ODES: Online Domain Adaptation with Expert Guidance for Medical Image Segmentation

Md Shazid Islam *, Sayak Nag, Arindam Dutta, Sk Miraj Ahmed**, Fahim Faisal Niloy, Shreyangshu Bera, and Amit K. Roy-Chowdhury

University of California, Riverside, CA, USA
{misla048,snag005,adutt020,sahme047,fnilo001,sbera004,amitrc}@ucr.edu

Abstract. Unsupervised domain adaptive segmentation typically relies on self-training using pseudo labels predicted by a pre-trained network on an unlabeled target dataset. However, noisy pseudo-labels present a major bottleneck in adapting a network to distribution shifts between source and target domains, particularly when data is coming in an online manner and adaptation is constrained to exactly one round of forward and backward passes. In this scenario, relying solely on inaccurate pseudo-labels can degrade segmentation quality, which is detrimental to medical image segmentation where accuracy and precision are of utmost priority. In this paper, we propose an approach to address this issue by incorporating expert guided active learning to enhance online domain adaptation, even without dedicated training data. We call our method **ODES**: Online Domain Adaptation with Expert Guidance for Medical Image Segmentation that adapts to each incoming batch of data in an online setup. However, acquiring annotations through active learning for all images in a batch often results in redundant data annotation and increases temporal overhead in online adaptation. We address this issue by proposing a novel image-pruning strategy that selects the most informative subset of images from the current batch for active learning. We also propose a novel acquisition function that enhances diversity of the selected samples for annotating. Our approach outperforms existing online adaptation approaches and produces competitive results compared to offline domain adaptive active learning methods. The code can be found at https://github.com/ShazidAraf/ODES

Keywords: Domain Adaptation, Active Learning, Deep Learning, Segmentation, Online Adaptation

1 Introduction

In recent years, deep learning-based models have shown impressive performance in medical image segmentation [1, 20]. However, their performance is attributed to the availability of fully annotated training data, which is expensive to acquire [9]. Furthermore, a model trained on one dataset might exhibit poor performance on another dataset due to domain shift [5]. In recent years, Unsupervised Domain Adaptation (UDA) [27, 30] has been proposed which leverages labeled source data and unlabeled

* Corresponding author
** The coauthor is now at Brookhaven National Laboratory, NY, USA; this work was done during his graduate studies at UCR.

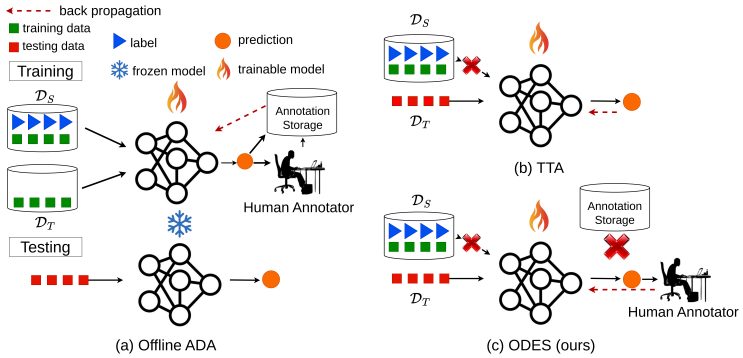

Fig. 1. Illustration of different domain adaptation setups. (a) illustrates the offline Active Domain Adaptation (ADA), where labelled source and unlabelled target domain data are used for training and an annotation storage is required to store the annotation from the active learner which is used later part of training. (b) shows test-time adaptation (TTA) setup. (c) illustrates our proposed setup ODES, where we do not allow any access to source data or any kind of data storage.

target data for self-training to reduce domain shifts. However, UDA assumes access to labeled source data, which is often restricted due to privacy concerns. On the other hand, the need for large unlabeled target data makes the approach tailored to offline adaptation only. In real-world healthcare, data arrives continuously without future patient information (online configuration), rendering offline UDA ineffective.

Offline UDA methods generally rely on self-training through pseudo-label refinement [27, 30]. However, pseudo labels have been shown to be very error-prone, which specifically hurts performance on classes that have limited samples in the training data [25], making them unacceptable for sensitive applications like medical image analysis. To address this issue, we incorporate Active Learning (AL) into the online adaptation strategy instead of solely relying on pseudo labels. AL allows for budgeted manual annotation of samples with the highest prediction uncertainty and is particularly suited for medical image analysis since an expert is usually involved during the image capturing process. While AL in offline UDA, known as Active Domain Adaptation (ADA), improves performance [22, 26], it still requires storing training data and annotations.

Our proposed approach is named **ODES**: Online Domain Adaptation with Expert Guidance for Medical Image Segmentation, an AL guided adaptation setup where we assume medical data arrives to the expert in an online streaming fashion and the expert has the scope of annotating an area [26]. To the best of our knowledge, *this is the first work incorporating AL in domain adaptation for medical image segmentation in an online setup*. In our setup, the model encounters a particular batch of data only once, eliminating the necessity for any data or annotation storage. Our setting is closely related to Test Time Adaptation methods (TTA) [12, 6, 8, 23]. Fig. 1 highlights the distinction between offline ADA, TTA and our setup.

Overview and Contributions. Existing TTA works are primarily motivated for real-time applications like autonomous driving, while ODES focuses on medical image analysis where real-time output is generally not a critical requirement. ODES operates in

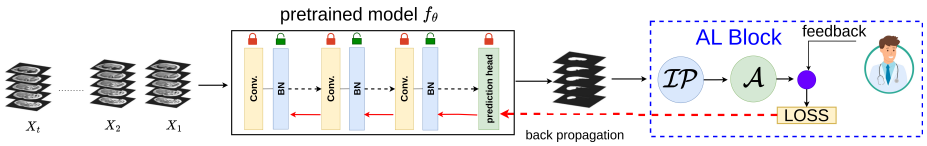

Fig. 2. The pre-trained model encounters a continuous stream of batched data from target domain. The model first predicts the pseudo-labels of the current batch. Following this, the AL block uses Image Pruning ($\mathcal{IP}$) (Sec 2) to prune the test batch and obtains a subset of $K\%$ most informative images in the batch. Next, pseudo labels of the selected images are passed to the Acquisition ($\mathcal{A}$) block (Sec. 2) which selects the most uncertain patches with the budgeted $b\%$ area in each of these images for annotation acquisition from an expert. Next, the Batch Normalization (BN) layers of this model are updated.

four sequential stages for each patient: 1) data collection, 2) inference, 3) acquisition, and 4) model update. After a medical image is collected, a forward pass through a pre-trained model provides the patient the segmentation result. Next uncertainty-guided AL is used to acquire annotations from an expert for the uncertain regions in the inferred result. The small wait time associated with AL is acceptable in the context of medical facilities since there exists a time interval in between imaging sessions of consecutive patients, known as *turnover time* [15]. In our setting, AL is leveraged during this turnover time, making it feasible for any practical medical setup without any disruption in existing workflows. Finally, the expert feedback through AL is used to update the model and deploy it to analyze the medical data of the next patient, and the same cycle is repeated. To further reduce the cost and time associated with annotation acquisition, we propose an innovative image pruning strategy to remove images in a batch with least informative value. Thus, the image pruning technique makes our approach further friendly for the online application setup by reducing both annotation time and the burden on the expert. We also propose a novel acquisition strategy that integrates uncertainty estimation with a diversity-aware sampling mechanism, ensuring that the selected samples are both informative and varied, thereby improving the efficiency of the active learning process.

Related Works. The ODES setup closely aligns with TTA due to their shared focus on online applicability. Existing TTA methods adapt using batch normalization updates [8, 23, 17, 3], teacher-student models [11, 24], generative models [2], and adaptive learning rates [28, 29]. In parallel, ADA methods have emerged to address the limitations of UDA. Approaches such as RIPU [26], LabOR [22] have utilized novel uncertainty guided acquisition and inconsistency masks in their AL strategy. However, all these ADA methods are designed for offline settings and are not directly applicable in online scenarios.

2 Methodology

In the ODES framework initially, a segmentation model f_θ is trained on a set of labeled source data $\mathcal{S}=\{(X_S^i,Y_S^i)\}_{i=1}^{M_s}\sim\mathcal{D}_s$ to segment total C number of classes, where $\mathcal{D}_s$ is the source domain data distribution. As shown in Fig. 2, following the setup of [23] the inference and adaptation process is continuous in nature whereby the model encounters a continuous stream of batches $\mathbf{X}_1 ... \rightarrow \mathbf{X}_t$. A batch can be represented by $\mathbf{X}_t=\{X_\mathcal{T}^j\}_{j=1}^{\mathcal{B}_t}$

where each $X_{\mathcal{T}}^{j} \in \mathcal{T}$ with $\mathcal{T}$ being the target domain having a different data distribution $\mathcal{D}_T$. The entire TTA process follows an infer, acquire, and update policy (Fig. 2).

First, the pre-trained model f_θ is used to infer on a batch $\mathbf{X}_t$ and obtain pseudo labels $\hat{\mathbf{P}}_t = \{\hat{P}_j\}_{j=1}^{\mathcal{B}_t}$. The active learner (expert) provides budgeted annotation based on these pseudo-labels. However, all the images do not exhibit the same amount of domain shift in a test batch. We hypothesize that annotations from the active learner can be efficiently utilized if the annotation is spent on the images with larger domain shifts instead of annotating all the images of the test batch.

Image Pruning. To select the images with larger domain shifts, we leverage the batch-normalization layer (BN) statistics of incoming test batches. BN statistics has been shown to be successful in quantifying domain shift [16]. The statistics of $\mathcal{D}_S$ are stored in the BN layer of the f_θ in terms of running mean and running variance. We compare the feature statistics for each $X_{\mathcal{T}}^{j} \in \mathbf{X}_t$ with the source statistics. With a domain shift, an abrupt change in feature statistics can be visible in terms of KL divergence [7]. Therefore, for each $X_{\mathcal{T}}^{j}$ in the batch we first augment (random horizontal and vertical flips, small rotations, and brightness adjustments) it to obtain $\tilde{X}_{\mathcal{T}}^{j}$.

Assuming the per-channel BN layers to exhibit a Gaussian distribution the divergence between the statistics of f_θ (approximated as $\mathcal{N}(\mu_{l_{c'}}^{\mathcal{S}}, (\sigma_{l_{c'}}^{\mathcal{S}})^2)$ and the BN statistics of $\tilde{X}_{\mathcal{T}}^{j}$ (approximated as $\mathcal{N}(\mu_{l_{c'}}^{\mathcal{T}_j}, (\sigma_{l_{c'}}^{\mathcal{T}_j})^2)$ is defined as

$$D(\mathcal{S}, \tilde{X}_{\mathcal{T}}^{j}) = \sum_{l} \sum_{c'} \mathrm{KL}\Big[\mathcal{N}\Big(\mu_{l_{c'}}^{\mathcal{S}}, (\sigma_{l_{c'}}^{\mathcal{S}})^2\Big), \mathcal{N}\Big(\mu_{l_{c'}}^{\mathcal{T}_j}, (\sigma_{l_{c'}}^{\mathcal{T}_j})^2\Big)\Big] \tag{1}$$

The higher the value of $D(\mathcal{S}, \tilde{X}_{\mathcal{T}}^{j})$ the greater the domain shift. Therefore, we select $K\%$ images from each $\mathbf{X}_t$ with the highest values of $D(\mathcal{S}, \tilde{X}_{\mathcal{T}}^{j})$ and remove the remaining resulting in a pruned batch, $\tilde{\mathbf{X}}_t$, with batch size $\tilde{\mathcal{B}}_t < \mathcal{B}_t$.

Acquisition Function. After obtaining the pruned test batch $\tilde{\mathbf{X}}_t$, the active learner annotates $b\%$ area of each image as square patches based on prediction uncertainty (entropy) [18] and regional impurity [26]. However, these functions lack diversity constraints, leading to redundant annotations of spatially close, similar patches, reducing AL efficacy. To address this, we propose a novel weighting strategy that prioritizes diverse samples in terms of spatial position and feature representation among the most uncertain selections.

a) Uncertainty Estimation. The uncertainty of a patch ($\mathcal{U}$) is defined as the average entropy ($\mathcal{H}$) of the pixels inside it which can be expressed by

$$\mathcal{U}(x,y) = \frac{1}{|A(x,y)|} \sum_{(u,v) \in A(x,y)} \mathcal{H}(u,v) \;\; ; \;\; \mathcal{H}(x,y) = -\sum_{c} \mathbf{P}(x,y,c) \log \mathbf{P}(x,y,c) \tag{2}$$

where $\mathbf{P}$ is the softmax output of the network prediction, (x,y) is the pixel coordinate, and c is the class. We consider a square area of $A(x,y)$ with center at (x,y).

b) Regional Impurity. Regional impurity ($\mathcal{P}$) [26] measures semantic mixing in a region, indicating impurity if multiple instances exist. It is computed for a square region around pixel (x,y) and expressed as:

$$\mathcal{P}(x,y) = -\sum_{c=1}^{C} \frac{|A^c(x,y)|}{|A(x,y)|} \log \frac{|A^c(x,y)|}{|A(x,y)|} \tag{3}$$

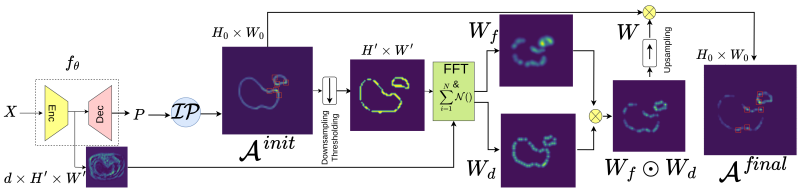

Fig. 3. The figure shows our weighting strategy: the pretrained model f_θ, which is composed of an Encoder (Enc) and Decoder (Dec) processes batch X to predict P. Image Pruning block ($\mathcal{IP}$) samples images with the largest domain shift. Red boxes highlight high-uncertainty regions of each of the sampled images. These patches are initially clustered ($\mathcal{A}^{init}$) closely. FFT sampling combined with Gaussian weighting maximizes their Euclidean and feature-space separation, resulting in a more diverse patch set in the final acquisition map ($\mathcal{A}^{final}$).

where $|A^c(x,y)|$ is the area corresponding to class c in the square region around pixel (x,y) and $|A(x,y)|$ is total area of the square. The initial acquisition map can be defined as $\mathcal{A}^{init}(x,y) = \mathcal{U}(x,y) \odot \mathcal{P}(x,y)$

Diversity Weightmap. Figure 3 illustrates our novel weighting strategy. First we downsample $\mathcal{A}^{init}$ to match the height (H') and width (W') of the d dimensional feature map. Then we obtain the high uncertainty samples by thresholding. We apply the Farthest First Traversal (FFT) algorithm [21] (Algorithm 2) twice to high-uncertainty samples: first using $dist =$ Euclidean distance to select the top-N spatially spread-out samples ($N = H' \times W' \times b\%$), and then using $dist =$ cosine distance to select the top-N feature-wise distant samples. For each of the selected samples, we construct a Gaussian centered at its feature location, and then aggregate all these Gaussians to form a combined diversity weighting map. For the Euclidean distance case, $W_d = \sum_{i=1}^{N} \mathcal{N}(\mu_{d_i}, \sigma^2_{d_i})$ where μ_{d_i} is a location of the sampled point and σ_{d_i} a hyperparameter which controls the spread of the Gaussian curve. Similarly for the feature distance case we obtain $W_f = \sum_{i=1}^{N} \mathcal{N}(\mu_{f_i}, \sigma^2_{f_i})$. Both of the W_d and W_f are normalized. The combined Gaussian weight map can be expressed by $W = upsample(W_d \odot W_f)$. W highlights the areas in $\mathcal{A}_{init}$ where both the Euclidean and feature distances are maximized for highly uncertain samples hence ensuring diversity which makes the AL procedure more effective. The final acquisition function is given as:

$$\mathcal{A}_j^{final}(x,y) = \mathcal{U}_j(x,y) \odot \mathcal{P}_j(x,y) \odot W_j(x,y) \text{ where } j \in \{1,2,...,\tilde{\mathcal{B}}_t\} \tag{4}$$

$$\mathcal{Q}^* = \{A(x^*,y^*)_j\}_{j=1}^{\tilde{\mathcal{B}}_t}, \text{ where } (x^*,y^*)_j = \left\{ \text{Top } \frac{|A_{img}|b\%}{|A(x,y)|} \text{ of } \underset{(x,y)\in R^2}{\text{argmax}} \mathcal{A}_j^{final}(x,y) \right\} \tag{5}$$

Here, the set $\mathcal{Q}^*$ comprises the top $b\%$ most informative patches selected for annotation from the pruned set of images. The term $|A(x,y)|$ represents the area of a single square patch, while $|A_{\text{img}}|$ denotes the total area of the image.

Algorithm 1 ODES
Image Selection Rate $K\%$ per batch, AL budget $b\%$ pixel per image

Require: Source pre-trained model f_θ
Compute BN statistics $\left\{(\mu^{\mathcal{S}}_{l_{c'}},(\sigma^{\mathcal{S}}_{l_{c'}})^2)\right\}$ for f_θ.
for each batch $t=1,2,...$ **do**
  Initialize divergence array DIV$=\{\}$
  for each image $j \in \left\{X^j_{\mathcal{T}}\right\}_{j=1}^{\mathcal{B}_t}$ **do**
    $X^j_{\mathcal{T}} \leftarrow$ Augmentation
    Compute BN statistics $\left\{(\mu^{\mathcal{T}_j}_{l_{c'}},(\sigma^{\mathcal{T}_j}_{l_{c'}})^2)\right\}$ for augmented $X^j_{\mathcal{T}}$
    Compute D using Eq. 1
    Store D in DIV
  end for
  $\tilde{\mathbf{X}}_t \leftarrow$ Select top-$K\%$ images from DIV
  Perform patch sampling on $\tilde{\mathbf{X}}_t$ (budget $b\%$) using Eq. 2,3, 4, 5.
  Update model using Eq. 6.
end for

Algorithm 2 Farthest-First Traversal (FFT)

Require: Set of high uncertainty $\mathcal{X}=\{x_1,x_2,...,x_n\}$, Number of points to sample N
Initialize $\mathcal{S} \leftarrow \{x_i\}$, where x_i has the highest value in $\mathcal{A}_{init}$
for $t=2$ to N **do**
  Compute distance $dist(x, \mathcal{S}) = \min_{s \in \mathcal{S}} dist(x,s), \quad \forall x \in \mathcal{X} \setminus \mathcal{S}$
  Select $x^* = \arg\max_{x \in \mathcal{X} \setminus \mathcal{S}} d(x,\mathcal{S})$
  Update $\mathcal{S} \leftarrow \mathcal{S} \cup \{x^*\}$
end for
return $\mathcal{S}$

Adapting the Model. After selecting $b\%$ area from each $\tilde{\mathbf{X}}_t$, their labels are obtained from the expert. f_θ is then adapted to $\mathcal{T}$ by updating BN layers using a supervised cross-entropy loss ($\mathcal{L}_{sup}$) and unsupervised continuity loss ($\mathcal{L}_{cont}$). In MRI, a batch consists of 2D images forming a 3D volumetric stack with smooth transitions between adjacent slices. Based on this, we define $\mathcal{L}_{cont}$ to minimize abrupt changes between successive slices. If $\mathbf{Y}(x,y,c)$ is the label assigned by the active learner at pixel (x,y) for class c and CE is the cross-entropy loss, the total loss can be expressed by

$$\begin{aligned} \mathcal{L}_{\text{sup}} &= -\frac{1}{|\mathcal{Q}^*|} \sum_{(x,y)\in\mathcal{Q}^*} \sum_{c=1}^{C} \mathbf{Y}(x,y,c) \log \mathbf{P}(x,y,c); \\ \mathcal{L}_{\text{cont}} &= \sum_{j=1}^{\mathcal{B}-1} CE(\hat{\mathbf{P}}^j, \hat{\mathbf{P}}^{j+1}); \quad \mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{cont}} \end{aligned} \tag{6}$$

The overall approach of ODES is shown in Algorithm 1.

3 Experiments and Results

Dataset and Adaptations. We use 4 datasets in our experiments, namely CHAOS [10], DUKE [14], BMC [13], and RUNMC [13]. Using these 4 datasets, we set up 3 adaptation scenarios: (a) T1-DUAL MRI of CHAOS in-phase (IP) $\rightarrow$ out-of-phase (OOP) CHAOS, where IP and OOP have been shown to exhibit domain shifts [8, 19]. (b) CHAOS T2-SPIR $\rightarrow$ DUKE. (c) BMC $\rightarrow$ RUNMC. The main classes for

Table 1. Comparison among different TTA Methods reported in terms of DSC. The best results are highlighted in red and the second best in blue. Here we consider b=1.

	CHAOS T1 (IP → OOP)					CHAOS → DUKE	BMC → RUNMC
Methods	Liver	L.Kidney	R.Kidney	Spleen	Mean	Liver	Prostate
Source only	87.77	37.97	18.92	67.31	52.99	26.28	65.32
TENT [23]	86.03±0.2	58.1±0.2	54.72±0.1	70.34±0.1	67.3±0.1	46.64±1.5	71.02±0.2
CoTTA [24]	86.38±0.1	52.8±0.1	58.27±0.1	71.06±0.2	67.1±0.1	52.58±0.8	73.87±0.2
F-TTA [8]	86.91±0.17	61.99±0.11	62.11±0.2	69.51±0.2	70.13±0.2	48.29±0.7	73.18±0.2
STTA [11]	85.91±1.3	51.64±1.8	58.44±1.2	70.71±1.3	67.13±1.4	53.36±1.4	73.51±0.7
SaTTCA [12]	87.90±0.1	64.85±0.1	65.47±0.1	74.51±0.1	73.19±0.1	58.39±0.7	74.91±0.3
TTAS [3]	86.74±0.2	61.68±0.2	58.13±0.2	71.07±0.2	69.41±0.2	49.39±1.2	72.81±0.2
ODES							
K = 100	88.90±0.1	72.41±0.1	74.55±0.1	78.56±0.1	78.61±0.1	71.65±0.2	79.34±0.1
K = 50	89.13±0.05	72.30±0.1	74.43±0.1	78.36±0.1	78.56±0.1	70.16±0.3	78.36±0.1
K = 10	88.41±0.1	71.45±0.1	70.22±0.1	77.12±0.1	76.80±0.1	66.11±0.2	77.47±0.1

segmentation in each experiment are (a) the liver, left kidney, right kidney and spleen, (b) liver and (c) prostate, respectively. We use Deeplabv3 [4] for the experiment with annotation budget 1%(b=1), and quantify performance using the Dice Score (DSC).

Table 2. Comparison with Offline ADA Method under the same annotation budget (b=1). The offline ADA method considers multiple forward passes during offline training, whereas ODES considers only one forward pass during adaptation.

	CHAOS T1 (IP → OOP)					CHAOS → DUKE	BMC → RUNMC
Methods	Liver	L.Kidney	R.Kidney	Spleen	Mean	Liver	Prostate
RIPU [26]	93.52±0.1	**87.42±0.1**	**86.71±0.2**	**86.95±0.1**	**88.65±0.1**	**80.12±0.2**	**89.01±0.1**
RIPU-SF [26]	**93.54±0.2**	85.69±0.1	84.29±0.3	86.69±0.2	87.55±0.2	78.48±0.1	88.16±0.1
LabOR [22]	92.14±0.1	86.45±0.1	84.41±0.1	84.76±0.1	86.94±0.1	78.77±0.2	86.95±0.2
ODES							
K = 100	91.16±0.1	81.38±0.2	83.32±0.1	80.72±0.1	84.15±0.1	77.71±0.3	84.29±0.2
K = 50	91.21±0.1	81.41±0.1	82.56±0.2	80.40±0.1	83.65±0.1	77.39±0.3	82.90±0.1
K = 10	91.15±0.2	82.05±0.1	79.12±0.1	80.51±0.1	83.2±0.1	76.01±0.2	79.62±0.2

3.1 Comparison with other methods

Comparison with online adaptation methods. As our problem setting is closely related to TTA, our baselines are some widely used state-of-the-art TTA methods. From Table 1 we observe that our method has outperformed all other TTA methods in all three adaptations. Adding very minimal annotation can significantly boost performance, specially in CHAOS → DUKE where wider range of variations (four distinct forms of contrasts) is seen in target domain. We observe that reducing annotations by 50% (K=100 to 50) and 90% (K=100 to 10) leads to only a 1.49% and 5.54% performance

Fig. 4. Visual comparison with different types of adaptations

drop, respectively, in CHAOS → DUKE, demonstrating a small performance loss despite a significant reduction in annotation cost. Same pattern is true for other adaptations too.

Comparison with offline adaptation methods. We compare ODES with ADA methods RIPU [26] and LabOR [22], considering two RIPU variants: standard (with source) and RIPU-SF (without source). As RIPU and LabOR are offline methods, they require a training set. So the target domain data is split into 80% training and 20% testing. In this experiment, ODES applies active learning only to the training split used in RIPU, with no AL in the test split. For a fair comparison, all results are reported on the test split in Table 2. Despite being an online method with a single pass constraint, ODES performs close to the offline ADA methods which are allowed to do multiple passes. For example, we observe performance gap of 2.79%,1.06%,2.66% from LabOR. We demonstrate some visual comparisons in Fig 4.

3.2 Ablation Studies

Impact of Proposed Image Pruning and Diversity Weighting. ODES proposed a novel image pruning strategy and diversity weighting strategy. In Table 3, we have shown different combinations of image pruning and diversity weighting to understand their impact. We observe the best performance was obtained when both strategies were involved.

Table 3. Image Pruning (I.P.) and Diversity Weighting (D.W.) with $K = 10$. I.P. not involved means K% were selected randomly from the batch instead of our proposed method.

I.P.	D.W.	IP → OOP	CHAOS → DUKE
✗	✗	74.32	63.59
✓	✗	75.88	65.46
✗	✓	75.41	64.77
✓	✓	**76.80**	**66.11**

Forgetting Analysis. We employ a cyclic evaluation to analyze whether catastrophic forgetting occurs in our approach. After one cycle of adaptation of all batches, we repeat the process with the adapted model. In Table 4, we observe increased DSC on the same batch of images. We observe that performance has enhanced in the second and third cycles of evaluation. If there were catastrophic forgetting, the performance would have degraded. The enhanced DSC indicates that catastrophic forgetting did not occur.

Table 4. Catastrophic Forgetting analysis. Mean DSC is reported for IP → OOP for each batch individually for multiple cycles.

	Incoming batch id →					
cycle	1	2	3	...	...	20
1	67.71	82.37	79.91	...	...	82.11
2	77.51	87.37	87.24	...	...	88.26
3	80.57	88.29	89.07	...	...	89.63

4 Conclusion

We introduce ODES, a novel framework for online domain adaptation in medical image segmentation that uses active learning on a streaming data. In order to reduce the annotation burden of the active learner, ODES utilizes a unique image-pruning strategy which not only mitigates the challenge of domain shift, but also makes the application more online-friendly. Also the diversity weighting strategy boosts the performance of AL. Through extensive experimentation, ODES has shown superior performance over existing TTA methods and also reaches close to the performance of offline adaptation.

Acknowledgments. This work was supported by NSF grants DMS 2029814 and OAC 2411453.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Asgari Taghanaki, S., Abhishek, K., Cohen, J.P., Cohen-Adad, J., Hamarneh, G.: Deep semantic segmentation of natural and medical images: a review. Artificial Intelligence Review **54**, 137–178 (2021)
2. Basak, H., Yin, Z.: Quest for clone: Test-time domain adaptation for medical image segmentation by searching the closest clone in latent space. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 555–566. Springer (2024)
3. Bateson, M., Lombaert, H., Ben Ayed, I.: Test-time adaptation with shape moments for image segmentation. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 736–745. Springer Nature Switzerland, Cham (2022)
4. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI) (2017)
5. Full, P.M., Isensee, F., Jäger, P.F., Maier-Hein, K.: Studying robustness of semantic segmentation under domain shift in cardiac mri. In: Statistical Atlases and Computational Models of the Heart. M&Ms and EMIDEC Challenges: 11th International Workshop, STACOM 2020, Held in Conjunction with MICCAI 2020. pp. 238–249. Springer (2021)
6. He, Y., Carass, A., Zuo, L., Dewey, B.E., Prince, J.L.: Autoencoder based self-supervised test-time adaptation for medical image analysis. Medical image analysis **72**, 102136 (2021)
7. Hershey, J.R., Olsen, P.A.: Approximating the kullback leibler divergence between gaussian mixture models. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07. vol. 4, pp. IV–317. IEEE (2007)

8. Hu, M., Song, T., Gu, Y., Luo, X., Chen, J., Chen, Y., Zhang, Y., Zhang, S.: Fully test-time adaptation for image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 251–260. Springer (2021)
9. Ibtehaz, N., Rahman, M.S.: Multiresunet: Rethinking the u-net architecture for multimodal biomedical image segmentation. Neural networks **121**, 74–87 (2020)
10. Kavur, A.E., Gezer, N.S., Barış, M., Aslan, S., Conze, P.H., Groza, V., Pham, D.D., Chatterjee, S., Ernst, P., Özkan, S., et al.: Chaos challenge-combined (ct-mr) healthy abdominal organ segmentation. Medical Image Analysis **69**, 101950 (2021)
11. Li, X., Fang, H., Wang, C., Liu, M., Duan, L., Xu, Y.: Cache-driven spatial test-time adaptation for cross-modality medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 146–156. Springer (2024)
12. Li, Z., Yang, J., Xu, Y., Zhang, L., Dong, W., Du, B.: Scale-aware test-time click adaptation for pulmonary nodule and mass segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 681–691. Springer (2023)
13. Liu, Q., Dou, Q., Yu, L., Heng, P.A.: Ms-net: multi-site network for improving prostate segmentation with heterogeneous mri data. IEEE transactions on medical imaging **39**(9), 2713–2724 (2020)
14. Macdonald, J.A., Zhu, Z., Konkel, B., Mazurowski, M., Wiggins, W., Bashir, M.: Duke liver dataset (mri) (Mar 2022). https://doi.org/10.5281/zenodo.6328447, `https://doi.org/10.5281/zenodo.6328447`
15. Mazzei, W.J.: Operating room start times and turnover times in a university hospital. Journal of clinical anesthesia **6**(5), 405–408 (1994)
16. Niloy, F.F., Ahmed, S.M., Raychaudhuri, D.S., Oymak, S., Roy-Chowdhury, A.K.: Effective restoration of source knowledge in continual test time adaptation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2091–2100 (2024)
17. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: International conference on machine learning. pp. 16888–16905. PMLR (2022)
18. Ozdemir, F., Peng, Z., Tanner, C., Fuernstahl, P., Goksel, O.: Active learning for segmentation by optimizing content information for maximal entropy. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. pp. 183–191. Springer (2018)
19. Ramalho, M., Herédia, V., de Campos, R.O., Dale, B.M., Azevedo, R.M., Semelka, R.C.: In-phase and out-of-phase gradient-echo imaging in abdominal studies: intra-individual comparison of three different techniques. Acta Radiologica **53**(4), 441–449 (2012)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. pp. 234–241. Springer (2015)
21. Rosenkrantz, D.J., Stearns, R.E., Lewis, II, P.M.: An analysis of several heuristics for the traveling salesman problem. SIAM journal on computing **6**(3), 563–581 (1977)
22. Shin, I., Kim, D.J., Cho, J.W., Woo, S., Park, K., Kweon, I.S.: Labor: Labeling only if required for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8588–8598 (2021)
23. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
24. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7201–7211 (2022)

25. Wei, C., Sohn, K., Mellina, C., Yuille, A., Yang, F.: Crest: A class-rebalancing self-training framework for imbalanced semi-supervised learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10857–10866 (2021)
26. Xie, B., Yuan, L., Li, S., Liu, C.H., Cheng, X.: Towards fewer annotations: Active learning via region impurity and prediction uncertainty for domain adaptive semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8068–8078 (2022)
27. Yang, C., Guo, X., Chen, Z., Yuan, Y.: Source free domain adaptation for medical image segmentation with fourier style mining. Medical Image Analysis **79**, 102457 (2022)
28. Yang, H., Chen, C., Jiang, M., Liu, Q., Cao, J., Heng, P.A., Dou, Q.: Dltta: Dynamic learning rate for test-time adaptation on cross-domain medical images. IEEE Transactions on Medical Imaging **41**(12), 3575–3586 (2022)
29. Zhang, Y., Huang, K., Chen, C., Chen, Q., Heng, P.A.: Satta: Semantic-aware test-time adaptation for cross-domain medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 148–158. Springer (2023)
30. Zhao, Z., Zhou, F., Xu, K., Zeng, Z., Guan, C., Zhou, S.K.: Le-uda: Label-efficient unsupervised domain adaptation for medical image segmentation. IEEE Transactions on Medical Imaging **42**(3), 633–646 (2023)