## 6 Solution to Eq. (10)

We focus on the optimization problem

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} \quad & \mathcal{L}(\theta, \mathbf{U}, \mathbf{V}) \\ \text{s.t.} \quad & \mathbf{U} \leq \mathbf{1}, \mathbf{V} \leq \mathbf{1}, \mathbf{U} + \mathbf{V} \geq \mathbf{1}. \end{aligned} \tag{12}$$

Importantly, three constraints are imposed on $\mathbf{U}$ and $\mathbf{V}$. The first two constraints establish an upper limit on mask fusion, while the third constraint ensures that at least one of the two masks actively contributes to the process. This optimization problem can be rewritten as

$$\begin{aligned} \min_{\mathbf{U},\mathbf{V}} & \sum_{c \in \mathcal{C}^t} \mathcal{L}_{\mathrm{b}} \left( \left[ \mathbf{U} \odot R(S(\mathbf{x}^t, \mathbf{y}^t), \beta) + \mathbf{V} \odot \mathbf{P} \right]_c, \sigma(f_{\theta^t}(\mathbf{x}^t))_c \right) \\ \text{s.t.} \quad & \mathbf{U} \leq \mathbf{1}, \mathbf{V} \leq \mathbf{1}, \mathbf{U} + \mathbf{V} \geq \mathbf{1}. \end{aligned} \tag{13}$$

Note that the logistic function $\sigma(\cdot)$ is an element-wise operator. Therefore, the optimization problem in Eq. (13) is element-wise independent. To simplify the notation, we represent the elements of each symbol using lowercase letters. Suppose $r = R(S(\mathbf{x}^t, \mathbf{y}^t), \beta)_{ci}$ and $f = \sigma(f_{\theta^t}(\mathbf{x}^t))_{ci}$ for $\forall c \in \mathcal{C}^t, i = (h, w) \in \mathcal{I}$, with $\mathcal{L}_{\mathrm{b}}$ being the binary cross-entropy loss, the optimization problem in Eq. (13) can be transformed into a series of element-wise subproblems as

$$\begin{aligned} \min_{u,v} \quad & \mathcal{L}(u, v) = \log(\frac{f}{1-f})(u \cdot r + v \cdot p) \\ \text{s.t.} \quad & u \leq 1, v \leq 1, u + v \geq 1. \end{aligned} \tag{14}$$

Given that the partial derivative of $\mathcal{L}(u, v)$ w.r.t. $u$ or $v$ is a constant, the change of $\mathcal{L}(u, v)$ is monotonic when $u$ or $v$ increases or decreases, which implies that the optimal solution of $u$ and $v$ must be achieved at the boundary. Therefore, the solution to Eq. (13) can be easily derived as:

$$\begin{aligned} \mathbf{U}_c &= \phi \left( \mathbb{I}(f_{\theta^t}(\mathbf{x}^t)_c > 0 \vee R(S(\mathbf{x}^t, \mathbf{y}^t), \beta)_c \leq \mathbf{P}_c) \right), \\ \mathbf{V}_c &= \phi \left( \mathbb{I}(f_{\theta^t}(\mathbf{x}^t)_c > 0 \vee R(S(\mathbf{x}^t, \mathbf{y}^t), \beta)_c > \mathbf{P}_c) \right). \end{aligned} \tag{15}$$

## 7 Algorithm Procedure of Teddy

We present the complete procedure of Teddy in Alg. 1.

## 8 More Experiment Results

Here we present more complete experiment results including the disjoint setting in Table 4. Besides, we provide ablation study on 10-5 VOC and COCO-to-VOC settings in Tables 5-6.

**Algorithm 1** The pseudo code of Teddy

**Input:** $\mathcal{X}^t$, composed by $\mathbf{x}^t$ with its corresponding image-level annotation $\mathbf{y}^t$ at step $t$, and model $f_{\theta^{t-1}}$ trained at step $t-1$.
**Output:** The predicted label $y = \{\arg\max_{c \in \mathcal{Y}^t} p_c^i\}_{i=1}^N$, $p_c^i$ is the model prediction of pixel $i$ for class $c$ and $\mathcal{Y}^t$ is the set of seen classes;
**while** *epoch in num_epoches* **do**
  **for** $(\mathbf{x}^t, \mathbf{y}^t)$ in $\mathcal{X}^t$ **do**
    Compute seed areas $S(\mathbf{x}^t, \mathbf{y}^t)$.
    Generate image-level prediction with Global Weighted Pooling and focal penalty based on the seed areas as $\gamma(S(\mathbf{x}^t, \mathbf{y}^t))$.
    Compute output from the previous model as $f_{\theta^{t-1}}(\mathbf{x}^t)$.
    Train seed areas with $\mathcal{L}_{cls}$ and $\mathcal{L}_{loc}$ according to Eq. (3, 4).
    **if** *epoch* $\geq 5$ **then**
      Compute model predictions $f_{\theta^t}(\mathbf{x}^t)$.
      Obtain the binarized predictions for old classes $R(f_{\theta^{t-1}}(\mathbf{x}^t), \alpha)$.
      Derive $\mathbf{P}$ based on the seed areas.
      Compute seed areas $S(\mathbf{x}^t, \mathbf{y}^t) = (\mathbf{1} - R(f_{\theta^{t-1}}(\mathbf{x}^t), \alpha))S(\mathbf{x}^t, \mathbf{y}^t)$ based on TME constraint.
      Compute $\mathbf{U}$ and $\mathbf{V}$ according to Eq. (11).
      Obtain the predictions for new classes $\mathbf{Z}$ according to Eq. (6).
      Obtain pixel-level supervision $\mathbf{G}$ for all classes according to Eq. (7).
      Train the segmentation model $f_\theta$ with $\mathcal{L}_{seg}$ according to Eq. (8).
    **end if**
  **end for**
**end while**

## 9 Sensitivity Analysis

We have conducted a sensitivity analysis for the hyper-parameters $\alpha$ and $\beta$, which govern the ratio used during the binarization of predictions based on SAM. The results of this analysis are presented in Tab. 7. Notably, the results exhibit a high degree of stability even when $\alpha$ varies over a wide range. This stability serves as validation for the accuracy and consistency of predictions for the old classes generated by the previous model.

## 10 Discussion on Semantic Foundation Model

Recently, many large models have been proposed in semantic segmentation. They achieve impressive performances, especially for their adeptness in weakly supervised and even zero-shot learning scenarios. Although it may seem that they can do anything in the computer vision community, this does not devalue research in specific segmentation challenges. These tasks often require fine-tuning of semantic foundation models for unique goals - a process demanding significant resources. Additionally, exploring specific scenarios can further advance these models, driving the evolution of research in the computer vision community.

**Table 4:** Results on 15-5 VOC, 10-10 VOC and COCO-to-VOC settings. "P" indicates pixel-level labels and "I" indicates image-level ones. The best method of utilizing image-level supervision is bold, and the best method using pixel level supervision is underlined. FT is a simple fine-tuning approach, deciding the lower bound of an incremental model, and Joint is trained on all the classed in one step, which can be served as the upper bound.

| Method | Sup | **15-5 VOC** | | | | | | **10-10 VOC** | | | | | | **COCO-to-VOC** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Disjoint | | | Overlap | | | Disjoint | | | Overlap | | | COCO | | |
| | | 1-15 | 16-20 | All | 1-15 | 16-20 | All | 1-10 | 11-20 | All | 1-10 | 11-20 | All | 1-60 | 61-80 | All |
| FT | P | 8.4 | 33.5 | 14.4 | 12.5 | 36.9 | 18.3 | 7.7 | 60.8 | 33.0 | 7.8 | 58.9 | 32.1 | 1.9 | 41.7 | 12.7 |
| LWF [34] | P | 39.7 | 33.3 | 38.2 | 67.0 | 41.8 | 61.0 | 63.1 | 61.1 | 62.2 | 70.7 | 63.4 | 67.2 | 36.7 | 49.0 | 40.3 |
| LWF-MC [47] | P | 41.5 | 25.4 | 37.6 | 59.8 | 22.6 | 51.0 | 52.4 | 42.5 | 47.7 | 53.9 | 43.0 | 48.7 | - | - | - |
| ILT [41] | P | 31.5 | 25.1 | 30.0 | 69.0 | 46.4 | 63.6 | 67.7 | 61.3 | 64.7 | 70.3 | 61.9 | 66.3 | 37.0 | 43.9 | 39.3 |
| CIL [30] | P | 42.6 | 35.0 | 40.8 | 14.9 | 37.3 | 20.2 | 37.4 | 60.6 | 48.8 | 38.4 | 60.0 | 48.7 | - | - | - |
| MiB [8] | P | 71.8 | 43.3 | 64.7 | 75.5 | 49.4 | 69.0 | 66.9 | 57.5 | 62.4 | 70.4 | 63.7 | 67.2 | 34.9 | 47.8 | 38.7 |
| PLOP [16] | P | 71.0 | 42.8 | 64.3 | 75.7 | 51.7 | 70.1 | 63.7 | 60.2 | 63.4 | 69.6 | 62.2 | 67.1 | 35.1 | 39.4 | 36.8 |
| SDR [42] | P | 73.5 | 47.3 | 67.2 | 75.4 | 52.6 | 69.9 | 67.5 | 57.9 | 62.9 | 70.5 | 63.9 | 67.4 | - | - | - |
| RECALL [39] | P | 69.2 | 52.9 | 66.3 | 67.7 | 54.3 | 65.6 | 64.1 | 56.9 | 61.9 | 66.0 | 58.8 | 63.7 | - | - | - |
| CAM [62] | I | 69.3 | 26.1 | 59.4 | 69.9 | 25.6 | 59.7 | 65.3 | 41.3 | 54.5 | 70.8 | 44.2 | 58.5 | 30.7 | 20.3 | 28.1 |
| SEAM [56] | I | 71.0 | 33.1 | 62.7 | 68.3 | 31.8 | 60.4 | 65.1 | 53.5 | 60.6 | 67.5 | 55.4 | 62.7 | 31.2 | 28.2 | 30.5 |
| SS [3] | I | 71.6 | 26.0 | 61.5 | 72.2 | 27.5 | 62.1 | 60.7 | 25.7 | 45.0 | 69.6 | 32.8 | 52.5 | 35.1 | 36.9 | 35.5 |
| EPS [32] | I | 72.4 | 28.5 | 65.2 | 69.4 | 34.5 | 62.1 | 64.2 | 54.1 | 60.6 | 69.0 | 57.0 | 64.3 | 34.9 | 38.4 | 35.8 |
| WILSON [7] | I | 73.6 | 43.8 | 67.3 | 74.2 | 41.7 | 67.2 | 64.5 | 54.3 | 60.8 | 70.4 | 57.1 | 65.0 | 39.8 | 41.0 | 40.6 |
| **Teddy** | I | **74.5** | **48.1** | **69.0** | **77.6** | **51.4** | **72.0** | **65.4** | **55.2** | **61.7** | **71.2** | **59.4** | **66.5** | **40.6** | **41.8** | **41.5** |
| Joint | P | 75.5 | 73.5 | 75.4 | 75.5 | 73.5 | 75.4 | 76.6 | 74.0 | 75.4 | 76.6 | 74.0 | 75.4 | - | - | - |

**Table 5:** Ablation study. "OB" stands for prediction binarization for old classes, "TME" stands for tendency-driven mutual exclusivity and "PF" stands for prediction fusion for new classes.

| Row | TME | | PF | **10-5 VOC** | | | **COCO-to-VOC** | | |
|---|---|---|---|---|---|---|---|---|---|
| | OB | w/o OB | | 1-10 | 11-20 | All | 1-60 | 61-80 | All |
| 1 | | | | 66.8 | 46.5 | 58.1 | 39.8 | 41.0 | 40.6 |
| 2 | | | ✓ | 67.0 | 50.7 | 60.3 | 40.6 | 39.7 | 40.9 |
| 3 | | ✓ | | 67.2 | 51.6 | 60.8 | 40.9 | 39.8 | 41.1 |
| 4 | ✓ | | | 67.6 | 52.0 | 61.2 | 41.1 | 39.8 | 41.3 |
| 5 | | ✓ | ✓ | 67.2 | 51.8 | 60.9 | 41.1 | 40.0 | 41.3 |
| 6 | ✓ | | ✓ | 68.9 | 51.7 | 61.7 | 40.6 | 41.8 | 41.5 |

Moreover, while models like SemanticSAM [33] are proficient in open-set segmentation, they assume all classes are known during one step training. The need for these models to adapt to unknown labels underscores the importance of incremental learning, enabling them to integrate new information naturally, akin to human learning. Given the high costs and efforts needed for pixel-level annotations, WILSS is highly motivated and relevant to the community.

**Table 6:** Effectiveness of prediction fusion for new classes through optimization.

| U | V | 10-5 VOC | | | COCO-to-VOC | | |
|---|---|---|---|---|---|---|---|
| | | 1-10 | 11-20 | All | 1-60 | 61-80 | All |
| 0.25 | 0.75 | 66.6 | 46.3 | 58.0 | 40.6 | 37.0 | 40.2 |
| 0.50 | 0.50 | 67.1 | 51.2 | 60.5 | 40.6 | 39.9 | 40.9 |
| 0.75 | 0.25 | 67.0 | 50.8 | 60.3 | 40.4 | 38.1 | 40.5 |
| Optimization | | 68.9 | 51.7 | 61.7 | 40.6 | 41.8 | 41.5 |

**Table 7:** Sensitivity analysis of Teddy on $\alpha$ and $\beta$ under 15-5 VOC setting.

| $\alpha$ | $\beta$ | Disjoint | | | Overlap | | |
|---|---|---|---|---|---|---|---|
| | | 1-15 | 16-20 | All | 1-15 | 16-20 | All |
| 0.9 | | 74.3 | 46.0 | 68.4 | 77.2 | 51.1 | 71.6 |
| 0.8 | | 74.5 | 48.1 | 69.0 | 77.1 | 51.3 | 71.6 |
| 0.7 | 0.5 | 74.5 | 48.1 | 69.0 | 77.1 | 51.1 | 71.5 |
| 0.6 | | 74.3 | 48.1 | 68.8 | 77.1 | 51.2 | 71.5 |
| 0.5 | | 74.1 | 47.9 | 68.6 | 77.3 | 49.3 | 71.3 |
| | 0.9 | 74.3 | 45.4 | 68.3 | 77.6 | 50.2 | 71.7 |
| | 0.8 | 74.4 | 45.9 | 68.4 | 77.0 | 50.7 | 71.4 |
| 0.8 | 0.7 | 74.4 | 46.0 | 68.5 | 76.9 | 50.5 | 71.3 |
| | 0.6 | 74.4 | 45.9 | 68.4 | 76.8 | 50.2 | 71.2 |
| | 0.5 | 74.5 | 48.1 | 69.0 | 77.6 | 51.4 | 72.0 |