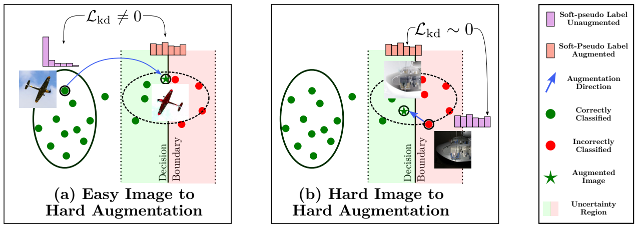
TeSLA: Test-Time Self-Learning With Automatic Adversarial Augmentation

Devavrat Tomar[1] Guillaume Vray[1] Behzad Bozorgtabar[1,2] Jean-Philippe Thiran[1,2]
[1]EPFL [2]CHUV
[1]{firstname}.{lastname}@epfl.ch

Abstract

*Most recent test-time adaptation methods focus on only classification tasks, use specialized network architectures, destroy model calibration or rely on lightweight information from the source domain. To tackle these issues, this paper proposes a novel **Te**st-time **S**elf-**L**earning method with automatic **A**dversarial augmentation dubbed **TeSLA** for adapting a pre-trained source model to the unlabeled streaming test data. In contrast to conventional self-learning methods based on cross-entropy, we introduce a new test-time loss function through an implicitly tight connection with the mutual information and online knowledge distillation. Furthermore, we propose a learnable efficient adversarial augmentation module that further enhances online knowledge distillation by simulating high entropy augmented images. Our method achieves state-of-the-art classification and segmentation results on several benchmarks and types of domain shifts, particularly on challenging measurement shifts of medical images. TeSLA also benefits from several desirable properties compared to competing methods in terms of calibration, uncertainty metrics, insensitivity to model architectures, and source training strategies, all supported by extensive ablations. Our code and models are available at https://github.com/devavratTomar/TeSLA.*



(a) Easy Image to Hard Augmentation
(b) Hard Image to Hard Augmentation

Figure 1. **Knowledge distillation with adversarial augmentations. (a)** *Easy* images with confident soft-pseudo labels and **(b)** *Hard* images with unconfident soft-pseudo labels are adversarially augmented and pushed to the uncertainty region (high entropy) near the decision boundary. The model is updated for **(a)** to match its output on the augmented views with non-augmented views of *Easy* test images using KL-Divergence $\mathcal{L}_{\text{kd}} \neq 0$, while not updated for **(b)** as $\mathcal{L}_{\text{kd}} \sim 0$ between *Hard* images and their augmented views.

1. Introduction

Deep neural networks (DNNs) perform exceptionally well when the training (*source*) and test (*target*) data follow the same distribution. However, distribution shifts are inevitable in real-world settings and propose a major challenge to the performance of deep networks after deployment. Also, access to the labeled training data may be infeasible at test time due to privacy concerns or transmission bandwidth. In such scenarios, **source-free domain adaptation** (**SFDA**) [1, 23, 25] and **test-time adaptation** (**TTA**) methods [19, 28, 39] aim to adapt the pre-trained source model to the unlabeled distributionally shifted target domain while easing access to source data. While SFDA methods have access to all full target data through multiple training epochs (offline setup), TTA methods usually process test images in an online streaming fashion and represent a more realistic domain adaptation. However, most of these methods are applied: (i) only to classification tasks, (ii) evaluated on the non-real-world domain shifts, e.g., the non-measurement shift; (iii) destroy model calibration—entropy minimizing with overconfident predictions [45] on incorrectly classified samples, and (iv) use specialized network architectures or rely on the source dataset feature statistics [28].

We address these issues by proposing a new test-time adaptation method with automatic adversarial augmentation called **TeSLA**, under which we further define realistic TTA protocols. Self-learning methods often supervise the model adaptation on the unlabeled test images using their predicted pseudo-labels. As the model can easily overfit on its own pseudo-labels, a weight-averaged teacher model (slowly updated by the student model) is employed for obtaining the pseudo-labels [41, 47]. The student model is then trained with cross-entropy ($\mathbb{CE}$) loss between the one-hot pseudo-labels and its predictions on the test images. In this paper, we instead propose to minimize flipped cross-entropy between the student model's predictions and the soft pseudo-labels (notice the reverse order) with the negative entropy of its marginalized predictions over the test images. In Sec. 3, we show that the proposed formulation is an equivalence to mutual information maximization implicitly corrected by the teacher-student knowledge distillation via pseudo-labels, yielding performance improvement on various test-time adaptation protocols and compared to the basic $\mathbb{CE}$ optimization (cf. ablation in Fig. 5-(1)).

Motivated by teacher-student knowledge distillation, another central tenet of our method is to assist the student model during adaptation in improving its performance on the hard-to-classify (high entropy) test images. For this purpose, we propose learning **automatic adversarial augmentations** (see Fig. 1) as a proxy for simulating images in the uncertainty region of the feature space. The model is then updated to ensure consistency between predictions of high entropy augmented images and soft-pseudo labels from the respective non-augmented versions. Consequently, the model is self-distilled on *Easy* test images with confident soft-pseudo labels (Fig. 1a). In contrast, the model update on *Hard* test images is discarded (Fig. 1b), resulting in better class-wise feature separation.

In summary, our contributions are: (i) we propose a novel test-time self-learning method based on *flipped* cross-entropy (f-$\mathbb{CE}$) through the tight connection with the mutual information between the model's predictions and the test images; (ii) we propose an efficient *plug-in* test-time automatic adversarial augmentation module used for online knowledge distillation from the teacher to the student network that consistently improves the performance of test-time adaptation methods, including ours; and (iii) TeSLA achieves new state-of-the-art results on several benchmarks, from common image corruption to realistic measurement shifts for classification and segmentation tasks. Furthermore, TeSLA outperforms existing TTA methods in terms of calibration and uncertainty metrics while making no assumptions about the network architecture and source domain's information, e.g., feature statistics or training strategy.

2. Related Work

In general, domain adaptation methods aim to distill knowledge from source data that are well-generalizable to target data and relax the assumption of i.i.d. between source and target datasets. To circumvent expensive and cumbersome annotation of new target data, **unsupervised domain adaptation (UDA)** emerges, and there has been a large corpus of UDAs [13, 17, 32, 51] to match the distribution of domains on both labeled source data and unlabeled target data. Nonetheless, the above-mentioned methods demand source data to achieve the domain adaptation process, which is often impractical in real-world scenarios, e.g., due to privacy restrictions. The above issue motivates research into **source-free domain adaptation (SFDA)** [1, 23, 25, 26] and **test-time training** or **adaptation (TTA)** [19, 28, 39, 45], which are more closely relevant to our problem setup.

SFDA approaches formulate domain adaptation through pseudo labeling [23, 26], target feature clustering [49], synthesizing extra training samples [24], or feature restoration [12]. SHOT [26] proposed feature clustering via information maximization while incorporating pseudo-labeling as additional supervision. BAIT [48] leverages the fixed source classifier as source anchors and uses them to achieve feature alignment between the source and target domain. CPGA [35] proposed the SFDA method based on matching feature prototypes between source and target domains. AaD-SFDA [50] proposed to optimize an objective by encouraging prediction consistency of local neighbors in feature space. Nevertheless, SFDA methods require apriori access to all target data in advance, and the current pseudo-labeling approach, e.g., SHOT [26], used offline pseudo-label refinement on a per-epoch basis. In a more realistic domain adaptation scenario, SFDA is still incompetent to perform inference and adaptation simultaneously.

TTA methods [4, 45] propose alleviating the domina shift by online (or streaming) adapting a model at test time. Still, we argue that there are ambiguities over the problem setup of TTA in the literature, particularly on whether sequential inference on target test data is feasible upon arrival [19, 45] or whether training objectives must be adapted [28, 40]. TTT [40] proposed fine-tuning the model parameters via rotation classification task as a proxy. On-target adaptation [44] used pseudo-labeling and contrastive learning to initialize the target-domain feature; each performed independently in their method. T3A [19] utilized pseudo-labeling to adjust the classifier prototype. More recently, TTAC [39] leveraged the clustering scheme to match target domain clusters to source domain categories. Nevertheless, existing TTA methods rely on specialized neural network architectures or only update a fraction of model weights yielding limited performance gain on the target data. For instance, TENT [45] proposed updating affine parameters in the batchnorm layers of convolutional neural networks (CNN), while AdaContrast [4] and SHOT [26] used an additional weight normalization classification layer with a projection head. In contrast, we show our method's superiority for various neural network architectures, including CNNs and vision transformers (ViTs) [11], without additional architectural requirements or different source model training strategies. Moreover, we update all model parameters, and our method is stable over a wide range of hyper-parameters, e.g., learning rate.

Test time augmentation methods [2, 29] are another popular line of domain adaptation research. GPS [29] learns optimal augmentation sub-policies by combining image transformations of RandAugment [10] that minimize calibrated log-likelihood loss [2] on the validation set. OptTTA [42] optimized the magnitudes of augmentation sub-policies using gradient descent to maximize mutual information and match feature statistics over augmented images. Though effective, these methods [29, 42, 46] are computationally expensive as all augmentation sub-policies need to be evaluated, making their real-time application difficult. Instead, our proposed adversarial augmentation policies can be learned online and are several orders faster than the existing learnable test-time augmentation strategies, e.g., [42].

3. Methodology

Overview. We first introduce our flipped cross-entropy (f-$\mathbb{CE}$) loss through the tight connection with the mutual information between the model's predictions and the test images. (Sec. 3.1). Using this equivalence, we derive our test-time loss function that implicitly incorporates teacher-student knowledge distillation. In Sec. 3.2, we propose to enhance the test-time teacher-student knowledge distillation by utilizing the consistency of the student model's predictions on the proposed adversarial augmentations with their corresponding refined soft-pseudo labels from the teacher model. We refine the soft-pseudo labels by averaging the teacher model's predictions on (1) weakly augmented test images; (2) nearest neighbors in the feature space. Furthermore, we propose an efficient online algorithm for learning adversarial augmentations (Sec. 3.3). Finally, in Sec. 3.4, we summarize our online test-time adaptation method, **TeSLA**, based on self-learning with the proposed automatic adversarial augmentation. The overall framework is shown in Fig. 2.

Problem setup. Given an existing model f_{θ_0}, parametrized by θ_0 and pre-trained on the inaccessible source data, we aim to improve its performance by updating all its parameters using the unlabeled streaming test data $\mathbf{X}$ during adaptation. Motivated by the concept of *mean teacher* [41] as the weight-averaged model over training steps that can yield a more accurate model, we build our online knowledge distillation framework upon teacher-student models. We utilize identical network architecture for teacher and student models. Each model $f = h \circ g$ comprises a backbone encoder $g : \mathbf{X} \rightarrow \mathbb{R}^D$ mapping unlabeled test image $\boldsymbol{x} \in \mathbf{X}$ to the feature representation $\mathbf{z} = g(\boldsymbol{x}) \in \mathbb{R}^D$ and a classifier (hypothesis) head $h : \mathbb{R}^D \rightarrow \mathbb{R}^K$ mapping $\mathbf{z}$ to the class prediction $\boldsymbol{y} \in \mathbb{R}^K$, where D and K denote the feature dimension and number of classes. The parameters of teacher θ_t and student θ_s are first initialized from the source model parameters θ_0. Then the teacher model's parameters θ_t are updated from the student model's parameters θ_s using an exponential moving average (EMA) with momentum coefficient α as: $\theta_t \leftarrow \alpha \cdot \theta_t + (1-\alpha) \cdot \theta_s$. Following [26], we freeze the classifier h's parameters and only update the encoder g's parameters during test-time adaptation.

3.1. Rationale Behind the f-CE Objective

We start by analyzing the rationale behind the proposed f-$\mathbb{CE}$ loss and its benefit for self-learning through the tight connection with the mutual information between the model's predictions and unlabeled test images. Before that, we first define the notations.

Notations. We define the random variables of unlabeled test images and the predictions from the student model f_s as $\mathbf{X}$ and $\mathbf{Y}$ and those of the teacher model f_t's soft pseudo-label predictions as $\hat{\mathbf{Y}}$. Furthermore, let $p_{\mathbf{Y}}$ denotes the marginal distribution over $\mathbf{Y}$, $p_{(\mathbf{Y},\mathbf{X})}$ be the joint distribution of $\mathbf{Y}$ and $\mathbf{X}$, and $p_{\mathbf{Y}|\mathbf{X}}$ be the conditional distribution of $\mathbf{Y}$ given $\mathbf{X}$. Then, the entropy of $\mathbf{Y}$ and the conditional entropy of $\mathbf{Y}$ given $\mathbf{X}$ can be defined as $\mathcal{H}(\mathbf{Y}) := \mathbb{E}_{p_{\mathbf{Y}}}\left[-\log p_{\mathbf{Y}}(\mathbf{Y})\right]$ and $\mathcal{H}(\mathbf{Y} \mid \mathbf{X}) := \mathbb{E}_{p_{(\mathbf{Y},\mathbf{X})}}\left[-\log p_{\mathbf{Y}|\mathbf{X}}(\mathbf{Y} \mid \mathbf{X})\right]$, respectively. Besides, let the flipped cross-entropy (f-$\mathbb{CE}$) given $\mathbf{X}$ be $\mathcal{H}\left(\mathbf{Y};\hat{\mathbf{Y}} \mid \mathbf{X}\right) := \mathbb{E}_{p_{(\mathbf{Y},\mathbf{X})}}\left[-\log p_{\hat{\mathbf{Y}}|\mathbf{X}}(\mathbf{Y} \mid \mathbf{X})\right]$.

Self-learning and mutual information. Based on the above definitions, the f-$\mathbb{CE}$ between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ conditioned over $\mathbf{X}$ has the tight connection with the mutual information $\mathcal{I}(\mathbf{Y};\mathbf{X})$ between unlabeled test images $\mathbf{X}$ and the predictions from the student model $\mathbf{Y}$ as follows:

$$\begin{aligned}\mathcal{H}\left(\mathbf{Y};\hat{\mathbf{Y}} \mid \mathbf{X}\right) &= \mathcal{H}(\mathbf{Y} \mid \mathbf{X}) + \mathcal{D}_{\text{KL}}\left(\mathbf{Y} \parallel \hat{\mathbf{Y}} \mid \mathbf{X}\right) \\ &= -\mathcal{I}(\mathbf{Y};\mathbf{X}) + \mathcal{H}(\mathbf{Y}) + \mathcal{D}_{\text{KL}}\left(\mathbf{Y} \parallel \hat{\mathbf{Y}} \mid \mathbf{X}\right)\end{aligned} \tag{1}$$

This implies that minimizing the f-$\mathbb{CE}$ ($\mathcal{H}\left(\mathbf{Y};\hat{\mathbf{Y}} \mid \mathbf{X}\right)$ in Eq. 1) and maximizing the entropy of class-marginal prediction $\mathcal{H}(\mathbf{Y})$ is equivalent to maximizing the mutual information between the test images and the student model's predictions $\mathcal{I}(\mathbf{Y};\mathbf{X})$ with a correction KL-divergence term $\mathcal{D}_{\text{KL}}\left(\mathbf{Y} \parallel \hat{\mathbf{Y}} \mid \mathbf{X}\right)$ involving soft-pseudo labels as:

$$\underbrace{\mathcal{H}\left(\mathbf{Y};\hat{\mathbf{Y}} \mid \mathbf{X}\right)}_{f\text{-}\mathbb{CE}} - \mathcal{H}(\mathbf{Y}) = -\underbrace{\mathcal{I}(\mathbf{Y};\mathbf{X})}_{\text{Mutual Info.}} + \mathcal{D}_{\text{KL}}\left(\mathbf{Y} \parallel \hat{\mathbf{Y}} \mid \mathbf{X}\right) \tag{2}$$

Thus, minimizing the left side of Eq. 2 with respect to the student model's parameters θ_s allows the student model to cluster test images using mutual information and criterion-corrected knowledge distillation from soft-pseudo labels. Using the above formulation, we define the following test-time objective to train the student model f_s on a batch of B test images $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_B\}$ with their corresponding soft-pseudo labels from teacher model $\hat{Y} = \{\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_B\}$.

$$\begin{aligned}\mathcal{L}_{\text{pl}}(X, \hat{Y}) = -\frac{1}{B}\sum_{i=1}^{B}\sum_{k=1}^{K} f_s(\boldsymbol{x}_i)_k \log((\hat{\boldsymbol{y}}_i)_k) \\ + \sum_{k=1}^{K} \hat{f}_s(X)_k \log(\hat{f}_s(X))_k\end{aligned} \tag{3}$$

where $\hat{f}_s(X) = \frac{1}{B}\sum_{i=1}^{B} f_s(\boldsymbol{x}_i)$ denotes the marginal class distribution over the batch of test images X.

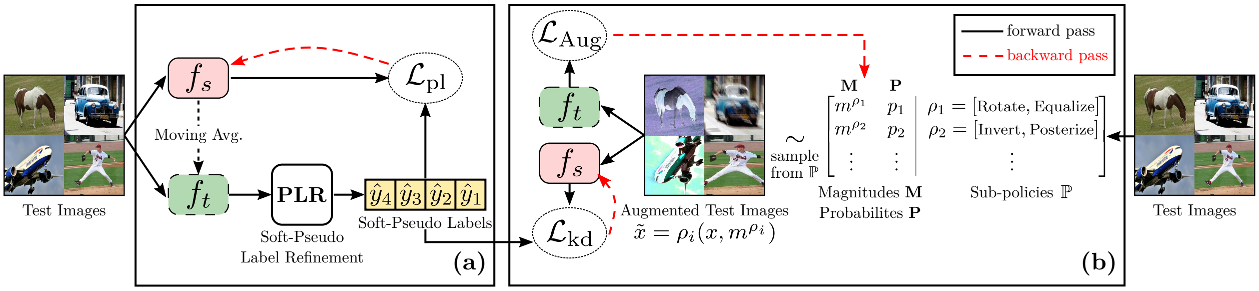

Test Images
f_s
$\mathcal{L}_{\text{pl}}$
Moving Avg.
f_t
PLR
$\hat{y}_4$ $\hat{y}_3$ $\hat{y}_2$ $\hat{y}_1$
Soft-Pseudo Labels
Soft-Pseudo
Label Refinement
(a)
$\mathcal{L}_{\text{Aug}}$
f_t
f_s
$\mathcal{L}_{\text{kd}}$
Augmented Test Images
$\tilde{x} = \rho_i(x, m^{\rho_i})$
$\sim$ sample from $\mathbb{P}$
M
P
m^{ρ_1} p_1
m^{ρ_2} p_2
$\rho_1 = $ [Rotate, Equalize]
$\rho_2 = $ [Invert, Posterize]
Magnitudes M
Probabilites P
Sub-policies $\mathbb{P}$
forward pass
backward pass
Test Images
(b)

Figure 2. **Overview of TeSLA Framework.** (a) The student model f_s is adapted on the test images by minimizing the proposed test-time objective $\mathcal{L}_{\text{pl}}$. The high-quality soft-pseudo labels required by $\mathcal{L}_{\text{pl}}$ are obtained from the exponentially weighted averaged teacher model f_t and refined using the proposed Soft-Pseudo Label Refinement (PLR) on the corresponding test images. The soft-pseudo labels are further utilized for teacher-student knowledge distillation via $\mathcal{L}_{\text{kd}}$ on the adversarially augmented views of the test images. (b) The adversarial augmentations are obtained by applying learned sub-policies sampled i.i.d from $\mathbb{P}$ using the probability distribution P with their corresponding magnitudes selected from M. The parameters M and P of the augmentation module are updated by the *unbiased gradient estimator* (Eq. 13) of the loss $\mathcal{L}_{\text{Aug}}$ computed on the augmented test images.

3.2. Knowledge Distillation via Augmentation

Self-learning from adversarial augmentations. As illustrated in Fig. 1, adversarial augmentations are learned by pushing their feature representations toward the decision boundary (*Expansion* phase), followed by updating the student model f_s to match its prediction on these augmented images with their corresponding soft-pseudo label (*Separation* phase), yielding better separation of features into their respective classes. In our method setup, we continually learn automatic augmentations of the test images (Sec. 3.3) that are **adversarial** to the current teacher model f_t and enforce consistency between the student model's f_s predictions on the augmented views and the corresponding soft-pseudo labels. Since we freeze the classifier module, f_s and f_t share the common decision boundary. Let $\tilde{\boldsymbol{x}}$ denote the learned adversarial augmented view of $\boldsymbol{x}$ and $\hat{\boldsymbol{y}}$ denote the corresponding soft-pseudo label. We minimize the following knowledge distillation loss $\mathcal{L}_{\text{kd}}$ using KL-divergence to distill knowledge from the teacher model to the student model:

$$\mathcal{L}_{\text{kd}}(\tilde{\boldsymbol{x}}, \hat{\boldsymbol{y}}) = \mathcal{D}_{\text{KL}}(\hat{\boldsymbol{y}} \| f_s(\tilde{\boldsymbol{x}})) \tag{4}$$

Soft pseudo label refinement (PLR). We refine the quality of soft pseudo-labels and the feature representations by averaging the teacher model's outputs on multiple weakly augmented image views $\rho_w(\boldsymbol{x})$ (via flipping, cropping augmentations) of the same test image $\boldsymbol{x}$ as follows:

$$\mathbf{z}_t, \boldsymbol{y}_t \leftarrow \mathbb{E}_{\mathbf{u} \in \rho_w(\boldsymbol{x})}[g_t(\mathbf{u}), h_t(g_t(\mathbf{u}))] \tag{5}$$

The refined *feature representations* $\mathbf{z}_t$ and the *softmaxed pseudo labels* $\boldsymbol{y}_t$ from the encoder g_t and the classifier h_t are further stored in an online memory queue $\mathbf{Q}$ of fixed size. The final soft pseudo-label is computed by averaging the refined soft pseudo-labels of n-nearest neighbors of the current test image in the feature space as follows:

$$\begin{gathered}\mathbf{Q}[\arg\max(\boldsymbol{y}_t)].\text{append}(\{\mathbf{z}_t, \boldsymbol{y}_t\}) \\ \hat{\boldsymbol{y}} = \frac{1}{n}\sum \mathcal{N}_{\mathbf{Q},n}(\mathbf{z}_t)\end{gathered} \tag{6}$$

where $\mathbf{Q}$ denotes an online memory of class-balanced queues (with size $|\mathbf{Q}[\arg\max(\boldsymbol{y}_t)]| \leq N_{\mathbf{Q}}$) of the refined feature representations and the softmaxed label predictions on the previously seen test images, and $\mathcal{N}_{\mathbf{Q},n}(\mathbf{z}_t)$ denotes soft labels of n-nearest neighbors of $\mathbf{z}_t$ from $\mathbf{Q}$.

3.3. Learning Adversarial Data Augmentation

This section presents our efficient online method for learning adversarial data augmentation for the current teacher f_t. We first introduce our adversarial augmentation search space. Then, we describe our differentiable strategy to optimize and sample adversarial augmentations that push the feature representations of the test images toward the uncertain region in the close vicinity of the decision boundary.

Policy search space. Let $\mathbb{O}$ be a set of all image transformation operations $\mathcal{O} : \mathbb{R}^{H \times W \times 3} \rightarrow \mathbb{R}^{H \times W \times 3}$ defined in our search space. In particular, we use image transformations Auto-Contrast, Equalize, Invert, Solarize, Posterize, Contrast, Brightness, Color, ShearX, ShearY, TranslateX, TranslateY, Rotate, and Sharpness. Each transformation $\mathcal{O}$ is specified with its magnitude parameter $m \in [0, 1]$. Since the output of some image operations may not be conditional on its magnitude (e.g., Equalize) or may not be differentiable (e.g., Posterize), for such image operations, we use the straight-through gradient estimate w.r.t. each pixel as: $\partial \mathcal{O}\left(\boldsymbol{x}_{ij}\right) / \partial m = \mathbf{1}$ for magnitude optimization.

We define a sub-policy ρ as a combination of N image operations (sub-policy dimension) from $\mathbb{O}$ that is applied sequentially to a given image $\boldsymbol{x}$ as:

$$\rho\left(\boldsymbol{x}, m^{\rho}\right) = \mathcal{O}_1^{\rho} \cdots \mathcal{O}_N^{\rho}\left(\boldsymbol{x}, m_N^{\rho}\right) \tag{7}$$

where $m^{\rho} = [m_1^{\rho}, \ldots, m_N^{\rho}]$ denotes the magnitude set of the sub-policy ρ. We also define $\mathbb{P} = \{\rho_1, \rho_2, \ldots\}$ as the set of all possible sub-policies.

Policy evaluation. We propose learnable adversarial augmentation, aiming to optimize and sample data augmentations as a proxy for simulating data in the uncertain region of the feature space. Given the teacher model f_t, a sub-policy ρ with magnitude m is evaluated for a test image $\boldsymbol{x}$ using the following adversarial objective:

$$\mathcal{L}_{\text{aug}}(\boldsymbol{x}, \rho) = \sum_{k=1}^{K} f_t\left(\tilde{\boldsymbol{x}}\right) \log\left(f_t\left(\tilde{\boldsymbol{x}}\right)\right) + \lambda_1 r(\tilde{\boldsymbol{x}}, \boldsymbol{x}) \tag{8}$$

where $\tilde{\boldsymbol{x}} = \rho\left(\boldsymbol{x}, m\right)$ denotes the adversarial augmented image and $r\left(\cdot, \cdot\right)$ regularizes augmentation severity. The hyperparameter λ_1 controls the augmentation severity. We optimize the adversarial augmentation loss $\mathcal{L}_{\text{aug}}$ with respect to the sub-policy ρ's parameters. Minimizing $\mathcal{L}_{\text{aug}}$ is equivalent to maximizing the entropy of the teacher model's prediction on the augmented image $\tilde{\boldsymbol{x}}$, thus pushing its feature representation towards the decision boundary (first term). For the augmentation regularization (second term), we use the mean squared distance function between the teacher encoder's L internal layers' activations of the adversarial augmentation and non-augmented versions of image $\boldsymbol{x}$ defined as:

$$r(\tilde{\boldsymbol{x}}, \boldsymbol{x}) = \frac{1}{L} \sum_{l=1}^{L} \left\| \mu_l(\tilde{\boldsymbol{x}}) - \mu_l(\boldsymbol{x}) \right\|^2 \tag{9}$$

where $\mu_l\left(\cdot\right)$ denotes the mean activation of the l^{th} layer of the teacher encoder g_t.

Policy optimization. Let $M = [m^{\rho_1}, \ldots, m^{\rho_{\|\mathbb{P}\|}}]$ denote the set of magnitudes m^{ρ} of all sub-policies $\rho \in \mathbb{P}$ and $P = [p_1, \ldots, p_{\|\mathbb{P}\|}]$ denote the probability of selecting a sub-policy from $\mathbb{P}$. The expected policy evaluation loss (Eq. 8) for an image $\boldsymbol{x}$ over the policy search space is given by:

$$\mathbb{E}[\mathcal{L}_{\text{aug}}(\boldsymbol{x})] = \sum_{i=1}^{\|\mathbb{P}\|} p_i \cdot \mathcal{L}_{\text{aug}}(\boldsymbol{x}, \rho_i) \tag{10}$$

Evaluating the gradient of $\mathbb{E}[\mathcal{L}_{\text{aug}}(\boldsymbol{x})]$ w.r.t. M and P can become computationally expensive as $\|\mathbb{P}\| \sim \binom{\|\mathbb{O}\|}{N}$. Thus, we use the following re-parameterization trick to estimate its *unbiased gradient*:

$$\nabla \mathbb{E}[\mathcal{L}_{\text{aug}}(x)] = \delta(\boldsymbol{x}, \mathbb{P}) = \sum_{i=1}^{\|\mathbb{P}\|} \nabla (p_i \cdot \mathcal{L}_{\text{aug}}(\boldsymbol{x}, \rho_i)) \tag{11}$$

$$= \sum_{i=1}^{\|\mathbb{P}\|} p_i (\nabla \mathcal{L}_{\text{aug}}(\boldsymbol{x}, \rho_i) + \mathcal{L}_{\text{aug}}(\boldsymbol{x}, \rho_i) \cdot \nabla \log p_i)$$

Thus, an unbiased estimator of $\delta(\boldsymbol{x}, \mathbb{P})$ can be written as:

$$\hat{\delta}(\boldsymbol{x}, \rho_i) = \nabla \mathcal{L}_{\text{aug}}(\boldsymbol{x}, \rho_i) + \mathcal{L}_{\text{aug}}(\boldsymbol{x}, \rho_i) \cdot \nabla \log p_i \tag{12}$$

where index i is sampled from the probability distribution P. For an online batch of B test images $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_B\}$, we apply augmentation sub-policies $\{\rho_{i_1}, \rho_{i_2} \ldots, \rho_{i_B}\}$ from $\mathbb{P}$ where $\{i_1, i_2, \ldots, i_B\}$ are sampled i.i.d from distribution P, and update the parameters P and M with the following stochastic gradient update rule:

$$[P, M] \leftarrow [P, M] - \frac{\gamma}{B} \sum_{j=1}^{B} \hat{\delta}(\boldsymbol{x}_j, \rho_{i_j}) \tag{13}$$

where γ is the learning rate. We set $\gamma = 0.1$ for all experiments without the need for hyperparameter tuning.

3.4. Self-Learning With Adversarial Augmentation

We minimize the following overall objective $\mathcal{L}_{\text{TeSLA}}$ for training the student model f_s on a batch of B test images $X = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_B\}$ and their adversarial augmented views $\tilde{X} = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_B\}$ with the corresponding refined soft-pseudo labels from the teacher model $\hat{Y} = \{\hat{\boldsymbol{y}}_1, \ldots, \hat{\boldsymbol{y}}_B\}$:

$$\mathcal{L}_{\text{TeSLA}}(X, \tilde{X}, \hat{Y}) = \mathcal{L}_{\text{pl}}(X, \hat{Y}) + \frac{\lambda_2}{B} \sum_{i=1}^{B} \mathcal{L}_{\text{kd}}(\tilde{\boldsymbol{x}}_i, \hat{\boldsymbol{y}}_i) \tag{14}$$

where λ_2 is a hyper-parameter for the knowledge distillation. The adversarial augmented views $\tilde{X}$ are obtained by sampling *i.i.d.* augmentation sub-policies from $\mathbb{P}$ for every test image in X with probability P and applying the corresponding magnitude from M. Concurrently, we also optimize M and P using the policy optimization mentioned in Sec. 3.3.

4. Experiments

4.1. Datasets and Experimental Settings

We evaluate and compare TeSLA against state-of-the-art (SOTA) test-time adaptation algorithms for both classification and segmentation tasks under three types of test-time distribution shifts resulting from (1) **common image corruption**, (2) **synthetic to real data transfer**, and (3) **measurement shifts** on medical images. The latter is characterized by a change in medical imaging systems, e.g., different scanners across hospitals or various staining techniques.

TTA protocols. We adopt the TTA protocols in [39] and categorize competing methods based on two factors: (i) source training objective and (ii) sequential or non-sequential inference. First, unlike [39], we use **Y** to indicate if access to the source domain's information, e.g., feature statistics is possible or if the source training objective is allowed to be modified; otherwise, we use **N**. Next, we use **O** to indicate a one-pass adaptation and evaluation protocol (one epoch) for sequential test data and **M** to show a multi-pass adaptation on all test data and inference after several epochs. Thus, we end up with four possible TTA protocols, namely **N-M**, **Y-M**, **N-O**, and **Y-O**. Unlike previous works [28, 39], TeSLA does not rely on any source domain feature distribution, yielding two realistic TTA protocols: **N-M** and **N-O**.

Hyperparameters for all experiments. We set $\lambda_1 = \lambda_2 = 1.0$ for all experiments (cf. ablation in **Appendix E**). For the adversarial augmentation module, we use sub-policy dimension $N = 2$ (except for VisDA-C [33], $N = 4/3$ for N-O/N-M protocols) (cf. **Appendix E**) and Adam optimizer [22]. For pseudo-label refinement (PLR), we use five weak augmentations composed of random flips and resize crop. More details of hyperparameters used for each experiment can be found in **Appendix A**.

Common image corruptions. To evaluate TeSLA's efficacy for the classification task, we use **CIFAR10-C/CIFAR100-C** [16] and large-scale **ImageNet-C** [16] datasets each containing 19 types of corruptions applied to the clean test set with five levels of severity. We perform validation on 4 out of 19 types of corruption (SPATTER, GAUSSIAN BLUR, SPECKLE NOISE, SATURATE) to select hyperparameters and test on the remaining 15 corruptions at the maximum severity level of 5. Following [28], we train the ResNet50 [15] on the clean CIFAR10/CIFAR100/ImageNet training set and adapt it to classify the unlabeled corrupted test set.

Synthetic to real data adaption. We use challenging and large-scale **VisDA-C**, and **VisDA-S** [33] datasets for evaluating synthetic-to-real data adaptation at test-time for classification and segmentation tasks, respectively. Following [4, 26], we adapt the ResNet101 network pre-trained on synthetic images to classify 12 vehicle classes on the photo-realistic images of VisDA-C. While for VisDA-S, we adapt the DeepLab-v3 [5] backbone pre-trained on synthetic GTA5 [36] to the Cityscapes [7] to segment 19 classes.

Measurement shifts on medical images. We access TeSLA on staining variations for the tissue type classification of hematoxylin & eosin (H&E) stained patches from colorectal cancer tissue slides. We use MobileNetV2 [38] trained on the source **Kather-19** dataset [20] and adapt it to the target **Kather-16** dataset [21] on four tissue categories: tumor, stroma, lymphocyte, and mucosa. We also evaluate TeSLA for variations in scanners and imaging protocols in multi-site medical images on two magnetic resonance imaging (MRI) datasets for the segmentation task, namely the **multi-site prostate MRI** [27] and **spinal cord grey matter segmentation (SCGM)** [34]. Following [42], we adapt the U-Net [37] from site 1 to sites {2,3,4} of the spinal cord dataset, and from sites {A,B} to sites {D,E,F} of the prostate dataset.

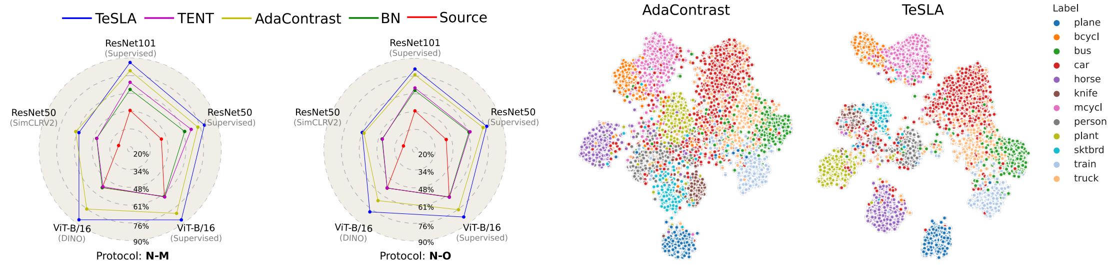
Competing baselines. We compare TeSLA with the following SFDA and TTA baselines, including direct inference of the trained source model on the target test data without adaptation (**Source**). We also implement the SFDA-based pseudo-labeling baselines from a basic pseudo-labeling approach (**PL**) to the **SHOT** approach [26] based on training the feature extraction module by adopting the information maximization loss to make globally diverse but individually certain predictions on the target domain. For recent TTA methods, we compare our method against **TTT++** [28], **AdaContrast** [4], **CoTTA** [47], basic **TENT** [45] and **BN** [18] as representative methods based on feature distribution alignment via stored statistics, contrastive learning, augmentation-averaged predictions, entropy minimization, and batch normalization statistic. Furthermore, we compare TeSLA to more recent methods based on test-time augmentation policy learning, **OptTTA** [42], and anchor clustering **TTAC** [39]. By default, [4, 18, 26] follow the N-M protocol, and we adapt them to the N-O protocol, while [42, 45, 47], by default, follow the N-O protocol and adapt them to the N-M protocol setting. Using source features distribution, TTAC and TTT++ follow Y-O and Y-M. We report error rates on classification tasks and Mean Intersection over Union (mIoU)/ Dice scores on segmentation tasks. For consistency across baselines and TTA protocols, we do not use any specialized model architectures (e.g., projection head, weight-normalized classifier layer) for SHOT [26] and AdaContrast [4].

Table 1. **Comparison of SOTA TTA methods under different protocols** evaluated on CIFAR-10/100-C, ImageNet-C, VisDA-C and Kather-16 datasets. We report the average error computed over 15 test corruptions for the common image corruption shifts. We also report *Class Avg.* in % error rates for synthetic-to-real and measurement shifts over 3 and 10 seeds, respectively.

Method	Protocol	**Common Image Corruptions**						**Syn-to-Real**		**Measurement Shift**	
		CIFAR10-C		CIFAR100-C		ImageNet-C		VisDA-C		Kather-16	
		O	M	O	M	O	M	O	M	O	M
Source	N	29.1		60.4		81.8		51.5		32.0	
BN [18, 30]	N	15.6	15.4	43.7	43.3	67.7	67.6	35.4	35.0	18.3	18.2
TENT [45]	N	14.1	12.9	39.0	36.5	57.4	54.2	33.5	29.3	16.2	12.0
SHOT [26]	N	13.9	14.2	39.2	38.7	68.7	68.2	29.4	24.5	14.7	12.0
AdaContrast [4]	N	-						23.1	20.2	-	
TTT++ [28]	Y	15.8	9.8	44.4	34.1	59.3	-	35.2	34.1	16.7	7.9
TTAC [39]	Y	13.4	**9.4**	41.7	33.6	58.7	-	32.2	31.1	9.6	5.5
TeSLA	N	12.5	9.7	38.2	32.9	55.0	**51.5**	**17.8**	**13.5**	**9.2**	3.3
TeSLA-s	Y	**12.1**	9.7	**37.3**	**32.6**	**53.1**	-	24.0	17.9	9.9	**3.1**

4.2. Results

Classification task. In Table 1, we summarize the average (Avg) classification error rates (%) of TeSLA against several SOTA **SFDA** and **TTA** baselines on the **common image corruptions** of CIFAR10-C, CIFAR100-C, ImageNet-C; **synthetic to real data** domain shifts of VisDA-C; and **measurement shifts** on the Kather-16 dataset. We also report per-class top-1 accuracies for VisDA-C and Kather-16 datasets and corruption-wise error rates on CIFAR-10-C/CIFAR-100-C and ImageNet-C datasets in **Appendix B**. TeSLA easily surpasses all competing methods under the protocols N-O and N-M on all the datasets. In particular, TeSLA outper-



(a) (b)

Figure 3. **Ablation experiments on the VisDA-C**. **(a) Source training strategy and model architecture.** TeSLA outperforms the competing TTA methods across varied source training strategies (Supervised, SimCLRV2 [6], and DINO [3]) and model architectures (ResNet50, ResNet101 [15], and ViT-B/16 [11]) using the same set of hyperparameters for both protocols (N-O/N-M). Each vertex represents the mean class avg accuracy over 3 seeds. **(b) t-SNE visualization** [43] comparison of feature embeddings for AdaContrast [4] and TeSLA.

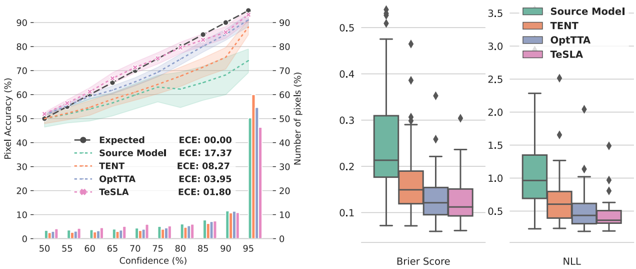


Figure 4. **Model calibration and uncertainty estimation.** Model calibration using reliability diagrams [31] and model uncertainty estimation using Brier Score and Negative Log Likelihood (NLL) of the Source Model, TENT [45], OptTTA [42] and, TeSLA on the prostate dataset for test-time adaptation.

Table 2. **Semantic segmentation results** (*Class Avg.* mIoU in %) on the VisDA-S (GTA5 → Cityscapes) test-time adaptation task.

Protocol		Source	BN [30]	TENT [45]	PL	CoTTA [47]	**TeSLA**
N	O	35.3	36.7	38.3	38.8	37.0	**44.5**
N	M		38.4	39.2	38.6	39.9	**46.0**

forms the second-best baseline under N-O/N-M protocol by a (%) margin 1.4/3.2 on CIFAR10-C, 0.8/3.6 on CIFAR100-C, 2.4/2.7 on ImageNet-C, 5.3/6.7 on VisDA-C, and 5.5/8.7 on the Kather-16 datasets. Despite not utilizing any source feature statistic, TeSLA (N-O/N-M) is either competitive or surpasses TTAC [39] and TTT++ [28] (Y-O/Y-M) that benefit from the source dataset feature statistics during adaptation. Including the global source feature alignment module of TTAC with TeSLA (**TeSLA-s**) could further improve its performance under (Y-O/Y-M) protocols for slight domain shifts (e.g., common image corruptions). We report the results of AdaContrast for VisDA-C only as the implementation for other datasets are not provided. We do not report TeSLA-s results on ImageNet-C under the Y-M protocol, as previous baselines were evaluated only Y-O protocol.

Segmentation task. Unlike TeSLA, several TTA methods e.g., AdaContrast [4], SHOT [26], TTT++ [28], TTAC [39], have *only* been evaluated on the classification task. Therefore, we compare TeSLA against the current SOTA **TTA** methods applied to the segmentation task on **synthetic-to-real data transfer** of the VisDA-S dataset in Table 2. TeSLA significantly outperforms the competing methods and achieves the best mIOU scores for both N-O and N-M protocols, beating the second-best baseline by a margin of +5.7% and +6.1%, respectively. In Table 3, we compare TeSLA against the recent SOTA test-time augmentation policy method, OptTTA [42] and other TTA baselines for the inter-site adaptation on two challenging MRI datasets mentioned in Sec. 4.1. TeSLA convincingly outperforms all other methods on severe measurement shifts across sites under **N-O** and **N-M** protocols.

Table 3. **Semantic segmentation results** (*Class Avg.* volume-wise *mean* Dice in %) on the cross-site spinal cord and prostate MRI test-time adaptation tasks.

Protocol		Source	BN [30]	TENT [45]	PL	OptTTA [42]	**TeSLA**
Spinal Cord	`Site{1} → Sites{2,3,4}`						
N	O	76.0±11.8	81.6±8.3	81.1±9.1	81.7±8.6	84.1±4.8	**85.3**±5.8
N	M		84.3±4.8	84.4±4.7	84.3±4.7	84.3±4.4	**85.4**±4.4
Prostate	`Sites{A,B} → Sites{D,E,F}`						
N	O	60.5±27.0	72.1±15.2	74.7±17.9	72.4±15.2	83.1±7.7	**83.5**±6.5
N	M		73.1±18.0	81.2±9.3	81.1±9.2	83.4±7.7	**84.3**±5.8

Computational cost for adversarial augmentations. As shown in Table 4, TeSLA learns the adversarial augmentation in an online manner, and thus its runtime is several orders faster than learnable test-time augmentation OptTTA [42], while comparable to that of static augmentation policies with an additional overhead of only 0.10 GPU hours/epoch on the VisDA-C dataset.

Test-time feature visualization. Fig. 3b compares t-SNE projection [43] of the encoder's features of AdaContrast [4] with TeSLA on the VisDA-C. TeSLA shows better inter-class separation for features than AdaContrast, as supported by an improved silhouette score from 0.149 to 0.271.

Model calibration and uncertainty estimation. Entropy minimization-based TTA methods [26, 45] explicitly make the model confident in their predictions, resulting in poor model calibration. For trusting the adapted model, it should

Table 4. **Performance comparison of adversarial augmentation** learned by TeSLA during adaptation against the prior art augmentation methods on runtime (GPU hours/epoch on NVIDIA GeForce RTX 3090) and *Class Avg.* accuracy in % on the VisDA-C dataset.

Augmentation Type	OptTTA	TeSLA RA [10]	TeSLA AA [8]	TeSLA $N = 2$	TeSLA $N = 3$
Runtime	4.50	0.05	0.06	0.16	0.18
Class Avg. Acc. (N-O)	-	80.2	81.2	81.5	82.2
Class Avg. Acc. (N-M)	-	84.7	85.7	86.3	86.5

output reliable confidence estimates matching its true underlying performance on the test images. Such a well-behaved model is characterized by expected **calibration error (ECE)** through **reliability diagram** [31], and uncertainty metrics of **brier score** and **negative log-likelihood (NLL)** [14]. In Fig. 4, we plot the reliability diagram (dividing the probability range $[0, 0.5]$ into 10 bins), and uncertainty metrics of the Source model, TENT [45] and OptTTA [42] against TeSLA for the inter-site model adaptation on the prostate dataset [27]. We observe that TeSLA achieves the best model calibration (lowest ECE of 1.80%) and lowest Brier and NLL scores of 0.12 and 0.24 ($p < 0.006$ against TENT).

4.3. Ablation Studies

In this section, we scrutinize the roles played by different components of TeSLA. All ablations are performed on the synthetic-to-real test data adaptation of the VisDA-C dataset.

Loss functions for self-learning. We compare TTA performance of our self-learning objective $\mathcal{L}_{\text{pl}}$ with the proposed flipped cross-entropy loss f-$\mathbb{CE}$ and the basic cross-entropy loss $\mathbb{CE}$ in Fig. 5-(1). We do not use the proposed augmentation module and pseudo-label refinement in this experiment. We observe that f-$\mathbb{CE}$ alone improves the accuracy of the Source model in the N-M protocol by +23.1% compared to +16.6% improvement by $\mathbb{CE}$. Using $\mathcal{L}_{\text{pl}}$ (Eq. 3) further improves the accuracy to a margin of +23.8%.

Contribution of individual components: PLR, $\mathcal{L}_{\text{pl}}$, and $\mathcal{L}_{\text{kd}}$. Our test-time objective $\mathcal{L}_{\text{TeSLA}}$ has three components–(i) self-learning loss $\mathcal{L}_{\text{pl}}$, (ii) pseudo-label refinement (PLR), and (iii) knowledge distillation $\mathcal{L}_{\text{kd}}$ with adversarial augmentation. In Fig. 5-(3), we study the effect of accruing individual components to $\mathcal{L}_{\text{TeSLA}}$ (starting from source trained model) for the test-time adaptation accuracy. We observe that $\mathcal{L}_{\text{pl}}$ alone improves the model's accuracy from 48.5% to 72.3% while adding PLR and $\mathcal{L}_{\text{kd}}$ further boosts it to 81.6% (+9.3) and 86.3% (+4.7), respectively.

PLR: ensembling on weak augmentations (Ens) and nearest neighbor averaging (NN). We also conduct an ablation on the two pseudo-label refinement approaches – (i) ensembling the teacher model's output on five weak augmentations. (ii) averaging soft pseudo labels of $n = 10$ nearest neighbors in the feature space (cf. sensitivity test in **Appendix E**). Fig. 5-(2) shows that combining both approaches gives better results than using them individually.

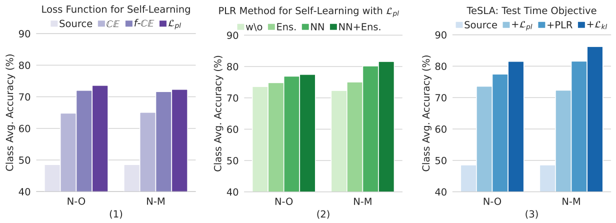


Figure 5. **Ablation experiments on the VisDA-C.** (*left to right*) (1) Ablation results for self-learning loss functions, including f-$\mathbb{CE}$, $\mathbb{CE}$, and $\mathcal{L}_{\text{pl}}$; (2) ablation on different PLR methods: ensembling predictions on 5 weak augmentations, averaging predictions of the 10 nearest neighbors in the feature space; (3) ablation on different components of TeSLA's objective function: $\mathcal{L}_{\text{pl}}$, PLR, and $\mathcal{L}_{\text{kd}}$.

Adversarial augmentation. Table 4 shows the benefit of using our adversarial augmentation ($N = 2, 3$) with TeSLA against (i) **RandAugment (RA)** [10], (ii) **AutoAugment (AA)** [9] (optimized for ImageNet), and (iii) OptTTA [42]. We achieve far better *Class Avg.* accuracy on the VisDA-C at the cost of slight runtime overhead compared to static augmentation policies for protocols (N-O, N-M). In addition, TeSLA augmentation (TeAA) consistently improves the performance of other TTA methods (Table 5).

Table 5. *Class Avg. accuracy (%)* of **TeSLA augmentation (TeAA)** with other TTA objectives under N-O protocol on VisDA-C dataset.

TENT	**+TeAA**	SHOT	**+TeAA**	AdaContrast	**+TeAA**
66.5	**72.0**	70.6	**73.1**	76.9	**79.1**

Source training strategies and model architectures. The ablation results (Fig. 3a) on the VisDA-C show that TeSLA surpasses competing TTA baselines for different source training strategies, including **Supervised** and self-supervised (**SimCLRV2** [6], **DINO** [3]). Moreover, we ablate TeSLA over different architectures, ranging from ResNet-50/101 [15] to the recent vision transformer ViT-B [11] using the same set of hyperparameters, showing the merits of no reliance on the network architecture or source training strategy.

Finally, in **Appendix E**, we provide detailed sensitivity results with respect to all hyperparameters.

5. Conclusion

We introduced TeSLA, a novel self-learning algorithm for test-time adaptation that utilizes automatic adversarial augmentation. TeSLA is agnostic to the model architecture and source training strategies and gives better model calibration than other TTA methods. Through extensive experiments on various domain shifts (measurement, common corruption, synthetic to real), we show TeSLA's superiority over previous TTA methods on classification and segmentation tasks. Note that TeSLA assumes class uniformly when implicitly maximizing mutual information and can be improved by incorporating prior, e.g., class label distribution statistics.

References

[1] Peshal Agarwal, Danda Pani Paudel, Jan-Nico Zaech, and Luc Van Gool. Unsupervised robust domain adaptation without source data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2009–2018, 2022. 1, 2

[2] Arsenii Ashukha, Alexander Lyzhov, Dmitry Molchanov, and Dmitry Vetrov. Pitfalls of in-domain uncertainty estimation and ensembling in deep learning. In *International Conference on Learning Representations*, 2019. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021. 7, 8

[4] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 295–305, 2022. 2, 6, 7

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 6

[6] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. *Advances in neural information processing systems*, 33:22243–22255, 2020. 7, 8

[7] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[8] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018. 8

[9] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 113–123, 2019. 8

[10] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 2, 8

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 7, 8

[12] Cian Eastwood, Ian Mason, Christopher K. I. Williams, and Bernhard Schölkopf. Source-free adaptation to measurement shift via bottom-up feature restoration. In *International Conference on Learning Representations*, 2022. 2

[13] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2

[14] Alvaro Gomariz, Tiziano Portenier, César Nombela-Arrieta, and Orcun Goksel. Probabilistic spatial analysis in quantitative microscopy with uncertainty-aware cell detection using deep bayesian regression of density maps. *arXiv preprint arXiv:2102.11865*, 2021. 8

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6, 7, 8

[16] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 6

[17] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. Pmlr, 2018. 2

[18] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015. 6

[19] Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. *Advances in Neural Information Processing Systems*, 34:2427–2440, 2021. 1, 2

[20] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019. 6

[21] Jakob Nikolas Kather, Frank Gerrit Zöllner, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Cleo-Aron Weis. Collection of textures in colorectal cancer histology, May 2016. 6

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] Jogendra Nath Kundu, Naveen Venkat, R Venkatesh Babu, et al. Universal source-free domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4544–4553, 2020. 1, 2

[24] Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, R Venkatesh Babu, et al. Towards inheritable models for open-set domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12376–12385, 2020. 2

[25] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9641–9650, 2020. 1, 2

[26] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for

unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020. 2, 3, 6, 7

[27] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE Transactions on Medical Imaging*, 2020. 6, 8

[28] Yuejiang Liu, Parth Kothari, Bastien Germain van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. TTT++: When does self-supervised test-time training fail or thrive? In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021. 1, 2, 6, 7

[29] Alexander Lyzhov, Yuliya Molchanova, Arsenii Ashukha, Dmitry Molchanov, and Dmitry Vetrov. Greedy policy search: A simple baseline for learnable test-time augmentation. In *Conference on Uncertainty in Artificial Intelligence*, pages 1308–1317. PMLR, 2020. 2

[30] Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. *arXiv preprint arXiv:2006.10963*, 2020. 6, 7

[31] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625–632, 2005. 7, 8

[32] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 2

[33] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge, 2017. 6

[34] Ferran Prados, John Ashburner, Claudia Blaiotta, Tom Brosch, Julio Carballido-Gamio, Manuel Jorge Cardoso, Benjamin N Conrad, Esha Datta, Gergely Dávid, Benjamin De Leener, et al. Spinal cord grey matter segmentation challenge. *Neuroimage*, 152:312–329, 2017. 6

[35] Zhen Qiu, Yifan Zhang, Hongbin Lin, Shuaicheng Niu, Yanxia Liu, Qing Du, and Mingkui Tan. Source-free domain adaptation via avatar prototype generation and adaptation. In *International Joint Conference on Artificial Intelligence*, 2021. 2

[36] Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *European Conference on Computer Vision (ECCV)*, volume 9906 of *LNCS*, pages 102–118. Springer International Publishing, 2016. 6

[37] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 6

[38] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 6

[39] Yongyi Su, Xun Xu, and Kui Jia. Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022. 1, 2, 5, 6, 7

[40] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020. 2

[41] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017. 1, 3

[42] Devavrat Tomar, Guillaume Vray, Jean-Philippe Thiran, and Behzad Bozorgtabar. OptTTA: Learnable test-time augmentation for source-free medical image segmentation under domain shift. In *Medical Imaging with Deep Learning*, 2022. 2, 3, 6, 7, 8

[43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[44] Dequan Wang, Shaoteng Liu, Sayna Ebrahimi, Evan Shelhamer, and Trevor Darrell. On-target adaptation. *arXiv preprint arXiv:2109.01087*, 2021. 2

[45] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021. 1, 2, 6, 7, 8

[46] Guotai Wang, Wenqi Li, Michael Aertsen, Jan Deprest, Sébastien Ourselin, and Tom Vercauteren. Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks. *Neurocomputing*, 338:34–45, 2019. 2

[47] 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. 1, 6, 7

[48] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Casting a bait for offline and online source-free domain adaptation. *arXiv preprint arXiv:2010.12427*, 2020. 2

[49] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8978–8987, 2021. 2

[50] Shiqi Yang, Yaxing Wang, Kai Wang, Shangling Jui, and Joost van de Weijer. Attracting and dispersing: A simple approach for source-free domain adaptation. *arXiv preprint arXiv:2205.04183*, 2022. 2

[51] Yifan Zhang, Ying Wei, Qingyao Wu, Peilin Zhao, Shuaicheng Niu, Junzhou Huang, and Mingkui Tan. Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Transactions on Image Processing*, 29:7834–7844, 2020. 2