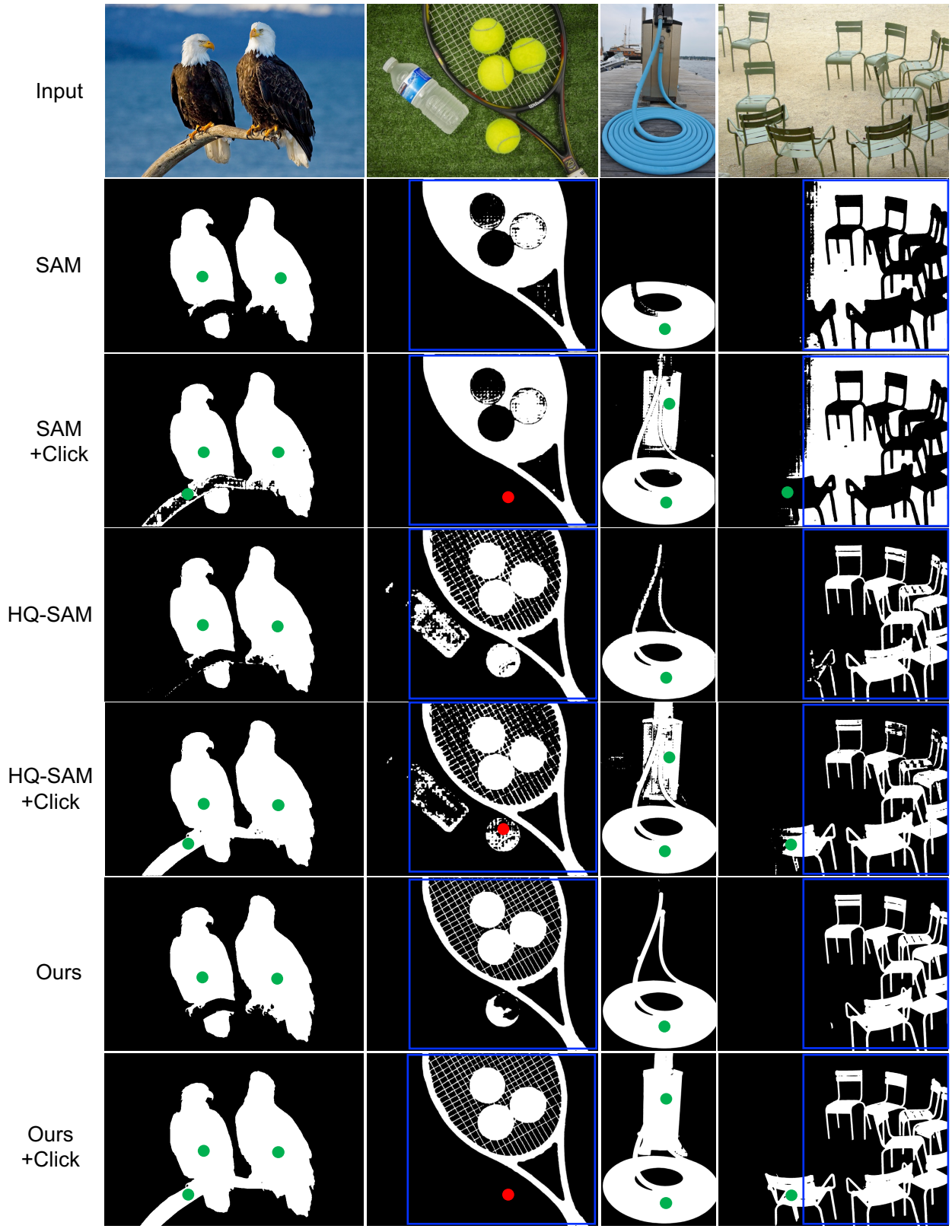

Figure 3. Qulitative comparisions between the proposed SAM2Refiner with SAM and HQ-SAM in prompting ablity. Points or bounding boxes are used as the first round prompts, and single positive/negative points are used as the second round prompts. Our SAM2Refiner exhibits robust prompting ablity while achieving superior performances. Here green points represent positive(foreground) click, red point represent negative(background) click, and blue box represent the bounding box prompt.