# Tendency-driven Mutual Exclusivity for Weakly Supervised Incremental Semantic Segmentation

Chongjie Si, Xuehui Wang, Xiaokang Yang, and Wei Shen[✉]

MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
{chongjiesi, wangxuehui, xkyang, wei.shen}@sjtu.edu.cn

Abstract. Weakly Incremental Learning for Semantic Segmentation (WILSS) leverages a pre-trained segmentation model to segment new classes using cost-effective and readily available image-level labels. A prevailing way to solve WILSS is the generation of seed areas for each new class, serving as a form of pixel-level supervision. However, a scenario usually arises where a pixel is concurrently predicted as an old class by the pre-trained segmentation model and a new class by the seed areas. Such a scenario becomes particularly problematic in WILSS, as the lack of pixel-level annotations on new classes makes it intractable to ascertain whether the pixel pertains to the new class or not. To surmount this issue, we propose an innovative, tendency-driven relationship of mutual exclusivity, meticulously tailored to govern the behavior of the seed areas and the predictions generated by the pre-trained segmentation model. This relationship stipulates that predictions for the new and old classes must not conflict whilst prioritizing the preservation of predictions for the old classes, which not only addresses the conflicting prediction issue but also effectively mitigates the inherent challenge of incremental learning - catastrophic forgetting. Furthermore, under the auspices of this tendency-driven mutual exclusivity relationship, we generate pseudo masks for the new classes, allowing for concurrent execution with model parameter updating via the resolution of a bi-level optimization problem. Extensive experiments substantiate the effectiveness of our framework, resulting in the establishment of new benchmarks and paving the way for further research in this field.

1 Introduction

Incremental learning (IL) emerges as a pivotal pathway for counteracting catastrophic forgetting [7, 20, 39] in deep learning paradigms. Predominantly in the area of image classification, numerous methodologies have been curated to fortify IL [29, 44, 46, 48, 56, 62]. Recently, semantic segmentation has witnessed a surge in methods tailored for incremental learning tasks (ILSS) [8, 16, 30, 41, 42]. These methods predominantly fall into regularization-based [8,16,41] or replay-based [58] categories, with the former prioritizing knowledge retention from previous training phases, and the latter employing additional samples to alleviate catastrophic forgetting. Nonetheless, a persistent hurdle is the dependency of ILSS approaches

on pixel-level annotations for new classes, which is a resource-intensive and time-consuming requirement.

To address this limitation, very recently, a novel task, weakly incremental learning for semantic segmentation (WILSS), has been proposed by [7]. In the WILSS paradigm, the initial training of a segmentation model takes place on a set of classes with pixel-level annotations. Thereafter, in the following incremental phases, WILSS shifts its focus exclusively to the utilization of image-level labels for new classes, with no access to old data. A prevalent way to tackle WILSS is the generation of seed areas for new classes to serve as pseudo pixel-level supervision. However, the seed areas are inevitably noisy, which might lead to a quandary: A pixel's label is predicted as belonging to an old class by the previous segmentation model but is simultaneously predicted as a new class by the seed area in the incremental phases. With only image-level labels available for the new classes, a formidable conundrum arises in determining whether this pixel's true label corresponds to the new classes or not. Given this scenario, a critical inquiry surfaces: how might we generate high-quality pseudo pixel-level labels for the new classes to address this conflict issue?

In this paper, we present Teddy, a novel approach designed to address the aforementioned challenge of conflicting predictions, as shown in Fig. 1. Specifically, the core of Teddy is the establishment of a unique tendency-driven relationship of mutual exclusivity (TME). This relationship is designed to harmonize the predictions between old and new classes. To elaborate, when a pixel is identified as belonging to an old class by the previous segmentation model, it is expected to be treated as background by the seed areas of new classes. TME not only effectively addresses the conflicting prediction issue, but also inherently strengthens the preservation of old class predictions, providing a robust countermeasure to the challenge of catastrophic forgetting.

Within the framework of Teddy, the TME relationship is meticulously crafted as a constraint for the optimization problem intrinsic to WILSS. To establish such constraint, we seek to binarize the predictions for the old classes and then produce predictions for new classes. The nexus between these old and new class predictions is deftly steered by the constraint of TME. Moreover, we leverage the capabilities of the Segment Anything Model (SAM) [27], to further bolsters this process. Finally, by formulating the optimization objective of WILSS with the TME constraint as a bi-level optimization problem, we are able to produce pixel-level pseudo labels for all classes. This process is executed concurrently with the updating of model parameters, ensuring an efficient and integrated approach.

The contributions of this paper are as follows:

- We propose a novel tendency-driven relationship of mutual exclusivity in WILSS, which effectively mitigates the conflict of the predictions generated by the pre-trained model and the seed areas.
- We propose a TME constrained bi-level optimization problem, through which we can simultaneously generate pixel-level pseudo labels for all classes and update the model parameters.

– Extensive experiments show that Teddy significantly outperforms state-of-the-art approaches in existing scenarios and data sets, establishing new benchmarks.

2   Related Work

### 2.1   Weakly Supervised Semantic Segmentation

Collecting precise pixel-wise annotations for fully supervised semantic segmentation is a resource-intensive and costly endeavor. In response to this challenge, researchers have introduced weakly-supervised semantic segmentation (WSSS) methods, which aims to train highly effective segmentation models using more cost-efficient forms of supervision, such as bounding boxes [12, 26, 43, 45], scribbles [34, 51, 52] and image-level labels [17, 19, 23, 25]. Due to its cost-effectiveness and widespread availability, image-level supervision has garnered the highest level of attention among various forms of weak supervision. The major of WSSS methods under image-level supervision [1, 2, 17, 19, 23, 25, 49] typically adheres to a three-stage learning process: initial Class Activation Map (CAM) [61] generation, followed by pseudo-mask generation, and the training of the segmentation model.

In contrast to the rapid advancements in pseudo-label generation techniques based on image-level supervision, the WSSS models learn from a predetermined set of classes, which can only tackle static scenario. Due to the growing attention for more intricate and demanding scenario of incremental learning, in this paper, we aim to extend a pre-trained segmentation model solely using image-level labels while accommodating the introduction of new classes over time.

2.2   Incremental Learning Semantic Segmentation

Incremental learning is devised to address the issue of catastrophic forgetting [7, 20, 39], which transpires when a model, in the process of acquiring knowledge regarding new classes over time, gradually loses its capacity to recall previously learned classes. Although an extensive investigation of incremental learning (IL) for image classification has been thoroughly conducted [6, 9, 14, 29, 44, 46, 48, 56, 60, 62], there exists only a limited number of methodologies proposed for addressing the incremental learning task for semantic segmentation (ILSS) [8, 16, 30, 41, 42]. The prevailing methods in ILSS can be broadly categorized into two groups, namely regularization-based and replay-based [58]. Regularization-based methods [8, 16, 41, 47] focus on preserving knowledge from prior training phases. For instance, PLOP [16] introduces a multi-scale pooling method that maintains spatial relationships across long and short distances at the feature level. Moreover, SDR [41] advocates enhancing class-conditional features by minimizing the discrepancy among features belonging to the same class. In replay-based methods, auxiliary samples are generated to counteract catastrophic forgetting. For instance, RECALL [38] tackles the forgetting problem by employing web-crawled images with pseudo labels.

Recently, a novel weakly incremental learning for semantic segmentation (WILSS) task was introduced by WILSON [7], wherein a pretrained segmentation model is employed to segment novel classes only with image-level labels, and without access to old data. As mentioned in Sec. 1, compared to ILSS, the problem of conflicting prediction becomes more pronounced in WILSS because of the absence of valid pixel-level supervisions. To address this issue, in this paper we aims to form a tendency-driven relationship of mutual exclusivity between the old and new predictions, which effectively tackles this issue and mitigates the catastrophic forgetting at the same time.
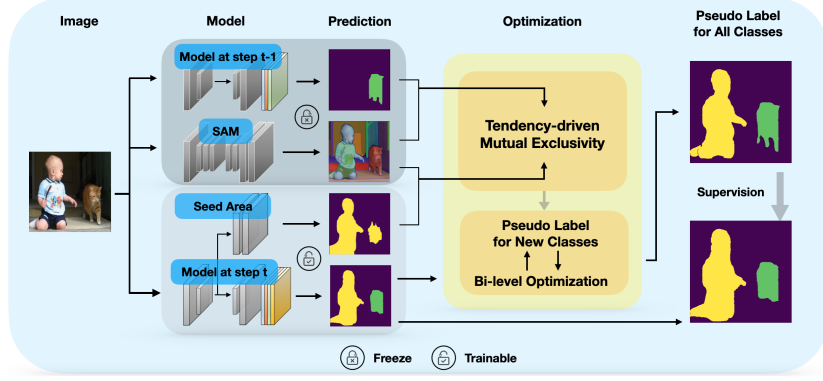
Fig. 1: Illustration of the proposed Teddy framework. A novel tendency-driven relationship of mutual exclusivity is proposed to regulate the interactions between the predictions produced by the seed areas and the previous model. Moreover, by solving a TME constrained bi-level optimization problem, we can generate pixel-level pseudo labels for all classes and update model parameters simultaneously.

2.3   Segment Anything Model

Recently, Segment Anything Model (SAM) [28], which is trained with 1B mask annotations and designed to distinguish a foreground region from a background one through some visual prompts (point, box, etc), has been proposed. SAM has surprising capabilities to segment objects accurately even if the objects are not involved in the training data set, which is ideally meaningful to produce class-agnostic masks for segmentation tasks. While SAM has been widely used [11, 36, 53, 54] in a relatively short period of time, it does not match the expected level of performance in many downstream tasks [37, 50, 59]. Therefore, how to harness the performance of SAM more effectively and imbue it with semantic information is a question worthy of profound consideration.

3   Method

Suppose $\mathcal{X}$ represents the input space, where each image $\mathbf{x} \in \mathcal{X}$ consists of a set of pixels denoted as $\mathcal{I}$, with a cardinality of $|\mathcal{I}| = H \times W = N$, without any loss of generality. Denote $\mathcal{Y}$ the label space. In the incremental segmentation setting, the training procedure is organized into multiple sequential steps. At each step t $(t > 0)$, we observe new classes $\mathcal{C}^t$, consequently expanding the label set to $\mathcal{Y}^t = \mathcal{Y}^{t-1} \cup \mathcal{C}^t$. Diverging from the original setting, in WILSS, pixel-level annotations are exclusively provided during the initial step $(t = 0)$. Subsequently, in the ensuing incremental steps, only cost-effective image-level labels are available, and access to prior training images is restricted. The primary objective of WILSS revolves around the adaptation of a model $f_\theta : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}| \times H \times W}$, which maps the input space to the pixel-level probability label space. Here, θ represents the parameters. The goal is to equip this model with the capacity to perform segmentation on new classes while retaining knowledge of the previously learned classes without incurring catastrophic forgetting.

3.1   Overview

In WILSS, the primary objective during step t $(t > 0)$ revolves around obtaining the optimal model parameters, symbolized as θ^t. To formalize this optimization task, we begin with the following loss function $\mathcal{L}$:

$$\min_{\theta^t} \quad \mathcal{L}(\theta^t | \theta^{t-1}; (\mathbf{x}^t, \mathbf{y}^t)). \tag{1}$$

Here, θ^{t-1} represents the parameters of the model at $t-1$ step, $\mathbf{x}^t$ denotes an image observed at step t and $\mathbf{y}^t \in \{0, 1\}^{|\mathcal{C}^t|}$ refers to the corresponding image-level label vector of that image. It is imperative to note that in the incremental steps, our access is confined solely to image-level labels and predictions for the previously learned classes produced by $f_{\theta^{t-1}}$. As discussed in the introduction, it is possible for $f_{\theta^{t-1}}$ to predict a pixel's label as belonging to a previously learned class, while the seed area concurrently predicts it as a new class.

To address this conflicting prediction issue, our proposed method, Teddy, aims to establish a novel tendency-driven relationship of mutual exclusivity (TME) between the seed areas representing new classes and the predictions for the old classes. To form the relationship, we first binarize the predictions of the old classes (Sec 3.2) which will used to form the TME constraint, and then generate seed areas for the new classes (Sec. 3.2). Subsequently, we elucidate how to form the TME and how it guides seed area refinement for the new classes (Sec. 3.2) as well as final pseudo pixel-level label generation for all the classes (Sec. 3.3). Finally, we demonstrate how to optimize the objective function with the constraint of TME (Sec. 3.4).

3.2   Tendency-driven Mutual Exclusivity

Prediction Binarization for Old Classes   Capitalizing on the fact that the model $f_{\theta^{t-1}}$ has been trained with pixel-level annotations, we anticipate that its

prediction for a given image $\mathbf{x}^t$, denoted as $f_{\theta^{t-1}}(\mathbf{x}^t) \in \mathbb{R}^{|\mathcal{Y}^{t-1}| \times H \times W}$, is more reliable than seed areas generated from image-level labels. We proceed to convert $f_{\theta^{t-1}}(\mathbf{x}^t)$ to its one-hot presentation $\delta(f_{\theta^{t-1}}(\mathbf{x}^t)) \in \{0,1\}^{|\mathcal{Y}^{t-1}| \times H \times W}$, where $\delta(\cdot)$ represents the *one-hot* function. Furthermore, we harness the power of SAM to generate m binary masks, each represented as $\mathbf{M}_i \in \{0,1\}^{H \times W}, i = 1, 2, ..., m$, to obtain the binarized predictions for old classes, which will be served as a part of TME. Specifically, $\mathbf{M}_i$ is considered as belonging to the c-th class if and only if

$$\frac{\sum(\delta(f_{\theta^{t-1}}(\mathbf{x}^t))_c \odot \mathbf{M}_i)}{\min\left(\sum(\delta(f_{\theta^{t-1}}(\mathbf{x}^t))_c), \sum(\mathbf{M}_i)\right)} > \alpha, \tag{2}$$

where $\odot$ denotes the element-wise Hadamard product between two matrices, $\sum(\cdot)$ denotes the sum of all elements in a matrix and $\alpha \geq 0.5$ is a hyper-parameter. It suggests that if the intersection area exceeds α times that of either set, mask $\mathbf{M}_i$ is regarded as belonging to the c-th class. To guarantee that $\mathbf{M}_i$ is classified into a single class, we set $\alpha \geq 0.5$. This prevents conflicts arising from masks overlapping with two old classes' predictions, each covering over α, leading to dual classification.

Based on the above assumption, we can have matrix $\mathbf{K} \in \{0,1\}^{m \times |\mathcal{Y}^{t-1}|}$, where $K_{ic} = 1$ means that mask $\mathbf{M}_i$ belongs to the c-th class, otherwise $K_{ic} = 0$. Finally, we obtain binarized predictions $\mathbf{R}^{t-1} \in \mathbb{R}^{|\mathcal{Y}^{t-1}| \times H \times W}$ for the old classes as $\mathbf{R}_c^{t-1} = \phi(\sum_{i=1}^m K_{ic}\mathbf{M}_i)$, where $\phi(\cdot)$ represents an element-wise operator, defined as $\phi(a) = \min(1, a)$ for a scalar a. To simplify notation, we introduce the function $R(\cdot)$ to denote this binarization process, i.e., $\mathbf{R}^{t-1} = R(f_{\theta^{t-1}}(\mathbf{x}^t), \alpha)$.

Seed Area Generation for New Classes   Drawing inspiration from the existing literature on WSSS [1, 2, 17, 19, 23, 25, 49], we can generate seed areas to produce pixel-level pseudo-supervision based on image-level labels for the new classes. The seed area generation process utilizes the features of the encoder within the segmentation model. For a given image $\mathbf{x}^t$ along with its corresponding image-level label $\mathbf{y}^t$, the *de facto* seed area generation method is class activation mapping (CAM) [61], which outputs a score map $S(\mathbf{x}^t, \mathbf{y}^t) \in \mathbb{R}^{|\mathcal{Y}^t| \times H \times W}$ for the image $\mathbf{x}^t$. Here following [7], we apply a function $\gamma(\cdot)$, which is a normalized Global Weighted Pooling [7] with the focal penalty term [7] which can better identify all visible parts of objects, on the score map to obtain the image-level prediction $\gamma(S(\mathbf{x}^t, \mathbf{y}^t)) \in \mathbb{R}^{|\mathcal{C}^t|}$. We then proceed to train the seed area by minimizing the classification loss $\mathcal{L}_{cls}$:

$$\min_{\theta^t} \quad \mathcal{L}_{cls} = \frac{1}{|\mathcal{C}^t|} \sum_{c \in \mathcal{C}^t} \mathcal{L}_{\mathrm{b}} \left\{ \mathbf{y}_c^t, \sigma(\gamma(S(\mathbf{x}^t, \mathbf{y}^t))_c) \right\}. \tag{3}$$

Here, $\sigma(\cdot)$ represents the logistic function and $\mathcal{L}_{\mathrm{b}}$ corresponds to the element-wise binary cross-entropy loss. Additionally, to inform seed area about the presence of old classes within the image and guide its attention towards alternative regions, we introduce another supervision provided by the segmentation model $f_{\theta^{t-1}}$, and

we aim to minimize the following objective function $\mathcal{L}_{loc}$:

$$\min_{\theta^t} \quad \mathcal{L}_{loc} = \frac{1}{|\mathcal{Y}^{t-1}|} \sum_{c \in \mathcal{Y}^{t-1}} \mathcal{L}_b \left\{ \sigma(f_{\theta^{t-1}}(\mathbf{x}^t))_c, \sigma(S(\mathbf{x}^t, \mathbf{y}^t))_c \right\}. \tag{4}$$
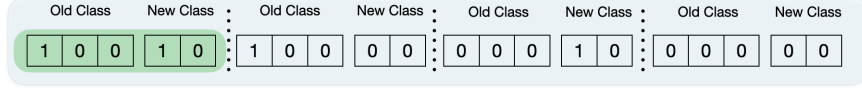
Fig. 2: A simple illustration for tendency-driven mutual exclusivity. In each case, the left vector represents $R(f_{\theta^{t-1}}(\mathbf{x}^t), \alpha)_i$, and the right one represents $\delta(S(\mathbf{x}^t, \mathbf{y}^t))_i$. TME will prevent the case under the green shadow while allowing the other three to occur.

TME Constraint As previously discussed, a pixel may be predicted as foreground by both $f_{\theta^{t-1}}(\mathbf{x}^t)$ and $S(\mathbf{x}^t, \mathbf{y}^t)$ in WILSS. To counteract this, as shown in Fig. 2, we aim to establish a relationship of mutual exclusivity between predictions for old and new classes. Specifically, we require that

$$\begin{aligned} &\|R(f_{\theta^{t-1}}(\mathbf{x}^t), \alpha)_i\| + \|\delta(S(\mathbf{x}^t, \mathbf{y}^t))_i\| \leq 1, \\ &\forall i = (h, w) \in \mathcal{I}. \end{aligned} \tag{5}$$

This constraint stipulates that for a given pixel, in the predictions of both new and old classes, at most one class is designated as foreground.

Delving deeper into the implications of this constraint, when a pixel is classified as an old class by $R(f_{\theta^{t-1}}(\mathbf{x}^t), \alpha)$, it is compelled to be classified as background by $\delta(S(\mathbf{x}^t, \mathbf{y}^t))$. In essence, the assignment of new labels is exclusively reserved for pixels predicted as background in the previous step. This constraint significantly preserves the predictions of the old classes and effectively mitigates the issue of conflicting prediction. Furthermore, the establishment of this mutual exclusive relationship hinges upon the predictions of the old classes, rendering it a tendency-driven form of mutual exclusivity. Moreover, Fig. 3 and the discussion in Sec. 4.6 demonstrate that the utilization TME is effective in addressing the issue of catastrophic forgetting. This relationship results in the generation of more accurate seed areas for the new classes, showcasing its potential to improve the quality and robustness of the model's performance when learning incrementally.

3.3 Pseudo Label Generation for All Classes

We proceed with the generation of pseudo masks for the new classes under the constraint of TME. Following [7], we can directly derive pseudo labels $\mathbf{P} \in \mathbb{R}^{|\mathcal{C}^t| \times H \times W}$ for the new classes, which is combination of one-hot and softmax forms of the seed areas $S(\mathbf{x}^t, \mathbf{y}^t)$. Subsequently, we can further utilize SAM to produce predictions for the new classes. These predictions are denoted as

Fig. 3: Qualitative results to show the effectiveness of TME. From left to right: image, binarized predictions for the old classes, seed areas without TME for the new classes, seed areas with TME for the new classes and refined seed areas with TME.

$R(S(\mathbf{x}^t, \mathbf{y}^t), \beta) \in \mathbb{R}^{|\mathcal{C}^t| \times H \times W}$, where $\beta \geq 0.5$ serves as a hyper-parameter. To fully amalgamating the advantages of both SAM and CAM, we fuse their predictions under the constraint of TME as

$$\mathbf{Z} = \mathbf{U} \odot R(S(\mathbf{x}^t, \mathbf{y}^t), \beta) + \mathbf{V} \odot \mathbf{P}$$
$$\text{s.t.} \quad \|R(f_{\theta^{t-1}}(\mathbf{x}^t), \alpha)_i\| + \|\delta(S(\mathbf{x}^t, \mathbf{y}^t))_i\| \leq 1, \tag{6}$$
$$\forall i = (h, w) \in \mathcal{I}.$$

Here, $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{|\mathcal{C}^t| \times H \times W}$ are two coefficient matrices that demands optimization, which will be discussed later.

To mitigate the risk of catastrophic forgetting, we use $f_{\theta^{t-1}}(\mathbf{x}^t)$ as the pixel-level pseudo label for the old classes. This substitution is made to counteract any potential bias introduced towards the new classes by the fused pseudo mask. The final pixel-level pseudo-label $\mathbf{G}$ is defined as:

$$\mathbf{G}_c = \begin{cases} \min\left(\sigma(f_{\theta^{t-1}}(\mathbf{x}^t))_c, \mathbf{Z}_c\right) & \text{if } c = \text{b}, \\ \mathbf{Z}_c & \text{if } c \in \mathcal{C}^t, \\ \sigma\left(f_{\theta^{t-1}}(\mathbf{x}^t)\right)_c & \text{if } c \in \mathcal{Y}^{t-1}, \end{cases} \tag{7}$$

where b represents the background class. We can then define the segmentation loss $\mathcal{L}_{seg}$ as:

$$\min_{\theta^t, \mathbf{U}, \mathbf{V}} \quad \mathcal{L}_{seg} = \frac{1}{|\mathcal{Y}^t|} \sum_{c \in \mathcal{Y}^t} \mathcal{L}_{\text{b}} \left\{ \mathbf{G}_c, \sigma(f_{\theta^t}(\mathbf{x}^t))_c \right\}$$
$$\text{s.t.} \quad \|R(f_{\theta^{t-1}}(\mathbf{x}^t), \alpha)_i\| + \|\delta(S(\mathbf{x}^t, \mathbf{y}^t))_i\| \leq 1, \tag{8}$$
$$\forall i = (h, w) \in \mathcal{I}.$$

3.4   TME Constrained Bi-level Optimization

Based on the preceding discussions, the objective function in Eq. (1) is finally formulated as

$$\min_{\theta^t, \mathbf{U}, \mathbf{V}} \quad \mathcal{L}(\theta^t | \theta^{t-1}; (\mathbf{x}^t, \mathbf{y}^t)) = \mathcal{L}_{cls} + \mathcal{L}_{loc} + \mathcal{L}_{seg}$$
$$\text{s.t.} \quad \|R(f_{\theta^{t-1}}(\mathbf{x}^t), \alpha)_i\| + \|\delta(S(\mathbf{x}^t, \mathbf{y}^t))_i\| \leq 1, \tag{9}$$
$$\forall i = (h, w) \in \mathcal{I}.$$

During training, we apply the TME constraint by converting the value of $S(\mathbf{x}^t, \mathbf{y}^t)_i$ into a zero vector if $\|R(f_{\theta^{t-1}}(\mathbf{x}^t), \alpha)_i\| = 1$ for any $i = (h, w) \in \mathcal{I}$. It is also imperative to determine the values of $\mathbf{U}$ and $\mathbf{V}$ with optimizing the network's parameters. We propose treating $\mathbf{U}$ and $\mathbf{V}$ as latent variables to be optimized concurrently with θ^t. By considering $\mathcal{L}$ as a function of θ^t, $\mathbf{U}$, and $\mathbf{V}$, we formulate the following bi-level optimization problem:

$$
\begin{aligned}
&\arg\min_{\theta^t} \quad \mathcal{L}(\theta^t, \mathbf{U}^*, \mathbf{V}^*); \\
&\text{subject to } \|R(f_{\theta^{t-1}}(\mathbf{x}^t), \alpha)_i\| + \|\delta(S(\mathbf{x}^t, \mathbf{y}^t))_i\| \leq 1, \\
&\qquad \forall i = (h, w) \in \mathcal{I}, \\
&\quad \mathbf{U}^*, \mathbf{V}^* = \arg\min_{\mathbf{U}, \mathbf{V}} \mathcal{L}(\theta^t, \mathbf{U}, \mathbf{V}).
\end{aligned}
\tag{10}
$$

It is essential to note that bi-level optimization problem has been proven strongly NP-hard [24]. Consequently, obtaining an exact solution for the problem (10) in polynomial time is impossible. Fortunately, the inner optimization problem in Eq. (10) has a closed-form solution, with the solution

$$
\begin{aligned}
\mathbf{U}_c &= \phi\left(\mathbb{I}(f_{\theta^t}(\mathbf{x}^t)_c > 0 \vee R(S(\mathbf{x}^t, \mathbf{y}^t), \beta)_c \leq \mathbf{P}_c)\right), \\
\mathbf{V}_c &= \phi\left(\mathbb{I}(f_{\theta^t}(\mathbf{x}^t)_c > 0 \vee R(S(\mathbf{x}^t, \mathbf{y}^t), \beta)_c > \mathbf{P}_c)\right).
\end{aligned}
\tag{11}
$$

Here, $\mathbb{I}(\cdot)$ represents the element-wise indicator function. For the detailed explanation on the solution process for $\mathbf{U}$ and $\mathbf{V}$, please refer to Supp. Sec. 6. Therefore, we employ an alternative optimization strategy to approximate the solution, in which $\mathbf{U}$ and $\mathbf{V}$ could be updated in an on-the-fly manner during each optimization iteration.

Moreover, the solution of $\mathbf{U}$ and $\mathbf{V}$ holds practical significance: when encountering a positive model prediction, an amalgamation of both masks occurs. However, if their cumulative value is minuscule, it's inferred as a model misjudgment. Conversely, for a negative model prediction, the lesser of the two masks is designated as the supervisory cue. If the smaller mask's value retains substantial magnitude, there exists room for further honing the model's prediction. This strategy weaves a symbiotic relationship between the model's forecasts and the dual-mask priors, culminating in the amplification of the model's performance.

The complete procedure of Teddy is shown in Alg. 1 in the Supp.

4 Experiments

4.1 Data sets

We present an extensive evaluation of Teddy on the Pascal VOC 2012 [18] and MS COCO [5]. Pascal VOC 2012 contains 1,464 training images and 1,449 validation ones with 20 classes. Following the standard method [7,31], Pascal VOC data set can be augmented with images from [21], leading to 10,582 images for training and 1,449 ones for validation. MS COCO is a large-scale data set consisting over 164K images across 80 classes. We also follow the training split and the annotation of [5] to deal with the overlapping annotation problem [35].

4.2   Protocols

Following previous works [7,8,38], two different incremental learning protocols are introduced: (1) **disjoint**, where images of each training step only contain current or previous seen classes; (2) **overlap**, where images of each training step contain any class, among which at least one is new. Obviously, the overlap scenario is more practical and challenging to the disjoint one, as in a real setting there isn't any oracle method to exclude future classes from the background. Hence, our experiments are primarily concentrated on the overlap setting. However, additional results pertaining to the disjoint setting are available in the supplementary materials.

We start to evaluate Teddy under two widely used **single-step** incremental learning scenarios: **15-5 VOC** setting, where 15 classes are learned in the pretraining step and another 5 classes are added in the second step, and **10-10 VOC** setting, in which 10 classes are first learned with another 10 being learned later. We also consider a more challenging single-step scenario, **COCO-to-VOC** [7]. Specifically, 60 classes of MS COCO which are not presented in VOC data set is learned in the first step, Afterwards, we learn 20 classes of Pascal VOC. Furthermore, we also consider **multi-step** settings: **10-2 VOC** setting, where 2 new classes are learned in 5 incremental steps and **10-5 VOC** setting, where 5 new classes are learned in each 2 incremental step. Following previous works [7,8,38], we report the results on the validation data set since the test set labels have not been publicly released.

4.3   Baselines

We compare our method with both supervised incremental learning and weakly supervised semantic segmentation methods, for WILSS is a novel setting proposed by WILSON [7]. For supervised incremental learning methods with pixel-wise annotations, the performance of eight representative state-of-the-art methods including LWF [33], LWF-MC [46], ILT [40], MIB [8], PLOP [16], CIL [30], SDR [41], and RECALL [38], are compared with. Moreover, following [7,8,38], we also compare with another two methods denoted as FT and Joint. 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. We refer the readers to [8] for more details on FT and Joint. As for weakly supervised semantic segmentation methods adopted to the incremental learning scenario, we compare Teddy with the reported results of the pseudo labels generated from four WSSS methods, including CAM [61], SEAM [55], SS [3] and EPS [32].

4.4   Implementation Details

We employ Deeplab V3 [10] for all the settings, with the backbone of ResNet-101 [22] for Pascal VOC and Wide-ResNet-38 [57] for COCO, both pretrained on ImageNet [13]. Similar to [7,8], we employ in-place activated batch normalization

Table 1: Results on different overlap 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 | Single-step | | | | | | | | | Multi-step | | | | | |
| | | 15-5 VOC | | | 10-10 VOC | | | COCO-to-VOC | | | 10-2 VOC | | | 10-5 VOC | | |
		1-15	16-20	All	1-10	11-20	All	1-60	61-80	All	1-10	11-20	All	1-10	11-20	All
FT	P	12.5	36.9	18.3	7.8	58.9	32.1	1.9	41.7	12.7	-	-	-	-	-	-
LWF [33]	P	67.0	41.8	61.0	<u>70.7</u>	63.4	67.2	36.7	<u>49.0</u>	<u>40.3</u>	-	-	-	-	-	-
LWF-MC [46]	P	59.8	22.6	51.0	53.9	43.0	48.7	-	-	-	-	-	-	-	-	-
ILT [40]	P	69.0	46.4	63.6	70.3	61.9	66.3	<u>37.0</u>	43.9	39.3	-	-	-	-	-	-
CIL [30]	P	14.9	37.3	20.2	38.4	60.0	48.7	-	-	-	-	-	-	-	-	-
MiB [8]	P	75.5	49.4	69.0	70.4	63.7	67.2	34.9	47.8	38.7	-	-	-	-	-	-
PLOP [16]	P	<u>75.7</u>	51.7	<u>70.1</u>	69.6	62.2	67.1	35.1	39.4	36.8	-	-	-	-	-	-
SDR [41]	P	75.4	52.6	69.9	70.5	<u>63.9</u>	<u>67.4</u>	-	-	-	-	-	-	-	-	-
RECALL [38]	P	67.7	<u>54.3</u>	65.6	66.0	58.8	63.7	-	-	-	-	-	-	-	-	-
CAM [61]	I	69.9	25.6	59.7	70.8	44.2	58.5	30.7	20.3	28.1	-	-	-	-	-	-
SEAM [55]	I	68.3	31.8	60.4	67.5	55.4	62.7	31.2	28.2	30.5	-	-	-	-	-	-
SS [3]	I	72.2	27.5	62.1	69.6	32.8	52.5	35.1	36.9	35.5	-	-	-	-	-	-
EPS [32]	I	69.4	34.5	62.1	69.0	57.0	64.3	34.9	38.4	35.8	-	-	-	-	-	-
WILSON [7]	I	74.2	41.7	67.2	70.4	57.1	65.0	39.8	41.0	40.6	38.7	22.4	32.5	66.8	46.5	58.1
Teddy	I	**77.6**	**51.4**	**72.0**	**71.2**	**59.4**	**66.5**	**40.6**	**41.8**	**41.5**	**50.3**	**32.0**	**43.1**	**68.9**	**51.7**	**61.7**
Joint	P	75.5	73.5	75.4	76.6	74.0	75.4	-	-	-	-	-	-	-	-	-

[4] to optimize memory usage for our experiments. The model is trained for 40 epochs with the batch size of 24 on two GPUs under distributed data parallel. The optimizer is SGD with a learning rate of 0.001, momentum 0.9 and weight decay 10^{-4}. Furthermore, the model is first warmed up for five epochs. After that, the model is trained with the learning rate decaying by using a polynomial schedule with a power of 0.9. We set $\alpha = 0.8$, $\beta = 0.5$ for all the settings. For SAM, we use ViT-H [15] as the backbone by default.

4.5   Results

15-5 VOC In this experimental setting, we introduce an additional 5 classes from the VOC data set after the initial stage. To ensure a fair comparison, we select the same classes as those in [7], which include: *plant, sheep, sofa, train, tv monitor.* As demonstrated in Table 1, Teddy attains a new state-of-the-art performance when compared to baselines trained with either pixel-level or image-level labels. It is noteworthy that Teddy even surpasses fully-supervised methods by 2.7%. In comparison to weak-supervised methods, Teddy outperforms WILSON by **7.1%** in general. Notably, for new classes Teddy outperforms WILSON by **23.3%**. Furthermore, Teddy demonstrates enhanced performance for old classes by **4.6%**, signifying that the phenomenon of catastrophic forgetting has been effectively mitigated.

10-10 VOC In this setting, 10 classes are introduced in the incremental step: *dining table, dog, horse, motorbike, person, plant, sheep, sofa, train, tv monitor.*

Table 2: 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	15-5 VOC			10-10 VOC			10-2 VOC		
	OB	w/o OB		1-15	16-20	All	1-10	11-20	All	1-10	11-20	All
1				74.1	41.5	67.0	70.4	57.0	65.0	38.8	22.3	32.5
2			✓	75.2	45.6	68.9	70.7	57.5	65.3	47.7	20.1	34.6
3		✓		76.8	46.0	70.1	71.0	58.6	66.0	50.2	23.7	38.5
4	✓			77.0	47.7	70.7	71.2	58.6	66.2	50.0	27.5	40.8
5		✓	✓	76.5	48.1	70.4	71.1	58.5	66.1	51.4	25.2	39.9
6	✓		✓	77.6	51.4	72.0	71.2	59.4	66.5	50.3	32.0	43.1

As shown in Table 1, Teddy achieves new SOTA performances and even surpasses some fully-supervised methods. Specifically, Teddy outperforms WILSON by 2.3% in general, and for new classes Teddy outperforms WILSON by **4.0%**.

COCO-to-VOC This setting presents a greater challenge. Initially, the base model is trained on 60 classes from the COCO data set, which do not overlap with the classes in VOC. Subsequently, in the second step, we train the additional 20 classes from the VOC data set using only image-level labels. The results are shown in Table 1, where we can observe that despite the heightened complexity and challenges intrinsic to this setup, Teddy achieves new SOTA performances. Notably, Teddy effectively preserves the previously acquired knowledge, epitomizing its prowess in counteracting catastrophic forgetting.

10-2 VOC and 10-5 VOC In these two settings, the issue of catastrophic forgetting exacerbates in these settings due to the increased number of tasks. Table 1 demonstrates that Teddy markedly surpasses WILSON. Notably, in the 10-2 VOC setting, Teddy exceeds WILSON's performance by a substantial margin of **32.6%**, and for the 10-5 VOC setting, Teddy outperforms WILSON by **6.2%**. The TME module effectively helps Teddy preserve the learnt knowledge, significantly reducing the impact of catastrophic forgetting, which leads to a substantial improvement.

4.6   Ablation Study

Tendency-driven Mutual Exclusivity   To assess the contribution of the TME in Teddy, we conduct comparisons between Teddy and that without TME. In this TME-excluded approach, the predictions for the new classes are directly refined, skipping the influence of TME. The comparative outcomes are presented in row 2 in Table 2. Evidently, comparing with rows 2 and 6, Teddy outperforms its degenerated version by **3.1, 1.2 and 8.5 mIoU**, respectively. Moreover, when comparing rows 1 and 3, the TME constraint can also boost a common WILSS model by **3.1, 1.0 and 6.0 mIoU**, respectively.

Table 3: Effectiveness of prediction fusion for new classes through optimization.

U	V	15-5 VOC			10-10 VOC			10-2 VOC		
		1-15	16-20	All	1-10	11-20	All	1-10	11-20	All
0.25	0.75	76.8	46.7	70.3	70.6	57.4	65.2	48.8	29.0	40.9
0.50	0.50	77.0	48.1	71.1	71.0	58.2	65.8	49.1	28.8	40.8
0.75	0.25	76.8	48.1	71.0	71.0	58.7	66.1	47.2	23.5	37.3
Optimization		77.6	51.4	72.0	71.2	59.4	66.5	50.3	32.0	43.1

Augmenting our empirical analysis, we proffer a series of qualitative insights to underscore the potency of TME. As shown in Fig. 3, it is obvious that the predictions with TME for the new classes evince greater precision than those without TME. The TME relationship not only encapsulates rich information of old classes but also effectively addresses the conflicting prediction issue and catastrophic forgetting, thereby resulting in superior performance.

Furthermore, the results of Teddy, when executed without prediction binarization for old classes (OB) in TME, are shown in row 5 of Table 2. Based on the results, it becomes apparent that incorporating prediction binarization into TME elevating the model's performance. By harnessing the prowess of SAM to execute binarization on predictions for old classes, we ensure that this information is effectively and accurately integrated into the TME constraint, which subsequently culminates in markedly improved outcomes.

Incorporation of SAM In Teddy, SAM takes part in the process of old class prediction binarization and new class prediction fusion. We further appraise the usefulness of SAM by proceeding without its incorporation. Comparing rows 3 and 4 in Table 2, we can find that the old class prediction binarization process can improve Teddy's performance by 0.6, 0.2 and 2.3 mIoU respectively. When comparing row 3 and 5, it is observed that the new class prediction fusion further boosts the performance by 0.3, 0.1 and 1.4 mIoU. This further demonstrates the effectiveness of SAM in these processes.

Prediction Fusion for New Classes Firstly, we evaluate the usefulness of prediction fusion (PF) for new classes. By omitting $R(S(\mathbf{x}^t, \mathbf{y}^t), \beta)$ and directly considering $\mathbf{P}$ under the TME constraint as the final mask for new classes, the performances are shown in row 4 in Table 2. An evident decrease across different settings underscores the efficacy of prediction fusion.

Furthermore, by configuring $\mathbf{U}$ and $\mathbf{V}$ as constants, we contrast the outcomes generated from various prediction fusion strategies. The results in Table 3 conspicuously indicate that coefficients derived from the optimization problem yield superior performance compared to other strategies. This observation underscores the value of our prediction fusion strategy in efficaciously merging predictions, thereby enhancing overall performance.

Qualitative Results In addition, we present qualitative results to further illustrate Teddy's superiority. As depicted in Fig. 4, it becomes evident that Teddy outperforms WILSON by producing more precise and well-defined segmentation results. Furthermore, Teddy exhibits accurate predictions for the old classes, serving as validation of its effectiveness in mitigating catastrophic forgetting.

Fig. 4: Qualitative results on 15-5 VOC setting for both new (sheep) and old classes. From left to right: image, WILSON [7], Teddy and the ground-truth.

5   Conclusion

In this paper, we introduce a novel framework called Teddy to address the intricate challenges posed by WILSS. Specifically, we propose a novel tendency-driven relationship of mutual exclusivity (TME) to regulate the interactions between the seed area and the predictions generated by the previous model, which not only resolves the conflicting prediction issue but also robustly combats catastrophic forgetting. Harnessing the strength of this relationship, we generate pseudo masks for new classes, allowing for concurrent execution with model parameter updating through a bi-level optimization strategy. Our extensive experiments, conducted on the Pascal VOC and MS COCO data sets, provide convincing evidence of the effectiveness and superiority of our framework, and validate that Teddy successfully addresses the issues of catastrophic forgetting and the conflicting prediction issue in WILSS.

Acknowledgements

References

1. Ahn, J., Cho, S., Kwak, S.: Weakly supervised learning of instance segmentation with inter-pixel relations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2209–2218 (2019)
2. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4981–4990 (2018)
3. Araslanov, N., Roth, S.: Single-stage semantic segmentation from image labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4253–4262 (2020)
4. Bulo, S.R., Porzi, L., Kontschieder, P.: In-place activated batchnorm for memory-optimized training of dnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5639–5647 (2018)
5. Caesar, H., Uijlings, J., Ferrari, V.: Coco-stuff: Thing and stuff classes in context. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1209–1218 (2018)
6. Castro, F.M., Marín-Jiménez, M.J., Guil, N., Schmid, C., Alahari, K.: End-to-end incremental learning. In: Proceedings of the European conference on computer vision (ECCV). pp. 233–248 (2018)
7. Cermelli, F., Fontanel, D., Tavera, A., Ciccone, M., Caputo, B.: Incremental learning in semantic segmentation from image labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4371–4381 (2022)
8. Cermelli, F., Mancini, M., Bulo, S.R., Ricci, E., Caputo, B.: Modeling the background for incremental learning in semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9233–9242 (2020)
9. Chaudhry, A., Dokania, P.K., Ajanthan, T., Torr, P.H.: Riemannian walk for incremental learning: Understanding forgetting and intransigence. In: Proceedings of the European conference on computer vision (ECCV). pp. 532–547 (2018)
10. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017)
11. Chen, T., Mai, Z., Li, R., Chao, W.l.: Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. arXiv preprint arXiv:2305.05803 (2023)
12. Dai, J., He, K., Sun, J.: Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1635–1643 (2015)
13. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
14. Dhar, P., Singh, R.V., Peng, K.C., Wu, Z., Chellappa, R.: Learning without memorizing. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5138–5146 (2019)

15. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
16. Douillard, A., Chen, Y., Dapogny, A., Cord, M.: Plop: Learning without forgetting for continual semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4040–4050 (2021)
17. Du, Y., Fu, Z., Liu, Q., Wang, Y.: Weakly supervised semantic segmentation by pixel-to-prototype contrast. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4320–4329 (2022)
18. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International journal of computer vision **88**, 303–338 (2010)
19. Fan, J., Zhang, Z., Song, C., Tan, T.: Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4283–4292 (2020)
20. French, R.M.: Catastrophic forgetting in connectionist networks. Trends in cognitive sciences **3**(4), 128–135 (1999)
21. Hariharan, B., Arbeláez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: 2011 international conference on computer vision. pp. 991–998. IEEE (2011)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
23. Huang, Z., Wang, X., Wang, J., Liu, W., Wang, J.: Weakly-supervised semantic segmentation network with deep seeded region growing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7014–7023 (2018)
24. Jeroslow, R.G.: The polynomial hierarchy and a simple model for competitive analysis. Mathematical programming **32**(2), 146–164 (1985)
25. Jiang, P.T., Hou, Q., Cao, Y., Cheng, M.M., Wei, Y., Xiong, H.K.: Integral object mining via online attention accumulation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2070–2079 (2019)
26. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 876–885 (2017)
27. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
28. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
29. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)
30. Klingner, M., Bär, A., Donn, P., Fingscheidt, T.: Class-incremental learning for semantic segmentation re-using neither old data nor old labels. In: 2020 IEEE 23rd international conference on intelligent transportation systems (ITSC). pp. 1–8. IEEE (2020)

31. Kolesnikov, A., Lampert, C.H.: Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14. pp. 695–711. Springer (2016)

32. Lee, S., Lee, M., Lee, J., Shim, H.: Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5495–5505 (2021)

33. Li, Z., Hoiem, D.: Learning without forgetting. IEEE transactions on pattern analysis and machine intelligence **40**(12), 2935–2947 (2017)

34. Lin, D., Dai, J., Jia, J., He, K., Sun, J.: Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3159–3167 (2016)

35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014)

36. Liu, Z., He, Y., Wang, W., Wang, W., Wang, Y., Chen, S., Zhang, Q., Yang, Y., Li, Q., Yu, J., Li, K., Chen, Z., Yang, X., Zhu, X., Wang, Y., Wang, L., Luo, P., Dai, J., Qiao, Y.: Interngpt: Solving vision-centric tasks by interacting with chatgpt beyond language. arXiv preprint arXiv:2305.05662 (2023)

37. Ma, J., Wang, B.: Segment anything in medical images. arXiv preprint arXiv:2304.12306 (2023)

38. Maracani, A., Michieli, U., Toldo, M., Zanuttigh, P.: Recall: Replay-based continual learning in semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7026–7035 (2021)

39. McCloskey, M., Cohen, N.J.: Catastrophic interference in connectionist networks: The sequential learning problem. In: Psychology of learning and motivation, vol. 24, pp. 109–165. Elsevier (1989)

40. Michieli, U., Zanuttigh, P.: Incremental learning techniques for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision workshops. pp. 0–0 (2019)

41. Michieli, U., Zanuttigh, P.: Continual semantic segmentation via repulsion-attraction of sparse and disentangled latent representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1114–1124 (2021)

42. Michieli, U., Zanuttigh, P.: Knowledge distillation for incremental learning in semantic segmentation. Computer Vision and Image Understanding **205**, 103167 (2021)

43. Oh, Y., Kim, B., Ham, B.: Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6913–6922 (2021)

44. Ostapenko, O., Puscas, M., Klein, T., Jahnichen, P., Nabi, M.: Learning to remember: A synaptic plasticity driven framework for continual learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11321–11329 (2019)

45. Papandreou, G., Chen, L.C., Murphy, K.P., Yuille, A.L.: Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In: Proceedings of the IEEE international conference on computer vision. pp. 1742–1750 (2015)

46. Rebuffi, S.A., Kolesnikov, A., Sperl, G., Lampert, C.H.: icarl: Incremental classifier and representation learning. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 2001–2010 (2017)
47. Roy, S., Volpi, R., Csurka, G., Larlus, D.: Rasp: Relation-aware semantic prior for weakly supervised incremental segmentation. arXiv preprint arXiv:2305.19879 (2023)
48. Shin, H., Lee, J.K., Kim, J., Kim, J.: Continual learning with deep generative replay. Advances in neural information processing systems **30** (2017)
49. Sun, G., Wang, W., Dai, J., Van Gool, L.: Mining cross-image semantics for weakly supervised semantic segmentation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. pp. 347–365. Springer (2020)
50. Tang, L., Xiao, H., Li, B.: Can sam segment anything? when sam meets camouflaged object detection. arXiv preprint arXiv:2304.04709 (2023)
51. Tang, M., Djelouah, A., Perazzi, F., Boykov, Y., Schroers, C.: Normalized cut loss for weakly-supervised cnn segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1818–1827 (2018)
52. Vernaza, P., Chandraker, M.: Learning random-walk label propagation for weakly-supervised semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7158–7166 (2017)
53. Wang, D., Zhang, J., Du, B., Tao, D., Zhang, L.: Scaling-up remote sensing segmentation dataset with segment anything model. arXiv preprint arXiv:2305.02034 (2023)
54. Wang, T., Zhang, J., Fei, J., Ge, Y., Zheng, H., Tang, Y., Li, Z., Gao, M., Zhao, S., Shan, Y., et al.: Caption anything: Interactive image description with diverse multimodal controls. arXiv preprint arXiv:2305.02677 (2023)
55. Wang, Y., Zhang, J., Kan, M., Shan, S., Chen, X.: Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12275–12284 (2020)
56. Wu, C., Herranz, L., Liu, X., Van De Weijer, J., Raducanu, B., et al.: Memory replay gans: Learning to generate new categories without forgetting. Advances in Neural Information Processing Systems **31** (2018)
57. Wu, Z., Shen, C., Van Den Hengel, A.: Wider or deeper: Revisiting the resnet model for visual recognition. Pattern Recognition **90**, 119–133 (2019)
58. Yu, C., Zhou, Q., Li, J., Yuan, J., Wang, Z., Wang, F.: Foundation model drives weakly incremental learning for semantic segmentation. arXiv preprint arXiv:2302.14250 (2023)
59. Yu, Y., Yang, X., Li, Q., Zhou, Y., Zhang, G., Yan, J., Da, F.: H2rbox-v2: Boosting hbox-supervised oriented object detection via symmetric learning. arXiv preprint arXiv:2304.04403 (2023)
60. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: International conference on machine learning. pp. 3987–3995. PMLR (2017)
61. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
62. Zhu, F., Zhang, X.Y., Wang, C., Yin, F., Liu, C.L.: Prototype augmentation and self-supervision for incremental learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5871–5880 (2021)