

# Test-time Adaptation of Pelvic Bone Segmentation Models via Dynamic Reliability-Guided

Ling Ren<sup>1</sup>, Chao Deng<sup>1\*</sup>, Ziming Wang<sup>1</sup>, Yuecong Xu<sup>2</sup>, and Kai Zheng<sup>3</sup>

<sup>1</sup> College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

<sup>2</sup> Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583

<sup>3</sup> State Key Laboratory Cultivation Base of Research, Prevention and Treatment for Oral Diseases, the Affiliated Stomatological Hospital of Nanjing Medical University, Nanjing 210029, China

**Abstract.** Reliable pelvic bone segmentation (PBS) from CT is essential for robot-assisted pelvic trauma surgery, yet deploying a source-trained model to a new hospital suffers from severe performance degradation due to cross-center domain shifts. While test-time adaptation (TTA) enables online model adaptation without accessing source data, existing methods show limited effectiveness for PBS, facing challenges including boundary degradation, anatomical inconsistency under domain shifts, and voxel-level class imbalance. To address these challenges, we propose a novel closed-loop dynamic Reliability-Guided TTA framework (ReGA) for PBS. Specifically, we introduce a pseudo-label reliability criterion termed Segmentation Inference Consistency Evaluation (SICE), which jointly measures region overlap and boundary deviation via dropout-based ensemble predictions. Based on SICE, a trust-weighted refinement module adaptively updates features to mitigate boundary errors in pseudo-labels. Furthermore, a confidence-weighted region-level contrastive learning strategy is proposed to enforce anatomical consistency. Finally, ReGA follows the teacher-student (TS) scheme to alleviate voxel-level class imbalance. Experiments on three heterogeneous 3D pelvic CT datasets demonstrate that ReGA consistently outperforms state-of-the-art TTA methods, enabling effective adaptation of the source-trained PBS model to unseen clinical domains. The code is available at <https://github.com/Ren-ling/ReGA>.

**Keywords:** Test-time adaptation · Segmentation · Pelvic bone CT.

## 1 Introduction

Reliable Pelvic Bone Segmentation (PBS) from CT is crucial for the pre-operative planning and intra-operative navigation of robot-assisted pelvic trauma surgery

---

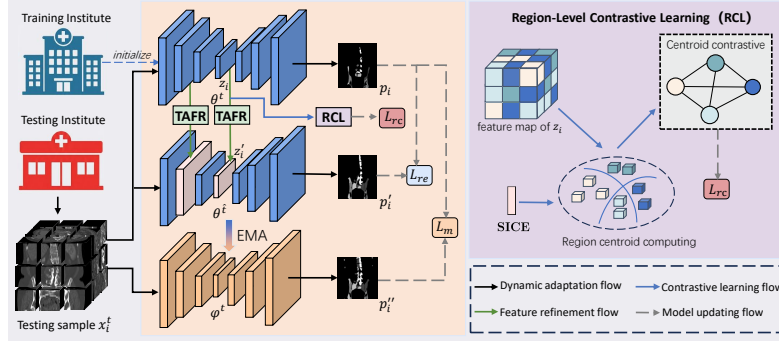
\* Corresponding author. [dengchao\\_neu@126.com](mailto:dengchao_neu@126.com)

[11, 13]. To avoid costly data annotation and repeated model training, deploying a well-trained PBS model to a new hospital is highly desirable. However, discrepancies in imaging devices and patient population heterogeneity introduce severe domain shifts, which lead to performance degradation of the deployed model in the new hospital. Unsupervised domain adaptation methods mitigate domain shift by knowledge transfer to an unlabeled target domain [2, 23] (i.e., a new hospital), improving deployment performances. However, these methods require full access to source domain data, which may raise privacy concerns [10].

To cope with the data privacy issue, source-free domain adaptation (SFDA) methods were introduced to avoid accessing any source data during adaptation [9]. Yet SFDA methods typically require sufficient target-domain samples collected offline [19]. In reality, it is usually difficult to obtain such pelvic CT data in a new hospital that aligns with current imaging protocols, as these data arrive in a case-by-case manner. Motivated by test-time adaptation (TTA) that performs continuous online adaptation during the test stage [15], we introduce it in the adaptation of the deployed PBS model.

Recently, several studies have explored TTA for medical image data [3, 4, 17, 22, 26]. Most of them achieve adaptation by modifying normalization layers of source-trained models [3, 4] or teacher-student (TS) schemes based on pseudo labels [17, 22, 26]. However, these methods exhibit performance degradation when applied in cross-domain PBS due to three major challenges. First, when deployed in a new hospital, a well-trained PBS model struggles to delineate boundaries between adjacent pelvic bones, which exhibit appearance variations across domains. Consequently, the prevailing TS schemes are confined by false prediction and error accumulation at these boundaries [26]. Second, as observed from the ground truth (GT), enforcing anatomical consistency is crucial for the reliable adaptation of the PBS model [1]. However, existing methods neglect the self-supervised abilities within the target domain to achieve this [17]. Third, PBS faces voxel-level class imbalance, in which the pelvic bones occupy only a small fraction of the volume. The class imbalance causes minor parameter updates during adaptation to be easily dominated by background voxels, impairing segmentation performance.

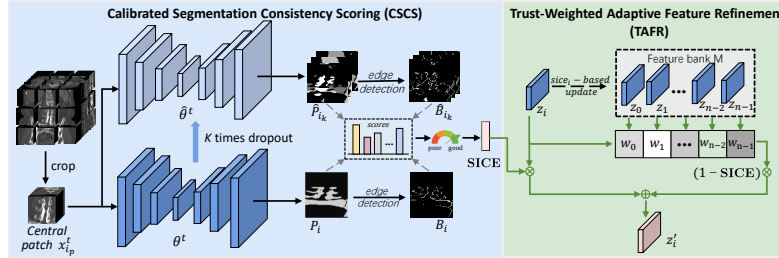
To cope with these challenges for effective TTA in PBS, we propose a novel closed-loop dynamic **Reliability-Guided TTA** framework (**ReGA**). Prior methods exhibit coarse boundary awareness in the target domain, so we propose a novel pseudo-label reliability criterion termed *Segmentation Inference Consistency Evaluation* (SICE). The SICE measures dropout-based ensemble segmentation consistency by considering region overlap and boundary deviation through Dice and Hausdorff distance metrics, further calibrated by confidence scores to improve assessing prediction quality. Based on SICE, a Trust-Weighted Adaptive Feature Refinement (TAFR) module selects high-SICE features from a memory bank, fusing them with the current testing sample to refine pseudo-labels. Furthermore, to enforce anatomical consistency under domain shifts, we propose a Region-Level Contrastive Learning (RCL) module that contrasts trust-aware re-



**Fig. 1.** Overall process of the proposed ReGA. During adaptation, the features of a testing sample  $x_i^t$  extracted by the encoder of the PBS model  $\theta^t$  (initialized at the training institute) are refined by TAFR (see Fig. 2). The refined features then replace those of the model  $\theta^{\hat{t}}$  copied from  $\theta^t$  to generate refined predictions. With the SICE score from CSCS (see Fig. 2) and the feature map of  $z_i$ , region centroids are computed for region-level contrastive learning. Finally,  $\theta^t$  is optimized by minimizing the refined pseudo-label loss  $L_{re}$ , mean teacher loss  $L_m$  and contrastive loss  $L_{rc}$ .

gional centroids. To this end, ReGA follows the TS scheme, which is commonly used in class imbalance segmentation and provides target-specific supervision.

In summary, our contributions are threefold. First, to the best of our knowledge, we propose the first TTA framework for PBS, named ReGA, which enables effective adaptation to unseen clinical domains. Second, we propose a closed-loop reliability-guided refinement scheme to address boundary degradation during on-line adaptation. In addition, a confidence-weighted region-level contrastive loss is introduced to enforce anatomical feature consistency across pelvic regions. Finally, extensive experiments on three 3D pelvic CT datasets demonstrate that ReGA outperforms state-of-the-art TTA methods.



**Fig. 2.** Overview of the CSCS and TAFR module. In CSCS, the test sample  $x_i^t$  is centrally cropped to obtain  $x_{ip}^t$ , which is passed through  $\theta^t$  with  $K$ -times dropout for multiple inferences. These inferences, along with the edges extracted using the Canny operator, are combined to compute SICE. In TAFR, latent features of high SICE are dynamically accumulated into a feature bank  $M$ . The current feature is then adaptively refined through weighted fusion with those from the bank, guided by SICE and cosine similarity, to achieve adaptation.

## 2 Methodology

Given a PBS model  $\theta^t$  trained on a source domain dataset  $\mathcal{D}^s = \{(x_i^s, y_i^s)\}_{i=1}^{N^s}$ , where  $x_i^s$  is a 3D tensor and  $y_i^s$  is a densely labeled 3D segmentation mask, with  $N^s$  representing the number of samples in the source domain. Let the unlabeled target domain dataset for testing be denoted as  $\mathcal{D}^t = \{x_i^t\}_{i=1}^{N^t}$  and  $N^t$  represents the number of samples in the target domain, which has a distribution shift from  $\mathcal{D}^s$ . The goal of this paper is to adapt  $\theta^t$  in an online manner within each batch  $B^j = \{x_i^t\}_{i=1}^B$  from  $\mathcal{D}^t$ , where  $j \in 1, \dots, N^t/B$  and  $B$  denotes the batch size. To achieve this, we propose ReGA for the adaptation of the PBS model on  $\mathcal{D}^t$ , as illustrated in Fig. 1.

### 2.1 Calibrated Segmentation Consistency Scoring

The performance of PBS models is commonly evaluated using the Dice coefficient and the Hausdorff distance. In real-world inference, where GT for test samples are unavailable, automated performance evaluation becomes crucial. Although Monte Carlo (MC) dropout [5] offers uncertainty estimates through multiple forward passes, these uncertainty maps do not directly reflect the segmentation accuracy, particularly of boundary delineation.

Motivated by prior studies [8, 26] that demonstrate a correlation between prediction variance under dropout and testing errors, we propose Calibrated Segmentation Consistency Scoring (CSCS) to calculate SICE to evaluate the reliability of pseudo-labels generated by the target model during adaptation. SICE introduces a boundary-oriented indicator to provide a more precise assessment of segmentation quality.

Specifically, let  $\hat{\theta}_k^t$  denote the target model  $\theta^t$  parameters with the  $k$ -th MC dropout. As shown in Fig. 2, given a test sample  $x_i^t$ , the central patch  $x_{i_p}^t$  is fed into  $\theta^t$  and its  $k$ -th dropout version  $\hat{\theta}_k^t$ , yielding predictions  $P_i = f(\theta^t, x_{i_p}^t)$  and  $\hat{P}_{i_k} = f(\hat{\theta}_k^t, x_{i_p}^t)$ , respectively. To explicitly characterize boundary discrepancies, edge maps are extracted from  $P_i$  and  $\hat{P}_{i_k}$  using the Canny operator [20], producing  $B_i = \text{Canny}(P_i)$  and  $\hat{B}_{i_k} = \text{Canny}(\hat{P}_{i_k})$ . The proposed SICE score is formulated as a weighted combination of the volumetric segmentation consistency ( $S_v$ ) and the boundary discrepancy ( $S_b$ ). The volumetric consistency metric is defined as

$$\mathcal{S}(x, \hat{x}) = \frac{1}{K \cdot L} \sum_{k=1}^K \sum_{l=0}^{L-1} \frac{2 \sum_{v \in \Omega} x_v^l \cdot \hat{x}_{k,v}^l}{\sum_{v \in \Omega} x_v^l + \sum_{v \in \Omega} \hat{x}_{k,v}^l}, \quad (1)$$

where  $\Omega$  represents the set of voxel indices, with  $v = (h, w, d) \in \Omega$  denoting the index of each voxel,  $K$  is the number of dropout forward passes and  $L$  is the number of classes.  $x_v^l$  and  $\hat{x}_{k,v}^l$  denote the predicted probability and those obtained by the  $k$ -th dropout version for the  $l$ -th class of the  $v$ -th voxel, respectively.

By instantiating  $\mathcal{S}(\cdot, \cdot)$  with the segmentation outputs, the volumetric segmentation consistency is obtained as

$$S_v = \mathcal{S}(P_i, \hat{P}_{i_k}). \quad (2)$$

Subsequently, the boundary consistency measure  $S_b$  is computed as

$$S_b = \max \left( P_\gamma(B_i, \hat{B}_{i_k}), P_\gamma(\hat{B}_{i_k}, B_i) \right), \quad (3)$$

where  $P_\gamma(B_i, \hat{B}_{i_k})$  denotes the  $\gamma$  percentile of the minimum distances from all voxels in the boundary voxel sets  $B_i$  to set  $\hat{B}_{i_k}$ , and  $P_\gamma(\hat{B}_{i_k}, B_i)$  denotes the calculation in the revise direction. Finally, the reliability score  $R_i$  for the current testing sample  $x_i^t$  is defined as

$$R_i = \lambda_1 S_v + \lambda_2 \exp(-S_b/\alpha), \quad (4)$$

where  $\lambda_1$  and  $\lambda_2$  weight region- and boundary-level consistency, respectively, and  $\alpha$  modulates the sensitivity to boundary deviations.

The dropout-based inference often exhibits high consensus in the interior regions of the target domain predictions, whereas discrepancies are primarily concentrated at the boundaries [18]. Consequently,  $S_v$  tends to overestimate the true segmentation performance. Thus, we introduce a calibrating coefficient  $\lambda \in (0, 1)$  to calibrate the estimation based on the overall confidence as follows

$$\lambda = 1 - \frac{1}{|\Omega| \log C} \sum_{v \in \Omega} \left( - \sum_{l=0}^L \bar{P}_{v,l} \log \bar{P}_{v,l} \right), \quad (5)$$

where  $\bar{P} = \frac{1}{K} \sum_{k=1}^K \hat{P}_{i_k}$  denotes the ensemble prediction map and  $\log C$  is the maximum entropy value used for normalization. The proposed SICE of the current testing sample  $x_i^t$  is formulated as  $SICE_i = \lambda \cdot R_i$ , which is used to select high-confidence samples and guide the subsequent adaptation process.

## 2.2 Trust-Weighted Adaptive Feature Refinement

As shown by the previous method [25], explicit feature alignment from well-predicted samples helps minimize the domain gap for other testing samples. Thus, we propose a trust-weighted refinement module with a dynamic feature bank that stores high-quality target domain representations and applies trust-weighted fusion for feature correction.

As described in Fig. 2, the dynamic feature bank  $M$  comprises  $\{Z_i\}_{i=0}^{n-1}$ , where  $Z_i = f(\theta_e^t, x_i^t)$  and  $M$  are updated using a first-in-first-out principle to ensure adaptability to the distribution of the incoming test sample. Specifically, when  $SICE_i$  exceeds the  $\tau$  percentile of  $SICE$  values of all previous samples before time step  $t$ ,  $Z_i$  is incorporated into the feature bank  $M$ .

For a new testing sample  $x_i^t$ , the cosine similarity between  $Z_i$  and the  $j_{th}$  feature  $Z_j$  in  $M$  is computed as  $\text{sim}(Z_i, Z_j)$ . Then  $Z_i$  is updated as a trust-weighted combination of  $Z_i$  and the feature  $Z_f$ , where  $Z_f = \sum_{j=0}^{n-1} W_j \cdot Z_j$  represents the reference feature from the dynamic bank based on the cosine similarity and  $W_j = \text{sim}(Z_i, Z_j) / \sum_j \text{sim}(Z_i, Z_j)$ . To encourage the retention of original features for well-predicted samples, while replacing those with poor predictions, the refined feature  $Z'_i$  is defined as

$$Z'_i = w(SICE_i) \cdot Z_i + (1 - w(SICE_i)) \cdot Z_f, \quad (6)$$

where  $w(\cdot)$  represents the normalization operation. The refined feature  $Z'_i$  is then sent to the student model  $\theta^t$  to obtain a refined prediction  $p'_i$ .

### 2.3 Region-Level Contrastive Learning

In the context of domain adaptation for segmentation models, prior works [21, 24] have utilized pseudo labels to perform contrastive learning for a more compact target feature distribution. With this intuition, we propose the RCL module to enforce anatomical consistency. Rather than relying on voxel-level representations, the region centroid is utilized to represent the entire region for reducing computational cost. Additionally, we dynamically assign weights to voxels based on prediction confidence and refined with the SICE score. Formally, the confidence-weighted centroid of class  $l$  is computed as

$$\mathbf{c}_l = \frac{\sum_{j \in \Omega} \mathbf{f}_j(Z_i) \cdot \mathbf{I}(\hat{y}_j = l) \cdot (1 - E_i) \cdot SICE_i(l)}{\sum_{j \in \Omega} \mathbf{I}(\hat{y}_j = l)}, \quad (7)$$

where  $\mathbf{f}_j$  is the feature map of  $Z_i$  at voxel  $j$ ,  $\mathbf{I}(\cdot)$  is the indicator function,  $\hat{y}_j$  is the prediction label from the target model  $\theta^t$  at voxel  $j$ ,  $E_i$  is the entropy-based uncertainty score, and  $SICE_i(l)$  denotes the SICE score for class  $l$ .

Then, the learning objective for anatomical consistency is formulated as

$$\mathcal{L}_{rc} = -\frac{1}{L} \sum_{l=0}^{L-1} \log \frac{\exp(\text{sim}(\mathbf{c}_l, \mathbf{c}_l))}{\sum_{m=0, m \neq l}^{L-1} \exp(\text{sim}(\mathbf{c}_l, \mathbf{c}_m))}, \quad (8)$$

where  $\text{sim}(\mathbf{u}, \mathbf{v}) = (\mathbf{u}^\top \mathbf{v}) / (\|\mathbf{u}\| \|\mathbf{v}\| \cdot T)$  denotes the cosine similarity scaled by a temperature parameter  $T$ .

### 2.4 Self-Adaptive Model Updating

Although TAFR contributes to producing generally reliable pseudo labels, the gradients can vary significantly across different testing batches, which causes instability during adaptation. In line with established protocol in TTA [15, 16], we adopt a mean teacher model to improve stability. At the time step  $t = 0$ , the mean teacher model  $\phi^t$  is initialized to be the same as the source pre-trained model  $\theta^t$ , i.e.,  $\phi^0 = \theta^0$ . Since the traditional mean teacher updated with a constant Exponential Moving Average (EMA) rate fails to effectively handle the dynamic changes in data quality during TTA, we propose a SICE-aware updating rule with an adaptive EMA rate as follows

$$\phi^{t+1} = (1 - SICE_i) \cdot \phi^t + SICE_i \cdot \theta^{t+1}, \quad (9)$$

where  $\theta^{t+1}$  are the updated student model at the current adaptation step  $t$ .

To handle potential noise in the refined prediction  $p'_i$ , the loss is weighted by  $SICE_i$  to suppress the contribution of poorly adapted samples. The total training loss for our proposed ReGA is defined as

$$L_{ReGA} = \frac{1}{B} \sum_{i=1}^B (L_m(p''_i, p_i) + SICE_i \cdot L_{re}(p'_i, p_i) + \beta L_{rc}), \quad (10)$$

where  $p''_i$  denotes teacher predictions,  $L_m$  and  $L_{re}$  correspond to the mean teacher and refined pseudo-label losses implemented with Dice and cross-entropy, and  $\beta$  is a tradeoff hyperparameter. For each batch, a back-propagation step is performed using  $L_{ReGA}$ , followed by a forward pass with the updated student model  $\theta^{t+1}$  to generate the final segmentation results.

**Table 1.** Comparison of different TTA methods for cross-domain PBS.

| Methods     | CLINIC→KITS19                  |                                  |                                 | CLINIC→MSD <sub>T10</sub>       |                                  |                                 |
|-------------|--------------------------------|----------------------------------|---------------------------------|---------------------------------|----------------------------------|---------------------------------|
|             | Dice $\uparrow$                | HD95 $\downarrow$                | ASD $\downarrow$                | Dice $\uparrow$                 | HD95 $\downarrow$                | ASD $\downarrow$                |
| Source Only | 19.7 $\pm$ 9.5                 | 112.7 $\pm$ 20.8                 | 67.2 $\pm$ 24.6                 | 13.8 $\pm$ 6.9                  | 149.3 $\pm$ 25.7                 | 56.3 $\pm$ 22.2                 |
| Target Only | 97.8 $\pm$ 4.6                 | 32.2 $\pm$ 16.5                  | 18.7 $\pm$ 12.1                 | 91.5 $\pm$ 5.1                  | 33.7 $\pm$ 13.5                  | 19.4 $\pm$ 9.2                  |
| TENT [15]   | 24.9 $\pm$ 7.9                 | 190.6 $\pm$ 66.1                 | 62.9 $\pm$ 27.4                 | 25.8 $\pm$ 12.7                 | 171.0 $\pm$ 53.3                 | 67.4 $\pm$ 25.9                 |
| InTEnt [4]  | 23.4 $\pm$ 6.2                 | 210.6 $\pm$ 87.1                 | 65.4 $\pm$ 30.2                 | 25.1 $\pm$ 10.0                 | 171.9 $\pm$ 61.3                 | 69.4 $\pm$ 25.6                 |
| VPTTA [3]   | 24.3 $\pm$ 6.1                 | 200.7 $\pm$ 72.6                 | 63.6 $\pm$ 31.1                 | 25.8 $\pm$ 12.6                 | 172.0 $\pm$ 53.6                 | 68.0 $\pm$ 29.8                 |
| CoTTA [16]  | 24.6 $\pm$ 7.6                 | 192.4 $\pm$ 63.0                 | 64.1 $\pm$ 29.0                 | 25.1 $\pm$ 11.6                 | 159.8 $\pm$ 62.0                 | 63.2 $\pm$ 30.4                 |
| TEGDA [26]  | 24.3 $\pm$ 10.0                | 187.7 $\pm$ 67.3                 | 63.4 $\pm$ 27.7                 | 25.8 $\pm$ 10.3                 | 168.0 $\pm$ 63.2                 | 62.1 $\pm$ 29.3                 |
| <b>Ours</b> | <b>32.1<math>\pm</math>7.3</b> | <b>160.4<math>\pm</math>42.8</b> | <b>57.9<math>\pm</math>29.5</b> | <b>26.4<math>\pm</math>12.0</b> | <b>143.9<math>\pm</math>70.1</b> | <b>54.2<math>\pm</math>35.4</b> |

### 3 Experiments and Results

**Datasets and Implementation Details.** We extensively evaluate the proposed ReGA on cross-domain PBS with three sub-datasets of the public dataset CTPelvic1K [12], including 1) MSD\_T10: 155 cases collected from the 10th sub-dataset of Medical Segmentation Decathlon [14], 2) KITS19: 44 pelvic CT scans comes from the Kits19 challenge [6], and 3) CLINIC: 103 pelvic CT scans collected from an orthopedic hospital without metal artifacts. The three datasets comprise four segmentation classes: sacrum, left hip, right hip, and lumbar spine.

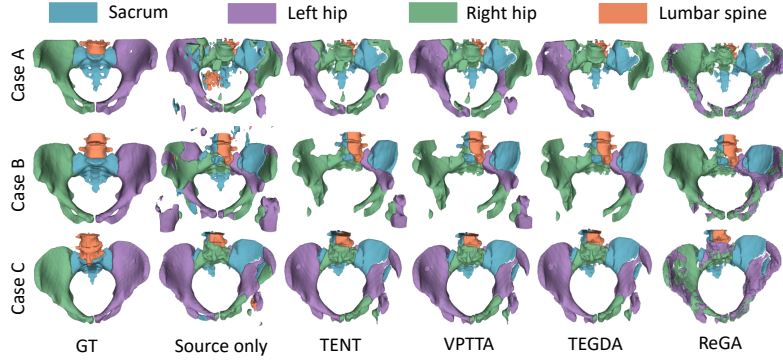
The source model was trained for 400 epochs using the cascaded 3D UNet from nnUNet [7], and the best validation checkpoint was adopted for adaptation. In ReGA, we set a dropout rate of 0.5, the batch size  $B = 1$ , the dropout number  $K = 5$ , the feature bank length  $M = 10$ ,  $\gamma = 95$  for calculating  $S_b$ ,  $\tau = 70$  for sample filtering in TAFR, and the tradeoff hyperparameter  $\beta = 0.01$ . The evaluation metrics are volume-level Dice coefficient, 95th percentile of Hausdorff Distance (HD95), and Average Symmetric Surface Distance (ASD). All models are trained on a single NVIDIA GeForce RTX 4090 24 GB GPU.

**Comparison with State-of-the-art TTA Methods.** We selected five state-of-the-art methods for detailed comparison, including normalization-based methods TENT [15], InTEnt [4], and VPTTA [3]; as well as TS schemes CoTTA [16] and TEGDA [26]. We also report results for the source-only model, obtained by applying the source pre-trained model directly to the target data. Table 1 presents the quantitative results for all comparative methods across different source and target domains. In the CLINIC→KITS19 setting, the source model achieves a Dice coefficient of only 19.7%. In contrast, state-of-the-art TTA methods range from 23.4% to 24.9%, while our ReGA method achieves the highest mean Dice coefficient of 32.1%, a significant improvement of 12.4% over the source model, and outperforms other methods. Moreover, our proposed ReGA achieves the lowest HD95 (160.4) and ASD (57.9) values, indicating improved boundary precision and structural consistency.

**Ablation Study.** We conduct an ablation study in Table 2 to evaluate the effectiveness of ReGA. Specifically, w/o  $L_m$ , w/o  $L_{re}$ , and w/o  $L_{rc}$  correspond to cases in which the mean-teacher loss, the TAFR module, and the RCL module are disabled during adaptation, respectively. Besides, the variant w/o  $B$  indicates

**Table 2.** Ablation studies of ReGA for cross-domain PBS.

| Variants     | CLINIC→KITS19                  |                                  |                                 | CLINIC→MSD <sub>T10</sub>       |                                  |                                 |
|--------------|--------------------------------|----------------------------------|---------------------------------|---------------------------------|----------------------------------|---------------------------------|
|              | DSC $\uparrow$                 | HD95 $\downarrow$                | ASD $\downarrow$                | DSC $\uparrow$                  | HD95 $\downarrow$                | ASD $\downarrow$                |
| w/o $L_m$    | 31.4 $\pm$ 5.4                 | 221.5 $\pm$ 99.3                 | 62.2 $\pm$ 37.4                 | 25.9 $\pm$ 12.9                 | 163.9 $\pm$ 51.7                 | 63.5 $\pm$ 25.4                 |
| w/o $L_{re}$ | 32.0 $\pm$ 6.1                 | 229.6 $\pm$ 105.8                | 61.9 $\pm$ 41.2                 | 25.4 $\pm$ 5.7                  | 168.9 $\pm$ 51.7                 | 66.3 $\pm$ 23.2                 |
| w/o $L_{rc}$ | 31.5 $\pm$ 6.1                 | 224.1 $\pm$ 102.5                | 64.8 $\pm$ 38.8                 | 25.6 $\pm$ 5.4                  | 170.7 $\pm$ 47.9                 | 68.4 $\pm$ 27.3                 |
| w/o $B$      | 31.1 $\pm$ 5.8                 | 225.1 $\pm$ 103.1                | 65.1 $\pm$ 37.1                 | 25.3 $\pm$ 12.8                 | 166.6 $\pm$ 54.9                 | 66.4 $\pm$ 26.8                 |
| w/o $E$      | 31.4 $\pm$ 6.0                 | 224.4 $\pm$ 102.7                | 65.0 $\pm$ 38.7                 | 25.1 $\pm$ 12.7                 | 167.6 $\pm$ 54.8                 | 67.9 $\pm$ 27.0                 |
| <b>Ours</b>  | <b>32.1<math>\pm</math>7.3</b> | <b>160.4<math>\pm</math>42.8</b> | <b>57.9<math>\pm</math>29.5</b> | <b>26.4<math>\pm</math>12.0</b> | <b>143.9<math>\pm</math>70.1</b> | <b>54.2<math>\pm</math>35.4</b> |

**Fig. 3.** Visualization of segmentation results for different TTA methods.

that  $S_b$  of the SICE score is removed, while w/o  $E$  denotes that entropy is used in place of SICE. Ablation results across the three datasets consistently show that each module incrementally enhances segmentation performance. From the obvious improvement in HD95 and ASD compared to the variant w/o  $B$  and w/o  $E$ , the proposed SICE can successfully evaluate pseudo-label quality, which contributes to the adaptation process.

**Visualization.** Fig. 3 presents qualitative comparisons across cases A and C on CLINIC→KITS19, and case B on CLINIC→MSD<sub>T10</sub>. Existing TTA methods (e.g., TENT, VPTTA, and TEGDA) struggle to maintain structural consistency and boundary accuracy under domain shifts. In contrast, ReGA better preserves anatomical integrity while achieving sharper boundary delineation. Notably, in Cases A and C, competing methods suffer from left–right inversion errors, while ReGA maintains anatomically consistent predictions. In Case B, ReGA demonstrates improved delineation of the right hip bone, better preserving the structural completeness of the pelvis.

## 4 Conclusion

In this paper, we propose ReGA, a dynamic reliability-guided test-time adaptation framework that addresses three key challenges in cross-domain PBS, i.e., boundary degradation, anatomical inconsistency, and voxel-level class imbalance. By combining pseudo-label reliability estimation with adaptive feature refine-

ment and anatomical consistency enforcement, ReGA achieves online adaptation for the PBS model deployment in new clinical settings. Experimental results on three heterogeneous 3D pelvic CT datasets demonstrate that ReGA outperforms current state-of-the-art methods, highlighting its potential for clinical applications. Future work will focus on refining the boundary extraction strategy to enhance its reliability on more complex anatomical structures.

**Acknowledgments.** The work was partially supported by the National Science Research Start-up Foundation of Nanjing University of Posts and Telecommunications under Grants NY221007 and NY222111, and was also partially supported by Guang-dong Bifang Intelligent Control Technology Co., Ltd.

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

## References

1. Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., Ben Ayed, I.: Source-free domain adaptation for image segmentation. *Medical Image Analysis* **82**(102617) (2022)
2. Cai, Z., Xin, J., You, C., Shi, P., Dong, S., Dvornek, N., Zheng, N., Duncan, J.: Style mixup enhanced disentanglement learning for unsupervised domain adaptation in medical image segmentation. *Medical Image Analysis* **101**(103440) (2025)
3. Chen, Z., Pan, Y., Ye, Y., Lu, M., Xia, Y.: Each test image deserves a specific prompt: continual test-time adaptation for 2d medical image segmentation. In: *CVPR*. pp. 11184–11193 (2024)
4. Dong, H., Konz, N., Gu, H., Mazurowski, M.: Medical image segmentation with intent: Integrated entropy weighting for single image test-time adaptation. In: *CVPR*. pp. 5046–5055 (2024)
5. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: *ICML*. pp. 1050–1059. PMLR (2016)
6. Heller, N., et al.: The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes. *arXiv preprint arXiv:1904.00445* (2019)
7. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods* **18**(2), 203–211 (2021)
8. Lee, T., Chottananurak, S., Gong, T., Lee, S.: Aetta: label-free accuracy estimation for test-time adaptation. In: *CVPR*. pp. 28643–28652 (2024)
9. Liang, J., Hu, D., Feng, J.: Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In: *International conference on machine learning*. pp. 6028–6039. PMLR (2020)
10. Litrico, M., Del Bue, A., Morerio, P.: Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In: *CVPR*. pp. 7640–7650 (2023)
11. Liu, J., Li, H., Zeng, B., Wang, H., Kikinis, R., Joskowicz, L., Chen, X.: An end-to-end geometry-based pipeline for automatic preoperative surgical planning of pelvic fracture reduction and fixation. *IEEE Transactions on Medical Imaging* **44**(1), 79–91 (2025)

12. Liu, P., et al.: Deep learning to segment pelvic bones: large-scale ct datasets and baseline models. *IJCARS* **16**(5), 749–756 (2021)
13. Liu, Y., et al.: Preoperative fracture reduction planning for image-guided pelvic trauma surgery: a comprehensive pipeline with learning. *Medical Image Analysis* **102**(103506) (2025)
14. Simpson, A., et al.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. arXiv preprint arXiv:1902.09063 (2019)
15. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
16. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: *CVPR*. pp. 7201–7211 (2022)
17. Wu, J., Gu, R., Lu, T., Zhang, S., Wang, G.: Upl-tta: Uncertainty-aware pseudo label guided fully test time adaptation for fetal brain segmentation. In: *IPMI 2023*. LNCS, vol. 13939, pp. 240–252. Springer (2023). [https://doi.org/10.1007/978-3-031-34048-2\\_19](https://doi.org/10.1007/978-3-031-34048-2_19)
18. Wu, J., et al.: Fpl+: Filtered pseudo label-based unsupervised cross-modality adaptation for 3d medical image segmentation. *IEEE Transactions on Medical Imaging* **43**(9), 3098–3109 (2024)
19. Yang, H., Chen, C., Jiang, M., Liu, Q., Cao, J., Heng, P., 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)
20. Yang, Y., Zhuang, J., Sun, G., Wang, R., Su, J.: Boundary-guided contrastive learning for semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging* **44**(7), 2973–2988 (2025)
21. Yu, Q., Xi, N., Yuan, J., Zhou, Z., Dang, K., Ding, X.: Source-free domain adaptation for medical image segmentation via prototype-anchored feature alignment and contrastive learning. In: *MICCAI 2023*. LNCS, vol. 14226, pp. 1–12. Springer (2023). [https://doi.org/10.1007/978-3-031-43990-2\\_1](https://doi.org/10.1007/978-3-031-43990-2_1)
22. Zhang, G., Qi, X., Wu, J., Yan, B., Wang, G.: Iplc+: Sam-guided iterative pseudo label correction for source-free domain adaptation in medical image segmentation. *IEEE Journal of Biomedical and Health Informatics* **29**(12), 9060–9072 (2025)
23. Zhang, X., et al.: Mapseg: Unified unsupervised domain adaptation for heterogeneous medical image segmentation based on 3d masked autoencoding and pseudo-labeling. In: *CVPR*. pp. 5851–5862 (2024)
24. Zhang, Y., Huang, K., Chen, C., Chen, Q., Heng, P.: Satta: semantic-aware test-time adaptation for cross-domain medical image segmentation. In: *MICCAI 2023*. LNCS, vol. 14221, pp. 160–171. Springer (2023). [https://doi.org/10.1007/978-3-031-43895-0\\_14](https://doi.org/10.1007/978-3-031-43895-0_14)
25. Zheng, B., et al.: Dual domain distribution disruption with semantics preservation: Unsupervised domain adaptation for medical image segmentation. *Medical Image Analysis* **97**(103275) (2024)
26. Zhou, Y., Wu, J., Liao, W., Zhang, S., Zhang, S., Wang, G.: Tegda: test-time evaluation-guided dynamic adaptation for medical image segmentation. In: *MICCAI 2025*. LNCS, vol. 15965, pp. 628–637. Springer (2026). [https://doi.org/10.1007/978-3-032-04978-0\\_60](https://doi.org/10.1007/978-3-032-04978-0_60)