Improving the Generalization of Segmentation Foundation Model under Distribution Shift via Weakly Supervised Adaptation

Haojie Zhang[1*]   Yongyi Su[1*]   Xun Xu[2†]   Kui Jia[3]

[1]South China University of Technology

[2]Institute for Infocomm Research, A*STAR

[3]School of Data Science, The Chinese University of Hong Kong, Shenzhen

Abstract

The success of large language models has inspired the computer vision community to explore image segmentation foundation model that is able to zero/few-shot generalize through prompt engineering. Segment-Anything (SAM), among others, is the state-of-the-art image segmentation foundation model demonstrating strong zero/few-shot generalization. Despite the success, recent studies reveal the weakness of SAM under strong distribution shift. In particular, SAM performs awkwardly on corrupted natural images, camouflaged images, medical images, etc. Motivated by the observations, we aim to develop a self-training based strategy to adapt SAM to target distribution. Given the unique challenges of large source dataset, high computation cost and incorrect pseudo label, we propose a weakly supervised self-training architecture with anchor regularization and low-rank finetuning to improve the robustness and computation efficiency of adaptation. We validate the effectiveness on 5 types of downstream segmentation tasks including natural clean/corrupted images, medical images, camouflaged images and robotic images. Our proposed method is task-agnostic in nature and outperforms pre-trained SAM and state-of-the-art domain adaptation methods on almost all downstream tasks with the same testing prompt inputs.

1. Introduction

The success of large language models is partially attributed to the strong capability of zero-shot generalization. This has called for numerous attention into developing foundation model for computer vision tasks where one of the key concern also lies in the generalization ability [26, 44, 48]. In particular, as a promptable image segmentation foundation model with strong zero-shot generalization, the Segment-Anything model (SAM) [27] was developed by training on billions of annotated masks. Despite the overwhelming size

---

* Equal contribution.

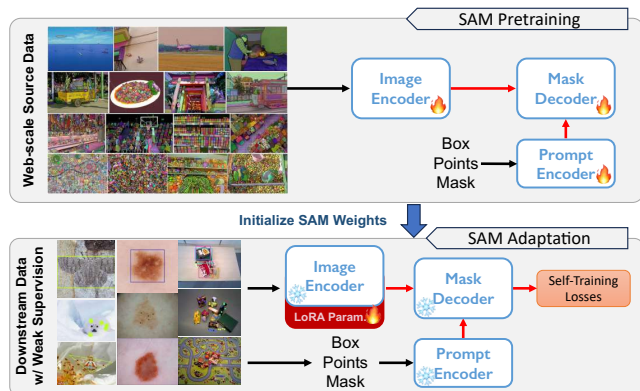† Correspondence to Xun Xu: alex.xun.xu@gmail.com.

Figure 1. Segment Anything Model was pre-trained on a large-scale dataset but exhibits awkward performance on diverse downstream segmentation tasks. We adapt SAM through weak supervision to enhance its generalization capabilities.

of dataset used for training, SAM was found to behave awkwardly on certain out-of-distribution downstream tasks, including camouflaged segmentation, medical segmentation, adversarial attacks, visual corruptions, etc. [5, 21]. This is likely caused by the distribution shift between the training data, i.e. the SA-1B dataset [27], and the challenging testing datasets. These observations motivate us to explore tangible solutions to make SAM more robust against real-world and more diverse downstream tasks.

Traditional paradigm towards improving model robustness and generalization often involves costly re-training. Methods are frequently customized to combat specific domain shifts. For example, domain randomization was developed to improve generalization across real testing domains [6], while adversarial training is commonly used to enhance robustness against attacks [58]. Applying these techniques to large foundation model is nonetheless impractical due to the enormous computing resources required for training the model on the web-scale training data. Therefore, instead of re-training the model, we opt for a more computation friendly paradigm by adapting or finetuning a

pre-trained foundation model to downstream dataset.

We specify three major challenges when adapting pre-trained foundation model to a new data distribution. First, traditional unsupervised domain adaptation [15] paradigm requires access to both source and target datasets which may not be feasible due to privacy issue and computation cost [34]. Second, updating all model weights for adaptation is often superior in performance, however, is constrained by the prohibitive memory cost due to the large model size [20]. Finally, unsupervised adaptation could be still very challenging due to the absence of label information on the target domain [60, 72, 75].

To tackle the above challenges, we propose a self-training based adaptation approach to synergize the exploitation of weak supervision and unlabeled data on target domain. Specifically, we first alleviate the dependence on source domain data by adopting a self-training based source-free domain adaptation strategy. The self-training approach first make segmentation predictions, a.k.a. pseudo labels, on the target domain data. The pseudo labels are used for supervising the update of segmentation model. As self-training is fragile due to incorrect pseudo labels, a.k.a. confirmation bias [2], we introduce a frozen source model as the anchor network to regularize the update of segmentation model. To further mitigate the high computation cost of updating the full model weights, we apply a low-rank weight decomposition to encoder weights and the back-propagation flows through the low-rank shortcut only. Finally, to further improve the effectiveness of source-free domain adaptation, weak supervisions, e.g. sparse point-wise annotation, on the target domain are introduced to provide stronger cues for domain adaptation. We reveal that the weak supervisions are naturally compatible with the prompt encoder within SAM. With the weak supervisions as prompt, we are endowed with more localized and less ambiguous pseudo predictions for self-training. The adapted model has demonstrated much stronger generalization capability on multiple downstream tasks.

We summarize the contributions of this work as follows.

- We are motivated by the generalization issue of segment-anything (SAM) model to diverse downstream segmentation tasks and propose a task-agnostic solution to adapt SAM through self-training with no access to source dataset.
- We exploit weak supervisions, including bounding box, point-wise annotation and coarse segmentation masks, to improve the effectiveness of adaptation. The weak supervision are fully compatible with the prompt encoder of SAM.
- Extensive experiments on 5 types of downstream instance segmentation tasks demonstrate the effectiveness of the proposed weakly supervised adaptation approach.

2. Related Work

Image Segmentation Foundation Model: The success of deep learning is attributed to the increasing size and neural networks and huge amount of training data. Recent successful computer vision models, e.g. image segmentation and object detection, follow a practice of finetuning from encoder network pre-trained on large image dataset, e.g. ImageNet [10]. Despite achieving impressive results on standard benchmark datasets, e.g. Psacal VOC [13] and COCO [37], the zero-shot and few-shot generalization ability is often limited. Inspired by the revolutionizing language/vision-language foundation model pre-trained on web-scale datasets [4, 48, 74], a new opportunity awaits for developing a generalizable vision foundation model. This motivates the emergence of vision foundation model pre-trained on huge dataset. SAM [27] and DINO v2 [44], to name a few. Among these foundation models, SAM stands out for its ability to enable zero-shot segmentation using prompt engineering. Therefore, we pay particular attention to SAM in this work. Since the inception of SAM, numerous attempts are made to validate the robustness of SAM under more challenging scenarios, such as medical images [42] [71], camouflaged objects [59], glass (transparent objects and mirrors) [18] and pose estimation [36]. Despite these early attempts to discover the weakness, there are very few matured solutions [5] to improve the generalization of SAM to challenging downstream tasks. In this work, we aim to fill this gap by proposing principled solutions to enhance the robustness of SAM against downstream tasks subject to significant distribution shift.

Source-Free Domain Adaptation: Unsupervised Domain Adaptation (UDA)[15, 24, 41, 61, 73] has emerged as a means to address the domain shift between source training data and target testing data. This enables models to be trained on cost-effective and easily annotated data, facilitating the transfer to more valuable but challenging-to-label datasets. However, UDA methods face limitations when source domain data is inaccessible due to privacy concerns or storage constraints. In response to this, Source-Free Domain Adaptation (SFDA)[28, 33, 34, 39, 65, 68] has been proposed to reduce the reliance on source data in domain adaptation (DA) methods. SHOT [34] represents an early effort to alleviate domain shift without access to source data, achieved through self-training on unlabeled target data. Numerous prior works [1, 7, 11, 26, 35, 47, 50] have explored the effectiveness of a teacher-student like self-training architecture in domain shift mitigation. The similar self-training framework [22] simultaneously trains on multiple sets of augmented data, utilizing teacher-student cycle consistency for mutual learning of the two networks. Other methods generate source-like samples [30, 40, 46], construct intermediate domains [65, 69], or perturb target domains [25, 32, 66, 70] to enhance model generalization.
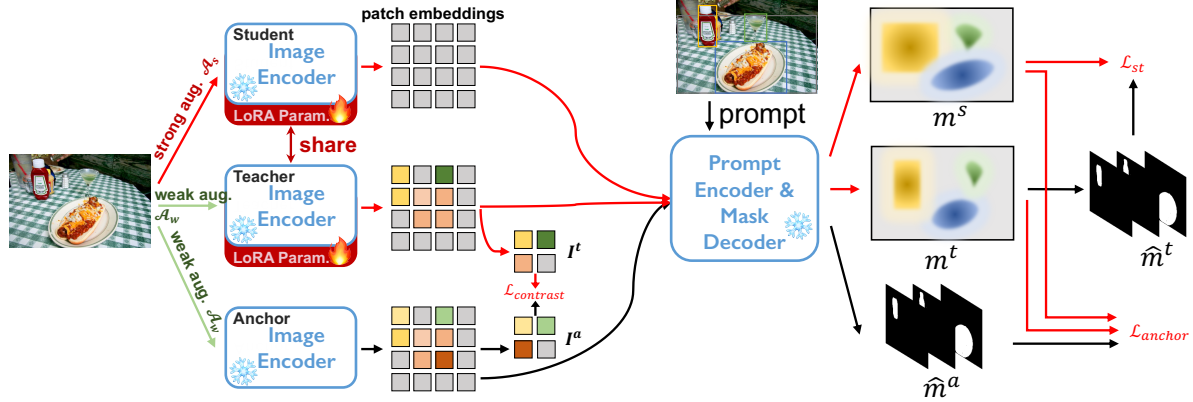
Figure 2. The proposed self-training architecture with anchor network regularization and contrastive loss regularization. Red arrows indicates the backpropagation flow.

Recent approaches [46, 64] leverage contrast loss between target samples and source domain prototypes. Recent attempts to source-free domain adaptation for image segmentation aims to minimize feature discrepancy between source and target domains [3, 29, 40]. In this work, we aim to adapt SAM to downstream tasks without accessing to the source domain data to avoid the high computation overhead and potential privacy issues for foundation models to be released in the future.

Weakly Supervised Domain Adaptive Segmentation: Unsupervised domain adaptation often follows a self-training paradigm [34], which is limited by the quality of pseudo labels. Instead of relying on the pseudo predictions, weakly supervised domain adaptive semantic segmentation (WDASS) exploits limited weak supervisions in the form of bounding box, points or coarse segmentation masks [8, 9, 45, 63]. Among these works, [45] employs both image and point labels, suggesting an adversarial alignment of features at the pixel level to address WDASS. In contrast, [63] relies on bounding boxes as weak labels and employs adversarial learning to achieve domain-invariant features. Additionally, [9] utilizes self-training and a boundary loss to enhance performance in WDASS with coarse labels. More recently, [8] utilizes weak label for aligning the source-target features and outperforms previous methods to achieve competitive performance compared to supervised learning. We observe that the common weak supervisions are inherently compatible with the SAM model as prompt input. Therefore, we propose to adapt SAM with weak supervisions in a seamless way.

3. Methodology

In this section, we first provide an overview of segment-anything model. Then, we introduce the self-training based adaptation framework. We elaborate how weak supervision could help establish correspondence between pseudo predictions for effective self-training. Lastly, we introduce the low memory footprint way of updating model weights.

3.1. Overview of Segment Anything Model

The segment-anything model consists of three major components, the encoder network $z = f(x; \Theta)$, prompt encoder $e = g(p; \Omega)$ and mask decoder $h(z, e; \Phi)$. The image encoder is pre-trained using masked autoencoder (MAE) [19]. The whole SAM model is further fine-tuned on the web-scale labeled training set, SA-1B [27], with 1.1B labeled masks. A combination of focal loss and dice loss are used for training SAM. During inference, the testing image x is first encoded by the image encoder $z = f(x; \Theta)$. Given encoded prompts e, a light-weight mask decoder makes three levels of segmentation predictions. In this work, we are motivated by the challenge of deploying SAM on many downstream tasks and propose to adapt SAM to downstream segmentation tasks with weak supervision without requiring access to source domain training data.

3.2. Source-Free Adaptation as Self-Training

Provided an unlabeled target domain dataset $\mathcal{D}_T = \{x_i\}$ and the pre-trained encoder network $f(x; \Theta)$, we adopt a teacher-student architecture for self-training. As illustrated in Fig. 2, we maintain three encoder networks, namely the anchor network $f(x; \Theta^a)$, the student network $f(x; \Theta^s)$ and the teacher network $f(x; \Theta^t)$, where the student and teacher networks share the weights $\Theta^s = \Theta^t$. For each sample x_i, we apply one random weak data augmentation $\mathcal{A}_w(x_i)$ as the input to anchor and teacher networks, and one random strong data augmentation $\mathcal{A}_s(x_i)$ as the input to student network. Through the three encoder networks, three feature maps are obtained respectively, as $\mathcal{F}^a = f(\mathcal{A}_w(x_i), \Theta^a)$, $\mathcal{F}^s = f(\mathcal{A}_s(x_i), \Theta^s)$ and $\mathcal{F}^t = f(\mathcal{A}_w(x_i), \Theta^t)$, where $\mathcal{F} \in \mathbb{R}^{D \times H \times W}$ and D refers to feature dimension. In the decoder network, given a number N_p of prompts, e.g. bounding boxes, points or coarse segmentation masks, a set of foreground masks for the instance segmentation results would be deduced, $\mathcal{Y}^a = \{y_j^a\}_{j=1 \cdots N_p}$, $\mathcal{Y}^s = \{y_j^s\}_{j=1 \cdots N_p}$, $\mathcal{Y}^t = \{y_j^t\}_{j=1 \cdots N_p}$, where each mask is then normalized between 0 and 1 by sigmoid function respectively, i.e. $m_j = Sigmoid(y_j), m_j \in [0, 1]^{H \times W}$. We
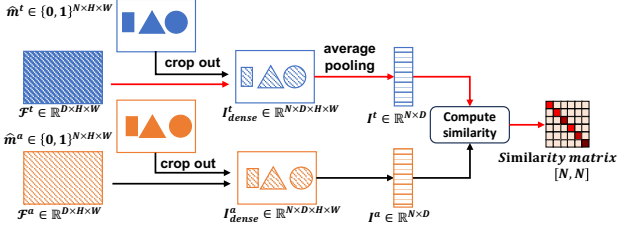
Figure 3. Illustration of contrastive loss between two views.

further denote a binarized segmentation mask with a threshold of 0.5 as $\hat{m}_j = \mathbb{1}(m_j > 0.5), \hat{m}_j \in \{0,1\}^{H \times W}$. Based on these notations, we elaborate three sets of adaptation objectives for self-training purpose.

Teacher-Student Self-Training Loss: We first introduce the Self-Training loss to update student/teacher network. Self-Training is widely used in semi-supervised learning [52] and recently is demonstrated to be very effective for source-free domain adaptation [35]. In this work, we continue the idea of self-training in classification task and perform focal loss [38] and dice loss [43] for the supervision of student's outputs with binarized teacher's predictions as follows, where γ controls the focus on hard training pixels and ϵ is a small value to avoid dividing by zero.

$$\mathcal{L}_{st}^{focal} = -\sum_{j=1\cdots N_p} \frac{1}{HW} \sum_{h,w} \mathbb{1}(\hat{m}_{jhw}^t = 1) \cdot (1 - m_{jhw}^s)^\gamma log(m_{jhw}^s) \quad (1)$$
$$+ \mathbb{1}(\hat{m}_{jhw}^t = 0) \cdot m_{jhw}^s{}^\gamma log(1 - m_{jhw}^s),$$

$$\mathcal{L}_{st}^{dice} = \sum_{j=1\cdots N_p} 1 - \frac{2\sum_{h,w} m_{jhw}^s \cdot \hat{m}_{jhw}^t + \epsilon}{\sum_{h,w} m_{jhw}^s + \sum_{h,w} \hat{m}_{jhw}^t + \epsilon}, \quad (2)$$

Anchored Loss for Robust Regularization: Training the networks with self-training loss alone is vulnerable to the accumulation of incorrect pseudo labels predicted by teacher network, a.k.a. confirmation bias [2]. Observations are made that the performance will drop after long iterations of self-training alone [56]. Existing source-free domain adaptation approaches often adopt additional constraints to prevent the negative impact of self-training, e.g. uniform distribution over predictions [35]. Without making such explicit assumptions, we propose to introduce the regularization by an anchor loss which minimizes the dice loss between an anchor network, which has frozen source model weights, and student and teacher networks respectively in Eq. 3. The frozen anchor network serves as the knowledge inherited from source domain and too much deviation between the source model and self-training updated model is discouraged to prevent the model from collapsing.

$$\mathcal{L}_{anchor} = \lambda_{stu}^{dice} \mathcal{L}^{dice}(m^s, \hat{m}^a) + \lambda_{tea}^{dice} \mathcal{L}^{dice}(m^t, \hat{m}^a). \quad (3)$$

Contrastive Loss for Encoder Output Regularization: The above two training objectives are performed in the output space of the decoder. As we empirically revealed in the

experiment section, updating encoder network is the most effective way to adapt SAM, it is necessary to apply regularization directly to the feature map as output from the encoder network. Specifically, as illustrated in Fig. 3, we first crop out each instance's feature I_j from the feature map based on its prediction mask in anchor and teacher branches.

$$I_j = \frac{\sum_{h,w} \overline{\mathcal{F}}_{hw} \cdot \hat{m}_{jhw}}{\sum_{h,w} \hat{m}_{jhw}}, \quad (4)$$

where $\overline{\mathcal{F}}$ is the L2 normalized feature map, i.e. $\overline{\mathcal{F}}_{hw} = \frac{\mathcal{F}_{hw}}{|\mathcal{F}_{hw}|}$. We further define the instance features, which are cropped out by the masks used the same prompt in two branches, as positive pairs, and the instances used different prompts as negative pairs to construct the contrast loss. The final contrastive loss is given as below, where τ is the temperature scaling.

$$\mathcal{L}_{contrast} = -log \frac{\sum_j exp(I_j^a \cdot I_j^t/\tau)}{\sum_j \sum_{j' \neq j} exp(I_j^a \cdot I_{j'}^t/\tau)} \quad (5)$$

Total Loss: We combine the above three loss functions as the final source-free domain adaptation loss.

$$\mathcal{L} = \lambda_{st}^{focal} \mathcal{L}_{st}^{focal} + \mathcal{L}_{st}^{dice} + \mathcal{L}_{anchor} + \mathcal{L}_{contrast}. \quad (6)$$

3.3. Prompt Generation for Self-Training

The SAM segmentation requires prompt inputs to disambiguate the granularity of segmentation. The prompt engineering can be implemented in both fully automated way and by human interaction.

Fully Automated Prompt for Self-Training: We first densely sample grid points with a spatial distance of 16 pixels as point prompt input. Initial stage segmentation masks $\{\hat{m}_j\}_{j=1\cdots N_{init}}$ are generated by the anchor network with the grid point prompts. We follow the fully automated segmentation introduced by SAM [27] to prune out masks with low IoU score and low stability score, followed by NMS suppression. To enable self-training and regularization between different branches, we further generate a fixed set of prompt $\{e_j\}_{j=1\cdots N_p}$ from $\{\hat{m}_j\}_{j=1\cdots N_{init}}$ as input to all three branches. As such, the segmentation outputs $\mathcal{Y}^a$, $\mathcal{Y}^s$ and $\mathcal{Y}^t$ are of the same length with exact one-to-one correspondence.

Weak Supervision as Prompt: Despite automatic segmentation can be enabled by sampling a grid of point prompts over the image and filtering out low quality and duplicated masks, the quality of segmentation is relatively poor and may contain many false positive predictions, making self-training less effective. Therefore, following prior weakly supervised domain adaptation works [9], we propose to use three types of weak supervisions, including bounding box, sparse point annotation and coarse segmentation mask. Fortunately, in the context of SAM, these weak supervisions perfect match with the prompt inputs which allow seamless integration of weak supervision for adapting SAM.

3.4. Low-Rank Weights Update

The enormous size of backbone network prohibits updating all model weights with large batchsize. However, many existing studies suggest the update of encoder network weights is an effective way of adapting pre-trained models [35, 54]. To enable updating the backbone network with larger batchsize, we opt for a computation friendly low-rank updating approach [20]. For each weight in the encoder network $\theta \in \mathbb{R}^{d_i \times d_o}$, we use a low-rank approximation $\omega = AB$ where $A \in \mathbb{R}^{d_i \times r}$ and $A \in \mathbb{R}^{r \times d_o}$ with r indicating the rank. We can achieve a compression rate of $r(d_i + d_o)/d_i \cdot d_o$. Only A and B are updated via backpropagation during adaptation to reduce memory footprint. At inference stage, the weight is reconstructed by combining the low rank reconstruction and original weight, $\theta = \theta + AB$.

4. Experiment

In this section, we first introduce the experiment settings and downstream tasks for adapting SAM model. We provide detailed comparisons to state-of-the-art methods and qualitative results. Finally, we analyze the effectiveness of individual components and specific designs for the network.

4.1. Datasets

The source domain training set, SA-1B, was mainly constructed by collecting from natural environments. In this work, we identified five types of downstream segmentation tasks, some of which feature a drastic distribution shift from SA-1B, for evaluation, covering natural clean images, natural corrupted images, medical images, camouflaged images and robotic images. We present the datasets evaluated for each type of downstream task in Tab. 1.

Datasplit: We divide each downstream dataset into non-overlapping training and testing sets following the ratio established by [8]. The adaptation step is implemented on the training set. After the model is adapted, we evaluate the model on the held-out testing set. For adaptation to the Camouflaged Objects, following the training protocol in [5], we use the dataset COD10K (the training set of camouflaged images) for training, and use the test set of CAMO, COD10K and the entire CHAMELEON dataset for performance evaluation.

4.2. Experiment Details

Segment-Anything Model: We adopt ViT-B [12] as the encoder network due to memory constraint. The standard prompt encoder and mask decoder are adopted.

Prompt Generation: We first provide details for generating prompt at both training stages. All training prompts are calculated from ground-truth segmentation mask, simulating the human interactions as weak supervision. Specifically, we extract the minimal bounding box that covers the whole instance segmentation mask as box prompt. The point prompt is created as randomly selecting 5 positive points within ground-truth instance segmentation mask and 5 negative points outside ground-truth mask. The coarse segmentation mask is simulated by fitting a polygon to the ground-truth mask. We choose the number of vertices as $P/20$ with P indicating the perimeter of mask. The minimal number of vertices is 3. We use the same way to generate testing prompt on the testing data. This practice guarantees fair evaluation of SAM model which requires prompt input for segmentation.

Evaluation Metrics: We report the mIoU as evaluation metrics. For each input prompt, the IoU is calculated between the ground-truth segmentation mask and predicted mask. The mIoU averages over the IoU of all instances.

Competing Methods: We evaluate multiple source-free domain adaptation approaches and one state-of-the-art weakly supervised domain adaptive segmentation approach. In specific, direct testing the pre-trained model (**Direct**) with fixed prompt inputs is prone to the distribution shift and may not perform well on target datasets with significant shift. **TENT** [62] is a vanilla test-time adaptation method which optimizes an entropy loss for adapting to target domain. **SHOT** [35] employs pseudo label and applies uniform distribution assumption for source-free domain adaptation. **Soft Teacher** [67] was originally developed for semi-supervised image segmentation. We adapt Soft Teacher for domain adaptation by keeping the self-training component. **TRIBE** [55] proposed a strong baseline for generic test-time adaptation on continual and class imbalanced domain shifts. We adapt TRIBE for domain adaptive segmentation by replacing the training losses. **DePT** [16] inserts visual prompts into a visual Transformer and adjusts these source-initialized prompts solely during the adaptation process without accessing the source data. **WDASS** [8] developed a weakly supervised domain adaptive segmentation method. We also evaluate the upperbound of adapting SAM by fine-tuning with ground-truth segmentation masks, which is dubbed as **Supervised**. Finally, we evaluate our weakly supervised domain adaptation method as **Ours**. All the competing methods use the same backbone, i.e. the ViT-B from SAM.

Hyperparameters: We finetune LoRA module of the ViT-B image encoder by Adam optimizer for all experiments. We set the batchsize to 4, distributed over four RTX3090 GPUs, and the learning rate to 0.0001 with a weight decay of 0.0001. For Self-Training loss, we set the hyperparameters γ and ϵ of $\mathcal{L}_{st}^{focal}$ and $\mathcal{L}_{st}^{dice}$ to 2 and 1 respectively. λ_{st}^{focal} is set to 20, which is the coefficients of $\mathcal{L}_{st}^{focal}$. For Anchor loss, the coefficients of two dice losses are denoted as λ_{stu}^{dice} and λ_{tea}^{dice}, both of which equal 0.5. For Contrast loss, we set the temperature τ to 0.3. We have set the low rank of the LoRA module for the image encoder to 4. We apply strong and weak data augmentations for self-training and choices for augmentation follows [53, 67].

Table 1. Dataset and datasplit used in this work.

	Natural Images		Corrupted Images	Medical Images		Camouflaged Objects			Robotic Images	
	COCO [37]	Pascal VOC [13]	COCO-C	kvasir-SEG [23]	ISIC [17]	CHAMELEON [51]	CAMO [31]	COD10K [14]	OCID [57]	OSD [49]
# Training	4246	2497	4246	834	900		4040		1972	56
# Testing	706	416	706	166	379	74	250	2026	328	55

Table 2. Adaptation results on COCO-C dataset using bounding box prompt.

Method	Brit	Contr	Defoc	Elast	Fog	Frost	Gauss	Glass	Impul	Jpeg	Motn	Pixel	Shot	Snow	Zoom	Avg
Direct	72.83	57.34	64.47	69.36	72.39	70.50	67.20	64.43	67.65	68.23	62.72	68.60	67.44	69.02	58.80	66.73
TENT	76.02	61.51	**67.48**	70.88	74.89	73.88	69.01	67.10	69.28	70.25	65.45	70.81	69.96	72.37	62.59	69.43
SHOT	73.84	59.09	65.91	69.57	73.98	72.51	68.30	66.09	68.61	69.45	64.56	70.48	68.77	71.03	60.17	68.16
Soft Teacher	73.90	**62.12**	65.41	71.32	72.16	73.27	68.84	67.49	68.73	70.18	**66.88**	69.79	70.08	73.33	64.88	69.23
TRIBE	76.40	60.86	66.19	72.72	75.08	75.14	70.34	66.66	70.83	72.42	65.94	70.24	**70.66**	74.22	64.56	70.15
DePT	69.15	57.26	59.08	66.80	58.73	66.75	66.75	62.74	65.65	66.39	61.66	66.65	67.57	66.62	58.21	64.42
WDASS	76.21	60.57	67.07	72.34	75.97	74.63	69.84	67.88	69.92	71.36	66.25	**71.99**	70.32	72.25	63.61	70.01
OURS	**78.50**	61.05	66.99	**73.93**	**77.09**	**76.10**	72.02	68.21	71.29	72.77	66.33	70.90	70.28	**75.07**	65.33	**71.05**
Supervised	78.86	74.81	72.04	74.32	78.01	77.14	73.43	72.12	74.08	75.30	71.39	75.15	74.25	76.34	68.04	74.35

4.3. Quantitative Evaluations

In this section, we report the quantitative evaluations on the 5 types of target datasets. For each dataset, we also report the adaptation results with weak supervision in the form of bounding box (box), sparse points (point) and coarse segmentation mask (poly).

Adaptation to Corrupted Images: Visual corruptions often occur due to sensor fault, bad weather conditions, etc. As seen from Tab. 2, without any adaptation, **Direct** testing is significantly worse than than the upperbound. With weakly supervisions, all methods could improve over direct testing, in particular, our proposed method outperforms all competing adaptation methods on 10 out of 15 types of corruptions.

Adaptation to Natural Images: We then present the results of adapting SAM to natural images in Tab. 3. For each type of weak supervision, we use the same type of prompt on the testing set. Despite the distribution gap between SA-1B and the target natural images, we observe significant performance gap between **Supervised** upperbound and **Direct** baseline, e.g. with box level supervision, there is $5 - 10\%$ gap in IoU. When weak supervision is provided, both state-of-the-art generic source-free domain adaptation methods and weakly supervised domain adaptive segmentation method improve the generalization on all three types of weak supervisions. Finally, our proposed weakly supervised method achieves a remarkable improvement over all competing methods. The results with box weak supervision is even approaching the fully supervised upperbound.

Adaptation to Medical Images: Segmentation for medical images is a major application of foundation models. Our empirical observations in Tab. 4 on two medical segmentation datasets suggest that direct applying pre-trained SAM is suboptimal. With weakly supervised adaptation, the segmentation accuracy is greatly improved. We observe the improvement to be particularly significant with point and box which are relatively easy to obtain due to the low labeling cost.

Adaptation to Camouflaged Objects: We further eval-

Table 3. Adaptation results on natural clean image datasets.

Method	COCO 2017			Pascal VOC		
	box	point	poly	box	point	poly
Direct	74.29	55.06	65.64	69.21	69.21	60.79
TENT	78.21	52.99	71.51	80.24	74.97	65.03
SHOT	75.18	58.46	69.26	79.80	74.26	63.38
Soft Teacher	75.94	43.36	68.27	72.93	56.09	62.20
TRIBE	77.56	49.56	70.99	78.87	69.21	65.39
DePT	71.00	37.35	63.27	74.09	42.99	59.94
WDASS	77.29	60.55	70.19	80.12	**76.15**	**66.98**
OURS	**80.12**	**62.09**	**72.33**	**80.27**	74.15	66.72
Supervised	81.50	69.77	73.39	81.23	76.98	71.32

Table 4. Adaptation results on medical image segmentation datasets.

Method	kvasir-SEG			ISIC		
	box	point	poly	box	point	poly
Direct	81.59	62.30	54.03	66.74	53.42	62.82
TENT	82.47	61.84	62.97	71.76	53.46	67.12
SHOT	82.30	63.76	61.34	71.99	55.99	66.86
Soft Teacher	84.12	73.53	58.15	75.74	54.95	72.29
TRIBE	85.05	73.03	64.61	72.61	50.36	67.99
DePT	81.91	52.06	61.55	78.43	46.79	72.75
WDASS	84.01	63.78	64.78	74.23	55.63	67.84
OURS	**85.47**	**75.23**	**67.40**	**80.01**	**62.12**	**75.36**
Supervised	85.89	77.54	81.64	81.62	79.81	80.26

uate on adapting SAM to camouflaged object detection. We evaluate on three camouflaged object detection datasets with results reported in Tab. 5. Preliminary studies revealed that SAM is particularly vulnerable to camouflaged objects [59] due to the low contrast between background and foreground. We make similar observations here that even with the help of prompt inputs at testing stage, **Direct** inference suffers a lot, achieving less than 66% mIoU. When weakly labeled data are used for adaptation, we observe very significant improvements, with point prompt only we improved by a staggering 35% in mIoU on the CHAMELEON dataset. Our proposed method is also consistently better than all competing methods on all types of weak supervisions.

Adaptation to Robotic Images: Finally, we evaluate adapting SAM to image collected from robotic environ-

Table 5. Adaptation results on camouflaged object datasets.

Method	CHAMELEON			CAMO			COD10K		
	box	point	poly	box	point	poly	box	point	poly
Direct	51.32	39.37	45.78	62.72	57.43	50.85	66.32	63.61	40.04
TENT	65.48	54.53	53.06	71.24	59.59	60.29	69.36	61.94	43.36
SHOT	68.60	62.47	54.36	71.61	62.78	58.72	69.09	65.25	42.38
Soft Teacher	65.92	44.17	46.72	62.30	48.64	51.26	66.32	50.04	32.27
TRIBE	71.00	52.80	54.99	66.00	61.97	60.54	67.84	63.62	42.75
DePT	54.48	33.46	42.47	55.44	33.07	48.63	59.32	34.06	35.51
WDASS	71.91	62.40	56.80	71.25	63.39	62.29	71.42	65.61	43.93
OURS	**75.94**	**74.00**	**66.83**	**73.42**	**65.55**	**62.90**	**71.93**	**70.55**	**45.87**
Supervised	78.05	85.86	68.38	79.17	77.01	67.12	78.06	78.44	64.90

ments where segmenting out the spatial extent of individual objects is essential for interactions. As seen from the results in Tab. 6, adapting SAM to robotic images with weak supervision further increase the segmentation accuracy. The improvement is more noticeable with point and polygon weak supervisions.

Table 6. Adaptation results on robotic image datasets.

Method	OCID			OSD		
	box	point	poly	box	point	poly
Direct	86.35	71.41	72.81	87.62	78.86	80.77
TENT	87.77	66.61	**77.53**	88.10	80.53	87.85
SHOT	88.06	74.39	76.25	88.09	80.52	87.86
Soft Teacher	84.98	68.46	73.75	90.41	80.49	87.00
TRIBE	86.77	67.86	76.50	90.42	**80.54**	87.84
DePT	82.00	56.52	70.92	81.84	69.06	82.50
WDASS	87.68	77.13	76.70	88.07	80.52	88.19
OURS	**88.09**	**80.14**	77.41	**92.11**	80.51	**89.72**
Supervised	91.24	89.22	79.23	92.14	82.41	90.83

4.4. Qualitative Evaluations

We present the qualitative comparisons of segmentation results for selected examples in Fig. 4. We make the following observations from the results. Evaluating pre-trained SAM (SAM) on COCO with both box and point prompt exhibits very fine-grained segmentation results. However, the model is overly sensitive to the high contrast regions, e.g. the hair for the girl and the toppings of pizza are missing from the segmentation mask. With the proposed weakly supervised adaptation approach, our method is able to generate more smoothed segmentation mask and the mask better reflects the semantic boundary. Similar observations are made from the medical image segmentation dataset. Without adapting the SAM model, segmentation results are either too conservative with low recall or expanding beyond the semantic boundary. With our weakly supervised adaptation, SAM is able to produce high fidelity results. We further present results on camouflaged segmentation task (CAMO). Without adaptation, SAM fails to segment out the spider's belly and confuses the snow background with the harp seal in the foreground. Our weakly adapted SAM successfully segments out the spider's belly and produces high quality mask for the harp seal. Finally, two examples from OCID dataset suggest that the pre-trained SAM tends to over-segment, e.g. missing the open whole for the tissue box, or confusing adjacent objects, e.g. merging the pen with the folder label. With weakly supervised adaptation SAM picks up the

Table 7. Ablation studies of the proposed weakly supervised adaptation method on COCO dataset.

Weak Sup.	Self-Train.	Anchor	Cont. Loss	box	point	poly
	original SAM			74.29	56.36	65.42
✓	✓			58.88	32.51	55.03
✓	✓	✓		79.65	57.25	70.49
✓	✓		✓	62.95	22.87	56.91
✓	✓	✓	✓	80.12	62.09	72.33
	✓	✓	✓	76.18	47.63	70.44

Table 8. Adaptation results on cross-prompt testing scenarios.

Tr. Weak. Sup.	box			point			polygon		
Te. Prompt	box	point	poly	box	point	poly	box	point	poly
Direct	74.29	54.76	66.75	74.29	54.64	66.75	74.29	54.76	65.64
TENT	78.21	54.83	68.82	52.99	52.99	68.58	77.08	53.90	71.51
WDASS	77.29	**58.78**	68.94	74.69	60.55	69.15	75.27	**55.27**	70.19
OURS	**80.12**	52.28	**72.07**	**76.49**	**62.09**	**71.02**	**78.57**	52.88	**72.33**

semantic meaning and produces segmentation results that better aligns with human intention.

4.5. Ablation Study

In this section, we analyze the effectiveness of individual components on COCO dataset. As presented in Tab. 7, when self-training is applied alone, we observe a significant performance drop ($74.29\% \rightarrow 58.88\%$ with box weak supervision), suggesting the severity of confirmation bias. This issue can be well remedied when the anchor loss regularization is applied. Self-training with anchored regularization already improves the performance of SAM after adaptation. Finally, the contrastive loss directly regularizes the encoder network's output and contributes with additional improvement of segmentation accuracy. We also investigate the final self-training architecture without any weak supervision, i.e. the pseudo label masks are generated with grid point prompt. The results after adaptation is slightly worse than with weak supervision but consistently better than without adaptation on both box and coarse mask testing prompts. Nevertheless, the results with point prompt performs even worse than without adaptation, suggesting the necessity of weakly supervised adaptation.

4.6. Additional Analysis

Generalization to Cross-Prompt Testing: We further investigate the effectiveness of adaptation by testing with all three types of prompts. Specifically, we compare with the best (WDASS) and second best (TENT) methods with results presented in Tab. 8. Apart from using point as testing prompt, our methods consistently improves the performance even though the testing prompt is different from training weak supervision.

Effectiveness of Updating Different Components: There are many options for updating the pre-trained SAM model. Given limited computing resources, choosing the appropriate components of network to update is crucial to optimize the generalization performance. In this section, we investigate several options of network components to update during adaptation. **Full finetune** means finetuning the whole encoder. **MaskDecoder** finetune the whole decoder

Figure 4. qualitative results on some selected examples. Three types of prompts at testing stage are visualized for reference.

of SAM without additional redundant learnable parameters. **LayerNorm** finetune the all layernorms of the SAM image encoder. **LoRA** finetunes the encoder network in through low-rank decomposition only. **EVP** use the Embedding Tune and the HFC Tune to tune the extracted features by image encoder. We present the investigation on the COCO dataset with bounding boxes as weak supervision in Tab. 9. In particular, we have also explored the impact of different combinations of components, e.g. MaskDecoder + LoRA, LoRA + EVP, and so on. The results suggest LoRA finetuning the encoder network alone yields the best performance.

5. Conclusion

We studies the generalization of Segment-Anything model for multiple downstream image segmentation tasks. A tangible solution with no access to source dataset and low memory cost was proposed to adapt SAM to downstream data in a source-free domain adaptation manner. The proposed method is naturally compatible with weak supervisions which could substantially improve the efficacy of adaptation. Extensive evaluations on 10 datasets from 5 types of downstream tasks suggest the proposed adapta-

Table 9. Adaptation results produced by finetuning different modules.

Finetuning Module	IoU
None	74.30
Full finetune	78.54
MaskDecoder	76.81
LayerNorm	79.30
Low-Rank Adaptation (LoRA)	**80.12**
Explicit Visual Prompt (EVP)	77.38
MaskDecoder + LoRA	77.28
MaskDecoder + EVP	76.83
LoRA + EVP	79.47
LoRA + LayerNorm	79.35
MaskDecoder + LoRA + EVP	77.31

tion method can significantly improve the generalization of SAM under various degrees of distribution shift.

References

[1] Waqar Ahmed, Pietro Morerio, and Vittorio Murino. Cleaning noisy labels by negative ensemble learning for source-free unsupervised domain adaptation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022. 2

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel E O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference on Neural Networks*, 2020. 2, 4

[3] Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. Source-free domain adaptation for image segmentation. *Medical Image Analysis*, 2022. 3

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 2020. 2

[5] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. *arXiv preprint arXiv:2304.09148*, 2023. 1, 2, 5

[6] Xiaoyu Chen, Jiachen Hu, Chi Jin, Lihong Li, and Liwei Wang. Understanding domain randomization for sim-to-real transfer. In *Int. Conf. Learn. Represent.*, 2022. 1

[7] Tong Chu, Yahao Liu, Jinhong Deng, Wen Li, and Lixin Duan. Denoised maximum classifier discrepancy for source-free unsupervised domain adaptation. In *AAAI conference on artificial intelligence*, 2022. 2

[8] Anurag Das, Yongqin Xian, Dengxin Dai, and Bernt Schiele. Weakly-supervised domain adaptive semantic segmentation with prototypical contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3, 5

[9] Anurag Das, Yongqin Xian, Yang He, Zeynep Akata, and Bernt Schiele. Urban scene semantic segmentation with low-cost coarse annotation. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023. 3, 4

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2009. 2

[11] Yuhe Ding, Lijun Sheng, Jian Liang, Aihua Zheng, and Ran He. Proxymix: Proxy-based mixup training with label refinery for source-free domain adaptation. *Neural Networks*, 2023. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5

[13] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. 2015. 2, 6

[14] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 6

[15] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, 2015. 2

[16] Yunhe Gao, Xingjian Shi, Yi Zhu, Hao Wang, Zhiqiang Tang, Xiong Zhou, Mu Li, and Dimitris N Metaxas. Visual prompt tuning for test-time domain adaptation. *arXiv preprint arXiv:2210.04831*, 2022. 5

[17] David Gutman, Noel CF Codella, Emre Celebi, Brian Helba, Michael Marchetti, Nabin Mishra, and Allan Halpern. Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (isbi) 2016, hosted by the international skin imaging collaboration (isic). *arXiv preprint arXiv:1605.01397*, 2016. 6

[18] Dongsheng Han, Chaoning Zhang, Yu Qiao, Maryam Qamar, Yuna Jung, SeungKyu Lee, Sung-Ho Bae, and Choong Seon Hong. Segment anything model (sam) meets glass: Mirror and transparent objects cannot be easily detected. *arXiv preprint arXiv:2305.00278*, 2023. 2

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 3

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 2, 5

[21] Yihao Huang, Yue Cao, Tianlin Li, Felix Juefei-Xu, Di Lin, Ivor W Tsang, Yang Liu, and Qing Guo. On the robustness of segment anything. *arXiv preprint arXiv:2305.16220*, 2023. 1

[22] Yifei Huang, Lijin Yang, and Yoichi Sato. Weakly supervised temporal sentence grounding with uncertainty-guided self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18908–18918, 2023. 2

[23] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, 2020. 6

[24] Mengmeng Jing, Lichao Meng, Jingjing Li, Lei Zhu, and Heng Tao Shen. Adversarial mixup ratio confusion for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 2022. 2

[25] Mengmeng Jing, Xiantong Zhen, Jingjing Li, and Cees Snoek. Variational model perturbation for source-free domain adaptation. *Advances in Neural Information Processing Systems*, 2022. 2

[26] Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence*, 2021. 1, 2

[27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2, 3, 4

[28] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[29] Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *IEEE/CVF International Conference on Computer Vision*, 2021. 3

[30] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021. 2

[31] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, 2019. 6

[32] Jingjing Li, Zhekai Du, Lei Zhu, Zhengming Ding, Ke Lu, and Heng Tao Shen. Divergence-agnostic unsupervised domain adaptation by adversarial attacks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[33] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[34] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, 2020. 2, 3

[35] Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2, 4, 5

[36] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation, 2023. 2

[37] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 2, 6

[38] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *IEEE/CVF International Conference on Computer Vision*, 2017. 4

[39] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems*, 2021. 2

[40] Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 3

[41] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, 2015. 2

[42] Jun Ma and Bo Wang. Segment anything in medical images. *arXiv preprint arXiv:2304.12306*, 2023. 2

[43] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D vision*, 2016. 4

[44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 2

[45] Sujoy Paul, Yi-Hsuan Tsai, Samuel Schulter, Amit K Roy-Chowdhury, and Manmohan Chandraker. Domain adaptive semantic segmentation using weak labels. In *European Conference on Computer Vision*, 2020. 3

[46] Zhen Qiu, Yifan Zhang, Hongbin Lin, Shuaicheng Niu, Yanxia Liu, Qing Du, and Mingkui Tan. Source-free domain adaptation via avatar prototype generation and adaptation. *arXiv preprint arXiv:2106.15326*, 2021. 2, 3

[47] Sanqing Qu, Guang Chen, Jing Zhang, Zhijun Li, Wei He, and Dacheng Tao. Bmd: A general class-balanced multicentric dynamic prototype strategy for source-free domain adaptation. In *European Conference on Computer Vision*, 2022. 2

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 1, 2

[49] Andreas Richtsfeld, Thomas Mörwald, Johann Prankl, Michael Zillich, and Markus Vincze. Segmentation of unknown objects in indoor environments. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012. 6

[50] Maohao Shen, Yuheng Bu, and Gregory Wornell. On the benefits of selectivity in pseudo-labeling for unsupervised multi-source-free domain adaptation. *arXiv preprint arXiv:2202.00796*, 2022. 2

[51] Przemysław Skurowski, Hassan Abdulameer, J Błaszczyk, Tomasz Depta, Adam Kornacki, and P Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished manuscript*, 2(6):7, 2018. 6

[52] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 2020. 4

[53] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 5

[54] Yongyi Su, Xun Xu, and Kui Jia. Revisiting real-time training: Sequential inference and adaptation by anchored clustering. *Advances in Neural Information Processing Systems*, 2022. 5

[55] Yongyi Su, Xun Xu, and Kui Jia. Towards real-world test-time adaptation: Tri-net self-training with balanced normalization. *arXiv preprint arXiv:2309.14949*, 2023. 5

[56] Yongyi Su, Xun Xu, Tianrui Li, and Kui Jia. Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering regularized self-training. *arXiv preprint arXiv:2303.10856*, 2023. 4

[57] Markus Suchi, Timothy Patten, David Fischinger, and Markus Vincze. Easylabel: A semi-automatic pixel-wise object annotation tool for creating robotic RGB-D datasets. In *International Conference on Robotics and Automation*, 2019. 6

[58] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[59] Lv Tang, Haoke Xiao, and Bo Li. Can sam segment anything? when sam meets camouflaged object detection. *arXiv preprint arXiv:2304.04709*, 2023. 2, 6

[60] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 2

[61] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 2

[62] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Int. Conf. Learn. Represent.*, 2021. 5

[63] Qi Wang, Junyu Gao, and Xuelong Li. Weakly supervised adversarial domain adaptation for semantic segmentation in urban scenes. *IEEE Transactions on Image Processing*, 2019. 3

[64] Rui Wang, Zuxuan Wu, Zejia Weng, Jingjing Chen, Guo-Jun Qi, and Yu-Gang Jiang. Cross-domain contrastive learning for unsupervised domain adaptation. *IEEE Transactions on Multimedia*, 2022. 3

[65] Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In *IEEE/CVF International Conference on Computer Vision*, 2021. 2

[66] Lin Xiong, Mao Ye, Dan Zhang, Yan Gan, Xue Li, and Yingying Zhu. Source data-free domain adaptation of object detector through domain-specific perturbation. *International Journal of Intelligent Systems*, 2021. 2

[67] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *IEEE/CVF International Conference on Computer Vision*, 2021. 5

[68] Shiqi Yang, Yaxing Wang, Joost Van De Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *IEEE/CVF International Conference on Computer Vision*, 2021. 2

[69] Shiqi Yang, Yaxing Wang, Luis Herranz, Shangling Jui, and Joost van de Weijer. Casting a bait for offline and online source-free domain adaptation. *Computer Vision and Image Understanding*, 2023. 2

[70] Dan Zhang, Mao Ye, Lin Xiong, Shuaifeng Li, and Xue Li. Source-style transferred mean teacher for source-data free object detection. In *ACM Multimedia Asia*. 2021. 2

[71] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023. 2

[72] Qiming Zhang, Jing Zhang, Wei Liu, and Dacheng Tao. Category anchor-guided unsupervised domain adaptation for semantic segmentation. *Advances in Neural Information Processing Systems*, 2019. 2

[73] Yabin Zhang, Bin Deng, Hui Tang, Lei Zhang, and Kui Jia. Unsupervised multi-class domain adaptation: Theory, algorithms, and practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2775–2792, 2020. 2

[74] Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024. 2

[75] Yang Zou, Zhiding Yu, Xiaofeng Liu, BVK Kumar, and Jinsong Wang. Confidence regularized self-training. In *IEEE/CVF International Conference on Computer Vision*, 2019. 2