WISH: Weakly Supervised Instance Segmentation using Heterogeneous Labels Supplementary Material

1. Details about Handling Class Tags

In the proposed heterogeneous setting, all types of weak labels must be mapped to the same representation space. To achieve this, we utilize SAM's pre-trained prompt latent space. However, unlike points or boxes, class tags do not explicitly provide spatial information. As discussed in Sec. 4.3, class tags require special handling, which is described in detail in this section.

While various approaches could convert the information embedded in class tags into spatial localization, this paper adopts a CAM-based method, widely used in weakly supervised segmentation. First, we generate CAMs using a CAM-head trained with image-level classification loss. The CAM of each class is scaled into the range of [0,1]. From the CAM of each class, multiple peak points are sampled using a local maximum filter (implemented in NumPy), following the approach used in S2C [18]. During peak sampling, local maxima with activation scores below a threshold of $\tau = 0.5$ are considered false activations and are discarded for robustness.

In some cases, multiple peaks may correspond to the same instance. To address this, we perform SAM inference using each peak as an input point prompt and obtain the corresponding SAM masks. If the maximum IoU (among SAM's three mask levels) between the masks exceeds a certain threshold, we hypothesize that these masks represent the same instance. The corresponding peaks are then merged. For merging, we select the peak with the highest SAM mask stability score, akin to the philosophy behind Non-Maximum Suppression (NMS). Note that this process allows for more than two peaks to be merged.

2. Exploring More Combinations

Given the prohibitive number of possible annotation-type combinations, we further explored interpolation between the heterogeneous tag-point setting (Table 3 in the main paper) and the homogeneous box setting. As shown in Table 1, we identified a mixed configuration that marginally outperforms both extremes.

In addition, we also conduct experiments considering the fixed time-budget. Prior work [2] reports a time-budget ratio of approximately 1:1.2 for tag versus point annotation on PASCAL. Although our empirical measurements differ slightly, we nonetheless replicated this ratio in Table 2. Under an equivalent time budget for tags and points, the heterogeneous tag-point setting loses its advantage. Importantly: (1) WISH already achieves state-of-the-art performance in homogeneous settings, and (2) the benefit of heterogeneous supervision may vary by practical context.

Table 1. Experiments on more combinations.

T	5,290	3,966	2,644	0
P	882	661	441	0
B	0	220	441	882
AP	47.3	**47.7**	46.9	45.1

Table 2. Experiments on the fixed time-budget.

T	10,582	7,055	3,527	0
P	0	2,939	5,879	8,818
AP	46.0	48.1	49.5	**50.6**

3. Hyperparameters

We adopted $\alpha = 2$ and $\gamma = 5$ from Mask2Former and set $\beta = 5$ heuristically, treating prompts similarly to segmentation masks. Table 3 shows robustness when varying β (2–5) on PASCAL with point supervision; $\beta = 4$ yields marginally better results, indicating prompt information sits between class and mask supervision.

Table 3. Impact of β.

β	2	3	4	5
AP	51.9	52.3	**52.8**	52.4