A Comprehensive Survey on Test-Time Adaptation under Distribution Shifts

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

Received: date / Accepted: date

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, testtime 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.

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 [1]. However, when the test distribution (target) differs from the training distribution (source), we face the prob-

lem of distribution shifts. Such a shift poses significant challenges for machine learning systems deployed in the wild, such as images captured by different cameras [2], road scenes of different cities [3], and imaging devices in different hospitals [4]. 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) [5] 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) [6] 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 [7, 8, 9]. 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 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

¹ Center for Research on Intelligent Perception and Computing & State Key Laboratory of Multimodal Artificial Intelligence Systems, CASIA, China

² School of Artificial Intelligence, University of Chinese Academy of Sciences, China

³ Nanjing University, China

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.

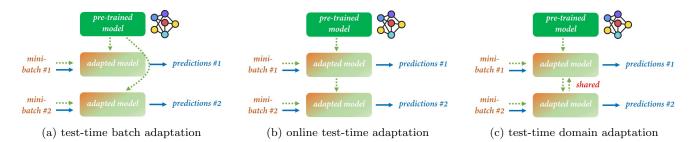


Fig. 1 The test-time adaptation (TTA) paradigm aims to adapt the pre-trained model to various types of unlabeled test data, including single mini-batch in (a), streaming data in (b), or an entire dataset in (c), before making predictions. During the adaptation process, either the model or the input data 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.

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, \dots, b_m during test time. Firstly, TTDA, also known as source-free domain adaptation [7, 10, 11], 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 [8, 12, 13, 14]. In other words, the predictions made for each mini-batch are independent of the predictions made for the other mini-batches. Thirdly, OTTA [9, 15, 16] 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.
- 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 [6, 17], our specific focus is on test-time adaptation where the availability of source data during adaptation is limited or non-existent. Two recent surveys [18, 19] 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 [20] 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.

2 Related Research Topics

2.1 Domain Adaptation

As a special case of transfer learning [21], DA [22] 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 [23]. There are two major assumptions of distribution shift [1]: 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., [6, 17]) 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 [24, 25], feature-level alignment [26, 27, 28]), outputlevel regularization [29, 30, 31], and class-prior estimation [32, 33, 34]. If it is feasible to generate training data from the source model [11], then the task of TTDA

 $^{^2}$ Such a single-sample adaptation corresponds to a batch size of 1, a.k.a., test-time instance adaptation.

can be tackled using conventional DA methods. Likewise, one relevant topic closely related to TTBA (batch size equals 1) is one-shot domain adaptation [35, 36], 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 [37, 38], 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) [39] is another special case of transfer learning where pre-trained models (source hypotheses) retain information about previously encountered tasks. Shallow HTL methods [40, 41, 42 typically assume that the optimal target hypothesis is closely associated with these source hypotheses, and subsequent methods [43, 44] extend this approach to a semi-supervised scenario where unlabeled target data are also utilized for training. Fine-tuning [45] 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 [46, 47] 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 [48, 49, 50] 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 [51] 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.q., [5, 52, 53]).

2.4 Self-Supervised Learning

Self-supervised learning [54] 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 [55], image inpainting [56], and image rotation [57]. Advanced pretext tasks like clustering [58, 59] and contrastive learning [60, 61] 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 [62], speech processing [63], and graph-structured data [64]. For TTA tasks, these self-supervised learning techniques can be utilized to help learn discriminative features [65] or act as an auxiliary task [8].

2.5 Semi-Supervised Learning

Semi-supervised learning [66] 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 [67, 68], which encourages the model to produce confident predictions; consistency regularization under input variations [69, 70] and model variations [71, 72], which forces networks to output similar predictions when inputs or models are perturbed; and graph-based regularization [73], 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 [74] employs data augmentation techniques [75] (e.g., geometric transformations and color space augmentations) on test images to boost prediction accuracy [76], estimate uncertainty [77], and enhance robustness [78, 79]. As a typical example, tencrop testing [76] computes the final prediction by averaging predictions from ten different scaled versions of a test image. Other popular aggregation strategies include selective augmentation [80] and learnable aggregation weights [74]. In addition to data variation, Monte Carlo (MC) dropout [81] 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 $\mathcal{D}_{\mathcal{S}} = \{(x_1, y_1), \dots, (x_{n_s}, y_{n_s})\}$ and one unlabeled target domain $\mathcal{D}_{\mathcal{T}} = \{x_1, \dots, x_{n_t}\}$ 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 [1] assumption $(p_{\mathcal{S}}(y|x) = p_{\mathcal{T}}(y|x))$. Other distribution shifts like prior shift [32] 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.[82] 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 [83, 84, 85] follow this learning mechanism and adapt the source classifier to unlabeled target features. To gain benefits from end-toend 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 [7, 65], model adaptation [11], and source-free domain adaptation [10], 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 $f_{\mathcal{S}}: \mathcal{X}_{\mathcal{S}} \to \mathcal{Y}_{\mathcal{S}}$ 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 [23] manner. Note that, all test data (target data) are required to be seen during adaptation.

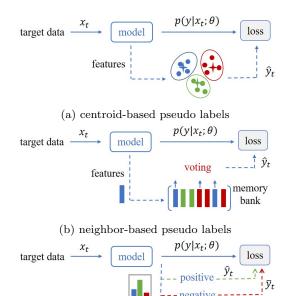
3.2 Taxonomy on TTDA Algorithms

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 [66] and employ various prevalent SSL techniques tailored for

 ${\bf Table~1}~{\bf A}$ taxonomy on TTDA methods with representative strategies.

Families	Model Rationales	Representative Strategies
pseudo- labeling	centroid-based neighbor-based complementary labels optimization-based	SHOT [7, 65], BMD [86] NRC [87], SSNLL [88] LD [89], ATP [90] ASL [91], KUDA [92]
consistency	data variations model variations both variations	G-SFDA [93], APA [94] SFDA-UR [95], FMML [96] AdaContrast [97], MAPS [98]
clustering	entropy minimization mutual information explicit clustering	ASFA [99], 3C-GAN [11] SHOT [7, 65], UMAD [100] ISFDA [101], SDA-FAS [102]
source estimation	data generation data translation data selection feature estimation	3C-GAN [11], DI [103] SFDA-IT [104], ProSFDA [105] SHOT++ [65], DaC [106] VDM-DA [107], CPGA [108]
self-supervision	auxiliary tasks	SHOT++ [65], StickerDA [109]



(c) complementary pseudo labels

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 .

unlabeled data during adaptation. A simple yet effective technique, pseudo-labeling [68], 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,

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

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 [110, 111, 112], most TTDA methods (e.g., SFIT [113]) consider only reliable pseudo labels using diverse filtering mechanisms. Fig. 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 [58], SHOT [7, 65] 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,

$$\begin{cases} m_c = \sum_x [p_\theta(y_c|x) \cdot g(x)] / \sum_x p_\theta(y_c|x), \ c \in [1, C], \\ \hat{y} = \arg\min_c d(g(x), m_c), \ \forall x \in \mathcal{D}_t, \end{cases}$$
(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 [106, 114, 108].

Twofer [115] identifies confident samples to build more accurate centroids, while BMD [86] 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 [116] 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 [117] 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 [86] 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 [88, 118, 119, 97, 120]. For instance, SSNLL [88] performs K-means clustering in the target domain and aggregates predictions of its neighbors within the same cluster. DIPE [118] diminishes label ambiguity by cor-

recting the pseudo label to the majority vote of its neighbors. In contrast, SFDA-APM [110] constructs an anchor set comprising only highly confident target samples and employs a point-to-set distance function to generate the pseudo labels. CAiDA [121] 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 [122], a few works [119, 97, 120, 123] 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 [120] sharpens the network output \bar{p} with the class frequency to avoid class imbalance, while NRC [87] devises a weighting scheme for neighbors during aggregation. Instead of using the soft pseudo label \hat{p} , AdaContrast [97] utilizes the hard pseudo label with the argmax operation.

Complementary pseudo labels. Motivated by the idea of negative learning [124], PR-SFDA [125] randomly chooses a label from the set $\{1, \ldots, C\} \setminus \{\hat{y}_i\}$ as the complementary label \bar{y}_i and thus optimizes the following loss function,

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

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 [89] develops a heuristic strategy to randomly select an informative complementary label with medium prediction scores. Besides, NEL [126] and PLUE [123] randomly select multiple complementary labels, except for the inferred pseudo label, and optimizes the multiclass variant of Eq. (3). ATP [90] further generates multiple complementary labels according to a pre-defined threshold on prediction scores.

Optimization-based pseudo labels. By leveraging the prior knowledge of the target label distribution like class balance [127], some TTDA methods [89, 95, 128] 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 [91] directly imposes the equipartition constraint on the pseudo labels \hat{p}_i and solves

the optimization problem below,

$$\begin{split} & \min_{\hat{p}_{i}} - \sum_{i} \sum_{c} \hat{p}_{ic} \log p_{\theta}(y_{c}|x_{i}) + \lambda \sum_{i} \sum_{c} \hat{p}_{ic} \log \hat{p}_{ic}, \\ & s.t. \ \forall i, c: \ \hat{p}_{ic} \in [0, 1], \ \sum_{c} \hat{p}_{ic} = 1, \ \sum_{i} \hat{p}_{ic} = \frac{n_{t}}{C}. \end{split}$$

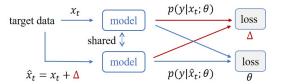
$$(4)$$

Likewise, IterNLL [129] provides a closed-form solution of $\{\hat{p}\}$ under the uniform prior assumption. KUDA [92] even introduces a hard constraint $\hat{p}_{ic} \in \{0,1\}$ and solves the zero-one programming problem. In addition, Re-CLIP [130] constructs the affinity graph and employs label propagation to produce closed-form pseudo labels.

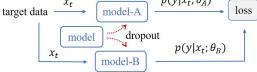
Ensemble-based pseudo labels. Rather than relying on a single noisy pseudo label, ISFDA [101] generates a secondary pseudo label to aid the primary one. Besides, ASL [91] and C-SFDA [131] adopt a weighted average of predictions under multiple random data augmentation, while ELR [132] ensembles historical predictions from previous training epochs. NEL [126] further aggregates logits under different data augmentation and trained models simultaneously. Inspired by a classic semi-supervised learning method [71], some TTDA methods [133, 134] maintain an EMA of predictions at different time steps as pseudo labels. Moreover, C-SFDA [131] maintains a mean teacher model [72] 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 [133, 135], a multi-head classifier [136], and models from both domains [113]. In particular, SFDA-VS [137] follows MC dropout [81] and obtains the final prediction through multiple forward passes.

Another line of ensemble-based TTDA methods [119, 91, 138] aims to integrate predictions from different labeling criteria using a weighted average. For example, e-SHOT-CE [119] utilizes both centroid-based and neighbor-based pseudo labels. Besides the weighting scheme, other approaches [108, 121, 118, 139] explore different labeling criteria in a cascade manner. For instance, DIPE [118] employs the neighbor-based labeling criterion with centroid-based pseudo labels.

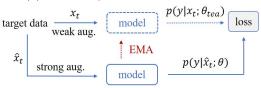
Learning with pseudo labels. Existing pseudolabeling-based TTDA methods have employed various robust divergence measures d_{pl} . Generally, most methods utilize the standard cross-entropy loss for all target samples with hard pseudo labels [7, 91] or soft pseudo labels [114, 140]. Note that several methods [120, 141] convert hard pseudo labels into soft pseudo labels using the label smoothing trick [142]. As pseudo labels are noisy, many TTDA methods incorporate an instancespecific weighting scheme into the standard cross-



(a) consistency under data variations t data $\xrightarrow{x_t}$ $\xrightarrow{\text{model-A}}$ $\xrightarrow{p(y|x_t;\theta_A)}$ $\xrightarrow{\text{los}}$



(b) consistency under model variations



(c) consistency under data & model variations

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 .

entropy loss, including hard weights [110, 113, 112], and soft weights [128, 137]. Besides, AUGCO [143] 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 [144], the inner product distance between the pseudo label and the prediction [87, 108], and a new discrepancy measure $\log(1-\hat{y}^Tp(y|x;\theta))$ [132]. Moreover, BMD [86] and OnDA [134] employ the symmetric cross-entropy loss to guide the self-labeling process. CATTAn [145] 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 [146, 66], 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. Fig. 3 illustrates three representative types of consistency training, which will be elaborated in the following part.

Consistency under data variations. Benefiting from advanced data augmentation techniques such as RandAugment [147], several prominent semi-supervised

learning methods [148, 70] unleash the power of consistency regularization over unlabeled data that can be effortlessly adopted in TTDA approaches. An exemplar of consistency regularization [70] is expressed as:

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

where $p_{\theta}(y|x_i) = p(y|x_i;\theta)$, and $\text{CE}(\cdot,\cdot)$ refers to crossentropy 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 virtual adversarial training (VAT) [69], which devises a smoothness constraint as follows,

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

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 [90] directly employs the same consistency regularization in Eq. (5), while other TTDA methods [149, 97, 106, 139] 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 [105] and SFDA-FSM [150] require learning the domain translation module first, and ProSFDA seeks featurelevel consistency under different augmentations at the same time. TeST [151] introduces a flexible mapping network to match features under two different augmentations. On the contrary, OSHT [152] 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 [11, 91] attempts to encourage consistency between target samples with their data-level neighbors, while APA [94] learns the neighbors in the feature space. Instead of generating the most divergent neighbor $x_i + \Delta_i$ according to the predictions, JN [153] devises a Jacobian norm regularization to control the smoothness in the neighborhood of the target sample. Furthermore, G-SFDA [93] discovers multiple neighbors from a memory bank and minimizes their inner product distances over the predictions. Moreover, Mixup [154] performs linear interpolations on two inputs and their corresponding labels, which can be treated as seeking consistency under data variation [133, 116, 139].

Consistency under model variations. Reducing model uncertainty [81] is also beneficial for learning ro-

bust features for TTDA tasks, on top of uncertainty measured with input change. Following MC dropout [81], FAUST [117] activates dropout in the model and performs multiple stochastic forward passes to estimate the epistemic uncertainty. SFDA-UR [95] appends multiple extra dropout layers behind the feature encoder and minimizes the mean squared error (MSE) between predictions as uncertainty. Further, ASFA [99] adds different perturbations to the intermediate features to promote predictive consistency. FMML [96] 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 [155] and output-level discrepancy [133, 156, 151]. Furthermore, the mean teacher framework [72] 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 [157] 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 [72] is enhanced by blending strong data augmentation techniques, and the discrepancy between predictions of the student and teacher models is minimized as follows,

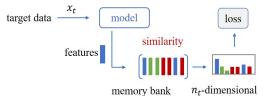
$$\mathcal{L}_{mt}^{con} = \mathbb{E}_{x \in \mathcal{D}_t} d_{mt}(p(y|x,\theta), p(y|\tau(x), \theta_{tea})), \tag{7}$$

where $\tau(\cdot)$ denotes the strong data augmentation, and d_{mt} denotes the divergence measure, e.g., the KL divergence [102, 113], the MSE loss [158], and the cross-entropy loss [88]. Besides, several methods [159, 4, 128, 160] 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 [159] matches the features extracted by different models with the MSE distance, while AdaContrast [97] and PLUE [123] learn semantically consistent features like MoCo [60].

Instead of strong data augmentations, LODS [160] and SFIT [113] use the style transferred image instead, MAPS [98] considers spatial transforms, and SMT [158] elaborates the domain-specific perturbation by averaging the target images. Different from model variations



(a) uncertainty minimization over network predictions



(b) clustering-promoting over network features

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

in the mean teacher scheme, OnTA [161] distills knowledge from the source model to the target model, while HCL [128] 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 [11, 138] 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 [162] and the MSE error [158, 163], whereas OSUDA [164] encourages the learned scaling and shifting parameters in BN layers to be consistent. Similarly, an explicit feature-level regularization [165] 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 [166] 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 [141, 118, 167, 95, 99] through modeling the consistency between different classifiers. Besides, SFDA-APM [110] 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 [146], 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 [7, 11] or promoting clustering among the target features [101, 108]. Fig. 4 illustrates these two representative types of clustering-based training, which will be elaborated in the following part.

Entropy minimization. ASFA [99] utilizes robust measures from information theory to encourage confident predictions for unlabeled target data. To achieve this, it minimizes the α -Tsallis entropy given by:

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

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 [11, 164, 95, 89, 136, 168, 151]. Besides, there exist numerous variations of standard entropy minimization. For instance, SFDA-VS [137] develops a nonlinear weighted entropy minimization loss that emphasizes low-entropy samples. TT-SFUDA [159] 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 [29, 169, 139], 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) [30] approximates the prediction diversity using the matrix rank, which is utilized by CDL [149]. Additionally, SI-SFDA [170] 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 [171] between the inputs and the discrete

labels as follows,

$$\max_{\theta} \mathcal{I}(\mathcal{X}_t, \hat{\mathcal{Y}}_t) = \mathcal{H}(\hat{\mathcal{Y}}_t) - \mathcal{H}(\hat{\mathcal{Y}}_t | \mathcal{X}_t) =$$

$$-\sum_{c=1}^{C} \bar{p}_{\theta}(y_c) \log \bar{p}_{\theta}(y_c) + \frac{1}{n_t} \sum_{i=1}^{n_t} \sum_{c=1}^{C} p_{\theta}(y_c | x_i) \log p_{\theta}(y_c | x_i),$$

$$(9)$$

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 [7] and SHOT++ [65] for image classification and then employed in plenty of TTDA methods [162, 157, 161, 153, 118, 123]. Instead of using the network prediction $p_{\theta}(y|x)$, GKD [114] employs the ensemble prediction based on its neighbors for mutual information maximization. DaC [106] and U-SFAN [172] introduce a balancing parameter between two terms in Eq. (9) to increase flexibility. In particular, U-SFAN [172] develops an uncertainty-guided entropy minimization loss by emphasizing low-entropy predictions, whereas ATP [90] encompasses the instance-wise uncertainty in both terms of Eq. (9). VMP [173] 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 [113, 87, 97, 109, 134, 141, 134, 145] to prevent class collapse. To better guide the learning process, a few works [174, 175] modify the mutual information regularization by substituting a reference class-ratio distribution in place of \mathcal{U} . Unlike AdaMI [168], which leverages the target class ratio as a prior, UMAD [100] utilizes the flattened label distribution within a mini-batch instead to mitigate the class imbalance problem, and AUGCO [143] maintains the moving average of the predictions as the reference distribution.

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 [176, 11, 177], data translation [113, 178, 179],

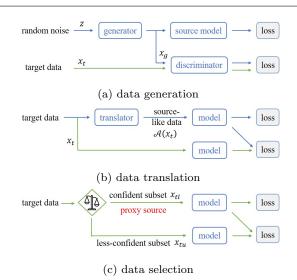


Fig. 5 Three representative types of source distribution estimation, where surrogate source data is obtained through generation, translation, and selection, respectively.

and data selection [65, 166, 120]. Fig. 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 [11] 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 [180] that is written as follows:

$$\min_{\theta_G} \max_{\theta_D} \mathbb{E}_{x_t \in \mathcal{X}_t} [\log D(x_t)] + \mathbb{E}_{y_t, z} [\log (1 - D(G(y_t, z)))]
- \lambda_s \mathbb{E}_{y_t, z} \sum_{c} \mathbb{1}(y_t = c) \log p(y_c | G(y_t, z), \theta),$$
(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 [176] disregards the last term in Eq. (10) to infer diverse target-like samples. On the other hand, SDDA [177] 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 [103] 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,

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

which is referred to as data impression of the source domain. Besides, SPGM [181] 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 [182, 20], SFDA-KTMA [169] 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,

$$\mathcal{L}_{bn} = \sum_{l} \sum_{i} \|\mu_{g,l}^{(i)} - \mu_{s,l}^{(i)}\|_{2} + \|\delta_{g,l}^{(i)^{2}} - \delta_{s,l}^{(i)^{2}}\|_{2}, \quad (12)$$

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 batchwise 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 [183], matching the BN statistics can aid in ensuring that the generated data resembles the source style. SFDA-FSM [150] further minimizes the L_2 -norm difference between intermediate features (a.k.a., the content loss [184]) to preserve the content knowledge of the target domain.

Data translation. SSFT-SSD [178] 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 [185] 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 [186], ProSFDA [105] 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 styletransferred image is obtained using spectrum mixup [187] between the target image and its perturbed image.

Another line of data translation methods [104, 113, 179] explicitly introduces an additional module \mathcal{A} to transfer target data to source-like style. In particular, SFDA-IT [104] 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 [104] employs entropy minimization over the fixed source model to promote semantic consistency. To improve the performance of style transfer, SFIT [113] further develops a variant of the style reconstruction loss [184] as follows,

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

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 [113] maintains the relationship of outputs between different networks. GDA [179] 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 [65, 166, 120, 90, 88, 188] 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++ [65] 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 [137, 169, 90] where the ratio or the number of selected samples per class is always kept same to prevent severe class imbalance. DaC [106] utilizes the maximum softmax probability instead of the entropy criterion. Furthermore, BETA [188] 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 [120, 189, 190] utilize the source model to partition the target domain before the intra-domain adaptation step. For each class separately, MTRAN [190] selects the low-entropy sample, ProxyMix [120] leverages the distance from target features to source class prototypes, and SAB [189] adopt the maximum prediction probability. To simulate the source domain more accurately, MTRAN [190] further applies the mixup augmentation technique after the dataset partition step. On the other hand, some TTDA methods [166, 167, 137, 88, 191, 141] do not fix the domain partition but alternately update the domain partition and learn the target model in the adaptation step. For instance, SSNLL [88] 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 [166] and SFDA-KTMA [169] split each mini-batch into two sets based on the criterion of entropy ranking, while D-MCD [191] employs

the classifier determinacy disparity and the agreement between different self-labeling strategies.

Feature estimation. In contrast to source data synthesis, previous works [108, 107, 192] provide a cost-effective alternative by simulating the source features. MAS³ [193] and LDAuCID [194] require learning a GMM over the source features before model adaptation, which may not hold in real-world scenarios. Instead, VDM-DA [107] constructs a proxy source domain by randomly sampling features from the following GMM,

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

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 [195] 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 [192] 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 [108] 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++ [65] and ProxyMix [120] follow a classic semi-supervised approach, MixMatch [196], to bridge the domain gap. Secondly, SDDA [177] adopts the widely-used domain adversarial alignment technique [27] that is formally written as:

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

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 [169, 137, 107]. Besides, a certain number of following methods [103, 178, 193] further employ advanced domain adversarial training strategies to achieve better adaptation. Thirdly, BAIT

[166] leverages the maximum classifier discrepancy [197] between two classifiers' outputs in an adversarial manner to achieve feature alignment, which has been followed by [141, 191]. Fourthly, some TTDA methods [192, 106, 189] explore the maximum mean discrepancy (MMD) [198] 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 [108, 106]. To model the instance-level alignment, MTRAN [190] 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 [57, 58, 59, 60, 61]. As mentioned earlier, the centroid-based pseudo labels are similar to the learning manner of DeepCluster [58]. Inspired by rotation prediction [57], SHOT++ [65] further comes up with a relative rotation prediction task and introduces an additional 4-way classification head during adaptation. Besides, OnTA [161] and CluP [156] exploit the self-supervised learning frameworks [60, 59] for learning discriminative features as initialization, respectively. TTT++ [165] learns an extra self-supervised branch using contrastive learning [61] in the source model, which facilitates the adaptation in the target domain with the same objective. FedICON [199] leverages unsupervised contrastive learning to guide the model to smoothly generalize to test data under intra-client heterogeneity. Recently, StickerDA [109] 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 [200] and POUF [201] 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 [138]. LA-VAE [202] exploits the variational auto-encoder to achieve latent feature alignment. In addition, the meta-learning mechanism is adopted in a few studies [203, 204] for the TTDA problem. A recent work [205] 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., $C_s = C_t$, and some TTDA algorithms [7, 128] are also validated in a relaxed partial-set setting [206], i.e., $C_t \subset C_s$. However, several TTDA works [7, 207, 152] consider the open-set learning scenario where the target label space C_t subsumes the source label space C_s . To allow more flexibility, open-partial-set domain adaptation [208] $(C_s \setminus C_t \neq \emptyset, C_t \setminus C_s \neq \emptyset,)$ is studied in TTDA methods [10, 140, 209]. Moreover, several recent studies [100, 210] 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 [7, 65, 211] extend the single-source TTDA algorithms by combining these adapted models together in the target domain. Besides, a couple of works [212, 121] 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 [135] 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 [126, 139] 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 [194, 134].

Unsupervised v.s. Semi-supervised. Some TTDA methods [149, 213] adapt the source model to the target domain with only a few labeled target samples and adequate unlabeled target samples. In these semisupervised learning scenarios, the standard classification loss over the labeled data could be readily incorporated to enhance the adaptation performance [65, 149]. 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 blackbox module through the cloud application programming interface (API). At an early time, IterLNL [129] treats this black-box TTDA problem as learning with noisy labels, and DINE [133] develops several structural regularizations within the knowledge distillation framework. These approaches inspire many recent black-box TTDA works [92, 96, 189, 188]. Beyond the deep learning framework, several shallow studies [82, 83] 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.[32] 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 [33, 214]. We refer interested readers to relevant literature [215]. A few methods such as ISFDA [101] and APA [94] pay attention to the class-imbalanced TTDA scenario where both data and label shifts are present.

Active TTDA. To improve the limited performance gains, MHPL [216] 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 [217, 155], 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 [218], the forgetting of source knowledge [93, 115], and the vulnerability to membership inference attack [219] and image-agnostic attacks (e.g., blended backdoor attack) [220].

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 [8] and one-sample generalization [221]), 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. [222] 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

³ https://openai.com/blog/openai-api/

 ${\bf Table~2}~{\rm A~taxonomy~on~TTBA~methods~with~representative~strategies.}$

Families	Representative Strategies
BN calibration	PredBN [225, 13], InstCal [226]
model optimization	TTT [8], GeOS [221], MEMO [12]
meta-learning	MLSR $[227]$, Full-OSHOT $[228]$
input adaptation	TPT [229], TTA-DAE [230]
dynamic inference	LAME [231], EMEA [232]

as in single-image models that are tested on real-time video data [223, 224]. 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 [7] when the batch size B is sufficiently large. Conversely, when the batch size B is equal to 1, TTBA degrades to TTIA [8]. 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.

4.2 Taxonomy on TTBA Algorithms

4.2.1 Batch Normalization Calibration

Normalization layers (e.g., batch normalization [233] and layer normalization [234]) 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,

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

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 $\mu_s = \mathbb{E}[\mathcal{X}_S]$ and variance $\sigma_s^2 = \mathbb{V}[\mathcal{X}_S]$) are typically approximated using

EMA over batch-level estimates $\{\hat{\mu}_k, \hat{\sigma}_k^2\}$,

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

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

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

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 [235], 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. PredBN [225], a pioneering TTBA method, substitutes the training BN statistics with those estimated per test batch.

PredBN+ [13] 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,

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

where the test statistics $\{\hat{\mu}_t, \hat{\sigma}_t^2\}$ are estimated via Eq. (18), and the hyper-parameter ρ_t controls the tradeoff between training and estimated test statistics. Moreover, TTN [236] presents an alternative solution that calibrates the estimation of the variance as follows,

$$\bar{\sigma}_t^2 = (1 - \rho_t) \cdot \sigma_s^2 + \rho_t \cdot \hat{\sigma}_t^2 + \rho_t (1 - \rho_t) (\hat{\mu}_t - \mu_s)^2.$$
 (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 [237] 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 [238] directly attempts to mix instance normalization (IN) [239] statistics estimated per instance with the training BN statistics. Instead of manually specifying a fixed value at test time, InstCal [226] 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 [240] 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 [221, 8, 241], 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 [242, 12, 243], which elaborately designs a task-specific objective for updating the pre-trained model.

Training with auxiliary tasks. Motivated by prior works [49, 244] in which incorporating self-supervision with supervised learning in a unified multi-task framework enhances adaptation and generalization, TTT [8] and OSHOT [241] 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,

$$\theta_e^*, \theta_c^*, \theta_s^* = \underset{\theta_e, \theta_c, \theta_s}{\arg\min} \sum_{i=1}^{n_s} \mathcal{L}_{pri}(x_i, y_i; \theta_e, \theta_c) + \mathcal{L}_{ssl}(x_i; \theta_e, \theta_s),$$
(21)

where \mathcal{L}_{pri} denotes the primary objective (e.g., crossentropy for classification tasks), and \mathcal{L}_{ssl} denotes the auxiliary SSL objective (e.g., rotation prediction [57] and solving jigsaw puzzles [49]). For each test instance x_t , TTT [8] first adjusts the feature encoder $f_e(\cdot; \theta_e)$ by optimizing the SSL objective,

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

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 [241] 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 [245, 246, 247]. Among them, TTT-MAE [247] is a recent extension of TTT that utilizes the transformer backbone and replaces the self-supervision with masked autoencoders [248].

To increase the dependency between the primary task and the auxiliary task, GeOS [221] further adds the features of the SSL head to the primary head. SR-TTT [249] 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 [250] 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 [242] proposes a selective entropy minimization objective for pixel-level semantic segmentation, while MALL [243] enforces edge consistency prior through a weighted normalized cut loss. Besides, MEMO [12] optimizes the entropy of the averaged prediction over multiple random augmentations of the input sample. PromptAlign [251] additionally handles the train-test distribution shift by matching the mean and variances of the test sample and the source dataset statistics. TTAS [252] further develops a class-weighted entropy objective, while SUTA [253] additionally incorporates minimum class confusion to reduce the uncertainty. A recent work [254] 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 [255, 256]. In particular, SCIO [257] develops a self-constrained optimization method to learn the coherent spatial structure. While adapting image models to a video input [223, 258], 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 [259], image denoising [260], generative modeling [261], and style transfer [262]. In addition, the model could be adapted to each instance by utilizing the generated data at test time. As an illustration, TTL-EQA [263] generates numerous synthetic question-answer pairs and subsequently leverages them to infer answers in the given context. ZSSR [264] trains a super-resolution network using solely down-sampled examples extracted from the test image itself.

4.2.3 Meta-Learning

MAML [51], a notable example of meta-learning [53], 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 [227, 228], and forward propagation [265, 266]. 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 [264], MLSR [227] develops a meta-learning method based on MAML for single-image superresolution. Concretely, the meta-objective w.r.t. the network parameter θ is shown as,

$$\min_{\theta} \sum_{i} \mathcal{L}(LR_{i}, HR_{i}; \theta - \alpha \nabla_{\theta} \mathcal{L}(LR_{i} \downarrow, LR_{i}; \theta)), \quad (23)$$

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 $LR_i \downarrow$ denotes the down-scaled version of the low-resolution input in the paired trained data (LR_i, 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,

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

then generates the high-resolution (HR) image as $f_{\theta_t}(\text{LR})$. Such a meta-learning mechanism based on self-supervised learning has been utilized by follow-up methods [267, 268, 269]. Among them, MetaVFI [270] further introduces self-supervised cycle consistency for video frame interpolation.

As an alternative, Full-OSHOT [228] 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:

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

and the definitions of variables are the same as OSHOT [241] 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 [271], 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 [271] and reconstruction [272, 273].

Forward propagation. Apart from the shared encoder $f_e(\theta_e)$ above, several other meta-learning methods exploit the normalization statistics [14, 274] or domain prototypes [265, 266] from the inner loop, allowing backward-free adaptation at inference time. Besides, some works incorporate extra meta-adjusters [275] or learnable prompts [276], 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 [277] 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 [278], SSGen [279] 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 [280, 281] 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 [230, 282, 283]. For example, TPT [229] freezes the pre-trained multimodal model and only learns the extra text prompt based on the marginal entropy of each instance. Another approach, CVP [284], optimizes the convolutional visual prompts in the input under the guidance of a self-supervised contrastive learning objective.

TTA-AE [285] additionally learns a set of autoencoders 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 [230] 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 [286] 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 [287] 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 [288] proposes mapping the target input onto the source data manifold using Fourier style transfer [187], serving as a pre-processor to the primary network. By contrast, TAF-Cal [282] further utilizes the average amplitude feature over the training data to perform Fourier style calibration [187] at both training and test phases, bridging the gap between training and test data. It is noteworthy that imposing a data manifold constraint [289, 250, 283, 290] can aid in achieving better alignment between the test data and unseen training data. Specifically, ITTP [289] trains a generative model over source features with target features projected onto

points in the source feature manifold for final inference. DDA [283] exploits the generative diffusion model for target data, while ESA [290] 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 [291, 292, 293, 294]. Among them, Anti-Adv [294] perturbs the test input to maximize the classifier's prediction confidence. Besides, SOAP [291] leverages self-supervisions like rotation prediction at both training and test phases and purifies adversarial test examples based on self-supervision only. SSRA [293] 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 [231] 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 [232, 295] learn the weights for each model, without making any changes to the models themselves. For example, EMEA [232] employs entropy minimization to update the ensemble coefficients before each model. GPR [296] 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 [8, 12] and batch adaptation [13, 223], 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., [221, 289, 232, 279, 282, 290, 295]) 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., [296, 297, 295]) 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 [8] 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 [12, 294] 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 [7, 11], a mini-batch [13, 14], or even a single instance [8, 12] 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 [8] 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_S on the source domain \mathcal{D}_S and a sequence of unlabeled minibatches $\{\mathcal{B}_1, \mathcal{B}_2, \cdots\}$, online test-time adaptation aims to leverage the labeled knowledge implied in f_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 [9], 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 [8] 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 [16] and EATA [298] investigate the continual testtime 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.

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) [233] are commonly employed in modern neural networks. Typically, BN

 ${f Table~3}$ A taxonomy on OTTA methods with representative strategies.

Families	Representative Strategies
BN calibration	DUA [299], DELTA [300]
entropy minimization	Tent [9], SAR [301]
pseudo-labeling	T3A [15], TAST [302]
consistency regularization	CFA [303], PETAL [304]
anti-forgetting regularization	CoTTA [16], EATA [298]

layers can encode domain-specific knowledge into normalization statistics [235]. A recent work [301] 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 [9] and RNCR [305] 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 [306] 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 [307] 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.

Similar to the running average estimation of BN statistics during training, ONDA [308] 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,

$$\mu_t = \rho \hat{\mu}_t + (1 - \rho)\mu_{t-1},$$

$$\sigma_t^2 = \rho \hat{\sigma}_t^2 + (1 - \rho)\frac{n_t}{n_t - 1}\sigma_{t-1}^2,$$
(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 [309] considers a heuristic weight through computing the distance between $\{\mu_{t-1}, \sigma_{t-1}\}$ and $\{\hat{\mu}_t, \hat{\sigma}_t\}$. EDTN [310] 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 [311] and DELTA [300] adopt the following reformulation of batch renormalization [312],

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

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 uti-

lize the test-time dataset-level running statistics via the moving average like Eq. (26).

For online adaptation with a single sample, MixNorm [313] mixes the estimated IN statistics with the exponential moving average BN statistics at test time. On the other hand, DUA [299] 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 [314] 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 outof-distribution samples. TN-SIB [315] 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 [9], 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 [314, 311]. Notably, VMP [173] reformulates Tent in a probabilistic framework by introducing perturbations into the model parameters by variational Bayesian inference. Several other methods [316, 317] also focus on minimizing the entropy at test time but utilize different combinations of learnable parameters. BACS [318] 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 [319] proposes minimizing the average entropy of predictions under different augmentations. FEDTHE+ [320] employs the same adaptation scheme as MEMO [12] that minimizes the entropy of the average prediction over different augmentations.

To avoid overfitting to non-reliable and redundant test data, EATA [298] 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 [321, 301]. CD-TTA [306] 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 [300] 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 [307], batch nuclear-norm maximization [305], maximum squares loss [306], and mutual information maximization [322, 323]. In addition, MuSLA [322] further considers the virtual adversarial training objective that enforces classifier consistency by adding a small perturbation to each sample. SAR [301] encourages the model to lie in a flat area of the entropy loss surface and optimizes the minimax entropy objective below,

$$\min_{\theta} \max_{\|\Delta_{\theta}\|_{2} \le \epsilon} \mathcal{H}(x; \theta + \Delta_{\theta}), \tag{28}$$

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 [324, 325] even employ entropy maximization for specific tasks, for example, AUTO [325] performs model updating for unknown samples at test time.

5.2.3 Pseudo-labeling

Unlike the unidirectional process of entropy minimization, many OTTA methods [326, 322, 231, 306] adopt pseudo labels generated at test time for model updates. Among them, MM-TTA [327] proposes a selective fusion strategy to ensemble predictions from multiple modalities. Besides, DLTTA [328] 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 [302] proposes a similar approach that reduces the difference between predictions from a prototype-based classifier and a neighborbased classifier. Notably, SLR+IT [329] 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 [330] 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 [331] theoretically analyzes the difference between hard and conjugate labels under gradient descent for a binary classification problem. Motivated by the idea of negative learning [124], ECL [332] further considers complementary labels from the least probable categories. Besides, T3A [15] 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 [72] framework, the pseudo labels under weak data augmentation obtained by the teacher network are known to be more stable. Built on this framework, RMT [333] pursues the teacherstudent consistency in predictions through a symmetric cross-entropy measure, while OIL [334] only exploits highly confident samples during consistency maximization. VDP [335] utilizes this framework to update visual domain prompts with the pre-trained model being frozen. Moreover, CoTTA [16] further employs multiple augmentations to refine the pseudo labels from the teacher network, which is also applied in other methods [304, 336, 337]. Inspired by maximum classifier discrepancy [197], AdaODM [338] proposes minimizing the prediction disagreement between two classifiers at test time to update the feature encoder.

Apart from the model variation above, several methods [319, 339, 340, 341, 342] also enforce the consistency of the corresponding predictions among different augmentations. In particular, SWR-NSP [323] introduces an additional nearest source prototype classifier at test time and minimizes the difference between predictions under two different augmentations. Besides, many methods [343, 344, 345, 326, 317] leverage the temporal coherence for video data and design a temporal consistency objective at test time. For example, TeCo [317] 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+ [320] pursues consistency in the feature space. Several other OTTA methods [346, 333, 341]) even pursue consistency between test features and source or target prototypes in the feature space. CFA [303] further proposes matching multiple central moments to achieve feature alignment. Furthermore, ACT-MAD [347] performs feature alignment by minimizing the discrepancy between the pre-computed training statistics and the estimates of test statistics. TTAC [341] calculates the online estimates of feature mean and variance at test time instead. Besides, CAFA [348] 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 [16, 298] 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 [326, 333, 344]. PAD [346] comes up with an alternative approach that keeps the relative relationship of irrelevant auxiliary data unchanged after test-time optimization. AUTO [325] 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 [9] only optimizes the affine parameters in the BN layers for test-time adaptation, and AUTO [325] updates the last feature block in the pre-trained model. SWR-NSP [323] divides the entire model parameters into shift-agnostic and shift-biased parameters and updates the former less and the latter more. Recently, VDP [335] fixes the pre-trained model but only optimizes the input prompts during adaptation.

Besides, CoTTA [16] proposes a stochastic restoration technique that randomly restores a small number of parameters to the initial weights in the pretrained model. PETAL [304] further selects parameters with smaller gradient norms in the entire model for restoration. By contrast, EATA [298] 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 [301] 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 [321] 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 [315, 349], Hebbian learning [316], and adversarial data augmentation [336]. TDA [350] 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 [9] that assumes the test data comes from a stationary distribution, dynamic OTTA assumes a dynamically changing distribution including continual OTTA [16], temporal OTTA [314], gradual OTTA [333], and practical OTTA [351]. A recent study [352] 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 [353, 354, 355] investigate changes in label distribution. Two interesting cases with online

feedback are studied in [354], *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 [7, 65, 106]. 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 [8, 13, 9, 12]. In addition, TTBA and OTTA methods are also evaluated in DG datasets such as VLCS, PACS, and Office-Home, as described in previous studies [221, 289, 15, 335].

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 [95, 169, 90]. 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 [226, 238]. OTTA methods further utilize Cityscapes-to-ACDC and Cityscapes-to-Foggy&Rainy Cityscapes for evaluation and comparison, as described in [16, 356].

6.3 Object Detection

Object detection is a fundamental computer vision task that involves locating instances of objects in images. While early TTA methods [357, 358] 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-

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

to-Cityscapes, Pascal-to-Clipart&Watercolor are commonly used by TTDA methods for evaluation and comparison, as detailed in [111, 128, 160, 151]. Additionally, datasets like VOC-to-Social Bikes and VOC-to-AMD are employed to evaluate TTBA methods [241, 228].

6.4 Beyond Vanilla Object Images

Medical images. Medical image analysis is another important downstream field of TTA methods, e.g., medical image classification [359, 360], medical image segmentation [285, 230], and medical image detection [4]. 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 [107], 3D semantic segmentation [361], and 3D object detection [362].

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 [273] and frame interpolation [270]. Besides, a few studies investigate the TTDA scheme for other video-based tasks including action recognition [363, 190, 317, 364], optical flow estimation [365] and object segmentation [366].

Multi-modal data. Researchers also develop different TTA methods for various multi-modal data, e.g., RGB and audio [367], RGB and depth [163, 327], RGB and motion [190], and image-text pairs [368]. Furthermore, the development of multi-modal pre-trained models such as CLIP [369] enables image classification through image-to-text matching, gaining popularity among recent TTA methods [251, 237, 337, 254].

Face and body data. Facial data is also an important application of TTA methods, such as face recognition [370], face anti-spoofing [203, 102, 179], and expression recognition [156]. For body data, TTA methods also pay attention to tasks such as pose estimation [245, 257, 98] and mesh reconstruction [343, 258].

6.5 Beyond Vanilla Recognition Problems

Low-level vision. TTA methods can be applied to low-level vision problems, *e.g.*, image super-resolution [227, 371], image deblurring [267], and image dehazing [268]. Besides, TTA is also introduced to image registration [372, 259], inverse problems [373, 374], and quality assessment [375].

Retrieval. Besides classification problems, TTA can also be applied to kinds of retrieval scenarios, *e.g.*, person re-identification [376, 280], sketch-to-image retrieval [272, 377], image-text matching [237], and fair image retrieval [378].

Generative modeling. TTA method can also vary the pre-trained generative model for style transfer and data generation [261, 262, 379].

Defense. Another interesting application is test-time adversarial defense [291, 292, 294], 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 [263], question answering [334], sentiment analysis [380], entity recognition [232], and aspect prediction [276]. In particular, a competition ⁵ has been launched under data sharing restrictions, comprising two NLP semantic tasks [381]: 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 [382] or node classification [256].

Speech processing. As far, there have been three TTA methods, *i.e.*, audio classification [383], speaker verification [266] and speech recognition [253].

Miscellaneous signals. TTA methods have been also validated on other types of signals, *e.g.*, radar signals [119], EEG signals [384], and vibration signals [385].

Reinforcement learning. Some TTA methods [246, 386] 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 [95, 90] and classification on Domain-Net, 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 [166] 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 [16, 298], the evaluation of each mini-batch is conducted before optimization on that mini-batch. This

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

manner differs from the standard evaluation protocol of OTTA [8] 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 [363], multi-modal data [327], and 3D point clouds [361], as well as regression problems like pose estimation [98].

Open-world adaptation. Existing TTA methods always follow the closed-set assumption; however, a growing number of TTDA methods [100, 209, 210] are beginning to explore model adaptation under an open-set setting. A recent OTTA method [325] 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 [217, 155] 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 [16, 298], leading to catastrophic forgetting. To reduce memory consumption while maintaining accuracy, recent works [321, 309] 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 [9], 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 [387] 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 [229, 388, 251, 237, 254, 389] developed that leverage vision-language models, such as CLIP [369], to enhance the zero-shot generalization. Meanwhile, some studies have focused on CLIP adaptation under the OTTA scenario [337, 350] as well as the TTDA setting [201, 130]. Additionally, several recent studies [388, 390] have explored leveraging large-scale generative models, such as Stable Diffusion [391], 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 [331] and auxiliary self-supervision [8]. One recent work [392] conducts an indepth 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 [393, 394, 395] are proposed to fairly evaluate various TTA methods. For example, the vision transformer (ViT) architecture is further employed for online TTA methods in [395], and a new dataset is developed to testify online TTA methods under continuously changing corruptions [394]. However, as there does not exist a labeled validation set, validation also remains a significant and unsolved issue for TTA methods. As noted in [396], 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 [393] 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 [397] in vectors of heterogeneous features is essential for industrial applications, and time series data [398] is predominant in

heterogeneous features is essential for industrial applications, and time series data [398] 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 [399, 400]. 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 [401], 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 [220]. 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 [402, 389].

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

The authors sincerely thank the editor and anonymous reviewers for their constructive comments on this work.

References

- Joaquin Quinonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. Dataset shift in machine learning. MIT Press, 2008.
- Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell. Adapting visual category models to new domains. In *Proc. ECCV*, pages 213–226, 2010.
- Yi-Hsin Chen, Wei-Yu Chen, Yu-Ting Chen, Bo-Cheng Tsai, Yu-Chiang Frank Wang, and Min Sun. No more discrimination: Cross city adaptation of road scene segmenters. In *Proc. ICCV*, pages 1992–2001, 2017.
- Xinyu Liu and Yixuan Yuan. A source-free domain adaptive polyp detection framework with style diversification flow. *IEEE Transactions on Medical Imaging*, 41(7):1897–1908, 2022.
- Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
- 6. Wouter M Kouw and Marco Loog. A review of domain adaptation without target labels. *IEEE Transactions*

- on Pattern Analysis and Machine Intelligence, 43(3):766–785, 2019.
- Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? source hypothesis transfer for unsupervised domain adaptation. In *Proc. ICML*, pages 6028–6039, 2020.
- Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *Proc. ICML*, pages 9229–9248, 2020.
- 9. Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *Proc. ICLR*, 2021.
- Jogendra Nath Kundu, Naveen Venkat, and R Venkatesh Babu. Universal source-free domain adaptation. In Proc. CVPR, pages 4544-4553, 2020.
- Rui Li, Qianfen Jiao, Wenming Cao, Hau-San Wong, and Si Wu. Model adaptation: Unsupervised domain adaptation without source data. In *Proc. CVPR*, pages 9641–9650, 2020.
- 12. Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *Proc. NeurIPS*, pages 38629–38642, 2022.
- Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Proc. NeurIPS*, pages 11539–11551, 2020.
- Marvin Zhang, Henrik Marklund, Nikita Dhawan, Abhishek Gupta, Sergey Levine, and Chelsea Finn. Adaptive risk minimization: Learning to adapt to domain shift. In *Proc. NeurIPS*, pages 23664–23678, 2021.
- Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Proc. NeurIPS*, pages 2427–2440, 2021.
- Qin Wang, Olga Fink, Luc Van Gool, and Dengxin Dai. Continual test-time domain adaptation. In Proc. CVPR, pages 7201–7211, 2022.
- 17. Garrett Wilson and Diane J Cook. A survey of unsupervised deep domain adaptation. ACM Transactions on Intelligent Systems and Technology, 11(5):1-46, 2020.
- Yuqi Fang, Pew-Thian Yap, Weili Lin, Hongtu Zhu, and Mingxia Liu. Source-free unsupervised domain adaptation: A survey. Neural Networks, page 106230, 2024.
- Jingjing Li, Zhiqi Yu, Zhekai Du, Lei Zhu, and Heng Tao Shen. A comprehensive survey on source-free domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- Yuang Liu, Wei Zhang, Jun Wang, and Jianyong Wang. Data-free knowledge transfer: A survey. arXiv preprint arXiv:2112.15278, 2021.
- 21. Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2009.
- Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. Machine Learning, 79:151–175, 2010.
- Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proc. ICML*, pages 200–209, 1999.
- Konstantinos Bousmalis, Nathan Silberman, David Dohan, Dumitru Erhan, and Dilip Krishnan. Unsupervised pixel-level domain adaptation with generative adversarial networks. In *Proc. CVPR*, pages 3722–3731, 2017.

- Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *Proc. ICML*, pages 1989–1998, 2018.
- Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *Proc. ICML*, pages 97–105, 2015.
- Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proc. ICML*, pages 1180–1189, 2015.
- Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In Proc. CVPR, pages 7167–7176, 2017.
- 29. Minghao Chen, Hongyang Xue, and Deng Cai. Domain adaptation for semantic segmentation with maximum squares loss. In *Proc. ICCV*, pages 2090–2099, 2019.
- Shuhao Cui, Shuhui Wang, Junbao Zhuo, Liang Li, Qingming Huang, and Qi Tian. Towards discriminability and diversity: Batch nuclear-norm maximization under label insufficient situations. In *Proc. CVPR*, pages 3941–3950, 2020.
- 31. Ying Jin, Ximei Wang, Mingsheng Long, and Jianmin Wang. Minimum class confusion for versatile domain adaptation. In *Proc. ECCV*, pages 464–480, 2020.
- Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. Neural Computation, 14(1):21–41, 2002.
- Zachary Lipton, Yu-Xiang Wang, and Alexander Smola.
 Detecting and correcting for label shift with black box predictors. In *Proc. ICML*, pages 3122–3130, 2018.
- Kamyar Azizzadenesheli, Anqi Liu, Fanny Yang, and Animashree Anandkumar. Regularized learning for domain adaptation under label shifts. In Proc. ICLR, 2019.
- Yawei Luo, Ping Liu, Tao Guan, Junqing Yu, and Yi Yang. Adversarial style mining for one-shot unsupervised domain adaptation. In *Proc. NeurIPS*, pages 20612–20623, 2020.
- 36. Thomas Varsavsky, Mauricio Orbes-Arteaga, Carole H Sudre, Mark S Graham, Parashkev Nachev, and M Jorge Cardoso. Test-time unsupervised domain adaptation. In Proc. MICCAI, pages 428–436, 2020.
- JH Moon, Debasmit Das, and CS George Lee. Multistep online unsupervised domain adaptation. In Proc. ICASSP, pages 41172–41576, 2020.
- 38. Luyu Yang, Mingfei Gao, Zeyuan Chen, Ran Xu, Abhinav Shrivastava, and Chetan Ramaiah. Burn after reading: Online adaptation for cross-domain streaming data. In *Proc. ECCV*, pages 404–422, 2022.
- Ilja Kuzborskij and Francesco Orabona. Stability and hypothesis transfer learning. In Proc. ICML, pages 942– 950, 2013.
- Jun Yang, Rong Yan, and Alexander G Hauptmann. Cross-domain video concept detection using adaptive svms. In *Proc. ACM-MM*, pages 188–197, 2007.
- 41. Tatiana Tommasi, Francesco Orabona, and Barbara Caputo. Learning categories from few examples with multi model knowledge transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(5):928–941, 2013.
- Sk Miraj Ahmed, Aske R Lejbolle, Rameswar Panda, and Amit K Roy-Chowdhury. Camera on-boarding for person re-identification using hypothesis transfer learning. In *Proc. CVPR*, pages 12144–12153, 2020.
- Shuang Ao, Xiang Li, and Charles Ling. Fast generalized distillation for semi-supervised domain adaptation. In Proc. AAAI, pages 1719–1725, 2017.

- 44. Arun Reddy Nelakurthi, Ross Maciejewski, and Jingrui He. Source free domain adaptation using an off-the-shelf classifier. In *Proc. IEEE BigData*, pages 140–145, 2018.
- Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Proc. NeurIPS, pages 3320–3328, 2014.
- Gido M van de Ven, Tinne Tuytelaars, and Andreas S Tolias. Three types of incremental learning. Nature Machine Intelligence, 4:1185–1197, 2022.
- Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. Sprompts learning with pre-trained transformers: An occam's razor for domain incremental learning. In *Proc.* NeurIPS, pages 5682–5695, 2022.
- Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy M Hospedales. Learning to generalize: meta-learning for domain generalization. In *Proc. AAAI*, pages 3490–3497, 2018.
- Fabio M Carlucci, Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Domain generalization by solving jigsaw puzzles. In *Proc. CVPR*, pages 2229–2238, 2019.
- Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In Proc. ICLR, 2020.
- Chelsea Finn, Pieter Abbeel, and Sergey Levine. Modelagnostic meta-learning for fast adaptation of deep networks. In *Proc. ICML*, pages 1126–1135, 2017.
- 52. Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, 2022.
- 53. Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. Meta-learning in neural networks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(9):5149–5169, 2021.
- 54. Longlong Jing and Yingli Tian. Self-supervised visual feature learning with deep neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4037–4058, 2020.
- Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Proc. ECCV*, pages 649–666, 2016
- Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proc. CVPR*, pages 2536–2544, 2016.
- Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *Proc. ICLR*, 2018.
- Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proc. ECCV*, pages 132–149, 2018.
- Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *Proc. NeurIPS*, pages 9912–9924, 2020.
- 60. Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. CVPR*, pages 9729–9738, 2020.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, pages 1597–1607, 2020.

62. Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proc. NAACL, pages 4171–4186, 2019.

- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for selfsupervised learning of speech representations. In *Proc.* NeurIPS, pages 12449–12460, 2020.
- Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. In *Proc. NeurIPS*, pages 5812–5823. 2020.
- 65. Jian Liang, Dapeng Hu, Yunbo Wang, Ran He, and Jiashi Feng. Source data-absent unsupervised domain adaptation through hypothesis transfer and labeling transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8602–8617, 2022.
- 66. Yanbei Chen, Massimiliano Mancini, Xiatian Zhu, and Zeynep Akata. Semi-supervised and unsupervised deep visual learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
- 67. Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Proc. NeurIPS*, pages 529–536, 2004.
- Dong-Hyun Lee. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Proc. ICML Workshops*, 2013.
- 69. Takeru Miyato, Shin-ichi Maeda, Masanori Koyama, and Shin Ishii. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8):1979–1993, 2018.
- 70. Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Proc. NeurIPS*, pages 596–608, 2020.
- Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In Proc. ICLR, 2017.
- Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. NeurIPS*, pages 1195–1204, 2017.
- Ahmet Iscen, Giorgos Tolias, Yannis Avrithis, and Ondrej Chum. Label propagation for deep semi-supervised learning. In *Proc. CVPR*, pages 5070–5079, 2019.
- Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In Proc. ICCV, pages 1214–1223, 2021.
- 75. Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1):1–48, 2019.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proc. CVPR, pages 770–778, 2016.
- Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. In *Proc.* UAI, pages 560–569, 2018.
- Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In Proc. ICLR, 2018.
- Juan C Pérez, Motasem Alfarra, Guillaume Jeanneret, Laura Rueda, Ali Thabet, Bernard Ghanem, and Pablo Arbeláez. Enhancing adversarial robustness via testtime transformation ensembling. In *Proc. ICCV*, pages 81–91, 2021.

80. Ildoo Kim, Younghoon Kim, and Sungwoong Kim. Learning loss for test-time augmentation. In *Proc. NeurIPS*, pages 4163–4174, 2020.

- Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proc. ICML*, pages 1050–1059, 2016.
- Boris Chidlovskii, Stephane Clinchant, and Gabriela Csurka. Domain adaptation in the absence of source domain data. In *Proc. KDD*, pages 451–460, 2016.
- 83. Stéphane Clinchant, Boris Chidlovskii, and Gabriela Csurka. Transductive adaptation of black box predictions. In *Proc. ACL*, pages 326–331, 2016.
- 84. Twan van Laarhoven and Elena Marchiori. Unsupervised domain adaptation with random walks on target labelings. arXiv preprint arXiv:1706.05335, 2017.
- Jian Liang, Ran He, Zhenan Sun, and Tieniu Tan. Distant supervised centroid shift: A simple and efficient approach to visual domain adaptation. In *Proc. CVPR*, pages 2975–2984, 2019.
- Sanqing Qu, Guang Chen, Jing Zhang, Zhijun Li, Wei He, and Dacheng Tao. Bmd: A general class-balanced multicentric dynamic prototype strategy for source-free domain adaptation. In *Proc. ECCV*, pages 165—-182, 2022.
- 87. Shiqi Yang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Exploiting the intrinsic neighborhood structure for source-free domain adaptation. In *Proc. NeurIPS*, pages 29393–29405, 2021.
- 88. Weijie Chen, Luojun Lin, Shicai Yang, Di Xie, Shiliang Pu, Yueting Zhuang, and Wenqi Ren. Self-supervised noisy label learning for source-free unsupervised domain adaptation. In *Proc. IROS*, pages 10185–10192, 2022.
- Fuming You, Jingjing Li, Lei Zhu, Zhi Chen, and Zi Huang. Domain adaptive semantic segmentation without source data. In *Proc. ACM-MM*, pages 3293– 3302, 2021.
- Yuxi Wang, Jian Liang, and Zhaoxiang Zhang. Source data-free cross-domain semantic segmentation: Align, teach and propagate. arXiv preprint arXiv:2106.11653, 2022.
- Hao Yan, Yuhong Guo, and Chunsheng Yang. Augmented self-labeling for source-free unsupervised domain adaptation. In Proc. NeurIPS Workshops, 2021.
- Tao Sun, Cheng Lu, and Haibin Ling. Prior knowledge guided unsupervised domain adaptation. In *Proc. ECCV*, pages 639–655, 2022.
- 93. Shiqi Yang, Yaxing Wang, Joost van de Weijer, Luis Herranz, and Shangling Jui. Generalized source-free domain adaptation. In *Proc. ICCV*, pages 8978–8987, 2021.
- 94. Tao Sun, Cheng Lu, and Haibin Ling. Domain adaptation with adversarial training on penultimate activations. In *Proc. AAAI*, 2023.
- Prabhu Teja Sivaprasad and François Fleuret. Uncertainty reduction for model adaptation in semantic segmentation. In Proc. CVPR, pages 9613–9623, 2021.
- 96. Qucheng Peng, Zhengming Ding, Lingjuan Lyu, Lichao Sun, and Chen Chen. Toward better target representation for source-free and black-box domain adaptation. arXiv preprint arXiv:2208.10531, 2022.
- Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In Proc. CVPR, pages 295–305, 2022.
- 98. Yuhe Ding, Jian Liang, Bo Jiang, Aihua Zheng, and Ran He. Maps: A noise-robust progressive learning approach for source-free domain adaptive keypoint de-

- tection. IEEE Transactions on Circuits and Systems for Video Technology, 34(3):1376–1387, 2024.
- 99. Kun Xia, Lingfei Deng, Wlodzislaw Duch, and Dongrui Wu. Privacy-preserving domain adaptation for motor imagery-based brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*, 69(11):3365–3376, 2022.
- 100. Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. Umad: Universal model adaptation under domain and category shift. arXiv preprint arXiv:2112.08553, 2021.
- 101. Xinhao Li, Jingjing Li, Lei Zhu, Guoqing Wang, and Zi Huang. Imbalanced source-free domain adaptation. In Proc. ACM-MM, pages 3330-3339, 2021.
- 102. Yuchen Liu, Yabo Chen, Wenrui Dai, Mengran Gou, Chun-Ting Huang, and Hongkai Xiong. Source-free domain adaptation with contrastive domain alignment and self-supervised exploration for face anti-spoofing. In Proc. ECCV, pages 511–528, 2022.
- 103. Gaurav Kumar Nayak, Konda Reddy Mopuri, Saksham Jain, and Anirban Chakraborty. 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, 2022.
- Yunzhong Hou and Liang Zheng. Source free domain adaptation with image translation. arXiv preprint arXiv:2008.07514, 2020.
- 105. Shishuai Hu, Zehui Liao, and Yong Xia. Prosfda: Prompt learning based source-free domain adaptation for medical image segmentation. arXiv preprint arXiv:2211.11514, 2022.
- 106. Ziyi Zhang, Weikai Chen, Hui Cheng, Zhen Li, Siyuan Li, Liang Lin, and Guanbin Li. Divide and contrast: Source-free domain adaptation via adaptive contrastive learning. In Proc. NeurIPS, pages 5137–5149, 2022.
- 107. Jiayi Tian, Jing Zhang, Wen Li, and Dong Xu. Vdm-da: Virtual domain modeling for source data-free domain adaptation. *IEEE Transactions on Circuits and Systems* for Video Technology, 32(6):3749–3760, 2022.
- 108. Zhen Qiu, Yifan Zhang, Hongbin Lin, Shuaicheng Niu, Yanxia Liu, Qing Du, and Mingkui Tan. Source-free domain adaptation via avatar prototype generation and adaptation. In Proc. IJCAI, pages 2921–2927, 2021.
- 109. Jogendra Nath Kundu, Suvaansh Bhambri, Akshay Kulkarni, Hiran Sarkar, Varun Jampani, and R. Venkatesh Babu. Concurrent subsidiary supervision for unsupervised source-free domain adaptation. In Proc. ECCV, pages 177–194, 2022.
- 110. Youngeun Kim, Donghyeon Cho, Kyeongtak Han, Priyadarshini Panda, and Sungeun Hong. Domain adaptation without source data. *IEEE Transactions on Artificial Intelligence*, 2(6):508–518, 2021.
- 111. Xianfeng Li, Weijie Chen, Di Xie, Shicai Yang, Peng Yuan, Shiliang Pu, and Yueting Zhuang. A free lunch for unsupervised domain adaptive object detection without source data. In *Proc. AAAI*, pages 8474–8481, 2021.
- 112. Cheng Chen, Quande Liu, Yueming Jin, Qi Dou, and Pheng-Ann Heng. Source-free domain adaptive fundus image segmentation with denoised pseudo-labeling. In *Proc. MICCAI*, pages 225–235, 2021.
- Yunzhong Hou and Liang Zheng. Visualizing adapted knowledge in domain transfer. In *Proc. CVPR*, pages 13824–13833, 2021.
- 114. Song Tang, Yuji Shi, Zhiyuan Ma, Jian Li, Jianzhi Lyu, Qingdu Li, and Jianwei Zhang. Model adaptation through hypothesis transfer with gradual knowledge distillation. In *Proc. IROS*, pages 5679–5685, 2021.

- 115. Chenxi Liu, Lixu Wang, Lingjuan Lyu, Chen Sun, Xiao Wang, and Qi Zhu. Twofer: Tackling continual domain shift with simultaneous domain generalization and adaptation. In Proc. ICLR, 2023.
- Jonghyun Lee, Dahuin Jung, Junho Yim, and Sungroh Yoon. Confidence score for source-free unsupervised domain adaptation. In *Proc. ICML*, pages 12365–12377, 2022.
- JoonHo Lee and Gyemin Lee. Feature alignment by uncertainty and self-training for source-free unsupervised domain adaptation. Neural Networks, 161:682

 –692, 2023.
- 118. Fan Wang, Zhongyi Han, Yongshun Gong, and Yilong Yin. Exploring domain-invariant parameters for source free domain adaptation. In *Proc. CVPR*, pages 7151– 7160, 2022.
- 119. Zhongping Cao, Zhenchang Li, Xuemei Guo, and Guoli Wang. Towards cross-environment human activity recognition based on radar without source data. IEEE Transactions on Vehicular Technology, 70(11):11843–11854, 2021.
- Yuhe Ding, Lijun Sheng, Jian Liang, Aihua Zheng, and Ran He. Proxymix: Proxy-based mixup training with label refinery for source-free domain adaptation. Neural Networks, 167:92–103, 2023.
- 121. Jiahua Dong, Zhen Fang, Anjin Liu, Gan Sun, and Tongliang Liu. Confident anchor-induced multi-source free domain adaptation. In *Proc. NeurIPS*, pages 2848– 2860, 2021.
- 122. Jian Liang, Dapeng Hu, and Jiashi Feng. Domain adaptation with auxiliary target domain-oriented classifier. In Proc. CVPR, pages 16632–16642, 2021.
- 123. Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for test-time adaptation. In Proc. CVPR, 2023.
- 124. Youngdong Kim, Junho Yim, Juseung Yun, and Junmo Kim. Nlnl: Negative learning for noisy labels. In Proc. ICCV, pages 101–110, 2019.
- 125. Xin Luo, Wei Chen, Yusong Tan, Chen Li, Yulin He, and Xiaogang Jia. Exploiting negative learning for implicit pseudo label rectification in source-free domain adaptive semantic segmentation. arXiv preprint arXiv:2106.12123, 2021
- 126. Waqar Ahmed, Pietro Morerio, and Vittorio Murino. Cleaning noisy labels by negative ensemble learning for source-free unsupervised domain adaptation. In *Proc.* WACV, pages 1616–1625, 2022.
- 127. Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proc. ECCV*, pages 289–305, 2018.
- 128. Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Model adaptation: Historical contrastive learning for unsupervised domain adaptation without source data. In *Proc. NeurIPS*, pages 3635–3649, 2021.
- Haojian Zhang, Yabin Zhang, Kui Jia, and Lei Zhang. Unsupervised domain adaptation of black-box source models. In *Proc. BMVC*, 2021.
- 130. Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, et al. Reclip: Refine contrastive language image pretraining with source free domain adaptation. In *Proc.* WACV, pages 2994–3003, 2024.
- 131. Nazmul Karim, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-pang Chiu, Supun Samarasekera, and Nazanin Rahnavard. C-sfda: A curriculum learning aided self-training framework for efficient source free do-

- main adaptation. In Proc. CVPR, 2023.
- 132. Li Yi, Gezheng Xu, Pengcheng Xu, Jiaqi Li, Ruizhi Pu, Charles Ling, A. Ian McLeod, and Boyu Wang. When source-free domain adaptation meets learning with noisy labels. In *Proc. ICLR*, 2023.
- 133. Jian Liang, Dapeng Hu, Jiashi Feng, and Ran He. Dine: Domain adaptation from single and multiple black-box predictors. In Proc. CVPR, pages 8003–8013, 2022.
- 134. Theodoros Panagiotakopoulos, Pier Luigi Dovesi, Linus Härenstam-Nielsen, and Matteo Poggi. Online domain adaptation for semantic segmentation in ever-changing conditions. In *Proc. ECCV*, pages 128–146, 2022.
- 135. Zongyao Li, Ren Togo, Takahiro Ogawa, and Miki Haseyama. Union-set multi-source model adaptation for semantic segmentation. In *Proc. ECCV*, pages 579–595, 2022.
- 136. Jogendra Nath Kundu, Akshay Kulkarni, Amit Singh, Varun Jampani, and R Venkatesh Babu. Generalize then adapt: Source-free domain adaptive semantic segmentation. In *Proc. ICCV*, pages 7046–7056, 2021.
- 137. Mucong Ye, Jing Zhang, Jinpeng Ouyang, and Ding Yuan. Source data-free unsupervised domain adaptation for semantic segmentation. In *Proc. ACM-MM*, pages 2233–2242, 2021.
- 138. Lin Xiong, Mao Ye, Dan Zhang, Yan Gan, and Yiguang Liu. Source data-free domain adaptation for a faster r-cnn. Pattern Recognition, 124:108436, 2022.
- 139. Vikash Kumar, Rohit Lal, Himanshu Patil, and Anirban Chakraborty. Conmix for source-free single and multitarget domain adaptation. In *Proc. WACV*, pages 4178– 4188, 2023.
- 140. Bin Deng, Yabin Zhang, Hui Tang, Changxing Ding, and Kui Jia. On universal black-box domain adaptation. arXiv preprint arXiv:2104.04665, 2021.
- 141. Qing Tian, Shun Peng, and Tinghuai Ma. Source-free unsupervised domain adaptation with trusted pseudo samples. ACM Transactions on Intelligent Systems and Technology, 14(2):1–17, 2023.
- 142. Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? In *Proc. NeurIPS*, pages 4694–4703, 2019.
- 143. Viraj Prabhu, Shivam Khare, Deeksha Kartik, and Judy Hoffman. Augco: Augmentation consistency-guided selftraining for source-free domain adaptive semantic segmentation. arXiv preprint arXiv:2107.10140, 2022.
- 144. Evgenia Rusak, Steffen Schneider, George Pachitariu, Luisa Eck, Peter Vincent Gehler, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. If your data distribution shifts, use self-learning. Transactions on Machine Learning Research, 2022.
- Kowshik Thopalli, Pavan Turaga, and Jayaraman J Thiagarajan. Domain alignment meets fully test-time adaptation. In *Proc. ACML*, pages 1006–1021, 2023.
- 146. Xiangli Yang, Zixing Song, Irwin King, and Zenglin Xu. A survey on deep semi-supervised learning. *IEEE Transactions on Knowledge and Data Engineering*, 2022.
- 147. Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In Proc. CVPR Workshops, 2020.
- 148. Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. In Proc. NeurIPS, pages 6256–6268, 2020.
- 149. Xiaodong Wang, Junbao Zhuo, Shuhao Cui, Shuhui Wang, and Yuejian Fang. Learning invariant representation with consistency and diversity for semi-supervised

- source hypothesis transfer. In $Proc.\ ICASSP,\ pages\ 5125-5129,\ 2024.$
- 150. Chen Yang, Xiaoqing Guo, Zhen Chen, and Yixuan Yuan. Source free domain adaptation for medical image segmentation with fourier style mining. *Medical Image Analysis*, 79:102457, 2022.
- 151. Samarth Sinha, Peter Gehler, Francesco Locatello, and Bernt Schiele. Test: Test-time self-training under distribution shift. In Proc. WACV, pages 2759–2769, 2023.
- Zeyu Feng, Chang Xu, and Dacheng Tao. Open-set hypothesis transfer with semantic consistency. *IEEE Transactions on Image Processing*, 30:6473

 –6484, 2021.
- 153. Weikai Li, Meng Cao, and Songcan Chen. Jacobian norm for unsupervised source-free domain adaptation. arXiv preprint arXiv:2204.03467, 2022.
- 154. Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *Proc. ICLR*, 2018.
- 155. Divya Kothandaraman, Sumit Shekhar, Abhilasha Sancheti, Manoj Ghuhan, Tripti Shukla, and Dinesh Manocha. Salad: source-free active label-agnostic domain adaptation for classification, segmentation and detection. In *Proc. WACV*, pages 382–391, 2023.
- Alessandro Conti, Paolo Rota, Yiming Wang, and Elisa Ricci. Cluster-level pseudo-labelling for source-free cross-domain facial expression recognition. In Proc. BMVC, 2022.
- 157. Qicheng Lao, Xiang Jiang, and Mohammad Havaei. Hypothesis disparity regularized mutual information maximization. In *Proc. AAAI*, pages 8243–8251, 2021.
- 158. Dan Zhang, Mao Ye, Lin Xiong, Shuaifeng Li, and Xue Li. Source-style transferred mean teacher for sourcedata free object detection. In ACM Multimedia Asia, pages 1–8, 2021.
- 159. Vibashan VS, Jeya Maria Jose Valanarasu, and Vishal M Patel. Target and task specific source-free domain adaptive image segmentation. arXiv preprint arXiv:2203.15792, 2022.
- 160. Shuaifeng Li, Mao Ye, Xiatian Zhu, Lihua Zhou, and Lin Xiong. Source-free object detection by learning to overlook domain style. In *Proc. CVPR*, pages 8014–8023, 2022.
- Dequan Wang, Shaoteng Liu, Sayna Ebrahimi, Evan Shelhamer, and Trevor Darrell. On-target adaptation. arXiv preprint arXiv:2109.01087, 2021.
- 162. Masato Ishii and Masashi Sugiyama. Source-free domain adaptation via distributional alignment by matching batch normalization statistics. arXiv preprint arXiv:2101.10842, 2021.
- 163. Sk Miraj Ahmed, Suhas Lohit, Kuan-Chuan Peng, Michael Jones, and Amit K Roy-Chowdhury. Crossmodal knowledge transfer without task-relevant source data. In Proc. ECCV, pages 111–127, 2022.
- 164. Xiaofeng Liu, Fangxu Xing, Chao Yang, Georges El Fakhri, and Jonghye Woo. Adapting off-the-shelf source segmenter for target medical image segmentation. In *Proc. MICCAI*, pages 549–559, 2021.
- 165. Yuejiang Liu, Parth Kothari, Bastien van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. Ttt++: When does self-supervised test-time training fail or thrive? In Proc. NeurIPS, pages 21808–21820, 2021.
- 166. Shiqi Yang, Yaxing Wang, Luis Herranz, Shangling Jui, and Joost van de Weijer. Casting a bait for offline and online source-free domain adaptation. Computer Vision and Image Understanding, 234:103747, 2023.

- 167. Haifeng Xia, Handong Zhao, and Zhengming Ding. Adaptive adversarial network for source-free domain adaptation. In Proc. ICCV, pages 9010–9019, 2021.
- 168. Mathilde Bateson, Hoel Kervadec, Jose Dolz, Hervé Lombaert, and Ismail Ben Ayed. Source-free domain adaptation for image segmentation. Medical Image Analysis, 82:102617, 2022.
- 169. Yuang Liu, Wei Zhang, and Jun Wang. Source-free domain adaptation for semantic segmentation. In Proc. CVPR, pages 1215–1224, 2021.
- 170. Yalan Ye, Ziqi Liu, Yangwuyong Zhang, Jingjing Li, and Hengtao Shen. Alleviating style sensitivity then adapting: Source-free domain adaptation for medical image segmentation. In *Proc. ACM-MM*, pages 1935–1944, 2022.
- 171. Yuan Shi and Fei Sha. Information-theoretical learning of discriminative clusters for unsupervised domain adaptation. In Proc. ICML, pages 1275–1282, 2012.
- 172. Subhankar Roy, Martin Trapp, Andrea Pilzer, Juho Kannala, Nicu Sebe, Elisa Ricci, and Arno Solin. Uncertainty-guided source-free domain adaptation. In Proc. ECCV, pages 537–555, 2022.
- 173. Mengmeng Jing, Xiantong Zhen, Jingjing Li, and Cees G. M. Snoek. Variational model perturbation for sourcefree domain adaptation. In *Proc. NeurIPS*, pages 17173– 17187, 2022.
- 174. Andreas Krause, Pietro Perona, and Ryan Gomes. Discriminative clustering by regularized information maximization. In *Proc. NeurIPS*, pages 775–783, 2010.
- 175. Weihua Hu, Takeru Miyato, Seiya Tokui, Eiichi Matsumoto, and Masashi Sugiyama. Learning discrete representations via information maximizing self-augmented training. In *Proc. ICML*, pages 1558–1567, 2017.
- 176. Pietro Morerio, Riccardo Volpi, Ruggero Ragonesi, and Vittorio Murino. Generative pseudo-label refinement for unsupervised domain adaptation. In *Proc. WACV*, pages 3130–3139, 2020.
- 177. Vinod K Kurmi, Venkatesh K Subramanian, and Vinay P Namboodiri. Domain impression: A source data free domain adaptation method. In *Proc. WACV*, pages 615–625, 2021.
- 178. Hao Yan, Yuhong Guo, and Chunsheng Yang. Sourcefree unsupervised domain adaptation with surrogate data generation. In Proc. BMVC, 2021.
- 179. Qianyu Zhou, Ke-Yue Zhang, Taiping Yao, Ran Yi, Kekai Sheng, Shouhong Ding, and Lizhuang Ma. Generative domain adaptation for face anti-spoofing. In *Proc.* ECCV, pages 335–356, 2022.
- Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.
- 181. Baoyao Yang, Andy J. Ma, and Pong C. Yuen. Revealing task-relevant model memorization for source-protected unsupervised domain adaptation. *IEEE Transactions on Information Forensics and Security*, 17:716–731, 2022.
- 182. Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proc. CVPR*, pages 8715–8724, 2020.
- 183. Yanghao Li, Naiyan Wang, Jiaying Liu, and Xiaodi Hou. Demystifying neural style transfer. In *Proc. IJCAI*, pages 2230–2236, 2017.
- 184. Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural net-

- works. In *Proc. CVPR*, pages 2414–2423, 2016.
- Roshni Sahoo, Divya Shanmugam, and John Guttag. Unsupervised domain adaptation in the absence of source data. In Proc. ICML Workshops, 2020.
- 186. Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Visual prompting: Modifying pixel space to adapt pre-trained models. arXiv preprint arXiv:2203.17274, 2022.
- Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proc. CVPR*, pages 4085–4095, 2020.
- 188. Jianfei Yang, Xiangyu Peng, Kai Wang, Zheng Zhu, Jiashi Feng, Lihua Xie, and Yang You. Divide to adapt: Mitigating confirmation bias for domain adaptation of black-box predictors. In Proc. ICLR, 2023.
- Chang Liu, Lihua Zhou, Mao Ye, and Xue Li. Selfalignment for black-box domain adaptation of image classification. *IEEE Signal Processing Letters*, 29:1709– 1713, 2022.
- 190. Yi Huang, Xiaoshan Yang, Ji Zhang, and Changsheng Xu. Relative alignment network for source-free multimodal video domain adaptation. In *Proc. ACM-MM*, pages 1652–1660, 2022.
- 191. Tong Chu, Yahao Liu, Jinhong Deng, Wen Li, and Lixin Duan. Denoised maximum classifier discrepancy for source free unsupervised domain adaptation. In Proc. AAAI, pages 472–480, 2022.
- 192. Ning Ding, Yixing Xu, Yehui Tang, Chao Xu, Yunhe Wang, and Dacheng Tao. Source-free domain adaptation via distribution estimation. In *Proc. CVPR*, pages 7212–7222, 2022.
- 193. Serban Stan and Mohammad Rostami. Unsupervised model adaptation for continual semantic segmentation. In Proc. AAAI, pages 2593–2601, 2021.
- 194. Mohammad Rostami. Lifelong domain adaptation via consolidated internal distribution. In *Proc. NeurIPS*, pages 11172–11183, 2021.
- 195. Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. In *Proc. ICLR*, 2018.
- 196. David Berthelot, Nicholas Carlini, Ian Goodfellow, Avital Oliver, Nicolas Papernot, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Proc. NeurIPS*, pages 5049–5059, 2019.
- 197. Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proc. CVPR*, pages 3723–3732, 2018.
- 198. Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(1):723–773, 2012.
- 199. Yue Tan, Chen Chen, Weiming Zhuang, Xin Dong, Lingjuan Lyu, and Guodong Long. Is heterogeneity notorious? taming heterogeneity to handle test-time shift in federated learning. In *Proc. NeurIPS*, 2023.
- 200. Korawat Tanwisuth, Xinjie Fan, Huangjie Zheng, Shujian Zhang, Hao Zhang, Bo Chen, and Mingyuan Zhou. A prototype-oriented framework for unsupervised domain adaptation. In *Proc. NeurIPS*, pages 17194–17208, 2021.
- 201. Korawat Tanwisuth, Shujian Zhang, Huangjie Zheng, Pengcheng He, and Mingyuan Zhou. Pouf: Promptoriented unsupervised fine-tuning for large pre-trained models. In *Proc. ICML*, pages 33816–33832, 2023.

202. Baoyao Yang, Hao-Wei Yeh, Tatsuya Harada, and Pong C Yuen. 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, 2021.

- Jingjing Wang, Jingyi Zhang, Ying Bian, Youyi Cai, Chunmao Wang, and Shiliang Pu. Self-domain adaptation for face anti-spoofing. In *Proc. AAAI*, pages 2746– 2754, 2021.
- 204. Ondrej Bohdal, Da Li, Shell Xu Hu, and Timothy Hospedales. Feed-forward source-free latent domain adaptation via cross-attention. In Proc. ICML Workshops, 2022.
- Aaditya Naik, Yinjun Wu, Mayur Naik, and Eric Wong. Do machine learning models learn common sense? arXiv preprint arXiv:2303.01433, 2023.
- 206. Jian Liang, Yunbo Wang, Dapeng Hu, Ran He, and Jiashi Feng. A balanced and uncertainty-aware approach for partial domain adaptation. In *Proc. ECCV*, pages 123–140, 2020.
- Jogendra Nath Kundu, Naveen Venkat, Ambareesh Revanur, and R Venkatesh Babu. Towards inheritable models for open-set domain adaptation. In *Proc. CVPR*, pages 12376–12385, 2020.
- 208. Kaichao You, Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Universal domain adaptation. In Proc. CVPR, pages 2720–2729, 2019.
- 209. Shiqi Yang, Yaxing Wang, Kai Wang, Shangling Jui, and Joost van de Weijer. One ring to bring them all: Model adaptation under domain and category shift. arXiv preprint arXiv:2206.03600, 2022.
- Sanqing Qu, Tianpei Zou, Florian Roehrbein, Cewu Lu, Guang Chen, Dacheng Tao, and Changjun Jiang. Upcycling models under domain and category shift. In Proc. CVPR, 2023.
- 211. Jogendra Nath Kundu, Akshay Kulkarni, Suvaansh Bhambri, Deepesh Mehta, Shreyas Kulkarni, Varun Jampani, and R Venkatesh Babu. Balancing discriminability and transferability for source-free domain adaptation. In Proc. ICML, pages 11710–11728, 2022.
- 212. Sk Miraj Ahmed, Dripta S Raychaudhuri, Sujoy Paul, Samet Oymak, and Amit K Roy-Chowdhury. Unsupervised multi-source domain adaptation without access to source data. In *Proc. CVPR*, pages 10103–10112, 2021.
- 213. Ning Ma, Jiajun Bu, Lixian Lu, Jun Wen, Sheng Zhou, Zhen Zhang, Jingjun Gu, Haifeng Li, and Xifeng Yan. Context-guided entropy minimization for semi-supervised domain adaptation. Neural Networks, 154:270-282, 2022.
- 214. Amr Alexandari, Anshul Kundaje, and Avanti Shrikumar. Maximum likelihood with bias-corrected calibration is hard-to-beat at label shift adaptation. In *Proc. ICML*, pages 222–232, 2020.
- 215. Tomáš Šipka, Milan Šulc, and Jiří Matas. The hitch-hiker's guide to prior-shift adaptation. In *Proc. WACV*, pages 1516–1524, 2022.
- Fan Wang, Zhongyi Han, Zhiyan Zhang, and Yilong Yin. Active source free domain adaptation. arXiv preprint arXiv:2205.10711, 2022.
- 217. Xinyao Li, Zhekai Du, Jingjing Li, Lei Zhu, and Ke Lu. Source-free active domain adaptation via energy-based locality preserving transfer. In *Proc. ACM-MM*, pages 5802–5810, 2022.
- Peshal Agarwal, Danda Pani Paudel, Jan-Nico Zaech, and Luc Van Gool. Unsupervised robust domain adap-

- tation without source data. In Proc.~WACV, pages 2009–2018, 2022.
- 219. Qiyuan An, Ruijiang Li, Lin Gu, Hao Zhang, Qingyu Chen, Zhiyong Lu, Fei Wang, and Yingying Zhu. A privacy-preserving unsupervised domain adaptation framework for clinical text analysis. arXiv preprint arXiv:2201.07317, 2022.
- 220. Lijun Sheng, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. Adaptguard: Defending against universal attacks for model adaptation. In *Proc. ICCV*, pages 19093–19103, 2023.
- 221. Antonio D'Innocente, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Learning to generalize one sample at a time with self-supervision. arXiv preprint arXiv:1910.03915, 2019.
- 222. Steven Wegmann, Francesco Scattone, Ira Carp, Larry Gillick, Robert Roth, and Jon Yamron. Dragon systems' 1997 broadcast news transcription system. In Proc. DARPA Broadcast News Transcription and Understanding Workshop, 1998.
- Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization. In *Proc. CVPR*, pages 2616– 2625, 2018.
- 224. Fatemeh Azimi, Sebastian Palacio, Federico Raue, Jörn Hees, Luca Bertinetto, and Andreas Dengel. Self-supervised test-time adaptation on video data. In Proc. WACV, pages 3439–3448, 2022.
- 225. Zachary Nado, Shreyas Padhy, D Sculley, Alexander D'Amour, Balaji Lakshminarayanan, and Jasper Snoek. Evaluating prediction-time batch normalization for robustness under covariate shift. In Proc. ICML Workshops, 2020.
- 226. Yuliang Zou, Zizhao Zhang, Chun-Liang Li, Han Zhang, Tomas Pfister, and Jia-Bin Huang. Learning instancespecific adaptation for cross-domain segmentation. In Proc. ECCV, pages 459–476, 2022.
- Seobin Park, Jinsu Yoo, Donghyeon Cho, Jiwon Kim, and Tae Hyun Kim. Fast adaptation to super-resolution networks via meta-learning. In *Proc. ECCV*, pages 754– 769, 2020.
- 228. Francesco Cappio Borlino, Salvatore Polizzotto, Barbara Caputo, and Tatiana Tommasi. Self-supervision & meta-learning for one-shot unsupervised cross-domain detection. Computer Vision and Image Understanding, 223:103549, 2022.
- 229. Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Proc. NeurIPS*, pages 14274– 14289, 2022.
- 230. Neerav Karani, Ertunc Erdil, Krishna Chaitanya, and Ender Konukoglu. Test-time adaptable neural networks for robust medical image segmentation. *Medical Image Analysis*, 68:101907, 2021.
- 231. Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proc. CVPR*, pages 8344–8353, 2022.
- Xinyi Wang, Yulia Tsvetkov, Sebastian Ruder, and Graham Neubig. Efficient test time adapter ensembling for low-resource language varieties. In EMNLP Findings, pages 730—737, 2021.
- Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. ICML*, pages 448–456, 2015.

- Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. In Proc. NeurIPS Workshops, 2016.
- 235. Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. In Proc. ICLR, 2017.
- 236. Hyesu Lim, Byeonggeun Kim, Jaegul Choo, and Sungha Choi. Ttn: A domain-shift aware batch normalization in test-time adaptation. In Proc. ICLR, 2023.
- 237. Yifei Zhou, Juntao Ren, Fengyu Li, Ramin Zabih, and Ser Nam Lim. Test-time distribution normalization for contrastively learned visual-language models. In *Proc.* NeurIPS, 2023.
- 238. Sherwin Bahmani, Oliver Hahn, Eduard Zamfir, Nikita Araslanov, Daniel Cremers, and Stefan Roth. Semantic self-adaptation: Enhancing generalization with a single sample. In Proc. ECCV Workshops, 2022.
- Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022, 2016.
- 240. Ansh Khurana, Sujoy Paul, Piyush Rai, Soma Biswas, and Gaurav Aggarwal. Sita: Single image test-time adaptation. arXiv preprint arXiv:2112.02355, 2021.
- Antonio D'Innocente, Francesco Cappio Borlino, Silvia Bucci, Barbara Caputo, and Tatiana Tommasi. Oneshot unsupervised cross-domain detection. In Proc. ECCV, pages 732–748, 2020.
- 242. Dequan Wang, Evan Shelhamer, Bruno Olshausen, and Trevor Darrell. Dynamic scale inference by entropy minimization. arXiv preprint arXiv:1908.03182, 2019.
- 243. Nikhil Reddy, Abhinav Singhal, Abhishek Kumar, Mahsa Baktashmotlagh, and Chetan Arora. Master of all: Simultaneous generalization of urban-scene segmentation to all adverse weather conditions. In *Proc. ECCV*, pages 51–69, 2022.
- 244. Yu Sun, Eric Tzeng, Trevor Darrell, and Alexei A Efros. Unsupervised domain adaptation through selfsupervision. arXiv preprint arXiv:1909.11825, 2019.
- 245. Jianfeng Zhang, Xuecheng Nie, and Jiashi Feng. Inference stage optimization for cross-scenario 3d human pose estimation. In *Proc. NeurIPS*, pages 2408–2419, 2020.
- 246. Nicklas Hansen, Rishabh Jangir, Yu Sun, Guillem Alenyà, Pieter Abbeel, Alexei A Efros, Lerrel Pinto, and Xiaolong Wang. Self-supervised policy adaptation during deployment. In *Proc. ICLR*, 2021.
- 247. Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei A Efros. Test-time training with masked autoencoders. In Proc. NeurIPS, pages 29374–29385, 2022.
- 248. Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. CVPR*, pages 16000–16009, 2022.
- 249. Fei Lyu, Mang Ye, Andy J Ma, Terry Cheuk-Fung Yip, Grace Lai-Hung Wong, and Pong C Yuen. Learning from synthetic ct images via test-time training for liver tumor segmentation. *IEEE Transactions on Medical Imaging*, 41(9):2510–2520, 2022.
- 250. Anindya Sarkar, Anirban Sarkar, and Vineeth N Balasubramanian. Leveraging test-time consensus prediction for robustness against unseen noise. In *Proc. WACV*, pages 1839–1848, 2022.
- 251. Jameel Hassan Abdul Samadh, Hanan Gani, Noor Hazim Hussein, Muhammad Uzair Khattak, Muzammal Naseer, Fahad Khan, and Salman Khan. Align your prompts: Test-time prompting with distri-

- bution alignment for zero-shot generalization. In *Proc. NeurIPS*, 2023.
- 252. Mathilde Bateson, Hervé Lombaert, and Ismail Ben Ayed. Test-time adaptation with shape moments for image segmentation. In *Proc. MICCAI*, pages 736–745, 2022.
- 253. Guan-Ting Lin, Shang-Wen Li, and Hung-yi Lee. Listen, adapt, better wer: Source-free single-utterance test-time adaptation for automatic speech recognition. In *Proc. Interspeech*, pages 2198–2202, 2022.
- 254. Shuai Zhao, Xiaohan Wang, Linchao Zhu, and Yi Yang. Test-time adaptation with clip reward for zero-shot generalization in vision-language models. In *Proc. ICLR*, 2024.
- 255. Quande Liu, Cheng Chen, Qi Dou, and Pheng-Ann Heng. Single-domain generalization in medical image segmentation via test-time adaptation from shape dictionary. In *Proc. AAAI*, pages 1756–1764, 2022.
- 256. Wei Jin, Tong Zhao, Jiayuan Ding, Yozen Liu, Jiliang Tang, and Neil Shah. Empowering graph representation learning with test-time graph transformation. In *Proc. ICLR*, 2023.
- 257. Zhehan Kan, Shuoshuo Chen, Zeng Li, and Zhihai He. Self-constrained inference optimization on structural groups for human pose estimation. In *Proc. ECCV*, pages 729–745, 2022.
- 258. Xueting Li, Sifei Liu, Shalini De Mello, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Online adaptation for consistent mesh reconstruction in the wild. In *Proc. NeurIPS*, pages 15009–15019, 2020.
- 259. Sunghwan Hong and Seungryong Kim. Deep matching prior: Test-time optimization for dense correspondence. In Proc. ICCV, pages 9907–9917, 2021.
- Sreyas Mohan, Joshua L Vincent, Ramon Manzorro, Peter Crozier, Carlos Fernandez-Granda, and Eero Simoncelli. Adaptive denoising via gaintuning. In *Proc.* NeurIPS, pages 23727–23740, 2021.
- David Bau, Hendrik Strobelt, William Peebles, Jonas Wulff, Bolei Zhou, Jun-Yan Zhu, and Antonio Torralba. Semantic photo manipulation with a generative image prior. ACM Transactions on Graphics, 38(4):1–11, 2019.
- 262. Sunwoo Kim, Youngjo Min, Younghun Jung, and Seungryong Kim. Controllable style transfer via test-time training of implicit neural representation. *Pattern Recognition*, 146:109988, 2024.
- 263. Pratyay Banerjee, Tejas Gokhale, and Chitta Baral. Self-supervised test-time learning for reading comprehension. In *Proc. NAACL*, pages 1200–1211, 2021.
- Assaf Shocher, Nadav Cohen, and Michal Irani. "zeroshot" super-resolution using deep internal learning. In Proc. CVPR, pages 3118–3126, 2018.
- 265. Abhimanyu Dubey, Vignesh Ramanathan, Alex Pentland, and Dhruv Mahajan. Adaptive methods for real-world domain generalization. In *Proc. CVPR*, pages 14340–14349, 2021.
- 266. Jangho Kim, Jun-Tae Lee, Simyung Chang, and Nojun Kwak. Variational on-the-fly personalization. In *Proc. ICML*, pages 11134–11147, 2022.
- Zhixiang Chi, Yang Wang, Yuanhao Yu, and Jin Tang. Test-time fast adaptation for dynamic scene deblurring via meta-auxiliary learning. In *Proc. CVPR*, pages 9137– 9146, 2021.
- 268. Huan Liu, Zijun Wu, Liangyan Li, Sadaf Salehkalaibar, Jun Chen, and Keyan Wang. Towards multi-domain single image dehazing via test-time training. In Proc. CVPR, pages 5831–5840, 2022.

 Chaerin Min, Taehyun Kim, and Jongwoo Lim. Metalearning for adaptation of deep optical flow networks. In Proc. WACV, pages 2145–2154, 2023.

- 270. Myungsub Choi, Janghoon Choi, Sungyong Baik, Tae Hyun Kim, and Kyoung Mu Lee. Test-time adaptation for video frame interpolation via meta-learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(12):9615–9628, 2021.
- 271. Ferran Alet, Maria Bauza, Kenji Kawaguchi, Nurullah Giray Kuru, Tomas Lozano-Perez, and Leslie Pack Kaelbling. Tailoring: Encoding inductive biases by optimizing unsupervised objectives at prediction time. In Proc. NeurIPS, pages 29206–29217, 2021.
- 272. Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch3t: Test-time training for zero-shot sbir. In Proc. CVPR, pages 7462–7471, 2022.
- 273. Huan Liu, Zhixiang Chi, Yuanhao Yu, Yang Wang, Jun Chen, and Jin Tang. Meta-auxiliary learning for future depth prediction in videos. In *Proc. WACV*, pages 5756–5765, 2023.
- Wenxuan Bao, Tianxin Wei, Haohan Wang, and Jingrui He. Adaptive test-time personalization for federated learning. In *Proc. NeurIPS*, 2023.
- 275. Zhishu Sun, Zhifeng Shen, Luojun Lin, Yuanlong Yu, Zhifeng Yang, Shicai Yang, and Weijie Chen. Dynamic domain generalization. In *Proc. IJCAI*, pages 1342–1348, 2022.
- 276. Eyal Ben-David, Nadav Oved, and Roi Reichart. Pada: Example-based prompt learning for on-the-fly adaptation to unseen domains. Transactions of the Association for Computational Linguistics, 10:414–433, 2022.
- 277. Seonguk Seo, Yumin Suh, Dongwan Kim, Geeho Kim, Jongwoo Han, and Bohyung Han. Learning to optimize domain specific normalization for domain generalization. In *Proc. ECCV*, pages 68–83, 2020.
- 278. Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proc. ICCV*, pages 1446–1455, 2019.
- 279. Zehao Xiao, Xiantong Zhen, Ling Shao, and Cees GM Snoek. Learning to generalize across domains on single test samples. In Proc. ICLR, 2022.
- 280. Boqiang Xu, Jian Liang, Lingxiao He, and Zhenan Sun. Mimic embedding via adaptive aggregation: Learning generalizable person re-identification. In *Proc. ECCV*, pages 372–388, 2022.
- 281. Mattia Segu, Alessio Tonioni, and Federico Tombari. Batch normalization embeddings for deep domain generalization. *Pattern Recognition*, 135:109115, 2023.
- 282. Xingchen Zhao, Chang Liu, Anthony Sicilia, Seong Jae Hwang, and Yun Fu. Test-time fourier style calibration for domain generalization. In *Proc. IJCAI*, pages 1721– 1727, 2022.
- 283. Jin Gao, Jialing Zhang, Xihui Liu, Trevor Darrell, Evan Shelhamer, and Dequan Wang. Back to the source: Diffusion-driven adaptation to test-time corruption. In Proc. CVPR, 2023.
- 284. Yun-Yun Tsai, Chengzhi Mao, Yow-Kuan Lin, and Junfeng Yang. Self-supervised convolutional visual prompts. arXiv preprint arXiv:2303.00198, 2023.
- 285. Yufan He, Aaron Carass, Lianrui Zuo, Blake E Dewey, and Jerry L Prince. Autoencoder based self-supervised test-time adaptation for medical image analysis. Medical Image Analysis, page 102136, 2021.

- 286. Hao Li, Han Liu, Dewei Hu, Jiacheng Wang, Hans Johnson, Omar Sherbini, Francesco Gavazzi, Russell D'Aiello, Adeline Vanderver, Jeffrey Long, Paulsen Jane, and Ipek Oguz. Self-supervised test-time adaptation for medical image segmentation. In Proc. MICCAI Workshops, 2022.
- 287. Gabriele Valvano, Andrea Leo, and Sotirios A Tsaftaris. Re-using adversarial mask discriminators for test-time training under distribution shifts. *Journal of Machine Learning for Biomedical Imaging*, 1:1–27, 2022.
- 288. Jan-Aike Termöhlen, Marvin Klingner, Leon J Brettin, Nico M Schmidt, and Tim Fingscheidt. Continual unsupervised domain adaptation for semantic segmentation by online frequency domain style transfer. In *Proc.* ITSC, pages 2881–2888, 2021.
- 289. Prashant Pandey, Mrigank Raman, Sumanth Varambally, and Prathosh AP. Generalization on unseen domains via inference-time label-preserving target projections. In *Proc. CVPR*, pages 12924–12933, 2021.
- 290. Zehao Xiao, Xiantong Zhen, Shengcai Liao, and Cees GM Snoek. Energy-based test sample adaptation for domain generalization. In Proc. ICLR, 2023.
- Changhao Shi, Chester Holtz, and Gal Mishne. Online adversarial purification based on self-supervision. In Proc. ICLR, 2021.
- 292. Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In Proc. ICML, pages 12062–12072, 2021.
- 293. Chengzhi Mao, Mia Chiquier, Hao Wang, Junfeng Yang, and Carl Vondrick. Adversarial attacks are reversible with natural supervision. In *Proc. ICCV*, pages 661– 671, 2021.
- 294. Motasem Alfarra, Juan C Pérez, Ali Thabet, Adel Bibi, Philip HS Torr, and Bernard Ghanem. Combating adversaries with anti-adversaries. In *Proc. AAAI*, pages 5992–6000, 2022.
- 295. Yi-Fan Zhang, Jindong Wang, Jian Liang, Zhang Zhang, Baosheng Yu, Liang Wang, Dacheng Tao, and Xing Xie. Domain-specific risk minimization for out-ofdistribution generalization. In *Proc. KDD*, pages 3409— 3421, 2023.
- 296. Vidit Jain and Erik Learned-Miller. Online domain adaptation of a pre-trained cascade of classifiers. In Proc. CVPR, pages 577–584, 2011.
- 297. Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *Proc. ICCV*, pages 7063–7072, 2019.
- 298. Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *Proc. ICML*, pages 16888–16905, 2022.
- 299. M Jehanzeb Mirza, Jakub Micorek, Horst Possegger, and Horst Bischof. The norm must go on: Dynamic unsupervised domain adaptation by normalization. In Proc. CVPR, pages 14765–14775, 2022.
- 300. Bowen Zhao, Chen Chen, and Shu-Tao Xia. Delta: Degradation-free fully test-time adaptation. In $Proc.\ ICLR$, 2023.
- 301. Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Zhiquan Wen, Yaofo Chen, Peilin Zhao, and Mingkui Tan. Towards stable test-time adaptation in dynamic wild world. In Proc. ICLR, 2023.
- 302. Minguk Jang, Sae-Young Chung, and Hye Won Chung. Test-time adaptation via self-training with nearest neighbor information. In Proc. ICLR, 2023.

- 303. Takeshi Kojima, Yutaka Matsuo, and Yusuke Iwasawa. Robustifying vision transformer without retraining from scratch by test-time class-conditional feature alignment. In *Proc. IJCAI*, pages 1009–1016, 2022.
- 304. Dhanajit Brahma and Piyush Rai. A probabilistic framework for lifelong test-time adaptation. In $Proc.\ CVPR,\ 2023.$
- 305. Minhao Hu, Tao Song, Yujun Gu, Xiangde Luo, Jieneng Chen, Yinan Chen, Ya Zhang, and Shaoting Zhang. Fully test-time adaptation for image segmentation. In *Proc. MICCAI*, pages 251–260, 2021.
- 306. Junha Song, Kwanyong Park, Inkyu Shin, Sanghyun Woo, and In So Kweon. Cd-tta: Compound domain test-time adaptation for semantic segmentation. arXiv preprint arXiv:2212.08356, 2022.
- 307. Fuming You, Jingjing Li, and Zhou Zhao. Test-time batch statistics calibration for covariate shift. arXiv preprint arXiv:2110.04065, 2021.
- 308. Massimiliano Mancini, Hakan Karaoguz, Elisa Ricci, Patric Jensfelt, and Barbara Caputo. Kitting in the wild through online domain adaptation. In *Proc. IROS*, pages 1103–1109, 2018.
- Junyuan Hong, Lingjuan Lyu, Jiayu Zhou, and Michael Spranger. Mecta: Memory-economic continual test-time model adaptation. In Proc. ICLR, 2023.
- 310. Shuoyuan Wang, Jindong Wang, Huajun Xi, Bob Zhang, Lei Zhang, and Hongxin Wei. 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, 2024.
- 311. Tao Yang, Shenglong Zhou, Yuwang Wang, Yan Lu, and Nanning Zheng. Test-time batch normalization. arXiv preprint arXiv:2205.10210, 2022.
- Sergey Ioffe. Batch renormalization: Towards reducing minibatch dependence in batch-normalized models. In Proc. NeurIPS, pages 1942–1950, 2017.
- 313. Xuefeng Hu, Gokhan Uzunbas, Sirius Chen, Rui Wang, Ashish Shah, Ram Nevatia, and Ser-Nam Lim. Mixnorm: Test-time adaptation through online normalization estimation. arXiv preprint arXiv:2110.11478, 2021.
- 314. Taesik Gong, Jongheon Jeong, Taewon Kim, Yewon Kim, Jinwoo Shin, and Sung-Ju Lee. Note: Robust continual test-time adaptation against temporal correlation. In *Proc. NeurIPS*, pages 27253–27266, 2022.
- 315. Jian Zhang, Lei Qi, Yinghuan Shi, and Yang Gao. Generalizable model-agnostic semantic segmentation via target-specific normalization. *Pattern Recognition*, 122:108292, 2022.
- 316. Yushun Tang, Ce Zhang, Heng Xu, Shuoshuo Chen, Jie Cheng, Luziwei Leng, Qinghai Guo, and Zhihai He. Neuro-modulated hebbian learning for fully test-time adaptation. In Proc. CVPR, 2023.
- 317. Chenyu Yi, Siyuan Yang, Yufei Wang, Haoliang Li, Yappeng Tan, and Alex Kot. Temporal coherent test-time optimization for robust video classification. In *Proc. ICLR*, 2023.
- Aurick Zhou and Sergey Levine. Bayesian adaptation for covariate shift. In *Proc. NeurIPS*, pages 914–927, 2021.
- 319. Prabhu Teja Sivaprasad and François Fleuret. Test time adaptation through perturbation robustness. In *Proc. NeurIPS Workshops*, 2021.
- 320. Liangze Jiang and Tao Lin. Test-time robust personalization for federated learning. In *Proc. ICLR*, 2023.

- Junha Song, Jungsoo Lee, In So Kweon, and Sungha Choi. Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In Proc. CVPR, 2023.
- 322. Hiroaki Kingetsu, Kenichi Kobayashi, Yoshihiro Okawa, Yasuto Yokota, and Katsuhito Nakazawa. Multi-step test-time adaptation with entropy minimization and pseudo-labeling. In *Proc. ICIP*, pages 4153–4157, 2022.
- 323. Sungha Choi, Seunghan Yang, Seokeon Choi, and Sungrack Yun. Improving test-time adaptation via shift-agnostic weight regularization and nearest source prototypes. In *Proc. ECCV*, pages 440–458, 2022.
- 324. Jogendra Nath Kundu, Siddharth Seth, Pradyumna YM, Varun Jampani, Anirban Chakraborty, and R Venkatesh Babu. Uncertainty-aware adaptation for self-supervised 3d human pose estimation. In Proc. CVPR, pages 20448–20459, 2022.
- 325. Puning Yang, Jian Liang, Jie Cao, and Ran He. Auto: Adaptive outlier optimization for online test-time ood detection. arXiv preprint arXiv:2303.12267, 2023.
- 326. Davide Belli, Debasmit Das, Bence Major, and Fatih Porikli. Online adaptive personalization for face antispoofing. In Proc. ICIP, pages 351–355, 2022.
- 327. Inkyu Shin, Yi-Hsuan Tsai, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Sparsh Garg, In So Kweon, and Kuk-Jin Yoon. Mm-tta: Multi-modal test-time adaptation for 3d semantic segmentation. In Proc. CVPR, pages 16928–16937, 2022.
- 328. Hongzheng Yang, Cheng Chen, Meirui Jiang, Quande Liu, Jianfeng Cao, Pheng Ann Heng, and Qi Dou. Dltta: Dynamic learning rate for test-time adaptation on crossdomain medical images. *IEEE Transactions on Medical Imaging*, 41(12):3575–3586, 2022.
- 329. Chaithanya Kumar Mummadi, Robin Hutmacher, Kilian Rambach, Evgeny Levinkov, Thomas Brox, and Jan Hendrik Metzen. Test-time adaptation to distribution shift by confidence maximization and input transformation. arXiv preprint arXiv:2106.14999, 2021.
- 330. Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and Zico Kolter. Test-time adaptation via conjugate pseudolabels. In Proc. NeurIPS, pages 6204–6218, 2022.
- 331. Jun-Kun Wang and Andre Wibisono. Towards understanding gd with hard and conjugate pseudo-labels for test-time adaptation. In *Proc. ICLR*, 2023.
- 332. Longbin Zeng, Jiayi Han, Liang Du, and Weiyang Ding. Rethinking precision of pseudo label: Test-time adaptation via complementary learning. *Pattern Recognition Letters*, 177:96–102, 2024.
- 333. Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In Proc. CVPR, 2023.
- 334. Hai Ye, Yuyang Ding, Juntao Li, and Hwee Tou Ng. Robust question answering against distribution shifts with test-time adaptation: An empirical study. In Proc. EMNLP Findings, 2022.
- 335. Yulu Gan, Xianzheng Ma, Yihang Lou, Yan Bai, Renrui Zhang, Nian Shi, and Lin Luo. Decorate the newcomers: Visual domain prompt for continual test time adaptation. In Proc. AAAI, 2023.
- 336. Devavrat Tomar, Guillaume Vray, Behzad Bozorgtabar, and Jean-Philippe Thiran. Tesla: Test-time self-learning with automatic adversarial augmentation. In Proc. CVPR, 2023.
- 337. Xiaosong Ma, Jie Zhang, Song Guo, and Wenchao Xu. Swapprompt: Test-time prompt adaptation for visionlanguage models. In *Proc. NeurIPS*, 2023.

338. Xin Zhang and Ying-Cong Chen. Adaptive domain generalization via online disagreement minimization. IEEE Transactions on Image Processing, 32:4247–4258, 2023.

- 339. Debasmit Das, Shubhankar Borse, Hyojin Park, Kambiz Azarian, Hong Cai, Risheek Garrepalli, and Fatih Porikli. Transadapt: A transformative framework for online test time adaptive semantic segmentation. In Proc. ICASSP, pages 1–5, 2023.
- 340. Jonathan Samuel Lumentut and In Kyu Park. 3d body reconstruction revisited: Exploring the test-time 3d body mesh refinement strategy via surrogate adaptation. In *Proc. ACM-MM*, pages 5923–5933, 2022.
- Yongyi Su, Xun Xu, and Kui Jia. Revisiting realistic test-time training: Sequential inference and adaptation by anchored clustering. In *Proc. NeurIPS*, pages 17543– 17555, 2022.
- 342. Junyang Chen, Xiaoyu Xian, Zhijing Yang, Tianshui Chen, Yongyi Lu, Yukai Shi, Jinshan Pan, and Liang Lin. Open-world pose transfer via sequential test-time adaption. In Proc. CVPR, 2023.
- 343. Shanyan Guan, Jingwei Xu, Yunbo Wang, Bingbing Ni, and Xiaokang Yang. Bilevel online adaptation for out-of-domain human mesh reconstruction. In *Proc. CVPR*, pages 10472–10481, 2021.
- 344. Yevhen Kuznietsov, Marc Proesmans, and Luc Van Gool. Towards unsupervised online domain adaptation for semantic segmentation. In *Proc. WACV Workshops*, pages 261–271, 2022.
- 345. Junho Kim, Inwoo Hwang, and Young Min Kim. Ev-tta: Test-time adaptation for event-based object recognition. In Proc. CVPR, pages 17745–17754, 2022.
- 346. Qilong Wu, Xiangyu Yue, and Alberto Sangiovanni-Vincentelli. Domain-agnostic test-time adaptation by prototypical training with auxiliary data. In Proc. NeurIPS Workshops, 2021.
- 347. Muhammad Jehanzeb Mirza, Pol Jané Soneira, Wei Lin, Mateusz Kozinski, Horst Possegger, and Horst Bischof. Actmad: Activation matching to align distributions for test-time-training. In *Proc. CVPR*, pages 24152–24161, 2023.
- 348. Sanghun Jung, Jungsoo Lee, Nanhee Kim, Amirreza Shaban, Byron Boots, and Jaegul Choo. Cafa: Classaware feature alignment for test-time adaptation. In *Proc. ICCV*, pages 19060–19071, 2023.
- 349. Chenyan Wu, Yimu Pan, Yandong Li, and James Z. Wang. Learning to adapt to online streams with distribution shifts. arXiv preprint arXiv:2303.01630, 2023.
- 350. Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In Proc. CVPR, 2024.
- Longhui Yuan, Binhui Xie, and Shuang Li. Robust testtime adaptation in dynamic scenarios. In *Proc. CVPR*, pages 15922–15932, 2023.
- 352. Robert A Marsden, Mario Döbler, and Bin Yang. Universal test-time adaptation through weight ensembling, diversity weighting, and prior correction. In *Proc. WACV*, pages 2555–2565, 2024.
- 353. Chunyu Yang and Jie Zhou. Non-stationary data sequence classification using online class priors estimation. Pattern Recognition, 41(8):2656–2664, 2008.
- 354. Amelie Royer and Christoph H Lampert. Classifier adaptation at prediction time. In *Proc. CVPR*, pages 1401–1409, 2015.
- 355. Ruihan Wu, Chuan Guo, Yi Su, and Kilian Q Weinberger. Online adaptation to label distribution shift. In

- $Proc.\ NeurIPS,\ pages\ 11340-11351,\ 2021.$
- 356. Riccardo Volpi, Pau de Jorge, Diane Larlus, and Gabriela Csurka. On the road to online adaptation for semantic image segmentation. In *Proc. CVPR*, pages 19184–19195, 2022.
- 357. Muhammad Abdullah Jamal, Haoxiang Li, and Boqing Gong. Deep face detector adaptation without negative transfer or catastrophic forgetting. In *Proc. CVPR*, pages 5608–5618, 2018.
- 358. Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In Proc. CVPR, pages 780–790, 2019.
- 359. Wenao Ma, Cheng Chen, Shuang Zheng, Jing Qin, Huimao Zhang, and Qi Dou. Test-time adaptation with calibration of medical image classification nets for label distribution shift. In *Proc. MICCAI*, pages 313–323, 2022.
- 360. Zhenbin Wang, Mao Ye, Xiatian Zhu, Liuhan Peng, Liang Tian, and Yingying Zhu. Metateacher: Coordinating multi-model domain adaptation for medical image classification. In Proc. NeurIPS, pages 20823–20837, 2022.
- 361. Cristiano Saltori, Evgeny Krivosheev, Stéphane Lathuilière, Nicu Sebe, Fabio Galasso, Giuseppe Fiameni, Elisa Ricci, and Fabio Poiesi. Gipso: Geometrically informed propagation for online adaptation in 3d lidar segmentation. In Proc. ECCV, pages 567–585, 2022.
- 362. Cristiano Saltori, Stéphane Lathuilière, Nicu Sebe, Elisa Ricci, and Fabio Galasso. Sf-uda^{3D}: Source-free unsupervised domain adaptation for lidar-based 3d object detection. In *Proc. 3DV*, pages 771–780, 2020.
- 363. Yuecong Xu, Jianfei Yang, Haozhi Cao, Keyu Wu, Wu Min, and Zhenghua Chen. Learning temporal consistency for source-free video domain adaptation. In Proc. ECCV, pages 147–164, 2022.
- Runhao Zeng, Qi Deng, Huixuan Xu, Shuaicheng Niu, and Jian Chen. Exploring motion cues for video testtime adaptation. In *Proc. ACM-MM*, pages 1840–1850, 2023.
- 365. Seyed Mehdi Ayyoubzadeh, Wentao Liu, Irina Kezele, Yuanhao Yu, Xiaolin Wu, Yang Wang, and Tang Jin. Test-time adaptation for optical flow estimation using motion vectors. *IEEE Transactions on Image Processing*, 32:4977–4988, 2023.
- 366. Juliette Bertrand, Giorgos Kordopatis Zilos, Yannis Kalantidis, and Giorgos Tolias. Test-time training for matching-based video object segmentation. In Proc. NeurIPS, 2023.
- Mirco Plananamente, Chiara Plizzari, and Barbara Caputo. Test-time adaptation for egocentric action recognition. In *Proc. ICIAP*, page 206–218, 2022.
- 368. Zhiquan Wen, Shuaicheng Niu, Ge Li, Qingyao Wu, Mingkui Tan, and Qi Wu. Test-time model adaptation for visual question answering with debiased selfsupervisions. *IEEE Transactions on Multimedia*, 26:2137– 2147, 2024.
- 369. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. ICML*, pages 8748–8763, 2021.
- 370. Taoshan Zhang, Youjun Xiang, Xianfeng Li, Zichun Weng, Zhen Chen, and Yuli Fu. Free lunch for cross-

- domain occluded face recognition without source data. In *Proc. ICASSP*, pages 2944–2948, 2022.
- 371. Zeshuai Deng, Zhuokun Chen, Shuaicheng Niu, Thomas Li, Bohan Zhuang, and Mingkui Tan. Efficient testtime adaptation for super-resolution with second-order degradation and reconstruction. In Proc. NeurIPS, 2023.
- 372. Wentao Zhu, Yufang Huang, Daguang Xu, Zhen Qian, Wei Fan, and Xiaohui Xie. Test-time training for deformable multi-scale image registration. In *Proc. ICRA*, pages 13618–13625, 2021.
- 373. Shady Abu Hussein, Tom Tirer, and Raja Giryes. Image-adaptive gan based reconstruction. In Proc. AAAI, pages 3121–3129, 2020.
- 374. Mohammad Zalbagi Darestani, Jiayu Liu, and Reinhard Heckel. Test-time training can close the natural distribution shift performance gap in deep learning based compressed sensing. In *Proc. ICML*, pages 4754–4776, 2022.
- 375. Jianzhao Liu, Xin Li, Shukun An, and Zhibo Chen. Source-free unsupervised domain adaptation for blind image quality assessment. arXiv preprint arXiv:2207.08124, 2022.
- 376. Ancong Wu, Wei-Shi Zheng, Xiaowei Guo, and Jian-Huang Lai. Distilled person re-identification: Towards a more scalable system. In *Proc. CVPR*, pages 1187–1196, 2019.
- 377. Soumava Paul, Aheli Saha, and Abhishek Samanta. Ttt-ucdr: Test-time training for universal cross-domain retrieval. arXiv preprint arXiv:2208.09198, 2022.
- 378. Fanjie Kong, Shuai Yuan, Weituo Hao, and Ricardo Henao. Mitigating test-time bias for fair image retrieval. In *Proc. NeurIPS*, 2023.
- 379. Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-Or. Mystyle: A personalized generative prior. ACM Transactions on Graphics, 41(6):1–10, 2022.
- 380. Bo Zhang, Xiaoming Zhang, Yun Liu, Lei Cheng, and Zhoujun Li. Matching distributions between model and data: Cross-domain knowledge distillation for unsupervised domain adaptation. In Proc. ACL, pages 5423– 5433, 2021.
- 381. Egoitz Laparra, Xin Su, Yiyun Zhao, Ozlem Uzuner, Timothy Miller, and Steven Bethard. Semeval-2021 task 10: Source-free domain adaptation for semantic processing. In *International Workshop on Semantic Evaluation* (SemEval), pages 348–356, 2021.
- 382. Yiqi Wang, Chaozhuo Li, Wei Jin, Rui Li, Jianan Zhao, Jiliang Tang, and Xing Xie. Test-time training for graph neural networks. arXiv preprint arXiv:2210.08813, 2022.
- 383. Malik Boudiaf, Tom Denton, Bart Van Merriënboer, Vincent Dumoulin, and Eleni Triantafillou. In search for a generalizable method for source free domain adaptation. In Proc. ICML, pages 2914–2931, 2023.
- 384. Pilhyeon Lee, Seogkyu Jeon, Sunhee Hwang, Minjung Shin, and Hyeran Byun. Source-free subject adaptation for eeg-based visual recognition. In *Proc. BCI*, pages 1–6, 2023
- 385. Jinyang Jiao, Hao Li, Tian Zhang, and Jing Lin. Source-free adaptation diagnosis for rotating machinery. *IEEE Transactions on Industrial Informatics*, 2022.
- 386. Ziyi Liu and Yongchun Fang. Learning adaptable risk-sensitive policies to coordinate in multi-agent general-sum games. In *Proc. ICONIP*, pages 27–40, 2023.
- 387. Moritz Hardt and Yu Sun. Test-time training on nearest neighbors for large language models. In *Proc. ICLR*,

- 2024.
- 388. Chun-Mei Feng, Kai Yu, Yong Liu, Salman Khan, and Wangmeng Zuo. Diverse data augmentation with diffusions for effective test-time prompt tuning. In *Proc. ICCV*, pages 2704–2714, 2023.
- 389. Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. In *Proc. ICLR*, 2024.
- 390. Mihir Prabhudesai, Tsung-Wei Ke, Alex Li, Deepak Pathak, and Katerina Fragkiadaki. Test-time adaptation of discriminative models via diffusion generative feedback. In *Proc. NeurIPS*, 2023.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. CVPR*, pages 10684–10695, 2022.
- 392. Shurui Gui, Xiner Li, and Shuiwang Ji. Active test-time adaptation: Theoretical analyses and an algorithm. In Proc. ICLR, 2024.
- 393. Yongcan Yu, Lijun Sheng, Ran He, and Jian Liang. Benchmarking test-time adaptation against distribution shifts in image classification. arXiv preprint arXiv:2307.03133, 2023.
- 394. Ori Press, Steffen Schneider, Matthias Kümmerer, and Matthias Bethge. Rdumb: A simple approach that questions our progress in continual test-time adaptation. In *Proc. NeurIPS*, 2023.
- 395. Zixin Wang, Yadan Luo, Liang Zheng, Zhuoxiao Chen, Sen Wang, and Zi Huang. In search of lost online test-time adaptation: A survey. arXiv preprint arXiv:2310.20199, 2023.
- 396. Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *Proc. ICML*, pages 42058–42080, 2023.
- 397. Vadim Borisov, Tobias Leemann, Kathrin Seßler, Johannes Haug, Martin Pawelczyk, and Gjergji Kasneci. Deep neural networks and tabular data: A survey. IEEE Transactions on Neural Networks and Learning Systems, 2022.
- 398. Mohamed Ragab, Emadeldeen Eldele, Wee Ling Tan, Chuan-Sheng Foo, Zhenghua Chen, Min Wu, Chee-Keong Kwoh, and Xiaoli Li. Adatime: A benchmarking suite for domain adaptation on time series data. ACM Transactions on Knowledge Discovery from Data, 2023.
- 399. Ravid Shwartz-Ziv and Amitai Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 81:84–90, 2022.
- 400. Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *Proc. NeurIPS*, pages 507–520, 2022.
- Birhanu Eshete. Making machine learning trustworthy. Science, 373(6556):743–744, 2021.
- 402. Eungyeup Kim, Mingjie Sun, Aditi Raghunathan, and J Zico Kolter. Reliable test-time adaptation via agreement-on-the-line. In Proc. NeurIPS Workshops, 2023.