This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Continuous Adaptation for Interactive Segmentation Using Teacher-Student Architecture

Barsegh Atanyan[1], Levon Khachatryan[1], Shant Navasardyan[1], Yunchao Wei[2], Humphrey Shi[1,3]
[1]Picsart AI Research (PAIR), [2]BJTU, [3]Georgia Tech
https://github.com/Picsart-AI-Research/Interactive-Segmentation-with-Continuous-Adapation

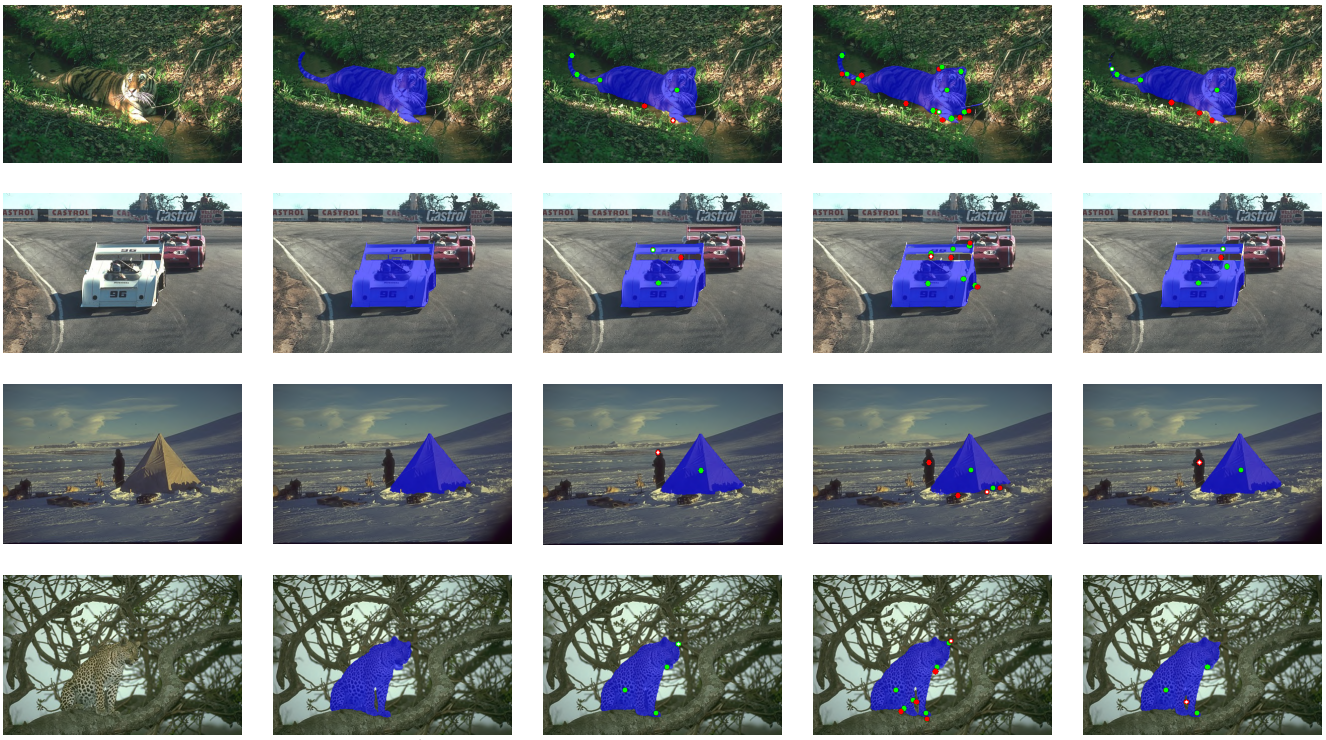

Figure 1. Comparison between frozen model FocalClick [11], the baseline and our method. Images are taken from Berkeley dataset. To illustrate the effectiveness of the proposed method to decrease catastrophic forgetting, the baseline and our method have been continuously adapted on DRIONS-DB [7] → Rooftop [49] → Berkeley [44]. Green and red points represent positive and negative clicks correspondingly. Automatic annotation has been done until reaching a target IOU of 95%.

Abstract

Interactive segmentation is the task of segmenting objects or regions of interest from images based on user annotations. While most current methods perform effectively on images from the same distribution as the training dataset, they suffer to generalize on unseen domains. To address this issue some approaches incorporate test-time adaptation techniques which, on the other hand, may lead to catastrophic forgetting (i.e. degrading the performance on the previously seen domains) when applied on datasets from various domains sequentially. In this paper, we propose a novel domain adaptation approach leveraging a teacher-student learning framework to tackle the catastrophic forgetting issue. Continuously updating the student and teacher models based on user clicks results in improved segmentation accuracy on unseen domains, while preserving comparable performance on previous domains. Our approach is evaluated on a sequence of datasets from unseen domains (i.e. medical, aerial images, etc.), and, after

adaptation, on the source domain demonstrating a significant decline of catastrophic forgetting (e.g. from 55% to 4% on Berkeley dataset).

1. Introduction

Interactive segmentation is the task of extracting masks for user specified objects or regions of interest. User annotations can be of the form of scribbles [4,6,16,17,24,28,33], clicks [5, 10, 11, 20, 22, 35, 36, 43, 47, 48, 60], bounding boxes [27, 46, 58, 61], polygons [1, 23, 37], *etc*. Applications of interactive segmentation vary from image editing (*e.g*. for editing selected objects, or removing them from the scene, etc.) to annotation of large datasets for image segmentation (*e.g*. for self-driving cars, aerial, or medical imaging, etc.).

Given the wide range of applications and the recent developments of deep learning methods for semantic, instance, or panoptic segmentation requiring large volumes of annotated data, there has been a growing interest towards interactive segmentation as well. As a result a large body of research [10, 11, 20, 22, 29, 35, 36, 43, 47, 48, 60, 61] has focused on developing new and improved deep learning based techniques for interactive segmentation.

Initial attempts of applying deep learning methods to interactive segmentation [60] incorporated user annotations as additional input to Fully Convolutional Networks (FCN) [39]. To better identify the target object, other works [36, 43, 61] explore new ways of utilizing user clicks. Still, click simulation for training did not reflect the iterative essence of the interactive task (all clicks were generated at once) in contrast to the practical application of interactive segmentation when people add clicks iteratively in erroneous regions based on current segmentation mask. To mimic human behaviour some works [11, 41, 48] generate clicks iteratively each time utilizing the segmentation result of previous iteration during training. Despite such improvements, most interactive segmentation methods failed to capture fine details in user annotated areas. To overcome this, new approaches emerged [11, 35] to apply local refinement around user clicks.

Nonetheless, the performance of these methods deteriorates on datasets from unseen domains. To address this issue, IA+SA [26] and RAIS [19] apply test-time adaptation to update the model based on user annotations which provide strong hints about the ground truth. These methods have shown promising results on adapting to a single dataset from an unseen domain. However, in most practical scenarios of large scale annotation the target domain is not stationary, meaning that datasets from diverse domains might become available for annotation. In such cases catastrophic forgetting becomes prominent in adaptation methods for interactive segmentation [19, 26]. To that end, it becomes crucial to update the model in a way that it accumulates knowledge from new datasets without forgetting the existing one. To accomplish this, we design a novel framework, consisting of teacher and student models, for continuous adaptation, applicable for any pre-trained off-the-shelf interactive segmentation model. Both the teacher and the student have the same architecture and the same initial knowledge by our design. During adaptation they constantly exchange their knowledge. Particularly, the teacher model shares its knowledge with the student before the adaptation on each dataset. Then the student model quickly gets adapted to the new dataset and exchanges the new knowledge with the teacher. The student model is updated based on user annotations, while the teacher is updated via an exponential moving average (EMA) of the student's parameters. Frequent updates of the student achieve improved performance on the current domain, while updating the teacher via EMA leads to a balanced knowledge accumulation from past and current domains, deteriorating forgetting.

Our contributions are summarized as.

- We propose a new method for continuous adaptation for interactive segmentation that alleviates catastrophic forgetting and can use any pre-trained off-the-shelf interactive segmentation model.
- As part of the proposed method, teacher and student models are updated using user-provided clicks and previous predictions to adapt to new domains.
- Through various adaptation scenarios we demonstrate the effectiveness of the proposed method both in adapting to new domain datasets such as Heart, Spleen, DRIONS-DB, Rooftop, but also we show how catastrophic forgetting is diminished compared to the baseline method.

2. Related Work

Interactive Segmentation. Before the recent breakthroughs in deep learning, traditional approaches [4, 6, 16, 17, 24, 27, 28, 46, 58] defined interactive segmentation as a graph cut optimization problem. For instance, some works [6, 46] treat user provided annotations as hard constraints for the optimization, or [16] use a random walker to get segmentation results. However, all these methods use only low-level features and do not yield high quality segmentation masks for images with complex structures or with similar foreground and background, hard textures. The introduction of FCNs for semantic segmentation [39] enabled researchers to apply deep learning based techniques to interactive segmentation as well. Since then various visual backbones [8, 9, 54, 59] have been used in interactive segmentation methods. The pioneering work using deep

Method	Backbone	GrabCut		Berkeley		DAVIS	
		NoC@85	NoC@90	NoC@85	NoC@90	NoC@85	NoC@90
RITM [48]	HRNet-18	**1.42**	1.54	**1.46**	2.26	4.36	5.74
PseudoClick [38]	HRNet-32	-	1.50	-	2.08	3.79	5.11
FocalClick [11]	SegFormerB3-S2	1.44	1.50	1.55	1.92	3.61	4.90
Ours$_{\text{FocalClick}}$	SegFormerB3-S2	**1.42**	**1.46**	1.53	**1.85**	**3.40**	**4.86**

Table 1. Evaluation results on benchmark datasets. NoC@85 and NoC@90 are used for denoting the average Number of Clicks required to obtain 85% and 90% IOU correspondingly.

Method	DRIONS-DB		Rooftop		
	NoC@85	NoC@90	NoC@80	NoC@85	NoC@90
IA+SA [26]	-	3.1	3.6	-	-
FocalClick [11]	4.75	6.23	1.79	2.22	2.86
Ours$_{\text{FocalClick}}$	**1.63**	**2.22**	**1.60**	**1.84**	**2.30**

Table 2. Evaluation results on datasets from different domains from the source. Note that IA+SA [26] and our method use test-time adaptation.

learning for interactive segmentation is DIOS [60], where user interactions are of the form of positive and negative clicks indicating the foreground object and the background. Clicks are transformed to Euclidean distance maps and are passed to the model together with the input image. DIOS established a well-defined strategy for random click simulation for training as well as an evaluation protocol for click based interactive segmentation methods. Most recent works use clicks as a method of interaction and have explored various representations - Gaussian or binary disks [48], superpixel-based or object-based guidance maps [42]. Besides, different methods utilize clicks with diverse purposes. FCA-Net [36] capitalizes the observation that the first click is the most important one and uses a separate First-Click Attention module to get improved results. DEXTR [43] uses four extreme points (left-most, top-most, right-most, bottom-most) to identify the object user wants to segment. IOG [61] takes one inside click near to the center of the foreground object and two more clicks at symmetrical corner locations (enabling to construct the bounding box) providing the opportunity for further corrections as well. Since during evaluation each next click is placed on the largest erroneous region, ITIS [41] proposed to generate several clicks iteratively during training to mimic natural annotation behaviour. RITM [48] emphasizes the importance of the usage of modern backbone models and large high-quality training datasets to significantly improve model performance. Several methods take the resulting segmentation mask of previous interactive step together with the image and click maps as input to the next step [11,14,35,41,48]. Meanwhile EdgeFlow [20] provides an edge mask as an additional input. RIS-Net [31], FocusCut [35] and FocalClick [11] focus on the interactive segmentation task from a localized perspective in order to refine the output. FCFI [57] exploits user clicks both for local feedback correction and global feedback integration into deep features. GPCIS [62] formulates the interactive segmentation task as a pixel-wise binary classification model based on a Gaussian process for each image. PseudoClick [38] uses automatically generated clicks to mimic human clicks for the enhancement of segmentation masks. PhraseClick [12] combines clicks and textual phrases to correctly locate and segment the target object. CLIPSeg [40] relies on visual or textual prompts to segment the target object. Recently released SAM [25] provides zero-shot capabilities for promptable segmentation based on clicks, bounding boxes, masks, text descriptions. BRS [22] and f-BRS [47] introduce a test-time optimization of the network inputs [22] or some auxiliary parameters [47] to ensure that user-provided clicks are labeled correctly. IA+SA [26] exploits user annotations to update the whole model test-time to adapt to specific images or new domains. RAIS [19] constructs a model with basic segmentation and adaptation modules and updates them test-time. Besides segmenting a single target object, methods have been proposed for full image annotation [2], human parsing [15], thin object segmentation [32].

Adaptation. Deep learning models in general, including semantic segmentation models, are trained on a specific source dataset, which is assumed to represent the distribution of the data that the model will be exposed to in the future. However, when the model is deployed in the real world, it may encounter data that comes from a different distribution than the one it was trained on. Various adaptation techniques have been designed to improve the performance of the model on new domains and increase its generalization capability, including but not limited to Unsu-

Method	Heart		Spleen	
	NoC@85	NoC@90	NoC@85	NoC@90
FocalClick [11]	5.75	8.40	3.73	4.29
Ours$_{\text{FocalClick}}$	**3.30**	**6.70**	**1.51**	**1.85**

Table 3. Evaluation results on Heart and Spleen datasets from The Medical Segmentation Decathlon [3]

Adaptation Sequence	Heart		Spleen	
	NoC@85	NoC@90	NoC@85	NoC@90
Heart $\longrightarrow$ Spleen	3.30	6.70	**1.46**	**1.83**
Spleen $\longrightarrow$ Heart	**2.95**	**6.00**	1.51	1.85

Table 4. Evaluation results on datasets from The Medical Segmentation Decathlon [3] using FocalClick [11] as pre-trained interactive segmentation model. Continuous adaptation has been applied on the specified sequences of datasets

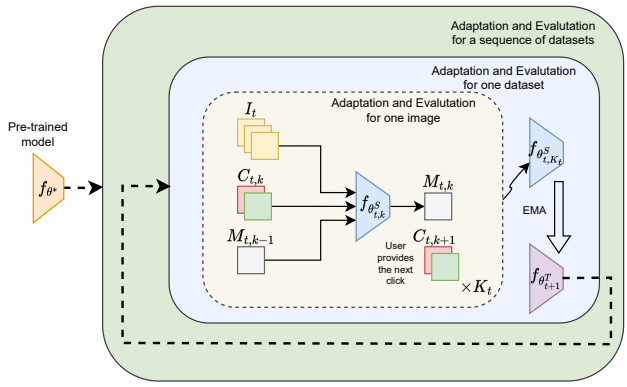

Figure 2. The overall architecture of our method. As part of it, the student model is used for interactive annotation and is adapted to the current dataset. Meanwhile the teacher model accumulates new knowledge balancing with the existing one via updates through exponential moving averages of student's parameters.

pervised Domain Adaptation [21, 30, 51, 52, 55], Test-Time Training [50], Test-Time Adaptation [53], Continual Learning [13, 56]. While some of these methods assume an access to target data during training, others use a subset of the source data for test-time adaptation. We believe that the above approaches put quite hard limitations as, first, in most practical situations the target domain is unknown during training, second, it is not feasible to keep the source data due to ethical, legal or just computational reasons. Moreover, the target domain should not necessarily be stationary and can change over time. Another approach for adaptation is to update the model test-time based on prediction entropy [30, 53] or self-training with pseudo-labels. CoTTA [56] proposes a framework for continual test-time adaptation that gives weight-averaged pseudo-labels and the prediction using several augmentations.

3. Method

We start this section with the problem formulation of continuous adaptation for interactive segmentation, followed by a subsection presenting our method.

3.1. Problem Formulation

Continuous adaptation of a pre-trained interactive segmentation model can be formally defined as follows. Let $f_{\theta^*}(x)$ be an interactive segmentation model with input x (image, user annotations, *etc.*), weights θ^* obtained after training on source dataset D_S. Also, there are L datasets $\mathcal{D} = \{D_1, D_2, \ldots, D_L\}$ from arbitrary domains, where each dataset D_l consists of n_l images, i.e. $D_l = \{I_1, I_2, \ldots, I_{n_l}\}$. However, these datasets are not available during training. Similarly, source dataset D_S is not available during annotation. The goal is to update pre-trained model parameters θ^* during annotation such that the performance is improved on dataset D_l while ensuring the performance does not degrade on previous datasets $D_S, D_1, D_2, \ldots D_{l-1}$.

3.2. Our Method

In this paper we design a continuous adaptation framework that enables to achieve the above formulated goals - adapt to new domains and preserve past knowledge. For that purpose we exploit teacher-student architecture, where the student is supposed to quickly adapt to a new domain, meanwhile the teacher accumulates the newly learned knowledge without catastrophic forgetting. The overview of our method can be found in Fig. 2.

Let the interactive segmentation model be $f_\theta(I, C, M)$, where I is the input RGB image, C is the 2 channel guidance map, including both positive and negative clicks, M is model's prediction at previous interactive step and θ represents model parameters. Note that I, C and M have the same spatial resolution $H \times W$. After a fully-supervised

Adaptation Sequence	**Baseline**$_{\text{FocalClick}}$ NoC@90	**Ours**$_{\text{FocalClick}}$ NoC@90
GrabCut	1.46	1.46
DRIONS-DB $\longrightarrow$ GrabCut	1.60	**1.46**
DRIONS-DB $\longrightarrow$ Rooftop $\longrightarrow$ GrabCut	1.72	**1.46**
Berkeley	1.85	1.85
DRIONS-DB $\longrightarrow$ Berkeley	2.72	**1.90**
DRIONS-DB $\longrightarrow$ Rooftop $\longrightarrow$ Berkeley	2.87	**1.93**
DAVIS	4.86	4.86
DRIONS-DB $\longrightarrow$ DAVIS	5.20	**4.79**
DRIONS-DB $\longrightarrow$ Rooftop $\longrightarrow$ DAVIS	5.23	**4.79**

Table 5. Evaluation results for the last dataset in the specified sequence of continuous adaptation using FocalClick [11] as pre-trained interactive segmentation model. The proposed method diminishes catastrophic forgetting.

Adaptation Sequence	**Baseline**$_{\text{RITM}}$ NoC@90	**Ours**$_{\text{RITM}}$ NoC@90
GrabCut	1.48	1.48
DRIONS-DB $\longrightarrow$ GrabCut	2.20	**1.48**
DRIONS-DB $\longrightarrow$ Rooftop $\longrightarrow$ GrabCut	2.10	**1.52**
Berkeley	2.27	2.27
DRIONS-DB $\longrightarrow$ Berkeley	4.39	**2.38**
DRIONS-DB $\longrightarrow$ Rooftop $\longrightarrow$ Berkeley	4.99	**2.37**
DAVIS	5.21	5.21
DRIONS-DB $\longrightarrow$ DAVIS	6.92	**5.24**
DRIONS-DB $\longrightarrow$ Rooftop $\longrightarrow$ DAVIS	7.63	**5.22**

Table 6. Evaluation results for the last dataset in the specified sequence of continuous adaptation using RITM [48] as pre-trained interactive segmentation model. The proposed method diminishes catastrophic forgetting.

training on source dataset D_S the optimized model parameters are θ^*.

At the beginning of adaptation, both teacher $f_{\theta^T}(I, C, M)$ and student $f_{\theta^S}(I, C, M)$ models are initialized as the pre-trained interactive segmentation model with weights θ^*, i.e. $\theta^T = \theta^*$ and $\theta^S = \theta^*$. To adapt on a dataset $D_l = \{I_1, I_2, ..., I_{n_l}\}$ we employ the following adaptation mechanism. For each image I_t after each click k a single gradient descent (GD) step is used to update the student model parameters $\theta^S_{t,k}$ based on L_C loss:

$$L_C(I_t, C_{t,k}, M_{t,k-1}, \theta^S_{t,k-1}, \theta^*) = L_{SCE}(M_{t,k-1}, C_{t,k}) + \gamma L_R(\theta^*, \theta^S_{t,k-1}), \tag{1}$$

where $C_{t,k} = [C^P_{t,k}, C^N_{t,k}]$ and $M_{t,k-1}$ denote the user annotations (positive and negative click maps) at the click k and the student model prediction after $k-1$ clicks respectively. $L_{SCE}(M_{t,k-1}, C_{t,k})$ denotes the sparse binary cross-entropy loss between the prediction and user annotations, i.e. pixels without annotation are not considered. $L_R(\theta^*, \theta^S_{t,k-1})$ is used to penalize large deviations from the pre-trained weights. γ is a hyper-parameter.

$$L_{SCE}(M_{t,k-1}, C_{t,k}) = \frac{1}{k}\sum_p [-C^P_{t,k} \odot \log M_{t,k-1} - C^N_{t,k} \odot \log(1 - M_{t,k-1})]_p \tag{2}$$

$$L_R(\theta^*, \theta^S_{t,k-1}) = \|\theta^* - \theta^S_{t,k-1}\|^2 \tag{3}$$

After there are no more corrections from the user and the interactive process has ended for I_t, we use a single GD step to update θ^S_{t,K_t} optimizing L_I loss:

$$L_I(I_t, C_{t,1:K_t}, M_{t,1:K_t}, \theta^S_{t,K_t}, \theta^*) = L_{SCE}(M_{t,K_t}, C_{t,K_t}) + L_{BCE}(M_{t,K_t}, M_{t,1}) + \gamma L_R(\theta^*, \theta^S_{t,K_t}), \tag{4}$$

where $C_{t,1:K_t} = \{C_{t,k}\}_{k=1}^{K_t}$, $M_{t,1:K_t} = \{M_{t,k}\}_{k=1}^{K_t}$ for K_t total number of clicks, and L_{BCE} is a regularization term based on the binary cross-entropy between the predicted segmentation maps $M_{t,1}$ and M_{t,K_t}:

$$L_{BCE}(M_{t,K_t}, M_{t,1}) = \frac{1}{HW}\sum_{p}[-M_{t,1} \odot \log M_{t,K_t} + (1 - M_{t,1}) \odot \log(1 - M_{t,K_t})]_p \quad (5)$$

After obtaining a final prediction for image I_t, we update teacher model as well. To achieve the initial objective for the teacher model and accumulate new knowledge without forgetting past one, teacher model is updated by the exponential moving average (EMA) using student model parameters:

$$\theta^T_{t+1} = \alpha\theta^T_t + (1 - \alpha)\theta^S_{t,K_t} \quad (6)$$

This way the adaptation is done iteratively for both student and teacher models updating student model parameters after each user click and updating teacher model parameters once for each image.

To adapt on dataset D_{l+1}, we keep teacher model as it is at the moment, i.e. updated using all images up to the last image of D_l. We reinitialize student model so that it has the same parameters as the teacher before the adaptation on D_{l+1}. This process is repeated for all available datasets. See Algorithm 1 for clarity. Note that annotation and adaptation happen in parallel. Implementation details can be found in the Supplementary Material.

Algorithm 1 Continuous Adaptation for interactive segmentation with teacher-student architecture

Require: Model $f_\theta(I, C, M)$, pre-trained weights θ^*, datasets $\mathcal{D}$, hyperparameters α, γ, learning rate λ

$\theta^T \leftarrow \theta^*$ ▷ Initialize teacher and student using the pre-trained weights
$\theta^S \leftarrow \theta^*$
for all $D_l \in \mathcal{D}$ **do**
  $\theta^S_{0,0} \leftarrow \theta^T$ ▷ Reinitialize student with teacher parameters for each new dataset
  for all $t = 1, 2, \ldots, n_l$ **do**
    $C_{t,0} \leftarrow \{\mathbf{0}, \mathbf{0}\}$
    $M_{t,0} \leftarrow \mathbf{0}$
    for all $k = 1, 2, \ldots, K_t$ **do**
      $C_{t,k} \leftarrow C_{t,k-1} \cup NewClick(I_t, M_{t,k-1})$
      $M_{t,k} \leftarrow f_{\theta^S}(I_t, C_{t,k}, M_{t,k-1})$
      $\theta^S_{t,k} \leftarrow \theta^S_{t,k-1} - -\lambda\frac{d}{d\theta^S}L_C(I_t, C_{t,k}, M_{t,k-1}, \theta^S_{t,k-1}, \theta^*; \gamma)$ ▷ Update student after each click
    end for
    $\theta^S_{t,K_t} \leftarrow \theta^S_{t,K_t} - -\lambda\frac{d}{d\theta^S}L_I(I_t, C_{t,1:K_t}, M_{t,1:K_t}, \theta^S_{t,K_t}, \theta^*; \gamma)$ ▷ Update student after each image
    $\theta^T \leftarrow \alpha\theta^T + (1 - \alpha)\theta^S_{t,K_t}$ ▷ Update teacher via EMA after each image
  end for
end for
return $\theta^T, \mathcal{M}$ ▷ Return teacher model parameters and all student predictions

4. Experiments

In this section we introduce how we have defined the baseline model. Then we evaluate the baseline and the proposed teacher-student architecture in different adaptation scenarios. We use FocalClick SegFormerB3-S2 [11] trained on COCO [34] + LVIS [18] as off-the-shelf pre-trained interactive segmentation model. First, we show that small improvements exist over the frozen model even for datasets that come from the same domain as the training dataset. Next, we confirm that the proposed adaptation method helps to increase the performance on datasets from various domains - aerial or medical images. To verify that continuous adaptation is superior to adaptation from the pre-trained model for each new dataset, we show that adaptation on similar domain dataset improves the performance on subsequent datasets. We demonstrate that the proposed adaptation setting decreases catastrophic forgetting on different sequences. Finally, we construct our method on top of RITM [48] and show that our continual adaptation mechanism can use arbitrary off-the-shelf pre-trained interactive segmentation model.

4.1. Constructing a baseline

We decide to construct a baseline model using IA+SA [26] adaptation approach, as a prominent work in test-time adaptation for interactive segmentation. Adaptation happens based on two defined mechanisms - Single Image Adaptation(IA) and Image Sequence Adaptation(SA). While the former is responsible to learn image specific details, the latter adapts the model to the large domain changes. Since ground truth masks are not available test time, user-provided clicks are used for sparse supervision. Also, two terms of regularization are used to ensure the pre-trained model does not forget the strong prior knowledge - dense supervision and important parameter change regularizer.

IA+SA uses a strong re-implementation of ITIS [41] as a pre-trained interactive segmentation model. However, as the implementation of IA+SA is not available and since the authors claim that the proposed adaptation mechanisms are orthogonal to architectural changes, we use current SOTA interactive segmentation model FocalClick [11] to construct a baseline method combining with IA+SA [26] adaptation approach.

In case of continuous adaptation we apply this baseline on a sequence of datasets. We demonstrate that catastrophic forgetting takes place for this approach. Through vari-

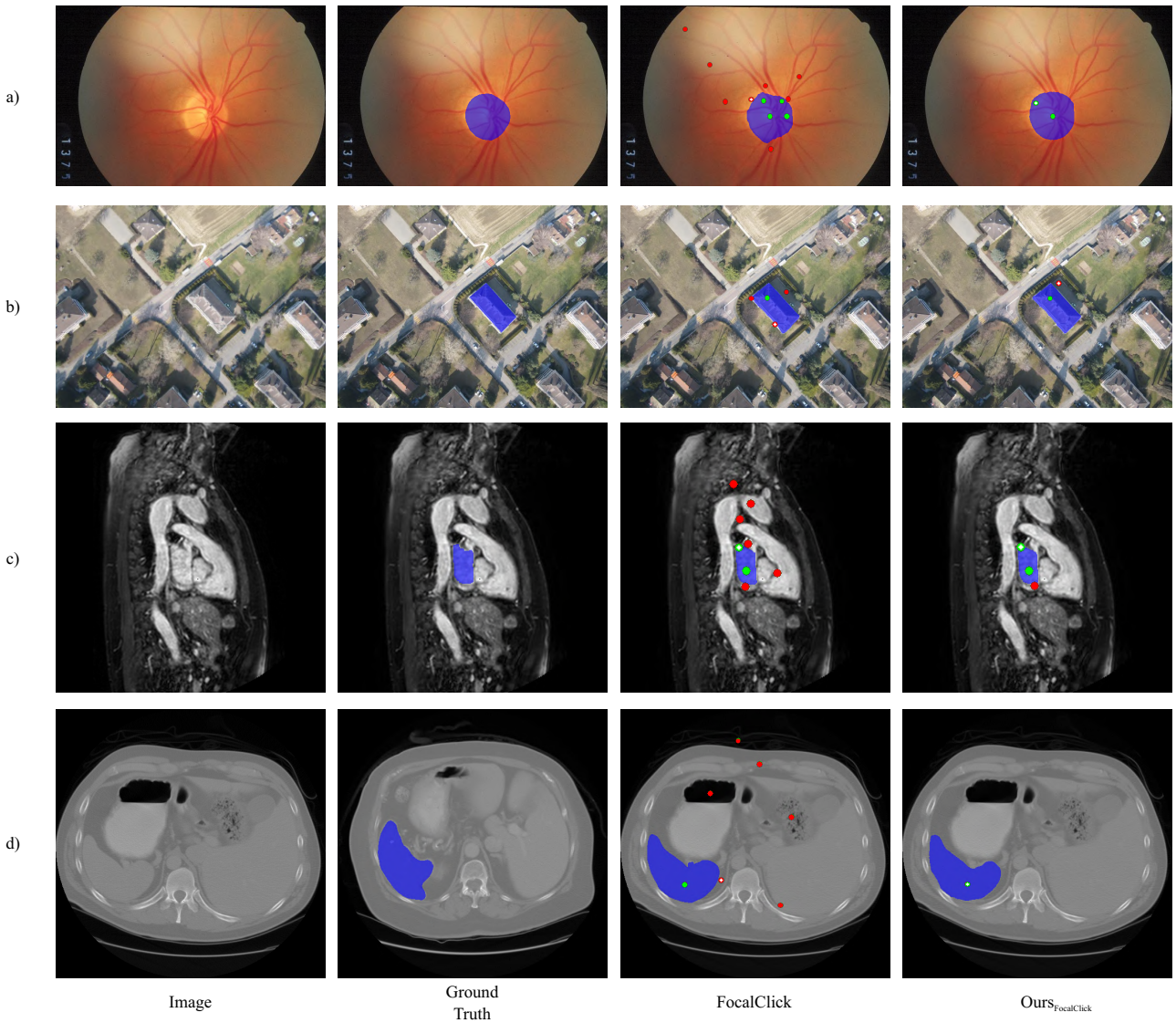

Figure 3. Comparison between frozen model FocalClick [11] and our method on diverse datasets from several domains. a) DRIONS-DB [7], b) Rooftop [49], c) Heart [3], d) Spleen [3]. Green and red points represent positive and negative clicks correspondingly. Automatic annotation has been done until reaching a target IOU of 95%

ous experiments we show that the proposed teacher-student framework for continuous adaptation of any pre-trained off-the-shelf interactive segmentation model helps to eliminate catastrophic forgetting.

4.2. Comparison with SOTA and baseline

Table 1 shows the comparison with SOTA interactive segmentation methods. We apply our method on three widely used benchmark datasets. GrabCut [46] and Berkeley [44] contain 50 and 100 samples correspondingly. Also, a subset with 345 frames from videos of DAVIS [45] dataset is used. It is the same subset proposed in Latent Diversity [29]. We can see that even in the absence of any domain changes, model updates help to achieve higher IOU with fewer clicks. Fig. 1 includes several qualitative results obtained by FocalClick [11], the baseline and our method.

To verify how effectively the proposed method adapts to large domain shifts, we use the same datasets as in IA+SA [26] from medical domain - DRIONS-DB [7] containing 110 images of eye fundus of different patients, and a dataset of 63 aerial images, Rooftop [49]. Results in Table 2 demonstrate how much adaptation helps to increase the performance on new domain datasets and decrease the required number of annotations to reach the target IOU. Par-

ticularly, for DRIONS-DB Number of Clicks required to obtain 90% IOU (NoC@90) is 2.22 for our method compared to 6.23 by FocalClick.

Besides DRIONS-DB and Rooftop, we evaluate our method on other datasets with large domain shift. We use Heart and Spleen datasets from The Medical Segmentation Decathlon [3]. These datasets contain 30 and 61 3D volumes of CT scans correspondingly. Since the proposed method is designed for 2D segmentation, we extract the slice with maximal ground truth mask area from each volume to obtain 30 Heart and 61 Spleen CT slices and their annotations. Again the proposed adaptation helps to achieve a significant performance improvement for these datasets as reported in Table 3. Several qualitative results are presented in Fig. 3. More examples can be found in the Supplementary Material.

To move on, in many practical scenarios several datasets from similar domains are used for annotation. Hence, it is reasonable to accumulate newly learned knowledge. To better illustrate the above point, we have applied our method to continuously adapt to two different datasets of CT scans. In one case, we adapted on Heart dataset, then on Spleen. On the other case, in the opposite order. Though there is a slight difference for Spleen dataset when adapted first or second, it might be because of already quite good performance on it. On the other hand, we can notice an average improvement about 10% or 0.7 clicks for Heart dataset when adapted after Spleen as reported in Table 4. The acquired knowledge during the adaptation on Spleen dataset boosts the performance on Heart scans.

Furthermore, we illustrate the effects of catastrophic forgetting doing continual adaptation first on one or two new domain datasets (DRIONS-DB, or DRIONS-DB and Rooftop) and then each of three benchmark datasets (GrabCut, Berkeley or DAVIS). As reported in Table 5, continuous adaptation of the baseline on the specified sequences yields significant forgetting for each of the benchmark datasets. Meanwhile applying the proposed teacher-student architecture eliminates effects of forgetting for GrabCut, Berkeley and provides an improvement for DAVIS.

We also construct the proposed continuous adaptation mechanism using another interactive segmentation model - RITM [48], to verify that any off-the-shelf pre-trained interactive segmentation model can be utilized. The baseline is designed in the same way as in previous experiments, except with one difference - RITM is used instead of FocalClick. We choose to use RITM with HRNet-18 backbone [9]. In Table 6 we demonstrate that the forgetting has been drastically decreased after applying teacher-student architecture for continuous adaptation for all benchmark datasets.

5. Conclusion

In this paper, we have presented a novel approach to interactive segmentation that leverages a teacher-student architecture for continuous adaptation. Our approach addresses the issue of catastrophic forgetting that arises when adaptation is done on several datasets from different domains sequentially. The proposed update rules for teacher and student models enable to achieve better results on benchmark datasets such as GrabCut, Berkeley and DAVIS, improve performance on different domain datasets by a large margin and decrease the effects of catastrophic forgetting significantly. We believe that our proposed method provides a promising direction for future research in the field of continuous adaptation for interactive segmentation, and can be applied to a wide range of applications such as general digital, aerial or medical imaging.

References

[1] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 859–868, 2018. 2

[2] Eirikur Agustsson, Jasper RR Uijlings, and Vittorio Ferrari. Interactive full image segmentation by considering all regions jointly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11622–11631, 2019. 3

[3] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 4, 7, 8

[4] Junjie Bai and Xiaodong Wu. Error-tolerant scribbles based interactive image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 392–399, 2014. 2

[5] Rodrigo Benenson, Stefan Popov, and Vittorio Ferrari. Large-scale interactive object segmentation with human annotators. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11700–11709, 2019. 2

[6] Yuri Y Boykov and M-P Jolly. Interactive graph cuts for optimal boundary & region segmentation of objects in nd images. In *Proceedings eighth IEEE international conference on computer vision. ICCV 2001*, volume 1, pages 105–112. IEEE, 2001. 2

[7] Enrique J Carmona, Mariano Rincón, Julián García-Feijoó, and José M Martínez-de-la Casa. Identification of the optic nerve head with genetic algorithms. *Artificial intelligence in medicine*, 43(3):243–259, 2008. 1, 7

[8] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolu-

tion, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[9] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2, 8

[10] Xi Chen, Zhiyan Zhao, Feiwu Yu, Yilei Zhang, and Manni Duan. Conditional diffusion for interactive segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7345–7354, 2021. 2

[11] Xi Chen, Zhiyan Zhao, Yilei Zhang, Manni Duan, Donglian Qi, and Hengshuang Zhao. Focalclick: towards practical interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1300–1309, 2022. 1, 2, 3, 4, 5, 6, 7

[12] Henghui Ding, Scott Cohen, Brian Price, and Xudong Jiang. Phraseclick: toward achieving flexible interactive segmentation by phrase and click. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 417–435. Springer, 2020. 3

[13] Arthur Douillard, Yifu Chen, Arnaud Dapogny, and Matthieu Cord. Plop: Learning without forgetting for continual semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4040–4050, 2021. 4

[14] Marco Forte, Brian Price, Scott Cohen, Ning Xu, and François Pitié. Getting to 99% accuracy in interactive segmentation. *arXiv preprint arXiv:2003.07932*, 2020. 3

[15] Yutong Gao, Liqian Liang, Congyan Lang, Songhe Feng, Yidong Li, and Yunchao Wei. Clicking matters: Towards interactive human parsing. *IEEE Transactions on Multimedia*, 2022. 3

[16] Leo Grady. Random walks for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 28(11):1768–1783, 2006. 2

[17] Varun Gulshan, Carsten Rother, Antonio Criminisi, Andrew Blake, and Andrew Zisserman. Geodesic star convexity for interactive image segmentation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3129–3136. IEEE, 2010. 2

[18] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 6

[19] Yuying Hao, Yi Liu, Juncai Peng, Haoyi Xiong, Guowei Chen, Shiyu Tang, Zeyu Chen, and Baohua Lai. Rais: Robust and accurate interactive segmentation via continual learning. *arXiv preprint arXiv:2210.10984*, 2022. 2, 3

[20] Yuying Hao, Yi Liu, Zewu Wu, Lin Han, Yizhou Chen, Guowei Chen, Lutao Chu, Shiyu Tang, Zhiliang Yu, Zeyu Chen, et al. Edgeflow: Achieving practical interactive segmentation with edge-guided flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1551–1560, 2021. 2, 3

[21] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016. 4

[22] Won-Dong Jang and Chang-Su Kim. Interactive image segmentation via backpropagating refinement scheme. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5297–5306, 2019. 2, 3

[23] Namgil Kim, Barom Kang, and Yeonok Cho. Split-gcn: Effective interactive annotation for segmentation of disconnected instance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 2

[24] Tae Hoon Kim, Kyoung Mu Lee, and Sang Uk Lee. Nonparametric higher-order learning for interactive segmentation. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3201–3208. IEEE, 2010. 2

[25] 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. *arXiv preprint arXiv:2304.02643*, 2023. 3

[26] Theodora Kontogianni, Michael Gygli, Jasper Uijlings, and Vittorio Ferrari. Continuous adaptation for interactive object segmentation by learning from corrections. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 579–596. Springer, 2020. 2, 3, 6, 7

[27] Victor Lempitsky, Pushmeet Kohli, Carsten Rother, and Toby Sharp. Image segmentation with a bounding box prior. In *2009 IEEE 12th international conference on computer vision*, pages 277–284. IEEE, 2009. 2

[28] Yin Li, Jian Sun, Chi-Keung Tang, and Heung-Yeung Shum. Lazy snapping. *ACM Transactions on Graphics (ToG)*, 23(3):303–308, 2004. 2

[29] Zhuwen Li, Qifeng Chen, and Vladlen Koltun. Interactive image segmentation with latent diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 577–585, 2018. 2, 7

[30] 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*, pages 6028–6039. PMLR, 2020. 4

[31] JunHao Liew, Yunchao Wei, Wei Xiong, Sim-Heng Ong, and Jiashi Feng. Regional interactive image segmentation networks. In *2017 IEEE international conference on computer vision (ICCV)*, pages 2746–2754. IEEE, 2017. 3

[32] Jun Hao Liew, Scott Cohen, Brian Price, Long Mai, and Jiashi Feng. Deep interactive thin object selection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 305–314, 2021. 3

[33] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3159–3167, 2016. 2

[34] 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

Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, pages 740–755. Springer, 2014. 6

[35] Zheng Lin, Zheng-Peng Duan, Zhao Zhang, Chun-Le Guo, and Ming-Ming Cheng. Focuscut: Diving into a focus view in interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2637–2646, 2022. 2, 3

[36] Zheng Lin, Zhao Zhang, Lin-Zhuo Chen, Ming-Ming Cheng, and Shao-Ping Lu. Interactive image segmentation with first click attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13339–13348, 2020. 2, 3

[37] Huan Ling, Jun Gao, Amlan Kar, Wenzheng Chen, and Sanja Fidler. Fast interactive object annotation with curve-gcn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5257–5266, 2019. 2

[38] Qin Liu, Meng Zheng, Benjamin Planche, Srikrishna Karanam, Terrence Chen, Marc Niethammer, and Ziyan Wu. Pseudoclick: Interactive image segmentation with click imitation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VI*, pages 728–745. Springer, 2022. 3

[39] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 2

[40] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7086–7096, 2022. 3

[41] Sabarinath Mahadevan, Paul Voigtlaender, and Bastian Leibe. Iteratively trained interactive segmentation. *arXiv preprint arXiv:1805.04398*, 2018. 2, 3, 6

[42] Soumajit Majumder and Angela Yao. Content-aware multi-level guidance for interactive instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11602–11611, 2019. 3

[43] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 616–625, 2018. 2, 3

[44] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, volume 2, pages 416–423. IEEE, 2001. 1, 7

[45] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 7

[46] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23(3):309–314, 2004. 2, 7

[47] Konstantin Sofiiuk, Ilia Petrov, Olga Barinova, and Anton Konushin. f-brs: Rethinking backpropagating refinement for interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8623–8632, 2020. 2, 3

[48] Konstantin Sofiiuk, Ilya A Petrov, and Anton Konushin. Reviving iterative training with mask guidance for interactive segmentation. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 3141–3145. IEEE, 2022. 2, 3, 5, 6, 8

[49] Xiaolu Sun, C Mario Christoudias, and Pascal Fua. Free-shape polygonal object localization. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 317–332. Springer, 2014. 1, 7

[50] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. 4

[51] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7472–7481, 2018. 4

[52] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2517–2526, 2019. 4

[53] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 4

[54] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 2

[55] Qin Wang, Dengxin Dai, Lukas Hoyer, Luc Van Gool, and Olga Fink. Domain adaptive semantic segmentation with self-supervised depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8515–8525, 2021. 4

[56] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 4

[57] Qiaoqiao Wei, Hui Zhang, and Jun-Hai Yong. Focused and collaborative feedback integration for interactive image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18643–18652, 2023. 3

[58] Jiajun Wu, Yibiao Zhao, Jun-Yan Zhu, Siwei Luo, and Zhuowen Tu. Milcut: A sweeping line multiple instance learning paradigm for interactive image segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 256–263, 2014. 2

[59] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34:12077–12090, 2021. 2

[60] Ning Xu, Brian Price, Scott Cohen, Jimei Yang, and Thomas S Huang. Deep interactive object selection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 373–381, 2016. 2, 3

[61] Shiyin Zhang, Jun Hao Liew, Yunchao Wei, Shikui Wei, and Yao Zhao. Interactive object segmentation with inside-outside guidance. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12234–12244, 2020. 2, 3

[62] Minghao Zhou, Hong Wang, Qian Zhao, Yuexiang Li, Yawen Huang, Deyu Meng, and Yefeng Zheng. Interactive segmentation as gaussion process classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19488–19497, 2023. 3