TEGDA: Test-time Evaluation-Guided Dynamic Adaptation for Medical Image Segmentation

Yubo Zhou[1], Jianghao Wu[1], Wenjun Liao[2], Shichuan Zhang[2], Shaoting Zhang[1,3], and Guotai Wang[1,3]

[1] University of Electronic Science and Technology of China, Chengdu, China
[2] Department of Radiation Oncology, Sichuan Cancer Hospital and Institute, University of Electronic Science and Technology of China, Chengdu, China
[3] Shanghai Artificial Intelligence Laboratory, Shanghai, China
guotai.wang@uestc.edu.cn

Abstract. Distribution shifts of medical images seriously limit the performance of segmentation models when applied in real-world scenarios. Test-Time Adaptation (TTA) has emerged as a promising solution for ensuring robustness on images from different institutions by tuning the parameters at test time without additional labeled training data. However, existing TTA methods are limited by unreliable supervision due to a lack of effective methods to monitor the adaptation performance without ground-truth, which makes it hard to adaptively adjust model parameters in the stream of testing samples. To address these limitations, we propose a novel Test-Time Evaluation-Guided Dynamic Adaptation (TEGDA) framework for TTA of segmentation models. In the absence of ground-truth, we propose a novel prediction quality evaluation metric based on Agreement with Dropout Inferences calibrated by Confidence (ADIC). Then it is used to guide adaptive feature fusion with those in a feature bank with high ADIC values to obtain refined predictions for supervision, which is combined with an ADIC-adaptive teacher model and loss weighting for robust adaptation. Experimental results on multi-domain cardiac structure and brain tumor segmentation demonstrate that our ADIC can accurately estimate segmentation quality on the fly, and our TEGDA obtained the highest average Dice and lowest average HD95, significantly outperforming several state-of-the-art TTA methods. The code is available at https://github.com/HiLab-git/TEGDA.

Keywords: Domain adaptation · Segmentation · Test-time evaluation.

1 Introduction

Despite the state-of-the-art performance of deep learning models on medical image segmentation, they usually require that the distribution of testing images be aligned with that of training data [13]. However, in clinical scenarios, there are large variations in imaging devices, protocols, and patient demographics, making test samples have an obvious distribution shift from training data, severely

Y. Zhou and J. Wu contributed equally to this work.

limiting performance at test time [3, 18]. To overcome this challenge, Test-Time Adaptation (TTA) is a promising solution as it can efficiently update the model during inference on a stream of testing samples without ground-truth labels.

Existing TTA methods mainly include back-propagation-free [15] and back-propagation-based methods [5, 21]. The first category including PTBN [15] and InTEnt [7] only updates Batch Normalization (BN) statistics based on the current testing image for adaptation. Despite the efficiency, their freezing of model parameters often limits the adaptability. The second category uses loss functions based on auxiliary tasks [9,18], entropy minimization [21] or pseudo-labels [22,24] to update model parameters for better adaptability. The auxiliary task-based methods are challenged by the gap between auxiliary task (e.g., reconstruction) [9] and segmentation task, while entropy minimization easily leads to over-confidence and incorrect predictions. In contrast, methods based on pseudo-labels are appealing as they can provide more target-related supervision for adaptation with the absence of ground-truth [22]. However, due to the domain gap between training and testing samples, the quality of pseudo-labels generated by model prediction is usually poor with a large variation, which adversely affects the performance. Therefore, it is critical to find a reliable metric to evaluate the prediction quality accurately during TTA and design adaptive strategies to leverage them to ensure reliable adaptation. Even though some metrics such as uncertainty based on prediction entropy [16], variance [25] or discrepancy [26] have been designed, they are often not well calibrated to accurately assess the prediction quality on testing samples with distribution shifts. In addition, how to leverage these evaluation metrics to help refine pseudo-labels and adapt the model update strategy has still rarely been explored for segmentation tasks.

To address these issues for pseudo-label-based TTA methods, we propose a novel Test-time Evaluation-Guided Dynamic Adaptation (TEGDA) framework. The contribution is three-fold. Firstly, we present a novel prediction quality evaluation metric based on Agreement with Dropout Inferences calibrated by Confidence (ADIC), where the Dice score between predictions by the model and its dropout version is leveraged to assess the robustness of the model on a testing sample, then it is further calibrated by the confidence to become highly relevant to the real Dice value between the prediction and its ground-truth. Secondly, we propose Adaptive Feature Fusion-based Refinement (AFFR) that adaptively fuses the feature of a sample with those with high ADIC values based on their similarity, leading to robust refined pseudo-labels. Thirdly, we introduce ADIC-guided Self-adaptive Model Updating (SMU) that consists of ADIC-aware pseudo-label loss weighting and ADIC-aware mean teacher to improve the stability of adaptation. Experiments on multi-domain 2D cardiac structure and 3D brain tumor segmentation demonstrate that the Pearson correlation coefficient between our ADIC and real Dice score is as high as 0.83, and TEGDA outperforms several state-of-the-art TTA methods on the two datasets, respectively.

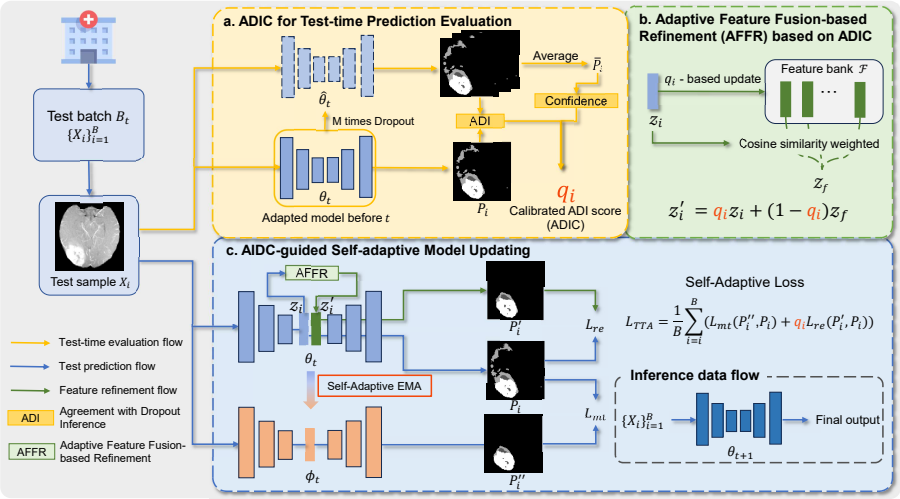

Fig. 1. Overview of our TEGDA for test-time adaptation of segmentation models.

2 Method

Given a segmentation model θ_s pre-trained on a source-domain dataset $\mathcal{D}_\mathcal{S}$, the unlabeled target-domain dataset for testing is denoted as $\mathcal{D}_\mathcal{T}$, where $\mathcal{D}_\mathcal{T}$ has a distribution shift from $\mathcal{D}_\mathcal{S}$. The goal of TTA is to adapt θ_s in real-time to each incoming test batch $B_t = \{X_i\}_{i=1}^B$ from $\mathcal{D}_\mathcal{T}$, where t is the batch index, B is the batch size, and X_i is a testing sample. Importantly, each batch undergoes model update and inference before moving on to the next, guaranteeing that the entire dataset is tested within a single epoch in TTA.

As depicted in Fig. 1, our TEGDA framework has three modules: 1) Test-time prediction quality evaluation via Agreement with Dropout Inferences calibrated by Confidence (ADIC); 2) Adaptive Feature Fusion-based Refinement (AFFR), which constructs a feature bank based on historical samples with high ADIC values and fuses the feature of a testing sample with those in the feature bank for pseudo-label refinement; and 3) ADIC-guided Self-adaptive Model Updating (SMU) that dynamically adjusts the weight of pseudo-labels based on ADIC values, and is combined with an ADIC-aware mean teacher for robust adaptation.

2.1 ADIC for Test-time Prediction Quality Evaluation

For segmentation tasks, Dice between the prediction and ground-truth is a common metric for evaluation. In the absence of ground-truth for testing samples, it is desirable to estimate the Dice values automatically. Though Monte Carlo (MC) Dropout [8] has been widely used for uncertainty estimation to assess the prediction quality, it is not specifically designed for segmentation tasks, and

cannot directly indicate Dice scores. To deal with this problem and inspired by previous works [10,12] that show prediction disagreement with dropouts can approximate the test error, we propose to use Agreement with Dropout Inference calibrated by Confidence (ADIC) as an estimation of the Dice score.

Specifically, let θ_t denote the adapted model parameter before the t-th batch, and $\hat{\theta}_t^m$ is a variant of θ_t with the m-th random dropout. For a test sample X, the predictions with θ_t and $\hat{\theta}_t^m$ are denoted as $P = f(\theta_t, X)$ and $\hat{P}^m = f(\hat{\theta}_t^m, X)$, respectively. The Agreement with Dropout Inference (ADI) score is:

$$ADI(X) = \frac{1}{MC} \sum_{m=1}^{M} \sum_{c=0}^{C-1} \frac{2 \sum_{n=0}^{N-1} P_{n,c} \cdot \hat{P}_{n,c}^m}{\sum_{n=0}^{N-1} P_{n,c} + \sum_{n=0}^{N-1} \hat{P}_{n,c}^m}, \tag{1}$$

where N denotes the pixel/voxel number in the image, and M is the number of dropout inferences. $P_{n,c}$ denotes the probability for class c at the n-th pixel in P, and $\hat{P}_{n,c}^m$ is the corresponding value obtained by the m-th dropout version.

Note that for segmentation models, dropout inference often yields high agreement between P and $\hat{P}^m$ in interior regions but minor discrepancy at borders in the target [25], leading ADI score to overestimate Dice. Therefore, we further introduce a factor $b \in (0, 1)$ to calibrate the ADI score, which is defined as the overall confidence of dropout-based predictions. The average prediction map is denoted as $\bar{P} = \frac{1}{M} \sum_{m=1}^{M} \hat{P}^m$, and the average pixel-wise entropy E_{avg} is:

$$E_{avg} = -\frac{1}{NC} \sum_{n=0}^{N-1} \sum_{c=0}^{C-1} \bar{P}_{n,c} \log \bar{P}_{n,c} \tag{2}$$

Then the overall confidence is $b = 1 - E_{avg}/logC$, where $logC$ is the maximal entropy value for normalization. The calibrated ADI score (ADIC) q is:

$$q(X) = b \cdot ADI(X) = (1 - \frac{E_{avg}}{logC}) \cdot ADI(X), \tag{3}$$

where q is the calibrated estimation of prediction quality, and used to select well-predicted samples and guide subsequent adaptation to the current sample.

2.2 Adaptive Feature Fusion-based Refinement based on ADIC

Since previous works have shown that utilizing robust features from well-predicted images can minimize the domain disparity for other test images [27], we propose an Adaptive Feature Fusion-based Refinement (AFFR) strategy, which uses ADIC to filter historically well-predicted samples to help adaptation.

Firstly, we maintain a dynamic feature bank $\mathcal{F}$ of fixed capacity L, updated with a first-in-first-out policy to remain adaptive to the recent distribution of test samples. Specifically, let θ_t^e denote the encoder of θ_t, the feature for the i-th image X_i is denoted as $z_i = f(\theta_t^e, X_i)$. Then z_i is pushed to $\mathcal{F}$ if $q(X_i)$ exceeds the τ percentile of all previous samples' ADIC values before time step t.

Secondly, for a new test sample, we use $sim(z_i, z_l)$ to denote the cosine similarity between z_i and the l-th feature z_l in $\mathcal{F}$. Then z_i is updated as a linear

combination of z_i and z_l $(l = 0, 1, ..., L-1)$. The weight of z_i is set as $q(X_i)$ so that preservation of the original feature is encouraged for well-predicted samples, and z_i is replaced by $z_f = \sum_{l=0}^{L-1} w_{i,l} \cdot z_l$ for poorly predicted ones, where $w_{i,l} = sim(z_i, z_l) / \sum_l sim(z_i, z_l)$.

$$z_i' = q_i z_i + (1 - q_i) z_f \tag{4}$$

where $q_i = q(X_i)$. z_i' is the refined feature of z_i, and sent to decoder θ_t^d of the adapted model before t-th step to obtain a refined prediction $P_i' = f(\theta_t^d, z_i')$, which is used as a refined pseudo-label for X_i.

2.3 ADIC-guided Self-adaptive Model Updating

Though AFFR can lead to relatively reliable pseudo labels, the gradient may fluctuate largely on different testing batches, leading to instability during adaptation. Following the common practice of TTA [20,22], we adopt a mean teacher model ϕ to improve stability. Considering that traditional mean teacher [19] with a constant Exponential Moving Average (EMA) rate (e.g., 0.9) is insufficient to accommodate the dynamic data quality changes in TTA, we introduce an ADIC-aware mean teacher with an adaptive EMA rate. Given the updated parameters of the student θ_{t+1} at time step t, we can get updated teacher ϕ_{t+1} as:

$$\phi_{t+1} = q_i \theta_{t+1} + (1 - q_i) \phi_t \tag{5}$$

Finally, to deal with potential noise in P_i', we also weight loss by q_i to suppress the contribution of poorly adapted samples. The total loss for our TEGDA is:

$$L_{TTA} = \frac{1}{B} \sum_{i=1}^{B} \left(L_{mt}(P_i'', P_i) + q_i L_{re}(P_i', P_i) \right) \tag{6}$$

where P_i'' is prediction from the teacher model. L_{mt} and L_{re} are the mean-teacher loss and refined pseudo-label loss, respectively, and they are implemented by the commonly used combination of Dice loss and cross-entropy loss. For each batch, a single back-propagation using L_{TTA} is conducted, followed by a forward propagation with the updated student θ_{t+1} to obtain the final predictions.

3 Experiment and Results

Datasets and Implementation Details We used two public multi-domain datasets for experiments: 1) **M&MS dataset** [4] for heart structure segmentation that consists of 345 3D MRI volumes collected from four different vendors (identified as Domain A, B, C, and D), with three segmentation classes: Left Ventricle (LV), Right Ventricle (RV), and Myocardium (Myo). The number of slices per volume ranges from 10 to 13. Due to the large slice spacing (9.2 to 10 mm), we used 2D U-Net [17] for slice-level segmentation, with each slice resized to 320×320. Domain A was set as the source domain, while Domain B, C, and

Table 1. Comparison between different TTA methods on the M&MS Dataset. The first and second sections represent volume-level Dice (%) and HD95 (mm), respectively. †denotes a significant improvement (p-value < 0.05) over the best existing method.

Method	Domain B				Domain C	Domain D
	LV	Myo	RV	Average	Average	Average
Source	85.64±9.24	69.82±10.72	75.68±18.70	77.05±11.22	69.17±14.96	79.93±10.13
TENT [21]	88.13±7.87	75.01±8.12	78.86±13.24	80.67±8.49	77.2±11.23	80.36±8.37
SAR [23]	88.09±7.88	74.95±8.16	79.04±13.04	80.69±8.42	77.09±11.36	80.36±8.6
CoTTA [22]	88.02±8.06	74.94±8.09	78.99±12.98	80.65±8.46	77.23±11.45	80.34±8.51
InTEnt [7]	86.12±8.93	70.64±10.43	75.98±17.96	77.58±10.83	70.16±14.6	80.17±9.97
VPTTA [5]	87.65±8.55	75.83±7.60	79.05±13.77	80.84±8.54	78.43±10.25	81.13±9.2
TEGDA	**89.71±6.86**†	**78.33±6.18**†	**83.30±9.50**†	**83.78±6.17**†	**79.34±10.42**†	**82.04±7.65**†
Source	16.18±22.55	20.95±23.16	8.68±15.31	15.27±14.74	26.7±21.62	16.64±17.9
TENT [21]	19.93±24.41	10.08±14.68	8.45±14.93	12.08±13.89	20.0±19.87	11.35±15.84
SAR [23]	17.78±24.96	13.13±18.47	11.19±19.51	14.7±15.74	25.08±20.89	13.66±14.16
CoTTA [22]	18.71±24.03	9.13±14.00	8.42±14.66	14.03±15.83	19.69±19.18	11.68±13.88
InTEnt [7]	16.19±23.76	21.16±24.06	9.74±19.54	15.09±14.24	25.75±21.01	15.06±15.93
VPTTA [5]	17.14±24.23	9.21±14.91	9.89±18.33	12.82±13.23	17.42±16.03	9.19±12.28
TEGDA	**11.38±21.12**†	**8.76±16.28**†	**6.04±12.99**†	**8.73±11.85**†	**16.76±17.7**†	**8.5±12.76**†

D as target domains independently. 2) **BraTS2023 dataset** [1, 2, 11, 14] that is sourced from different patient groups, and focuses on brain tumor segmentation including Whole Tumor (WT), Tumor Core (TC), and Enhanced Tumor (ET). Adult glioma dataset (BraTS-GLI) [1, 2, 14] including 1251 patients was set as the source domain, and pediatric brain tumor dataset (BraTS-PED) [11] including 99 patients was set as the target domain. Each patient includes volumetric images in four modalities resampled to a uniform isotropic resolution ($1mm^3$). We resized them to a size of $4 \times 128 \times 128 \times 128$, and employed the 3D U-Net [6].

For both datasets, images were linearly normalized to $[0, 1]$. The source model was trained in the source domain for 200 epochs using the Dice loss and Adam optimizer with a learning rate of 1.0×10^{-3}, and the checkpoint performed best in validation was used for adaptation. During adaptation, we set the batch size to 10 for M&MS and 1 for BraTS2023, employing a single epoch with Adam optimizer and a learning rate of 1.0×10^{-4}. In TEGDA, we set a dropout rate of 0.5, the dropout number $M = 10$, the feature bank length $L = 10$ and the percentile $\tau = 90$ for sample filtering. The experiments were conducted with PyTorch 1.8.1 and a GeForce RTX 3090. We used volume-level Dice similarity and 95 percentile of Hausdorff Distance (HD95) for evaluation.

Comparison with State-of-the-art TTA Methods Our TEGDA was compared with five state-of-the-art TTA methods: 1) **TENT** [21] that updates parameters of BN layers by entropy minimizing; 2) **SAR** [23] that uses sharpness-aware entropy minimization; 3) **CoTTA** [22] that uses a mean-teacher with test-time augmentation-based pseudo-labels; 4) **InTEnt** [7] that uses an ensemble of multiple adapted models based on different estimates of the target domain

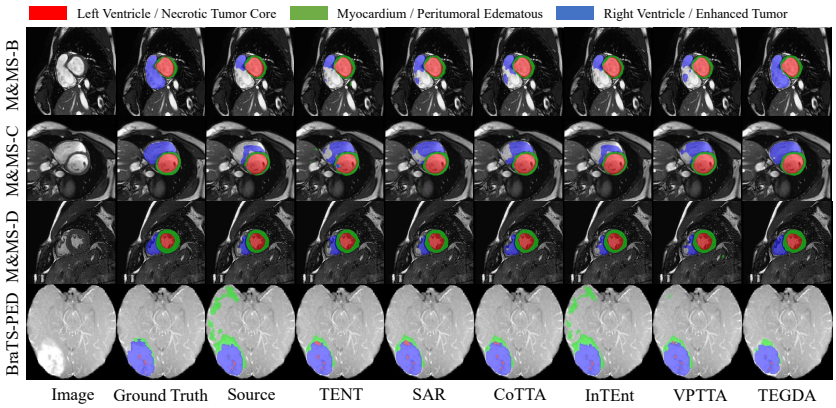

Fig. 2. Qualitative comparison of different TTA methods on two datasets.

Table 2. Comparison between different TTA methods on the BraTS2023 dataset. †denotes a significant improvement (p-value < 0.05) over the best existing method.

Method	Dice (%)				HD95 (mm)
	WT	TC	ET	Average	Average
Source	78.56±21.78	29.65±30.27	51.29±37.43	53.17±22.46	24.05±28.61
TENT [21]	**85.14±14.06**	32.08±30.47	51.76±35.97	56.33±20.86	19.71±24.83
SAR [23]	84.99±14.67	31.90±30.45	49.23±36.24	55.38±21.42	19.79±25.29
CoTTA [22]	84.89±14.80	31.95±30.49	50.22±36.39	55.69±21.21	20.01±25.15
InTEnt [7]	79.27±20.92	29.86±30.25	51.55±37.42	53.56±22.23	23.43±27.00
VPTTA [5]	84.28±14.4	31.15±29.88	51.82±36.26	55.75±20.47	22.59±26.30
TEGDA	84.35±17.08	**35.19±31.86**†	**55.26±36.59**†	**58.27±19.74**†	**16.28±24.59**†

statistics; 5) **VPTTA** [5] that adapts by image-specific prompts in the frequency domain. 'Source' means inference with the source model without adaptation.

Table 1 shows the quantitative results on the M&MS dataset. The source model obtained an average Dice of 77.05% on Domain B, while that of existing methods ranged from 77.58% to 80.84%. Our TEGDA got the highest Dice of 83.78%, which is 6.73 percentage points higher than the baseline and significantly better than the existing methods. Besides, TEGDA obtained the highest Dice and lowest HD95 on all three domains, outperforming the other methods.

Table 2 further shows the comparison of these methods for adapting 3D segmentation models on the BraTS2023 dataset. It's worth noting that there is a significant domain shift between BraTS-GLI and BraTS-PED due to the complex structure of brain tumors and the appearance difference between different age groups, resulting in poor performance of the source model. Nevertheless, our TEGDA outperformed existing methods, improved average Dice from 56.33% to 58.27% compared with the best existing method, and obtained the lowest HD95.

The qualitative results shown in Fig. 2 demonstrate that existing methods often result in under-segmentation or over-segmentation, while our TEGDA is superior in accurately delineating the target regions in different domains.

Table 3. Ablation study of TEGDA on different datasets. AFFR=◇ means using entropy for selecting well-predicted samples instead of ADIC. SMU=◇ means setting EMA rate and weight of L_{re} to fixed values (0.9 and 1.0, respectively).

L_{mt}	L_{re}	AFFR	SMU	M&MS Domain B			BraTS-PED		
				LV	Myo	RV	WT	ET	TC
				85.64±9.24	69.82±10.72	75.68±18.70	78.56±21.78	29.65±30.27	51.29±37.43
✓				87.97±8.00	75.49±8.41	78.96±13.05	84.07±16.10	31.37±31.25	49.64±37.28
	✓	◇	◇	87.62±7.92	76.08±7.72	81.08±10.73	78.73±22.60	35.27±32.20	52.68±37.21
✓	✓	◇	◇	88.12±7.62	76.57±7.43	81.56±10.45	79.75±21.04	**35.59±32.21**	53.37±36.98
✓	✓	✓	◇	89.32±7.15	77.41±7.13	82.63±10.21	83.27±17.57	35.34±31.91	53.26±36.75
✓	✓	✓	✓	**89.71±6.86**	**78.33±6.18**	**83.30±9.50**	**84.35±17.08**	35.19±31.86	**55.26±36.59**

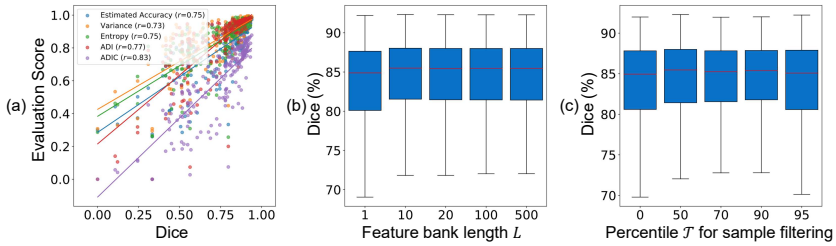

Fig. 3. (a) Comparison between ADIC and other prediction quality evaluation methods, where r is Pearson correlation coefficient between the evaluation score and real Dice; (b) Effect of feature bank length L on average Dice. (c) Effect of percentile τ for sample filtering on average Dice. All subfigures show results on Domain B of M&MS.

Ablation Study Quantitative results of the ablation study of our TEGDA framework on the two datasets are detailed in Table 3. The baseline method was using the source model for inference. We gradually added different components including L_{mt}, L_{re}, AFFR and SMU. It can be observed on both datasets that the incorporation of each component contributed to performance improvement.

Besides, for estimating the Dice with the absence of ground-truth on testing images, we compare our ADIC with ADI and some alternative methods, including Entropy [16], Variance [25] and Estimated Accuracy [12] based on MC dropout. The evaluation scores of these methods were linear normalized to [0,1] for a more intuitive comparison. Fig. 3 (a) shows that our ADIC is highly correlated with the real Dice compared to the other four methods, with a correlation coefficient of 0.83. In addition, our TEGDA only has two main hyper-parameters, i.e., the feature bank length L and percentile τ for sample filtering. Fig. 3 (b) shows that the adaptation performance is insensitive to L when $L \geq 10$, and Fig. 3 (c) indicates that either excessively high or excessively low τ will impact performance, and $\tau = 90$ performs best for AFFR.

4 Conclusion

We proposed a novel Test-Time Evaluation-Guided Dynamic Adaptation (TEGDA) framework for medical image segmentation, addressing the critical issue of unreliable supervision of existing TTA methods. TEGDA leverages a novel metric based on Agreement with Dropout Inference calibrated by Confidence (ADIC) to reliably evaluate the prediction quality. ADIC is used to adaptively fuse features of a sample with those with high prediction quality for refinement, and ADIC-aware pseudo-label loss weighting and ADIC-aware mean teacher are used for robust model adaptation. Experiments on both 2D and 3D datasets demonstrated its robustness and superiority over existing TTA methods. In the future, it is of interest to extend TEGDA to different applications in medical images.

Acknowledgements. This work was supported by the Natural Science Foundation of Sichuan Province under grant 2025ZNSFSC0455.

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

References

1. Baid, U., Ghodasara, S., Mohan, S., Bilello, M., Calabrese, E., Colak, E., Farahani, K., Kalpathy-Cramer, J., Kitamura, F.C., Pati, S., et al.: The RSNA-ASNR-MICCAI BRATS 2021 benchmark on brain tumor segmentation and radiogenomic classification. arXiv:2107.02314 (2021)
2. Bakas, S., Akbari, H., Sotiras, A., Bilello, M., Rozycki, M., Kirby, J.S., Freymann, J.B., Farahani, K., Davatzikos, C.: Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. Scientific data **4**(1), 170117 (2017)
3. Basak, H., Yin, Z.: Quest for clone: test-time domain adaptation for medical image segmentation by searching the closest clone in latent space. In: MICCAI. pp. 555–566. Springer (2024)
4. Campello, V.M., Gkontra, P., Izquierdo, C., Martin-Isla, C., Sojoudi, A., Full, P.M., Maier-Hein, K., Zhang, Y., He, Z., Ma, J., et al.: Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. IEEE Transactions on Medical Imaging **40**(12), 3543–3554 (2021)
5. 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)
6. Çiçek, Ö., Abdulkadir, A., Lienkamp, S.S., Brox, T., Ronneberger, O.: 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: MICCAI. pp. 424–432. Springer (2016)
7. Dong, H., Konz, N., Gu, H., Mazurowski, M.A.: Medical image segmentation with InTEnt: integrated entropy weighting for single image test-time adaptation. In: CVPR Workshops. pp. 5046–5055 (2024)
8. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: representing model uncertainty in deep learning. In: ICML. pp. 1050–1059. PMLR (2016)

9. 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)
10. Jiang, Y., Nagarajan, V., Baek, C., Kolter, J.Z.: Assessing generalization of SGD via disagreement. arXiv:2106.13799 (2021)
11. Kazerooni, A.F., Khalili, N., Liu, X., Haldar, D., Jiang, Z., Anwar, S.M., Albrecht, J., Adewole, M., Anazodo, U., Anderson, H., et al.: The brain tumor segmentation (BRATS) challenge 2023: Focus on pediatrics (CBTN-CONNECT-DIPGR-ASNR-MICCAI BraTS-PEDs). arXiv:2305.17033 (2023)
12. Lee, T., Chottananurak, S., Gong, T., Lee, S.J.: AETTA: label-free accuracy estimation for test-time adaptation. In: CVPR. pp. 28643–28652 (2024)
13. Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I.: A survey on deep learning in medical image analysis. Medical Image Analysis **42**, 60–88 (2017)
14. Menze, B.H., Jakab, A., Bauer, S., Kalpathy-Cramer, J., Farahani, K., Kirby, J., Burren, Y., Porz, N., Slotboom, J., Wiest, R., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2014)
15. Nado, Z., Padhy, S., Sculley, D., D'Amour, A., Lakshminarayanan, B., Snoek, J.: Evaluating prediction-time batch normalization for robustness under covariate shift. arXiv:2006.10963 (2020)
16. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: ICML. pp. 16888–16905. PMLR (2022)
17. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI. pp. 234–241. Springer (2015)
18. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: ICML. pp. 9229–9248. PMLR (2020)
19. Tarvainen, A., Valpola, H.: Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In: NeurIPS. p. 1195–1204 (2017)
20. Tomar, D., Vray, G., Bozorgtabar, B., Thiran, J.P.: Tesla: test-time self-learning with automatic adversarial augmentation. In: CVPR. pp. 20341–20350 (2023)
21. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: fully test-time adaptation by entropy minimization. In: ICLR (2021)
22. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: CVPR. pp. 7201–7211 (2022)
23. Wang, W., Zhong, Z., Wang, W., Chen, X., Ling, C., Wang, B., Sebe, N.: Dynamically instance-guided adaptation: A backward-free approach for test-time domain adaptive semantic segmentation. In: CVPR. pp. 24090–24099 (2023)
24. 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. pp. 237–249 (2023)
25. Wu, J., Guo, D., Wang, G., Yue, Q., Yu, H., Li, K., Zhang, S.: FPL+: filtered pseudo label-based unsupervised cross-modality adaptation for 3D medical image segmentation. IEEE Transactions on Medical Imaging **43**(9), 3098–3109 (2024)
26. 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)

27. Zheng, B., Zhang, R., Diao, S., Zhu, J., Yuan, Y., Cai, J., Shao, L., Li, S., Qin, W.: Dual domain distribution disruption with semantics preservation: Unsupervised domain adaptation for medical image segmentation. Medical Image Analysis **97**, 103275 (2024)