Single Image Test-Time Adaptation via Multi-View Co-Training

Smriti Joshi[1], Richard Osuala[1], Lidia Garrucho[1], Kaisar Kushibar[1], Dimitri Kessler[1], Oliver Diaz[1,2], and Karim Lekadir[1,3]

[1] Departament de Matemàtiques i Informàtica, Universitat de Barcelona, Spain
smriti.joshi@ub.edu
[2] Computer Vision Center, Bellaterra, Spain
[3] Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluís Companys 23, Barcelona, Spain

Abstract. Test-time adaptation enables a trained model to adjust to a new domain during inference, making it particularly valuable in clinical settings where such on-the-fly adaptation is required. However, existing techniques depend on large target domain datasets, which are often impractical and unavailable in medical scenarios that demand per-patient, real-time inference. Moreover, current methods commonly focus on two-dimensional images, failing to leverage the volumetric richness of medical imaging data. Bridging this gap, we propose a Patch-Based Multi-View Co-Training method for Single Image Test-Time adaptation. Our method enforces feature and prediction consistency through uncertainty-guided self-training, enabling effective volumetric segmentation in the target domain with only a single test-time image. Validated on three publicly available breast magnetic resonance imaging datasets for tumor segmentation, our method achieves performance close to the upper bound supervised benchmark while also outperforming all existing state-of-the-art methods, on average by a Dice Similarity Coefficient of 3.75%. We will publicly share our accessible codebase, readily integrable with the popular nnUNet framework, at https://github.com/smriti-joshi/muvi.git.

Keywords: Unsupervised Domain Adaptation · Segmentation · Test-Time Training · Breast MRI

1 Introduction

Domain shift in medical imaging presents a significant and well-known challenge that restricts the real-world applicability of automated methods due to the distribution gap between training and testing domains. This issue is attributed to differences in acquisition hardware, field strength, reconstruction software, and imaging protocols. The differences may be exhibited in the overall quality and appearance, signal intensities and contrast, resolution and slice thickness, and field of view and acquisition plane of the images. While clinicians are able to readily cope with domain shifts and account for them in their assessments, it is detrimental for deep learning-based automatic methods. Typically, transfer

learning [12][27] and unsupervised domain adaptation methods [11][3] are used to alleviate the domain shift problem. However, both approaches rely on retraining with large target domain datasets at every domain shift occurrence. This requires knowledge of the specific type and magnitude of the shift, along with annotations that are consistent and free from inter-observer variability [10]. For these reasons, test-time domain adaptation methods have gained momentum in recent years, as they enable the model to adapt dynamically to various types of unforeseen domain shifts during inference in an unsupervised manner [13].

The existing test-time adaptation methods can be roughly divided into the following four categories. *Auxiliary-Task* based approaches, first introduced in Test-Time Training (TTT) [24], have a Y-shape model architecture for pretraining, with a shared feature extractor between the main task (e.g. classification, segmentation) and a self-supervised proxy task, i.e. rotation prediction. At test time, only the auxiliary task is trained to adjust the feature extractor to target domain [16,21]. While effective, these methods rely on substantial modifications to the model architectures for training with source domain data. *Normalization* based methods reduce dependence on source training by adapting the network through modifications to normalization layers. For instance, PTN [18] uses the test batch statistics for batch normalization (BN) layers during inference. BNAdapt [23] adjusts the pretrained BN layers by weighted averaging of source and target domain statistics. Tent [25] and FSEG [7] update the affine parameters of BN layers through entropy minimization and contour regularization objectives respectively. This category of methods heavily relies on BN and particularly, a large batch size to estimate the target domain statistics. InTent [2] tackles this limitation in single-image setups by sampling between training and test BN statistics in an entropy-weighted manner. While its backpropagation-free design improves efficiency, it also limits the extent of model adaptation. Next, *Self-Training* based methods train during test-time via pseudolabels. For example, SHOT [14] adapts the feature extractor through self-training while keeping the classifier frozen. MEMO [28] and AugMix [5] use prediction consistency between augmentations. These methods are often sensitive to label noise and can accumulate errors. Finally, *Prototype-based* methods adapt by reducing the distance to class prototypes in feature space [9,14]. However, they rely on consistent region-of-interest shapes, limiting their effectiveness in cases with high anatomical variability. Current approaches often require large batches of data, rely on shape priors, are typically validated on organ-specific tasks, and notably fail to exploit the full three-dimensional (3D) context, despite operating on volumetric modalities.

Fang et al. [4] distill information between different views of 3D images to guide the learning process for an image synthesis task. A similar approach pretrains swin transformers [26] through a cross attention block between different views. Inspired by these works, we introduce a novel *source-free* method to adapt the network during test-time on a *single image* for *shape-variable* tumor segmentation task through *Multi-View Co-Training*. In summary, our contributions include:

1. Designing a novel source-free test-time domain adaptation method based on uncertainty-informed patch-based multi-view co-training.
2. Conducting a comprehensive evaluation on three heterogeneous, publicly available breast MRI datasets for tumor segmentation, where it surpasses previous state-of-the-art methods and achieves performance close to the supervised benchmark.
3. Examining the impact of normalization layers on reducing generalization error in the single image test-time adaption setting.

2 Methods

2.1 Problem Definition

Given a model $M(\theta_s)$ with parameters θ_s trained to learn function $f : X_s \to Y_s$ on source data s, our goal is to adapt $M(\theta_s)$ to the target domain t by learning the new function $k : X_t \to Y_t$, without access to the original source data nor the ground truth labels of the target samples. Here X is a 3D tensor and the ground truth Y is a densely-labeled voxel-wise 3D segmentation mask. Note that $M(\theta_s)$ does not learn continually and is reset to θ_s for each $x_t \in X_t$, as is demonstrated in Figure 1(a). Our decision to work on single images is motivated by the realistic scenario in hospitals where inference is needed on-demand and per-patient. Even if a delay in collecting the incoming batch is acceptable, a single target distribution cannot be guaranteed in the case of a single-center multi-scanner setup, rendering many existing techniques ineffective. Finally, medical imaging data consists of high-resolution volumes, requiring GPU processing with large VRAM capacities, making large batches expensive and resource-intensive.

2.2 Test Time Adaptation

Given target data $x_t \in X_t$ and model $M(\theta_s)$ having learnt f_{θ_s}, we use a combination of three distinct techniques to learn function k_{θ_t}.

Modulation parameters and BN statistics: All model parameters θ are adapted in a single epoch. We apply batch normalization in two steps: (i) normalization by current batch statistics, $\hat{x} = (x - \mu)/(\sqrt{\sigma^2 + \epsilon})$ where μ and σ are mean and variance of the batch, and ϵ is a small constant for numerical stability and (ii) transformation by learnable parameters γ and β as $y = \gamma\hat{x} + \beta$. In multiple domain-shift scenarios such as changes in acquisition protocol or scanner types, BN statistics from the source network trained with diverse data can still provide a reasonable estimate of the target distribution as opposed to modifying them based on patches of a single image, which can adversely affect model performance. Therefore, we keep the original mean μ_s and variance σ_s obtained during pre-training while adapting parameters γ and β through gradient propagation of the loss function.

Multi-view Co-Training: Building upon intuition gathered from previous works [4,26], we propose patch-based multi-view training as a measure to adapt images during test time, thereby effectively mitigating the impact of diverse and

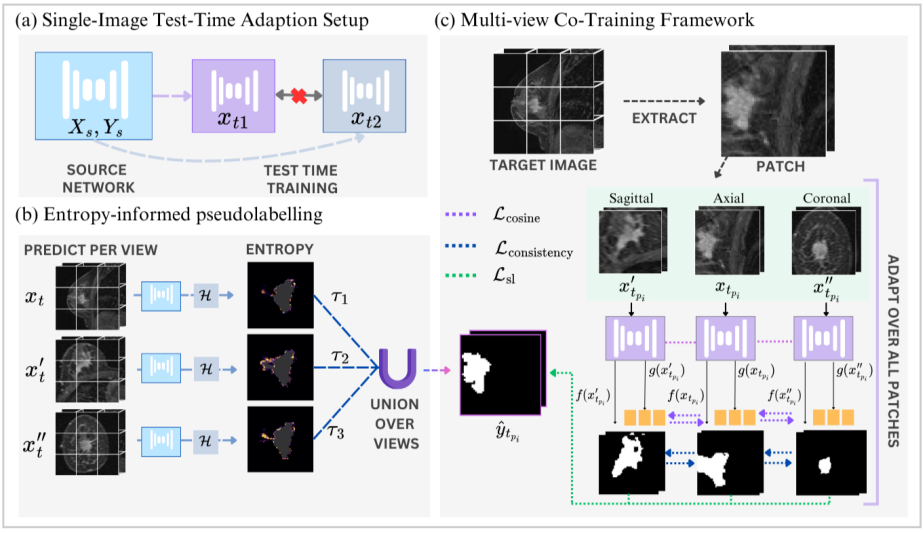

Fig. 1: Pipeline of the proposed method MuVi. (a) Our setup where the source network is adapted independently for each target sample x_t, (b) computing pseudolabel through entropy-threshold union of prediction from each view, (c) Adaptation via patch-based training through our multi-view consistency framework.

a priori unknown domain shifts. To this end, for a test case x_t, we obtain the corresponding overlapping patches $\{x_{t_{p_1}}, x_{t_{p_2}}, \ldots, x_{t_{p_n}}\}$, where n is the total number of patches. Each patch x_{tpi} is permuted to obtain corresponding representations from two alternate views: $x'_{tpi} = \pi_1(x_{tpi})$ and $x''_{tpi} = \pi_2(x_{tpi})$, where π_1 and π_2 denote view-specific permutation functions. We denote these three views collectively as $v \in \{x_{tpi}, x'_{tpi}, x''_{tpi}\}$, representing the original patch and its two transformed versions. The self-learning objective, $\mathcal{L}_{\text{sl}}$, is then minimized for each of these patches with pseudolabel $\hat{y}_{tp_i}$ (obtained as explained in the next section) as follows:

$$\mathcal{L}_{\text{sl}} = \sum_{v \in \{x_{t_{p_i}}, x'_{t_{p_i}}, x''_{t_{p_i}}\}} \left[\mathcal{L}_{\text{DICE}}(f(v), \hat{y}_{t_{p_i}}) + \mathcal{L}_{\text{CE}}(f(v), \hat{y}_{t_{p_i}})\right],$$

where DICE and CE stand for Dice Loss and Cross Entropy Loss, respectively. We further integrate a consistency constraint between views through $\mathcal{L}_{\text{consistency}}$, by minimizing the same objective as above but between predictions from transformed views $x'_{t_{p_i}}$, $x''_{t_{p_i}}$ and the original view $x_{t_{p_i}}$. Additionally, we add a consistency term between feature embeddings of different views to ensure alignment independent of the pseudolabels. Specifically, we minimize

$$\mathcal{L}_{\text{cosine}} = 1 - \cos(g(x_{t_{p_i}}), g(x'_{t_{p_i}})) + 1 - \cos(g(x_{t_{p_i}}), g(x''_{t_{p_i}})),$$

where $g(\cdot)$ represents the feature extractor of the network. Consequently, the final loss is then given by $\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{sl}} + \lambda_2 \mathcal{L}_{\text{consistency}} + \lambda_3 \mathcal{L}_{\text{cosine}}$, where we

define $\lambda_1, \lambda_2, \lambda_3$ to calibrate the relative importance of each loss. In our baseline implementation of this framework, we weigh all losses equally.

Entropy-guided Self-Training: To introduce complementary information from axial, coronal, and sagittal views, we compute an uncertainty-informed combination of corresponding predictions. This entropy-based pseudolabel is computed at the image level prior to training as opposed to the aforementioned patch-level multi-view co-training formulation. Specifically, the per-pixel entropy $H(z)$ is calculated using the model's predicted probability $z \in [0, 1]$ for the foreground class at each pixel, where $H(z) = -z \log_2(z) - (1-z) \log_2(1-z)$. Then, the prediction corresponding to each view is subjected to an entropy threshold τ, representing the minimum confidence required for accepting the prediction. These thresholds are empirically set on the validation, with values explored within the range of $[0.1, 0.6]$. A higher threshold (0.4) is applied to the view with the highest resolution, while a stricter threshold (0.2) is used for the other planes. The final entropy-based pseudolabel is given by the respective union as:

$$\hat{y} = \bigcup_{v \in \{x_t, x_t', x_t''\}} \{j \mid H(\sigma(f(v(j))) < \tau_v\},$$

where σ is the sigmoid function and j refers to the image pixel.

3 Experiments and Results

3.1 Dataset and Implementation

Dataset: To train and validate our proposed method, we extract the first phase of the T1-weighted dynamic contrast-enhanced (DCE) sequence from three breast MRI datasets [6][1]: 1) *Duke-Breast-Cancer-MRI* [22], collected between 2000-2014 and includes MRI scans of 922 patients acquired at a single center in the United States. In this work, we only use the first phase of the DCE series from 254 cases for which the segmentation masks are available from [6]. The DCE MRIs are acquired in the axial plane; 2) TCGA-BRCA [15], collected from 1999 - 2004, contains MRI scans of 80 patients from four different centers in the United States. The acquisition planes of the DCE MRIs are either sagittal or axial; 3) ISPY1 [19], collected from 2002 - 2006, contains MRI scans of 161 patients from a single center in the United States, with DCE MRIs acquired in the sagittal plane. Additional dataset statistics are shown in Table 1. We adopt the Duke-Breast-Cancer-MRI as the source training dataset due to it having the highest number of cases, alongside varied protocols and scanners.

Setup: We train with unilateral breasts due to a lack of bilateral tumor annotations in the datasets. For each dataset, 30 cases are reserved for testing, and the remaining cases are used for model training with an 80:20 train-validation split. We use the 3D UNet from nnUNet [8] as the base framework for source training. This decision is motivated by the need to establish a strong baseline, utilizing the framework's robust pre-processing, prediction, and post-processing pipeline. The following modifications are made to the original pipeline: (i) the model is

Table 1: **Overview of datasets used in this work.** Each dataset, containing variable vendors, scanners, and protocols, is treated as an independent domain.

Dataset	Vendors	Scanners	Pixel Spacing			Slice Thickness		
			mean	std	median	mean	std	median
Duke-Breast-Cancer MRI	GE, Siemens	7	0.73	0.11	0.70	1.07	0.15	1.00
ISPY1	GE, Siemens, Philips	7	0.76	0.14	0.78	2.42	0.54	2.30
TCGA-BRCA	GE	3	0.68	0.10	0.66	2.17	0.29	2.00

trained for 500 epochs, and (ii) the normalization layers are changed from *Instance Normalization* to *Batch Normalization*. The models are evaluated using three widely used segmentation metrics [17]: Dice Similarity Coefficient (DSC), Hausdorff Distance (HD) and Average Asymmetric Surface Distance (ASD). All models are trained on a single NVIDIA GeForce RTX 4080 16 GB GPU.

3.2 Results

The results of the experiments on two datasets, TCGA-BRCA and ISPY1, are demonstrated in Tables 2 and 3 respectively, where our method shows the largest improvement of 5.57% and 4.13% over the baseline in terms of DSC.

Baselines and Supervised Bounds: We compute two baselines, (i) anisotropic patch size of [112, 224, 96] (as computed by nnUNet preprocessing pipeline based on the input data characteristics) and (ii) isotropic patch size of [128, 128, 128]. By providing spatial consistency and avoiding directional biases through an isotropic input, nnUNet improves on the target domain test set by an average of 7.4% DSC. We use this baseline for performing test time adaptation with all methods. We also train supervised networks for computing two upper bounds, target only and source plus target, the latter achieving better performance, likely because of its larger training cohort.

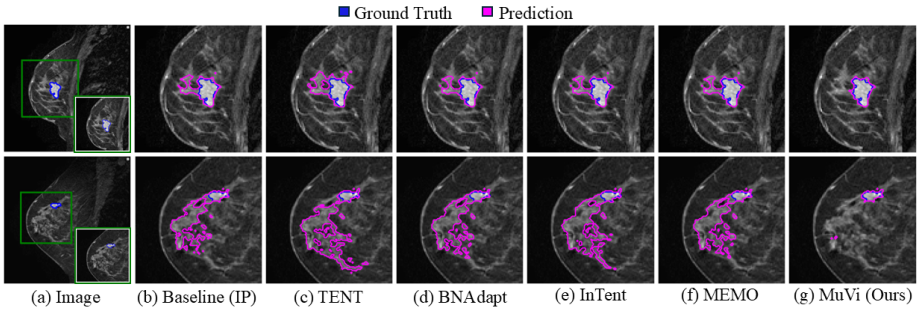

Fig. 2: Qualitative segmentation results from different methods. Our method localizes the tumor precisely while removing the misidentified breast tissue.

Table 2: Results of the mentioned methods on **TCGA-BRCA Dataset**. T refers to target only, S + T refers to source and target data combined. Standard deviation showcases the variability between patients in the test set. AP and IP refer to 3D UNet with anisotropic and isotropic patches. Best results in bold.

Metric	**DSC ↑**	**HD↓**	**ASD↓**
Upper Bound			
Supervised (T)	0.6648 ± 0.2558	22.4165 ± 30.8837	16.1690 ± 26.1061
Supervised (S + T)	0.7008 ± 0.2616	15.4179 ± 37.9761	11.6789 ± 29.1357
Baseline			
3D UNet (AP)	0.5925 ± 0.3139	16.2708 ± 30.6949	14.4209 ± 29.4008
3D UNet (IP)	0.6254 ± 0.269	16.6877 ± 21.1614	12.5212 ± 16.7173
Test-Time Adaptation			
PTN [18]	0.5187 ± 0.2764	35.2485 ± 32.1232	26.5875 ± 24.8109
Tent [25]	0.5187 ± 0.2764	35.2492 ± 32.1239	26.5879 ± 24.8103
BNAdapt [23]	0.6418 ± 0.2509	17.9102 ± 22.42	13.3762 ± 17.5268
InTent [2]	0.6228 ± 0.2689	18.8721 ± 21.9771	14.3614 ± 17.3471
MEMO [28]	0.6435 ± 0.2497	17.787 ± 22.4028	13.2727 ± 17.5068
MuVi (ours)	**0.6811 ± 0.2269**	**15.9166 ± 25.6073**	**11.45655 ± 20.1114**
Ablation: MuVi			
w/o source BN statistics	0.3813 ± 0.3262	57.1691 ± 48.2155	52.2985 ± 47.3137
w/o entropy-weighted labels	0.6341 ± 0.2652	22.7417 ± 32.4716	17.9036 ± 27.7267
w/o consistency	0.6745 ± 0.227	17.1451 ± 28.8524	12.8935 ± 23.8826

Comparison with Other Methods: We selected five state-of-the-art methods for detailed comparison on the clinically highly relevant tumor segmentation task. Auxiliary task-based methods (dependence on architecture) as well as prototype-based methods (dependence on shape) are excluded. We select the most popular state-of-the-art normalization-based methods, namely PTN [18], Tent [25], BNAdapt [23], and InTent [2]. We also include the self-training based method MEMO [28], which is closest to our approach since it adds consistency between augmentations during test time training. BNAdapt, InTent and MEMO are tested for single image adaptation in their original implementation and corresponding settings are used for comparison. Table 2 and 3 show that our method outperforms existing techniques, demonstrating the effectiveness of our proposed entropy-based pseudolabels informed by three views and reduced reliance on BN layers. In Figure 2, we further provide a visual comparison, demonstrating how our method effectively distinguishes the tumor from surrounding breast tissue, accurately localizing its boundaries.

Ablation of Entropy Weighted Pseudolabel and Consistency: As shown in Table 2, directly averaging predictions from all views without accounting for uncertainty still improves on the baseline. However, since the baseline is not trained on multiple views, selectively propagating only the confident labels from complementary lower-resolution views is essential for achieving substantial improvements. In our case, this approach leads to a 5.4% increase in DSC and an approximately 7-point reduction in HD. Enforcing consistency across views enhances feature robustness to varying perspectives and pixel resolutions, resulting in further refinement of results in just one epoch.

Table 3: Results of the mentioned methods on **ISPY1 Dataset**. T Refers to target only, S + T refers to source and target data combined. Standard deviation showcases the variability between patients in the test set. AP and IP refer to 3D UNet with anisotropic and isotropic patches. Best results in bold.

Metric	DSC ↑	HD ↓	ASD ↓
Upper Bound			
Supervised (T)	0.7010 ± 0.2324	9.7768 ± 18.0119	8.1711 ± 15.7302
Supervised (S + T)	0.7248 ± 0.2191	8.8684 ± 14.5787	6.3187 ± 9.8694
Baseline			
3D UNet (AP)	0.5426 ± 0.3136	28.1163 ± 58.3449	21.1092 ± 43.7914
3D UNet (IP)	0.6586 ± 0.2304	12.8004 ± 25.9020	10.5284 ± 24.7581
Test-Time Adaptation			
PTN [18]	0.6284 ± 0.2188	19.9790 ± 27.8741	14.3163 ± 19.3467
Tent [25]	0.6284 ± 0.2188	19.9792 ± 27.8760	14.3162 ± 19.3471
BNAdapt [23]	0.6559 ± 0.2238	13.5064 ± 25.9386	10.9942 ± 24.9634
InTent [2]	**0.6606 ± 0.2287**	10.0681 ± 14.7725	7.9732 ± 12.4688
MEMO [28]	0.6586 ± 0.2254	12.8885 ± 26.0073	10.6552 ± 25.0241
MuVi (ours)	0.6588 ± 0.2362	**8.7271 ± 14.9046**	**6.9601 ± 12.0732**
Experiment: Instance Normalization			
3D UNet (AP)	0.5823 ± 0.2563	21.3682 ± 31.9623	16.1128 ± 23.8279
3D UNet (IP)	0.6564 ± 0.2625	9.3165 ± 16.4794	7.5948 ± 13.8091
MEMO [28]	0.6568 ± 0.2623	9.2749 ± 16.4329	7.5669 ± 13.7739
MuVi (ours)	**0.7001 ± 0.2349**	**6.1374 ± 11.2560**	**5.1850 ± 9.7773**

Effect of Normalization: Emphasizing the importance of using source domain BN statistics during testing, we observe in metrics obtained from Tent [25] and PTN [18] that relying on test image statistics worsens performance as a single image is insufficient for approximating the target domain distribution. InTent [2], which is proposed to alleviate this very issue, also shows only a slight improvement in ISPY1 dataset while performing worse on TCGA-BRCA dataset. SAR [20] argues that BN layers can be ineffective in real-world scenarios and instead adopts Layer and Group Normalization. Taking a similar direction, we evaluate our method using Instance Normalization (IN) layers, a method well-suited for small batches and style variations such as in medical images. We compare our method to only MEMO [28] as other methods explicitly depend on the BN layers.

As shown in Table 3, BN and IN baselines perform similarly on the test data. In contrast to MEMO which hardly improves on the baseline, our method, not bound by source BN stats, now achieves a notable 4.37% increase in DSC and, surprisingly, HD and ASD that are even lower than the supervised benchmark.

4 Discussion and Conclusion

In this work, we propose a novel method for source-free test-time adaptation of 3D medical images. Through a multi-view co-training framework, we introduce complementary volumetric information from axial, sagittal and coronal views by an entropy-informed self-training objective. Our method is label-free, adapts on a single image and needs only one epoch of training. We further demonstrate

that a simple measure of using an isotropic patch size notably improves generalization by mitigating directional biases. We also investigate the role of batch normalization, a key component in current literature that relies on large batches for adaptation, a requirement that is often difficult to meet in medical settings. Our findings suggest that batch-agnostic methods, i.e. Instance Normalization, offer greater stability in mitigating the effects of unseen domain shifts. We extensively validate our method on three publicly available multi-site multi-vendor breast-MRI datasets for tumor segmentation, demonstrating improvements over existing state-of-the-art methods.

Acknowledgments. This project has received funding from European research and innovation programme under grant agreement No 101057699 (RadioVal) and Horizon 2020 under grant agreement No 952103 (EuCanImage). The work has also been supported by FUTURE-ES (PID2021-126724OB-I00) and AIMED (PID2023-146786OB-I00) from the Ministry of Science and Innovation of Spain.

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

References

1. Clark, K., Vendt, B., Smith, K., Freymann, J., Kirby, J., Koppel, P., Moore, S., Phillips, S., Maffitt, D., Pringle, M., et al.: The cancer imaging archive (tcia): maintaining and operating a public information repository. Journal of digital imaging **26**, 1045–1057 (2013)
2. Dong, H., Konz, N., Gu, H., Mazurowski, M.A.: 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. pp. 5046–5055 (2024)
3. Dorent, R., Kujawa, A., Ivory, M., Bakas, S., Rieke, N., Joutard, S., Glocker, B., Cardoso, J., Modat, M., Batmanghelich, K., et al.: Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. Medical Image Analysis **83**, 102628 (2023)
4. Fang, C., Wang, L., Zhang, D., Xu, J., Yuan, Y., Han, J.: Incremental cross-view mutual distillation for self-supervised medical ct synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20677–20686 (2022)
5. Fleuret, F., et al.: Test time adaptation through perturbation robustness. In: NeurIPS 2021 Workshop on Distribution Shifts: Connecting Methods and Applications (2021)
6. Garrucho, L., Kushibar, K., Reidel, C.A., Joshi, S., Osuala, R., Tsirikoglou, A., Bobowicz, M., del Riego, J., Catanese, A., Gwoździewicz, K., et al.: A large-scale multicenter breast cancer DCE-MRI benchmark dataset with expert segmentations. arXiv preprint arXiv:2406.13844 (2024)
7. Hu, M., Song, T., Gu, Y., Luo, X., Chen, J., Chen, Y., Zhang, Y., Zhang, S.: Fully test-time adaptation for image segmentation. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24. pp. 251–260. Springer (2021)

8. 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)
9. Iwasawa, Y., Matsuo, Y.: Test-time classifier adjustment module for model-agnostic domain generalization. Advances in Neural Information Processing Systems **34**, 2427–2440 (2021)
10. Joshi, S., Osuala, R., Garrucho, L., Tsirikoglou, A., del Riego, J., Gwoździewicz, K., Kushibar, K., Diaz, O., Lekadir, K.: Leveraging epistemic uncertainty to improve tumour segmentation in breast mri: an exploratory analysis. In: Medical Imaging 2024: Image Processing. vol. 12926, pp. 292–300. SPIE (2024)
11. Joshi, S., Osuala, R., Martín-Isla, C., Campello, V.M., Sendra-Balcells, C., Lekadir, K., Escalera, S.: nn-unet training on cyclegan-translated images for cross-modal domain adaptation in biomedical imaging. In: International MICCAI Brainlesion Workshop. pp. 540–551. Springer (2021)
12. Kushibar, K., Valverde, S., Gonzalez-Villa, S., Bernal, J., Cabezas, M., Oliver, A., Llado, X.: Supervised domain adaptation for automatic sub-cortical brain structure segmentation with minimal user interaction. Scientific reports **9**(1), 6742 (2019)
13. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. International Journal of Computer Vision **133**(1), 31–64 (2025)
14. 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)
15. Lingle, W., et al.: The cancer genome atlas breast invasive carcinoma collection (tcga-brca)(version 3)[data set]. cancer imag. arch.(2016)
16. Liu, Y., Kothari, P., Van Delft, B., Bellot-Gurlet, B., Mordan, T., Alahi, A.: Ttt++: When does self-supervised test-time training fail or thrive? Advances in Neural Information Processing Systems **34**, 21808–21820 (2021)
17. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., et al.: Metrics reloaded: recommendations for image analysis validation. Nature methods pp. 1–18 (2024)
18. Nado, Z., Padhy, S., Sculley, D., D'Amour, A., Lakshminarayanan, B., Snoek, J.: Evaluating prediction-time batch normalization for robustness under covariate shift. arXiv preprint arXiv:2006.10963 (2020)
19. Newitt, D., Hylton, N., et al.: Multi-center breast dce-mri data and segmentations from patients in the i-spy 1/acrin 6657 trials. Cancer Imaging Arch **10**(7) (2016)
20. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. arXiv preprint arXiv:2302.12400 (2023)
21. Osowiechi, D., Hakim, G.A.V., Noori, M., Cheraghalikhani, M., Ben Ayed, I., Desrosiers, C.: Tttflow: Unsupervised test-time training with normalizing flow. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2126–2134 (2023)
22. Saha, A., Harowicz, M.R., Grimm, L.J., Weng, J., Cain, E.H., Kim, C.E., Ghate, S.V., Walsh, R., Mazurowski, M.A.: Dynamic contrast-enhanced magnetic resonance images of breast cancer patients with tumor locations [data set]. The Cancer Imaging Archive (2021). https://doi.org/https://doi.org/10.7937/TCIA.e3sv-re93
23. Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., Bethge, M.: Improving robustness against common corruptions by covariate shift adaptation. Advances in neural information processing systems **33**, 11539–11551 (2020)

24. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., Hardt, M.: Test-time training with self-supervision for generalization under distribution shifts. In: International conference on machine learning. pp. 9229–9248. PMLR (2020)
25. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
26. Wang, Y., Li, Z., Mei, J., Wei, Z., Liu, L., Wang, C., Sang, S., Yuille, A.L., Xie, C., Zhou, Y.: Swinmm: masked multi-view with swin transformers for 3d medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 486–496. Springer (2023)
27. Zhang, L., Mohamed, A.A., Chai, R., Guo, Y., Zheng, B., Wu, S.: Automated deep learning method for whole-breast segmentation in diffusion-weighted breast mri. Journal of Magnetic Resonance Imaging **51**(2), 635–643 (2020)
28. Zhang, M., Levine, S., Finn, C.: Memo: Test time robustness via adaptation and augmentation. Advances in neural information processing systems **35**, 38629–38642 (2022)