International Journal of Computer Vision (2025) 133:31–64
https://doi.org/10.1007/s11263-024-02181-w

A Comprehensive Survey on Test-Time Adaptation Under Distribution Shifts

Jian Liang[1,2] · Ran He[1,2] · Tieniu Tan[1,2,3]

Received: 16 October 2023 / Accepted: 4 July 2024 / Published online: 18 July 2024

Abstract
Machine learning methods strive to acquire a robust model during the training process that can effectively generalize to test samples, even in the presence of distribution shifts. However, these methods often suffer from performance degradation due to unknown test distributions. Test-time adaptation (TTA), an emerging paradigm, has the potential to adapt a pre-trained model to unlabeled data during testing, before making predictions. Recent progress in this paradigm has highlighted the significant benefits of using unlabeled data to train self-adapted models prior to inference. In this survey, we categorize TTA into several distinct groups based on the form of test data, namely, test-time domain adaptation, test-time batch adaptation, and online test-time adaptation. For each category, we provide a comprehensive taxonomy of advanced algorithms and discuss various learning scenarios. Furthermore, we analyze relevant applications of TTA and discuss open challenges and promising areas for future research. For a comprehensive list of TTA methods, kindly refer to https://github.com/tim-learn/awesome-test-time-adaptation.

Keywords Transfer learning · Domain adaptation · Distribution shift · Source-free domain adaptation · Model adaptation · Test-time training · Test-time adaptation

1 Introduction

Traditional machine learning methods assume that the training and test data are drawn independently and identically (i.i.d.) from the same distribution (Quinonero-Candela et al., 2008). However, when the test distribution (target) differs from the training distribution (source), we face the problem of *distribution shifts*. Such a shift poses significant challenges for machine learning systems deployed in the wild, such as images captured by different cameras (Saenko et al., 2010), road scenes of different cities (Chen et al., 2017), and imaging devices in different hospitals (Liu & Yuan, 2022). As a result, the research community has developed a variety of generalization or adaptation techniques to improve model robustness against distribution shifts. For instance, *domain generalization* (DG) (Zhou et al., 2022) aims to learn a model using data from one or multiple source domains that can generalize well to any out-of-distribution target domain. On the other hand, *domain adaptation* (DA) (Kouw & Loog, 2019) follows the transductive learning principle to leverage knowledge from a labeled source domain to an unlabeled target domain.

This survey primarily focuses on the paradigm of *test-time adaptation* (TTA), which involves adapting a pre-trained model from the source domain to unlabeled data in the target domain before making predictions (Liang et al., 2020; Sun et al., 2020; Wang et al., 2021). While DG operates solely during the training phase, TTA has the advantage of being able to access test data from the target domain during the test phase. This enables TTA to enhance recognition performance through adaptation with the available test data. Additionally, DA typically necessitates access to both labeled data

Communicated by Hong Liu.

✉ Jian Liang
liangjian92@gmail.com

Ran He
rhe@nlpr.ia.ac.cn

Tieniu Tan
tnt@nlpr.ia.ac.cn

[1] New Laboratory of Pattern Recognition (NLPR) and State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS) State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA, Beijing, China

[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China

[3] Nanjing University, Nanjing, China

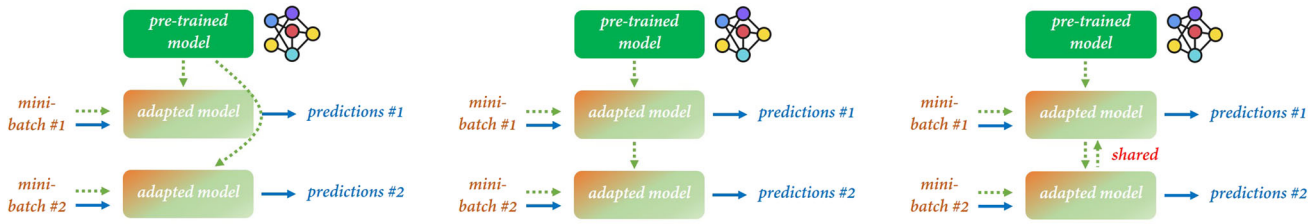

Fig. 1 The test-time adaptation (TTA) paradigm aims to adapt the pre-trained model to various types of unlabeled test data, including single mini-batch in (**a**), streaming data in (**b**), or an entire dataset in (**c**), before making predictions. During the adaptation process, either the model or the input data can be altered to improve performance against distribution shifts. The dotted green arrow indicates the test-time training phase before inference, while the blue arrow denotes pure inference

from the source domain and (unlabeled) data from the target domain simultaneously, which can be prohibitive in privacy-sensitive applications such as medical data. In contrast, TTA only requires access to the pre-trained model from the source domain, making it a secure and practical alternative solution.

Based on the characteristics of the test data,[1] TTA methods can be categorized into three distinct cases in Fig. 1: *test-time domain adaptation* (TTDA), *test-time batch adaptation* (TTBA), and *online test-time adaptation* (OTTA). For a better illustration, let us consider a scenario where there are m unlabeled mini-batches denoted as $b_1, \cdots, b_m$ during test time. Firstly, TTDA, also known as source-free domain adaptation (Liang et al., 2020; Kundu et al., 2020; Li et al., 2020), utilizes all m test batches for multi-epoch adaptation before generating final predictions. Secondly, TTBA individually adapts the pre-trained model to one[2] or a few instances (Sun et al., 2020; Zhang et al., 2022; Schneider et al., 2020; Zhang et al., 2021). In other words, the predictions made for each mini-batch are independent of the predictions made for the other mini-batches. Thirdly, OTTA (Wang et al., 2021; Iwasawa & Matsuo, 2021; Wang et al., 2022) adapts the pre-trained model to the target data $\{b_1, \cdots, b_m\}$ in an online manner, where each mini-batch can only be observed once. Importantly, the knowledge learned from previously observed mini-batches can facilitate adaptation to the current mini-batch. It is worth emphasizing that OTTA methods can be applied to TTDA with multiple epochs, and TTBA methods can be applied to OTTA with the assumption of knowledge reuse.

In this survey, we for the first time define the broad concept of *test-time adaptation* and consider the three aforementioned topics (*i.e.*, TTDA, TTBA, and OTTA) as its special cases. Subsequently, we thoroughly review the advanced algorithms for each topic and present a summary of various applications related to TTA. Our contributions can be summarized into three key aspects.

1. To our knowledge, this is the first survey that provides a systematic overview of three distinct topics within the broad test-time adaptation paradigm.
2. We propose a novel taxonomy of existing methods and provide a clear definition for each topic. We hope this survey will help readers gain a deeper understanding of the advancements in each area.
3. We analyze various applications related to the TTA paradigm in Sec. 6, and provide an outlook of recent emerging trends and open problems in Sec. 7 to shed light on future research directions.

Comparison with previous surveys While our survey contributes to the broader research area of DA, which has been previously reviewed in other works such as Kouw and Loog (2019), Wilson and Cook (2020), our specific focus is on test-time adaptation where the availability of source data during adaptation is limited or non-existent. Two recent surveys (Fang et al., 2024; Li et al., 2024) have focused on source-free domain adaptation which is a particular topic extremely similar to TTDA discussed in our survey. Even within the specific topic, we provide a novel taxonomy that encompasses a wider range of related papers. Another survey (Liu et al., 2021) considers source-free domain adaptation as an instance of data-free knowledge transfer, which shares some overlap with our survey. However, we unify TTDA and several related topics from the perspective of model adaptation under distribution shifts. We believe that it is a novel and pivotal contribution to the field of transfer learning.

[1] In this survey, we use the terms “test data" and “target data" interchangeably to refer to the data used for adaptation at test time.

[2] Such a single-sample adaptation corresponds to a batch size of 1, *a.k.a.*, test-time instance adaptation.

2 Related Research Topics

2.1 Domain Adaptation

As a special case of transfer learning (Pan & Yang, 2009), DA (Ben-David et al., 2010) typically leverages labeled data from a source domain to learn a classifier for an unlabeled target domain with a different distribution, in a transductive learning manner (Joachims, 1999). There are two major assumptions of distribution shift (Quinonero-Candela et al., 2008): *covariate shift* in which the input features cause the labels; and *label shift* in which the output labels cause the features. We briefly introduce a few popular techniques and refer the reader to the existing literature on DA (*e.g.*, Kouw and Loog 2019, Wilson and Cook 2020) for further information. DA methods rely on the existence of source data to bridge the domain gap, and existing techniques can be broadly divided into four categories, *i.e.*, input-level translation (Bousmalis et al., 2017; Hoffman et al., 2018), feature-level alignment (Long et al., 2015; Ganin & Lempitsky, 2015; Tzeng et al., 2017)), output-level regularization (Chen et al., 2019; Cui et al., 2020; Jin et al., 2020), and class-prior estimation (Saerens et al., 2002; Lipton et al., 2018; Azizzadenesheli et al., 2019). If it is feasible to generate training data from the source model (Li et al., 2020), then the task of TTDA can be tackled using conventional DA methods. Likewise, one relevant topic closely related to TTBA (batch size equals 1) is *one-shot domain adaptation* (Luo et al., 2020; Varsavsky et al., 2020), which entails adapting to a single unlabeled instance while still necessitating the source domain during adaptation. Moreover, OTTA is closely related to *online domain adaptation* (Moon et al., 2020; Yang et al., 2022), which involves adapting to an unlabeled target domain with streaming data that is promptly deleted after adaptation.

2.2 Hypothesis Transfer Learning

Hypothesis transfer learning (HTL) (Kuzborskij & Orabona, 2013) is another special case of transfer learning where pre-trained models (source hypotheses) retain information about previously encountered tasks. Shallow HTL methods (Yang et al., 2007; Tommasi et al., 2013; Ahmed et al., 2020) typically assume that the optimal target hypothesis is closely associated with these source hypotheses, and subsequent methods (Ao et al., 2017; Nelakurthi et al., 2018) extend this approach to a semi-supervised scenario where unlabeled target data are also utilized for training. Fine-tuning (Yosinski et al., 2014) is a typical example of a deep HTL method that may update a partial set of parameters in the source model. Despite HTL methods assuming no explicit access to the source domain or any knowledge about the relatedness of the source and target distributions, they still require a certain quantity of labeled data in the target domain. Another related topic is *domain-incremental learning* (van de Ven et al., 2022; Wang et al., 2022) which tackles the same type of problem but in diverse contexts. However, such an incremental learning task focuses more on the anti-forgetting ability after learning a supervised task.

2.3 Domain Generalization

DG (Li et al., 2018; Carlucci et al., 2019; Gulrajani & Lopez-Paz, 2020) aims to learn a model from one or multiple different but related domains that can generalize well on unseen testing domains. Researchers often devise specialized training techniques to enhance the generalization capability of the pre-trained model, which can be compatible with the studied TTA paradigm. Notably, MAML (Finn et al., 2017) is a representative approach that learns the initialization of a model's parameters to achieve optimal fast learning on a new task using a small number of samples and gradient steps. Such a meta-learning strategy offers a straightforward solution for TTA without the incorporation of test data in the meta-training stage. For further information, we refer the reader to existing literature (*e.g.*, Zhou et al. 2022, Wang et al. 2022, Hospedales et al. 2021).

2.4 Self-Supervised Learning

Self-supervised learning (Jing & Tian, 2020) is a learning paradigm that focuses on how to learn from unlabeled data by obtaining supervisory signals from the data itself through pretext tasks that leverage its underlying structure. Early pretext tasks in the computer vision field include image colorization (Zhang et al., 2016), image inpainting (Pathak et al., 2016), and image rotation (Gidaris et al., 2018). Advanced pretext tasks like clustering (Caron et al., 2018, 2020) and contrastive learning (He et al., 2020; Chen et al., 2020) have achieved remarkable success, even exceeding the performance of their supervised counterparts. Self-supervised learning is also popular in other fields like natural language processing (Kenton & Toutanova, 2019), speech processing (Baevski et al., 2020), and graph-structured data (You et al., 2020). For TTA tasks, these self-supervised learning techniques can be utilized to help learn discriminative features (Liang et al., 2022) or act as an auxiliary task (Sun et al., 2020).

2.5 Semi-Supervised Learning

Semi-supervised learning (Chen et al., 2022) is another learning paradigm concerned with leveraging unlabeled data to reduce the reliance on labeled data. A common objective for semi-supervised learning methods comprises two terms: a supervised loss over labeled data and an unsupervised loss over unlabeled data. Regarding the latter term, there are three

typical cases: self-training (Grandvalet & Bengio, 2004; Lee, 2013), which encourages the model to produce confident predictions; consistency regularization under input variations (Miyato et al., 2018; Sohn et al., 2020) and model variations (Laine & Aila, 2017; Tarvainen & Valpola, 2017), which forces networks to output similar predictions when inputs or models are perturbed; and graph-based regularization (Iscen et al., 2019), which seeks local smoothness by maximizing the pairwise similarities between nearby data points. For TTA tasks, these semi-supervised learning techniques can be easily integrated to unsupervisedly update the pre-trained model during adaptation.

2.6 Test-Time Augmentation

Test-time augmentation (Shanmugam et al., 2021) employs data augmentation techniques (Shorten & Khoshgoftaar, 2019) (*e.g.*, geometric transformations and color space augmentations) on test images to boost prediction accuracy (He et al., 2016), estimate uncertainty (Smith & Gal, 2018), and enhance robustness (Guo et al., 2018; Pérez et al., 2021). As a typical example, ten-crop testing (He et al., 2016) computes the final prediction by averaging predictions from ten different scaled versions of a test image. Other popular aggregation strategies include selective augmentation (Kim et al., 2020) and learnable aggregation weights (Shanmugam et al., 2021). In addition to data variation, Monte Carlo (MC) dropout (Gal & Ghahramani, 2016) enables dropout within the network during testing and performs multiple forward passes with the same input data to estimate the model uncertainty. Generally, test-time augmentation techniques do not explicitly consider distribution shifts but can be advantageous for TTA methods.

3 Test-Time Domain Adaptation

3.1 Problem Definition

Definition 1 (Domain) A domain $\mathcal{D}$ is a joint distribution $p(x, y)$ defined on the input–output space $\mathcal{X} \times \mathcal{Y}$, where random variables $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ denote the input data and the label (output), respectively.

In a well-studied DA problem, the domain of interest is called the target domain $p_{\mathcal{T}}(x, y)$ and the domain with labeled data is called the source domain $p_{\mathcal{S}}(x, y)$. The label y can either be discrete (in a classification task) or continuous (in a regression task). Unless otherwise specified, $\mathcal{Y}$ is a C-cardinality label set, and we usually have one labeled source domain $\mathcal{D}_{\mathcal{S}} = \{(x_1, y_1), \ldots, (x_{n_s}, y_{n_s})\}$ and one unlabeled target domain $\mathcal{D}_{\mathcal{T}} = \{x_1, \ldots, x_{n_t}\}$ under data distribution shifts: $\mathcal{X}_{\mathcal{S}} = \mathcal{X}_{\mathcal{T}}, p_{\mathcal{S}}(x) \neq p_{\mathcal{T}}(x)$, including the *covariate shift* (Quinonero-Candela et al., 2008) assumption ($p_{\mathcal{S}}(y|x) = p_{\mathcal{T}}(y|x)$). Other distribution shifts like *prior shift* (Saerens et al., 2002) are further discussed in Sec. 3.3. Typically, the unsupervised domain adaptation (UDA) paradigm aims to leverage supervised knowledge in $\mathcal{D}_{\mathcal{S}}$ to help infer the label of each target sample in $\mathcal{D}_{\mathcal{T}}$.

Chidlovskii et al. (2016) for the first time consider performing domain adaptation with no access to source data. Specifically, they propose three scenarios for feature-based domain adaptation with: source classifier with accessible models and parameters, source classifier as a black-box model, and source class means as representatives. This new setting utilizes all the test data to adjust the classifier learned from the training data, which could be regarded as a broad test-time adaptation scheme. Several methods (Clinchant et al., 2016; van Laarhoven & Marchiori, 2017; Liang et al., 2019) follow this learning mechanism and adapt the source classifier to unlabeled target features. To gain benefits from end-to-end representation learning, researchers are more interested in generalization with deep models. Such a setting without access to source data during adaptation is termed as source data-absent domain adaptation (Liang et al., 2020, 2022), model adaptation (Li et al., 2020), and source-free domain adaptation (Kundu et al., 2020), respectively. For the sake of simplicity, we utilize the term *test-time domain adaptation* and give a unified definition.

Definition 2 (Test-Time Domain Adaptation, TTDA) Given a well-trained classifier $f_{\mathcal{S}} : \mathcal{X}_{\mathcal{S}} \rightarrow \mathcal{Y}_{\mathcal{S}}$ on the source domain $\mathcal{D}_{\mathcal{S}}$ and an unlabeled target domain $\mathcal{D}_{\mathcal{T}}$, *test-time domain adaptation* aims to leverage the labeled knowledge implied in $f_{\mathcal{S}}$ to infer labels of all the samples in $\mathcal{D}_{\mathcal{T}}$, in a transductive learning (Joachims, 1999) manner. Note that, all test data (target data) are required to be seen during adaptation (Table 1).

3.2 Taxonomy on TTDA Algorithms

3.2.1 Pseudo-Labeling

To adapt a pre-trained model to an unlabeled target domain, a majority of TTDA methods take inspiration from the semi-supervised learning (SSL) field (Chen et al., 2022) and employ various prevalent SSL techniques tailored for unlabeled data during adaptation. A simple yet effective technique, pseudo-labeling (Lee, 2013), aims to assign a class label $\hat{y} \in \mathbb{R}^C$ for each unlabeled sample x in $\mathcal{X}_t$ and optimize the following supervised learning objective to guide the learning process,

$$\min_{\theta} \mathbb{E}_{\{x,\hat{y}\}\in\mathcal{D}_t}\ w_{pl}(x) \cdot d_{pl}(\hat{y}, p(y|x;\theta)), \tag{1}$$

Table 1 A taxonomy on TTDA methods with representative strategies

Families	Model rationales	Representative strategies
Peudo-labeling	Centroid-based	SHOT (Liang et al., 2020, 2022), BMD (Qu et al., 2022)
	Nighbor-based	NRC (Yang et al., 2021), SSNLL (Chen et al., 2022)
	Complementary labels	LD (You et al., 2021), ATP (Wang et al., 2022)
	Optimization-based	ASL (Yan et al., 2021), KUDA (Sun et al., 2022)
Consistency	Data variations	G-SFDA (Yang et al., 2021), APA (Sun et al., 2023)
	Model variations	SFDA-UR (Sivaprasad & Fleuret, 2021), FMML (Peng et al., 2022)
	Both variations	AdaContrast (Chen et al., 2022), MAPS (Ding et al., 2024)
Clustering	Entropy minimization	ASFA (Xia et al., 2022), 3C-GAN (Li et al., 2020)
	Mutual information	SHOT (Liang et al., 2020, 2022), UMAD (Liang et al., 2021)
	Explicit clustering	ISFDA (Li et al., 2021), SDA-FAS (Liu et al., 2022)
Source estimation	Data generation	3C-GAN (Li et al., 2020), DI (Nayak et al., 2022)
	Data translation	SFDA-IT (Hou & Zheng, 2020), ProSFDA (Hu et al., 2022)
	Data selection	SHOT++ (Liang et al., 2022), DaC (Zhang et al., 2022)
	Feature estimation	VDM-DA (Tian et al., 2022), CPGA (Qiu et al., 2021)
Self-supervision	Auxiliary tasks	SHOT++ (Liang et al., 2022), StickerDA (Kundu et al., 2022)

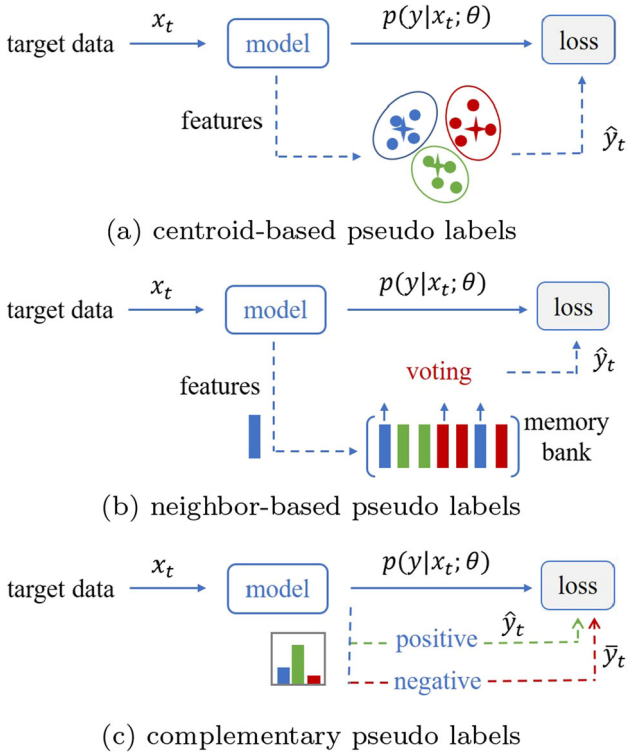

Fig. 2 Three representative types of pseudo-labeling, where θ represents the model parameters, and $\hat{y}_t$ (or $\bar{y}_t$) denotes the pseudo label of the instance x_t

where $w_{pl}(x)$ denotes the real-valued weight associated with each pseudo-labeled sample $\{x, \hat{y}\}$, and $d_{pl}(\cdot)$ denotes the divergence between the predicted label probability distribution and the pseudo label probability $\hat{y}$, *e.g.*, $-\sum_c \hat{y}_c \log[p(y|x;\theta)]_c$ if using the cross entropy as the divergence measure. Since the pseudo labels of target data are inevitably inaccurate under domain shift, there exist three different solutions: (1) improving the quality of pseudo labels via denoising; (2) filtering out inaccurate pseudo labels with $w_{pl}(\cdot)$; and (3) developing a robust divergence measure $d_{pl}(\cdot,\cdot)$ for pseudo-labeling. To reduce the effects of noisy pseudo labels based on the argmax operation (Kim et al., 2021; Li et al., 2021; Chen et al., 2021), most TTDA methods (*e.g.*, SFIT (Hou and Zheng 2021)) consider only reliable pseudo labels using diverse filtering mechanisms. Figure 2 illustrates three representative types of pseudo-labeling, which will be elaborated in the following part.

Centroid-based pseudo labels Inspired by a classic self-supervised approach, DeepCluster (Caron et al., 2018), SHOT (Liang et al., 2020, 2022) resorts to target-specific clustering for denoising the pseudo labels. The key idea is to obtain target-specific class centroids based on the network predictions and the target features and then derive the unbiased pseudo labels via the nearest centroid classifier. Formally, the class centroids and pseudo labels are updated as follows,

$$\begin{cases} m_c = \sum_x [p_\theta(y_c|x) \cdot g(x)] / \sum_x p_\theta(y_c|x), \ c \in [1, C], \\ \hat{y} = \arg\min_c d(g(x), m_c), \ \forall x \in \mathcal{D}_t, \end{cases} \tag{2}$$

where $p_\theta(y_c|x) = [p(y|x;\theta)]_c$ denotes the probability associated with the c-th class, and $g(x)$ denotes the feature of input x. m_c denotes the c-th class centroid, and $d(\cdot,\cdot)$ denotes the cosine distance function. As class centroids always contain robust discriminative information and meanwhile weaken the category imbalance problem, this label

refinery is prevalent in follow-up TTDA studies (Zhang et al., 2022; Tang et al., 2021; Qiu et al., 2021).

Twofer (Liu et al., 2023) identifies confident samples to build more accurate centroids, while BMD (Qu et al., 2022) posits that a coarse centroid may not effectively represent ambiguous data and instead employs K-means clustering to discover multiple prototypes for each class. Additionally, CoWA-JMDS (Lee et al., 2022) performs Gaussian Mixture Modeling (GMM) in the target feature space to obtain the log-likelihood and pseudo label of each sample. Apart from hard pseudo labels, FAUST (Lee & Lee, 2023) explores soft pseudo labels based on the class centroids, *e.g.*, $[\hat{y}]_c = \frac{exp(-d(g(x),m_c)/\tau)}{\sum_c exp(-d(g(x),m_c)/\tau)}$, where τ denotes the temperature. In contrast, BMD (Qu et al., 2022) employs the exponential moving average (EMA) technique to dynamically accumulate the class centroids in mini-batches.

Neighbor-based pseudo labels Another prevalent label denoising technique is to generate pseudo labels by incorporating the predictions of neighboring labels, relying on the assumption of local smoothness (Chen et al., 2022; Wang et al., 2022; Cao et al., 2021; Chen et al., 2022; Ding et al., 2023). For instance, SSNLL (Chen et al., 2022) performs K-means clustering in the target domain and aggregates predictions of its neighbors within the same cluster. DIPE (Wang et al., 2022) diminishes label ambiguity by correcting the pseudo label to the majority vote of its neighbors. In contrast, SFDA-APM (Kim et al., 2021) constructs an anchor set comprising only highly confident target samples and employs a point-to-set distance function to generate the pseudo labels. CAiDA (Dong et al., 2021) proposes a greedy chain-search strategy to find its nearest neighbor in the anchor set, interpolates its nearest anchor to the target feature, and uses the prediction of the synthetic feature instead.

Inspired by neighborhood aggregation (Liang et al., 2021), a few works (Cao et al., 2021; Chen et al., 2022; Ding et al., 2023; Litrico et al., 2023) maintain a memory bank storing both features and predictions of the target data $\{g(x_i), q_i\}_{i=1}^{n_t}$, allowing online refinement of pseudo labels. Typically, the refined pseudo label is obtained through $\hat{p}_i = \frac{1}{m}\sum_{j\in\mathcal{N}_i} q_j$, where $\mathcal{N}_i$ denotes the indices of m nearest neighbors of $g(x_i)$ in the memory bank. Specifically, ProxyMix (Ding et al., 2023) sharpens the network output $\bar{p}$ with the class frequency to avoid class imbalance, while NRC (Yang et al., 2021) devises a weighting scheme for neighbors during aggregation. Instead of using the soft pseudo label $\hat{p}$, AdaContrast (Chen et al., 2022) utilizes the hard pseudo label with the argmax operation.

Complementary pseudo labels Motivated by the idea of negative learning (Kim et al., 2019), PR-SFDA (Luo et al., 2021) randomly chooses a label from the set $\{1, \ldots, C\}\backslash\{\hat{y}_i\}$ as the complementary label $\bar{y}_i$ and thus optimizes the following loss function,

$$\min_{\theta} -\sum\nolimits_{i=1}^{n_t}\sum\nolimits_{c=1}^{C} \mathbb{1}(\bar{y}_i = c)\log(1 - p_\theta(y_c|x_i)), \tag{3}$$

where $\hat{y}_i$ denotes the inferred hard pseudo label. $\bar{y}$ is referred to as a negative pseudo label, indicating that the given input does not belong to this label. The probability of correctness is $\frac{C-1}{C}$ for the complementary label $\bar{y}_i$, providing correct information even from wrong labels $\hat{y}_i$. LD (You et al., 2021) develops a heuristic strategy to randomly select an informative complementary label with medium prediction scores. Besides, NEL (Ahmed et al., 2022) and PLUE (Litrico et al., 2023) randomly select multiple complementary labels, except for the inferred pseudo label, and optimizes the multi-class variant of Eq. (3). ATP (Wang et al., 2022) further generates multiple complementary labels according to a predefined threshold on prediction scores.

Optimization-based pseudo labels By leveraging the prior knowledge of the target label distribution like class balance (Zou et al., 2018), some TTDA methods (You et al., 2021; Sivaprasad & Fleuret, 2021; Huang et al., 2021) vary the threshold for each class so that a certain proportion of points per class are selected. Such a strategy helps avoid the 'winner-takes-all' dilemma where the pseudo labels come from several major categories, potentially deteriorating the following training process. Furthermore, ASL (Yan et al., 2021) directly imposes the equi-partition constraint on the pseudo labels $\hat{p}_i$ and solves the optimization problem below,

$$\begin{aligned} &\min_{\hat{p}_i} -\sum\nolimits_i\sum\nolimits_c \hat{p}_{ic}\log p_\theta(y_c|x_i) + \lambda\sum\nolimits_i\sum\nolimits_c \hat{p}_{ic}\log\hat{p}_{ic}, \\ &s.t.\ \forall i, c:\ \hat{p}_{ic}\in[0,1],\ \sum\nolimits_c \hat{p}_{ic} = 1,\ \sum\nolimits_i \hat{p}_{ic} = \frac{n_t}{C}. \end{aligned} \tag{4}$$

Likewise, IterNLL (Zhang et al., 2021) provides a closed-form solution of $\{\hat{p}\}$ under the uniform prior assumption. KUDA (Sun et al., 2022) even introduces a hard constraint $\hat{p}_{ic} \in \{0, 1\}$ and solves the zero–one programming problem. In addition, ReCLIP (Hu et al. 2024) constructs the affinity graph and employs label propagation to produce closed-form pseudo labels.

Ensemble-based pseudo labels Rather than relying on a single noisy pseudo label, ISFDA (Li et al., 2021) generates a secondary pseudo label to aid the primary one. Besides, ASL (Yan et al., 2021) and C-SFDA (Karim et al., 2023) adopt a weighted average of predictions under multiple random data augmentation, while ELR (Yi et al., 2023) ensembles historical predictions from previous training epochs. NEL (Ahmed et al., 2022) further aggregates logits under different data augmentation and trained models simultaneously. Inspired by a classic semi-supervised learning method (Laine & Aila, 2017), some TTDA methods (Liang et al., 2022; Panagiotakopoulos et al., 2022) maintain an EMA of predictions at

different time steps as pseudo labels. Moreover, C-SFDA (Karim et al., 2023) maintains a mean teacher model (Tarvainen & Valpola, 2017) that generates pseudo labels for the current student network. Additionally, other methods attempt to generate pseudo labels based on predictions from various models, *e.g.*, multiple source models (Liang et al., 2022; Li et al., 2022), a multi-head classifier (Kundu et al., 2021), and models from both domains (Hou & Zheng, 2021). In particular, SFDA-VS (Ye et al., 2021) follows MC dropout (Gal & Ghahramani, 2016) and obtains the final prediction through multiple forward passes.

Another line of ensemble-based TTDA methods (Cao et al., 2021; Yan et al., 2021; Xiong et al., 2022) aims to integrate predictions from different labeling criteria using a weighted average. For example, e-SHOT-CE (Cao et al., 2021) utilizes both centroid-based and neighbor-based pseudo labels. Besides the weighting scheme, other approaches (Qiu et al., 2021; Dong et al., 2021; Wang et al., 2022; Kumar et al., 2023) explore different labeling criteria in a cascade manner. For instance, DIPE (Wang et al., 2022) employs the neighbor-based labeling criterion with centroid-based pseudo labels.

Learning with pseudo labels Existing pseudo-labeling-based TTDA methods have employed various robust divergence measures d_{pl}. Generally, most methods utilize the standard cross-entropy loss for all target samples with hard pseudo labels (Liang et al., 2020; Yan et al., 2021) or soft pseudo labels (Tang et al., 2021; Deng et al., 2021). Note that several methods (Ding et al., 2023; Tian et al., 2023) convert hard pseudo labels into soft pseudo labels using the label smoothing trick (Müller et al., 2019). As pseudo labels are noisy, many TTDA methods incorporate an instance-specific weighting scheme into the standard cross-entropy loss, including hard weights (Kim et al., 2021; Hou & Zheng, 2021; Chen et al., 2021), and soft weights (Huang et al., 2021; Ye et al., 2021). Besides, AUGCO (Prabhu et al., 2022) considers the class-specific weight in the cross-entropy loss to mitigate label imbalance. In addition to the cross-entropy loss, alternative choices include the generalized cross entropy (Rusak et al., 2022), the inner product distance between the pseudo label and the prediction (Yang et al., 2021; Qiu et al., 2021), and a new discrepancy measure $\log(1-\hat{y}^T p(y|x;\theta))$ (Yi et al., 2023). Moreover, BMD (Qu et al., 2022) and OnDA (Panagiotakopoulos et al., 2022) employ the symmetric cross-entropy loss to guide the self-labeling process. CATTAn (Thopalli et al., 2023) exploits the negative log-likelihood ratio between correct and competing classes.

3.2.2 Consistency Training

Consistency regularization, a prevailing strategy in recent semi-supervised learning literature (Yang et al., 2022; Chen et al., 2022), is primarily built on the smoothness assumption or the manifold assumption. It aims to enforce consistent network predictions or features under variations in the input data space or the model parameter space. Moreover, another line of consistency training methods attempts to match the statistics of different domains even without the source data. Figure 3 illustrates three representative types of consistency training, which will be elaborated in the following part.

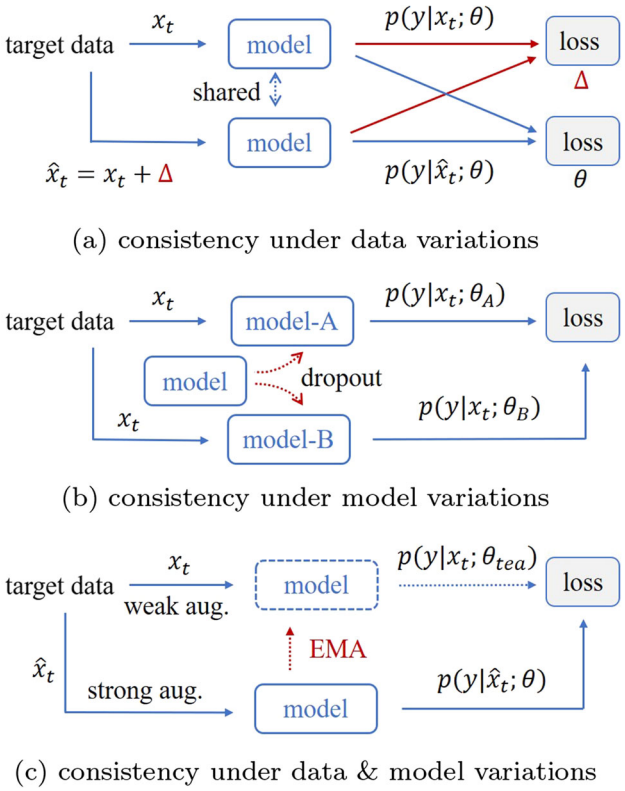

Fig. 3 Three representative types of consistency training, where $\hat{x}_t$ represents the data variant of x_t, and θ_A (or θ_B and θ_{tea}) denotes the model variant of θ

Consistency under data variations Benefiting from advanced data augmentation techniques such as RandAugment (Cubuk et al., 2020), several prominent semi-supervised learning methods (Xie et al., 2020; Sohn et al., 2020) unleash the power of consistency regularization over unlabeled data that can be effortlessly adopted in TTDA approaches. An exemplar of consistency regularization (Sohn et al., 2020) is expressed as:

$$\mathcal{L}_{fm}^{con} = \frac{1}{n_t}\sum_{i=1}^{n_t} \mathrm{CE}\left(p_{\tilde{\theta}}(y|x_i),\, p_\theta(y|\hat{x}_i)\right), \tag{5}$$

where $p_\theta(y|x_i) = p(y|x_i;\theta)$, and $\mathrm{CE}(\cdot,\cdot)$ refers to cross-entropy between two distributions. Besides, $\hat{x}_i$ represents the variant of x_i under another augmentation transformation, and $\tilde{\theta}$ is a fixed copy of current network parameters θ. Another representative consistency regularization is vir-

tual adversarial training (VAT) (Miyato et al., 2018), which devises a smoothness constraint as follows,

$$\mathcal{L}_{vat}^{con} = \frac{1}{n_t}\sum_{i=1}^{n_t}\max_{\|\Delta_i\|\leq\epsilon}[\mathrm{KL}(p_{\tilde{\theta}}(y|x_i)\;||\;p_{\theta}(y|x_i+\Delta_i))], \quad (6)$$

where Δ_i is a perturbation that disperses the prediction most within an intensity range of ϵ for the target data x_i, and KL denotes the Kullback–Leibler divergence.

ATP (Wang et al., 2022) directly employs the same consistency regularization in Eq. (5), while other TTDA methods (Wang et al., 2024; Chen et al., 2022; Zhang et al., 2022; Kumar et al., 2023) replace $p_{\tilde{\theta}}(y|x_i)$ with hard pseudo labels for target data under weak augmentation, followed by a cross-entropy loss for target data under strong augmentation. Note that, many of these hard labels are obtained using the label denoising techniques mentioned earlier. Apart from strong augmentations, ProSFDA (Hu et al., 2022) and SFDA-FSM (Yang et al., 2022) require learning the domain translation module first, and ProSFDA seeks feature-level consistency under different augmentations at the same time. TeST (Sinha et al., 2023) introduces a flexible mapping network to match features under two different augmentations. On the contrary, OSHT (Feng et al., 2021) maximizes the mutual information between the predictions of two different transformed inputs to retain the semantic information as much as possible.

Following the objective in Eq. (6), another line of TTDA methods (Li et al., 2020; Yan et al., 2021) attempts to encourage consistency between target samples with their data-level neighbors, while APA (Sun et al., 2023) learns the neighbors in the feature space. Instead of generating the most divergent neighbor $x_i + \Delta_i$ according to the predictions, JN (Li et al., 2022) devises a Jacobian norm regularization to control the smoothness in the neighborhood of the target sample. Furthermore, G-SFDA (Yang et al., 2021) discovers multiple neighbors from a memory bank and minimizes their inner product distances over the predictions. Moreover, Mixup (Zhang et al., 2018) performs linear interpolations on two inputs and their corresponding labels, which can be treated as seeking consistency under data variation (Liang et al., 2022; Lee et al., 2022; Kumar et al., 2023).

Consistency under model variations Reducing model uncertainty (Gal & Ghahramani, 2016) is also beneficial for learning robust features for TTDA tasks, on top of uncertainty measured with input change. Following MC dropout (Gal & Ghahramani, 2016), FAUST (Lee & Lee, 2023) activates dropout in the model and performs multiple stochastic forward passes to estimate the epistemic uncertainty. SFDA-UR (Sivaprasad & Fleuret, 2021) appends multiple extra dropout layers behind the feature encoder and minimizes the mean squared error (MSE) between predictions as uncertainty. Further, ASFA (Xia et al., 2022) adds different perturbations to the intermediate features to promote predictive consistency. FMML (Peng et al., 2022) offers another form of model variation by network slimming and sought predictive consistency across different networks.

Another consistency regularization requires the existence of both the source and target models and thus minimizes the difference across different models, such as feature-level discrepancy (Kothandaraman et al., 2023) and output-level discrepancy (Liang et al., 2022; Conti et al., 2022; Sinha et al., 2023). Furthermore, the mean teacher framework (Tarvainen & Valpola, 2017) is also utilized to form a strong teacher model and a learnable student model. The teacher and the student models share the same architecture, and the weights of the teacher model θ_{tea} are gradually updated by $\theta_{tea} = (1-\eta)\theta_{tea} + \eta\theta$, where θ denotes the weights of the student model, and η is the momentum coefficient. Therefore, the mean teacher model is regarded as a temporal ensemble of student models with more accurate predictions. In reality, a few TTDA methods including (Lao et al., 2021) consider the multi-head classifier and promote consistent predictions by different heads.

Consistency under data & model variations In reality, data variation and model variation could be integrated into a unified framework. For example, the mean teacher framework (Tarvainen & Valpola, 2017) is enhanced by blending strong data augmentation techniques, and the discrepancy between predictions of the student and teacher models is minimized as follows,

$$\mathcal{L}_{mt}^{con} = \mathbb{E}_{x\in\mathcal{D}_t} d_{mt}(p(y|x,\theta),\, p(y|\tau(x),\theta_{tea})), \quad (7)$$

where $\tau(\cdot)$ denotes the strong data augmentation, and d_{mt} denotes the divergence measure, *e.g.*, the KL divergence (Liu et al., 2022; Hou & Zheng, 2021), the MSE loss (Zhang et al., 2021), and the cross-entropy loss (Chen et al., 2022). Besides, several methods (Vibashan et al., 2022; Liu & Yuan, 2022; Huang et al., 2021; Li et al., 2022) attempt to extract useful information from the teacher and employ task-specific loss functions to seek consistency. Apart from the output-level consistency, TT-SFUDA (Vibashan et al., 2022) matches the features extracted by different models with the MSE distance, while AdaContrast (Chen et al., 2022) and PLUE (Litrico et al., 2023) learn semantically consistent features like MoCo (He et al., 2020).

Instead of strong data augmentations, LODS (Li et al., 2022) and SFIT (Hou & Zheng, 2021) use the style transferred image instead, MAPS (Ding et al., 2024) considers spatial transforms, and SMT (Zhang et al., 2021) elaborates the domain-specific perturbation by averaging the target images. Different from model variations in the mean teacher scheme, OnTA (Wang et al., 2021) distills knowledge from the source model to the target model, while HCL (Huang

et al., 2021) promotes feature-level consistency among the current model and historical model.

Miscellaneous consistency regularizations To prevent excessive deviation from the original source model, a flexible strategy is adopted by a few TTDA methods (Li et al., 2020; Xiong et al., 2022) by establishing a parameter-based regularization term $\|\theta_s - \theta\|_2^2$, where θ_s is the fixed source weight. Another line of research focuses on matching the batch normalization (BN) statistics (*i.e.*, the mean and the variance), across models with different measures, such as the KL divergence (Ishii & Sugiyama, 2021) and the MSE error (Zhang et al., 2021; Ahmed et al., 2022), whereas OSUDA (Liu et al., 2021) encourages the learned scaling and shifting parameters in BN layers to be consistent. Similarly, an explicit feature-level regularization (Liu et al., 2021) is devised to match the first and second-order moments of features in different domains.

As for the network architecture in the target domain, a unique design termed dual-classifier is utilized to seek robust domain-invariant representations. For example, BAIT (Yang et al., 2023) introduces an extra C-dimensional classifier to the source model, forming a dual-classifier model with a shared feature encoder. During adaptation in the target domain, the shared feature encoder and the new classifier are trained with the classifier from the source domain head fixed. Such a training scheme has also been utilized by many TTDA methods (Tian et al., 2023; Wang et al., 2022; Xia et al., 2021; Sivaprasad & Fleuret, 2021; Xia et al., 2022) through modeling the consistency between different classifiers. Besides, SFDA-APM (Kim et al., 2021) develops a self-training framework that optimizes the shared feature encoder and two classification heads with different pseudo-labeling losses, respectively.

3.2.3 Clustering-Based Training

Except for the pseudo-labeling paradigm, nearly all semi-supervised learning algorithms rely on the cluster assumption (Yang et al., 2022), which asserts that the decision boundary should not cross high-density regions, but instead lie in low-density regions. As a result, another popular category of TTDA approaches favors low-density separation by reducing the uncertainty of the target network predictions (Liang et al., 2020; Li et al., 2020) or promoting clustering among the target features (Li et al., 2021; Qiu et al., 2021). Figure 4 illustrates these two representative types of clustering-based training, which will be elaborated in the following part.

Entropy minimization ASFA (Xia et al., 2022) utilizes robust measures from information theory to encourage confident predictions for unlabeled target data. To achieve this,

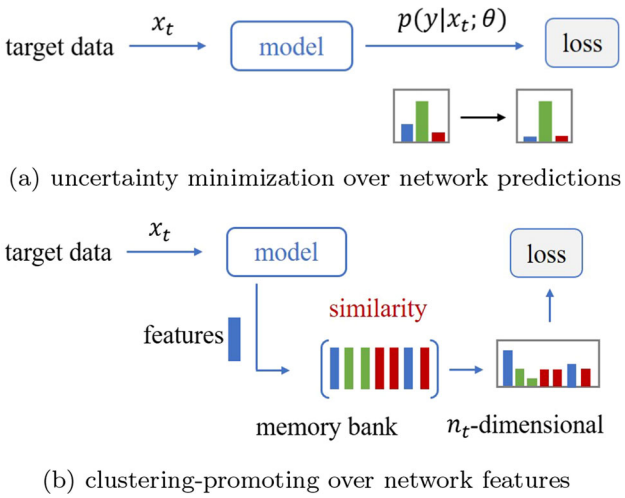

Fig. 4 Two representative types of clustering-based training, where similarity is obtained based on a feature memory bank

it minimizes the α-Tsallis entropy given by:

$$\mathcal{L}_{tsa} = \frac{1}{n_t}\sum_{i=1}^{n_t}\frac{1}{\alpha-1}[1-\sum_{c=1}^{C} p_\theta(y_c|x_i)^\alpha], \tag{8}$$

where $\alpha > 0$ is called the entropic index. Note that, as α approaches 1, the Tsallis entropy converges to the standard Shannon entropy, given by $\mathcal{H}(p_\theta(y|x_i)) = \sum_c p_\theta(y_c|x_i) \log p_\theta(y_c|x_i)$. In practice, the conditional Shannon entropy $\mathcal{H}(p_\theta(y|x))$ has been widely used in TTDA methods (Li et al., 2020; Liu et al., 2021; Sivaprasad & Fleuret, 2021; You et al., 2021; Kundu et al., 2021; Bateson et al., 2022; Sinha et al., 2023). Besides, there exist numerous variations of standard entropy minimization. For instance, SFDA-VS (Ye et al., 2021) develops a nonlinear weighted entropy minimization loss that emphasizes low-entropy samples. TT-SFUDA (Vibashan et al., 2022) focuses on the entropy of the ensemble predictions under multiple augmentations.

When α is set to 2, the Tsallis entropy in Eq. (8) is equivalent to the maximum squares loss (Chen et al., 2019; Liu et al., 2021; Kumar et al., 2023), given by $\sum_c p_\theta(y_c|x_i)^2$. Compared to the Shannon entropy, the gradient of the maximum squares loss increases linearly, preventing easy samples from dominating the training process in the high probability region. Building on this, Batch Nuclear-norm Maximization (BNM) (Cui et al., 2020) approximates the prediction diversity using the matrix rank, which is utilized by CDL (Wang et al., 2024). Additionally, SI-SFDA (Ye et al., 2022) pays attention to the class confusion matrix and minimizes the inter-class confusion to ensure that no samples are ambiguously classified into two classes at the same time.

Mutual information maximization Another favorable clustering-based regularization is mutual information maximization, which aims to maximize the mutual information (Shi

& Sha, 2012) between the inputs and the discrete labels as follows,

$$\begin{aligned}\max_{\theta} \mathcal{I}(\mathcal{X}_t, \hat{\mathcal{Y}}_t) &= \mathcal{H}(\hat{\mathcal{Y}}_t) - \mathcal{H}(\hat{\mathcal{Y}}_t|\mathcal{X}_t) \\ &= -\sum_{c=1}^{C} \bar{p}_\theta(y_c) \log \bar{p}_\theta(y_c) \\ &\quad + \frac{1}{n_t} \sum_{i=1}^{n_t} \sum_{c=1}^{C} p_\theta(y_c|x_i) \log p_\theta(y_c|x_i), \end{aligned} \tag{9}$$

where $\bar{p}_\theta(y_c) = \frac{1}{n_t} \sum_i p_\theta(y_c|x_i)$ denotes the c-th element in the estimated class label distribution. Intuitively, increasing the extra diversity term $\mathcal{H}(\hat{\mathcal{Y}}_t)$ promotes uniform distribution of target labels, circumventing the degenerate solution where each sample is assigned to the same class. Such a regularization is initially introduced in SHOT (Liang et al., 2020) and SHOT++ (Liang et al., 2022) for image classification and then employed in plenty of TTDA methods (Ishii & Sugiyama, 2021; Lao et al., 2021; Wang et al., 2021; Li et al., 2022; Wang et al., 2022; Litrico et al., 2023). Instead of using the network prediction $p_\theta(y|x)$, GKD (Tang et al., 2021) employs the ensemble prediction based on its neighbors for mutual information maximization. DaC (Zhang et al., 2022) and U-SFAN (Roy et al., 2022) introduce a balancing parameter between two terms in Eq. (9) to increase flexibility. In particular, U-SFAN (Roy et al., 2022) develops an uncertainty-guided entropy minimization loss by emphasizing low-entropy predictions, whereas ATP (Wang et al., 2022) encompasses the instance-wise uncertainty in both terms of Eq. (9). VMP (Jing et al., 2022) further provides a probabilistic framework based on Bayesian neural networks and integrates mutual information into the likelihood function.

It is worth noting that the diversity term can be rewritten as $\mathcal{H}(\hat{\mathcal{Y}}_t) = -\mathrm{KL}(\bar{p}_\theta(y)||\mathcal{U}) + \log C$, where $\bar{p}_\theta(y)$ denotes the average label distribution in the target domain, and $\mathcal{U}$ is a C-dimensional uniform vector. This term alone has also been employed in numerous TTDA methods (Hou & Zheng, 2021; Yang et al., 2021; Chen et al., 2022; Kundu et al., 2022; Panagiotakopoulos et al., 2022; Tian et al., 2023; Panagiotakopoulos et al., 2022; Thopalli et al., 2023) to prevent class collapse. To better guide the learning process, a few works (Krause et al., 2010; Hu et al., 2017) modify the mutual information regularization by substituting a reference class-ratio distribution in place of $\mathcal{U}$. Unlike AdaMI (Bateson et al., 2022), which leverages the target class ratio as a prior, UMAD (Liang et al., 2021) utilizes the flattened label distribution within a mini-batch instead to mitigate the class imbalance problem, and AUGCO (Prabhu et al., 2022) maintains the moving average of the predictions as the reference distribution.

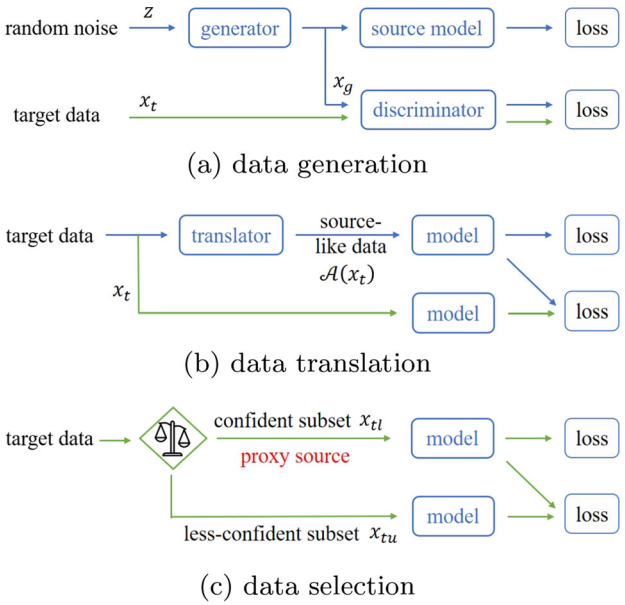

Fig. 5 Three representative types of source distribution estimation, where surrogate source data is obtained through generation, translation, and selection, respectively

3.2.4 Source Distribution Estimation

Another favored family of TTDA approaches compensates for the absence of source data by inferring data from the pre-trained model, transforming the challenging TTDA problem into a well-studied DA problem. Existing source estimation approaches could be categorized into three groups: data generation from random noises (Morerio et al., 2020; Li et al., 2020; Kurmi et al., 2021), data translation (Hou & Zheng, 2021; Yan et al., 2021; Zhou et al., 2022), and data selection (Liang et al., 2022; Yang et al., 2023; Ding et al., 2023). Figure 5 illustrates three representative types of source distribution estimation, which will be elaborated in the following part.

Data generation To generate valid target-style source samples, 3C-GAN (Li et al., 2020) introduces a data generator $G(\cdot;\theta_G)$ conditioned on randomly sampled labels, along with a binary discriminator $D(\cdot;\theta_D)$. The optimization objective is similar to the conditional GAN (Mirza & Osindero, 2014) that is written as follows:

$$\begin{aligned}\min_{\theta_G} \max_{\theta_D} \; & \mathbb{E}_{x_t \in \mathcal{X}_t}[\log D(x_t)] + \mathbb{E}_{y_t,z}[\log(1 - D(G(y_t, z)))] \\ & - \lambda_s \mathbb{E}_{y_t,z} \sum\nolimits_c \mathbb{1}(y_t = c) \log p(y_c|G(y_t, z), \theta), \end{aligned} \tag{10}$$

where z is a random noise vector, y_t is a pre-defined label, $\lambda_s > 0$ is a balancing parameter, and θ denotes the parameters of the pre-trained prediction model. By alternately optimizing θ_G and θ_D, the resulting class conditional generator G can

generate multiple surrogate labeled source instances for the subsequent domain alignment step, *i.e.*, $\mathcal{D}_g = \{x_i, y_i\}_{i=1}^{n_g}$, where $x_i = G(y_i, z)$ and n_g is the number of generated samples. PLR (Morerio et al., 2020) disregards the last term in Eq. (10) to infer diverse target-like samples. On the other hand, SDDA (Kurmi et al., 2021) maximizes the log-likelihood of generated data x_g and employs two different domain discriminators, *i.e.*, a data-level GAN discriminator and a feature-level domain discriminator.

In addition to adversarial training, DI (Nayak et al., 2022) performs Dirichlet modeling with the source class similarity matrix and then optimizes the noisy input to match its output with the sampled softmax vector q as,

$$x_g = \arg\min_x \mathrm{CE}(q, p_\theta(y|x)) \tag{11}$$

which is referred to as data impression of the source domain. Besides, SPGM (Yang et al., 2022) first estimates the target distribution using GMM and then constrains the generated data to be derived from the target distribution.

Motivated by recent advances in data-free knowledge distillation (Yin et al., 2020; Liu et al., 2021), SFDA-KTMA (Liu et al., 2021) exploits the moving average statistics of activations stored in BN layers of the pre-trained source model and imposes the following BN matching constraint on the generator,

$$\mathcal{L}_{bn} = \sum\nolimits_l \sum\nolimits_i \|\mu_{g,l}^{(i)} - \mu_{s,l}^{(i)}\|_2 + \|{\delta_{g,l}^{(i)}}^2 - {\delta_{s,l}^{(i)}}^2\|_2, \tag{12}$$

where B is the size of a mini-batch, $\mu_{s,l}^{(i)}$ and ${\delta_{s,l}^{(i)}}^2$ represent the corresponding running mean and variance stored in the source model, and $\mu_{g,l}^{(i)} = \frac{1}{B}\sum_z f_l^{(i)}(x_g)$ and ${\delta_{g,l}^{(i)}}^2 = \frac{1}{B}\sum_z (f_l^{(i)}(x_g) - \mu_{g,l}^{(i)})^2$ denote the batch-wise mean and variance estimates of the i-th feature channel at the l-th layer for synthetic data from the generator, respectively. As indicated in Li et al. (2017), matching the BN statistics can aid in ensuring that the generated data resembles the source style. SFDA-FSM (Yang et al., 2022) further minimizes the L_2-norm difference between intermediate features (*a.k.a.*, the content loss Gatys et al. 2016) to preserve the content knowledge of the target domain.

Data translation SSFT-SSD (Yan et al., 2021) initializes x_g as $x_t \in \mathcal{X}_t$ and directly performs optimization on the input space with the gradient of the L_2-norm regularized cross-entropy loss being zero. On the contrary, SFDA-TN (Sahoo et al., 2020) optimizes a learnable data transformation network that maps target data to the source domain such that the maximum class probability is maximized. Inspired by the success of visual prompts (Bahng et al., 2022), ProSFDA (Hu et al., 2022) adds a learnable image perturbation to all target data, enabling the BN statistics to be aligned with those stored in the source model. Besides, the style-transferred image is obtained using spectrum mixup (Yang & Soatto, 2020) between the target image and its perturbed image.

Another line of data translation methods (Hou & Zheng, 2020, 2021; Zhou et al., 2022) explicitly introduces an additional module $\mathcal{A}$ to transfer target data to source-like style. In particular, SFDA-IT (Hou & Zheng, 2020) optimizes the translator with the style matching loss in Eq. (12) as well as the feature-level content loss, with the source model frozen. Furthermore, SFDA-IT (Hou & Zheng, 2020) employs entropy minimization over the fixed source model to promote semantic consistency. To improve the performance of style transfer, SFIT (Hou & Zheng, 2021) further develops a variant of the style reconstruction loss (Gatys et al., 2016) as follows,

$$\mathcal{L}_{style} = \|g(x)g(x)^T - g(\mathcal{A}(x))g(\mathcal{A}(x))^T\|_2, \tag{13}$$

where $g(x) \in \mathcal{R}^{n_c \times HW}$ denotes the reshaped feature map, and H, W and n_c represent the feature map height, width, and the number of channels, respectively. The channel-wise self correlations $g(x)g(x)^T$ are also known as the Gram matrix. Additionally, SFIT (Hou & Zheng, 2021) maintains the relationship of outputs between different networks. GDA (Zhou et al., 2022) also relies on BN-based style matching and entropy minimization but further enforces the phase consistency and the feature-level consistency between the original image and the stylized image to preserve the semantic content.

Data selection In addition to synthesizing source samples through data generation or data translation, another family of TTDA methods (Liang et al., 2022; Yang et al., 2023; Ding et al., 2023; Wang et al., 2022; Chen et al., 2022; Yang et al., 2023) selects source-like samples from the target domain as surrogate source data, greatly reducing computational costs. Typically, the whole target domain is divided into two splits, *i.e.*, a labeled subset $\hat{\mathcal{X}}_{tl}$ and an unlabeled subset $\hat{\mathcal{X}}_{tu}$, where the labeled subset acts as the inaccessible source domain. Based on the network outputs of the adapted model in the target domain, SHOT++ (Liang et al., 2022) makes the first attempt towards data selection by selecting low-entropy samples in each class for an extra intra-domain alignment step. Such an adapt-and-divide strategy has been adopted in later works (Ye et al., 2021; Liu et al., 2021; Wang et al., 2022) where the ratio or the number of selected samples per class is always kept same to prevent severe class imbalance. DaC (Zhang et al., 2022) utilizes the maximum softmax probability instead of the entropy criterion. Furthermore, BETA (Yang et al., 2023) constructs a two-component GMM over all the target features to separate the confident subset $\hat{\mathcal{X}}_{tl}$ from the less confident subset $\hat{\mathcal{X}}_{tu}$.

Apart from the adapted target model, a few approaches (Ding et al., 2023; Liu et al., 2022; Huang et al., 2022) utilize the source model to partition the target domain before the intra-domain adaptation step. For each class separately, MTRAN (Huang et al., 2022) selects the low-entropy sample, ProxyMix (Ding et al., 2023) leverages the distance from target features to source class prototypes, and SAB (Liu et al., 2022) adopt the maximum prediction probability. To simulate the source domain more accurately, MTRAN (Huang et al., 2022) further applies the mixup augmentation technique after the dataset partition step. On the other hand, some TTDA methods (Yang et al., 2023; Xia et al., 2021; Ye et al., 2021; Chen et al., 2022; Chu et al., 2022; Tian et al., 2023) do not fix the domain partition but alternately update the domain partition and learn the target model in the adaptation step. For instance, SSNLL (Chen et al., 2022) follows the small loss trick for noisy label learning and assigns samples with small loss to the labeled subset at the beginning of each epoch. On top of the global division, BAIT (Yang et al., 2023) and SFDA-KTMA (Liu et al., 2021) split each mini-batch into two sets based on the criterion of entropy ranking, while D-MCD (Chu et al., 2022) employs the classifier determinacy disparity and the agreement between different self-labeling strategies.

Feature estimation In contrast to source data synthesis, previous works (Qiu et al., 2021; Tian et al., 2022; Ding et al., 2022) provide a cost-effective alternative by simulating the source features. MAS^3 (Stan & Rostami, 2021) and LDAuCID (Rostami, 2021) require learning a GMM over the source features before model adaptation, which may not hold in real-world scenarios. Instead, VDM-DA (Tian et al., 2022) constructs a proxy source domain by randomly sampling features from the following GMM,

$$p_v(z) = \sum\nolimits_{c=1}^{C} \pi_c \, \mathcal{N}(z|\mu_c, \Sigma_c), \tag{14}$$

where z denotes the virtual domain feature, and $p_v(z)$ is the distribution of the virtual domain in the feature space. For each Gaussian component, $\pi_c \geq 0$ represents the mixing coefficient satisfying $\sum_c \pi_c = 1$, and μ_c, Σ_c represent the mean and the covariance matrix, respectively. Specifically, μ_c is approximated by the L_2-normalized class prototype (Chen et al., 2018) that corresponds to the c-th row of weights in the source classifier, and a class-agnostic covariance matrix is heuristically determined by pairwise distances among different class prototypes. To incorporate relevant knowledge from the target domain, SFDA-DE (Ding et al., 2022) further selects confident pseudo-labeled target samples and re-estimates the mean and covariance over these source-like samples as an alternative. In contrast, CPGA (Qiu et al., 2021) trains a prototype generator from conditional noises to generate multiple avatar feature prototypes for each class, encouraging that class prototypes are intra-class compact and inter-class separated.

Virtual domain alignment Once the source distribution is estimated, it is essential to seek virtual domain alignment between the proxy source domain and the target domain for knowledge transfer. We review a variety of virtual domain alignment techniques as follows. Firstly, SHOT++ (Liang et al., 2022) and ProxyMix (Ding et al., 2023) follow a classic semi-supervised approach, MixMatch (Berthelot et al., 2019), to bridge the domain gap. Secondly, SDDA (Kurmi et al., 2021) adopts the widely-used domain adversarial alignment technique (Ganin & Lempitsky, 2015) that is formally written as:

$$\begin{aligned} &\min_{\theta_H} \max_{\theta_D} \; \mathbb{E}_{x_t \in \mathcal{X}_p}[\log D(H(x_t))] \\ &+\mathbb{E}_{x_t \in \mathcal{X}_t}[\log(1 - D(H(x_t)))], \end{aligned} \tag{15}$$

where H and D respectively represent the feature encoder and the binary domain discriminator, and $\mathcal{X}_p$ denotes the proxy source domain. Due to its simplicity, the domain adversarial training strategy has also been utilized in the following works (Liu et al., 2021; Ye et al., 2021; Tian et al., 2022). Besides, a certain number of following methods (Nayak et al., 2022; Yan et al., 2021; Stan & Rostami, 2021) further employ advanced domain adversarial training strategies to achieve better adaptation. Thirdly, BAIT (Yang et al., 2023) leverages the maximum classifier discrepancy (Saito et al., 2018) between two classifiers' outputs in an adversarial manner to achieve feature alignment, which has been followed by Tian et al. (2023), Chu et al. (2022). Fourthly, some TTDA methods (Ding et al., 2022; Zhang et al., 2022; Liu et al., 2022) explore the maximum mean discrepancy (MMD) (Gretton et al., 2012) and propose various conditional variants to reduce the difference of features across domains. In addition, features from different domains could be also aligned through contrastive learning between source prototypes and target samples (Qiu et al., 2021; Zhang et al., 2022). To model the instance-level alignment, MTRAN (Huang et al., 2022) reduces the difference between features from the target data and its corresponding variant in the virtual source domain.

3.2.5 Self-Supervised Learning

Self-supervised learning is a learning paradigm tailored to learn feature representation from unlabeled data based on pretext tasks (Gidaris et al., 2018; Caron et al., 2018, 2020; He et al., 2020; Chen et al., 2020). As mentioned earlier, the centroid-based pseudo labels are similar to the learning manner of DeepCluster (Caron et al., 2018). Inspired by rotation prediction (Gidaris et al., 2018), SHOT++ (Liang et al., 2022) further comes up with a relative rotation prediction task and introduces an additional 4-way classification head

during adaptation. Besides, OnTA (Wang et al., 2021) and CluP (Conti et al., 2022) exploit the self-supervised learning frameworks (He et al., 2020; Caron et al., 2020) for learning discriminative features as initialization, respectively. TTT++ (Liu et al., 2021) learns an extra self-supervised branch using contrastive learning (Chen et al., 2020) in the source model, which facilitates the adaptation in the target domain with the same objective. FedICON (Tan et al., 2023) leverages unsupervised contrastive learning to guide the model to smoothly generalize to test data under intra-client heterogeneity. Recently, StickerDA (Kundu et al., 2022) designs three self-supervised objectives such as sticker location, and optimizes the sticker intervention-based pretext task with the auxiliary classification head in both the source training and target adaptation phases.

Remarks In addition, some remaining TTDA methods have not been covered in the previous discussions. PCT (Tanwisuth et al., 2021) and POUF (Tanwisuth et al., 2023) treat the weights in the classifier layer as source prototypes, and develop an optimal transport-based feature alignment strategy between target features and source prototypes. Besides, target prototypes could also be considered representative labeled data, and such a prototypical augmentation helps correct the classifier with pseudo-labeling (Xiong et al., 2022). LA-VAE (Yang et al., 2021) exploits the variational autoencoder to achieve latent feature alignment. In addition, the meta-learning mechanism is adopted in a few studies (Wang et al., 2021; Bohdal et al., 2022) for the TTDA problem. A recent work (Naik et al., 2023) even generates common sense rules and adapts models to the target domain to reduce rule violations.

3.3 Learning Scenarios of TTDA Algorithms

Closed-set *v.s.* Open-set Most existing TTDA methods focus on a closed-set scenario, *i.e.*, $\mathcal{C}_s = \mathcal{C}_t$, and some TTDA algorithms (Liang et al., 2020; Huang et al., 2021) are also validated in a relaxed partial-set setting (Liang et al., 2020), *i.e.*, $\mathcal{C}_t \subset \mathcal{C}_s$. However, several TTDA works (Liang et al., 2020; Kundu et al., 2020; Feng et al., 2021) consider the open-set learning scenario where the target label space $\mathcal{C}_t$ subsumes the source label space $\mathcal{C}_s$. To allow more flexibility, open-partial-set domain adaptation (You et al., 2019) ($\mathcal{C}_s \backslash \mathcal{C}_t \neq \emptyset, \mathcal{C}_t \backslash \mathcal{C}_s \neq \emptyset$,) is studied in TTDA methods (Kundu et al., 2020; Deng et al., 2021; Yang et al., 2022). Moreover, several recent studies (Liang et al., 2021; Qu et al., 2023) even develop a unified framework for both open-set and open-partial-set scenarios.

Single-source *v.s.* Multi-source To fully transfer knowledge from multiple source models, prior TTDA methods (Liang et al., 2020, 2022; Kundu et al., 2022) extend the single-source TTDA algorithms by combining these adapted models together in the target domain. Besides, a couple of works (Ahmed et al., 2021; Dong et al., 2021) are elaborately designed for adaptation with multiple source models. While each source domain typically shares the same label space with the target domain, UnMSMA-MiFL (Li et al., 2022) considers a union-set multi-source scenario where the union set of the source label spaces is the same as the target label space.

Single-target *v.s.* Multi-target Several TTDA methods (Ahmed et al., 2022; Kumar et al., 2023) also validate the effectiveness of their proposed methods for multi-target domain adaptations where multiple unlabeled target domains exist at the same time. It is worth noting that each target domain may come in a streaming manner, thus the model is successively adapted to different target domains (Rostami, 2021; Panagiotakopoulos et al., 2022).

Unsupervised *v.s.* Semi-supervised Some TTDA methods (Wang et al., 2024; Ma et al., 2022) adapt the source model to the target domain with only a few labeled target samples and adequate unlabeled target samples. In these semi-supervised learning scenarios, the standard classification loss over the labeled data could be readily incorporated to enhance the adaptation performance (Liang et al., 2022; Wang et al., 2024).

White-box *v.s.* Black-box Sharing a model with all the parameters may not be flexible for adjustment if the model turns out to have harmful applications. [3] In this case, the source model is accessible as a black-box module through the cloud application programming interface (API). At an early time, IterLNL (Zhang et al., 2021) treats this black-box TTDA problem as learning with noisy labels, and DINE (Liang et al., 2022) develops several structural regularizations within the knowledge distillation framework. These approaches inspire many recent black-box TTDA works (Sun et al., 2022; Peng et al., 2022; Liu et al., 2022; Yang et al., 2023). Beyond the deep learning framework, several shallow studies (Chidlovskii et al., 2016; Clinchant et al., 2016) focus on the black-box TTDA problem with the target features and their predictions available.

Data *v.s.* Label shifts Different from TTDA methods that narrowly focus on adaptation under data distribution change $p_{\mathcal{S}}(x) \neq p_{\mathcal{T}}(x)$, another family of TTA methods studies label distribution change, $p_{\mathcal{S}}(y) \neq p_{\mathcal{T}}(y)$. For instance, Saerens et al. (2002) propose a well-known prior adaptation framework that adapts an off-the-shelf classifier to a new label distribution with unlabeled data at test time, followed by Lipton et al. (2018), Alexandari et al. (2020). We refer interested readers to relevant literature (Šipka et al., 2022). A few methods such as ISFDA (Li et al., 2021) and APA (Sun et al., 2023) pay attention to the class-imbalanced TTDA scenario where both data and label shifts are present.

[3] https://openai.com/blog/openai-api/

Active TTDA To improve the limited performance gains, MHPL (Wang et al., 2022) introduces a new setting, active TTDA, where a few target data can be selected to be labeled by human annotators. This active setting is also studied by other methods (Li et al., 2022; Kothandaraman et al., 2023), and the key point lies in how to select valuable target samples for labeling.

Miscellaneous TTDA scenarios In addition, researchers also focus on other aspects of TTDA, *e.g.*, the robustness against adversarial attacks (Agarwal et al., 2022), the forgetting of source knowledge (Yang et al., 2021; Liu et al., 2023), and the vulnerability to membership inference attack (An et al., 2022) and image-agnostic attacks (*e.g.*, blended backdoor attack) (Sheng et al., 2023).

4 Test-Time Batch Adaptation

During the testing phase, it is possible that there may exist a single instance or instances from different distributions. This situation necessitates the development of techniques that can adapt off-the-shelf models to individual instances. To be concise, we refer to this learning scheme as *test-time instance adaptation* (*a.k.a.*, standard test-time training Sun et al. 2020 and one-sample generalization D'Innocente et al. 2019), which can be viewed as a special case of test-time domain adaptation ($n_t = 1$).

4.1 Problem Definition

Definition 3 (*Test-Time Instance Adaptation, TTIA*) Given a classifier $f_{\mathcal{S}}$ learned on the source domain $\mathcal{D}_{\mathcal{S}}$, and an unlabeled target instance $x_t \in \mathcal{D}_{\mathcal{T}}$ under distribution shift, *test-time instance adaptation* aims to leverage the labeled knowledge implied in $f_{\mathcal{S}}$ to infer the label of x_t adaptively.

To the best of our knowledge, the concept *test-time adaptation* is first introduced by Wegmann et al. (1998) in 1998, where the speaker-independent acoustic model is adapted to a new speaker with unlabeled data at test time. However, this differs from the definition of *test-time instance adaptation* mentioned earlier, as it involves using a few instances instead of a single instance for personalized adaptation. This scenario is frequently encountered in real-world applications, such as in single-image models that are tested on real-time video data (Brahmbhatt et al., 2018; Azimi et al., 2022). To avoid ambiguity, we further introduce a generalized learning scheme, *test-time batch adaptation*, and give its definition as follows.

Definition 4 (*Test-Time Batch Adaptation, TTBA*) Given a classifier $f_{\mathcal{S}}$ learned on the source domain $\mathcal{D}_{\mathcal{S}}$, and a mini-batch of unlabeled target instances $\{x_t^1, x_t^2, \cdots, x_t^B\} (B \geq 1)$ from $\mathcal{D}_{\mathcal{T}}$ under distribution shift, *test-time batch adaptation* aims to leverage the labeled knowledge implied in $f_{\mathcal{S}}$ to infer the label of each instance at the same time.

It is important to acknowledge that the inference of each instance is not independent, but rather influenced by the other instances in the mini-batch. Test-Time Batch Adaptation (TTBA) can be considered a form of TTDA (Liang et al., 2020) when the batch size B is sufficiently large. Conversely, when the batch size B is equal to 1, TTBA degrades to TTIA (Sun et al., 2020). Typically, these schemes assume no access to the source data or the ground-truth labels of data on the target distribution. In the following, we provide a taxonomy of TTBA (including TTIA) algorithms, as well as the learning scenarios (Table 2).

4.2 Taxonomy on TTBA Algorithms

4.2.1 Batch Normalization Calibration

Normalization layers (*e.g.*, batch normalization Ioffe and Szegedy 2015 and layer normalization Ba et al. 2016) are considered essential components of modern neural networks. For example, a batch normalization (BN) layer calculates the mean and variance for each activation over the training data $\mathcal{X}_{\mathcal{S}}$, and normalizes each incoming sample x_s as follows,

$$\hat{x}_s = \gamma \cdot \frac{x_s - \mathbb{E}[X_{\mathcal{S}}]}{\sqrt{\mathbb{V}[X_{\mathcal{S}}] + \epsilon}} + \beta, \tag{16}$$

where γ and β denote the scale and shift parameters (*a.k.a.*, the learnable affine transformation parameters), and ϵ is a small constant introduced for numerical stability. The BN statistics (*i.e.*, the mean $\mu_s = \mathbb{E}[\mathcal{X}_{\mathcal{S}}]$ and variance $\sigma_s^2 = \mathbb{V}[\mathcal{X}_{\mathcal{S}}]$) are typically approximated using EMA over batch-level estimates $\{\hat{\mu}_k, \hat{\sigma}_k^2\}$,

$$\mu_s \leftarrow (1-\rho) \cdot \mu_s + \rho \cdot \hat{\mu}_k, \ \sigma_s^2 \leftarrow (1-\rho) \cdot \sigma_s^2 + \rho \cdot \hat{\sigma}_k^2, \tag{17}$$

where ρ is the momentum, k denotes the training step, and the statistics over the k-th mini-batch $\{x_i\}_{i=1}^{B_s}$ are

$$\hat{\mu}_k = \frac{1}{B_s} \sum\nolimits_i x_i, \ \hat{\sigma}_k^2 = \frac{1}{B_s} \sum\nolimits_i (x_i - \mu_k)^2, \tag{18}$$

where B_s denotes the batch size at training time. During inference, the BN statistics estimated at training time are frozen for each test sample. AdaBN (Li et al., 2017), a seminal work in the DA literature, suggests that the statistics in the BN layers represent domain-specific knowledge. To bridge the domain gap, AdaBN replaces the training BN statistics with new statistics estimated over the entire target domain.

Table 2 A taxonomy on TTBA methods with representative strategies

Families	Representative strategies
BN calibration	PredBN (Nado et al., 2020; Schneider et al., 2020), InstCal (Zou et al., 2022)
Model optimization	TTT (Sun et al., 2020), GeOS (D'Innocente et al., 2019), MEMO (Zhang et al., 2022)
Meta-learning	MLSR (Park et al., 2020), Full-OSHOT (Borlino et al., 2022)
Input adaptation	TPT (Shu et al., 2022), TTA-DAE (Karani et al., 2021)
Dynamic inference	LAME (Boudiaf et al., 2022), EMEA (Wang et al., 2021)

PredBN (Nado et al., 2020), a pioneering TTBA method, substitutes the training BN statistics with those estimated per test batch.

PredBN+ (Schneider et al., 2020) adopts the running averaging strategy for BN statistics during training and suggests mixing the BN statistics per batch with the training statistics $\{\mu_s, \sigma_s^2\}$ as,

$$\begin{aligned}\bar{\mu}_t &= (1-\rho_t)\cdot\mu_s + \rho_t\cdot\hat{\mu}_t,\ \bar{\sigma}_t^2 \\ &= (1-\rho_t)\cdot\sigma_s^2 + \rho_t\cdot\hat{\sigma}_t^2,\end{aligned} \tag{19}$$

where the test statistics $\{\hat{\mu}_t, \hat{\sigma}_t^2\}$ are estimated via Eq. (18), and the hyper-parameter ρ_t controls the trade-off between training and estimated test statistics. Moreover, TTN (Lim et al., 2023) presents an alternative solution that calibrates the estimation of the variance as follows,

$$\bar{\sigma}_t^2 = (1-\rho_t)\cdot\sigma_s^2 + \rho_t\cdot\hat{\sigma}_t^2 + \rho_t(1-\rho_t)(\hat{\mu}_t-\mu_s)^2. \tag{20}$$

Instead of using the same value for different BN layers, TTN optimizes the interpolating weight ρ_t during the post-training phase using labeled source data. Alternatively, DN (Zhou et al., 2023) proposes subtracting the mean of embeddings within each mini-batch before inference.

Typically, methods that rectify BN statistics may suffer from limitations when the batch size B is small, particularly when $B = 1$. SaN (Bahmani et al., 2022) directly attempts to mix instance normalization (IN) (Ulyanov et al., 2016) statistics estimated per instance with the training BN statistics. Instead of manually specifying a fixed value at test time, InstCal (Zou et al., 2022) introduces an additional module during training to learn the interpolating weight between IN and BN statistics, allowing the network to dynamically adjust the importance of training statistics for each test instance. By contrast, AugBN (Khurana et al., 2021) expands a single instance to a batch of instances using random augmentation, then estimates the BN statistics using the weighted average over these augmented instances.

4.2.2 Model Optimization

Another family of TTBA methods involves adjusting the parameters of a pre-trained model for each unlabeled test instance (batch). These methods are generally divided into two main categories: (1) training with auxiliary tasks (D'Innocente et al., 2019; Sun et al., 2020; D'Innocente et al., 2020), which introduces an additional self-supervised learning task in the primary task during both training and test phases, and (2) fine-tuning with unsupervised objectives (Wang et al., 2019; Zhang et al., 2022; Reddy et al., 2022), which elaborately designs a task-specific objective for updating the pre-trained model.

Training with auxiliary tasks Motivated by prior works (Carlucci et al., 2019; Sun et al., 2019) in which incorporating self-supervision with supervised learning in a unified multi-task framework enhances adaptation and generalization, TTT (Sun et al., 2020) and OSHOT (D'Innocente et al., 2020) are two pioneering works that leverage the same self-supervised learning (SSL) task at both training and test phases, to implicitly align features from the training domain and the test instance. Specifically, they adopt a common multi-task architecture, comprising the primary classification head $h_c(\cdot;\theta_c)$, the SSL head $h_s(\cdot;\theta_s)$, and the shared feature encoder $f_e(\cdot;\theta_e)$. The following joint objective of TTT or OSHOT is optimized at the training stage,

$$\begin{aligned}\theta_e^*, \theta_c^*, \theta_s^* &= \arg\min_{\theta_e,\theta_c,\theta_s} \sum_{i=1}^{n_s} \mathcal{L}_{pri}(x_i, y_i; \theta_e, \theta_c) \\ &\quad + \mathcal{L}_{ssl}(x_i; \theta_e, \theta_s),\end{aligned} \tag{21}$$

where $\mathcal{L}_{pri}$ denotes the primary objective (*e.g.*, cross-entropy for classification tasks), and $\mathcal{L}_{ssl}$ denotes the auxiliary SSL objective (*e.g.*, rotation prediction Gidaris et al. 2018 and solving jigsaw puzzles Carlucci et al. 2019). For each test instance x_t, TTT (Sun et al., 2020) first adjusts the feature encoder $f_e(\cdot;\theta_e)$ by optimizing the SSL objective,

$$\theta_e(x_t) = \arg\min_{\theta_e} \mathcal{L}_{ssl}(x_t; \theta_s^*, \theta_e), \tag{22}$$

then obtains the prediction with the adjusted model as $\hat{y} = h_c(f_e(x;\theta_e(x_t)); \theta_c^*)$. By contrast, OSHOT (D'Innocente et al., 2020) modifies the parameters of both the feature encoder and the SSL head according to the SSL objective at test time. Generally, many follow-up methods adopt the same auxiliary

training strategy by developing various self-supervisions for different applications (Zhang et al., 2020; Hansen et al., 2021; Gandelsman et al., 2022). Among them, TTT-MAE (Gandelsman et al., 2022) is a recent extension of TTT that utilizes the transformer backbone and replaces the self-supervision with masked autoencoders (He et al., 2022).

To increase the dependency between the primary task and the auxiliary task, GeOS (D'Innocente et al., 2019) further adds the features of the SSL head to the primary head. SR-TTT (Lyu et al., 2022) does not follow the Y-shaped architecture but instead utilizes an explicit connection between the primary task and the auxiliary task. Specifically, SR-TTT takes the output of the primary task as the input of the auxiliary task. TTCP (Sarkar et al., 2022) follows the same pipeline as TTT, but it leverages a test-time prediction ensemble strategy by identifying augmented samples that the SSL head could correctly classify.

Training-agnostic fine-tuning To avoid modifying training with auxiliary tasks in the source domain, the other methods focus on developing unsupervised objectives solely for optimizing the model at test time. DIEM (Wang et al., 2019) proposes a selective entropy minimization objective for pixel-level semantic segmentation, while MALL (Reddy et al., 2022) enforces edge consistency prior through a weighted normalized cut loss. Besides, MEMO (Zhang et al., 2022) optimizes the entropy of the averaged prediction over multiple random augmentations of the input sample. PromptAlign (Samadh et al., 2023) additionally handles the train-test distribution shift by matching the mean and variances of the test sample and the source dataset statistics. TTAS (Bateson et al., 2022) further develops a class-weighted entropy objective, while SUTA (Lin et al., 2022) additionally incorporates minimum class confusion to reduce the uncertainty. A recent work (Zhao et al., 2024) develops a reinforcement learning approach that updates the model parameters via policy gradient to maximize the expected reward.

Self-supervised consistency regularization under various input variations is also favorable in customizing the pre-trained model for each test input (Liu et al., 2022; Jin et al., 2023). In particular, SCIO (Kan et al., 2022) develops a self-constrained optimization method to learn the coherent spatial structure. While adapting image models to a video input (Brahmbhatt et al., 2018; Li et al., 2020), ensuring temporal consistency between adjacent frames is a crucial aspect of the unsupervised learning objective. Many other methods directly update the model with the unlabeled objectives tailored to specific tasks, *e.g.*, image matching (Hong & Kim, 2021), image denoising (Mohan et al., 2021), generative modeling (Bau et al., 2019), and style transfer (Kim et al., 2024). In addition, the model could be adapted to each instance by utilizing the generated data at test time. As an illustration, TTL-EQA (Banerjee et al., 2021) generates numerous synthetic question-answer pairs and subsequently leverages them to infer answers in the given context. ZSSR (Shocher et al., 2018) trains a super-resolution network using solely down-sampled examples extracted from the test image itself.

4.2.3 Meta-Learning

MAML (Finn et al., 2017), a notable example of meta-learning (Hospedales et al., 2021), learns a meta-model that can be quickly adapted to perform well on a new task using a small number of samples and gradient steps. Such a learning paradigm is typically well-suited for test-time adaptation, where we can update the meta-model using an unlabeled objective over a few test data. There exist two distinct categories: backward propagation (Park et al., 2020; Borlino et al., 2022), and forward propagation (Dubey et al., 2021; Kim et al., 2022). The latter category does not alter the trained model but includes the instance-specific information in the dynamical neural network.

Backward propagation Inspired by the pioneering work (Shocher et al., 2018), MLSR (Park et al., 2020) develops a meta-learning method based on MAML for single-image super-resolution. Concretely, the meta-objective *w.r.t.* the network parameter θ is shown as,

$$\min_{\theta} \sum_i \mathcal{L}(\mathrm{LR}_i, \mathrm{HR}_i; \theta - \alpha \nabla_\theta \mathcal{L}(\mathrm{LR}_i \downarrow, \mathrm{LR}_i; \theta)), \tag{23}$$

where $\mathcal{L}(A, B; \theta) = \|f_\theta(A) - B\|_2^2$ is the loss function, α is the learning rate of gradient descent, and $\mathrm{LR}_i \downarrow$ denotes the down-scaled version of the low-resolution input in the paired trained data $(\mathrm{LR}_i, \mathrm{HR}_i)$. At inference time, MLSR first adapts the meta-learned network to the low-resolution test image and its down-sized image (LR$\downarrow$) using the parameter θ^* learned in Eq. (23) as initialization,

$$\theta_t \leftarrow \theta^* - \alpha \nabla_\theta \mathcal{L}(\mathrm{LR} \downarrow, \mathrm{LR}; \theta^*), \tag{24}$$

then generates the high-resolution (HR) image as $f_{\theta_t}(\mathrm{LR})$. Such a meta-learning mechanism based on self-supervised learning has been utilized by follow-up methods (Chi et al., 2021; Liu et al., 2022; Min et al., 2023). Among them, MetaVFI (Choi et al., 2021) further introduces self-supervised cycle consistency for video frame interpolation.

As an alternative, Full-OSHOT (Borlino et al., 2022) proposes a meta-auxiliary learning approach that optimizes the shared encoder with an inner auxiliary task, providing a better initialization for the subsequent primary task:

$$\min_{\theta_e, \theta_c} \sum_i \mathcal{L}_{pri}(x_i, y_i; \theta_e - \alpha \nabla_{\theta_e} \mathcal{L}_{ssl}(x_i; \theta_e, \theta_s), \theta_c), \tag{25}$$

and the definitions of variables are the same as OSHOT (D'Innocente et al., 2020) in Eq. (21). After the meta-training phase, the parameters (θ_e, θ_s) are updated for each test sample according to the auxiliary self-supervised objective. This learning paradigm is also known as meta-tailoring (Alet et al., 2021), where $\mathcal{L}_{ssl}$ in the inner loop affects the optimization of $\mathcal{L}_{pri}$ in the outer loop. Subsequent methods exploit various self-supervisions in the inner loop, including contrastive learning (Alet et al., 2021) and reconstruction (Sain et al., 2022; Liu et al., 2023).

Forward propagation Apart from the shared encoder $f_e(\theta_e)$ above, several other meta-learning methods exploit the normalization statistics (Zhang et al., 2021; Bao et al., 2023) or domain prototypes (Dubey et al., 2021; Kim et al., 2022) from the inner loop, allowing backward-free adaptation at inference time. Besides, some works incorporate extra meta-adjusters (Sun et al., 2022) or learnable prompts (Ben-David et al., 2022), by taking the instance embedding as input, to dynamically generate a small subset of parameters in the network, which are optimized at the training phase. DSON (Seo et al., 2020) proposes to fuse IN with BN statistics by linearly interpolating the means and variances, incorporating the instance-specific information in the trained model. Following another popular meta-learning framework (Li et al., 2019), SSGen (Xiao et al., 2022) suggests episodically dividing the training data into meta-train and meta-test to learn the meta-model, which is subsequently applied to the entire training data for final test-time inference. It is also employed by Xu et al. (2022), Segu et al. (2023) where multiple source domains are involved during training.

4.2.4 Input Adaptation

In contrast to model-level optimization, which updates pre-trained models for input data, another line of TTBA methods focuses on changing input data for pre-trained models (Karani et al., 2021; Zhao et al., 2022; Gao et al., 2023). For example, TPT (Shu et al., 2022) freezes the pre-trained multimodal model and only learns the extra text prompt based on the marginal entropy of each instance. Another approach, CVP (Tsai et al., 2023), optimizes the convolutional visual prompts in the input under the guidance of a self-supervised contrastive learning objective.

TTA-AE (He et al., 2021) additionally learns a set of auto-encoders in each layer of the trained model at training time. It is posited that unseen inputs have larger reconstruction errors than seen inputs, thus a set of domain adaptors is introduced at test time to minimize the reconstruction loss. Similarly, TTA-DAE (Karani et al., 2021) only learns an image-to-image translator (*a.k.a.*, input adaptor) for each input so that the frozen training-time denoising auto-encoder could well reconstruct the network output. TTO-AE (Li et al., 2022) follows the Y-shaped architecture of TTT and optimizes both the shared encoder and the additional input adaptor to minimize reconstruction errors in both heads. Instead of auxiliary auto-encoders, AdvTTT (Valvano et al., 2022) leverages a discriminator that is adversarially trained to distinguish real from predicted network outputs, so that the prediction output for each adapted test input satisfies the adversarial output prior.

OST (Termöhlen et al., 2021) proposes mapping the target input onto the source data manifold using Fourier style transfer (Yang & Soatto, 2020), serving as a pre-processor to the primary network. By contrast, TAF-Cal (Zhao et al., 2022) further utilizes the average amplitude feature over the training data to perform Fourier style calibration (Yang & Soatto, 2020) at both training and test phases, bridging the gap between training and test data. It is noteworthy that imposing a data manifold constraint (Pandey et al., 2021; Sarkar et al., 2022; Gao et al., 2023; Xiao et al., 2023) can aid in achieving better alignment between the test data and unseen training data. Specifically, ITTP (Pandey et al., 2021) trains a generative model over source features with target features projected onto points in the source feature manifold for final inference. DDA (Gao et al., 2023) exploits the generative diffusion model for target data, while ESA (Xiao et al., 2023) updates the target feature by energy minimization through Langevin dynamics.

In addition to achieving improved recognition results against domain shifts, a certain number of TTBA methods also explore input adaptation for the purpose of test-time adversarial defense (Shi et al., 2021; Yoon et al., 2021; Mao et al., 2021; Alfarra et al., 2022). Among them, Anti-Adv (Alfarra et al., 2022) perturbs the test input to maximize the classifier's prediction confidence. Besides, SOAP (Shi et al., 2021) leverages self-supervisions like rotation prediction at both training and test phases and purifies adversarial test examples based on self-supervision only. SSRA (Mao et al., 2021) only exploits the self-supervised consistency under different augmentations at test time to remove adversarial noises in the attacked data.

4.2.5 Dynamic Inference

LAME (Boudiaf et al., 2022) utilizes neighbor consistency to enforce consistent assignments on neighboring points in the feature space, without modifying the pre-trained model. Upon multiple pre-trained models learned from the source data, a few works (Wang et al., 2021; Zhang et al., 2023) learn the weights for each model, without making any changes to the models themselves. For example, EMEA (Wang et al., 2021) employs entropy minimization to update the ensemble coefficients before each model. GPR (Jain & Learned-Miller, 2011) is one of the early works that only adjusts the network predictions instead of the pre-trained model. In particular, it

bootstraps the more difficult faces in an image from the more easily detected faces and adopts Gaussian process regression to encourage smooth predictions for similar patches.

4.3 Learning Scenarios of TTBA Algorithms

Instance *v.s.* Batch As defined above, test-time adaptation could be divided into two cases: instance adaptation (Sun et al., 2020; Zhang et al., 2022) and batch adaptation (Schneider et al., 2020; Brahmbhatt et al., 2018), according to whether a single instance or a batch of instances exist at test time.
Single *v.s.* Multiple In contrast to vanilla test-time adaptation that utilizes the pre-trained model from one single source domain, some works (*e.g.*, D'Innocente et al. 2019, Pandey et al. 2021, Wang et al. 2021, Xiao et al. 2022, Zhao et al. 2022, Xiao et al. 2023, Zhang et al. 2023) are interested in domain generalization problems where multiple source domains exist.
White-box *v.s.* Black-box A majority of TTBA methods focus on adapting white-box models to test instances, while some other works (*e.g.*, Jain and Learned-Miller 2011, Chen et al. 2019, Zhang et al. 2023) do not have access to the parameters of the pre-trained model (black-box) and instead adjust the predictions according to generic structural constraints.
Customized *v.s.* On-the-fly Most existing TTA methods require training one or more customized models in the source domain, *e.g.*, TTT (Sun et al., 2020) employs a Y-shaped architecture with an auxiliary head. However, it may be not allowed to train the source model in a customized manner for some real-world applications. Other works (Zhang et al., 2022; Alfarra et al., 2022) do not rely on customized training in the source domain but develop flexible techniques for adaptation with on-the-fly models.

5 Online Test-Time Adaptation

Previously, we have considered various test-time adaptation scenarios where pre-trained source models are adapted to a domain (Liang et al., 2020; Li et al., 2020), a mini-batch (Schneider et al., 2020; Zhang et al., 2021), or even a single instance (Sun et al., 2020; Zhang et al., 2022) at test time. However, offline test-time adaptation typically requires a certain number of samples to form a mini-batch or a domain, which may be infeasible for streaming data scenarios where data arrives continuously and in a sequential manner. To reuse past knowledge like online learning, TTT (Sun et al., 2020) employs an online variant that does not optimize the model episodically for each input but instead retains the optimized model for the last input.

5.1 Problem Definition

Definition 5 (Online Test-Time Adaptation, OTTA) Given a well-trained classifier $f_{\mathcal{S}}$ on the source domain $\mathcal{D}_{\mathcal{S}}$ and a sequence of unlabeled mini-batches $\{\mathcal{B}_1, \mathcal{B}_2, \cdots\}$, *online test-time adaptation* aims to leverage the labeled knowledge implied in $f_{\mathcal{S}}$ to infer labels of samples in $\mathcal{B}_i$ under distribution shift, in an online manner. In other words, the knowledge learned in previously seen mini-batches could be accumulated for adaptation to the current mini-batch.

The above definition corresponds to the problem addressed in Tent (Wang et al., 2021), where multiple mini-batches are sampled from a new data distribution that is distinct from the source data distribution. Besides, it also encompasses the online test-time instance adaptation problem, as introduced in TTT-Online (Sun et al., 2020) when the batch size equals 1. However, samples at test time may come from a variety of different distributions, leading to new challenges such as error accumulation and catastrophic forgetting. To address this issue, CoTTA (Wang et al., 2022) and EATA (Niu et al., 2022) investigate the continual test-time adaptation problem that adapts the pre-trained source model to the continually changing test data. Such a non-stationary adaptation problem could be also viewed as a special case of the definition above, when each mini-batch may come from a different distribution (Table 3).

5.2 Taxonomy on OTTA Algorithms

5.2.1 Batch Normalization Calibration

As noted in the previous section, normalization layers such as batch normalization (BN) (Ioffe & Szegedy, 2015) are commonly employed in modern neural networks. Typically, BN layers can encode domain-specific knowledge into normalization statistics (Li et al., 2017). A recent work (Niu et al., 2023) further investigates the effects of different normalization layers under the test-time adaptation setting. In the following, we mainly focus on the BN layer due to its widespread usage in existing methods.

Tent (Wang et al., 2021) and RNCR (Hu et al., 2021) propose replacing the fixed BN statistics (*i.e.*, mean and variance $\{\mu_s, \sigma_s^2\}$) in the pre-trained model with the estimated ones $\{\hat{\mu}_t, \hat{\sigma}_t^2\}$ from the t-th test batch. CD-TTA (Song et al., 2022) develops a switchable mechanism that selects the most similar one from multiple BN branches in the pre-trained model using the Bhattacharya distance. Besides, Core (You et al., 2021) calibrates the BN statistics by interpolating between the fixed source statistics and the estimated ones at test time, namely, $\mu_t = \rho\hat{\mu}_t + (1-\rho)\mu_s$, $\sigma_t = \rho\hat{\sigma}_t + (1-\rho)\sigma_s$, where $\rho \in [0, 1]$ is a momentum hyper-parameter.

Table 3 A taxonomy on OTTA methods with representative strategies

Families	Representative Strategies
BN calibration	DUA (Mirza et al., 2022), DELTA (Zhao et al., 2023)
Entropy minimization	Tent (Wang et al., 2021), SAR (Niu et al., 2023)
Pseudo-labeling	T3A (Iwasawa & Matsuo, 2021), TAST (Jang et al., 2023)
Consistency regularization	CFA (Kojima et al., 2022), PETAL (Brahma & Rai, 2023)
Anti-forgetting regularization	CoTTA (Wang et al., 2022), EATA (Niu et al., 2022)

Similar to the running average estimation of BN statistics during training, ONDA (Mancini et al., 2018) proposes initializing the BN statistics $\{\mu_0, \sigma_0^2\}$ as $\{\mu_s, \sigma_s^2\}$ and updating them for the t-th test batch,

$$\begin{aligned} \mu_t &= \rho\hat{\mu}_t + (1-\rho)\mu_{t-1}, \\ \sigma_t^2 &= \rho\hat{\sigma}_t^2 + (1-\rho)\frac{n_t}{n_t-1}\sigma_{t-1}^2, \end{aligned} \tag{26}$$

where n_t denotes the number of samples in the batch, and ρ is a momentum hyper-parameter. Instead of a constant value for ρ, MECTA (Hong et al., 2023) considers a heuristic weight through computing the distance between $\{\mu_{t-1}, \sigma_{t-1}\}$ and $\{\hat{\mu}_t, \hat{\sigma}_t\}$. EDTN (Wang et al., 2024) further introduces a straightforward layer-wise strategy to set the momentum hyper-parameters for different layers.

To decouple the gradient backpropagation and the selection of BN statistics, GpreBN (Yang et al., 2022) and DELTA (Zhao et al., 2023) adopt the following reformulation of batch re-normalization (Ioffe, 2017),

$$\hat{x}_t = \gamma \cdot \frac{\frac{x_t - \hat{\mu}_t}{\hat{\sigma}_t} \cdot sg(\hat{\sigma}_t) + sg(\hat{\mu}_t) - \mu}{\sigma} + \beta, \tag{27}$$

where $sg(\cdot)$ denotes the stop-gradient operation, and $\{\gamma, \beta\}$ are the affine parameters in the BN layer. To obtain stable BN statistics $\{\mu, \sigma^2\}$, these methods utilize the test-time dataset-level running statistics via the moving average like Eq. (26).

For online adaptation with a single sample, MixNorm (Hu et al., 2021) mixes the estimated IN statistics with the exponential moving average BN statistics at test time. On the other hand, DUA (Mirza et al., 2022) adopts a decay strategy for the weighting hyper-parameter ρ and forms a small batch from a single image to stabilize the online adaptation process. To obtain more accurate estimates of test-time statistics, NOTE (Gong et al., 2022) maintains a class-balanced memory bank that is utilized to update the BN statistics using an exponential moving average. Additionally, NOTE proposes a selective mixing strategy that only calibrates the BN statistics for detected out-of-distribution samples. TN-SIB (Zhang et al., 2022) also leverages a memory bank that provides samples with similar styles to the test sample, to accurately estimate BN statistics.

5.2.2 Entropy Minimization

Entropy minimization is a widely used technique to handle unlabeled data. A pioneering approach, Tent (Wang et al., 2021), proposes minimizing the mean entropy over the test batch to update the affine parameters $\{\gamma, \beta\}$ of BN layers in the pre-trained model, followed by various subsequent methods (Gong et al., 2022; Yang et al., 2022). Notably, VMP (Jing et al., 2022) reformulates Tent in a probabilistic framework by introducing perturbations into the model parameters by variational Bayesian inference. Several other methods (Tang et al., 2023; Yi et al., 2023) also focus on minimizing the entropy at test time but utilize different combinations of learnable parameters. BACS (Zhou & Levine, 2021) incorporates the entropy regularization for unlabeled data in the approximate Bayesian inference algorithm, and samples multiple model parameters to obtain the marginal probability for each sample. In addition, TTA-PR (Sivaprasad & Fleuret, 2021) proposes minimizing the average entropy of predictions under different augmentations. FEDTHE+ (Jiang & Lin, 2023) employs the same adaptation scheme as MEMO (Zhang et al., 2022) that minimizes the entropy of the average prediction over different augmentations.

To avoid overfitting to non-reliable and redundant test data, EATA (Niu et al., 2022) develops a sample-efficient entropy minimization strategy that identifies samples with lower entropy values than the pre-defined threshold for model updates, which is also adopted by follow-up methods (Song et al., 2023; Niu et al., 2023). CD-TTA (Song et al., 2022) leverages the similarity between feature statistics of the test sample and source running statistics as sample weights, instead of using discrete weights $\{0, 1\}$. Besides, DELTA (Zhao et al., 2023) derives a class-wise re-weighting approach that associates sample weights with corresponding pseudo labels to mitigate bias towards dominant classes.

There exist many alternatives to entropy minimization for adapting models to unlabeled test samples including class confusion minimization (You et al., 2021), batch nuclear-norm maximization (Hu et al., 2021), maximum squares loss (Song et al., 2022), and mutual information maximization (Kingetsu et al., 2022; Choi et al., 2022). In addition, MuSLA (Kingetsu et al., 2022) further considers the virtual adversarial training objective that enforces classifier consistency by adding a small perturbation to each sample. SAR (Niu et al.,

2023) encourages the model to lie in a flat area of the entropy loss surface and optimizes the minimax entropy objective below,

$$\min_{\theta} \max_{\|\Delta_\theta\|_2 \leq \epsilon} \mathcal{H}(x; \theta + \Delta_\theta), \tag{28}$$

where $\mathcal{H}(\cdot)$ denotes the entropy function, and Δ_θ denotes the weight perturbation in a Euclidean ball with radius ϵ. Moreover, a few methods (Kundu et al., 2022; Yang et al., 2023) even employ entropy maximization for specific tasks, for example, AUTO (Yang et al., 2023) performs model updating for unknown samples at test time.

5.2.3 Pseudo-Labeling

Unlike the unidirectional process of entropy minimization, many OTTA methods (Belli et al., 2022; Kingetsu et al., 2022; Boudiaf et al., 2022; Song et al., 2022) adopt pseudo labels generated at test time for model updates. Among them, MM-TTA (Shin et al., 2022) proposes a selective fusion strategy to ensemble predictions from multiple modalities. Besides, DLTTA (Yang et al., 2022) obtains soft pseudo labels by averaging the predictions of its nearest neighbors in a memory bank, and subsequently optimizes the symmetric KL divergence between the model outputs and these pseudo labels. TAST (Jang et al., 2023) proposes a similar approach that reduces the difference between predictions from a prototype-based classifier and a neighbor-based classifier. Notably, SLR+IT (Mummadi et al., 2021) develops a negative log-likelihood ratio loss instead of the commonly used cross-entropy loss, providing non-vanishing gradients for highly confident predictions.

Conjugate-PL (Goyal et al., 2022) presents a way of designing unsupervised objectives for TTA by leveraging the convex conjugate function. The resulting objective resembles self-training with specific soft labels, referred to as conjugate pseudo labels. A recent work (Wang & Wibisono, 2023) theoretically analyzes the difference between hard and conjugate labels under gradient descent for a binary classification problem. Motivated by the idea of negative learning (Kim et al., 2019), ECL (Zeng et al., 2024) further considers complementary labels from the least probable categories. Besides, T3A (Iwasawa & Matsuo, 2021) proposes merely adjusting the classifier layer by computing class prototypes using online unlabeled data and classifying each unlabeled sample based on its distance to these prototypes.

5.2.4 Consistency Regularization

In the classic mean teacher (Tarvainen & Valpola, 2017) framework, the pseudo labels under weak data augmentation obtained by the teacher network are known to be more stable. Built on this framework, RMT (Döbler, 2023) pursues the teacher-student consistency in predictions through a symmetric cross-entropy measure, while OIL (Ye et al., 2022) only exploits highly confident samples during consistency maximization. VDP (Gan et al., 2023) utilizes this framework to update visual domain prompts with the pre-trained model being frozen. Moreover, CoTTA (Wang et al., 2022) further employs multiple augmentations to refine the pseudo labels from the teacher network, which is also applied in other methods (Brahma & Rai, 2023; Tomar et al., 2023; Ma et al., 2023). Inspired by maximum classifier discrepancy (Saito et al., 2018), AdaODM (Zhang & Chen, 2023) proposes minimizing the prediction disagreement between two classifiers at test time to update the feature encoder.

Apart from the model variation above, several methods (Sivaprasad & Fleuret, 2021; Das et al., 2023; Lumentut & Park, 2022; Su et al., 2022; Chen et al., 2023) also enforce the consistency of the corresponding predictions among different augmentations. In particular, SWR-NSP (Choi et al., 2022) introduces an additional nearest source prototype classifier at test time and minimizes the difference between predictions under two different augmentations. Besides, many methods (Guan et al., 2021; Kuznietsov et al., 2022; Kim et al., 2022; Belli et al., 2022; Yi et al., 2023) leverage the temporal coherence for video data and design a temporal consistency objective at test time. For example, TeCo (Yi et al., 2023) encourages adjacent frames to have semantically similar features to increase the robustness against corruption at test time.

In contrast to constraints in the prediction space, FEDTHE+ (Jiang & Lin, 2023) pursues consistency in the feature space. Several other OTTA methods (Wu et al., 2021; Döbler, 2023; Su et al., 2022)) even pursue consistency between test features and source or target prototypes in the feature space. CFA (Kojima et al., 2022) further proposes matching multiple central moments to achieve feature alignment. Furthermore, ACT-MAD (Mirza et al., 2023) performs feature alignment by minimizing the discrepancy between the pre-computed training statistics and the estimates of test statistics. TTAC (Su et al., 2022) calculates the online estimates of feature mean and variance at test time instead. Besides, CAFA (Jung et al., 2023) uses the Mahalanobis distance to achieve low intra-class variance and high inter-class variance for test data.

5.2.5 Anti-forgetting Regularization

Previous studies (Wang et al., 2022; Niu et al., 2022) find that the model optimized by TTA methods suffers from severe performance degradation (named forgetting) on original training samples. To mitigate the forgetting issue, a natural solution is to keep a small subset of training data that is further learned at test time as regularization (Belli et al., 2022; Döbler, 2023; Kuznietsov et al., 2022). PAD (Wu et al., 2021) comes up with an alternative approach that keeps the relative relationship of irrelevant auxiliary data unchanged

after test-time optimization. AUTO (Yang et al., 2023) maintains a memory bank to store easily recognized samples for replay and prevents overfitting towards unknown samples at test time.

Another anti-forgetting solution lies in using merely a few parameters for test-time model optimization. For example, Tent (Wang et al., 2021) only optimizes the affine parameters in the BN layers for test-time adaptation, and AUTO (Yang et al., 2023) updates the last feature block in the pre-trained model. SWR-NSP (Choi et al., 2022) divides the entire model parameters into shift-agnostic and shift-biased parameters and updates the former less and the latter more. Recently, VDP (Gan et al., 2023) fixes the pre-trained model but only optimizes the input prompts during adaptation.

Besides, CoTTA (Wang et al., 2022) proposes a stochastic restoration technique that randomly restores a small number of parameters to the initial weights in the pre-trained model. PETAL (Brahma & Rai, 2023) further selects parameters with smaller gradient norms in the entire model for restoration. By contrast, EATA (Niu et al., 2022) introduces an importance-aware Fisher regularizer to prevent excessive changes in model parameters. The importance is estimated from test samples with generated pseudo labels. SAR (Niu et al., 2023) proposes a sharpness-aware and reliable optimization scheme, which removes samples with large gradients and encourages model weights to lie in a flat minimum. Further, EcoTTA (Song et al., 2023) presents a self-distilled regularization by forcing the output of the test model to be close to that of the pre-trained model.

Remarks There are several other solutions for the OTTA problem, *e.g.*, meta-learning (Zhang et al., 2022; Wu et al., 2023), Hebbian learning (Tang et al., 2023), and adversarial data augmentation (Tomar et al., 2023). TDA (Karmanov et al., 2024) further provides a training-free solution by leveraging a dynamic memory bank that stores pseudo labels and features from previous samples.

5.3 Learning Scenarios of OTTA Algorithms

Stationary *v.s.* Dynamic In contrast to vanilla OTTA (Wang et al., 2021) that assumes the test data comes from a stationary distribution, dynamic OTTA assumes a dynamically changing distribution including continual OTTA (Wang et al., 2022), temporal OTTA (Gong et al., 2022), gradual OTTA (Döbler, 2023), and practical OTTA (Yuan et al., 2023). A recent study (Marsden et al., 2024) delves into the realm of universal OTTA, a more complex setting where both domain non-stationarity and temporal correlation may coexist, with the specific test-time scenario often remaining unknown.

Data *v.s.* Label shifts While the majority of OTTA methods concentrate on shifts in data distribution, some approaches (Yang & Zhou, 2008; Royer & Lampert, 2015; Wu et al., 2021) investigate changes in label distribution. Two interesting cases with online feedback are studied in Royer and Lampert (2015), *i.e.*, online feedback (the correct label is revealed to the system after prediction) and bandit feedback (the decision made by the system is correct or not is revealed).

Other differences between OTTA methods are the same as TTBA, *i.e.*, **instance *v.s.* batch**, **customized *v.s.* on-the-fly**, and **single *v.s.* multiple**.

6 Applications [4]

6.1 Image Classification

The most common application of test-time adaptation is multi-class image classification. Firstly, TTDA methods are commonly evaluated and compared on widely used DA datasets, including Digits, Office, Office-Home, VisDA-C, and DomainNet, as described in previous studies (Liang et al., 2020, 2022; Zhang et al., 2022). Secondly, TTBA and OTTA methods consider natural distribution shifts in object recognition datasets, *e.g.*, corruptions in CIFAR-10-C, CIFAR-100-C, and ImageNet-C, natural renditions in ImageNet-R, misclassified real-world samples in ImageNet-A, and unknown distribution shifts in CIFAR−10.1, as detailed in previous studies (Sun et al., 2020; Schneider et al., 2020; Wang et al., 2021; Zhang et al., 2022). In addition, TTBA and OTTA methods are also evaluated in DG datasets such as VLCS, PACS, and Office-Home, as described in previous studies (D'Innocente et al., 2019; Pandey et al., 2021; Iwasawa & Matsuo, 2021; Gan et al., 2023).

6.2 Semantic Segmentation

Semantic segmentation aims to categorize each pixel of the image into a set of semantic labels, which is a critical module in autonomous driving. Many domain adaptive semantic segmentation datasets, such as GTA5-to-Cityscapes, SYNTHIA-to-Cityscapes, and Cityscapes-to-Cross-City, are commonly adopted to evaluate TTDA methods, as depicted in (Sivaprasad and Fleuret 2021; Liu et al. 2021; Wang et al. 2022). In addition to these datasets, BDD100k, Mapillary, and WildDash2, and IDD are also used to conduct comparisons for TTBA and OTTA methods, as shown in (Zou et al. 2022; Bahmani et al. 2022). OTTA methods further utilize Cityscapes-to-ACDC and Cityscapes-to-Foggy&Rainy Cityscapes for evaluation and comparison, as described in (Wang et al. 2022; Volpi et al. 2022).

[4] A table of commonly used datasets across various TTA applications is also provided in the GitHub repository.

6.3 Object Detection

Object detection is a fundamental computer vision task that involves locating instances of objects in images. While early TTA methods (Jamal et al., 2018; RoyChowdhury et al., 2019) focus on binary tasks such as pedestrian and face detection, lots of current efforts are devoted to generic multi-class object detection. Typically, many domain adaptive object detection tasks including Cityscapes-to-BDD100k, Cityscapes-to-Foggy Cityscapes, KITTI-to-Cityscapes, Sim10k-to-Cityscapes, Pascal-to-Clipart&Watercolor are commonly used by TTDA methods for evaluation and comparison, as detailed in (Li et al. 2021; Huang et al. 2021; Li et al. 2022; Sinha et al. 2023). Additionally, datasets like VOC-to-Social Bikes and VOC-to-AMD are employed to evaluate TTBA methods (D'Innocente et al., 2020; Borlino et al., 2022).

6.4 Beyond Vanilla Object Images

Medical images Medical image analysis is another important downstream field of TTA methods, *e.g.*, medical image classification (Ma et al., 2022; Wang et al., 2022), medical image segmentation (He et al., 2021; Karani et al., 2021), and medical image detection (Liu & Yuan, 2022). Among them, medical segmentation attracts the most attention in this field.
3D point clouds Nowadays, 3D sensors have become a crucial component of perception systems. Many tasks for 2D images have been adapted for LiDAR point clouds, such as 3D object classification (Tian et al., 2022), 3D semantic segmentation (Saltori et al., 2022), and 3D object detection (Saltori et al., 2020).
Videos As mentioned above, TTBA and OTTA methods can address how to efficiently adapt an image model to real-time video data for problems such as depth prediction (Liu et al., 2023) and frame interpolation (Choi et al., 2021). Besides, a few studies investigate the TTDA scheme for other video-based tasks including action recognition (Xu et al., 2022; Huang et al., 2022; Yi et al., 2023; Zeng et al., 2023), optical flow estimation (Ayyoubzadeh et al., 2023) and object segmentation (Bertrand et al., 2023).
Multi-modal data Researchers also develop different TTA methods for various multi-modal data, *e.g.*, RGB and audio (Plananamente et al., 2022), RGB and depth (Ahmed et al., 2022; Shin et al., 2022), RGB and motion (Huang et al., 2022), and image-text pairs (Wen et al., 2024). Furthermore, the development of multi-modal pre-trained models such as CLIP (Radford et al., 2021) enables image classification through image-to-text matching, gaining popularity among recent TTA methods (Samadh et al., 2023; Zhou et al., 2023; Ma et al., 2023; Zhao et al., 2024).
Face and body data Facial data is also an important application of TTA methods, such as face recognition (Zhang et al., 2022), face anti-spoofing (Wang et al., 2021; Liu et al., 2022; Zhou et al., 2022), and expression recognition (Conti et al., 2022). For body data, TTA methods also pay attention to tasks such as pose estimation (Zhang et al., 2020; Kan et al., 2022; Ding et al., 2024) and mesh reconstruction (Guan et al., 2021; Li et al., 2020).

6.5 Beyond Vanilla Recognition Problems

Low-level vision TTA methods can be applied to low-level vision problems, *e.g.*, image super-resolution (Park et al., 2020; Deng et al., 2023), image deblurring (Chi et al., 2021), and image dehazing (Liu et al., 2022). Besides, TTA is also introduced to image registration (Zhu et al., 2021; Hong & Kim, 2021), inverse problems (Hussein et al., 2020; Darestani et al., 2022), and quality assessment (Liu et al., 2022).
Retrieval Besides classification problems, TTA can also be applied to kinds of retrieval scenarios, *e.g.*, person re-identification (Wu et al., 2019; Xu et al., 2022), sketch-to-image retrieval (Sain et al., 2022; Paul et al., 2022), image-text matching (Zhou et al., 2023), and fair image retrieval (Kong et al., 2023).
Generative modeling TTA method can also vary the pre-trained generative model for style transfer and data generation (Bau et al., 2019; Kim et al., 2024; Nitzan et al., 2022).
Defense Another interesting application is test-time adversarial defense (Shi et al., 2021; Yoon et al., 2021; Alfarra et al., 2022), which tries to generate robust predictions for possible perturbed samples.

6.6 Natural Language Processing (NLP)

The TTA paradigm is also studied in tasks of the NLP field, such as reading comprehension (Banerjee et al., 2021), question answering (Ye et al., 2022), sentiment analysis (Zhang et al., 2021), entity recognition (Wang et al., 2021), and aspect prediction (Ben-David et al., 2022). In particular, a competition [5] has been launched under data sharing restrictions, comprising two NLP semantic tasks (Laparra et al., 2021): negation detection and time expression recognition.

6.7 Beyond CV and NLP

Graph data For graph data (*e.g.*, social networks), TTA methods are evaluated and compared on either graph classification (Wang et al., 2022) or node classification (Jin et al., 2023).
Speech processing As far, there have been three TTA methods, *i.e.*, audio classification (Boudiaf et al., 2023), speaker verification (Kim et al., 2022) and speech recognition (Lin et al., 2022).

[5] https://competitions.codalab.org/competitions/26152

Miscellaneous signals TTA methods have been also validated on other types of signals, *e.g.*, radar signals (Cao et al., 2021), EEG signals (Lee et al., 2023), and vibration signals (Jiao et al., 2022).

Reinforcement learning Some TTA methods (Hansen et al., 2021; Liu & Fang, 2023) also address the generalization of reinforcement learning policies across different environments.

6.8 Evaluation

As the name suggests, TTA methods should evaluate the performance of test data after test-time optimization immediately. However, there are different protocols for evaluating TTA methods in the field, making a rigorous evaluation protocol important. Firstly, some TTDA works, particularly for domain adaptive semantic segmentation (Sivaprasad & Fleuret, 2021; Wang et al., 2022) and classification on DomainNet, adapt the source model to an unlabeled target set and evaluate the performance on the test set that shares the same distribution as the target set. However, this in principle violates the setting of TTA, although the performance on the test set is always consistent with that of the target set. *We suggest that such SFDA methods report the performance on the target set at the same time.* Secondly, some TTDA works such as BAIT (Yang et al., 2023) offer an online variant, but such online TTDA methods differ from OTTA in that the evaluation is conducted after one full epoch. *We suggest online TTDA methods change the name to "one-epoch TTDA" to avoid confusion with OTTA methods.* Thirdly, for continual TTA methods (Wang et al., 2022; Niu et al., 2022), the evaluation of each mini-batch is conducted before optimization on that mini-batch. This manner differs from the standard evaluation protocol of OTTA (Sun et al., 2020) where optimization is conducted ahead of evaluation. *We suggest that continual TTA methods follow the same protocol as vanilla OTTA methods.*

7 Emerging Trends and Open Problems

7.1 Emerging Trends

Diverse downstream fields Even most existing efforts in the TTA field have been devoted to visual tasks such as image classification and semantic segmentation, a growing number of TTA methods are now focusing on other understanding problems over video data (Xu et al., 2022), multi-modal data (Shin et al., 2022), and 3D point clouds (Saltori et al., 2022), as well as regression problems like pose estimation (Ding et al., 2024).

Open-world adaptation Existing TTA methods always follow the closed-set assumption; however, a growing number of TTDA methods (Liang et al., 2021; Yang et al., 2022; Qu et al., 2023) are beginning to explore model adaptation under an open-set setting. A recent OTTA method (Yang et al., 2023) further focuses on the performance of out-of-distribution detection tasks at test time. Besides, for large distribution shifts, it is challenging to perform effective knowledge transfer by relying solely on unlabeled target data, thus several recent works (Li et al., 2022; Kothandaraman et al., 2023) also introduce active learning to involve humans in the loop.

Memory-efficient continual adaptation In real-world applications, test samples may come from a continually changing environment (Wang et al., 2022; Niu et al., 2022), leading to catastrophic forgetting. To reduce memory consumption while maintaining accuracy, recent works (Song et al., 2023; Hong et al., 2023) propose different memory-friendly OTTA solutions for resource-limited end devices.

On-the-fly adaptation The majority of existing TTA methods require a customized pre-trained model from the source domain, bringing the inconvenience for instant adaptation. Thus, fully test-time adaptation (Wang et al., 2021), which allows adaptation with an on-the-fly model, has attracted increasing attention.

Foundation models Large language models like GPT have attracted widespread attention due to their surprisingly strong ability in various tasks. Given a query to a language model, a recent work (Hardt & Sun, 2024) performs test-time training by fine-tuning the model based on its retrieved nearest neighbors. Over the past two years, there has been a growing number of TTBA methods (Shu et al., 2022; Feng et al., 2023; Samadh et al., 2023; Zhou et al., 2023; Zhao et al., 2024; Yoon et al., 2024) developed that leverage vision-language models, such as CLIP (Radford et al., 2021), to enhance the zero-shot generalization. Meanwhile, some studies have focused on CLIP adaptation under the OTTA scenario (Ma et al., 2023; Karmanov et al., 2024) as well as the TTDA setting (Tanwisuth et al., 2023; Hu et al., 2024). Additionally, several recent studies (Feng et al., 2023; Prabhudesai et al., 2023) have explored leveraging large-scale generative models, such as Stable Diffusion (Rombach et al., 2022), for developing TTA methods.

7.2 Open Problems

Theoretical analysis While most existing works focus on developing effective TTA methods to obtain better empirical performance, the theoretical analysis of when and why TTA works remains an open problem. Several TTA methods have provided theoretical results on specific designs under linear models such as gradient descent with pseudo-labels (Wang & Wibisono, 2023) and auxiliary self-supervision (Sun et al., 2020). One recent work (Gui et al., 2024) conducts an in-depth theoretical analysis based on learning theories and mainly explores how can significant distribution shifts

be effectively addressed under the online TTA setting. We believe that more rigorous analyses, especially on deep learning models, can provide deeper insights and inspire the development of new TTA methods.

Benchmark and validation Recently, several new benchmarks (Yu et al., 2023; Press et al., 2023; Wang et al., 2023) are proposed to fairly evaluate various TTA methods. For example, the vision transformer (ViT) architecture is further employed for online TTA methods in Wang et al. (2023), and a new dataset is developed to testify online TTA methods under continuously changing corruptions (Press et al., 2023). However, as there does not exist a labeled validation set, validation also remains a significant and unsolved issue for TTA methods. As noted in Zhao et al. (2023), evaluations of TTA methods have often been conducted unfairly. Existing studies frequently determine hyper-parameters through grid search on the test data, which is not feasible in real-world applications. To address this issue, a recent benchmark (Yu et al., 2023) has proposed a fixed validation strategy with a predetermined online batch order. It selects the optimal hyper-parameters based on the first one of the adaptation tasks for all the tasks. In the future, a benchmark can be built where a labeled validation set and an unlabeled test set exist at test time, providing a more realistic evaluation scenario for TTA methods.

New applications Tabular data (Borisov et al., 2022) in vectors of heterogeneous features is essential for industrial applications, and time series data (Ragab et al., 2023) is predominant in real-world applications like healthcare and manufacturing. So far, limited prior work has explored TTA in the context of tabular or time series data, despite their importance and prevalence in real-world scenarios. When it comes to adapting to tabular data, deep learning models have generally underperformed compared to tree-based models such as XGBoost and random forests (Shwartz-Ziv & Armon, 2022; Grinsztajn et al., 2022). Therefore, it would be interesting to investigate how TTA methods developed primarily for deep learning models can be applied and perform when used with tree-based models for tabular data scenarios.

Trustworthiness Current TTA methods focus more on robustness under distribution shifts while ignoring other goals of trustworthy machine learning (Eshete, 2021), *e.g.*, fairness, security, privacy, and explainability. Regarding class-wise fairness, the adapted model's performance may vary considerably across different categories in the target domain. However, existing TTA methods have not thoroughly investigated the worst-class accuracy for classification tasks. As for security, in the TTDA setting, the source provider could potentially be a malicious actor who inserts backdoors into the pre-trained model (Sheng et al., 2023). This could enable the attacker to then target the model adapted by the end user using the same embedded backdoor triggers. Furthermore, another important issue with existing TTA methods is their tendency towards overconfidence, which undermines the reliability of their predictions (Kim et al., 2023; Yoon et al., 2024).

8 Conclusion

Learning to adapt a pre-trained model to unlabeled data under distribution shifts is an emerging and critical problem in the field of machine learning. This survey provides a comprehensive review of three related topics: test-time domain adaptation, test-time batch adaptation, and online test-time adaptation. These topics are unified as a broad learning paradigm of test-time adaptation (TTA). For each topic, we first introduce its definition and a new taxonomy of advanced algorithms. Additionally, we provide a review of applications related to test-time adaptation, as well as an outlook of emerging research trends and open problems. We believe that this survey will assist both newcomers and experienced researchers in better understanding the current state of research in TTA under distribution shifts.

Acknowledgements We sincerely thank the editor and anonymous reviewers for their constructive comments on this work. We also thank Lijun Sheng for his valuable feedback on this work. This work was funded by the Beijing Nova Program (No. Z211100002121108), the Young Elite Scientists Sponsorship Program by CAST (2023QNRC001), and the National Natural Science Foundation of China under (No. 62276256).

References

Agarwal, P., Paudel, D. P., Zaech, J.-N., & Van Gool, L. (2022) Unsupervised robust domain adaptation without source data. In *Proceedings of WACV* (pp. 2009–2018).

Ahmed, S. K. M., Lejbolle, A. R., Panda, R., & Roy-Chowdhury, A. K. (2020). Camera on-boarding for person re-identification using hypothesis transfer learning. In *Proceedings of CVPR* (pp. 12144–12153).

Ahmed, S. K. M., Lohit, S., Peng, K.-C., Jones, M., & Roy-Chowdhury, A. K. (2022). Cross-modal knowledge transfer without task-relevant source data. In *Proceedings of ECCV* (pp. 111–127).

Ahmed, W., Morerio, P., & Murino, V. (2022). Cleaning noisy labels by negative ensemble learning for source-free unsupervised domain adaptation. In *Proceedings of WACV* (pp. 1616–1625).

Ahmed, S. K. M., Raychaudhuri, D. S., Paul, S., Oymak, S., & Roy-Chowdhury, A. K. (2021). Unsupervised multi-source domain adaptation without access to source data. In *Proceedings of CVPR* (pp. 10103–10112).

Alet, F., Bauza, M., Kawaguchi, K., Kuru, N. G., Lozano-Perez, T., & Kaelbling, L. P. (2021). Tailoring: Encoding inductive biases by optimizing unsupervised objectives at prediction time. In *Proceedings of NeurIPS* (pp. 29206–29217).

Alexandari, A., Kundaje, A., & Shrikumar, A. (2020). Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *Proceedings of ICML* (pp. 222–232).

Alfarra, M., Pérez, J. C., Thabet, A., Bibi, A., Torr, P. H. S., & Ghanem, B. (2022). Combating adversaries with anti-adversaries. In *Proceedings of AAAI* (pp. 5992–6000).

An, Q., Li, R., Gu, L., Zhang, H., Chen, Q., Lu, Z., Wang, F., & Zhu, Y. (2022). A privacy-preserving unsupervised domain adaptation framework for clinical text analysis. arXiv:2201.07317.

Ao, S., Li, X., & Ling, C. (2017). Fast generalized distillation for semi-supervised domain adaptation. In *Proceedings of AAAI* (pp. 1719–1725).

Ayyoubzadeh, S. M., Liu, W., Kezele, I., Yu, Y., Wu, X., Wang, Y., & Jin, T. (2023). Test-time adaptation for optical flow estimation using motion vectors. *IEEE Transactions on Image Processing, 32*, 4977–4988.

Azimi, F., Palacio, S., Raue, F., Hees, J., Bertinetto, L., & Dengel, A. (2022). Self-supervised test-time adaptation on video data. In *Proceedings of WACV* (pp. 3439–3448).

Azizzadenesheli, K., Liu, A., Yang, F., & Anandkumar, A. (2019). *Regularized learning for domain adaptation under label shifts*. In *Proceedings of ICLR*.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). *Layer normalization*. In *Proceedings of NeurIPS workshops*.

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proceedings of NeurIPS* (pp. 12449–12460).

Bahmani, S., Hahn, O., Zamfir, E., Araslanov, N., Cremers, D., & Roth, S. (2022). *Semantic self-adaptation: Enhancing generalization with a single sample*. In *Proceedings of ECCV workshops*.

Bahng, H., Jahanian, A., Sankaranarayanan, S., & Isola, P. (2022). Visual prompting: Modifying pixel space to adapt pre-trained models. arXiv:2203.17274.

Banerjee, P., Gokhale, T., & Baral, C. (2021). Self-supervised test-time learning for reading comprehension. In *Proceedings of NAACL* (pp. 1200–1211).

Bao, W., Wei, T., Wang, H., & He, J. (2023). *Adaptive test-time personalization for federated learning*. In *Proceedings of NeurIPS*.

Bateson, M., Lombaert, H., & Ayed, I. B. (2022). Test-time adaptation with shape moments for image segmentation. In *Proceedings of MICCAI* (pp. 736–745).

Bateson, M., Kervadec, H., Dolz, J., Lombaert, H., & Ayed, I. B. (2022). Source-free domain adaptation for image segmentation. *Medical Image Analysis, 82*, 102617.

Bau, D., Strobelt, H., Peebles, W., Wulff, J., Zhou, B., Zhu, J.-Y., & Torralba, A. (2019). Semantic photo manipulation with a generative image prior. *ACM Transactions on Graphics, 38*(4), 1–11.

Belli, D., Das, D., Major, B., & Porikli, F. (2022). Online adaptive personalization for face anti-spoofing. In *Proceedings of ICIP* (pp. 351–355).

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning, 79*, 151–175.

Ben-David, E., Oved, N., & Reichart, R. (2022). Pada: Example-based prompt learning for on-the-fly adaptation to unseen domains. *Transactions of the Association for Computational Linguistics, 10*, 414–433.

Berthelot, D., Carlini, N., Goodfellow, I., Oliver, A., Papernot, N., & Raffel, C. (2019). Mixmatch: A holistic approach to semi-supervised learning. In *Proceedings of NeurIPS* (pp. 5049–5059).

Bertrand, J., Zilos, G. K., Kalantidis, Y., & Tolias, G. (2023). *Test-time training for matching-based video object segmentation*. In *Proceedings of NeurIPS*.

Bohdal, O., Li, D., Hu, S. X., & Hospedales, T. (2022). *Feed-forward source-free latent domain adaptation via cross-attention*. In *Proceedings of ICML workshops*.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2022). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Borlino, F. C., Polizzotto, S., Caputo, B., & Tommasi, T. (2022). Self-supervision & meta-learning for one-shot unsupervised cross-domain detection. *Computer Vision and Image Understanding, 223*, 103549.

Boudiaf, M., Denton, T., Van Merriënboer, B., Dumoulin, V., & Triantafillou, E. (2023). In search for a generalizable method for source free domain adaptation. In *Proceedings of ICML* (pp. 2914–2931).

Boudiaf, M., Mueller, R., Ayed, I. B., & Bertinetto, L. (2022). Parameter-free online test-time adaptation. In *Proceedings of CVPR* (pp. 8344–8353).

Bousmalis, K., Silberman, N., Dohan, D., Erhan, D., & Krishnan, D. (2017). Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proceedings of CVPR* (pp. 3722–3731).

Brahma, D., & Rai, P. (2023). *A probabilistic framework for lifelong test-time adaptation*. In *Proceedings of CVPR*.

Brahmbhatt, S., Gu, J., Kim, K., Hays, J., & Kautz, J. (2018). Geometry-aware learning of maps for camera localization. In *Proceedings of CVPR* (pp. 2616–2625).

Cao, Z., Li, Z., Guo, X., & Wang, G. (2021). Towards cross-environment human activity recognition based on radar without source data. *IEEE Transactions on Vehicular Technology, 70*(11), 11843–11854.

Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B., & Tommasi, T. (2019). Domain generalization by solving jigsaw puzzles. In *Proceedings of CVPR* (pp. 2229–2238).

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *Proceedings of ECCV* (pp. 132–149).

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. In *Proceedings of NeurIPS* (pp. 9912–9924).

Chen, Y.-H., Chen, W.-Y., Chen, Y.-T., Tsai, B.-C., Wang, Y.-C.F., & Sun, M. (2017). No more discrimination: Cross city adaptation of road scene segmenters. In *Proceedings of ICCV* (pp. 1992–2001).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. In *Proceedings of ICML* (pp. 1597–1607).

Chen, W., Lin, L., Yang, S., Xie, D., Pu, S., Zhuang, Y., & Ren, W. (2022). Self-supervised noisy label learning for source-free unsupervised domain adaptation. In *Proceedings of IROS* (pp. 10185–10192).

Chen, C., Liu, Q., Jin, Y., Dou, Q., & Heng, P.-A. (2021). Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In *Proceedings of MICCAI* (pp. 225–235).

Chen, W.-Y., Liu, Y.-C., Kira, Z., Wang, Y.-C.F., & Huang, J.-B. (2018). *A closer look at few-shot classification*. ICLR: In Proceedings of

Chen, Y., Mancini, M., Zhu, X., & Akata, Z. (2022). Semi-supervised and unsupervised deep visual learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Chen, Y., Schmid, C., & Sminchisescu, C. (2019). Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proceedings of ICCV* (pp. 7063–7072).

Chen, D., Wang, D., Darrell, T., & Ebrahimi, S. (2022). Contrastive test-time adaptation. In *Proceedings of CVPR* (pp. 295–305).

Chen, J., Xian, X., Yang, Z., Chen, T., Lu, Y., Shi, Y., Pan, J., & Lin, L. (2023). *Open-world pose transfer via sequential test-time adaption*. In *Proceedings of CVPR*.

Chen, M., Xue, H., & Cai, D. (2019). Domain adaptation for semantic segmentation with maximum squares loss. In *Proceedings of ICCV* (pp. 2090–2099).

Chi, Z., Wang, Y., Yu, Y., & Tang, J. (2021). Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning. In *Proceedings of CVPR* (pp. 9137–9146).

Chidlovskii, B., Clinchant, S., & Csurka, G. (2016). Domain adaptation in the absence of source domain data. In *Proceedings of KDD* (pp. 451–460).

Choi, S., Yang, S., Choi, S., & Yun, S. (2022). Improving test-time adaptation via shift-agnostic weight regularization and nearest source prototypes. In *Proceedings of ECCV* (pp. 440–458).

Choi, M., Choi, J., Baik, S., Kim, T. H., & Lee, K. M. (2021). Test-time adaptation for video frame interpolation via meta-learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(12), 9615–9628.

Chu, T., Liu, Y., Deng, J., Li, W., & Duan, L. (2022). Denoised maximum classifier discrepancy for source free unsupervised domain adaptation. In *Proceedings of AAAI* (pp. 472–480).

Clinchant, S., Chidlovskii, B., & Csurka, G. (2016). Transductive adaptation of black box predictions. In *Proceedings of ACL* (pp. 326–331).

Conti, A., Rota, P., Wang, Y., & Ricci, E. (2022). *Cluster-level pseudo-labelling for source-free cross-domain facial expression recognition*. In *Proceedings of BMVC*.

Cubuk, E. D., Zoph, B., Shlens, J., & Le, Q. V. (2020). *Randaugment: Practical automated data augmentation with a reduced search space*. In *Proceedings of CVPR workshops*.

Cui, S., Wang, S., Zhuo, J., Li, L., Huang, Q., & Tian, Q. (2020). Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proceedings of CVPR* (pp. 3941–3950).

Darestani, M. Z., Liu, J., & Heckel, R. (2022). Test-time training can close the natural distribution shift performance gap in deep learning based compressed sensing. In *Proceedings of ICML* (pp. 4754–4776).

Das, D., Borse, S., Park, H., Azarian, K., Cai, H., Garrepalli, R., & Porikli, F. (2023). Transadapt: A transformative framework for online test time adaptive semantic segmentation. In *Proceedings of ICASSP* (pp. 1–5).

Deng, Z., Chen, Z., Niu, S., Li, T., Zhuang, B., & Tan, M. (2023). *Efficient test-time adaptation for super-resolution with second-order degradation and reconstruction*. In *Proceedings of NeurIPS*.

Deng, B., Zhang, Y., Tang, H., Ding, C., & Jia, K. (2021). On universal black-box domain adaptation. arXiv:2104.04665.

Ding, N., Xu, Y., Tang, Y., Xu, C., Wang, Y., & Tao, D. (2022). Source-free domain adaptation via distribution estimation. In *Proceedings of CVPR* (pp. 7212–7222).

Ding, Y., Liang, J., Jiang, B., Zheng, A., & He, R. (2024). Maps: A noise-robust progressive learning approach for source-free domain adaptive keypoint detection. *IEEE Transactions on Circuits and Systems for Video Technology, 34*(3), 1376–1387.

Ding, Y., Sheng, L., Liang, J., Zheng, A., & He, R. (2023). Proxymix: Proxy-based mixup training with label refinery for source-free domain adaptation. *Neural Networks, 167*, 92–103.

D'Innocente, A., Borlino, F. C., Bucci, S., Caputo, B., & Tommasi, T. (2020). One-shot unsupervised cross-domain detection. In *Proceedings of ECCV* (pp. 732–748).

D'Innocente, A., Bucci, S., Caputo, B., & Tommasi, T. (2019). Learning to generalize one sample at a time with self-supervision. arXiv:1910.03915.

Döbler, M., Marsden, R. A., & Yang, B. (2023). Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of CVPR*.

Dong, J., Fang, Z., Liu, A., Sun, G., & Liu, T. (2021). Confident anchor-induced multi-source free domain adaptation. In *Proceedings of NeurIPS* (pp. 2848–2860).

Dubey, A., Ramanathan, V., Pentland, A., & Mahajan, D. (2021). Adaptive methods for real-world domain generalization. In *Proceedings of CVPR* (pp. 14340–14349).

Eshete, B. (2021). Making machine learning trustworthy. *Science, 373*(6556), 743–744.

Fang, Y., Yap, P.-T., Lin, W., Zhu, H., & Liu, M. (2024). Source-free unsupervised domain adaptation: A survey. *Neural Networks*, 106230.

Feng, C.-M., Yu, K., Liu, Y., Khan, S., & Zuo, W. (2023). Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proceedings of ICCV* (pp. 2704–2714).

Feng, Z., Xu, C., & Tao, D. (2021). Open-set hypothesis transfer with semantic consistency. *IEEE Transactions on Image Processing, 30*, 6473–6484.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of ICML* (pp. 1126–1135).

Gal, Y., & Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of ICML* (pp. 1050–1059).

Gan, Y., Ma, X., Lou, Y., Bai, Y., Zhang, R., Shi, N., & Luo, L. (2023). *Decorate the newcomers: Visual domain prompt for continual test time adaptation*. In *Proceedings of AAAI*.

Gandelsman, Y., Sun, Y., Chen, X., & Efros, A. A. (2022). Test-time training with masked autoencoders. In *Proceedings of NeurIPS* (pp. 29374–29385).

Ganin, Y., & Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *Proceedings of ICML* (pp. 1180–1189).

Gao, J., Zhang, J., Liu, X., Darrell, T., Shelhamer, E., & Wang, D. (2023). *Back to the source: Diffusion-driven adaptation to test-time corruption*. In *Proceedings of CVPR*.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of CVPR* (pp. 2414–2423).

Gidaris, S., Singh, P., & Komodakis, N. (2018). *Unsupervised representation learning by predicting image rotations*. In *Proceedings of ICLR*.

Gong, T., Jeong, J., Kim, T., Kim, Y., Shin, J., & Lee, S.-J. (2022). Note: Robust continual test-time adaptation against temporal correlation. In *Proceedings of NeurIPS* (pp. 27253–27266).

Goyal, S., Sun, M., Raghunathan, A., & Kolter, Z. (2022). Test-time adaptation via conjugate pseudo-labels. In *Proceedings of NeurIPS* (pp. 6204–6218).

Grandvalet, Y., & Bengio, Y. (2004). Semi-supervised learning by entropy minimization. In *Proceedings of NeurIPS* (pp. 529–536).

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., & Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research, 13*(1), 723–773.

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In *Proceedings of NeurIPS* (pp. 507–520).

Guan, S., Xu, J., Wang, Y., Ni, B., & Yang, X. (2021). Bilevel online adaptation for out-of-domain human mesh reconstruction. In *Proceedings of CVPR* (pp. 10472–10481).

Gui, S., Li, X., & Ji, S. (2024). *Active test-time adaptation: Theoretical analyses and an algorithm*. In *Proceedings of ICLR*.

Gulrajani, I., & Lopez-Paz, D. (2020). In search of lost domain generalization. In *Proceedings of ICLR*.

Guo, C., Rana, M., Cisse, M., & van der Maaten, L. (2018). *Countering adversarial images using input transformations*. In *Proceedings of ICLR*.

Hansen, N., Jangir, R., Sun, Y., Alenyà, G., Abbeel, P., Efros, A. A., Pinto, L., & Wang, X. (2021). *Self-supervised policy adaptation during deployment*. In *Proceedings of ICLR*.

Hardt, M., & Sun, Y. (2024). *Test-time training on nearest neighbors for large language models*. In *Proceedings of ICLR*.

He, Y., Carass, A., Zuo, L., Dewey, B. E., & Prince, J. L. (2021). Autoencoder based self-supervised test-time adaptation for medical image analysis. *Medical Image Analysis*, 102136.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., & Girshick, R. (2022). Masked autoencoders are scalable vision learners. In *Proceedings of CVPR* (pp. 16000–16009).

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In *Proceedings of CVPR* (pp. 9729–9738).
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of CVPR* (pp. 770–778).
Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A., & Darrell, T. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of ICML* (pp. 1989–1998).
Hong, S., & Kim, S. (2021). Deep matching prior: Test-time optimization for dense correspondence. In *Proceedings of ICCV* (pp. 9907–9917).
Hong, J., Lyu, L., Zhou, J., & Spranger, M. (2023). *Mecta: Memory-economic continual test-time model adaptation*. In *Proceedings of ICLR*.
Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2021). Meta-learning in neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 5149–5169.
Hou, Y., & Zheng, L. (2020). Source free domain adaptation with image translation. arXiv:2008.07514.
Hou, Y., & Zheng, L. (2021). Visualizing adapted knowledge in domain transfer. In *Proceedings of CVPR* (pp. 13824–13833).
Hu, S., Liao, Z., & Xia, Y. (2022). Prosfda: Prompt learning based source-free domain adaptation for medical image segmentation. arXiv:2211.11514.
Hu, W., Miyato, T., Tokui, S., Matsumoto, E., & Sugiyama, M. (2017). Learning discrete representations via information maximizing self-augmented training. In *Proceedings of ICML* (pp. 1558–1567).
Hu, M., Song, T., Gu, Y., Luo, X., Chen, J., Chen, Y., Zhang, Y., & Zhang, S. (2021). Fully test-time adaptation for image segmentation. In *Proceedings of MICCAI* (pp. 251–260).
Hu, X., Uzunbas, G., Chen, S., Wang, R., Shah, A., Nevatia, R., & Lim, S.-N. (2021). Mixnorm: Test-time adaptation through online normalization estimation. arXiv:2110.11478.
Hu, X., Zhang, K., Xia, L., Chen, A., Luo, J., Sun, Y., Wang, K., Qiao, N., Zeng, X., & Sun, M. et al. (2024) Reclip: Refine contrastive language image pre-training with source free domain adaptation. In *Proceedings of WACV* (pp. 2994–3003).
Huang, J., Guan, D., Xiao, A., & Lu, S. (2021). Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In *Proceedings of NeurIPS* (pp. 3635–3649).
Huang, Y., Yang, X., Zhang, J., & Xu, C. (2022). Relative alignment network for source-free multimodal video domain adaptation. In *Proceedings of ACM-MM* (pp. 1652–1660).
Hussein, S. A., Tirer, T., & Giryes, R. (2020). Image-adaptive gan based reconstruction. In *Proceedings of AAAI* (pp. 3121–3129).
Ioffe, S. (2017). Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In *Proceedings of NeurIPS* (pp. 1942–1950).
Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of ICML* (pp. 448–456).
Iscen, A., Tolias, G., Avrithis, Y., & Chum, O. (2019). Label propagation for deep semi-supervised learning. In *Proceedings of CVPR* (pp. 5070–5079).
Ishii, M., & Sugiyama, M. (2021). Source-free domain adaptation via distributional alignment by matching batch normalization statistics. arXiv:2101.10842.
Iwasawa, Y., & Matsuo, Y. (2021). Test-time classifier adjustment module for model-agnostic domain generalization. In *Proceedings of NeurIPS* (pp. 2427–2440).
Jain, V., & Learned-Miller, E. (2011). Online domain adaptation of a pre-trained cascade of classifiers. In *Proceedings of CVPR* (pp. 577–584).
Jamal, M. A., Li, H., & Gong, B. (2018). Deep face detector adaptation without negative transfer or catastrophic forgetting. In *Proceedings of CVPR* (pp. 5608–5618).
Jang, M., Chung, S.-Y., & Chung, H. W. (2023). *Test-time adaptation via self-training with nearest neighbor information*. In *Proceedings of ICLR*.
Jiang, L., & Lin, T. (2023). *Test-time robust personalization for federated learning*. In *Proceedings of ICLR*.
Jiao, J., Li, H., Zhang, T., & Lin, J. (2022). Source-free adaptation diagnosis for rotating machinery. *IEEE Transactions on Industrial Informatics*.
Jin, Y., Wang, X., Long, M., & Wang, J. (2020). Minimum class confusion for versatile domain adaptation. In *Proceedings of ECCV* (pp. 464–480).
Jin, W., Zhao, T., Ding, J., Liu, Y., Tang, J., & Shah, N. (2023). *Empowering graph representation learning with test-time graph transformation*. In *Proceedings of ICLR*.
Jing, M., Zhen, X., Li, J., & Snoek, C. G. M. (2022). Variational model perturbation for source-free domain adaptation. In *Proceedings of NeurIPS* (pp. 17173–17187).
Jing, L., & Tian, Y. (2020). Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(11), 4037–4058.
Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *Proceedings of ICML* (pp. 200–209).
Jung, S., Lee, J., Kim, N., Shaban, A., Boots, B., & Choo, J. (2023). Cafa: Class-aware feature alignment for test-time adaptation. In *Proceedings of ICCV* (pp. 19060–19071).
Kan, Z., Chen, S., Li, Z., & He, Z. (2022). Self-constrained inference optimization on structural groups for human pose estimation. In *Proceedings of ECCV* (pp. 729–745).
Karani, N., Erdil, E., Chaitanya, K., & Konukoglu, E. (2021). Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis, 68*, 101907.
Karim, N., Mithun, N. C., & Rajvanshi, A., et al. (2023) C-sfda: A curriculum learning aided self-training framework for efficient source free domain adaptation. In *Proceedings of CVPR*.
Karmanov, A., Guan, D., Lu, S., Saddik, A. E., & Xing, E. (2024). *Efficient test-time adaptation of vision-language models*. In *Proceedings of CVPR*.
Kenton, J.D.M.-W.C., & Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL* (pp. 4171–4186).
Khurana, A., Paul, S., Rai, P., Biswas, S., & Aggarwal, G. (2021). Sita: Single image test-time adaptation. arXiv:2112.02355.
Kim, J., Hwang, I., & Kim, Y. M. (2022). Ev-tta: Test-time adaptation for event-based object recognition. In *Proceedings of CVPR* (pp. 17745–17754).
Kim, I., Kim, Y., & Kim, S. (2020). Learning loss for test-time augmentation. In *Proceedings of NeurIPS* (pp. 4163–4174).
Kim, J., Lee, J.-T., Chang, S., & Kwak, N. (2022). Variational on-the-fly personalization. In *Proceedings of ICML* (pp. 11134–11147).
Kim, E., Sun, M., Raghunathan, A., & Kolter, J. Z. (2023). *Reliable test-time adaptation via agreement-on-the-line*. In *Proceedings of NeurIPS workshops*.
Kim, Y., Yim, J., Yun, J., & Kim, J. (2019). Nlnl: Negative learning for noisy labels. In *Proceedings of ICCV* (pp. 101–110).
Kim, Y., Cho, D., Han, K., Panda, P., & Hong, S. (2021). Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence, 2*(6), 508–518.
Kim, S., Min, Y., Jung, Y., & Kim, S. (2024). Controllable style transfer via test-time training of implicit neural representation. *Pattern Recognition, 146*, 109988.
Kingetsu, H., Kobayashi, K., Okawa, Y., Yokota, Y., & Nakazawa, K. (2022). Multi-step test-time adaptation with entropy minimization and pseudo-labeling. In *Proceedings of ICIP* (pp. 4153–4157).

Kojima, T., Matsuo, Y., & Iwasawa, Y. (2022). Robustifying vision transformer without retraining from scratch by test-time class-conditional feature alignment. In *Proceedings of IJCAI* (pp. 1009–1016).

Kong, F., Yuan, S., Hao, W., & Henao, R. (2023). *Mitigating test-time bias for fair image retrieval*. In *Proceedings of NeurIPS*.

Kothandaraman, D., Shekhar, S., Sancheti, A., Ghuhan, M., Shukla, T., & Manocha, D. (2023). Salad: Source-free active label-agnostic domain adaptation for classification, segmentation and detection. In *Proceedings of WACV* (pp. 382–391).

Kouw, W. M., & Loog, M. (2019). A review of domain adaptation without target labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(3), 766–785.

Krause, A., Perona, P., & Gomes, R. (2010). Discriminative clustering by regularized information maximization. In *Proceedings of NeurIPS* (pp. 775–783).

Kumar, V., Lal, R., Patil, H., & Chakraborty, A. (2023). Conmix for source-free single and multi-target domain adaptation. In *Proceedings of WACV* (pp. 4178–4188).

Kundu, J. N., Bhambri, S., Kulkarni, A., Sarkar, H., Jampani, V., & Babu, R. V. (2022). Concurrent subsidiary supervision for unsupervised source-free domain adaptation. In *Proceedings of ECCV* (pp. 177–194).

Kundu, J. N., Kulkarni, A., Bhambri, S., Mehta, D., Kulkarni, S., Jampani, V., & Babu, R. V. (2022). Balancing discriminability and transferability for source-free domain adaptation. In *Proceedings of ICML* (pp. 11710–11728).

Kundu, J. N., Kulkarni, A., Singh, A., Jampani, V., & Babu, R. V. (2021). Generalize then adapt: Source-free domain adaptive semantic segmentation. In *Proceedings of ICCV* (pp. 7046–7056).

Kundu, J. N., Seth, S., Pradyumna, Y. M., Jampani, V., Chakraborty, A., & Babu, R. V. (2022). Uncertainty-aware adaptation for self-supervised 3d human pose estimation. In *Proceedings of CVPR* (pp. 20448–20459).

Kundu, J. N., Venkat, N., & Babu, R. V. (2020). Universal source-free domain adaptation. In *Proceedings of CVPR* (pp. 4544–4553).

Kundu, J. N., Venkat, N., Revanur, A., & Babu, R. V. (2020). Towards inheritable models for open-set domain adaptation. In *Proceedings of CVPR* (pp. 12376–12385).

Kurmi, V. K., Subramanian, V. K., & Namboodiri, V. P. (2021). Domain impression: A source data free domain adaptation method. In *Proceedings of WACV* (pp. 615–625).

Kuzborskij, I., & Orabona, F. (2013). Stability and hypothesis transfer learning. In *Proceedings of ICML* (pp. 942–950).

Kuznietsov, Y., Proesmans, M., & Van Gool, L. (2022). Towards unsupervised online domain adaptation for semantic segmentation. In *Proceedings of WACV workshops* (pp. 261–271).

Laine, S., & Aila, T. (2017). *Temporal ensembling for semi-supervised learning*. In *Proceedings of ICLR*.

Lao, Q., Jiang, X., & Havaei, M. (2021). Hypothesis disparity regularized mutual information maximization. In *Proceedings of AAAI* (pp. 8243–8251).

Laparra, E., Su, X., Zhao, Y., Uzuner, O., Miller, T., & Bethard, S. (2021). Semeval-2021 task 10: Source-free domain adaptation for semantic processing. In *International workshop on semantic evaluation (SemEval)* (pp. 348–356).

Lee, D.-H. (2013). *Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks*. In *Proceedings of ICML workshops*.

Lee, P., Jeon, S., Hwang, S., Shin, M., & Byun, H. (2023). Source-free subject adaptation for eeg-based visual recognition. In *Proceedings of BCI* (pp. 1–6).

Lee, J., Jung, D., Yim, J., & Yoon, S. (2022). Confidence score for source-free unsupervised domain adaptation. In *Proceedings of ICML* (pp. 12365–12377).

Lee, J., & Lee, G. (2023). Feature alignment by uncertainty and self-training for source-free unsupervised domain adaptation. *Neural Networks, 161*, 682–692.

Li, W., Cao, M., & Chen, S. (2022). Jacobian norm for unsupervised source-free domain adaptation. arXiv:2204.03467.

Li, X., Chen, W., Xie, D., Yang, S., Yuan, P., Pu, S., & Zhuang, Y. (2021). A free lunch for unsupervised domain adaptive object detection without source data. In *Proceedings of AAAI* (pp. 8474–8481).

Li, X., Du, Z., Li, J., Zhu, L., & Lu, K. (2022). Source-free active domain adaptation via energy-based locality preserving transfer. In *Proceedings of ACM-MM* (pp. 5802–5810).

Li, R., Jiao, Q., Cao, W., Wong, H.-S., & Wu, S. (2020). Model adaptation: Unsupervised domain adaptation without source data. In *Proceedings of CVPR* (pp. 9641–9650).

Li, X., Li, J., Zhu, L., Wang, G., & Huang, Z. (2021). Imbalanced source-free domain adaptation. In *Proceedings of ACM-MM* (pp. 3330–3339).

Li, X., Liu, S., De Mello, S., Kim, K., Wang, X., Yang, M.-H., & Kautz, J. (2020) Online adaptation for consistent mesh reconstruction in the wild. In *Proceedings of NeurIPS* (pp. 15009–15019).

Li, H., Liu, H., Hu, D., Wang, J., Johnson, H., Sherbini, O., Gavazzi, F., D'Aiello, R., Vanderver, A., Long, J., Jane, P., & Oguz, I. (2022). *Self-supervised test-time adaptation for medical image segmentation*. In *Proceedings of MICCAI workshops*.

Li, Z., Togo, R., Ogawa, T., & Haseyama, M. (2022). Union-set multi-source model adaptation for semantic segmentation. In *Proceedings of ECCV* (pp. 579–595).

Li, Y., Wang, N., Liu, J., & Hou, X. (2017). Demystifying neural style transfer. In *Proceedings of IJCAI* (pp. 2230–2236).

Li, Y., Wang, N., Shi, J., Liu, J., & Hou, X. (2017). *Revisiting batch normalization for practical domain adaptation*. In *Proceedings of ICLR*.

Li, D., Yang, Y., Song, Y.-Z., & Hospedales, T. M. (2018). Learning to generalize: meta-learning for domain generalization. In *Proceedings of AAAI* (pp. 3490–3497).

Li, S., Ye, M., Zhu, X., Zhou, L., & Xiong, L. (2022). Source-free object detection by learning to overlook domain style. In *Proceedings of CVPR* (pp. 8014–8023).

Li, J., Yu, Z., Du, Z., Zhu, L., & Shen, H. T. (2024). A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, D., Zhang, J., Yang, Y., Liu, C., Song, Y.-Z., & Hospedales, T.M. (2019). Episodic training for domain generalization. In *Proceedings of ICCV* (pp. 1446–1455).

Liang, J., He, R., Sun, Z., & Tan, T. (2019). Distant supervised centroid shift: A simple and efficient approach to visual domain adaptation. In *Proceedings of CVPR* (pp. 2975–2984).

Liang, J., Hu, D., & Feng, J. (2020). Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In *Proceedings of ICML* (pp. 6028–6039).

Liang, J., Hu, D., & Feng, J. (2021). Domain adaptation with auxiliary target domain-oriented classifier. In *Proceedings of CVPR* (pp. 16632–16642).

Liang, J., Hu, D., Feng, J., & He, R. (2021). Umad: Universal model adaptation under domain and category shift. arXiv:2112.08553.

Liang, J., Hu, D., Feng, J., & He, R. (2022). Dine: Domain adaptation from single and multiple black-box predictors. In *Proceedings of CVPR* (pp. 8003–8013).

Liang, J., Wang, Y., Hu, D., He, R., & Feng, J. (2020). A balanced and uncertainty-aware approach for partial domain adaptation. In *Proceedings of ECCV* (pp. 123–140).

Liang, J., Hu, D., Wang, Y., He, R., & Feng, J. (2022). Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(11), 8602–8617.

Lim, H., Kim, B., Choo, J., & Choi, S. (2023). *Ttn: A domain-shift aware batch normalization in test-time adaptation*. In *Proceedings of ICLR*.

Lin, G.-T., Li, S.-W., & Lee, H.-y. (2022). Listen, adapt, better wer: Source-free single-utterance test-time adaptation for automatic speech recognition. In *Proceedings of Interspeech* (pp. 2198–2202).

Lipton, Z., Wang, Y.-X., & Smola, A. (2018). Detecting and correcting for label shift with black box predictors. In *Proceedings of ICML* (pp. 3122–3130).

Litrico, M., Bue, A. D., & Morerio, P. (2023). *Guiding pseudo-labels with uncertainty estimation for test-time adaptation*. In *Proceedings of CVPR*.

Liu, Z., & Fang, Y. (2023). Learning adaptable risk-sensitive policies to coordinate in multi-agent general-sum games. In *Proceedings of ICONIP* (pp. 27–40).

Liu, Y., Chen, Y., Dai, W., Gou, M., Huang, C.-T., & Xiong, H. (2022). Source-free domain adaptation with contrastive domain alignment and self-supervised exploration for face anti-spoofing. In *Proceedings of ECCV* (pp. 511–528).

Liu, Q., Chen, C., Dou, Q., & Heng, P.-A. (2022). Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. In *Proceedings of AAAI* (pp. 1756–1764).

Liu, H., Chi, Z., Yu, Y., Wang, Y., Chen, J., & Tang, J. (2023). Meta-auxiliary learning for future depth prediction in videos. In *Proceedings of WACV* (pp. 5756–5765).

Liu, Y., Kothari, P., van Delft, B., Bellot-Gurlet, B., Mordan, T., & Alahi, A. (2021). Ttt++: When does self-supervised test-time training fail or thrive? In *Proceedings of NeurIPS* (pp. 21808–21820).

Liu, J., Li, X., An, S., & Chen, Z. (2022). Source-free unsupervised domain adaptation for blind image quality assessment. arXiv:2207.08124.

Liu, C., Wang, L., Lyu, L., Sun, C., Wang, X., & Zhu, Q. (2023). *Twofer: Tackling continual domain shift with simultaneous domain generalization and adaptation*. In *Proceedings of ICLR*.

Liu, H., Wu, Z., Li, L., Salehkalaibar, S., Chen, J., & Wang, K. (2022). Towards multi-domain single image dehazing via test-time training. In *Proceedings of CVPR* (pp. 5831–5840).

Liu, X., Xing, F., Yang, C., El Fakhri, G., & Woo, J. (2021). Adapting off-the-shelf source segmenter for target medical image segmentation. In *Proceedings of MICCAI* (pp. 549–559).

Liu, Y., Zhang, W., & Wang, J. (2021). Source-free domain adaptation for semantic segmentation. In *Proceedings of CVPR* (pp. 1215–1224).

Liu, Y., Zhang, W., Wang, J., & Wang, J. (2021). Data-free knowledge transfer: A survey. arXiv:2112.15278.

Liu, X., & Yuan, Y. (2022). A source-free domain adaptive polyp detection framework with style diversification flow. *IEEE Transactions on Medical Imaging, 41*(7), 1897–1908.

Liu, C., Zhou, L., Ye, M., & Li, X. (2022). Self-alignment for black-box domain adaptation of image classification. *IEEE Signal Processing Letters, 29*, 1709–1713.

Long, M., Cao, Y., Wang, J., & Jordan, M. (2015). Learning transferable features with deep adaptation networks. In *Proceedings of ICML* (pp. 97–105).

Lumentut, J. S., & Park, I. K. (2022). 3d body reconstruction revisited: Exploring the test-time 3d body mesh refinement strategy via surrogate adaptation. In *Proceedings of ACM-MM* (pp. 5923–5933).

Luo, X., Chen, W., Tan, Y., Li, C., He, Y., & Jia, X. (2021). Exploiting negative learning for implicit pseudo label rectification in source-free domain adaptive semantic segmentation. arXiv:2106.12123.

Luo, Y., Liu, P., Guan, T., Yu, J., & Yang, Y. (2020). Adversarial style mining for one-shot unsupervised domain adaptation. In *Proceedings of NeurIPS* (pp. 20612–20623).

Lyu, F., Ye, M., Ma, A. J., Yip, T.C.-F., Wong, G.L.-H., & Yuen, P. C. (2022). Learning from synthetic CT images via test-time training for liver tumor segmentation. *IEEE Transactions on Medical Imaging, 41*(9), 2510–2520.

Ma, W., Chen, C., Zheng, S., Qin, J., Zhang, H., & Dou, Q. (2022). Test-time adaptation with calibration of medical image classification nets for label distribution shift. In *Proceedings of MICCAI* (pp. 313–323).

Ma, X., Zhang, J., Guo, S., & Xu, W. (2023). *Swapprompt: Test-time prompt adaptation for vision-language models*. In *Proceedings of NeurIPS*.

Ma, N., Bu, J., Lu, L., Wen, J., Zhou, S., Zhang, Z., Gu, J., Li, H., & Yan, X. (2022). Context-guided entropy minimization for semi-supervised domain adaptation. *Neural Networks, 154*, 270–282.

Mancini, M., Karaoguz, H., Ricci, E., Jensfelt, P., & Caputo, B. (2018). Kitting in the wild through online domain adaptation. In *Proceedings of IROS* (pp. 1103–1109).

Mao, C., Chiquier, M., Wang, H., Yang, J., & Vondrick, C. (2021). Adversarial attacks are reversible with natural supervision. In *Proceedings of ICCV* (pp. 661–671).

Marsden, R. A., Döbler, M., & Yang, B. (2024). Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. In *Proceedings of WACV* (pp. 2555–2565).

Min, C., Kim, T., & Lim, J. (2023). Meta-learning for adaptation of deep optical flow networks. In *Proceedings of WACV* (pp. 2145–2154).

Mirza, M., & Osindero, S. (2014). Conditional generative adversarial nets. arXiv:1411.1784.

Mirza, M. J., Micorek, J., Possegger, H., & Bischof, H. (2022). The norm must go on: Dynamic unsupervised domain adaptation by normalization. In *Proceedings of CVPR* (pp. 14765–14775).

Mirza, M. J., Soneira, P. J., Lin, W., Kozinski, M., Possegger, H., & Bischof, H. (2023). Actmad: Activation matching to align distributions for test-time-training. In *Proceedings of CVPR* (pp. 24152–24161).

Miyato, T., Maeda, S.-I., Koyama, M., & Ishii, S. (2018). Virtual adversarial training: A regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 41*(8), 1979–1993.

Mohan, S., Vincent, J.L., Manzorro, R., Crozier, P., Fernandez-Granda, C., & Simoncelli, E. (2021). Adaptive denoising via gaintuning. In *Proceedings of NeurIPS* (pp. 23727–23740).

Moon, J. H., Das, D., Lee, C. S. G. (2020). Multi-step online unsupervised domain adaptation. In *Proceedings of ICASSP* (pp. 41172–41576).

Morerio, P., Volpi, R., Ragonesi, R., & Murino, V. (2020). Generative pseudo-label refinement for unsupervised domain adaptation. In *Proceedings of WACV* (pp. 3130–3139).

Müller, R., Kornblith, S., & Hinton, G. E. (2019). When does label smoothing help? In *Proceedings of NeurIPS* (pp. 4694–4703).

Mummadi, C. K., Hutmacher, R., Rambach, K., Levinkov, E., Brox, T., & Metzen, J. H. (2021). Test-time adaptation to distribution shift by confidence maximization and input transformation. arXiv:2106.14999.

Nado, Z., Padhy, S., Sculley, D., D'Amour, A., Lakshminarayanan, B., & Snoek, J. (2020). *Evaluating prediction-time batch normalization for robustness under covariate shift*. In *Proceedings of ICML workshops*.

Naik, A., Wu, Y., Naik, M., & Wong, E. (2023). Do machine learning models learn common sense? arXiv:2303.01433.

Nayak, G. K., Mopuri, K. R., Jain, S., & Chakraborty, A. (2022). Mining data impressions from deep models as substitute for the unavailable training data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(11), 8465–8481.

Nelakurthi, A. R., Maciejewski, R., & He, J. (2018). Source free domain adaptation using an off-the-shelf classifier. In *Proceedings of IEEE BigData* (pp. 140–145).

Nitzan, Y., Aberman, K., He, Q., Liba, O., Yarom, M., Gandelsman, Y., Mosseri, I., Pritch, Y., & Cohen-Or, D. (2022). Mystyle: A per-

sonalized generative prior. *ACM Transactions on Graphics, 41*(6), 1–10.
Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., & Tan, M. (2022). Efficient test-time model adaptation without forgetting. In *Proceedings of ICML* (pp. 16888–16905).
Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., & Tan, M. (2023). *Towards stable test-time adaptation in dynamic wild world*. In *Proceedings of ICLR*.
Panagiotakopoulos, T., Dovesi, P. L., Härenstam-Nielsen, L., & Poggi, M. (2022). Online domain adaptation for semantic segmentation in ever-changing conditions. In *Proceedings of ECCV* (pp. 128–146).
Pandey, P., Raman, M., Varambally, S., & Prathosh A. P. (2021) Generalization on unseen domains via inference-time label-preserving target projections. In *Proceedings of CVPR* (pp. 12924–12933).
Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359.
Park, S., Yoo, J., Cho, D., Kim, J., & Kim, T. H. (2020). Fast adaptation to super-resolution networks via meta-learning. In *Proceedings of ECCV* (pp. 754–769).
Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. A. (2016). Context encoders: Feature learning by inpainting. In *Proceedings of CVPR* (pp. 2536–2544).
Paul, S., Saha, A., & Samanta, A. (2022). Ttt-ucdr: Test-time training for universal cross-domain retrieval. arXiv:2208.09198.
Peng, Q., Ding, Z., Lyu, L., Sun, L., & Chen, C. (2022). Toward better target representation for source-free and black-box domain adaptation. arXiv:2208.10531.
Pérez, J. C., Alfarra, M., Jeanneret, G., Rueda, L., Thabet, A., Ghanem, B., & Arbeláez, P. (2021). Enhancing adversarial robustness via test-time transformation ensembling. In *Proceedings of ICCV* (pp. 81–91).
Plananamente, M., Plizzari, C., & Caputo, B. (2022). Test-time adaptation for egocentric action recognition. In *Proceedings of ICIAP* (pp. 206-218).
Prabhu, V., Khare, S., Kartik, D., & Hoffman, J. (2022). Augco: Augmentation consistency-guided self-training for source-free domain adaptive semantic segmentation. arXiv:2107.10140.
Prabhudesai, M., Ke, T.-W., Li, A., Pathak, D., & Fragkiadaki, K. (2023). *Test-time adaptation of discriminative models via diffusion generative feedback*. In *Proceedings of NeurIPS*.
Press, O., Schneider, S., Kümmerer, M., & Bethge, M. (2023). *Rdumb: A simple approach that questions our progress in continual test-time adaptation*. In *Proceedings of NeurIPS*.
Qiu, Z., Zhang, Y., Lin, H., Niu, S., Liu, Y., Du, Q., & Tan, M. (2021). Source-free domain adaptation via avatar prototype generation and adaptation. In *Proceedings of IJCAI* (pp. 2921–2927).
Qu, S., Chen, G., Zhang, J., Li, Z., He, W., & Tao, D. (2022). Bmd: A general class-balanced multicentric dynamic prototype strategy for source-free domain adaptation. In *Proceedings of ECCV* (pp. 165—182).
Qu, S., Zou, T., Roehrbein, F., Lu, C., Chen, G., Tao, D., & Jiang, C. (2023). *Upcycling models under domain and category shift*. In *Proceedings of CVPR*.
Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2008). *Dataset shift in machine learning*. MIT Press.
Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., & Clark, J. et al. (2021) Learning transferable visual models from natural language supervision. In *Proceedings of ICML* (pp. 8748–8763).
Ragab, M., Eldele, E., Tan, W. L., Foo, C.-S., Chen, Z., Wu, M., Kwoh, C.-K., & Li, X. (2023). Adatime: A benchmarking suite for domain adaptation on time series data. *ACM Transactions on Knowledge Discovery from Data*.
Reddy, N., Singhal, A., Kumar, A., Baktashmotlagh, M., & Arora, C. (2022). Master of all: simultaneous generalization of urban-scene segmentation to all adverse weather conditions. In *Proceedings of ECCV* (pp. 51–69).
Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of CVPR* (pp. 10684–10695).
Rostami, M. (2021). Lifelong domain adaptation via consolidated internal distribution. In *Proceedings of NeurIPS* (pp. 11172–11183).
Roy, S., Trapp, M., Pilzer, A., Kannala, J., Sebe, N., Ricci, E., & Solin, A. (2022). Uncertainty-guided source-free domain adaptation. In *Proceedings of ECCV* (pp. 537–555).
RoyChowdhury, A., Chakrabarty, P., Singh, A., Jin, S., Jiang, H., Cao, L., & Learned-Miller, E. (2019). Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of CVPR* (pp. 780–790).
Royer, A., & Lampert, C. H. (2015). Classifier adaptation at prediction time. In *Proceedings of CVPR* (pp. 1401–1409).
Rusak, E., Schneider, S., Pachitariu, G., Eck, L., Gehler, P. V., Bringmann, O., Brendel, W., & Bethge, M. (2022). If your data distribution shifts, use self-learning. *Transactions on Machine Learning Research*.
Saenko, K., Kulis, B., Fritz, M., & Darrell, T. (2010). Adapting visual category models to new domains. In *Proceedings of ECCV* (pp. 213–226).
Saerens, M., Latinne, P., & Decaestecker, C. (2002). Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation, 14*(1), 21–41.
Sahoo, R., Shanmugam, D., & Guttag, J. (2020). *Unsupervised domain adaptation in the absence of source data*. In *Proceedings of ICML Workshops*.
Sain, A., Bhunia, A. K., Potlapalli, V., Chowdhury, P. N., Xiang, T., & Song, Y.-Z. (2022). Sketch3t: Test-time training for zero-shot sbir. In *Proceedings of CVPR* (pp. 7462–7471).
Saito, K., Watanabe, K., Ushiku, Y., & Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of CVPR* (pp. 3723–3732).
Saltori, C., Krivosheev, E., Lathuilière, S., Sebe, N., Galasso, F., Fiameni, G., Ricci, E., & Poiesi, F. (2022). Gipso: Geometrically informed propagation for online adaptation in 3D lidar segmentation. In *Proceedings of ECCV* (pp. 567–585).
Saltori, C., Lathuiliére, S., Sebe, N., Ricci, E., & Galasso, F. (2020). Sf-uda^{3D}: Source-free unsupervised domain adaptation for lidar-based 3d object detection. In *Proceedings of 3DV* (pp. 771–780).
Samadh, J. H. A., Gani, H., Hussein, N. H., Khattak, M. U., Naseer, M., Khan, F., & Khan, S. (2023). *Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization*. In *Proceedings of NeurIPS*.
Sarkar, A., Sarkar, A., & Balasubramanian, V. N. (2022). Leveraging test-time consensus prediction for robustness against unseen noise. In *Proceedings of WACV* (pp. 1839–1848).
Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., & Bethge, M. (2020). Improving robustness against common corruptions by covariate shift adaptation. In *Proceedings of NeurIPS* (pp. 11539–11551).
Segu, M., Tonioni, A., & Tombari, F. (2023). Batch normalization embeddings for deep domain generalization. *Pattern Recognition, 135*, 109115.
Seo, S., Suh, Y., Kim, D., Kim, G., Han, J., & Han, B. (2020). Learning to optimize domain specific normalization for domain generalization. In *Proceedings of ECCV* (pp. 68–83).
Shanmugam, D., Blalock, D., Balakrishnan, G., & Guttag, J. (2021). Better aggregation in test-time augmentation. In *Proceedings of ICCV* (pp. 1214–1223).
Sheng, L., Liang, J., He, R., Wang, Z., & Tan, T. (2023). Adaptguard: Defending against universal attacks for model adaptation. In *Proceedings of ICCV* (pp. 19093–19103).

Shi, Y., & Sha, F. (2012). Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In *Proceedings of ICML* (pp. 1275–1282).
Shi, C., Holtz, C., & Mishne, G. (2021). *Online adversarial purification based on self-supervision*. In *Proceedings of ICLR*.
Shin, I., Tsai, Y.-H., Zhuang, B., Schulter, S., Liu, B., Garg, S., Kweon, I. S., & Yoon, K.-J. (2022). Mm-tta: Multi-modal test-time adaptation for 3d semantic segmentation. In *Proceedings of CVPR* (pp. 16928–16937).
Shocher, A., Cohen, N., & Irani, M. (2018). "Zero-shot" super-resolution using deep internal learning. In *Proceedings of CVPR* (pp. 3118–3126).
Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data, 6*(1), 1–48.
Shu, M., Nie, W., De-An Huang, Yu, Z., Goldstein, T., Anandkumar, A., & Xiao, C. (2022). Test-time prompt tuning for zero-shot generalization in vision-language models. In *Proceedings of NeurIPS* (pp. 14274–14289).
Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion, 81*, 84–90.
Sinha, S., Gehler, P., Locatello, F., & Schiele, B. (2023). Test: Test-time self-training under distribution shift. In *Proceedings of WACV* (pp. 2759–2769).
Šipka, T., Šulc, M., & Matas, J. (2022). The hitchhiker's guide to prior-shift adaptation. In *Proceedings of WACV* (pp. 1516–1524).
Sivaprasad, P. T., & Fleuret, F. (2021). *Test time adaptation through perturbation robustness*. In *Proceedings of NeurIPS workshops*.
Sivaprasad, P. T., & Fleuret, F. (2021). Uncertainty reduction for model adaptation in semantic segmentation. In *Proceedings of CVPR* (pp. 9613–9623).
Smith, L., & Gal, Y. (2018). Understanding measures of uncertainty for adversarial example detection. In *Proceedings of UAI* (pp. 560–569).
Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., & Li, C.-L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proceedings of NeurIPS* (pp. 596–608).
Song, J., Lee, J., Kweon, I. S., & Choi, S. (2023). *Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization*. In *Proceedings of CVPR*.
Song, J., Park, K., Shin, I., Woo, S., & Kweon, I. S. (2022). Cd-tta: Compound domain test-time adaptation for semantic segmentation. arXiv:2212.08356.
Stan, S., & Rostami, M. (2021). Unsupervised model adaptation for continual semantic segmentation. In *Proceedings of AAAI* (pp. 2593–2601).
Su, Y., Xu, X., & Jia, K. (2022). Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering. In *Proceedings of NeurIPS* (pp. 17543–17555).
Sun, T., Lu, C., & Ling, H. (2022). Prior knowledge guided unsupervised domain adaptation. In *Proceedings of ECCV* (pp. 639–655).
Sun, T., Lu, C., & Ling, H. (2023). *Domain adaptation with adversarial training on penultimate activations*. In *Proceedings of AAAI*.
Sun, Z., Shen, Z., Lin, L., Yu, Y., Yang, Z., Yang, S., & Chen, W. (2022). Dynamic domain generalization. In *Proceedings of IJCAI* (pp. 1342–1348).
Sun, Y., Tzeng, E., Darrell, T., & Efros, A. A. (2019) Unsupervised domain adaptation through self-supervision. arXiv:1909.11825.
Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A., & Hardt, M. (2020). Test-time training with self-supervision for generalization under distribution shifts. In *Proceedings of ICML* (pp. 9229–9248).
Tan, Y., Chen, C., Zhuang, W., Dong, X., Lyu, L., & Long, G. (2023). *Is heterogeneity notorious? taming heterogeneity to handle test-time shift in federated learning*. In *Proceedings of NeurIPS*.
Tang, S., Shi, Y., Ma, Z., Li, J., Lyu, J., Li, Q., & Zhang, J. (2021). Model adaptation through hypothesis transfer with gradual knowledge distillation. In *Proceedings of IROS* (pp. 5679–5685).
Tang, Y., Zhang, C., Xu, H., Chen, S., Cheng, J., Leng, L., Guo, Q., & He, Z. (2023). *Neuro-modulated Hebbian learning for fully test-time adaptation*. In *Proceedings of CVPR*.
Tanwisuth, K., Fan, X., Zheng, H., Zhang, S., Zhang, H., Chen, B., & Zhou, M. (2021). A prototype-oriented framework for unsupervised domain adaptation. In *Proceedings of NeurIPS* (pp. 17194–17208).
Tanwisuth, K., Zhang, S., Zheng, H., He, P., & Zhou, M. (2023). Pouf: Prompt-oriented unsupervised fine-tuning for large pre-trained models. In *Proceedings of ICML* (pp. 33816–33832).
Tarvainen, A., & Valpola, H. (2017). Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proceedings of NeurIPS* (pp. 1195–1204).
Termöhlen, J.-A., Klingner, M., Brettin, L. J., Schmidt, N. M., & Fingscheidt, T. (2021). Continual unsupervised domain adaptation for semantic segmentation by online frequency domain style transfer. In *Proceedings of ITSC* (pp. 2881–2888).
Thopalli, K., Turaga, P., & Thiagarajan, J. J. (2023). Domain alignment meets fully test-time adaptation. In *Proceedings of ACML* (pp. 1006–1021).
Tian, Q., Peng, S., & Ma, T. (2023). Source-free unsupervised domain adaptation with trusted pseudo samples. *ACM Transactions on Intelligent Systems and Technology, 14*(2), 1–17.
Tian, J., Zhang, J., Li, W., & Xu, D. (2022). Vdm-da: Virtual domain modeling for source data-free domain adaptation. *IEEE Transactions on Circuits and Systems for Video Technology, 32*(6), 3749–3760.
Tomar, D., Vray, G., Bozorgtabar, B., & Thiran, J.-P. (2023). *Tesla: Test-time self-learning with automatic adversarial augmentation*. In *Proceedings of CVPR*.
Tommasi, T., Orabona, F., & Caputo, B. (2013). Learning categories from few examples with multi model knowledge transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 36*(5), 928–941.
Tsai, Y.-Y., Mao, C., Lin, Y.-K., & Yang, J. (2023). Self-supervised convolutional visual prompts. arXiv:2303.00198.
Tzeng, E., Hoffman, J., Saenko, K., & Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of CVPR* (pp. 7167–7176).
Ulyanov, D., Vedaldi, A., & Lempitsky, V. (2016). Instance normalization: The missing ingredient for fast stylization. arXiv:1607.08022.
Valvano, G., Leo, A., & Tsaftaris, S. A. (2022). Re-using adversarial mask discriminators for test-time training under distribution shifts. *Journal of Machine Learning for Biomedical Imaging, 1*, 1–27.
van de Ven, G. M., Tuytelaars, T., & Tolias, A. S. (2022). Three types of incremental learning. *Nature Machine Intelligence, 4*, 1185–1197.
van Laarhoven, T., & Marchiori, E. (2017). Unsupervised domain adaptation with random walks on target labelings. arXiv:1706.05335.
Varsavsky, T., Orbes-Arteaga, M., Sudre, C. H., Graham, M. S., Nachev, P., & Cardoso, M. J. (2020). Test-time unsupervised domain adaptation. In *Proceedings of MICCAI* (pp. 428–436).
Vibashan, V. S., Valanarasu, J. M. J., & Patel, V. M. (2022). Target and task specific source-free domain adaptive image segmentation. arXiv:2203.15792.
Volpi, R., de Jorge, P., Larlus, D., & Csurka, G. (2022). On the road to online adaptation for semantic image segmentation. In *Proceedings of CVPR* (pp. 19184–19195).
Wang, J.-K., & Wibisono, A. (2023). *Towards understanding gd with hard and conjugate pseudo-labels for test-time adaptation*. In *Proceedings of ICLR*.
Wang, Q., Fink, O., Van Gool, L., & Dai, D. (2022). Continual test-time domain adaptation. In *Proceedings of CVPR* (pp. 7201–7211).

Wang, F., Han, Z., Gong, Y., & Yin, Y. (2022). Exploring domain-invariant parameters for source free domain adaptation. In *Proceedings of CVPR* (pp. 7151–7160).
Wang, F., Han, Z., Zhang, Z., & Yin, Y. (2022). Active source free domain adaptation. arXiv:2205.10711.
Wang, Y., Huang, Z., & Hong, X. (2022). S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. In *Proceedings of NeurIPS* (pp. 5682–5695).
Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., & Yu, P. (2022). Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*.
Wang, Y., Li, C., Jin, W., Li, R., Zhao, J., Tang, J., & Xie, X. (2022). Test-time training for graph neural networks. arXiv:2210.08813.
Wang, Y., Liang, J., & Zhang, Z. (2022). Source data-free cross-domain semantic segmentation: Align, teach and propagate. arXiv:2106.11653.
Wang, D., Liu, S., Ebrahimi, S., Shelhamer, E., & Darrell, T. (2021). On-target adaptation. arXiv:2109.01087.
Wang, Z., Luo, Y., Zheng, L., Chen, Z., Wang, S., & Huang, Z. (2023). In search of lost online test-time adaptation: A survey. arXiv:2310.20199.
Wang, D., Shelhamer, E., Liu, S., Olshausen, B., & Darrell, T. (2021). *Tent: Fully test-time adaptation by entropy minimization*. In *Proceedings of ICLR*.
Wang, D., Shelhamer, E., Olshausen, B., & Darrell, T. (2019). Dynamic scale inference by entropy minimization. arXiv:1908.03182.
Wang, X., Tsvetkov, Y., Ruder, S., & Neubig, G. (2021). Efficient test time adapter ensembling for low-resource language varieties. In *EMNLP findings* (pp. 730—737).
Wang, Z., Ye, M., Zhu, X., Peng, L., Tian, L., & Zhu, Y. (2022). Metateacher: Coordinating multi-model domain adaptation for medical image classification. In *Proceedings of NeurIPS* (pp. 20823–20837).
Wang, J., Zhang, J., Bian, Y., Cai, Y., Wang, C. & Pu, S. (2021). Self-domain adaptation for face anti-spoofing. In *Proceedings of AAAI* (pp. 2746–2754).
Wang, X., Zhuo, J., Cui, S., Wang, S., & Fang, Y. (2024). Learning invariant representation with consistency and diversity for semi-supervised source hypothesis transfer. In *Proceedings of ICASSP* (pp. 5125–5129).
Wang, S., Wang, J., Xi, H., Zhang, B., Zhang, L., & Wei, H. (2024). Optimization-free test-time adaptation for cross-person activity recognition. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 7*(4), 1–27.
Wegmann, S., Scattone, F., Carp, I., Gillick, L., Roth, R., & Yamron, J. (1998). *Dragon systems' 1997 broadcast news transcription system*. In *Proceedings of DARPA broadcast news transcription and understanding workshop*.
Wen, Z., Niu, S., Li, G., Wu, Q., Tan, M., & Wu, Q. (2024). Test-time model adaptation for visual question answering with debiased self-supervisions. *IEEE Transactions on Multimedia, 26*, 2137–2147.
Wilson, G., & Cook, D. J. (2020). A survey of unsupervised deep domain adaptation. *ACM Transactions on Intelligent Systems and Technology, 11*(5), 1–46.
Wu, R., Guo, C., Su, Y., & Weinberger, K. Q. (2021). Online adaptation to label distribution shift. In *Proceedings of NeurIPS* (pp. 11340–11351).
Wu, C., Pan, Y., Li, Y., & Wang, J. Z. (2023). Learning to adapt to online streams with distribution shifts. arXiv:2303.01630.
Wu, Q., Yue, X., & Sangiovanni-Vincentelli, A. (2021). *Domain-agnostic test-time adaptation by prototypical training with auxiliary data*. In *Proceedings of NeurIPS workshops*.
Wu, A., Zheng, W.-S., Guo, X., & Lai, J.-H. (2019). Distilled person re-identification: Towards a more scalable system. In *Proceedings of CVPR* (pp. 1187–1196).
Xia, H., Zhao, H., & Ding, Z. (2021). Adaptive adversarial network for source-free domain adaptation. In *Proceedings of ICCV* (pp. 9010–9019).
Xia, K., Deng, L., Duch, W., & Wu, D. (2022). Privacy-preserving domain adaptation for motor imagery-based brain-computer interfaces. *IEEE Transactions on Biomedical Engineering, 69*(11), 3365–3376.
Xiao, Z., Zhen, X., Liao, S., & Snoek, C. G. M. (2023). *Energy-based test sample adaptation for domain generalization*. In *Proceedings of ICLR*.
Xiao, Z., Zhen, X., Shao, L., & Snoek, C. G. M. (2022). *Learning to generalize across domains on single test samples*. In *Proceedings of ICLR*.
Xie, Q., Dai, Z., Hovy, E., Luong, T., & Le, Q. (2020). Unsupervised data augmentation for consistency training. In *Proceedings of NeurIPS* (pp. 6256–6268).
Xiong, L., Ye, M., Zhang, D., Gan, Y., & Liu, Y. (2022). Source data-free domain adaptation for a faster R-CNN. *Pattern Recognition, 124*, 108436.
Xu, B., Liang, J., He, L., & Sun, Z. (2022). Mimic embedding via adaptive aggregation: Learning generalizable person re-identification. In *Proceedings of ECCV* (pp. 372–388).
Xu, Y., Yang, J., Cao, H., Wu, K., Min, W., & Chen, Z. (2022). Learning temporal consistency for source-free video domain adaptation. In *Proceedings of ECCV* (pp. 147–164).
Yan, H., Guo, Y., & Yang, C. (2021). *Augmented self-labeling for source-free unsupervised domain adaptation*. In *Proceedings of NeurIPS workshops*.
Yan, H., Guo, Y., & Yang, C. (2021). *Source-free unsupervised domain adaptation with surrogate data generation*. In *Proceedings of BMVC*.
Yang, Y., & Soatto, S. (2020). FDA: Fourier domain adaptation for semantic segmentation. In *Proceedings of CVPR* (pp. 4085–4095).
Yang, L., Gao, M., Chen, Z., Xu, R., Shrivastava, A., & Ramaiah, C. (2022). Burn after reading: Online adaptation for cross-domain streaming data. In *Proceedings of ECCV* (pp. 404–422).
Yang, P., Liang, J., Cao, J., & He, R. (2023). Auto: Adaptive outlier optimization for online test-time ood detection. arXiv:2303.12267.
Yang, J., Peng, X., Wang, K., Zhu, Z., Feng, J., Xie, L., & You, Y. (2023). *Divide to adapt: Mitigating confirmation bias for domain adaptation of black-box predictors*. In *Proceedings of ICLR*.
Yang, X., Song, Z., King, I., & Xu, Z. (2022). A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*.
Yang, S., van de Weijer, J., Herranz, L., & Jui, S. (2021). Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In *Proceedings of NeurIPS* (pp. 29393–29405).
Yang, S., Wang, Y., van de Weijer, J., Herranz, L., & Jui, S. (2021). Generalized source-free domain adaptation. In *Proceedings of ICCV* (pp. 8978–8987).
Yang, S., Wang, Y., Wang, K., Jui, S., & van de Weijer, J. (2022). One ring to bring them all: Model adaptation under domain and category shift. arXiv:2206.03600.
Yang, J., Yan, R., & Hauptmann, A. G. (2007). Cross-domain video concept detection using adaptive svms. In *Proceedings of ACM-MM* (pp. 188–197).
Yang, T., Zhou, S., Wang, Y., Lu, Y., & Zheng, N. (2022). Test-time batch normalization. arXiv:2205.10210.
Yang, H., Chen, C., Jiang, M., Liu, Q., Cao, J., Heng, P. A., & Dou, Q. (2022). Dltta: Dynamic learning rate for test-time adaptation on cross-domain medical images. *IEEE Transactions on Medical Imaging, 41*(12), 3575–3586.
Yang, C., Guo, X., Chen, Z., & Yuan, Y. (2022). Source free domain adaptation for medical image segmentation with fourier style mining. *Medical Image Analysis, 79*, 102457.

Yang, B., Ma, A. J., & Yuen, P. C. (2022). Revealing task-relevant model memorization for source-protected unsupervised domain adaptation. *IEEE Transactions on Information Forensics and Security, 17*, 716–731.

Yang, S., Wang, Y., Herranz, L., Jui, S., & van de Weijer, J. (2023). Casting a bait for offline and online source-free domain adaptation. *Computer Vision and Image Understanding, 234*, 103747.

Yang, B., Yeh, H.-W., Harada, T., & Yuen, P. C. (2021). Model-induced generalization error bound for information-theoretic representation learning in source-data-free unsupervised domain adaptation. *IEEE Transactions on Image Processing, 31*, 419–432.

Yang, C., & Zhou, J. (2008). Non-stationary data sequence classification using online class priors estimation. *Pattern Recognition, 41*(8), 2656–2664.

Ye, H., Ding, Y., Li, J., & Ng, H. T. (2022). *Robust question answering against distribution shifts with test-time adaptation: An empirical study*. In *Proceedings of EMNLP findings*.

Ye, Y., Liu, Z., Zhang, Y., Li, J., & Shen, H. (2022). Alleviating style sensitivity then adapting: Source-free domain adaptation for medical image segmentation. In *Proceedings of ACM-MM* (pp. 1935–1944).

Ye, M., Zhang, J., Ouyang, J., & Yuan, D. (2021). Source data-free unsupervised domain adaptation for semantic segmentation. In *Proceedings of ACM-MM* (pp. 2233–2242).

Yi, L., Xu, G., Xu, P., Li, J., Pu, R., Ling, C., McLeod, A. I., & Wang, B. (2023). *When source-free domain adaptation meets learning with noisy labels*. In *Proceedings of ICLR*.

Yi, C., Yang, S., Wang, Y., Li, H., Tan, Y.-P., & Kot, A. (2023). *Temporal coherent test-time optimization for robust video classification*. In *Proceedings of ICLR*.

Yin, H., Molchanov, P., Alvarez, J. M., Li, Z., Mallya, A., Hoiem, D., Jha, N. K., & Kautz, J. (2020). Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of CVPR* (pp. 8715–8724).

Yoon, J., Hwang, S. J., & Lee, J. (2021). Adversarial purification with score-based generative models. In *Proceedings of ICML* (pp. 12062–12072).

Yoon, H. S., Yoon, E., Tee, J. T. J., Hasegawa-Johnson, M., Li, Y., & Yoo, C. D. (2024). *C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion*. In *Proceedings of ICLR*.

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? In *Proceedings of NeurIPS* (pp. 3320–3328).

You, Y., Chen, T., Sui, Y., Chen, T., Wang, Z., & Shen, Y. (2020). Graph contrastive learning with augmentations. In *Proceedings of NeurIPS* (pp. 5812–5823).

You, F., Li, J., & Zhao, Z. (2021). Test-time batch statistics calibration for covariate shift. arXiv:2110.04065.

You, F., Li, J., Zhu, L., Chen, Z., & Huang, Z. (2021). Domain adaptive semantic segmentation without source data. In *Proceedings of ACM-MM* (pp. 3293–3302).

You, K., Long, M., Cao, Z., Wang, J., & Jordan, M. I. (2019). Universal domain adaptation. In *Proceedings of CVPR* (pp. 2720–2729).

Yu, Y., Sheng, L., He, R., & Liang, J. (2023). Benchmarking test-time adaptation against distribution shifts in image classification. arXiv:2307.03133.

Yuan, L., Xie, B., & Li, S. (2023). Robust test-time adaptation in dynamic scenarios. In *Proceedings of CVPR* (pp. 15922–15932).

Zeng, R., Deng, Q., Xu, H., Niu, S., & Chen, J. (2023). Exploring motion cues for video test-time adaptation. In *Proceedings of ACM-MM* (pp. 1840–1850).

Zeng, L., Han, J., Liang, D., & Ding, W. (2024). Rethinking precision of pseudo label: Test-time adaptation via complementary learning. *Pattern Recognition Letters, 177*, 96–102.

Zhang, Z., Chen, W., Cheng, H., Li, Z., Li, S., Lin, L., & Li, G. (2022). Divide and contrast: Source-free domain adaptation via adaptive contrastive learning. In *Proceedings of NeurIPS* (pp. 5137–5149).

Zhang, H., Cisse, M., Dauphin, Y. N., & Lopez-Paz, D. (2018). *mixup: Beyond empirical risk minimization*. In *Proceedings of ICLR*.

Zhang, R., Isola, P., & Efros, A. A. (2016). Colorful image colorization. In *Proceedings of ECCV* (pp. 649–666).

Zhang, M., Levine, S., & Finn, C. (2022). Memo: Test time robustness via adaptation and augmentation. In *Proceedings of NeurIPS* (pp. 38629–38642).

Zhang, M., Marklund, H., Dhawan, N., Gupta, A., Levine, S., & Finn, C. (2021). Adaptive risk minimization: Learning to adapt to domain shift. In *Proceedings of NeurIPS* (pp. 23664–23678).

Zhang, J., Nie, X., & Feng, J. (2020). Inference stage optimization for cross-scenario 3d human pose estimation. In *Proceedings of NeurIPS* (pp. 2408–2419).

Zhang, Y.-F., Wang, J., Liang, J., Zhang, Z., Yu, B., Wang, L., Tao, D., & Xie, X. (2023). Domain-specific risk minimization for out-of-distribution generalization. In *Proceedings of KDD* (pp. 3409–3421).

Zhang, T., Xiang, Y., Li, X., Weng, Z., Chen, Z., & Fu, Y. (2022). Free lunch for cross-domain occluded face recognition without source data. In *Proceedings of ICASSP* (pp. 2944–2948).

Zhang, D., Ye, M., Xiong, L., Li, S., & Li, X. (2021). Source-style transferred mean teacher for source-data free object detection. In *ACM Multimedia Asia* (pp. 1–8).

Zhang, H., Zhang, Y., Jia, K., & Zhang, L. (2021). *Unsupervised domain adaptation of black-box source models*. In *Proceedings of BMVC*.

Zhang, B., Zhang, X., Liu, Y., Cheng, L., & Li, Z. (2021). Matching distributions between model and data: Cross-domain knowledge distillation for unsupervised domain adaptation. In *Proceedings of ACL* (pp. 5423–5433).

Zhang, X., & Chen, Y.-C. (2023). Adaptive domain generalization via online disagreement minimization. *IEEE Transactions on Image Processing, 32*, 4247–4258.

Zhang, J., Qi, L., Shi, Y., & Gao, Y. (2022). Generalizable model-agnostic semantic segmentation via target-specific normalization. *Pattern Recognition, 122*, 108292.

Zhao, B., Chen, C., & Xia, S.-T. (2023). *Delta: Degradation-free fully test-time adaptation*. In *Proceedings of ICLR*.

Zhao, H., Liu, Y., Alahi, A., & Lin, T. (2023). On pitfalls of test-time adaptation. In *Proceedings of ICML* (pp. 42058–42080).

Zhao, X., Liu, C., Sicilia, A., Hwang, S. J., & Fu, Y. (2022). Test-time fourier style calibration for domain generalization. In *Proceedings of IJCAI* (pp. 1721–1727).

Zhao, S., Wang, X., Zhu, L., & Yang, Y. (2024). *Test-time adaptation with clip reward for zero-shot generalization in vision-language models*. In *Proceedings of ICLR*.

Zhou, A., & Levine, S. (2021). Bayesian adaptation for covariate shift. In *Proceedings of NeurIPS* (pp. 914–927).

Zhou, K., Liu, Z., Qiao, Y., Xiang, T., & Loy, C. C. (2022). Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Zhou, Y., Ren, J., Li, F., Zabih, R., & Lim, S. N. (2023). *Test-time distribution normalization for contrastively learned visual-language models*. In *Proceedings of NeurIPS*.

Zhou, Q., Zhang, K.-Y., Yao, T., Yi, R., Sheng, K., Ding, S., & Ma, L. (2022). Generative domain adaptation for face anti-spoofing. In *Proceedings of ECCV* (pp. 335–356).

Zhu, W., Huang, Y., Xu, D., Qian, Z., Fan, W., & Xie, X. (2021). Test-time training for deformable multi-scale image registration. In *Proceedings of ICRA* (pp. 13618–13625).

Zou, Y., Yu, Z., Kumar, B. V. K., & Wang, J. (2018). Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of ECCV* (pp. 289–305).

Zou, Y., Zhang, Z., Li, C.-L., Zhang, H., Pfister, T., & Huang, J.-B. (2022). Learning instance-specific adaptation for cross-domain segmentation. In *Proceedings of ECCV* (pp. 459–476).