Semantic-aware SAM for Point-Prompted Instance Segmentation

Supplementary Material

7. Segmentation branch

Multi-mask Proposals Supervision. We utilize SOLOv2[46] as our segmentation branch for its efficiency. Alongside using $Mask_{prm}$ from the box_{prm} mapping for segmentation supervision, we integrated a proposal filter, which ranks proposals in $H \in \mathbb{R}^{N\times M}$ based on PRM scores, extracting initial masks to yield $Mask_{sam}$. Combined with $Mask_{prm}$, these guide the segmentation network. The loss function is as follows:

$$\begin{aligned}\mathcal{L}_{mask} = \mathcal{L}_{Dice}(Mask_{pre}, Mask_{prm}) \\ +\gamma \cdot \mathcal{L}_{Dice}(Mask_{pre}, Mask_{sam})\end{aligned} \tag{11}$$

where $\mathcal{L}_{Dice}$ is the Dice Loss [37], γ is set as 0.25.

Inference. For SAPNet's inference process, only the segmentation branch is retained after training, which is identical to the original instance segmentation model[46]. Given an input image, mask predictions are directly produced via an efficient Matrix-NMS technique. The pseudo-mask selection procedures of PSM, PRM, and other MIL-based modules only introduce computational overhead during training; they are entirely cost-free during inference.

8. Visualization and ablation studies on COCO

Implementation Details. On COCO and VOC2012SBD datasets, the initial learning rates were 1.5×10^{-2} and 2×10^{-3}, respectively, reduced by a factor of 10 at the 8th and 10th epochs. In PDG, the exponential factor d was set to 0.015. For NPG, the IoU threshold T_{neg1} and T_{neg2} were 0.3 and 0.5, respectively. For BMS, the k was set to 3. From the mask bag, we selected masks with the highest, medium, and lowest scores outside of $Mask_{prm}$ for MPS to accelerate the convergence of segmentation. Training spanned 12 epochs. Single-scale evaluation (1333×800) was used for the 1x schedule. For the 3x schedule, a multi-scale training approach was adopted, with the image's short side resized between 640 and 800 pixels (in 32-pixel increments) and MPS will be removed . Inference was conducted using single-scale evaluation.

Visualization. The proposed SAPNet significantly mitigates SAM's semantic ambiguity. Fig. 5 reveals SAM's limitations in discerning semantic significance from point prompts (the category indicated by each point). With SAPNet's refinement, each object is equipped with a mask that accurately represents its semantic category. The fidelity of masks selected by SAPNet is distinctly higher than that of top-1 masks from SAM.

To illustrate the advantages of SAPNet over the single-MIL approach, Fig. 6 and Fig. 7 contrast the outcomes of the local segmentation problems and the scenario involving proximate objects of the same class, respectively. With the chosen masks encompassing the entirety of the objects, post-selection and refinement via SAPNet markedly alleviate the local segmentation issues. Furthermore, by incorporating point distance guidance, SAPNet achieves commendable segmentation results even with the adjacent objects of the same class, successfully isolating the masks pertinent to each specific object.

Algorithm 2 Box Mining Strategy

Input: T_{min1},T_{min1},$box_{select}$from PRM stage,initial positive bagB^+,image I.
Output: Final bounding box for PRM box_{prm}.

for $i \in N$,N is the number of object in image I **do**
  $G_i \leftarrow select(B_i, top_k)$,$B_i \in B^+$;
  $count = 0$;
  if $proposal_j^w \cdot proposal_j^h > box_{select}^h \cdot box_{select}^w$ for each $proposal_j \in G_i$ **then**
    $iou = IOU(proposal_j, box_{select})$;
    if $iou > T_{min1}$ **then**
      box_{prm} = $(proposal_j + iou \cdot box_{select})/(iou + 1)$;
      $T_{min1} = iou$
      $count+ = 1$;
    else if $count == 0$ and $box\_select \in proposal_j$ and $iou > T_{min2}$ **then**
      box_{prm} = $(proposal_j \cdot iou + box_{select})/(iou + 1)$;
      $T_{min2} = iou$
    end if
  end if
end for

Fig. 8 shows additional instance segmentation results of SAPNet on the COCO dataset, demonstrating superior segmentation performance both in the case of individual large objects and in denser scenes. Our method exhibits outstanding capabilities in segmenting singular large targets as well as operating effectively in complex, crowded environments.

Ablation studies for negative proposals. In Tab. 8, we evaluate the effect of different threshold settings on SAPNet's final segmentation performance when the negative examples are generated in the NPG on the COCO dataset. We find that the two threshold pairs have less effect on the final segmentation performance with different settings, and the module is robust to hyperparameters.

T_{neg1}	T_{neg2}	AP	AP_{50}	AP_{75}	AP^s	AP^m	AP^l
0.3	0.3	30.9	51.3	32.0	12.2	34.7	47.4
0.3	0.5	**31.2**	**51.8**	**32.3**	**12.6**	**35.1**	**47.8**
0.5	0.3	30.8	51.1	32.0	12.1	34.7	47.3
0.5	0.5	31.1	51.8	32.4	12.4	35.2	47.4

Table 8. Constraints in negative proposals generation.

9. Detection and segmentation performance of SAPNet

Detection performance on COCO17 and VOC2012SBD. As shown in Tab. 9, utilizing the COCO and VOC datasets, we conduct an extensive comparison of the proposed methodology against a range of detection methods, encompassing fully, image-level, and point annotation. On the COCO dataset, our method demonstrates a notable improvement over the current SOTA P2BNet [9], with a substantial increment of 10.4 AP and culminating in a score of (32.5 AP *vs* 22.1 AP). Moreover, under a training schedule extended to 3x, the detection efficacy of SAPNet equates to that of the fully-annotated FPN [33]. Also, under a 3x training schedule, within the detection of the VOC dataset, our approach exceeds the previous SOTA by 8.0 AP_{50}, approximating 91% efficacy of the fully-annotated FPN. We observe that the image-level methods significantly underperform that of the point-annotated methods on the challenging COCO dataset, achieving only 36% of the performance of the fully-annotated approaches. That distinctly accentuates the advantageous that point-prompted methods, optimizing the trade-off between the economy of annotation efforts and the robustness of detection performance.

Method	Sup.	BB.	COCO17(AP)	VOC12(AP_{50})
FPN [33]	$\mathcal{B}$	R-50	37.4	75.3
CASD [19]	$\mathcal{I}$	VGG-16	12.8	53.6
CASD [19]	$\mathcal{I}$	R-50	13.9	56.8
OD-WSCL [40]	$\mathcal{I}$	VGG-16	13.6	56.2
OD-WSCL [40]	$\mathcal{I}$	R-101	14.4	-
UFO^2 [38]	$\mathcal{P}$	VGG-16	13.0	41.0
UFO^{2‡} [38]	$\mathcal{P}$	R-50	13.2	38.6
P2BNet-FR [9]	$\mathcal{P}$	R-50	22.1	48.3
SAPNet(*ours*)	$\mathcal{P}$	R-50	**32.5**	**64.8**
SAPNet(*ours*)*	$\mathcal{P}$	R-101	**37.2**	**68.5**

Table 9. Object detection performance on COCO 2017 and VOC 2012 $test$ sets. The evaluation metric are AP and AP_{50} respectively.

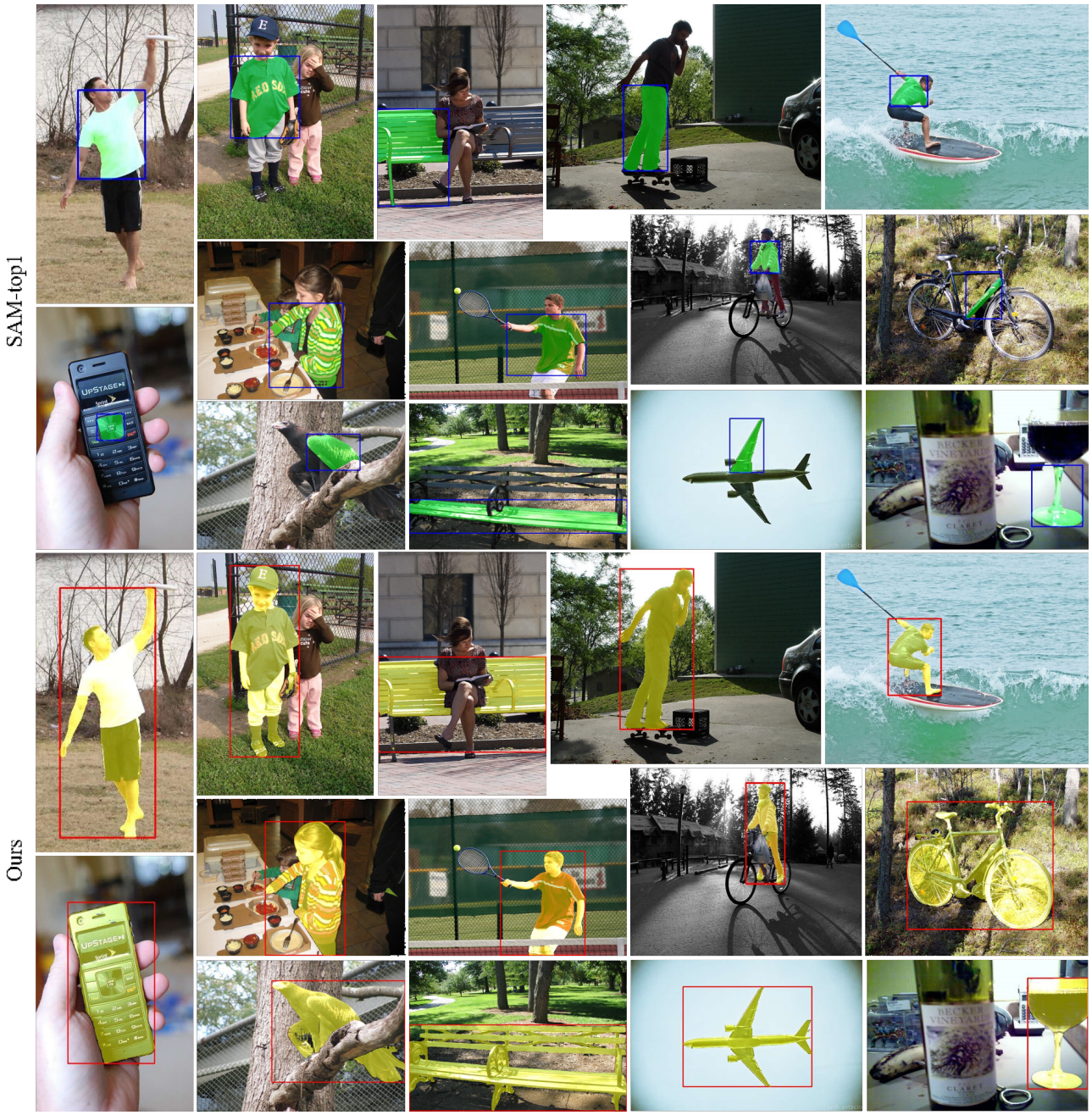

Figure 5. Visualization comparison between SAM-top1 and SAPNet on COCO 2017 dataset about semantic ambiguity, showing SAM's segmentation outcomes top-1 in green masks and our results in yellow masks. The blue and red bounding boxes highlight the respective mask boundaries.

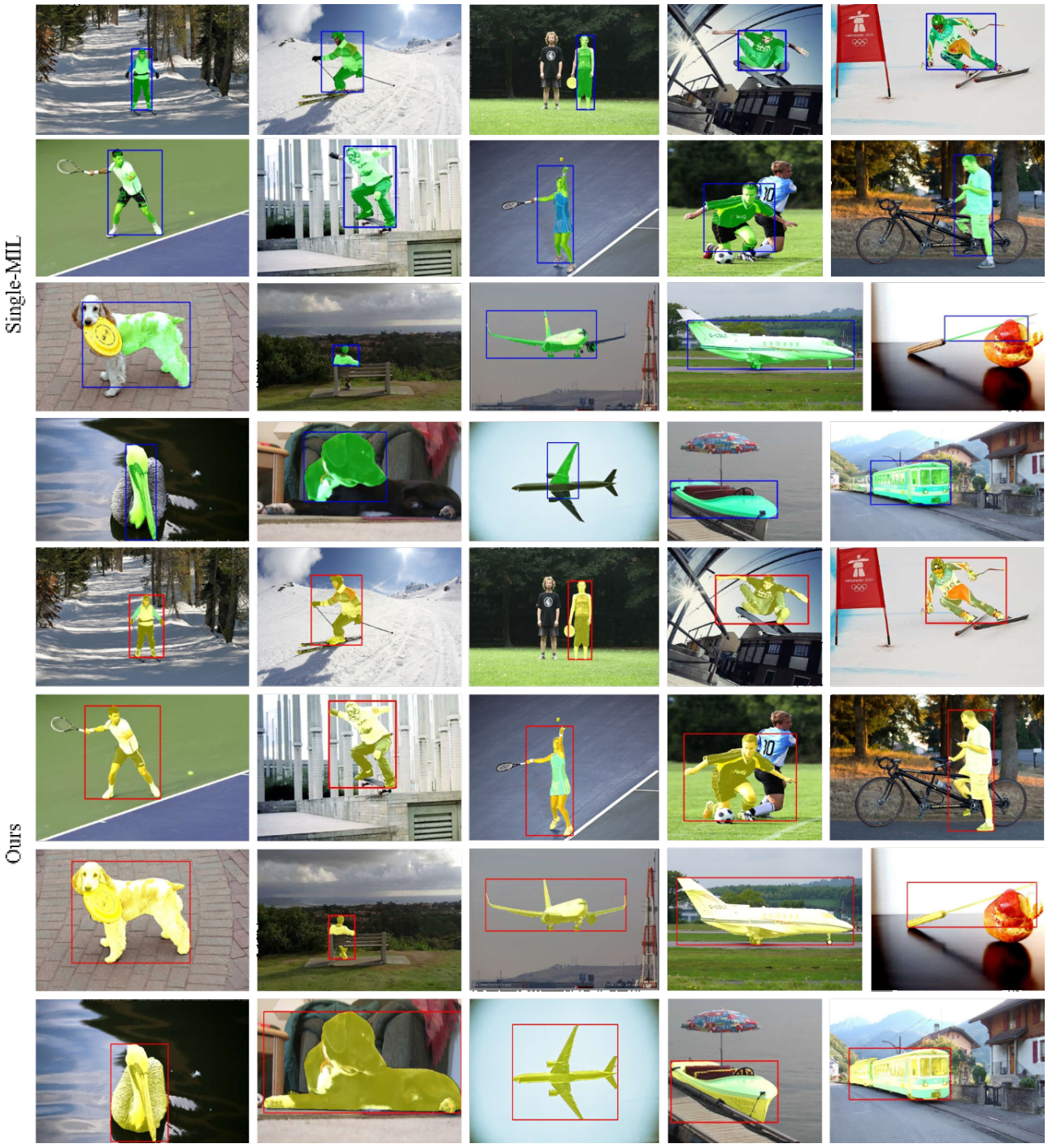

Figure 6. Visualization comparison between the single-MIL and SAPNet on COCO 2017 dataset about local segmentation, showing single-MIL's outcomes in green masks and our results in yellow masks. The blue and red bounding boxes highlight the respective mask boundaries.

Figure 7. Visualization of the segmentation problem for similar neighboring objects, showing single-MIL's outcomes in green masks and our results in yellow masks. The blue and red bounding boxes highlight the respective mask boundaries.

Figure 8. Visualization of instance segmentation results utilizing the Resnet-101-FPN backbone. The model is trained on the COCO train2017 dataset.