Efficient Deformable Convolutional Prompt for Continual Test-Time Adaptation in Medical Image Segmentation

Shiyu Liu[1], **Daoqiang Zhang**[2], **Xiaoke Hao**[1]

[1]School of Artificial Intelligence, Hebei University of Technology, Tianjin, 300401, China
[2]College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, 211106, China
202222802030@stu.hebut.edu.cn, dqzhang@nuaa.edu.cn, haoxiaoke@scse.hebut.edu.cn

Abstract

The domain gap resulting from mismatches in acquisition details like protocol and scanner between training and test data hinders the deployment of the trained model in clinical practice. To address this issue, Continual test-time adaptation (CTTA) has been proposed to adapt the source model to continually changing unlabeled domains without accessing the source data. Existing methods learn an image-level visual prompt for target domains and inject the trainable prompt into the input space. However, they either combine the input with a prompt of equal scale or determine the prompt injection position through complex strategies such as uncertainty estimation or Fourier Transform. These approaches substantially increase the number of trainable parameters and computational burden, especially in high-dimensional medical imaging data. To overcome these challenges, we propose the Efficient Deformable Convolutional Prompt (EDCP), which leverages the inductive bias of convolution to reduce trainable parameters compared to standard prompts. We further enhance convolution by making it deformable, addressing fine-grained domain shifts at the pixel level through an offset branch. To improve training efficiency and balance parameters between the convolution and offset branches, we decompose the offset transformation into two parts, storing one in an offset bank that also serves as a domain indicator. This bank accelerates training by skipping test images similar to those already stored. Prompt updates are guided by layer-wise alignment of source-target statistics without unfreezing batch normalization layers. Extensive experiments demonstrate the superiority of our method in 2D and 3D medical image segmentation tasks.

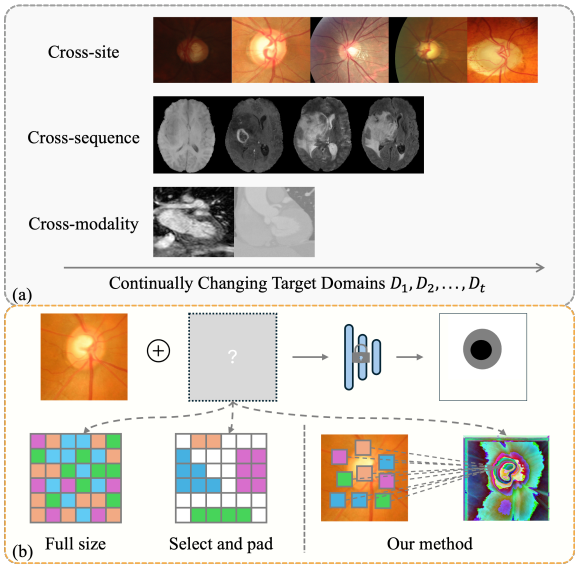

Figure 1: (a) Illustrations of three different domain shifts in medical images. (b) Our goal is to adapt the source model to continually changing domains in the input space by injecting visual prompts. Existing prompt-based methods either inject the prompt at the same scale as the input or use complex strategies to select the injection position, padding the rest with constant values (white square). In contrast, we leverage the inherent advantages of convolutional inductive bias and extend it to be deformable, significantly reducing the trainable parameters in the prompt.

Introduction

Semantic segmentation is the fundamental step toward anatomical structure recognition in medical image analysis. The segmentation model can achieve remarkable performance when the model is trained with a supervised signal and tested with data of consistent distribution. However, the model suffers from performance degradation when there exists a domain gap between the source and target data distribution. Unfortunately, this domain gap commonly exists in real scenarios, as the weather changes from sunny to rainy to foggy, but it is more severe in medical imaging. For example, as shown in Fig. 1(a), the same tissue (i.e., fundus) is acquired with different imaging scanners and setups across the medical centers. Besides, magnetic resonance imaging (MRI) is a widely adopted radiation-free imaging technique and also bears heterogeneity in cross-sequence (e.g., the contrast varies for the same brain tumor in T1, T1ce, T2, and FLAIR sequence). Moreover, computed tomography (CT) and MRI provide significantly different signals for the same region (i.e., heart) due to their different imaging principles.

To address this issue, a line of work focuses on unsupervised domain adaptation (UDA), which typically requires

access to entire labeled source domain data and large-scale unlabeled target domain data to perform distribution alignment. However, this setting is problematic in clinical practice for two reasons. Firstly, accessing source data at test time is difficult, because data sharing across hospitals is prohibitive due to the privacy concern. Secondly, UDA aligns the source and target domain batch by batch and repeats several epochs before the inference phase. However, due to the scarcity of medical images, the test data usually arrive one by one, making it very time-consuming to assemble a batch, not to mention that the inference stage requires training through multiple epochs. Moreover, in clinical practice, there is a strong need for immediate predictions on a single test sample to facilitate timely diagnosis and treatment for patients. Therefore, it would be more convenient if a source pre-trained model could quickly and continuously adapt to each test sample at inference time, without requiring the source data.

This motivates another line of work, test-time adaptation (TTA), which deals with the above issues by exploring the distributional information gradually from each target domain test sample without using source data. TENT (Wang et al. 2021) proposes to update only the affine parameters in batch normalization (BN) layers by minimizing the entropy of model predictions on the test data. MABN (Wu et al. 2024) constructs an auxiliary branch to perform self-supervised learning (SSL) and meta-learning to supervise BN updates. However, these TTA methods require the target domains to be stationary and neglect the continually changing target domains in most real-world scenarios. CoTTA (Wang et al. 2022)first proposes a continual test-time adaptation (CTTA) approach to tackle the above issues, specifically, CoTTA refines the unreliable pseudo label by using weight-averaged and augmentation-averaged prediction and preserves the source knowledge by stochastically restoring a small number of neurons. VDP (Gan et al. 2023) proposes to learn an image-level visual prompt while having the source model parameters frozen, eliminating the update of the entire model only the prompt injected into the input space is learnable. However, it still needs to update the network with an exponential moving average (EMA) between the student and teacher model and deploy SSL tasks. VPTTA (Chen et al. 2024) proposes to inject the prompt into the low-frequency of the input which is associated with style texture, but it needs to perform the Fast Fourier Transform (FFT) first to inject the prompt and resume the input by the inverse FFT. They both suffer from complicated strategies for prompt updating and injection.

To tackle the above problems, we first introduce the concept of our Efficient Deformable Convolutional Prompt (EDCP), as shown in Fig. 1(b), instead of adopting the prompt having the same scale of input or determining the prompt injection position by performing complicated SSL tasks or computational intensive transformation of input, we leverage the inherent advantages of inductive bias in convolution, only a few trainable parameters can provide a powerful prompt. We further extend its representative capability in low-parameter conditions by attaching each position of the parameter with an offset which is predicted by the offset branch, enabling the convolution kernel to be deformable. We then propose a new framework for the efficient training of our EDCP. Specifically, we observe a significant parameter imbalance, with several times more parameters in the offset branch compared to the convolution. We design a parameter balance strategy, offset transformation decompose, which decreases the number of parameters needed for offset prediction and stabilizes the training of EDCP. Furthermore, the supervision for the training of the prompt is provided by the structure similarity between the adapted image and augmented image and layer-wise source-target statistics alignment without unfreezing the source pre-trained model. Moreover, we designed an offset bank to store the transformation matrices from the second part of the decomposition in the offset branch, which can be considered an implicit domain indicator. By comparing the transformation matrix of the current iteration with those in the offset bank, we can skip training on samples with high similarity, significantly speeding up the training process while maintaining its representational capability, an approach we refer to as fast inference. Overall, our method is both parameter-efficient and training-efficient for adapting the frozen source model in the input space. Our contributions are summarized as follows:

- We propose a parameter-efficient visual prompt, EDCP, to address domain shift by using offsets as implicit domain indicators. We can obtain representative prompts for input space adaptation without requiring complex prompt selection strategies.
- We present a novel prompt training acceleration strategy by skipping training samples within similar domains, which significantly reduces the training time for CTTA.
- Our approach outperforms most state-of-the-art methods on extensive experiments, covering both 2D and 3D medical segmentation tasks. It proves that our approach is practical for both performance and inference speed.

Related Work

Test-time Adaptation

Test-time adaptation (TTA) aims to fine-tune the source pre-trained model by the target data in the test phase and it can be divided into two main categories, model-based and parameter-free. For the model-based method, most of them align the distribution by updating the BN layers in the source model. Tent (Wang et al. 2021) updates the affine parameters in each BN layer by minimizing the entropy of the prediction. MABN (Wu et al. 2024) obtains supervision by enforcing the alignment of the auxiliary and main branches via meta-learning. SANTA (Chakrabarty, Sreenivas, and Biswas 2023) modifies the affine parameters using source anchoring based self-distillation. TTN (Lim et al. 2023) presents a new normalization technique that interpolates the standardization statistics by adjusting the importance of BN and transductive BN. The other is the parameters-free method. (Valanarasu et al. 2024) proposes to equip each convolutional block with an adaptive BN to adapt the features with respect to a domain code which is prior knowledge of medical images. InTEnt (Dong et al. 2024) integrates predictions made with various

estimates of target domain statistics and weights them with their entropy.

Continual test-time adaptation (CTTA) is proposed to address the performance degradation of TTA under the setting of continuously changing target domains. CoTTA (Wang et al. 2022) first proposes to address the sequence of shifts in TTA and refine the pseudo labels by weight average and augmentation average. To avoid catastrophic forgetting, it randomly restores a small portion of neurons in each iteration. DAT (Ni et al. 2024) tackles the changing domains by selecting and updating domain-specific parameters and task-relevant parameters in the model according to the test data distribution. DLTTA (Yang et al. 2022) presents a novel dynamic learning rate adjustment approach to deal with a variety of scale distribution shifts in the sequence. However, we try to solve this problem at the input level to minimize the computation overhead on the updating of each BN layer.

Prompt Learning

Prompt learning is first proposed in natural language processing (NLP) to devise additional instructions for input text to fine-tune the large language model to various downstream tasks. Inspired by its success in NLP, several methods have involved it in computer vision. SVDP (Yang et al. 2024) proposes a novel approach to adaptively allocate trainable parameters on the pixels and designs the domain prompt updating strategy to optimize the prompt differently. CVP (Tsai, Mao, and Yang 2024) first introduces convolution kernel to visual prompt, which demands fewer trainable parameters compared to conventional prompt, facilitating the lightweight input space adaptation. VDP (Gan et al. 2023) devises domains-specific prompts and domain-agnostic prompts and further optimizes them with a homeostasis-based adaptation strategy, significantly relieving catastrophic forgetting and error accumulation. VPTTA (Chen et al. 2024) presents a low-frequency prompt for each test image to align the statistics in the BN layers. However, the above methods either involve complex strategies to obtain prompts and their injection locations or are only effective in TTA settings, ignoring continuously changing domains. In contrast, our proposed EDCP is both parameter-efficient and training-efficient in CTTA settings, requiring no complex strategies for prompt acquisition and significantly accelerating the prompt training phase.

Method

Problem Definition

Given a model f_θ trained on the source domain and to be tested on continual changing unlabeled target domains $D_1, D_2, ..., D_t$,where $D_j = (x_i^t)_{i=1}^{N_j}$, and N_j is the scale of the target domain. Our goal is to continuously adjust the incoming data of new distribution online to adapt to the source model. Considering the clinical practice, where data arrives at the subject level and timely prediction results are required, we assume a batch size of 1 for all experiments.

Our approach has three components: a deformable vision prompt, a source-target alignment strategy for updating the prompt, and a fast inference method to speed up the training phase. The overview of our method is illustrated in Fig. 2.

Deformable Convolutional Prompt

Design of the prompt First, we designed a convolution prompt by directly applying convolution to the input. This approach leverages the inductive bias of convolution, with the only trainable parameters being the convolution kernels. However, we found that under the CTTA setting, using convolution alone is insufficient to handle significant domain shifts. Domain shifts are reflected not only in pixel value changes but also in offsets in various directions (e.g., the contours and positions of objects undergo substantial changes, as shown in Figure 1(a)). Inspired by (Dai et al. 2017), we propose adding an additional branch to provide offsets for convolution sampling locations. Therefore, our deformable convolution prompt addresses domain shifts at the pixel level by accounting for both value changes and offsets. More specifically, the input of the source pre-trained model can be formalized as

$$\widetilde{X_i^t} = X_i^t + \lambda \mathrm{Conv}(X_i^t, \mathrm{Offset}(X_i^t)) \tag{1}$$

where $X_i^t \in \mathbb{R}^{C\times H\times W}$ is the original unlabeled input, $\widetilde{X_\iota^t}$ is the adapted input after prompt injection, $\mathrm{Conv}(\cdot,\cdot)$ denotes the convolution with input and sampling offset, $\mathrm{Offset}(\cdot)$ is the offset encoder, which transforms from $\mathbb{R}^{C\times H\times W}$ to $\mathbb{R}^{2\times K\times K\times H\times W}$, providing offsets in the x and y directions for the convolution, where 2 represents the offset directions and K is the kernel size.

Transformation decomposition However, we find that the convolution branch and offset branch suffer from severe parameter imbalance; specifically, the offset encoder has many times more parameters than the convolution, which severely hinders prompt training on a single sample. To address this issue, we decompose the transformation into two parts,

$$\mathrm{Offset}(\cdot) = f_d \odot f_g \tag{2}$$

where $f_d : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^{C\times H\times W}$ represents the group convolution process applied to the input independently of the channels, and $f_g : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^{2\times K\times K\times H\times W}$ denotes the layer-wise fusion within the feature map. After this decomposition, we achieve a balanced parameter proportion between the two branches. The parameters of f_g, denoted as $\mathcal{A}_g \in \mathbb{R}^{2\times C\times K\times K}$, are considered as implicit domain indicators. We further extend this to an offset bank to enable efficient training of our EDCP in the fast inference section.

Source-target Alignment

Input alignment Instead of aligning predictions at the output level, which requires a forward pass through the model, we align distribution statistics at the input level. Specifically, we augment the original image and enforce structural consistency between the adapted augmented input and the adapted original input:

$$\mathcal{L}_{in}(h(\tilde{X}_i^t), \tilde{X}_i^t) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)} \tag{3}$$

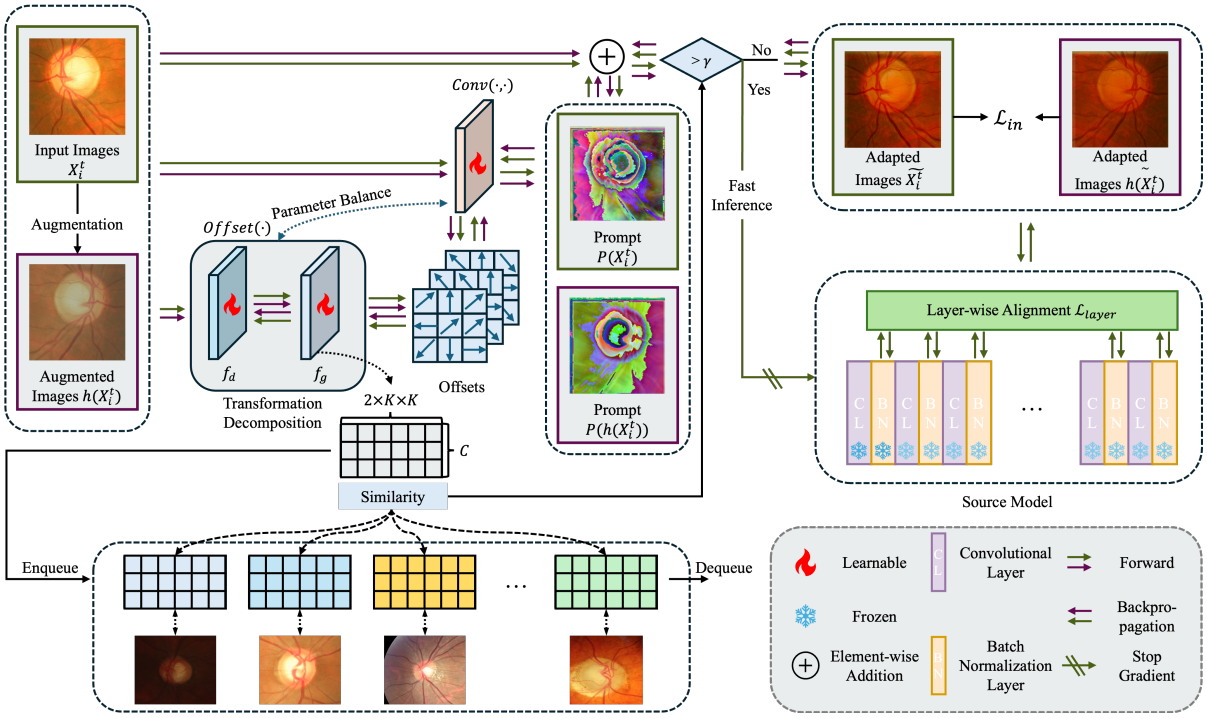

Figure 2: Overview of our EDCP. The inputs are original image and image with augmentation. First, we address the parameter imbalance between the two branches by decomposing the transformation into two stages. Next, we store the transformation matrix from the second stage in an offset bank, where each transformation matrix is considered an implicit domain indicator. We obtain prompts by combining offsets with convolution, and these prompts are then added to the original images to produce the adapted image. We determine whether to train on the current sample by measuring the similarity between the current transformation matrix and those in the offset bank: if the similarity is below a threshold γ, we perform source-target alignment at both the input level and layer-wise. Otherwise, we skip the training phase for the current sample and proceed directly to inference, without performing gradient backpropagation.

where $\tilde{X}_i^t$ is the adapted original input, and $h(\cdot)$ represents random augmentation. μ_x and μ_y are the means of the two inputs, σ_x and σ_y are their variances,$\sigma_{x,y}$ is the covariance, and c_1 and c_2 are constants. This method helps bridge the gap between related domains, which is the fundamental goal of prompt learning.

Layer-wise alignment The distribution shift largely exists in the BN statistics between the source and target samples. Inspired by (Chen et al. 2024), we propose guiding prompt updates by measuring the differences between the BN statistics of the target samples and those stored in each layer of the source model:

$$\mathcal{L}_{layer} = \frac{1}{j}\sum_{j=1}^{N} |\mu_s^j - \mu_t^j| + |\sigma_s^j - \sigma_t^j| \tag{4}$$

where μ_s, σ_s are the mean and variance from the source domain, μ_t, σ_t are from the target domain. j denotes the j-th BN layer in the source model, and N represents the total number of BN layers. The overall objective function $\mathcal{L}L$ is formulated as:

$$\mathcal{L} = \mathcal{L}_{in} + \mathcal{L}_{layer} \tag{5}$$

Fast Inference

In the CTTA setting, the model adapts for each input; however, not all updates are necessary. Based on our observations, the extent of prompt updates decreases within the same domain but increases when switching domains. We hypothesize that updates are only beneficial when there is a significant domain change. Therefore, we design an offset bank composed of the historical updates of $\mathcal{A}_g$ in Eq. (2). We measure domain change by estimating point-wise offset similarity with each $\mathcal{A}$ in the offset bank B:

$$Sim(\mathcal{A}_g, B) = \frac{1}{S \times K^2}\sum_{i=1}^{S}\sum_{j=1}^{K^2} \frac{\mathcal{A}_g^j \cdot \mathcal{A}_i^j}{\|\mathcal{A}_g^j\| \cdot \|\mathcal{A}_i^j\|} \tag{6}$$

where i represents the i-th $\mathcal{A}$ in the offset bank, j denotes the offset position, S denotes bank size, and K^2 denotes the total number of points to be measured. $\mathcal{A}_g^j \in \mathcal{R}^{2\times C}$ denotes the offset vector in the current iteration for kernel position j. When the similarity exceeds the threshold γ, we alternate the training process with direct inference by removing the augmentation and the calculations in Eq. (3) and Eq. (4). By ignoring similar samples within the domain, we significantly reduce training time and also prevent prompt overfitting to

Algorithm 1: Training of EDCP

Require: Fozen source model f_θ, deformable convolution prompt f_d, offset bank B, bank size S, similarity threshold γ
Input: For each time step i, current test image X_i^t
Output: Segmentation result M_i

Obtain $h(X_i^t)$ by augmenting X_i^t
Obtain adapted input $\tilde{X}_i^t$ and $h(\tilde{X}_i^t)$ via their prompts by Eq. (1)
Calculate offset similarity Sim with members in offset bank B by Eq. (6)
if $Sim < \gamma$ or $len(B) < S$ **then**
    // Train
    Calculate $\mathcal{L}_{in}$ by Eq. (3)
    Only forward unaugmented input $f_\theta(\tilde{X}_i^t)$ and calculate layer-wise alignment loss $\mathcal{L}_{layer}$ by Eq. (4)
    Backward and update f_d
else
    // Bypass the train
end if
// Inference
Obtain $\tilde{X}_i^t$ using updated prompt f_d
Obtain segmentation result by $M_i = f_\theta(\tilde{X}_i^t)$
Update the offset bank B
return M_i

the same domain. The overall process of EDCP is summarized in Algorithm 1.

Experiments

Datasets and Evaluation Metrics

The OD/OC segmentation dataset. The dataset contains five public fundus datasets collected from different medical centers with inconsistent statistical characteristics, denoted as domain A (RIM-ONE-r3) (Fumero et al. 2011), B (REFUGE) (Orlando et al. 2020), C (ORIGA) (Zhang et al. 2010), D (REFUGE-Validation) (Orlando et al. 2020), and E (Drishti-GS) (Sivaswamy et al. 2014). There are 159, 400, 650, 800, and 101 RGB images from these datasets. Following (Chen et al. 2024), we crop the region centering at OD with the size of 800x800 and further resize it to 512x512. The Dice score metric (DSC) is used for evaluation.

The brain tumor segmentation dataset. We evaluate our method with the widely used multimodal Brain Tumor Segmentation (BraTS 2020) dataset (Bakas et al. 2019), which comprises multi-contrast MRI exams with four sequences: T1, T1ce, T2, and FLAIR. Following the challenge, we segment three tumor regions for evaluation: (1) whole tumor (WT), including all tumor tissues, (2) tumor core (TC), composed of the enhancing tumor, necrotic, and non-enhancing tumor core, and (3) enhancing tumor (ET). The BraTS 2020 includes 369 training cases, each case containing four sequences. We use DSC for performance quantification.

The hole heart segmentation dataset. We use the 40 independent scans (20 CT and 20 MRI) of cardiac regions from the Multi-Modality Whole Heart Segmentation (MMWHS) Challenge 2017 (Zhuang 2018) with the labels of ascending aorta (AA), left atrium blood cavity (LA), left ventricle bold cavity (LV), and myocardium of the left ventricle (MYO). Similarly, we adopt DSC for our performance evaluation.

Implementation Details

In the CTTA setting, we select each domain as the source domain one at a time, with all other domains serving as continuously changing target domains. Specifically, for cross-sites, we choose the data from any one site as the source domain (e.g., A), and the other domains as continuously changing target domains (e.g., B→C→D→E). Similarly, for cross-sequence, we select a sequence (e.g., T1) as the source domain, with the other sequences as target domains (T1ce→T2→FLAIR). For cross-modality, we choose any modality as the source domain (e.g., MRI) and another as the target domain (e.g., CT). In the source-training phase, we train a ResUNet-34 as our baseline for 2D segmentation like OC/OD segmentation and brain tumor segmentation tasks and train a 3DUnet for 3D cardiac segmentation. We utilize the Adam optimizer with a learning rate of 0.05 for all the tasks. The hyperparameters λ (injection rate of prompt), K (size of offset bank), γ (similarity threshold) are set to 0.01, 10, and 0.95 for all segmentation tasks.

Experimental Results

We evaluate our proposed method on three continual test time adaptation datasets with different kinds of domain shift, i.e., cross-site, cross-sequence, and cross-modality. 'Source Only' denotes the source model directly tests without adaptation. TENT-continual (Wang et al. 2021) is a method based on entropy minimization and we skip the model reset to facilitate the CTTA setting. CoTTA (Wang et al. 2022) refines pseudo-label by source anchor model and exponential moving average (EMA) updated source model with the consistency between original and augmented prediction. SAR (Niu et al. 2023) achieves robust adaptation by filtering noisy samples and encouraging model weight to go to a flat minimum. The above methods optimize the BN layers in the model, however, VPTTA (Chen et al. 2024) proposes to inject the visual prompt into the low-frequency domain of input without unfreezing the BN layers. Visualization of the segmentation results is shown in Fig. 4.

Cross-Site Segmentation of OD/OC We compare the other methods with our proposed method on the OD/OC segmentation task, as shown in Table 1. It can be observed that all the methods are effective in the cross-site domain shift, improving the 'Source Only' baseline significantly. Our method achieves the best performance over the others, more specifically, it not only shows competitive performance in domains where most methods have a good performance but also performs well in domains where other methods exhibit suboptimal results.

Cross-Sequence Segmentation of Brain Tumor We conduct the comparison on the brain tumor segmentation as shown in Table 2. This is a challenging task, the reasons

Methods	Domain A		Domain B		Domain C		Domain D		Domain E		Average
	OC	OD	OC	OD	OC	OD	OC	OD	OC	OD	
Source Only	50.81	71.91	68.50	82.46	54.02	73.50	48.41	66.25	49.93	71.05	63.68
TENT-continual	62.93	72.47	**72.23**	82.21	56.30	75.15	50.21	64.87	50.05	77.11	66.35
CoTTA	55.92	66.15	69.03	82.15	56.22	72.10	50.23	66.85	56.27	76.58	65.15
SAR	61.27	77.46	69.03	80.46	53.30	70.83	**55.84**	**69.58**	57.28	**80.06**	67.51
VPTTA	61.34	76.28	69.92	82.93	55.00	73.26	53.59	68.89	53.89	75.42	67.05
Ours	**63.75**	**77.81**	72.12	**84.05**	**58.29**	**75.45**	54.39	68.89	**57.30**	78.27	**69.03**

Table 1: Performance comparison (DSC % in mean) for the OD/OC segmentation task. The best and second-best results are highlighted in **bold** and underline, respectively.

Methods	T1			T1ce			T2			FLAIR			Average
	WT	TC	ET	WT	TC	ET	WT	TC	ET	WT	TC	ET	
Source Only	29.47	20.11	17.20	42.14	23.86	14.58	31.74	19.47	10.91	34.87	**26.49**	17.68	24.04
TENT-continual	11.36	19.09	**28.43**	17.56	13.91	23.01	1.28	10.24	19.42	11.93	15.65	14.17	15.50
CoTTA	5.34	14.87	25.90	9.88	14.38	**25.26**	0.86	11.29	**22.10**	2.93	10.17	20.48	13.62
SAR	22.05	23.00	26.01	41.95	22.98	14.94	32.46	19.62	11.38	34.23	26.36	17.25	24.35
VPTTA	25.75	14.70	16.16	21.16	14.94	13.13	20.16	10.98	4.38	21.35	14.06	9.32	15.51
Ours	**44.94**	**25.87**	24.06	**47.49**	**25.22**	17.86	**38.99**	**22.95**	14.00	**34.96**	26.17	**22.64**	**28.76**

Table 2: Performance comparison (DSC % in mean) for the brain tumor segmentation task. The best and second-best results are highlighted in **bold** and underline, respectively.

include: (1) the source model is pre-trained under the uni-sequence condition, which limits the representative capability of the model, leading to unsatisfied performance for the 'Source Only' model, (2) the domain shift between the different domains is extremely large (e.g., Whole Tumor is clear in T1ce, but blur in T1) and (3) the test sample is several times larger than train samples, the model is prone to have error accumulation and catastrophic forgetting. All the methods suffer from performance degradation, even inferior to the 'Source Only' baseline. SAR still achieves a marginal improvement through its robust sample filter and model weight optimization strategies. However, our method can continuously outperform nearly all the methods in each sequence. We contribute to its offset bank, which can avoid both error accumulation and catastrophic forgetting.

Cross-Modality Segmentation of Cardiac We also evaluate our method on the 3D cardiac segmentation task, as shown in Table 3. This is another challenging adaptation task, only 20 samples are available for the adaptation stage. As expected, all methods failed to improve the 'source-only' model due to the large number of trainable parameters and the small sample size, leading to underfitting. However, our approach, along with VPTTA, involves a very small number of parameters, allowing the prompt to be sufficiently trained with only a few samples. Furthermore, our deformable convolutional prompt and parameter decomposition greatly enhance the prompt's representative capability while reducing the number of parameters, resulting in a further improvement over VPTTA.

Comparison in a long-term continual test-time adaptation As shown in Fig. 3, we evaluate long-term adaptation performance by testing the source model, pre-trained on each domain of the OD/OC segmentation task, over five

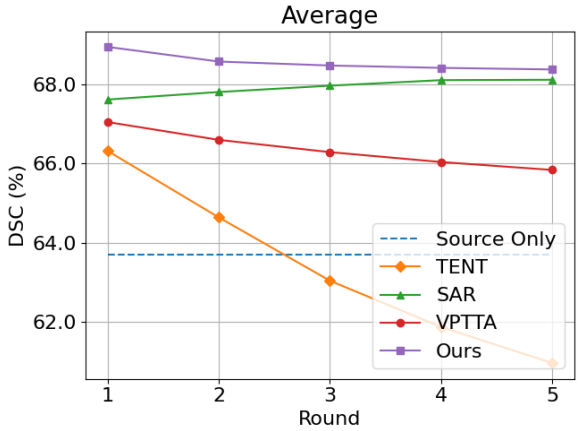

Figure 3: Performance comparison on the OD/OC segmentation task for long-term continual test-time adaptation, where the average performance is calculated across five domains (A, B, C, D, and E).

rounds. Performance trends reflect catastrophic forgetting and error accumulation. TENT shows severe degradation, while VPTTA exhibits moderate decline due to its statistics warm-up strategy. SAR achieves a slight improvement in average DSC, as it applies a complex regularization strategy on model weights to mitigate error accumulation. In contrast, our method experiences a minor decrease, which progressively stabilizes over rounds.

Ablation Study

We conduct the ablation experiment on the OD/OC tasks to prove the effectiveness of the proposed framework in Ta-

Methods	MR				CT				Average
	AA	LA	LV	MYO	AA	LA	LV	MYO	
Source Only	52.30	79.01	86.70	71.41	45.33	35.51	71.73	69.00	63.87
TENT-continual	52.79	79.17	86.74	70.65	45.33	35.52	71.76	69.01	63.87
CoTTA	52.32	79.01	86.69	71.31	45.33	35.51	71.76	69.01	63.87
SAR	52.30	79.01	86.70	71.41	45.33	35.51	71.73	69.00	63.87
VPTTA	51.90	79.56	86.52	71.62	45.69	35.71	**72.14**	68.99	64.02
Ours	**58.19**	**83.54**	**88.45**	**74.51**	**46.68**	**35.93**	72.14	**70.08**	**66.19**

Table 3: Performance comparison (DSC % in mean) for the cardiac segmentation task. The best and second-best results are highlighted in **bold** and underline, respectively.

DCP	TD	STA	FI	DSC	Time	Parameters
				63.68	1.0	0
✓				61.88	2.1	588
✓		✓		64.04	2.4	588
✓	✓	✓		68.16	2.6	186
✓	✓	✓	✓	**69.03**	**1.6**	**186**

Table 4: Ablation studies of our framework. DCP represents the deformable convolutional prompt, TD represents the transformation decomposition, STA denotes the source-target alignment, and FI refers to the fast inference.

Methods	RT	Params	Avg. F.	Avg. B.	Mem. (MB)
Source	1.0	0	1	0	2496
TENT	1.7	21184	2	1	3195
CoTTA	2.0	21184	3	1	3625
SAR	2.3	21184	3	2	3259
VPTTA	2.4	**75**	2	1	3635
Ours	**1.5**	186	2	**0.03**	**2901**

Table 5: Efficiency comparison with other methods. RT denotes relative training time, Params denotes trainable parameters, 'Avg. F.' denotes average forward times, 'Avg. B.' denotes average backward times, and 'Mem.' denotes GPU memory.

ble 4. To be noticed, we utilize entropy minimization as our default optimization if the model is adapted without our STA. The results demonstrate that: (1) Using DCP only with entropy minimization hurts performance because entropy alignment is coarse-grained, addressing only the prediction level. In contrast, when using STA, which adds only a small amount of training time, performance improves significantly, as STA considers both the prompt-level alignment and the statistical distribution differences at each BN layer of the model; (2) the TD helps address the parameter imbalance issue and significantly reduces the number of parameters, resulting in a substantial improvement in model performance. (3) the FI further reduces training time and serves as regularization to prevent the prompt from overfitting to a single domain, thereby enhancing model performance.

Efficiency Analysis

To compare the efficiency of our method with the others, we conduct experiments on the OD/OC segmentation task,

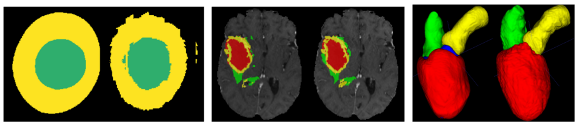

Figure 4: Example segmentation results of the proposed method for OD/OC, brain tumor, and cardiac segmentation. Left: ground truth; right: our results.

as shown in Table 5. Methods based on the update of the BN layer have a significantly larger number of parameters compared to prompt-based methods. Additionally, TENT requires only 2 forward passes (i.e., one for optimization and one for inference) and one backward pass, resulting in less training time. CoTTA, due to the use of the EMA model and source anchor model, has increased memory usage and also requires an additional forward pass for augmented data. While SAR shows strong robustness in the aforementioned segmentation tasks, it relies on two backward passes and two steps of gradient optimization, which greatly increases computation time. Although VPTTA has the fewest parameters, it requires the Fast Fourier Transform for prompt injection, leading to the highest computational time and memory usage. Our proposed method also has a low parameter count and, by skipping a large number of samples during the fast inference phase, achieves an average of only 0.03 backward passes, making it superior to other methods in terms of training time and memory usage.

Conclusion

In this paper, we have presented an efficient prompt-based method, EDCP, for continual test-time adaptation. Specifically, we introduced a deformable convolution prompt with a small number of parameters to adapt the model in the input space. Additionally, we addressed the parameter imbalance issue between the offset branch and the convolution branch by decomposing the offset transformation into two parts, with the second part stored in an offset bank as an implicit domain indicator, which is then used to accelerate inference. Moreover, we designed source-target alignment at both the input and layer levels without unfreezing the BN layers. Finally, we incorporated fast inference to skip the training of samples highly similar to those in the offset bank. Extensive experiments on various medical domain shift benchmarks demonstrate the superiority of our method.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China under Grant No. 62276088.

References

Bakas, S.; Reyes, M.; Jakab, A.; Bauer, S.; Rempfler, M.; Crimi, A.; Shinohara, R. T.; Berger, C.; Ha, S. M.; Rozycki, M.; et al. 2019. Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge. arXiv:1811.02629.

Chakrabarty, G.; Sreenivas, M.; and Biswas, S. 2023. SANTA: Source Anchoring Network and Target Alignment for Continual Test Time Adaptation. *Transactions on Machine Learning Research*.

Chen, Z.; Pan, Y.; Ye, Y.; Lu, M.; and Xia, Y. 2024. Each test image deserves a specific prompt: Continual test-time adaptation for 2d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11184–11193.

Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; and Wei, Y. 2017. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, 764–773.

Dong, H.; Konz, N.; Gu, H.; and Mazurowski, M. A. 2024. Medical Image Segmentation with InTEnt: Integrated Entropy Weighting for Single Image Test-Time Adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5046–5055.

Fumero, F.; Alayón, S.; Sanchez, J. L.; Sigut, J.; and Gonzalez-Hernandez, M. 2011. RIM-ONE: An open retinal image database for optic nerve evaluation. In *2011 24th international symposium on computer-based medical systems (CBMS)*, 1–6. IEEE.

Gan, Y.; Bai, Y.; Lou, Y.; Ma, X.; Zhang, R.; Shi, N.; and Luo, L. 2023. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 7595–7603.

Lim, H.; Kim, B.; Choo, J.; and Choi, S. 2023. TTN: A Domain-Shift Aware Batch Normalization in Test-Time Adaptation. arXiv:2302.05155.

Ni, J.; Yang, S.; Xu, R.; Liu, J.; Li, X.; Jiao, W.; Chen, Z.; Liu, Y.; and Zhang, S. 2024. Distribution-Aware Continual Test-Time Adaptation for Semantic Segmentation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 3044–3050. IEEE.

Niu, S.; Wu, J.; Zhang, Y.; Wen, Z.; Chen, Y.; Zhao, P.; and Tan, M. 2023. Towards Stable Test-Time Adaptation in Dynamic Wild World. arXiv:2302.12400.

Orlando, J. I.; Fu, H.; Breda, J. B.; Van Keer, K.; Bathula, D. R.; Diaz-Pinto, A.; Fang, R.; Heng, P.-A.; Kim, J.; Lee, J.; et al. 2020. Refuge challenge: A unified framework for evaluating automated methods for glaucoma assessment from fundus photographs. *Medical image analysis*, 59: 101570.

Sivaswamy, J.; Krishnadas, S.; Joshi, G. D.; Jain, M.; and Tabish, A. U. S. 2014. Drishti-gs: Retinal image dataset for optic nerve head (onh) segmentation. In *2014 IEEE 11th international symposium on biomedical imaging (ISBI)*, 53–56. IEEE.

Tsai, Y.-Y.; Mao, C.; and Yang, J. 2024. Convolutional visual prompt for robust visual perception. *Advances in Neural Information Processing Systems*, 36.

Valanarasu, J. M. J.; Guo, P.; Vibashan, V.; and Patel, V. M. 2024. On-the-fly test-time adaptation for medical image segmentation. In *Medical Imaging with Deep Learning*, 586–598. PMLR.

Wang, D.; Shelhamer, E.; Liu, S.; Olshausen, B.; and Darrell, T. 2021. Tent: Fully Test-time Adaptation by Entropy Minimization. arXiv:2006.10726.

Wang, Q.; Fink, O.; Van Gool, L.; and Dai, D. 2022. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7201–7211.

Wu, Y.; Chi, Z.; Wang, Y.; Plataniotis, K. N.; and Feng, S. 2024. Test-time domain adaptation by learning domain-aware batch normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15961–15969.

Yang, H.; Chen, C.; Jiang, M.; Liu, Q.; Cao, J.; Heng, P. A.; and Dou, Q. 2022. Dltta: Dynamic learning rate for test-time adaptation on cross-domain medical images. *IEEE Transactions on Medical Imaging*, 41(12): 3575–3586.

Yang, S.; Wu, J.; Liu, J.; Li, X.; Zhang, Q.; Pan, M.; Gan, Y.; Chen, Z.; and Zhang, S. 2024. Exploring sparse visual prompt for domain adaptive dense prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 16334–16342.

Zhang, Z.; Yin, F. S.; Liu, J.; Wong, W. K.; Tan, N. M.; Lee, B. H.; Cheng, J.; and Wong, T. Y. 2010. Origa-light: An online retinal fundus image database for glaucoma analysis and research. In *2010 Annual international conference of the IEEE engineering in medicine and biology*, 3065–3068. IEEE.

Zhuang, X. 2018. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE transactions on pattern analysis and machine intelligence*, 41(12): 2933–2946.