This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Medical Image Segmentation with InTEnt: Integrated Entropy Weighting for Single Image Test-Time Adaptation

Haoyu Dong Nicholas Konz Hanxue Gu
Maciej A. Mazurowski
Duke University
2127 Campus Drive, Durham, NC 27708
{haoyu.dong151, nicholas.konz, hanxue.gu, maciej.mazurowski}@duke.edu

Abstract

Test-time adaptation (TTA) refers to adapting a trained model to a new domain during testing. Existing TTA techniques rely on having multiple test images from the same domain, yet this may be impractical in real-world applications such as medical imaging, where data acquisition is expensive and imaging conditions vary frequently. Here, we approach such a task, of adapting a medical image segmentation model with only a single unlabeled test image. Most TTA approaches, which directly minimize the entropy of predictions, fail to improve performance significantly in this setting, in which we also observe the choice of batch normalization (BN) layer statistics to be a highly important yet unstable factor due to only having a single test domain example. To overcome this, we propose to instead integrate over predictions made with various estimates of target domain statistics between the training and test statistics, weighted based on their entropy statistics. Our method, validated on 24 source/target domain splits across 3 medical image datasets surpasses the leading method by 2.9% Dice similarity score on average.

1. Introduction

Deep neural networks have demonstrated impressive performance when source (training) and target (test) images are drawn from the same distribution. Unfortunately, this assumption often fails in real-world applications, where target data may be corrupted naturally (*e.g.*, with weather changes or sensor degradation [17]) or acquired differently (*e.g.*, MRIs taken with different scanners or under different protocols [18]). Trained models can be sensitive to these shifts, resulting in performance degradation, known as the *domain shift* problem [6, 29].

Early work [32] solves this problem by learning auxiliary tasks during training, which can be sub-optimal since the training pipeline is altered. Fully Test-time Adaptation (TTA) methods instead propose to adapt models solely using target domain data and have achieved significant improvements in robustness to domain shift [23, 24, 35]. Typically, model parameters are updated to minimize the entropy of model predictions on test images, as a proxy for minimizing the cross-entropy given that the target labels are unknown [35]. However, recent works have observed that these improvements have occurred only within certain conditions, namely that target images (1) are available in a relatively large quantity and (2) can arrive continuously, *i.e.*, in an online fashion [4, 26, 38]. The first condition further implicitly assumes that all images in the same batch are from the same domain and have balanced class information, and the second condition unavoidably favors target data that arrived later. Both conditions bring restrictions to real-world usage.

In this paper, we consider an extreme case of TTA, where a model only has access to **a single target image** during adaptation. This setting is called **S**ingle **I**mage **T**est-**T**ime **A**daptation (SITTA or SITA [15]), which we summarize and compare to related settings in Table 2. SITTA avoids

Method	SC	Che.	Ret.	Avg. ↑
UNet	57.8	82.9	53.3	64.0
MEMO	59.1	85.3	54.1	65.5
TEnt	57.7	93.0	58.9	68.7
SAR	57.5	93.0	58.9	68.4
FSeg	57.8	93.1	**58.9**	68.7
SITA	61.3	90.5	56.7	68.7
InTEnt	**64.5**	**94.1**	58.6	**71.6**

Table 1. Average Dice similarity coefficient (repeated 10 times, average of all source/target domain splits) of different TTA methods on three datasets. The leading performance is highlighted. **Our method outperforms the SOTA on two datasets and surpasses the leading method by** 2.9% **on average.**

Setting	Source data	Target data	Train Objective	Test Objective	Online
Fine-tuning	-	X^t, Y^t	$L(X^t, Y^t)$	-	✓
Test-time training	X^s, Y^s	X^t	$L(X^t, Y^t)+L(X^s, X^t)$	-	✓
Test-time adaptation (TTA)	-	X^t	-	$L(X^t)$	✓
Continual TTA	-	x_i^t	-	$L(x_i^t)$	✓
Single Image TTA (ours)	-	x_i^t	-	$L(x_i^t)$	✗

Table 2. **Comparison of different TTA settings and the data available in each.** X^t and X^s refer to a batch of images from the target and source domains, respectively. x_i^t refers to a single image from the target domain. "Online" refers to whether information from prior test images is accessible for a new prediction.

the above-mentioned assumption naturally and is especially relevant to medical image analysis, where obtaining additional images from the same domain can be expensive, time-consuming, or even infeasible due to medical image privacy concerns and scanner setting inhomogeneities [6, 21]. We focus on segmentation because it is a common yet challenging task in medical image analysis.

Our extensive experiments reveal that existing TTA methods, which typically optimize learnable batch normalization layer parameters (scale and shift) for the target domain, fail to alter network performance significantly. Instead, we find that batch norm. layer *statistics* (mean and standard deviation), hereafter referred to simply as "statistics", play a crucial role in model adaptation, aligning with recent observations [30]. However, the best choice of statistics, is highly variable between different domain shifts, due to the instability of relying on only a single target domain image. To address these challenges, we propose a novel method for creating an ensemble of several possible adapted models constructed using different estimates of the target domain statistics. Rather than simply selecting the model with the lowest prediction entropy, we integrate all models' predictions. We explore and combine various integration strategies, including simple averaging, weighted averaging based on entropy or entropy sharpness (a concept recently discovered by [26] to be informative during TTA), and others. This ensembling approach is robust when relying on only a single test domain image because it does not require iterative optimization of model parameters. We also incorporate a novel approach of equally balancing the entropy contributions of predicted foreground and background pixels that is specifically designed for segmentation, rather than treating all pixel predictions equally. Moreover, our method can be integrated with other TTA methods as there is no limitation on the models to be integrated from. Our method is named **InTEnt**: **In**tegrated **T**est-time **Ent**ropy Weighting for Single Image Adaptation, summarized in Figure 1. InTEnt achieves superior performance over existing methods in a variety of medical image domain shift settings, as shown in Table 1.

Contributions. Our main contributions are the following:

1. We demonstrate the importance of batch normalization layer statistic selection for adapting models to a single test image, and use this to generate an ensemble of possible adapted models.
2. To address the variability of the optimal batch norm. statistic choice for different domain shift settings, we propose a simple yet effective strategy of integrating the predictions of the different adapted models, weighted by their prediction entropy.
3. Our method achieves an average performance of 71.6% Dice similarity coefficient (DSC) for 24 different domain shift settings across 3 datasets, while other approaches give at most 68.7% DSC.

2. Related Work

2.1. Single Image Test-Time Adaptation

Test-time adaptation (TTA) aims at fine-tuning model parameters during test time, using only test data [20]. In this work, we broaden this application by considering the challenging case of only having a single test image, focusing on medical image segmentation where such a constraint is very realistic.

Certain previous works approach single image TTA by learning extra information during training [10, 22]. For example, [14] introduced a denoising autoencoder to correct test predictions; [33] proposed to pre-train a domain encoder that can simulate target image domain information; [5] learned a diffusion model that projects target domain images back to the source domain. Although effective, these methods utilize auxiliary networks during training, removing the possibility of adapting arbitrary pre-trained models to target downstream tasks. Another direction is to augment the test image before adaptation, which increases the robustness and correctness in estimating from a single test image [5, 15, 37]. However, these methods are sensitive to the choice of augmentation function, and we find that they lead to sub-optimal performance for segmentation.

2.2. Test-time Adaptation with Prediction Entropy

TEnt [35] first found that minimizing prediction entropy during test time can improve network performance. We

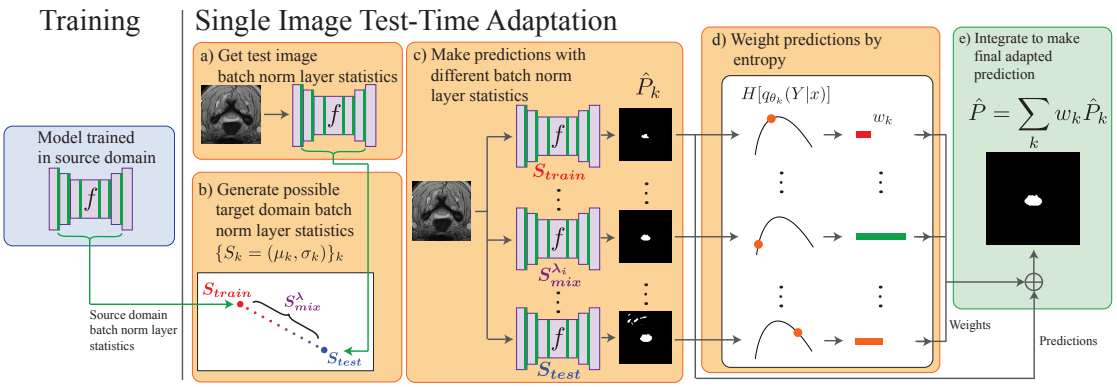

Figure 1. **Summary of our method for single-image test time adaptation of a segmentation model (Algorithm 1).** Note that segmentation probability map predictions $\hat{P}_k$ and $\hat{P}$ are rounded to binary masks for visualization.

claim that the effectiveness of this approach relies on two aspects: (1) the model is well-calibrated and (2) the estimation of the test domain prediction entropy is accurate. A model is well-calibrated if its predicted probabilities are representative of the true correctness likelihood [7], *i.e.*, a predicted probability closer to 0 or 1 (lower entropy) should be more likely to be correct. Thus, entropy minimization results directly in more accurate predictions. However, it is infeasible to precisely calibrate the model without access to the training data.

The second aspect is that prediction entropy can only be reliably estimated when test images arrive in a batch and are from the same domain, allowing for a faithful estimation of the target domain statistics via entropy minimization [4, 26, 38]. To solve this problem in the single-image TTA setting, we propose to instead integrate predictions made with different target domain statistic estimates, as only using a single estimate found via entropy minimization leads to inconsistent results due to the difficulty of accurate estimation with just one image.

3. Method

3.1. Single Image Test-Time Adaptation for Segmentation

In this section, we will review test-time adaptation (TTA) and its connection to prediction entropy. For single image TTA (SITTA), we have a pre-trained model f with parameters θ, and a *single* test image $x \in \mathbb{R}^{N_C \times H \times W}$ with N_C channels, which has an unknown corresponding label Y sampled from some test domain conditional probability distribution $p_{\text{test}}(Y|x)$ [36]. For binary segmentation, an image label $Y \in \{0,1\}^{H \times W}$ is a segmentation mask of $H \times W$ pixels, and f outputs a predicted probability map $\hat{P} \in [0,1]^{H \times W}$. The goal of TTA is to determine the optimal model parameters θ that maximize the likelihood $p(y^{i,j}|x,\theta)$ of the model prediction for each pixel (i,j), given the unknown pixel label sampled from $p_{\text{test}}(y^{i,j}|x)$.

Let $q_\theta(y^{i,j}|x)$ denote the model's prediction distribution for a given pixel, *i.e.*, f's predicted probability for that pixel to contain the object of interest. Maximizing the pixel likelihood is equivalent to minimizing the cross-entropy between the predicted and true distributions, $H(q_\theta, p_{\text{test}}) = -\mathbb{E}_{y^{i,j} \sim q_\theta(y^{i,j}|x)} \ln p_{\text{test}}(y^{i,j}|x)$ [8]. Writing $q_\theta(y^{i,j}|x)$ and $p_{\text{test}}(y^{i,j}|x)$ as q_θ and p_{test} for brevity, the cross-entropy can be decomposed into

$$\begin{aligned} H(q_\theta, p_{\text{test}}) &= -\mathbb{E}_{y^{i,j} \sim q_\theta} \ln p_{\text{test}} \\ &= -\mathbb{E}_{y^{i,j} \sim q_\theta} [\ln p_{\text{test}} - \ln q_\theta + \ln q_\theta] \\ &= -\mathbb{E}_{y^{i,j} \sim q_\theta} \ln q_\theta + \mathbb{E}_{y^{i,j} \sim q_\theta} \ln \frac{q_\theta}{p_{\text{test}}} \\ &= H[q_\theta] + D_{\text{KL}}[q_\theta || p_{\text{test}}], \end{aligned} \tag{1}$$

where $H[q_\theta] = -\mathbb{E}_{y^{i,j} \sim q_\theta} \ln q_\theta$ is the entropy of the predictive distribution, and D_{KL} is the Kullback-Leibler Divergence between q_θ and p_{test}.

Without access to the target labels, it is impossible to evaluate the $p_{\text{test}} = p_{\text{test}}(y^{i,j}|x)$ term in D_{KL}, so that minimizing the prediction entropy $H[q_\theta]$ would be the only feasible option. If we assume that the predictions for different pixels are independent [34], the predictive and true distributions for the entire image mask Y can be written as the product of individual pixel probabilities, as $q_\theta(Y|x) = \prod_{i=1}^{H} \prod_{j=1}^{W} q_\theta(y^{i,j}|x)$ and $p_{\text{test}}(Y|x) = \prod_{i=1}^{H} \prod_{j=1}^{W} p_{\text{test}}(y^{i,j}|x)$, and similar for the likelihood $p(Y|x,\theta) = \prod_{i=1}^{H} \prod_{j=1}^{W} p(y^{i,j}|x,\theta)$. Then, minimizing the mask prediction entropy can be accomplished by minimizing the sum (or equivalently, the average) of pixel prediction entropies,

$$\begin{aligned} H[q_\theta(Y|x)] &= -\mathbb{E}_{Y \sim q_\theta(Y|x)} \ln q_\theta(Y|x) \\ &= -\mathbb{E}_{Y \sim q_\theta(Y|x)} \sum_{i=1}^{H} \sum_{j=1}^{W} \ln q_\theta(y^{i,j}|x). \end{aligned} \tag{2}$$

Foreground-Background-Balanced Entropy Weighting Despite pixel predictions being independent, they can contribute differently to the final quality of the predicted mask. For example, given one mask prediction with moderately low entropy across all pixels, and another with zero entropy for background predictions and high entropy for foreground predictions, the former would result in more faithful predictions yet a lower overall entropy if averaged across all pixels. Thus, we propose a new strategy to balance the importance of foreground and background predictions. Specifically, we define the predicted foreground entropy as

$$H_{FG}[q_\theta(Y|x)] = -\sum_{i,j \in S} q_\theta(y^{i,j}|x) \ln q_\theta(y^{i,j}|x), \quad \text{where } S = \{(i,j) \mid q_\theta(y^{i,j}|x) \geq 0.5\}, \tag{3}$$

with the background entropy $H_{BG}[q_\theta(Y|x)]$ defined similarly with the complement of S. We then use the average of H_{FG} and H_{BG} as the final weight for a given model prediction.

3.2. Adapting Models via Batch Normalization Layers

Formally, a Batch Normalization (BN) layer [12] can be expressed as

$$BN(h) = \gamma(h - \mu_{train})/\sigma_{train} + \beta, \tag{4}$$

where h is the input feature map, $\{\gamma, \beta\}$ are scale and shift parameters learned during training, and $S_{train} := \{\mu_{train}, \sigma_{train}\}$ are the tracked mean and variance of the source domain. The mean and variance, hereafter referred to as "statistics", are computed per channel, *i.e.*, over the batch and spatial dimension. When domain shift occurs, the test domain statistics $S_{test} := \{\mu_{test}, \sigma_{test}\}$ can differ from the tracked ones, leading to sub-optimal performance. While other methods optimize γ, β at test time with gradient descent to minimize prediction entropy [26, 35] or customized objectives [10], we find that in the SITTA setting, optimization leads to minor changes in the final prediction. Therefore, we propose to instead modify the *statistics*, with the following scheme.

We can freely interpolate between the training and test statistics S_{train} and S_{test} with $\lambda \in (0, 1)$ to obtain mixed statistics

$$S^\lambda_{mix} := \lambda \times S_{train} + (1 - \lambda) \times S_{test}. \tag{5}$$

Instead of selecting a single λ, we sample evenly from $(0, 1)$ with a step size hyperparameter C to create a range of mixed statistics to consider. By default, we use $C = 0.2$, which creates $\lambda \in \{0.2, 0.4, 0.6, 0.8\}$. We use each of the training, test, and different mixed statistics to define an ensemble of adapted models. Figure 1c) visualizes how varying these statistics will affect model predictions, and we include a more detailed visualization in the Experiment section.

3.3. Integrating over Adapted Models

With our proposed strategy of adapting f to the test domain via the modification of batch norm. statistics, we can obtain multiple predictions for a test image x by using each of the statistics

$$S_k \in \{S_{train}, S^{\lambda=C,\ldots,1-C}_{mix}, S_{test}\} \tag{6}$$

to define a set $\mathcal{F}$ of models. A simple solution would be to use the model f_k out of $\mathcal{F}$ that results in the prediction with minimum entropy, but we found this to be less stable and robust, due to relying on a single image from the target domain for entropy estimation.

Instead, we take a Bayesian approach [1, 9] of integrating over (the predictions of) all adapted models. First, let $\hat{P}_k$ be the prediction (segmentation probability map) of model $f_k \in \mathcal{F}$ with adapted parameters θ_k (**note:** here we write θ_k to include the adapted batch norm. layer statistics, although these aren't learnable). We could weight each $\hat{P}_k$ by the posterior probability of the model parameters θ_k, but this is unknown. Instead, we weight each prediction by the likelihood $p(Y|x, \theta_k)$, as it is proportional to the posterior given that the evidence and prior distributions are unknown/intractable. This model weighting results in an optimal prediction of

$$\hat{P} = \sum_{f_k \in \mathcal{F}} \hat{P}_k p(Y|x, \theta_k). \tag{7}$$

As we cannot fully evaluate the likelihood of a model without the ground truth label for x, we can approximate it using the prediction entropy as in Eq. (1), with

$$p(Y|x, \theta_k) = e^{-H(q_{\theta_k}, p_{\text{test}})} = e^{-H(q_{\theta_k})} e^{-D_{\text{KL}}(q_{\theta_k} || p_{\text{test}})} \underset{\sim}{\propto} e^{-H(q_{\theta_k})}, \tag{8}$$

where we have written q_{θ_k} and p_{test} short-hand for the predictive and true segmentation distributions $q_{\theta_k}(Y|x)$ and $p_{\text{test}}(Y|x)$, respectively. In other words, models that have lower balanced segmentation prediction entropy: $w_k := -H_{FG}[q_{\theta_k}(Y|x)] + H_{BG}[q_{\theta_k}(Y|x)]$ (Eq. (3)) are weighted higher. Lastly, we normalize w_k by $w'_k = w_k / [\max(\{w_k\}_{\forall k}) - \min(\{w_k\}_{\forall k})]$ to assign higher weights to predictions with lower entropy. This is the integration strategy that we use for our final algorithm (performance shown in Table 1). We also compare a wide range of entropy-based prediction weighting strategies in Sec. 4.5. We will next introduce entropy sharpness, a recent concept that is also potentially usable as a weighting strategy.

3.4. Minimizing Prediction Entropy Sharpness

As recent TTA literature [26] found that prediction entropy $H[q_{\theta_k}(Y|x)]$ can be unstable when estimated from a

Algorithm 1 **In**tegrated **T**est-time **Ent**ropy Weighting for Single Image Adaptation for Segmentation

Input: Test image $x \in \mathbb{R}^{N_C \times H \times W}$, source domain-trained segmentation model $f : \mathbb{R}^{N_C \times H \times W} \to [0,1]^{H \times W}$.

Create ensemble of adapted models by modifying batch norm statistics:
$\mathcal{F} = \{f_k : f \text{ with batch norm. stats } S_k \text{ (Eq. (6))}\}$
Predict segmentation probability maps: $\hat{P}_k = f_k(x)$
Weight each model by its prediction entropy (Eq. (3)):
$w_k = -H_{FG}[q_{\theta_k}(Y|x)] - H_{BG}[q_{\theta_k}(Y|x)]$
Normalize weights:
$w'_k = w_k / [\max(\{w_k\}_{\forall k}) - \min(\{w_k\}_{\forall k})]$
$\{w_k\}_{\forall k} = \text{softmax}(\{w'_k\}_{\forall k})$
Obtain integrated segmentation prediction:
$\hat{P} = \sum_{f_k \in \mathcal{F}} w_k \hat{P}_k$

small number of test images, we also evaluate an alternative strategy to weight models according to prediction entropy *sharpness* with respect to model parameters.

The *sharpness* of the prediction entropy of a model is defined as the entropy's highest possible sensitivity with respect to a small perturbation ϵ to the model parameters. Formally, finding model parameters that give minimum entropy sharpness is a joint optimization problem

$$\min_{\theta} \max_{||\epsilon||_2 \leq \rho} H[q_{\theta_k+\epsilon}(Y|x)] \tag{9}$$

[26], for some small constant ρ (0.1 by default), where $H[q_{\theta_k+\epsilon}(Y|x)]$ is the prediction entropy of the model evaluated with parameters $\theta_k + \epsilon$ on the test image x (Eq. (2)). If a first-order Taylor approximation is used for the inner optimization, a closed-form solution

$$\hat{\epsilon}(\theta) = \frac{\rho \, \text{sign}\left(\nabla_\theta H[q_\theta(Y|x)]\right) |\nabla_\theta H[q_\theta(Y|x)]|}{\|\nabla_\theta H[q_\theta(Y|x)]\|_2} \tag{10}$$

is possible [2]. We can then easily estimate the prediction entropy sharpness of some adapted model $f_k \in \mathcal{F}$ as

$$\text{sharp}(f_k; x) = H[q_{\theta_k+\hat{\epsilon}(\theta_k)}(Y|x)] - H[q_{\theta_k}(Y|x)]. \tag{11}$$

Returning to our model-averaging scheme of the previous section, we can give high weight w_k to the prediction of an adapted model f_k if it has low entropy sharpness, to obtain a final integrated prediction $\hat{P} = \sum_{f_k \in \mathcal{F}} w_k \hat{P}_k$. For our case of single image TTA for segmentation, the sharpness (Eq. (11)) simplifies to

$$\text{sharp}(f_k; x) = \sum_{i=1}^{H} \sum_{j=1}^{W} \hat{P}_k^{i,j} \ln \hat{P}_k^{i,j} - \hat{P}_{\theta_k+\hat{\epsilon}(\theta_k)}^{i,j} \ln \hat{P}_{\theta_k+\hat{\epsilon}(\theta_k)}^{i,j}, \tag{12}$$

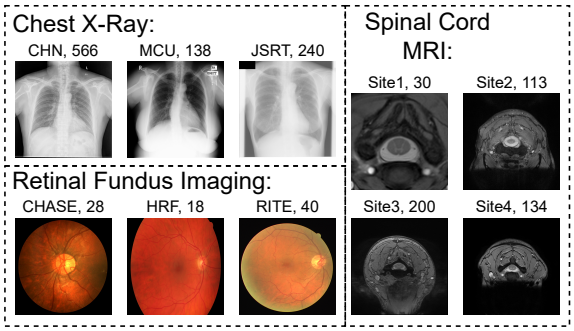

Figure 2. **Overview of the datasets used in this paper.** Above each example image, we list its domain and the total number of images from this domain.

where $\hat{P}_{\theta_k+\hat{\epsilon}(\theta_k)}^{i,j}$ is the (i,j) entry of the predicted segmentation probability map of f_k for x given parameters $\theta_k + \hat{\epsilon}(\theta_k)$. We could then define model weights as $w_k := -\text{sharp}(f_k; x)$, which we will later compare to our strategy.

3.5. Summary

We summarize our method for single image test-time adaptation for segmentation in Figure 1 and Algorithm 1. Beginning with some segmentation model trained on source domain data and a single test image of an unknown domain that we wish to adapt the model to, we first use our batch norm. statistic modification scheme (Eq. (6)) to create an ensemble of possible adapted models. By default, we weigh each model according to its segmentation prediction entropy before integrating over all models to obtain a final prediction, but we will also evaluate additional weighting strategies, including via entropy sharpness. We name our method **InTEnt**, or **In**tegrated **T**est-time **Ent**ropy Weighting for Single Image Adaptation.

Our method takes about 0.06 seconds to compute for a single test image, consisting of 6 times forward (Eq. (6) with $C = 0.2$) where a single forward takes 0.01 second on an NVIDIA RTX A6000. Note that computation cost is not our primary concern given that we only have one image to perform inference on in the single-image TTA setting.

4. Experiments and Results

4.1. Setup

Datasets. We evaluate our proposed method on three medical image segmentation tasks with publicly available multi-institution/domain datasets. Grouped by {modality}/{object of interest}, these are: (1) **Spinal Cord (SC) MRI slices**/gray matter: Spinal Cord Gray Matter Segmentation Challenge Dataset [28]; (2) **Retinal (RET.) Fundus Imaging**/blood vessel: CHASE [3], RITE [11], and HRF [27]; (3) **Chest (CHE.) X-ray**/lung: CHN, MCU [13] and JSRT [31]. Figure 2 summarizes the domains in

λ	Method	Spinal Cord				Chest			Retinal			Avg.
		Site1	Site2	Site3	Site4	CHN	MCU	JSRT	CHASE	HRF	RITE	
1.0	UNet	49.0	72.9	34.0	75.3	90.7	**80.4**	77.5	**46.4**	57.7	55.9	**64.0**
	+TEnt	48.2	72.7	**34.3**	75.9	90.7	80.1	76.8	45.9	57.1	55.3	63.7
	+SAR	**49.8**	**73.5**	32.7	74.8	90.3	80.3	**79.3**	46.8	**58.2**	**56.3**	64.2
	+FSeg	48.2	72.8	34.1	**75.9**	**90.7**	80.1	76.8	45.8	57.1	55.3	63.7
	+MEMO	47.7	72.5	33.8	75.4	90.0	80.3	75.8	45.8	57.0	55.3	63.4
0.5	UNet	62.2	70.4	43.8	77.3	95.7	91.5	93.2	54.8	59.5	61.5	71.0
	+TEnt	61.9	70.3	**44.4**	78.4	95.7	91.6	93.4	54.6	59.1	61.6	71.1
	+SAR	62.3	71.1	41.5	76.3	95.5	90.9	92.9	54.9	59.6	61.5	70.7
	+FSeg	61.9	70.3	44.3	78.4	**95.7**	**91.7**	**93.4**	54.5	59.1	61.5	71.1
	+MEMO	61.6	70.0	44.0	78.4	95.6	91.3	93.4	54.5	59.0	61.5	70.9
	+SITA	**63.4**	**71.4**	41.6	78.1	95.8	91.3	93.2	**55.4**	**60.2**	**61.6**	**71.2**
0.0	UNet	56.0	65.8	47.3	63.2	**95.8**	93.4	89.8	**57.5**	**58.3**	61.0	68.8
	+TEnt	53.2	65.7	47.5	64.5	95.3	93.5	90.3	57.4	58.2	61.3	68.7
	+SAR	54.9	66.2	47.0	61.6	95.1	93.1	89.3	57.5	58.3	60.7	68.4
	+FSeg	53.3	65.7	47.5	64.5	95.3	**93.5**	90.3	57.5	58.2	**61.3**	68.7
	+MEMO	53.8	65.4	**47.8**	64.5	95.2	93.3	90.4	57.3	58.2	**61.3**	68.7
	+SITA	**56.0**	**67.1**	44.5	**67.6**	95.4	93.2	**90.4**	57.8	58.2	60.8	**69.1**

Table 3. **The performance of UNet with various TTA methods given different batch norm. layer statistic choices defined by λ**, given as Dice segmentation similarity score averaged over 10 repeated experiments. Models are tested on all target domains from the same dataset. The highest score in each choice is highlighted.

each modality and the number of images from each domain. To evaluate various domain adaptation methods in a given modality, we train a segmentation model on images from a single domain and adapt and evaluate the model for one of the other domains. During evaluation, model parameters are reset to their source domain setting for each test image, as we consider the offline setting.

Implementation Details. We center crop the input images to 144×144 for Spinal Cord [19], resize input images to 256×256 for Fundus, and 128×128 for Chest, following prior works. All images are further normalized to $[0, 1]$. We use an improved version of the UNet architecture [25] for the segmentation model, which includes additional attention layers and a middle block between the encoder and decoder. The model is trained with equally weighted binary cross entropy (BCE) and Dice coefficient losses, optimized using Adam [16] with a learning rate of 10^{-4} and momentum of 0.9. Batch size is set to 10. During training, the batch norm. layer statistics are updated via an exponentially moving average with a step size of 0.1. Segmentation predictions are evaluated with the Dice similarity score with respect to the target mask. 80% of the images are randomly selected for training and the rest is used for validation. We train for 200 epochs, with early-stopping criteria for when the (source domain) validation score is not improved after 20 epochs. All experiments are repeated 10 times with the same train/validation split. The average performance is reported. Code and trained models will be made publicly available upon acceptance.

Competing methods. We compare our method to several recent TTA approaches that can be extended to the SITTA setting. TEnt [35], SAR [26], and FSeg [10] propose to minimize entropy, entropy sharpness, and Regional Nuclear-Norm loss, respectively, by updating normalization layer parameters, which all use test image batch norm. statistics ($\lambda = 0$ in Eq. (5)). SITA [15] is another TTA strategy that takes the batch norm. statistics of different augmentations of the test image, and uses the average of all statistics to make the final prediction, using $\lambda = 0.8$. We also evaluate SITA with their additional proposed strategy "OP" for finding the optimal statistics interpolated between the train and test domains using majority voting on minimum entropy. Finally, MEMO [37] combines both strategies by computing the average prediction entropy given a set of transformed versions of the test image, using $\lambda = 15/16$. To adapt these methods to the (offline) SITTA setting, we reduce the test batch size to 1 and reset the model parameters after each adaptation. All other hyperparameters follow the settings of the respective original paper.

4.2. Performance Comparison to Existing Methods

Table 1 shows the average performance of existing TTA methods and our method across different datasets, averaged over all source/target domain splits. Our method achieves the leading performance on both the spinal cord and chest dataset, surpassing the runner-up methods by 3.2% and 1.0% Dice similarity score (DSC), respectively. The method is also on par with SOTA on the fundus dataset

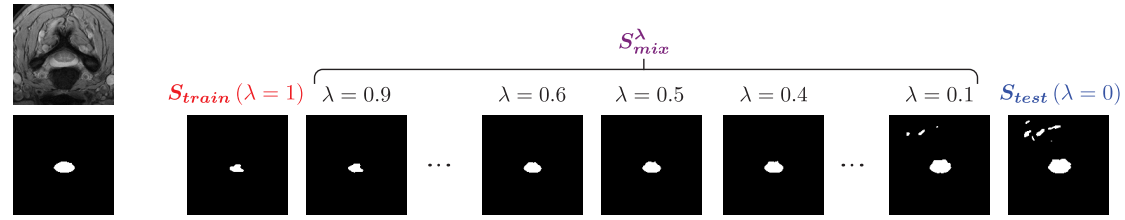

Figure 3. The effect on model prediction when using different domain batch norm. layer statistics.

Method	BN. stat. Strategy	Spinal Cord				Chest			Retinal			Avg.
		Site1	Site2	Site3	Site4	CHN	MCU	JSRT	CHASE	HRF	RITE	
UNet	$\lambda = 1.0$	49.0	72.9	34.0	75.3	90.7	80.4	77.5	46.4	57.7	55.9	64.0
	$\lambda = 0.8$	57.8	72.4	37.7	77.3	94.1	86.0	91.1	50.9	59.0	59.5	68.6
	$\lambda = 0.6$	61.6	71.1	41.6	77.8	95.4	90.1	93.3	53.8	59.4	61.2	70.5
	$\lambda = 0.4$	62.0	69.7	45.8	76.4	95.8	92.4	92.8	55.7	59.5	61.7	71.2
	$\lambda = 0.2$	59.7	68.0	49.6	72.3	95.7	93.5	91.6	56.9	59.1	61.6	70.8
	$\lambda = 0.0$	56.0	65.8	47.3	63.2	95.8	93.4	89.8	57.5	58.3	61.0	68.8
InTEnt	Average	61.4	70.5	46.1	78.7	95.6	91.7	93.3	54.7	59.9	61.8	71.4
	Entropy	61.4	70.5	46.6	78.8	95.6	91.9	93.4	54.7	59.9	61.7	71.5
	Ent.-Min	52.2	69.4	47.6	76.5	95.0	**93.5**	92.2	53.2	58.1	56.3	69.4
	Ent.-Topk	59.6	70.1	**49.5**	78.6	95.7	93.4	93.5	54.4	59.0	59.5	71.3
	Ent.-Norm	60.7	70.2	49.2	**79.2**	95.8	92.9	93.7	54.7	59.7	61.0	71.7
	Ent.-Baln	**62.2**	**71.4**	45.2	79.2	**95.9**	92.9	**94.1**	53.8	**59.9**	**62.2**	71.7
	Sharpness	61.1	70.5	46.8	77.6	95.6	92.3	93.3	**55.7**	59.7	61.7	71.4

Table 4. **Top:** Baseline UNet performance with different batch norm. statistics S^{λ}_{mix} (Eq. (5)), averaged over all domain shifts. **Bottom:** Integrated performance of the top block using different integration strategies.

(0.3% lower DSC).

4.3. Importance of the Choice of Batch Norm. Statistics

Interestingly, from the previous experiments we observe that the performance of TEnt [35], SAR [26], and FSeg [10] are quite similar despite their different optimization objectives. Further investigation on these methods (TEnt, SAR, and FSeg) reveals that all methods utilize the test image batch norm. (BN) statistics, *i.e.*, $\lambda = 0$ (Eq. (5)). To further study the relationship between BN statistics and adapted network performance, we evaluate all competing methods and the baseline model with different selections of statistics, *i.e.*, choices of λ. Note that SITA is not applicable when $\lambda = 1$ since it only alters the test statistics. The results are shown in Table 3, with the details of individual source/target domain performances in Appendix Section 2.1. Note that we report each method's performance with its default λ in Table 1, *e.g.* the performance of SAR, FSeg, and SITA in Table 1 aligns with that in Table 3 when $\lambda = 0$.

By comparing all methods for a fixed λ/statistic, we observe that the gains by each method are small in the single image segmentation TTA setting: usually $< 1\%$ change in Dice similarity score (DSC), with the best being $+4.8\%$ DSC (Chest X-Ray, JSRT→MCU, $\lambda = 1$), and the worst being -3.2% DSC (SC, Site 1→ Site 2, $\lambda = 0$). The rank of these competing methods is also not consistent, especially when examined at the individual source/target domain level, further showing their instability. Instead, the effect of the specific domain shift and statistics used is far greater. Altering the hyperparameters (*e.g.*, iteration count, learning rate) of the TTA methods could potentially amplify their effects, but this could also worsen cases where the method degrades performance. As such, we use the default hyperparameters recommended by each paper.

It could be the case that there is some optimal λ/mixture of source and test domain statistics for general single-image (segmentation) TTA, but our experiments do not support this. As shown in Table 3, the choice of optimal statistics can vary greatly even for different domain shifts within the same dataset, for both TTA-adapted models and UNet. For example, in the spinal cord dataset, when "site 1" becomes the source domain, a mix of train/test ($\lambda = 0.5$) is favored, yet when training on "site 3", models favor the test statistics

($\lambda = 0.0$). This was our motivation for instead *integrating* over predictions made with a variety of statistics.

4.4. Integrating InTEnt with Existing Methods

Since InTEnt only integrates over different models and does not involve any optimization or augmentation, it can also be combined with other TTA methods to further improve the performance (implementation strategies are described in Appendix Section 1). As a proof of concept, we conduct experiments on settings where the additional methods lead to noticeable improvements. Based on Table 3, we integrate InTEnt with SITA [15] on Spinal Cord - Site1/2 and with FSeg [10] on Site4 since the improvements of these methods are over 0.6% DSC across all choices of λ. The experiments show that InTEnt's performance increases from $62.2\% \rightarrow 64.2\%, 71.4\% \rightarrow 72.0\%, 79.2\% \rightarrow 80.1\%$ on Spinal Cord - Site 1,2,4, respectively. These results suggest that InTEnt could serve as a foundation for enhancing the performance of other TTA methods.

4.5. Ablation Study

After creating an ensemble of adapted models f_k using different statistics (Eq. (6)), our default strategy for integrating over all models' predictions $\hat{P}_k$ is to weight each prediction by its balanced entropy between foreground and background, normalize the weights, and take a weighted average of the predictions (Algorithm 1). We first present visually an example of how model predictions change to the change of λ/batch norm. statistic. In Figure 3, we observe that the un-adapted model ($\lambda = 1$) fails to segment the gray matter fully and achieves an ideal prediction when the $\lambda \simeq 0.5$, or about an even mix between train and test statistics. However, the model becomes over-confident and incorrectly segments the non-gray matter regions when the statistics come mainly from the given test image ($\lambda \simeq 0$). Next, we evaluate a range of modifications to this strategy:

1. "Average": average all predictions with equal weights.
2. "Entropy": use the exact prediction entropy to weight, as $w_k = -H[q_{\theta_k}(Y|x)]$ (Eq. (2)).
3. "Ent.-Min": the predictions with the minimum entropy are used as the final prediction.
4. "Ent.-TopK": the top-k predictions with the minimum entropy as averaged. We set $K = 2$. This is also the "optimal prior" (OP) method proposed in SITA.
5. "Ent.-Norm": a normalization is applied to ensure the maximum difference among the entropies is 1.
6. "Ent.-Baln": the entropy is computed separately for fore/background (Eq. (3)). We select this strategy when compared externally and the details are shown in Algorithm 1.
7. "Sharpness": use entropy sharpness (Eq. (12)) to weight, as $w_k = -\text{sharp}(f_k; x)$, followed by the same weight normalization as in Algorithm 1.

We show the performance of our method using these different integration strategies, alongside the baseline model performance given the different batch norm. statistic / values of λ being integrated over, in Table 4. The detailed performances of individual source/target domains are shown in Appendix Section 2.2. First, we observe that the "Ent.-Min" strategy usually leads to the worst performance among all weighting strategies, demonstrating the instability in relying on a single prediction/statistic. To be noted, this strategy still gives an average performance of 69.4% DSC, which is higher than the leading competing methods. The novel concept of entropy sharpness ("Sharpness") results in the best average performance on the Retinal dataset, yet this trend is not universal. Although the "Ent.-Baln" strategy, which our method uses, surpasses other strategies in most scenarios, the difference in performance between the weighting strategies is small. This may be caused by the variability of the relation between entropy and prediction correctness, where the root issue is trying to use a single data point to estimate prediction entropy.

We also conduct an ablation study of the effect on the numbers of models to be integrated over. In this case, we fix the integration strategy and alter the choice of C, which controls the step size during interpolation (Eq. (5)). The results are shown in Appendix Section 3 due to space limitation and we do not observe a significant change when adjusting C.

In general, InTEnt is robust against the change of the integration strategy and choice of C; a desired property since test label information is also infeasible during real-world application. We argue that the main contribution of this work is to explore the importance and necessity of batch norm. statistic selection in TTA. Integrating predictions given multiple statistics is but one solution, and we hope our work can inspire further research in this direction.

5. Conclusion

Single-image test-time adaptation is attractive for medical image segmentation due to common imaging domain inhomogeneity issues, and the expense and difficulty of acquiring new target domain images. It also benefits the general TTA setting when applied to real-world scenarios. However, relying on only a single target domain image to perform adaptation comes with its difficulties and surprises; for example, using solely the test image batch norm. statistics is not always optimal. Our proposed method, **InTEnt**, stabilizes adapted model predictions by integrating over predictions made with multiple possible estimations of the target domain statistics. We hope that our study motivates further research in segmentation SITTA for medical imaging and beyond, especially regarding the importance of the choice of normalization layer statistics.

References

[1] James O Berger, Brunero Liseo, and Robert L Wolpert. Integrated likelihood methods for eliminating nuisance parameters. *Statistical science*, pages 1–22, 1999. 4

[2] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *International Conference on Learning Representations*, 2020. 5

[3] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 59(9):2538–2548, 2012. 5

[4] Yulu Gan, Yan Bai, Yihang Lou, Xianzheng Ma, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7595–7603, 2023. 1, 3

[5] Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven adaptation to test-time corruption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11786–11796, 2023. 2

[6] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021. 1, 2

[7] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 3

[8] Yaoshiang Ho and Samuel Wookey. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling. *IEEE access*, 8:4806–4813, 2019. 3

[9] Jennifer A Hoeting, David Madigan, Adrian E Raftery, and Chris T Volinsky. Bayesian model averaging: a tutorial (with comments by m. clyde, david draper and ei george, and a rejoinder by the authors. *Statistical science*, 14(4):382–417, 1999. 4

[10] Minhao Hu, Tao Song, Yujun Gu, Xiangde Luo, Jieneng Chen, Yinan Chen, Ya Zhang, and Shaoting Zhang. Fully test-time adaptation for image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2021. 2, 4, 6, 7, 8

[11] Qiao Hu, Michael D Abràmoff, and Mona K Garvin. Automated separation of binary overlapping trees in low-contrast color retinal images. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*, pages 436–443. Springer, 2013. 5

[12] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 2015. 4

[13] Stefan Jaeger, S. Candemir, S. Antani, Yi-Xiang J. Wang, P. Lu, and G. Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4 6:475–7, 2014. 5

[14] Neerav Karani, Ertunc Erdil, Krishna Chaitanya, and Ender Konukoglu. Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68: 101907, 2021. 2

[15] Ansh Khurana, S. Paul, Piyush Rai, Soma Biswas, and Gaurav Aggarwal. Sita: Single image test-time adaptation. *ArXiv*, abs/2112.02355, 2021. 1, 2, 6, 8

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 6

[17] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard L. Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts. *ArXiv*, abs/2012.07421, 2020. 1

[18] Nicholas Konz and Maciej A Mazurowski. Reverse engineering breast mris: Predicting acquisition parameters directly from images. In *Medical Imaging with Deep Learning*, 2023. 1

[19] Haoliang Li, YuFei Wang, Renjie Wan, Shiqi Wang, Tie-Qiang Li, and Alex Kot. Domain generalization for medical imaging classification with linear-dependency regularization. *Advances in neural information processing systems*, 33:3118–3129, 2020. 6

[20] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *arXiv preprint arXiv:2303.15361*, 2023. 2

[21] Jin Liu, Yi Pan, Min Li, Ziyue Chen, Lu Tang, Chengqian Lu, and Jianxin Wang. Applications of deep learning to mri images: A survey. *Big Data Mining and Analytics*, 1(1):1–18, 2018. 2

[22] Quande Liu, Cheng Chen, Qi Dou, and Pheng-Ann Heng. Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1756–1764, 2022. 2

[23] Xiaofeng Liu, Fangxu Xing, Chao Yang, Georges El Fakhri, and Jonghye Woo. Adapting off-the-shelf source segmenter for target medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 549–559, 2021. 1

[24] Wenao Ma, Cheng Chen, Shuang Zheng, Jin Qin, Huimao Zhang, and Qi Dou. Test-time adaptation with calibration of medical image classification nets for label distribution shift. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2022. 1

[25] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 6

[26] Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In *The*

Eleventh International Conference on Learning Representations, 2023. 1, 2, 3, 4, 5, 6, 7

[27] J Odstrčilík, Jiri Jan, J Gazárek, and R Kolář. Improvement of vessel segmentation by matched filtering in colour retinal images. In *World Congress on Medical Physics and Biomedical Engineering, September 7-12, 2009, Munich, Germany: Vol. 25/11 Biomedical Engineering for Audiology, Ophthalmology, Emergency & Dental Medicine*, pages 327–330. Springer, 2009. 5

[28] F. Prados, J. Ashburner, Claudia Blaiotta, T. Brosch, J. Carballido-Gamio, M. Cardoso, Benjamin N. Conrad, Esha Datta, G. Dávid, B. Leener, S. Dupont, P. Freund, C. Wheeler-Kingshott, Francesco Grussu, R. Henry, B. Landman, E. Ljungberg, Bailey Lyttle, S. Ourselin, N. Papinutto, S. Saporito, R. Schlaeger, Seth A. Smith, P. Summers, R. Tam, M. Yiannakas, A. Zhu, and J. Cohen-Adad. Spinal cord grey matter segmentation challenge. *Neuroimage*, 152: 312 – 329, 2017. 5

[29] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence. Dataset shift in machine learning. 2009. 1

[30] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. *Advances in neural information processing systems*, 33:11539–11551, 2020. 2

[31] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi. Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *AJR. American journal of roentgenology*, 174 1:71–4, 2000. 5

[32] Yu Sun, X. Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International Conference on Machine Learning*, 2019. 1

[33] Jeya Maria Jose Valanarasu, Pengfei Guo, Vibashan VS, and Vishal M. Patel. On-the-fly test-time adaptation for medical image segmentation. In *Medical Imaging with Deep Learning*, 2023. 2

[34] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Matthieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2517–2526, 2019. 3

[35] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 1, 2, 4, 6, 7

[36] Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7201–7211, 2022. 3

[37] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. *Advances in Neural Information Processing Systems*, 35: 38629–38642, 2022. 2, 6

[38] Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML*, 2023. 1, 3