Test-Time Adaptation via Conjugate Pseudolabels

Anonymous Author(s)
Affiliation
Address
email

Abstract

Test-time adaptation (TTA) refers to adapting neural networks to distribution shifts, with access to only the unlabeled test samples from the new domain at test-time. Prior TTA methods optimize over unsupervised objectives such as the entropy of model predictions in TENT [42], but it is unclear what exactly makes a good TTA loss. In this paper, we start by presenting a surprising phenomenon: if we attempt to *meta-learn* the "best" possible TTA loss over a wide class of functions, then we recover a function that is *remarkably* similar to (a temperature-scaled version of) the softmax-entropy employed by TENT. This only holds, however, if the classifier we are adapting is trained via cross-entropy loss; if the classifier is trained via squared loss, a different "best" TTA loss emerges. To explain this phenomenon, we analyze test-time adaptation through the lens of the training losses's *convex conjugate*. We show that under natural conditions, this (unsupervised) conjugate function can be viewed as a good local approximation to the original supervised loss and indeed, it recovers the "best" losses found by meta-learning. This leads to a generic recipe that can be used to find a good TTA loss for *any* given supervised training loss function of a general class. Empirically, our approach consistently dominates other TTA alternatives over a wide range of domain adaptation benchmarks. Our approach is particularly of interest when applied to classifiers trained with *novel* loss functions, e.g., the recently-proposed PolyLoss [19] function, where it differs substantially from (and outperforms) an entropy-based loss. Further, we show that our conjugate based approach can also be interpreted as a kind of self-training using a very specific soft label, which we refer to as the *conjugate pseudolabel*. Overall, our method provides a broad framework for better understanding and improving test-time adaptation.

1 Introduction

Modern deep networks perform exceeding well on new test inputs that are close to the training distribution. However, this performance dramatically decreases on test inputs drawn from a different distribution. While there is a large body of work on improving the robustness of models, most robust training methods are highly specialized to the setting they cater to. For e.g., they assume pre-specified perturbations, subpopulations, and spurious correlations, or access to unlabeled data from the target distribution, and most methods offer close to no improvement on general distribution shifts beyond what they were trained for [9, 16].

In practice, it is often cumbersome (or even impossible) to precisely characterize all possible distribution shifts a model could encounter and then train accordingly. Instead, a model already trained on some source data must be able to *adapt* to the new test data from a different target domain, while making predictions i.e. test-time adaptation. This setting of *test-time adaptation* (TTA) has gained interest in recent years [5, 39, 42, 46]. TTA is typically accomplished by updating the source model parameters via a few steps of optimization on an *unsupervised objective* involving the new test sample from the target distribution. The choice of this unsupervised objective, which we call TTA loss,

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

governs how successful the adaptation process is. [39] uses a self-supervised objective on the test sample, [42] uses the entropy of model predictions, and several follow-ups have proposed variants or alternatives [31, 46]. However, it remains unclear as to how to choose or guide the selection of this TTA loss, and thus far the choice of these losses has remained largely heuristic in nature.

In this work, we begin by presenting a set of intriguing experiments where we attempt to *learn* the "best" TTA loss for a given source network and distribution shift. We parameterize the TTA loss by another neural network whose parameters are learnt via meta-learning [3, 7] where we differentiate through the adaptation process to find the TTA loss that achieves the best adaptation on distribution shifts. Surprisingly, we ultimately learn a TTA loss that looks *remarkably* similar to (a temperature-scaled version of) the softmax-entropy loss, which was already proposed by [42, 46]. Why did we recover the commonly used softmax-entropy loss despite the fact that the procedure is capable of learning a very general class of losses and the meta-learning process could potentially *specialize* to the distribution shift of interest? Furthermore, we find that this pattern only holds when the loss used to train the source classifier is the cross-entropy loss; when a different loss such as squared loss is used instead, the meta-learning procedure recovers a TTA loss that itself looks more like a negative squared error, and is very different from the softmax-entropy loss.

In order to explain this phenomenon, we propose to consider TTA through the lens of the *convex conjugate* function. Specifically, given a hypothesis function $h(x)$ and label y, several common losses (cross-entropy and the squared loss amongst them, but not limited to these) can be written in the form $\mathcal{L}(h(x), y) = f(h(x)) - y^T h(x)$ for some function f. In these cases, we show that "natural" TTA loss for such classifiers is precisely the (negation of) the convex conjugate evaluated at the gradient of h, $\mathcal{L}_{\text{TTA}}(x) = -f^*(\nabla f(h(x))$, where f^* is the convex conjugate of f. This framework not only explains the results of our meta-learning experiments, but also justifies why some specific choices of TTA loss in the previous literature work well (e.g., this framework recovers TENT's choice of softmax-entropy for cross-entropy-trained classifier). Moreover, it also provides a broad framework for what the TTA loss should be when the source model is trained using various different loss functions (for example the recently-proposed PolyLoss [19, 23]) as is becoming increasingly common in machine learning. Further, we show that our proposed conjugate adaptation loss is in fact a kind of self-training with pseudo-labels [35], a classic approach in machine learning. Various formulations of the pseudo-label have been proposed in the literature, and our conjugate analysis provides a general recipe for the "correct" choice of soft pseudo-labels given by $\hat{y}(x) = \nabla f(h(x))$. We thus refer to these as *conjugate pseudo-labels* (Conjugate PL's), and believe our work provides a broad framework for understanding adaptation with unlabeled data in general.

Finally, we empirically verify the effectiveness of our proposed conjugate adaptation loss across several datasets and training losses, such as cross-entropy and squared loss, along with the recently-proposed PolyLoss [19] (which itself has shown higher standard test accuracy on a wide range of vision tasks). Over *all* models, datasets and training losses, we find our proposed conjugate pseudo-labeling consistently outperforms prior TTA losses and improves TTA performance over the current state of the art.

2 Background and preliminaries.

Test-time adaptation. We are interested in mapping an input $x \in \mathbb{R}^d$ to a label $y \in \mathcal{Y}$. We learn a model $h_\theta : \mathbb{R}^d \mapsto \mathbb{R}^{|\mathcal{Y}|}$ parameterized by θ that maps an input x to predictions $h_\theta(x)$. We assume access to a trained source model and adapt at test-time over the test input, before making the final prediction. This is the standard test-time adaptation (TTA) setting [39, 42]. During TTA, we update the model parameters on an unsupervised objective $\mathcal{L}(x, h_\theta)$. For example, in TENT [42], this loss is the entropy of the softmax-normalized predictions of the model. At each time step of adaptation, we observe a batch of test inputs and we take a gradient step towards optimizing the TTA loss on this test batch. As is standard, we measure the average online performance of models across all steps (number of test batch inputs seen) in the adaptation process.

Meta learning the loss function. In order to explore the existence of different TTA losses, we employ the meta-learning procedure where we attempt to *learn* the TTA loss. We use a similar procedure as prior work on meta-learning loss functions [3, 29] and parameterize the loss function via a neural network $m_\phi : \mathbb{R}^{|\mathcal{Y}|} \mapsto \mathbb{R}$ that takes in the model predictions/logits and outputs a loss value. We want to learn parameter ϕ such that when we update θ via the loss function m_ϕ, our final performance is optimal. In order to do so, let x be the unlabeled test samples to adapt to, and y be the

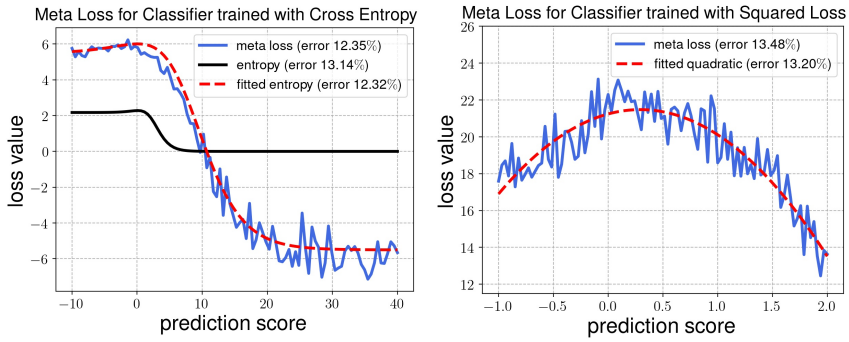
corresponding labels. We update θ and ϕ alternatively as follows.

$$\theta^{t+1} \leftarrow \theta^t - \alpha \frac{\partial m_{\phi^t}(h_{\theta^t}(x))}{\partial \theta^t}, \quad \phi^{t+1} \leftarrow \phi^t - \beta \frac{\partial \mathcal{L}(h_{\theta^{t+1}}(x'), y')}{\partial \phi^t}, \tag{1}$$

where $\mathcal{L}$ is some supervised surrogate loss function such as cross-entropy. Please refer to Appendix A3 for further details regarding meta-learning setup. Note that the meta-learning process above assumes access to labels y of test inputs. In this paper, we do *not* propose meta-learning the TTA loss as an approach. Rather, we use meta-learning to explore what the "best" TTA losses look like. We discuss our findings from this exploration in the next section.

3 Test-time Adaptation via Meta-Learnt Losses

The objective used in TENT is the softmax-entropy of the model predictions which essentially makes the classifier more confident in its current predictions. The same can be achieved by various other loss formulations such as those mentioned in [31]. With so many possible choices for the loss function, what should we use for TTA? In this section, we attempt to answer this empirically and present some intriguing observations.



(a) (b)

Figure 1: Visualization of meta loss (blue) by varying one input prediction score. (a) For cross-entropy loss trained model, the learnt meta loss can be approximated with a scaled softmax-entropy function (dashed red). (b) When the source model is trained with a squared loss for classification, the learnt meta loss (blue) can be fitted closely with a quadratic function (dashed red), shown in Figure 11a.

Experiment 1. We learn the TTA loss parameterized by a neural network via meta-learning as described in Section 2. Our source classifier is a ResNet-26 trained on CIFAR-10 and we adapt to distribution shifts in CIFAR-10-C. We use the 4 labeled validation noises in CIFAR-10-C to learn the meta-loss network parameters and we denote the resulting learnt loss function by meta-TTA loss. We then adapt the source classifier to the test set of 15 corruptions by optimizing the meta-TTA loss.

Observations. First, we find that TTA using meta-TTA loss performs better than TENT (12.35% vs 13.14%), suggesting that there are better TTA losses than previous losses based on softmax-entropy.

However, on examining this meta-TTA loss, we find a surprising observation. Figure 11b (blue curve) visualizes the learnt meta-loss over model predictions as we vary a single class prediction with the rest fixed. Qualitatively, the learnt meta-loss looks very similar to softmax-entropy in one dimension. In fact, we can fit it closely with a scaled softmax-entropy function (dashed red curve): $\alpha \cdot \mathcal{H}(\text{softmax}(h_\theta(x)/T))$, where α is a magnitude parameter and T is a temperature scaler. We want to test if the meta-loss is basically learning the softmax-entropy function. Hence, we perform test-time adaptation with the fitted softmax-entropy function instead (dashed red curve) and achieve an error of 12.32%, essentially recovering the performance of meta-TTA.

Despite the ability to represent many different loss functions and potentially specialize to the CIFAR-10-C setting, the meta-loss procedure gave back the standard entropy objective. *Do we always recover a loss that looks like softmax-entropy?*

Experiment 2. In an attempt to isolate when we get back the entropy objective, we vary several things. We tried different architectures for the source classifier, different losses $\mathcal{L}$ during the meta-learning process (1) and different training losses for the source classifier.

Results. We observed that we consistently recovered the temperature scaled softmax-entropy function in all cases *except* when we varied the training loss for the source classifier. On using the squared loss function [15], a strikingly different meta-TTA loss emerges. Figure 11a (blue curve) shows the learnt meta-loss (13.48% error) for this network. Here again, the meta-TTA loss outperforms entropy (14.57%) but it is not simply due to a scaling factor. The loss now looks like the negative squared error (red curve). Like previously, we tried fitting a quadratic loss directly to the meta loss in Figure 11a, and this time we even slightly outperformed the meta-TTA loss.

To summarize, we used a meta-learning procedure to search for the "best" TTA loss, where the loss itself was parameterized by a neural network that could potentially represent arbitrarily complex loss functions. However, we ended up with loss functions displaying remarkable structure: across different architectures and different variants of meta-learning, for a classifier trained with cross-entropy, the meta-TTA loss was temperature scaled softmax-entropy and for a classifier trained with squared loss, the meta-TTA loss was a negative squared loss. This is interesting from both a practical and conceptual standpoint where the "best" TTA loss depends on the loss used to train the source classifier in a clean fashion. We attempt to understand and explain this phenomenon in the next section.

4 Conjugate Pseudo Labels

Results in the previous section raise an obvious question: *why* does softmax-entropy as used in TENT seem to be the "best" possible test time adaptation loss for classifiers trained via cross-entropy (at least, best in the sense that meta-learning consistently recovers something which essentially mimics softmax-entropy, even though meta-loss is parameterized by a neural network and hence could learn much more complex functions specific to the model and the particular shift)? And why, alternatively, does a quadratic TTA loss seem to perform best when the classifier is trained via squared loss?

In this section, we offer an explanation of this phenomenon via the construct of the convex conjugate function [1]. As we will see, our method recovers softmax-entropy *and* quadratic loss as the "natural" objectives for classifiers trained via cross-entropy and squared loss respectively. Furthermore, for classifiers trained via *other* loss functions, as is becoming increasingly common in deep learning, our approach naturally suggests corresponding test-time adaptation losses, which we show in the next section to comparatively outperform alternatives. Thus, we argue that our framework overall provides a compelling recipe for specifying the "correct" method for TTA for a large class of possible losses.

4.1 Losses and the convex conjugate

We begin by formally considering loss functions between a hypothesis output $h_\theta(x)$ (e.g., the logit outputs of a classifier, or the direct prediction of a regressor) and target y that take the following form

$$\mathcal{L}(h_\theta(x), y) = f(h_\theta(x)) - y^T h_\theta(x) \tag{2}$$

for some function f; when there is no risk of confusion, we will use h in place of $h_\theta(x)$ for simplicity of notation. While not every loss can be expressed in such a form, this captures a wide variety of common losses (possibly scaled by a constant value). For example, cross-entropy loss corresponds to the choice $f(h) = \log \sum_i \exp(h_i)$ and where y denotes a one-hot encoding of the class label; similarly, squared loss corresponds to the choice $f(h) = \frac{1}{2}\|h\|_2^2$.

When training an over-parameterized classifier, we can roughly view the training process as (approximately) attaining the *minimum* over hypotheses h for each training example

$$\min_\theta \frac{1}{t} \sum_{i=1}^{t} \mathcal{L}(h_\theta(x_i), y_i) \approx \frac{1}{t} \sum_{i=1}^{t} \min_h \mathcal{L}(h, y_i) \tag{3}$$

where t is the number of training samples. However, in the case of losses in the form (2), the minimization over h in this form represents a very specific and well-known optimization problem: it is known as the *convex conjugate* [1] of the function f

$$\min_h \mathcal{L}(h, y) = \min_h \{f(h) - y^T h\} = -f^\star(y) \tag{4}$$

where $f^\star$ denotes the convex conjugate of f. $f^\star$ is a convex function in y (and indeed, is convex regardless of whether or not f is convex). Furthermore, for the case that f is convex differentiable, the optimality condition of this minimization problem is given by $\nabla f(h^{\text{opt}}) = y$, so we also have that

$$f^\star(y) = f^\star(\nabla f(h^{\text{opt}})) \tag{5}$$

where h^{opt} refers to the optimal classifier (used interchangeably with $h_{\theta^{\text{opt}}}$). Putting this all together, we can state (admittedly, in a rather informal manner) that under the assumption that θ^{opt} is chosen so as to approximately minimize the empirical loss on the source data in the over-parameterized setting, we have that for t inputs

$$\frac{1}{t}\sum_{i=1}^{t}\mathcal{L}(h_{\theta^{\text{opt}}}(x_i), y_i) \approx \frac{1}{t}\sum_{i=1}^{t} -f^{\star}(\nabla f(h_{\theta^{\text{opt}}}(x_i))) \tag{6}$$

i.e., the empirical loss can be approximated by the (negative) conjugate applied to the gradient of the f, at least in a region close to the optimal θ^{opt} that minimizes the empirical loss. But the later expression has the notable benefit that it does not require any label y_i in order to compute the loss, and thus can be used as a basis for TTA on target domain of the hypothesis function $h_{\theta^{\text{opt}}}$.

Definition 1 (conjugate adaptation loss) *Consider a loss function that takes the form given in 2, used for training a hypothesis h_θ in the over-parameterized regime. We define the conjugate adaptation loss $\mathcal{L}^{conj}(h_\theta(x)) : \mathbb{R}^{|\mathcal{Y}|} \mapsto \mathbb{R}$ as follows.*

$$\mathcal{L}^{conj}(h_\theta(x)) = -f^{\star}(\nabla f(h_\theta(x))) = f(h_\theta(x)) - \nabla f(h_\theta(x))^{\top} h_\theta(x). \tag{7}$$

4.2 Recovery of existing test-time adaptation strategies

Cross-entropy The interesting aspect to this formalism is that when applied to classifiers trained with cross-entropy, it recovers exactly the TENT approach to TTA : minimizing the softmax-entropy of $h_\theta(x)$. And indeed, this loss was *also* recovered when using meta-learning to learn the "optimal" test-time adaptation loss. To see this, note that for cross-entropy, we have that $f(h) = \log \sum_i \exp(h_i)$, giving the optimality condition $y = \nabla f(h^{\text{opt}}) = \frac{\exp(h^{\text{opt}})}{\sum_i \exp(h_i^{\text{opt}})}$ and the conjugate function

$$f^{\star}(y) = \begin{cases} \sum_i y_i \log y_i & \text{if } \sum_i y_i = 1 \\ \infty & \text{otherwise} \end{cases}. \tag{8}$$

In other words,

$$\mathcal{L}^{\text{conj}}(h_\theta(x)) = -f^{\star}(\nabla f(h_\theta(x))) = -\sum_i \frac{\exp(h_i)}{\sum_j \exp(h_j)} \log \frac{\exp(h_i)}{\sum_j \exp(h_j)} \tag{9}$$

i.e. softmax-entropy of the model prediction, which is exactly the TTA loss that TENT uses.

Squared loss For the squared loss, we have that $f(h) = \frac{1}{2}\|h\|_2^2$, leading to the optimality condition $y = h$ and conjugate function $f^{\star}(y) = \frac{1}{2}\|y\|_2^2$. Hence, the adaptation loss in this case would be simply given by $\mathcal{L}^{\text{conj}}(h_\theta(x)) = -f^{\star}(\nabla f(h_\theta(x))) = -\frac{1}{2}\|h\|_2^2$ which is also what we observed in the meta-learning experiments discussed in section 3.

4.3 Conjugate pseudo-labels

We now emphasize that by the nature of our approximations, there is an additional simple interpretation of the conjugate loss: it is also equal to the original loss (2) applied to the "psuedo-labels" $y_\theta^{\text{CPL}}(x) = \nabla f(h_\theta(x))$, where CPL refers to conjugate pseudo-labels, i.e.,

$$\mathcal{L}^{\text{conj}}(h_\theta(x)) = -f^{\star}(\nabla f(h_\theta(x))) = f(h_\theta(x)) - \nabla f(h_\theta(x))^{T} h_\theta(x) = \mathcal{L}(h_\theta(x), \nabla f(h_\theta(x))). \tag{10}$$

This property is known as the Fenchel-Young inequality, that is $f(x) + f^{\star}(u) \geq x^T u$ holding with equality when $u = \nabla f(x)$. In other words, our conjugate adaptation loss is precisely equivalent to self-training under the specific soft pseudo-labels given by $y = \nabla f(h_\theta(x))$. And indeed, for many cases, this may be a more convenient form to compute than explicitly computing the conjugate function at all. For this reason, we refer to our method as that of *conjugate pseudo-labels*.

In the case of cross-entropy loss, this approach then corresponds exactly to self-training using labels given by the softmax applied to the current hypothesis. We must emphasize, however, that while our conjugate formulation indeed has this "simple" form for the case of cross-entropy loss, the real advantage comes in that it provides the "correct" pseudo-label for use with *other* losses, which may result in pseudo-labels different from the "common" softmax operation.

Example: conjugate pseudo-labels for PolyLoss. PolyLoss [19] is a recently-proposed simple alternative to cross-entropy loss than has been shown to improve performance across a wide variety of compute tasks. This loss is given by the form

$$\mathcal{L}_{\text{poly}}(h_\theta(x), y) = \mathcal{L}_{\text{ce}}(h_\theta(x), y) + \epsilon \cdot y^T(1 - \text{softmax}(h_\theta(x))) \tag{11}$$

We note that this can be put exactly into our conjugate form (equation 2) by writing the loss in a slightly more involved fashion, which we refer to as the *expanded conjugate form*

$$\mathcal{L}_{\text{poly}}(h_\theta(x), y) = f(h_\theta(x)) - y^T g(h_\theta(x)). \tag{12}$$

where f is the log-sum-exp function as before, and $g(h) = h - \epsilon(1 - \text{softmax}(h))$. In order to formally put this into the form of the previous loss function (equation 2), we can simply define an alternative hypothesis as the function $h'_\theta(x) = g(h_\theta(x))$, and then define PolyLoss in the conjugate form as

$$\mathcal{L}_{\text{poly}}(h'_\theta(x), y) = f(g^{-1}(h'_\theta(x))) - y^T h'_\theta(x). \tag{13}$$

Typically, however, it is easier to simply operate on the expanded conjugate form, which yields the optimality condition for the pseudo-label $\nabla f(h^{\text{opt}}) = \mathsf{D}g(h^{\text{opt}})y_\theta^{\text{CPL}}(x)$, where D is the Jacobian operator. For the case of PolyLoss, this leads to the conjugate pseudo-label of the following form: $y_\theta^{\text{CPL}}(x) = (I + \epsilon\text{diag}(z) - \epsilon zz^T)^{-1}z, \quad z \equiv \text{softmax}(h_\theta(x))$.

Test-time adaptation. Finally, we note that the above discussion doesn't actually address any topics related to test-time adaptation to OOD data, but merely provides a generic characterization of a self-training procedure for generic loss functions of the form (2). However, the application to TTA on OOD data is fairly straightforward: as long as the learnt source parameters θ is a reasonable approximation to the true optimal θ^{opt} on the shifted domain, self-training with the conjugate pseudo-labels provides a reasonable proxy for fine-tuning the network on the true OOD loss. We emphasize that, common to most approaches for TTA , there are still some amount of design decisions that must be put in place; these are detailed in Section 5.1. In practice, we observe OOD generalization typically benefits (across all baselines) from an additional "temperature" scaling, i.e., applying the TTA loss to $h_\theta(x)/T$ for some fixed temperature T, although it requires a held-out validation dataset for tuning T. However, we should emphasize that truly unsupervised TTA would require making an informed guess for the value of these hyper-parameters. The full procedure for test time adaptation via conjugate pseudo-labels is shown in Algorithm 1.

Algorithm 1 Conjugate pseudo-labeling (Conjugate PL)

Input: Source classifier θ_0 trained using loss $\mathcal{L}(h_\theta(x), y) = f(h_\theta(x)) - h_\theta(x)^\top y$.
N batches of test data $D_{\text{test}} = [x_1, x_2, \ldots, x_N]$
Hyperparams: learning rate η and temperature T.

Let $\bar{h}_\theta(x) \stackrel{\text{def}}{=} h_\theta(x)/T$ be the temperature scaled predictor.

Let $y_\theta^{\text{CPL}}(x)$ denote the conjugate pseudo-label function $y_\theta^{\text{CPL}}(x) = \nabla(f(\bar{h}_\theta(x)))$.

for $n = 0, 1, \ldots N-1$ **do**

$\theta_{n+1} = \theta_n - \eta\nabla\mathcal{L}\Big(\bar{h}_\theta(x_n), y_\theta^{\text{CPL}}(x_n)\Big)$ [Self-training with conjugate pseudo-labels]

5 Experiments

In this section, we empirically evaluate the effectiveness and generality of the proposed conjugate pseudo-labeling procedure (Algorithm 1) for test-time adaptation on a variety of datasets.

5.1 Setup

Datasets We evaluate on the three common corruption benchmarks: adapting a classifier trained on CIFAR-10 to CIFAR-10-C, CIFAR-100 to CIFAR-100-C and ImageNet to ImageNet-C [11]. Following the previous works [39, 42], we report the error averaged across corruptions at the highest severity for CIFAR-10/100-C and averaged across corruptions and severity level for ImageNet-C. We also evaluate on three domain adaptation datasets: adapting a classifier trained on SVHN to MNIST, an ImageNet classifier to ImageNet-R [12] and adapting from synthetic to real data in VISDA-C [30].

Models and Training Losses Following previous works on TTA[39, 42], we use ResNet-26 [10] as the source classifier architecture for CIFAR-10/100 experiments, ResNet-18 for SVHN to MNIST and a ResNet-50 for ImageNet and source synthetic data on VisDA-C. We consider source classifiers trained via the following loss functions: the de-facto cross-entropy, recently proposed polyloss [19] and squared loss [15].

Baselines. Our proposed conjugate pseudo-label is the classic approach of self-training with a specific form of pseudo-labels. In self-training, we replace the label y with a pseudo-label $\tilde{y}(x)$ and adapt by optimizing the loss function $\mathcal{L}(h_\theta(x), \tilde{y}(x))$. Note that we could either instantaneously update the pseudo-labels using the current classifier, or generate pseudo-labels once with just the source classifier. Instantaneous updates have been shown to work better for domain adaptation [6, 31], and we perform instantaneous updates for all methods.

While we propose using $\tilde{y}(x) = \nabla f(h_\theta(x))$ (See Section 4.3), we compare to the standard pseudo-labels used in the literature:

- (i) the "hard" pseudo-label (hard PL) where $\tilde{y}(x) = \arg\max_i \left(h_\theta(x)\right)_i$ is the most likely class as predicted by h_θ. As is common in the self-training literature, we perform confidence thresholding.
- (ii) The "soft" pseudo-label (soft PL) where $\tilde{y}(x)$ is obtained by applying a softmax function to the model predictions $h_\theta(x)$.

We also compare with the following recently proposed test-time adaptation methods.

- Entropy Minimization (ENT) [42] minimizes the entropy of model predictions.
- Robust Pseudo-Label [31] where we minimize a robust classification loss, $\mathcal{L}_{\text{rpl}} = q^{-1}(1 - p(i|x)^q)$ where $i = \text{argmax}_j p(j|x)$ and $q \in [0, 1]$.
- MEMO [46] minimizes entropy of a model's outputs across different augmentations of a test input. We implement a *batch version*, where we see multiple test points at once, for fair comparisons.

TTA methodology Following [42] and [32], we fine-tune the learnable scale and shift parameters of the batch normalization layers across all adaptation losses. For each batch, batch normalization statistics is also updated, as suggested in [34]. We report performance at the end of one round of test-time adaptation over the entire test set.

We tune the learning rate (LR) and temperature (T) on the validation noises in the corruption benchmark by grid-search. LR is selected from $\{1e^{-1}, 1e^{-2}, \ldots 1e^{-4}\}$ and T from $\{1, 2 \ldots 5\}$. All the experiments have been performed on A6000 GPU's. On domain adaptation benchmarks, where there is no held-out target domain, we set T to be 1 and use the LR suggested by [5, 42]. We use the same hyperparameter tuning protocol across all methods. We single out temperature as a very important hyperparameter, as we discuss in the results below.

Dataset	Temperature (T)	Hard PL	Robust PL	MEMO	Conjugate PL (**ENT**)
CIFAR-10-C	✗	13.95 (±0.06)	13.97 (±0.04)	**12.60** (±0.04)	13.07 (±0.05)
	✓	13.95 (±0.06)	12.85 (±0.04)	**12.51** (±0.01)	**12.51** (±0.03)
CIFAR-100-C	✗	45.22 (±0.4)	39.80 (±0.18)	**38.52** (±0.16)	41.15 (±0.25)
	✓	45.22 (±0.4)	36.37 (±0.10)	37.38 (±0.06)	**36.10** (±0.07)
ImageNet-C	✗	**45.43** (±0.05)	45.68 (±0.01)	48.91(±0.03)	45.82(±0.01)
	✓	45.43 (±0.05)	45.61 (±0.01)	48.91(±0.04)	**45.36**(±0.01)

Table 1: Mean errors when adapting to corruptions using a source classifier trained via cross-entropy loss. Here, conjugate pseudo-labeling becomes softmax-entropy minimization. With the right temperature scaling, softmax-entropy minimization matches or outperforms other approaches. Prior reported gains of other methods over softmax-entropy minimization disappear when we use temperature scaling. For additional context, the source classifier errors without adaptation are: CIFAR-10-C (29.54%), CIFAR-100-C (62.26%), ImageNet-C (61.89%)

5.2 Results on classifiers trained with cross-entropy

We study the effectiveness of our proposed conjugate pseudo-labels when the source classifier is trained via cross-entropy loss. In this case, baselines Softmax PL and ENT are the same as Conjugate PL. Thus we omit them in our results. Table 1, reports the performance of various TTA methods.

When the source classifier is trained via cross-entropy, our conjugate pseudo-label algorithm exactly corresponds to entropy minimization with an additional temperature scaling. Entropy minimization as

proposed in prior work [42] does not tune the temperature parameter, and some newer objectives such as robust PL or MEMO outperform vanilla entropy minimization. For example, on CIFAR-100-C, vanilla ENT obtaines 41.15% average error, while robust PL improves this to 39.80% and MEMO to 38.52%. However, with the right temperature scaling, entropy minimization obtains 36.10% error which outperforms the newer objectives (with and without temperature scaling). A similar observation holds for CIFAR-10-C and ImageNet-C as well. Essentially, the gains over vanilla entropy minimization vanish when we do temperature scaling, and entropy minimization (i.e. conjugate pseudo-labeling corresponding to cross-entropy) turns out to be the best objective after all.

5.3 Results on classifiers trained with polyloss and squared loss

In the case of cross-entropy, conjugate pseudo-labeling reduces to the familiar notion of entropy minimization. We now explore the performance of our method on different loss functions where the conjugate pseudo-labels differ substantially from entropy minimization (section 4.3, section 4.2). Table 2 presents the results on the corruption benchmarks and Table 3 presents the results on the other domain adaptation datasets for source classifiers trained with PolyLoss.

Dataset	T	Hard PL	Robust PL	ENT	MEMO	Softmax PL	Conjugate PL (**Ours**)
CIFAR-10-C	✗	13.81(±0.12)	14.23(±0.02)	13.46(±0.06)	13.23(±0.07)	14.64(±0.11)	**13.02**(±0.09)
	✓	13.81(±0.12)	12.45(±0.05)	12.23(±0.06)	12.33(±0.04)	12.26(±0.04)	**12.08**(±0.05)
CIFAR-100-C	✗	40.47(±0.05)	42.86(±0.11)	40.12(±0.08)	39.90(±0.05)	41.00(±0.11)	**38.17**(±0.17)
	✓	40.47(±0.05)	39.80(±0.08)	38.23(±0.05)	39.23(±0.04)	37.04(±0.06)	**36.83**(±0.08)
ImageNet-C	✗	45.44(±0.21)	46.27(±0.03)	46.10(±0.03)	48.21(±0.05)	44.63(±0.03)	**44.01**(±0.01)
	✓	45.44(±0.21)	46.27(±0.03)	45.50(±0.02)	48.21(±0.04)	44.45(±0.03)	**44.01**(±0.01)

Table 2: Mean errors when adapting to corruptions using a source classifier trained via recently proposed Poly-1 Loss [19]. Conjugate pseudo-labeling consistently outperforms all previous approaches. For additional context, source classifier errors without adaptation : CIFAR-10-C (30.22%), CIFAR-100-C (63.91%) and ImageNet-C (62.18%).

Dataset	Source Error	Hard PL	Robust PL	Entropy	Softmax PL	Conjugate PL **Ours**
SVHN → MNIST	28.33	20.21	19.73	14.28	16.54	**10.73**
SVHN → USPS	31.58	23.32	26.12	23.12	24.07	**21.62**
SVHN → MNISTM	61.69	50.73	51.35	49.33	50.47	**47.59**
ImageNet-R	64.19	58.52	59.46	58.25	56.62	**55.63**
VisDA-C	58.13	40.43	45.44	44.11	39.63	**38.42**

Table 3: Target error when adapting models trained via polyloss on source domains across different domain adaptation benchmarks. Conjugate pseudo-labeling offers consistent and substantial gains over previous approaches across three datasets.

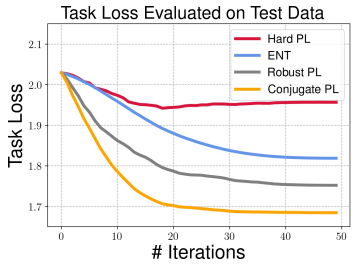


Figure 2: Task Loss (PolyLoss $\epsilon = 6$) evaluated on CIFAR-10-C test data during test-time adaptation.

First, we note that, across all datasets in Table 2 and Table 3, our conjugate PL approach outperforms all other TTA losses. With polyloss classifiers, entropy minimization is no longer the best method—on CIFAR-100-C, entropy minimization achieves 38.23% error while our conjugate PL achieves 36.83%. We see similar *consistent gains* on CIFAR-10-C, ImageNet-C, ImageNet-R and VisDA-C. On digit adaptation tasks from SVHN to MNIST/USPS/MNISTM, where there is a larger shift between source and target, the gains are especially pronounced. Figure 2 compares how the task loss (polyloss $\epsilon = 6$) on the test data decreases as we adapt the model through conjugate PL and other baselines. We use CIFAR-10-C as an example. Observe that our proposed conjugate PL indeed reduces the task loss the most among other baselines.

Furthermore, on CIFAR-10-C and ImageNet-C, we find that adapting polyloss classifiers via conjugate PL improves the performance over all methods applied to cross-entropy trained source classifiers. For e.g., on ImageNet-C, the performance improves from 45.34% to 44.01%. However, this is only true when using the proposed conjugate PL. If we just did softmax-entropy minimization (even with temperature scaling), the final adapted performance of a polyloss classifier (45.5%) is in fact worse than that of a cross-entropy classifier (45.34%). Our results suggest that as we develop new training

losses that improve the source classifiers, it is important to adapt via conjugate pseudo-labeling to reap the maximum gains.

Similarly, we experiment with the case when the source classifier is trained using squared loss on the CIFAR-10 and CIFAR-100 datasets, and observe consistent gains using the proposed conjugate pseudo-labels over the baselines. For example, on CIFAR-10-C, TTA using conjugate PL gives and error of 12.87%, outperforming baselines like ENT (13.24%) and Softmax PL (31.81%). Table 5 in Appendix A.7 shows the detailed results.

Comparing Table 1 and Table 2, we see that the relative ordering between the various baselines differs. This is further evidence that the adaptation loss has to depend on the training loss, and we believe our conjugate pseudo-label approach captures this appropriately by offering consistent gains across the various settings we experimented with.

6 Related Works

Source-Free Domain Adaptation These approaches [18, 20, 22] in general consider an apriori access to unlabeled data from the target domain. SHOT [22] proposes information maximization on the unlabeled data while [45] promotes feature clustering. [18, 20] use generative modeling to estimate the underlying source distributions for enforcing feature invariance. These approaches require offline optimization on unlabeled target dataset with carefully designed objectives and custom training pipelines. Test-time adaptation (TTA) adapts the model at test-time in an online manner based on its own predictions. Although TTA hinges on the optimization setup [31, 34, 42], previous works are mostly focused on proposing new *unsupervised* objectives. [39] optimizes over the task loss from a self-supervised learning problem, but it alters the training of source classifier. Closest to our setting, is TENT ([42]), which proposes to minimize the softmax-entropy of model predictions at test time. Several follow ups like [5, 28, 31, 36, 46] propose alternative TTA objectives, e.g. robust pseudo-labelling [31], augmentations [28, 46] and contrastive learning [5, 41]. However, most of these are based on heuristically chosen TTA objectives. In this paper, we provide a principled way of designing unsupervised objectives for TTA .

Pseudo-Labels and Self Training [6, 17, 43] provide theoretical insights into how self-training with pseudo-labels can help under distribution shift. Our work provides a general framework for the choice of soft pseudo-labels based on the conjugate analysis of the source training objective. Some prior works like [6, 13, 21, 25, 47, 48] have documented the improvement in performance when using instantaneous pseudo-labels over pre-computed pseudo-labels, and thus lend further support to the benefits of our proposed conjugate pseudo-labeling approach.

Distributionally Robust Training For completeness, we briefly discuss different robust training methods. These approaches towards adaptation alter the source training under various set of assumptions in order to learn more robust and transferable representations. [14, 33, 44] consider risk minimization under various formulations of worst-case loss. [2] proposes a new objective for learning invariant correlations. [4, 24] propose reweighing the training samples. A host of approaches like [26, 27, 37, 40] try to achieve feature alignment by matching distribution statistics, notably maximum mean discrepancy. Similarly, [8] tries to align feature representations in a common space using discriminator based approach whereas [38] tries to achieve the same using self-supervised learning. However, retraining with source data might not be feasible due to data privacy and efficiency issues.

7 Conclusion and Limitations

In this paper, we proposed a general method of conjugate pseudo-labeling that derives an appropriate TTA loss for a given source classifier. Across a variety of different training losses and distribution shifts, our method offered consistent gains over alternatives. This conjugate formalism was inspired by an intriguing set of meta-learning experiments that suggested that these conjugate pseudo-labels are somehow the "best" adaptation loss. However, the role of several heuristics like updating only the batch normalization layers and temperature scaling which seems to help all the baselines too, is not well understood. An immediate extension of our work includes broadening the class of loss functions considered beyond equation 2. Similarly, the meta-learning setup can be extended to include intermediate representations and consider learning functions over a batch of input while accounting for their interactions. The experimental investigation in this work focused on adapting in the presence of distribution shift, but the notion of conjugate pseudo-labels is more general, and is worth investigating even in the standard semi-supervised learning setting.

References

[1] https://en.wikipedia.org/wiki/Convex_conjugate.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019. URL https://arxiv.org/abs/1907.02893.

[3] Sarah Bechtle, Artem Molchanov, Yevgen Chebotar, Edward Grefenstette, Ludovic Righetti, Gaurav Sukhatme, and Franziska Meier. Meta-learning via learned loss. *arXiv preprint arXiv:1906.05374*, 2019.

[4] Jonathon Byrd and Zachary C. Lipton. What is the effect of importance weighting in deep learning?, 2018. URL https://arxiv.org/abs/1812.03372.

[5] Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation, 2022. URL https://arxiv.org/abs/2204.10377.

[6] Yining Chen, Colin Wei, Ananya Kumar, and Tengyu Ma. Self-training avoids using spurious features under domain shift, 2020. URL https://arxiv.org/abs/2006.10032.

[7] Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks, 2017. URL https://arxiv.org/abs/1703.03400.

[8] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. 2015. doi: 10.48550/ARXIV.1505.07818. URL https://arxiv.org/abs/1505.07818.

[9] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization, 2020. URL https://arxiv.org/abs/2007.01434.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://arxiv.org/abs/1512.03385.

[11] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations, 2019. URL https://arxiv.org/abs/1903.12261.

[12] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *ICCV*, 2021.

[13] Yosuke Higuchi, Niko Moritz, Jonathan Le Roux, and Takaaki Hori. Advancing momentum pseudo-labeling with conformer and initialization strategy, 2021. URL https://arxiv.org/abs/2110.04948.

[14] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers?, 2016. URL https://arxiv.org/abs/1611.02041.

[15] Like Hui and Mikhail Belkin. {EVALUATION} {of} {neural} {architectures} {trained} {with} {square} {loss} {vs} {cross}-{entropy} {in} {classification} {tasks}. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=hsFN92eQEla.

[16] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. Wilds: A benchmark of in-the-wild distribution shifts, 2020. URL https://arxiv.org/abs/2012.07421.

[17] Ananya Kumar, Tengyu Ma, and Percy Liang. Understanding self-training for gradual domain adaptation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5468–5479. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/kumar20c.html.

[18] Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 615–625, 2021. doi: 10.1109/WACV48630.2021.00066.

[19] Zhaoqi Leng, Mingxing Tan, Chenxi Liu, Ekin Dogus Cubuk, Jay Shi, Shuyang Cheng, and Dragomir Anguelov. Polyloss: A polynomial expansion perspective of classification loss functions. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=gSdSJoenupI.

[20] Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9638–9647, 2020. doi: 10.1109/CVPR42600.2020.00966.

[21] Xinzhe Li, Qianru Sun, Yaoyao Liu, Qin Zhou, Shibao Zheng, Tat-Seng Chua, and Bernt Schiele. Learning to self-train for semi-supervised few-shot classification. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/bf25356fd2a6e038f1a3a59c26687e80-Paper.pdf.

[22] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation, 2020. URL https://arxiv.org/abs/2002.08546.

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection, 2017. URL https://arxiv.org/abs/1708.02002.

[24] Evan Zheran Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information, 2021. URL https://arxiv.org/abs/2107.09044.

[25] Hong Liu, Jianmin Wang, and Mingsheng Long. Cycle self-training for domain adaptation, 2021. URL https://arxiv.org/abs/2103.03571.

[26] Mingsheng Long, Jianmin Wang, Guiguang Ding, Jiaguang Sun, and Philip S. Yu. Transfer feature learning with joint distribution adaptation. In *2013 IEEE International Conference on Computer Vision*, pages 2200–2207, 2013. doi: 10.1109/ICCV.2013.274.

[27] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I. Jordan. Learning transferable features with deep adaptation networks, 2015. URL https://arxiv.org/abs/1502.02791.

[28] Chaithanya Kumar Mummadi, Robin Hutmacher, Kilian Rambach, Evgeny Levinkov, Thomas Brox, and Jan Hendrik Metzen. Test-Time Adaptation to Distribution Shift by Confidence Maximization and Input Transformation. *arXiv e-prints*, art. arXiv:2106.14999, June 2021.

[29] Junhyuk Oh, Matteo Hessel, Wojciech M. Czarnecki, Zhongwen Xu, Hado P van Hasselt, Satinder Singh, and David Silver. Discovering reinforcement learning algorithms. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1060–1070. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/0b96d81f0494fde5428c7aea243c9157-Paper.pdf.

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

[31] Evgenia Rusak, Steffen Schneider, George Pachitariu, Luisa Eck, Peter Vincent Gehler, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. If your data distribution shifts, use self-learning, 2022. URL https://openreview.net/forum?id=1oEvY1a67c1.

[32] Evgenia Rusak, Steffen Schneider, George Pachitariu, Luisa Eck, Peter Vincent Gehler, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. If your data distribution shifts, use self-learning, 2022. URL https://openreview.net/forum?id=1oEvY1a67c1.

[33] Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

[34] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 11539–11551. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/85690f81aadc1749175c187784afc9ee-Paper.pdf.

[35] H. Scudder. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory*, 11(3):363–371, 1965. doi: 10.1109/TIT.1965.1053799.

[36] Prabhu Teja Sivaprasad and François Fleuret. Test time adaptation through perturbation robustness. *CoRR*, abs/2110.10232, 2021. URL https://arxiv.org/abs/2110.10232.

[37] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation, 2016. URL https://arxiv.org/abs/1612.01939.

[38] Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A. Efros. Unsupervised domain adaptation through self-supervision, 2019. URL https://arxiv.org/abs/1909.11825.

[39] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei A. Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts, 2019. URL https://arxiv.org/abs/1909.13231.

[40] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance, 2014. URL https://arxiv.org/abs/1412.3474.

[41] Dequan Wang, Shaoteng Liu, Sayna Ebrahimi, Evan Shelhamer, and Trevor Darrell. On-target adaptation. *CoRR*, abs/2109.01087, 2021. URL https://arxiv.org/abs/2109.01087.

[42] 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. URL https://openreview.net/forum?id=uXl3bZLkr3c.

[43] Sang Michael Xie, Ananya Kumar, Robbie Jones, Fereshte Khani, Tengyu Ma, and Percy Liang. In-n-out: Pre-training and self-training using auxiliary information for out-of-distribution robustness, 2020. URL https://arxiv.org/abs/2012.04550.

[44] Ziyu Xu, Chen Dan, Justin Khim, and Pradeep Ravikumar. Class-weighted classification: Trade-offs and robust approaches, 2020. URL https://arxiv.org/abs/2005.12914.

[45] Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation, 2021. URL https://arxiv.org/abs/2108.01614.

[46] Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation, 2021. URL https://arxiv.org/abs/2110.09506.

[47] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. Domain adaptation for semantic segmentation via class-balanced self-training, 2018. URL https://arxiv.org/abs/1810.07911.

[48] Yang Zou, Zhiding Yu, Xiaofeng Liu, B. V. K. Vijaya Kumar, and Jinsong Wang. Confidence regularized self-training, 2019. URL https://arxiv.org/abs/1908.09822.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We add the limitations of our work in the end. See Section 7
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 7. We do not foresee any specific negative societal impact of our work, apart from general guidelines valid for deep learning in general.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will be submitting the code along with the supplementary material. We will be adding a detailed readme for replicating the experiments.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] All the implementation details like the gridsearch and validation data have been mentioned in the paper, see section 5.1. Further, we will be adding exact hyper-parameters used for all the datasets in the appendix
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] All the results in the paper have the standard deviations reported, obtained by running the experiment 5 times with random seeds.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 5.1
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We will be adding the code to replicate experiments in this paper
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A] We use all the publicly available datasets
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]