ACTIVE TEST-TIME ADAPTATION VIA SHIFT-AWARE QUERYING AND CLASS-CONDITIONAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Representations trained on one distribution can encode both predictive information that remains stable across distributions and distribution-specific associations that become unreliable under distribution shift. Test-time adaptation should reduce reliance on these associations while preserving predictive information, but unlabeled target data provides limited supervision for this objective. We propose Active Contrastive Test-Time Adaptation (ACTTA), which combines selective label acquisition with class-conditional representation alignment. ACTTA queries diverse target samples whose representations deviate from class-conditional source statistics. A supervised contrastive objective encourages representation alignment to reduce distribution-specific associations, while cross-entropy maintains class discrimination. Under an idealized data-generating process, our analysis characterizes conditions under which class-conditional invariance excludes distribution-varying information while preserving label information. Experiments on four vision and language benchmarks spanning temporal, geographic, and synthetic shifts show that ACTTA achieves the highest performance among the evaluated methods. Further evaluations demonstrate improved generalization to unseen distributions and preservation of performance in the training distribution.

1 INTRODUCTION

The lack of robustness of machine learning models under data distribution shifts remains a major obstacle to their reliable real-world deployment (Koh et al., 2021; Yao et al., 2022). Test-time adaptation (TTA) addresses this challenge by adapting a pre-trained model to the test distribution using incoming unlabeled test data during inference (Adachi et al., 2025; Iwasawa & Matsuo, 2021; Wang et al., 2021). In many prediction tasks, the underlying data-generating process contains stable causal mechanisms, where certain causal features exert an invariant influence on the target (Huang et al., 2023; Sun et al., 2021), while non-causal features may be only incidentally associated with the target in the training distribution. Under distribution shift, such non-causal associations may change, whereas causal relationships tend to remain stable. This motivates learning representations that preserve label information while remaining invariant to changes in non-causal features (Kaur et al., 2023; Lv et al., 2022). Nevertheless, existing TTA methods largely overlook features that are not stably associated with the label and inadvertently reinforce spurious correlations, contributing to error accumulation during adaptation (Niu et al., 2023), loss of generalization (Lim et al., 2026), and forgetting of the training distribution (Niu et al., 2022).

Learning representations that remain predictive across distributions is particularly challenging in TTA. Representations of the pretrained model may rely on feature–label associations that are predictive in the training distribution but unstable under distribution shift. Maintaining predictive performance therefore requires reducing such information while preserving label-relevant information. Achieving this objective is difficult with only unlabeled test inputs, as there is no direct supervision for assessing whether feature–label associations remain reliable. Although existing TTA methods DeYO Lee et al. (2024) and TACT Liu et al. (2025) attempt to reduce reliance on non-causal features, they require prior knowledge to design data augmentations that simulate changes in causal or non-causal features during the data-generating process. However, when such prior knowledge is unavailable or when the features cannot be easily simulated through augmentation, these methods become ineffective.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

To address these limitations, we propose Active Contrastive Test-time Adaptation (ACTTA), an algorithm that incorporates active learning, allowing the model to query labels of a small number of test samples. These labels provide evidence about predictive relationships that remain valid under distribution shift, enabling the model to better preserve label-predictive information shared across distributions. In contrast to prior work, which incorporates active learning into TTA primarily by selecting uncertain or representative samples (Gui et al., 2024; Lee et al., 2025; Tan et al., 2026; Wang & Ding, 2025; Zhao et al., 2026), our approach prioritizes samples whose representations reveal substantial deviations from the source distribution and are therefore particularly informative for adaptation.

The representations learned by a model for prediction capture salient features of the input, with training representations typically concentrated within a bounded region of the representation space (Aghajanyan et al., 2021; Hinton & Salakhutdinov, 2006). When a test sample's representation lies outside this region, it indicates the presence of features that deviate from those seen during training. Such out-of-distribution representations serve as signals of distribution shift (Hendrycks et al., 2019; Hsu et al., 2020), and querying their labels provides supervision where training feature–label associations may be unreliable. To identify such samples, we pre-compute the mean and covariance of training representations for each class, and use them to compute the Mahalanobis distance (Mahalanobis, 2018) between a test representation and the training representations of its predicted class. The samples with the largest distance are selected as candidates for annotation. To ensure that the selected samples cover a broad range of the distribution, we apply a diversity-based selection strategy among these candidates, selecting those that are most different from previously selected samples measured by cosine similarity.

To reduce the impact of distribution-varying features, we aim to learn an invariant representation across distributions by minimizing the distance between the training and test representations of the same class in representation space. Existing methods typically reduce the distance by matching the test distribution to the training distribution in the fixed representation space (Jung et al., 2023; Niu et al., 2024; Zhang et al., 2024), which may preserve spurious correlations already encoded in the training representations. In contrast, we propose to update the representation space with a learned transformation applied to both synthetic training and test representations to encourage same-class alignment across distributions. To encourage representations of the same class to cluster together and those of different classes to remain distinct in the updated space, we employ supervised contrastive learning between test representations and synthetic training representations generated from precomputed class-conditional statistics. The transformation is implemented as a low-rank residual adapter, enabling parameter-efficient updates given limited available samples at test time. As the representation space evolves, the classifier is jointly adapted to the new space.

Under an idealized data-generating process, our theoretical analysis explains how class-conditional invariance can exclude non-causal information while preserving label information. Empirical evaluations on four benchmark datasets demonstrate that ACTTA achieves better TTA performance compared to SOTA methods. We also evaluate the adapted model on out-of-distribution domains as well as the original training distribution. Model adapted with ACTTA achieves better generalization to unseen domains, and suffers less performance degradation on the training distribution compared to other methods. Overall, these results suggest that ACTTA reduces distribution-specific associations while maintaining the predictive performance.

2 Related Work

Many TTA methods have been proposed to improve test-time performance using only unlabeled inference data. One line encourages confident predictions via entropy minimization (Duan et al., 2025; Hu et al., 2025b; Wang et al., 2021; Zhang et al., 2025) or self-training with pseudo-labels (Goyal et al., 2022; Hu et al., 2025a; Kang et al., 2023; Sun et al., 2024; Wang et al., 2022). A second line aligns the test distribution to the training distribution in representation space by matching feature statistics (Niu et al., 2024; You et al., 2025; Zhang et al., 2024) or minimizing discrepancy measures (Chakrabarty et al., 2023; Döbler et al., 2023; Jung et al., 2023; Lai et al., 2026). However, matching toward the training distribution preserves spurious correlations already encoded in training representations, even in class-conditional variants (Jung et al., 2023). A third line adjusts the classifier using nearest-neighbour information (Iwasawa & Matsuo, 2021; Zhang et al., 2023), and a fourth

enforces prediction consistency across augmented views (Chen et al., 2022a; Nguyen et al., 2023; Niu et al., 2025; Zhang et al., 2022). Despite these advances, TTA methods remain susceptible to error accumulation (Gong et al., 2023; Niu et al., 2023), catastrophic forgetting (Niu et al., 2022; Wang et al., 2022), and loss of generalization after over-adaptation (Lim et al., 2026; Press et al., 2023). These are consequences of inadvertently reinforcing spurious patterns that are not stably associated with the prediction target.

A related objective in out-of-distribution generalization is to learn predictive relationships that remain stable across environments (Arjovsky et al., 2019; Chen et al., 2022b; Huang et al., 2023; Lv et al., 2022; Mahajan et al., 2021; Sun et al., 2021). However, these methods operate at training time and cannot correct a deployed model. Two TTA methods, DeYO (Lee et al., 2024) and TACT (Liu et al., 2025) address reliance on features that are not stably associated with the prediction target at test time through designed augmentations. DeYO Lee et al. (2024) uses augmentations that destroy features that are stably associated with the prediction target to filter samples for more reliable model updates. TACT Liu et al. (2025) uses augmentations that alter non-causal features to remove their semantics from the representations. Both require prior knowledge of the causal structure to design appropriate augmentations, limiting applicability when such knowledge is unavailable. Our method removes this requirement by querying labels for a small number of test samples.

Several active TTA methods select informative samples for label annotation, guided by uncertainty or representativeness. SimATTA (Gui et al., 2024) clusters high-entropy samples and queries cluster centers. EATTA (Wang & Ding, 2025) selects samples most sensitive to feature perturbation. BiTTA (Lee et al., 2025) queries only binary correctness feedback via Monte Carlo uncertainty sampling. Further work addresses universal and continual settings (Tan et al., 2026; Zhao et al., 2026), and studies how to leverage annotated samples for hyperparameter selection (Li et al., 2024). In contrast, we select samples whose representations deviate most from the training representations, as they are good indicators of distribution shifts. In addition, we use cosine similarity for diversity-based selection, which is more computationally efficient than clustering.

3 METHODOLOGY

We study test-time adaptation for classification, where a pretrained model is adapted to a test distribution that differs from its training distribution. Test samples arrive sequentially in batches. For each batch, the model produces predictions and updates online. The pretrained model $f \coloneqq h \circ g$ comprises a feature extractor g, which maps an input x to a representation $z = g(x) \in \mathbb{R}^d$, and a classifier h that makes predictions using z. Representations learned by the pretrained model can encode both information that is stable across distributions and distribution-specific associations (Izmailov et al., 2022; Kirichenko et al., 2023). Our objective is to update the learned representation to reduce distribution-specific associations while maintaining the predictive performance.

To achieve this goal, we study an active test-time adaptation algorithm. During adaptation, the model is allowed to select a small number of samples from the test batch, query their labels, and use them as supervision. Following existing active TTA methods Gui et al. (2024); Lee et al. (2025); Tan et al. (2026); Wang & Ding (2025); Zhao et al. (2026), we assume a fixed annotation budget per batch. Additionally, a first-in first-out memory bank of fixed size is available to store previously selected samples and their corresponding labels for training.

Our proposed algorithm, Active Contrastive Test-Time Adaptation (ACTTA), consists of two steps: (1) sample selection that identifies test samples most likely to be exposed to distribution shifts, and (2) adaptation that reduces distribution-varying features by learning representations that are invariant across distributions via supervised contrastive learning. Figure 1 gives an overview of our method.

3.1 ACTIVE SAMPLE SELECTION VIA REPRESENTATION DEVIATION

As suggested by several works, performance degrades under distribution shift because the model's representations of test data drift from those of the training data, causing the learned classifier to make incorrect predictions Jung et al. (2023); Niu et al. (2024); Wang et al. (2023); You et al. (2025); Zhang et al. (2024). During training, a model learns to map inputs to representations that capture useful features for prediction. Prior work have shown that these learned representations do not uniformly occupy the representation space. Instead, they lie on a low-dimensional manifold, leaving most of the

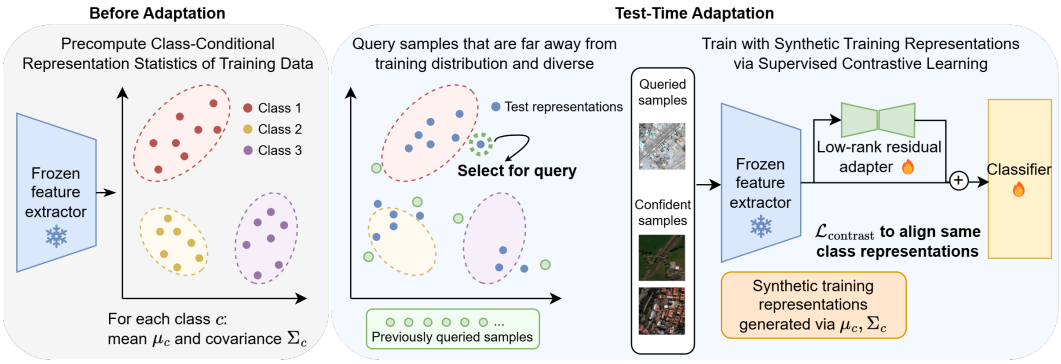


Figure 1: Overview of ACTTA. Class-conditional statistics of training representations are precomputed. Samples that deviate from training representations and are diverse from previous queries are selected for annotation. A low-rank residual adapter is trained via supervised contrastive learning using test samples and synthetic training representations to learn representations that remain predictive across distributions.

space empty Aghajanyan et al. (2021); Hinton & Salakhutdinov (2006). If a test representation lies in a sparsely populated or unseen region, the test sample likely contains unusual feature combinations or entirely new attributes. Therefore, test samples whose representations deviate far from the training distribution should be selected for label annotation, as they are good indicators of distribution shift.

Measuring the deviation solely against the global training representations can be misleading, as different classes often occupy distinct, well-separated regions of the representation space. Therefore, we compute the distance of a test representation to the training representations of its predicted class using Mahalanobis distance (Hsu et al., 2020; Jung et al., 2023; Mahalanobis, 2018). Given the mean μ and covariance matrix Σ of a distribution, the Mahalanobis distance D of a point m to the distribution is defined as $D\left(m;\mu,\Sigma\right)=\sqrt{(m-\mu)^\top\Sigma^{-1}(m-\mu)}$. Deviations in low-variance directions are more emphasized in the distance calculation than in high-variance directions.

To compute the Mahalanobis distance, we precompute the statistics of the training representations. For each class c, given training representations $\{z_i\}_{i=1}^{N_c}$ belonging to this class, we calculate the mean $\mu_c=\frac{1}{N_c}\sum_i z_i$ and covariance $\Sigma_c=\frac{1}{N_c}\sum_i(z_i-\mu_c)(z_i-\mu_c)^\top$, and store these statistics for use during test-time adaptation. We additionally compute the mean within-class training distance $\bar{D}_c=\frac{1}{N_c}\sum_i D\left(z_i;\mu_c,\Sigma_c\right)$, which serves as a normalization factor to account for the variability of the tightness of the clusters formed by representations of different classes.

The feature extractor remains frozen during adaptation, so that the precomputed statistics remain accurate. For a test representation z with predicted class $\hat{c}$, we compute the normalized Mahalanobis distance $\tilde{D}_{\hat{c}}(z)$ using equation 1. A large $\tilde{D}_{\hat{c}}(z)$ indicates that z lies far from the training representations of class $\hat{c}$.

$$\tilde{D}_{\hat{c}}(z)=D\left(z;\mu_{\hat{c}},\Sigma_{\hat{c}}\right)/\bar{D}_{\hat{c}}. \tag{1}$$

To ensure broad coverage of the data distribution, we apply a diversity-based selection strategy among samples with the largest deviations. For each test batch, we compute the normalized Mahalanobis distance per sample and retain samples at or above the τ_d-quantile as candidates for annotation. For example, $\tau_d=0.9$ retains approximately the highest 10% of deviation scores. τ_d balances selecting high-deviation samples with maintaining diversity. Among these candidates, we iteratively select samples that differ the most from previously selected samples stored in the memory bank.

At each time point t, for each candidate representation, we compute its cosine similarity with representations in the memory bank $\mathcal{M}_t=\{(z_1,y_1),(z_2,y_2),\ldots\}$, and take the maximum value as the similarity score shown in equation 2.

$$\text{Similarity}(z)=\max_{(z',y')\in\mathcal{M}_t} z^\top z'/\left(\|z\|\,\|z'\|\right). \tag{2}$$

We query the label of the candidate with the smallest similarity score, and add its representation and label to the memory bank. This selection process is repeated among the remaining candidates until the annotation budget is reached.

3.2 Class-Conditional Representation Alignment

The pretrained feature extractor can learn representations that encode both predictive information that is stable across distributions and distribution-specific associations (Izmailov et al., 2022; Kirichenko et al., 2023). To reduce distribution-specific associations, we propose to learn a transformation on the representations extracted by the feature extractor to enforce invariance across training and test distributions while preserving class discrimination In doing so, information that is unstable across distributions are reduced, while information that are stable are preserved.

To learn invariant representations, we need access to training representations, which are unavailable at test time. We synthesize training representations by leveraging the precomputed mean and covariance matrix. Following existing work Jung et al. (2023); Lee et al. (2020), we assume that the representations of each class c follow a Gaussian distribution $\mathcal{N}(\mu_c, \Sigma_c)$. From the eigendecomposition $\Sigma_c = U_c \Lambda_c U_c^\top$, where U_c is the matrix eigenvectors, and Λ_c is the diagonal matrix of eigenvalues, we generate synthetic representations for class c using Equation equation 3.

$$\tilde{z} = \mu_c + U_c \Lambda_c^{1/2} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{3}$$

We adopt supervised contrastive learning Khosla et al. (2020) to learn invariant representations by pulling representations of the same-class samples closer and pushing representations of different-class samples apart. The objective implicitly maximizes mutual information between same-class representations across distributions Chen et al. (2022b), and theoretically guarantees reduced distances between same-class representations across distributions Wang & Isola (2020).

A residual adaptor φ is introduced to update the representation as $z^\varphi = z + \varphi(z)$. To implement φ in a parameter-efficient manner, inspired by Low-Rank Adaptation (Hu et al., 2022), it consists of two linear layers with a non-linear ReLU activation in between. The inner dimension of the linear layers φ_d is much smaller than the representation dimension. To address the cold-start problem, where limited labeled data is collected in early batches, we initialize the final layer of φ with near-zero weights. This ensures that $z^\varphi \approx z$ at initialization, allowing φ to learn gradually without introducing abrupt changes to the representations.

As contrastive learning benefits from a large number of representation pairs (Chen et al., 2020), we augment the annotated samples in the memory bank with confident samples from the current test batch. For each test sample, we compute the prediction entropy $H = -\sum_c p_c \log p_c$, where p_c is the predicted probability of class c. A sample is considered confident if its H is below a threshold, defined as the τ_e percentile of entropies over the batch. These samples are then assigned pseudo-labels based on their predictions. Given batch $\mathcal{B}_{\text{test}}$ formed by the memory bank and confident samples, we generate a synthetic training batch $\mathcal{B}_{\text{train}}$ of equal size and identical class distribution using equation 3. We then perform supervised contrastive learning on the combined batch of $\mathcal{B} = \mathcal{B}_{\text{test}} \cup \mathcal{B}_{\text{train}}$, where positive pairs are samples that share the same label.

For every sample i in $\mathcal{B}$, let $P(i) = \{j \in \mathcal{B} : j \neq i, y_j = y_i\}$ denote its positive set. The supervised contrastive loss is

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{|\mathcal{B}|} \sum_{i=1}^{|\mathcal{B}|} \frac{1}{|P(i)|} \sum_{j \in P(i)} \log \frac{\exp\left(\text{sim}(z_i^\varphi, z_j^\varphi)/\kappa\right)}{\sum_{k \in \mathcal{B}} \exp\left(\text{sim}(z_i^\varphi, z_k^\varphi)/\kappa\right)}, \tag{4}$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and κ is a temperature hyperparameter.

The classifier h makes predictions using the updated representation z^φ. To preserve class discrimination as the representation changes, we include a cross-entropy loss ℓ_{CE} on the combined batch $\mathcal{B}$. The overall objective combines the cross-entropy and contrastive losses $\mathcal{L} = \ell_{\text{CE}} + \mathcal{L}_{\text{contrast}}$, which is minimized to update the residual adapter φ and classifier h.

4 Theoretical Analysis

We use the following analysis to explain how ACTTA's contrastive and cross-entropy objectives encourage reducing distribution-specific associations while maintaining predictive performance. Let X, Y, X_C, X_{NC}, and E denote the input, label, causal features, non-causal features, and environment, respectively. We consider the data-generating process in Figure 2:

$$Y = \psi_Y(X_C), \quad X = \psi_X(X_C, X_{NC}), \quad X_{NC} = \psi_{NC}(E), \quad X_C \perp E, \tag{5}$$

where the functions are deterministic and measurable. The causal features determine the label, while the environment determines the non-causal features. We consider the ideal case where we have infinitely many independent samples from the data-generating process. The above assumptions imply $Y \perp E$ and $Y \perp X_{NC}$.

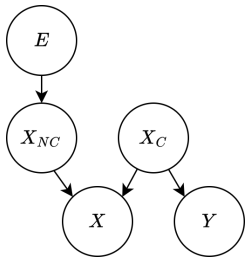


Figure 2: Data-generating process used in the analysis.

Let $\phi : \mathcal{X} \to \mathbb{R}^d$ be a representation function.

Definition 1 (Distribution-invariant representation)**.** *ϕ is distribution-invariant if $I(\phi(X); E) = 0$, meaning that the representation and environment are independent.*

Proposition 1. *Given $X_{NC} = \psi_{NC}(E)$, for any distribution-invariant ϕ, $I(\phi(X); X_{NC}) = 0$.*

Proposition 2. *Given $X_C \perp E$ and $Y = \psi_Y(X_C)$, for all ϕ such that $I(\phi(X); E \mid Y) = 0$, ϕ is distribution-invariant.*

Together, these results show that independence from the environment within each class implies that the representation contains no non-causal information under equation 5. ACTTA's contrastive objective encourages this within-class alignment by pulling same-class source and target representations together through a shared transformation. The global minimizer of the contrastive loss would collapse same-class representations to identical points Wang & Isola (2020), achieving $I(z^\varphi; E \mid Y) = 0$ for the observed class-conditional distributions. This provides a motivation for reducing distribution-specific information. Proofs are in Appendices A.1 and A.2.

However, invariance alone does not ensure that the representation remains predictive. Proposition 3 characterizes when a distribution-invariant representation retains the full predictive information of the causal feature.

Proposition 3. *Under equation 5, for any ϕ satisfying $I(\phi(X); E \mid Y) = 0$, $I(\phi(X); Y) = I(X_C; Y)$ if and only if $H(Y \mid \phi(X), E) = 0$.*

The result states that the representation retains all label information in the causal features when the label is fully determined by the representation and environment. Refer to Appendix A.3 for the proof. ACTTA's cross-entropy objective encourages accurate predictions from z^φ, supporting preservation of label information. Since $H(Y \mid z^\varphi, E) \leq H(Y \mid z^\varphi)$, making the label predictable from the representation also supports the proposition's condition. Thus, the contrastive objective encourages same-class alignment across distributions, while cross-entropy maintains class discrimination in the updated space.

5 Performance Study

We study the performance of ACTTA on vision and language datasets covering various types of distribution shifts, including temporal, geographic, and synthetic shifts. Compared to prior works that primarily benchmark on image data, our experiments offer a broader evaluation TTA methods 268 across diverse, real-world distribution shifts

Datasets. The datasets we studied are summarized below:

- **FMoW** (Christie et al., 2018) from Wilds (Koh et al., 2021) is a satellite imagery dataset where the task is to classify the functional purpose of land use into 62 categories. Distribution shift arises from variation in geographic region and acquisition year. The test set contains 22,108 images.
- **iWildCam** (Beery et al., 2020) from Wilds (Koh et al., 2021) is a wildlife camera-trap dataset requiring species identification across 182 categories. Distribution shift arises from deployment at different camera trap locations. The test set contains 42,791 images.

- **HuffPost** (Rishabh & Grover, 2021) from Wild-Time (Yao et al., 2022) is a news headline classification dataset with 11 categories drawn from the HuffPost archive. Distribution shift arises from the temporal change of the style or content of events covered in the news. The test set contains 24,397 headlines.
- **Waterbirds** (Wah et al., 2011) curated by (Sagawa* et al., 2020) is a synthetic binary classification dataset that places waterbirds and landbirds on water or land backgrounds to introduce a spurious correlation between bird type and background in the training data, which breaks in the test set. The test set contains 5,794 images.

Baselines. We compare ACTTA with Active TTA methods and TTA methods that do not actively query labels at test time. The baselines are summarized below:

Active TTA baselines. **SimATTA** (Gui et al., 2024) queries centers of clusters of representations of high-entropy test samples, and balances the learning of high-entropy sample and low-entropy training-distribution-like samples. **EATTA** (Wang & Ding, 2025) selects perturbation-sensitive samples for supervised updates, and unsupervisedly updates confident low-entropy samples. The two objectives are balanced via gradient-norm-based debiasing. **BiTTA** (Lee et al., 2025) reduces annotation cost by querying only binary correctness feedback for samples chosen by Monte Carlo uncertainty sampling. Reinforcement learning is used to balance adaptation on annotated samples with agreement-based confident samples.

Non-Active TTA baselines. **T3A** (Iwasawa & Matsuo, 2021) updates the classification prototypes with confident test representations. **Tent** (Wang et al., 2021) performs entropy minimization to encourage confident predictions. **SAR** (Niu et al., 2023) filters high-entropy or noisy samples and includes sharpness-aware minimization and model reset to reduce error accumulation during adaptation. **CAFA** (Jung et al., 2023) aligns test representations to the training distribution by minimizing the class-conditional Mahalanobis distance. **Ada-ReAlign** (Zhang et al., 2024) maintains a mixture of feature prototypes and adaptively realigns representations as the test distribution evolves. **GOLD** (Lai et al., 2026) identifies a subspace in the representation space which enables efficient adaptation and preserves generalization. **DeYO** (Lee et al., 2024) identifies confident samples using only causal features via augmentations that destroy those features, then applies confidence-reweighted entropy minimization for adaptation. **TACT** (Liu et al., 2025) applies augmentations that alter non-causal features to update representations and classification prototypes.

Model architecture. We study ACTTA on the architectures used by the data publishers. DenseNet-121 (Huang et al., 2017) is used for FMoW, ResNet-50 (He et al., 2016) is used for iWildCam and Waterbirds, and DistilBERT (Sanh et al., 2019) is used for HuffPost. We use the published pretrained model when available. Appendix B.1 provides more details on the models we studied.

Hyperparameters and model selection. We perform test-time adaptation with a test batch size of 64. Following (Wang & Ding, 2025), we set the annotation budget to 1 sample per batch and the memory bank size to 300 samples for all active TTA methods. We tune three hyperparameters for ACTTA: the distance percentile threshold $\tau_d \in \{0.9, 0.7, 0.5\}$, the entropy percentile threshold $\tau_e \in \{0.2, 0.4, 0.6\}$, and the inner dimension of the residual adaptor $\varphi_d \in \{16, 64, 128\}$. The contrastive temperature κ is set to 1. We use SGD for optimization with momentum 0.9 (Wang et al., 2021). The learning rate for all methods is searched in $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$. The hyperparameters are searched within the ranges specified in their respective papers for baseline methods. Appendix B.2 provides the details. Following the protocol recommended in Zhao et al. (2023), we employ oracle selection to choose the best-performing hyperparameters.

5.1 Test-Time Adaptation Performance

Following the evaluation protocol of each dataset, we report the worst accuracy across geographical regions for FMoW, macro F1 for iWildCam, average accuracy for HuffPost, and worst-group accuracy for Waterbirds, whose data are grouped by background and bird type. GOLD, DeYO, and TACT rely on data augmentations during adaptation, and lack instructions on augmentations for text inputs. As a result, these methods are unable to adapt on the HuffPost dataset. Table 1 shows the test-time adaptation performance, reporting the mean and standard deviation over three experimental runs. ACTTA achieves the highest accuracy on all four datasets, suggesting that it is an effective strategy for improving TTA performance under diverse distribution shifts. Non-active TTA methods yield only marginal improvement over no TTA baseline, highlighting the limitations of existing methods

Table 1: Test-time adaptation performance (%). The best performance of each dataset is in bold, and the second-best is underlined.

	Method	FMoW	iWildCam	HuffPost	Waterbirds
	no TTA	31.16	32.29	70.35	60.28
non active TTA	T3A	32.65 ± 0.20	26.11 ± 0.17	69.96 ± 0.05	68.43 ± 0.45
	Tent	30.90 ± 0.47	33.07 ± 0.66	70.62 ± 0.03	53.32 ± 0.48
	SAR	30.84 ± 0.43	33.27 ± 0.79	70.59 ± 0.02	50.47 ± 0.22
	CAFA	30.85 ± 0.44	32.47 ± 0.22	70.69 ± 0.02	52.65 ± 0.13
	Ada-ReAlign	30.52 ± 0.27	33.07 ± 0.26	69.81 ± 0.05	52.34 ± 0.34
	GOLD	32.51 ± 0.40	32.29 ± 0.09	–	62.36 ± 0.70
	DeYO	31.17 ± 0.50	33.31 ± 0.69	–	71.50 ± 0.76
	TACT	31.53 ± 0.41	23.54 ± 0.67	–	62.93 ± 0.44
active TTA	SimATTA	$\underline{38.28 \pm 0.48}$	$\underline{37.58 \pm 0.48}$	$\underline{70.77 \pm 0.18}$	66.20 ± 1.90
	EATTA	34.12 ± 1.01	34.57 ± 0.12	70.72 ± 0.11	59.40 ± 0.53
	BiTTA	32.46 ± 1.23	35.26 ± 0.06	70.05 ± 0.17	$\underline{75.08 \pm 1.21}$
	ACTTA	$\mathbf{39.49 \pm 0.38}$	$\mathbf{40.31 \pm 1.15}$	$\mathbf{71.14 \pm 0.07}$	$\mathbf{76.86 \pm 0.24}$

Table 2: Post-adaptation performance on a held-out validation distribution (%).

Method	FMoW	iWildCam	HuffPost	Average
no TTA	50.19	36.44	71.41	52.68
SimATTA	$\underline{49.91 \pm 3.50}$	$\underline{36.32 \pm 0.28}$	71.20 ± 0.50	52.48
EATTA	47.45 ± 3.35	34.36 ± 1.47	$\underline{71.29 \pm 0.11}$	51.07
BiTTA	43.38 ± 5.00	34.51 ± 0.58	70.93 ± 0.27	49.61
ACTTA	$\mathbf{52.10 \pm 0.68}$	$\mathbf{36.60 \pm 0.70}$	$\mathbf{71.63 \pm 0.04}$	**53.44**

under large distribution shifts. In contrast, active TTA methods consistently outperform non-active methods, demonstrating that querying for a small number of labels can provide critical feedback to effectively guide the adaptation.

5.2 Generalization to Unseen Distributions

A model that reduces distribution-specific associations with the prediction target in the learned representation should generalize better to new distributions that were not seen during adaptation. To verify this, after adapting each model on the test split, we evaluate its performance on a held-out validation split whose distribution is different from the test split and the training split. Waterbirds is excluded from the analysis as there is no validation split of a different distribution. Table 2 reports these post-adaptation generalization results. ACTTA achieves the best average accuracy, even outperforming the base model. The improvement stems from its contrastive alignment that explicitly reduces distribution-variant components of the representation, leading to representations that transfer more robustly across distributions.

5.3 Performance of Adapted Models on Original Training Distribution

Adaptation to the test distribution risks degrading performance on the original training distribution. To assess this effect, we evaluate the model on the training distribution after adaptation. We test it on the in-distribution test split of FMoW, iWildCam and HuffPost, and the training split of Waterbirds since an in-distribution test split is not available. Table 3 shows the results. As expected, all methods exhibit some drop in in-distribution performance relative to the no-TTA baseline. Nevertheless, ACTTA best preserves in-distribution accuracy overall.

5.4 Ablation Study

We perform ablation studies on the sample selection strategy and the composition of the training samples used in the adaptation process. Table 4(a) compares our selection strategy, which integrates distance and diversity criteria, with random selection, as well as variants that employ each criterion

Table 3: Post-adaptation performance on the training distribution (%).

Method	FMoW	iWildCam	HuffPost	Waterbirds	Average
no TTA	**58.33**	**47.38**	**81.21**	**100.00**	**71.73**
SimATTA	51.50 ± 3.88	46.65 ± 0.26	79.58 ± 1.19	**100.00 ± 0.00**	69.43
EATTA	52.53 ± 5.72	43.70 ± 1.63	80.97 ± 0.13	99.86 ± 0.11	69.27
BiTTA	48.81 ± 9.38	43.44 ± 1.50	79.57 ± 1.75	88.95 ± 7.86	65.19
ACTTA	55.48 ± 0.29	47.00 ± 0.61	79.72 ± 0.45	99.93 ± 0.06	70.53

Table 4: Ablation study on sample selection and training sample composition.

(a) Sample selection strategy.

Distance	Diversity	FMoW	HuffPost
Random Selection		38.66 ± 0.92	70.95 ± 0.04
✓		36.39 ± 0.56	71.07 ± 0.05
	✓	37.72 ± 1.34	70.93 ± 0.08
✓	✓	**39.49 ± 0.38**	**71.14 ± 0.07**

(b) Training sample composition.

Queried Test	Conf. Test	Syn. Train	FMoW	HuffPost
No TTA			31.16	70.35
✓	✓		39.39 ± 0.52	**71.14 ± 0.04**
✓		✓	38.21 ± 0.51	70.96 ± 0.18
✓	✓	✓	**39.49 ± 0.38**	**71.14 ± 0.07**

in isolation. Our full selection strategy outperforms all alternatives, highlighting the importance of both criteria. Table 4(b) presents the performance when either no confident test samples are included in training, or loss of synthetic training representations does not contribute to the model update. Excluding confident test samples leads to a drop in TTA performance. While omitting synthetic training representations does not substantially hurt TTA performance, Table 5 shows that it causes a significant drop in post-adaptation performance on both the held-out validation split with unseen distribution and the original training distribution for FMoW. Including both achieves the best TTA performance while also improving generalization and reducing performance drop on training distribution, confirming their roles in learning invariant representations.

Table 5: Post-adaptation performance under different training sample compositions.

Queried Test	Conf. Test	Syn. Train	FMoW		HuffPost	
			Unseen Distribution	Training Distribution	Unseen Distribution	Training Distribution
No TTA			50.19	58.33	71.41	81.21
✓	✓		49.16 ± 1.52	50.85 ± 1.55	71.54 ± 0.03	**79.89 ± 0.15**
✓		✓	49.98 ± 1.90	55.47 ± 0.39	**71.63 ± 0.17**	79.67 ± 0.52
✓	✓	✓	**52.10 ± 0.68**	55.48 ± 0.29	**71.63 ± 0.04**	79.72 ± 0.45

6 Conclusion and Future Work

We presented ACTTA, an active test-time adaptation method designed to reduce distribution-specific associations while preserving predictive information. ACTTA combines source-relative, diversity-based label acquisition with a shared transformation of source and target representations. Contrastive learning encourages class-conditional alignment, while cross-entropy maintains class discrimination. Our analysis explains the compatibility of invariance and label-information preservation under idealized assumptions. Across four vision and language benchmarks, ACTTA achieves the highest performance among the evaluated methods, with better generalization to the unseen distribution and preservation of performance on the original distribution. However, ACTTA's effectiveness depends on the frozen feature extractor capturing sufficient predictive information. If the representation output by the frozen feature extractor contains limited predictive information, the adaptation may be limited in its ability to learn representations that predictive while reducing distribution-varying features. Future work should address this limitation.

AI use statement

In this work, we used generative AI tools for formulate mathematical claims, assist in the writing of proofs, propose or refine hypotheses, and design or provide feedback on research methodology or experiments; We have not used generative AI tools for help develop theoretical models or conceptual

frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, implement methods, and interpret results; and generate synthetic data sets, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis are not applicable to this work. We have reviewed all AI-assisted work. We checked the correcteness of the mathematical claims and proofs, and verified that the proposed hypotheses are consistent with the experimental results. . We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Kazuki Adachi, Shin'ya Yamaguchi, Atsutoshi Kumagai, and Tomoki Hamagami. Test-time adaptation for regression by subspace alignment. In *International Conference on Learning Representations*, 2025.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pp. 7319–7328, 2021.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Sara Beery, Elijah Cole, and Arvi Gjoka. The iwildcam 2020 competition dataset. *arXiv preprint arXiv:2004.10340*, 2020.

Goirik Chakrabarty, Manogna Sreenivas, and Soma Biswas. SANTA: Source anchoring network and target alignment for continual test time adaptation. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=V7guVYzvE4.

Dian Chen, Dequan Wang, Trevor Darrell, and Sayna Ebrahimi. Contrastive test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 295–305, 2022a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Yongqiang Chen, Yonggang Zhang, Yatao Bian, Han Yang, MA KAILI, Binghui Xie, Tongliang Liu, Bo Han, and James Cheng. Learning causally invariant representations for out-of-distribution generalization on graphs. In *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=A6AFK_JwrIW.

Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Mario Döbler, Robert A Marsden, and Bin Yang. Robust mean teacher for continual and gradual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7704–7714, 2023.

Dexin Duan, Rui Xu, Peilin Liu, and Fei Wen. Lifelong test-time adaptation via online learning in tracked low-dimensional subspace. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=NFvAa2hNzH.

Taesik Gong, Yewon Kim, Taeckyung Lee, Sorn Chottananurak, and Sung-Ju Lee. SoTTA: Robust test-time adaptation on noisy data streams. In *Advances in Neural Information Processing Systems*, 2023.

Sachin Goyal, Mingjie Sun, Aditi Raghunathan, and J Zico Kolter. Test time adaptation via conjugate pseudo-labels. *Advances in Neural Information Processing Systems*, 35:6204–6218, 2022.

Shurui Gui, Xiner Li, and Shuiwang Ji. Active test-time adaptation: Theoretical analyses and an algorithm. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=YHUGlwTzFB.

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

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyxCxhRcY7.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10951–10960, 2020.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Yihao Hu, Congyu Qiao, Xin Geng, and Ning Xu. Selective label enhancement learning for test-time adaptation. In *International Conference on Learning Representations*, 2025a.

Zixuan Hu, Yichun Hu, Xiaotong Li, SHIXIANG TANG, and LINGYU DUAN. Beyond entropy: Region confidence proxy for wild test-time adaptation. In *International Conference on Machine Learning*, 2025b.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Zhuo Huang, Xiaobo Xia, Li Shen, Bo Han, Mingming Gong, Chen Gong, and Tongliang Liu. Harnessing out-of-distribution examples via augmenting content and style. In *International Conference on Learning Representations*, 2023.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems*, volume 34, pp. 2427–2440, 2021.

Pavel Izmailov, Polina Kirichenko, Nate Gruver, and Andrew Gordon Wilson. On feature learning in the presence of spurious correlations. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=wKhUPzqVap6.

Sanghun Jung, Jungsoo Lee, Nanhee Kim, Amirreza Shaban, Byron Boots, and Jaegul Choo. Cafa: Class-aware feature alignment for test-time adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19060–19071, 2023.

Juwon Kang, Nayeong Kim, Donghyeon Kwon, Jungseul Ok, and Suha Kwak. Leveraging proxy of training data for test-time adaptation. In *International Conference on Machine Learning*, pp. 15737–15752. PMLR, 2023.

Jivat Neet Kaur, Emre Kiciman, and Amit Sharma. Modeling the data-generating process is necessary for out-of-distribution generalization. In *International Conference on Learning Representations*, 2023.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in neural information processing systems*, 33:18661–18673, 2020.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Zb6c8A-Fghk.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pp. 5637–5664. PMLR, 2021.

Guannan Lai, Da-Wei Zhou, Zhenguo Li, and Han-Jia Ye. The golden subspace: Where efficiency meets generalization in continual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Dongha Lee, Sehun Yu, and Hwanjo Yu. Multi-class data description for out-of-distribution detection. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1362–1370, 2020.

Jonghyun Lee, Dahuin Jung, Saehyung Lee, Junsung Park, Juhyeon Shin, Uiwon Hwang, and Sungroh Yoon. Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In *International Conference on Learning Representations*, 2024.

Taeckyung Lee, Sorn Chottananurak, Junsu Kim, Jinwoo Shin, Taesik Gong, and Sung-Ju Lee. Test-time adaptation with binary feedback. In *International Conference on Machine Learning*, 2025.

Yushu Li, Yongyi Su, Xulei Yang, Kui Jia, and Xun Xu. Exploring human-in-the-loop test-time adaptation by synergizing active learning and model selection. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=P09rAv8UH7.

Taejun Lim, Joong-Won Hwang, and Kibok Lee. When and where to reset matters for long-term test-time adaptation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=0JayjvOKxt.

Yingnan Liu, Rui Qiao, Mong-Li Lee, and Wynne Hsu. Test-time adaptation by causal trimming. In *Advances in Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=zFGdHL9pcD.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8046–8056, 2022.

Divyat Mahajan, Shruti Tople, and Amit Sharma. Domain generalization using causal matching. In *International Conference on Machine Learning*, pp. 7313–7324. PMLR, 2021.

Prasanta Chandra Mahalanobis. On the generalized distance in statistics. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, 80:S1–S7, 2018.

A. Tuan Nguyen, Thanh Nguyen-Tang, Ser-Nam Lim, and Philip H.S. Torr. Tipi: Test time adaptation with transformation invariance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24162–24171, 06 2023.

Shuaicheng Niu, Jiaxiang Wu, Yifan Zhang, Yaofo Chen, Shijian Zheng, Peilin Zhao, and Mingkui Tan. Efficient test-time model adaptation without forgetting. In *International Conference on Machine Learning*, pp. 16888–16905. PMLR, 2022.

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 *International Conference on Learning Representations*, 2023.

Shuaicheng Niu, Chunyan Miao, Guohao Chen, Pengcheng Wu, and Peilin Zhao. Test-time model adaptation with only forward passes. In *International Conference on Machine Learning*, 2024.

Shuaicheng Niu, Guohao Chen, Peilin Zhao, Tianyi Wang, Pengcheng Wu, and Zhiqi Shen. Self-bootstrapping for versatile test-time adaptation. In *International Conference on Machine Learning*, 2025.

Ori Press, Steffen Schneider, Matthias Kuemmerer, and Matthias Bethge. RDumb: A simple approach that questions our progress in continual test-time adaptation. In *Advances in Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=VfP6VTVsHc.

Misra Rishabh and Jigyasa Grover. Sculpting data for ml: The first act of machine learning, 2021.

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

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Haopeng Sun, Lumin Xu, Sheng Jin, Ping Luo, Chen Qian, and Wentao Liu. PROGRAM: PROtotype GRAph model based pseudo-label learning for test-time adaptation. In *International Conference on Learning Representations*, 2024.

Xinwei Sun, Botong Wu, Xiangyu Zheng, Chang Liu, Wei Chen, Tao Qin, and Tie-Yan Liu. Recovering latent causal factor for generalization to distributional shifts. In *Advances in Neural Information Processing Systems*, volume 34, pp. 16846–16859, 2021.

Jiayao Tan, Fan Lyu, Chenggong Ni, Fuyuan Hu, Wei Feng, and Rui Yao. Exposing mixture and annotating confusion for active universal test-time adaptation. In *International Conference on Learning Representations*, 2026.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations*, 2021.

Guowei Wang and Changxing Ding. Effortless active labeling for long-term test-time adaptation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 25633–25642, 2025.

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

Shuai Wang, Daoan Zhang, Zipei Yan, Jianguo Zhang, and Rui Li. Feature alignment and uniformity for test time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20050–20060, 2023.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 9929–9939, 2020.

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei W Koh, and Chelsea Finn. Wild-time: A benchmark of in-the-wild distribution shift over time. In *Advances in Neural Information Processing Systems*, 2022.

Linjing You, Jiabao Lu, and Xiayuan Huang. Test-time correlation alignment. In *International Conference on Machine Learning*, 2025.

Marvin Zhang, Sergey Levine, and Chelsea Finn. Memo: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 38629–38642, 2022.

Qingyang Zhang, Yatao Bian, Xinke Kong, Peilin Zhao, and Changqing Zhang. COME: Test-time adaption by conservatively minimizing entropy. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=506BjJ1ziZ.

Yifan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, and Tieniu Tan. Adanpc: Exploring non-parametric classifier for test-time adaptation. In *International Conference on Machine Learning*, pp. 41647–41676. PMLR, 2023.

Zhen-Yu Zhang, Zhiyu Xie, Huaxiu Yao, and Masashi Sugiyama. Test-time adaptation in non-stationary environments via adaptive representation alignment. In *Advances in Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=0EfUYVMrLv.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *International Conference on Machine Learning*, pp. 42058–42080. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhao23d.html.

Kewei Zhao, Guangle Song, Chenyu Ge, Wenqi Hu, and Xinfeng Liu. Active test-time adaptation for continual medical image classification. *Pattern Recognition*, pp. 113288, 2026.

A Details of Theoretical Analysis

The following results use the data-generating process in equation 5. The assumptions $X_C \perp E$ and $Y = \psi_Y(X_C)$ imply that the label distribution is the same across environments, while $X_{NC} = \psi_{NC}(E)$ means that the non-causal features are determined by the environment.

A.1 Distribution Invariance and Non-Causal Features

Proposition 1. Given $X_{NC} = \psi_{NC}(E)$, for any distribution-invariant ϕ, $I(\phi(X); X_{NC}) = 0$.

Proof. As the non-causal feature X_{NC} is obtained from E through a deterministic measurable function ψ_{NC}, by the data processing inequality, we have

$$I(\phi(X); E) \geq I(\phi(X); \psi_{NC}(E)) = I(\phi(X); X_{NC}). \tag{6}$$

As mutual information is non-negative,

$$I(\phi(X); E) \geq I(\phi(X); X_{NC}) \geq 0. \tag{7}$$

By Definition 1, distribution invariance gives $I(\phi(X); E) = 0$.

Therefore, $I(\phi(X); X_{NC}) = 0$. □

A.2 Class-Conditional Invariance

Proposition 2. Given $X_C \perp E$ and $Y = \psi_Y(X_C)$, for all ϕ such that $I(\phi(X); E \mid Y) = 0$, ϕ is distribution-invariant.

Proof. As X_C is independent of E, and Y is obtained from X_C by a deterministic measurable function ψ_Y, the label is also independent of the environment, i.e., $Y \perp E$.

Let $Z = \phi(X)$. Given $I(Z; E \mid Y) = 0$,

$$p(z, e \mid y) = p(z \mid y)p(e \mid y). \tag{8}$$

Given $Y \perp E$,

$$p(e \mid y) = p(e). \tag{9}$$

By Equation equation 8 and equation 9, the joint distribution $p(z, e)$ can be written as

$$\begin{aligned} p(z,e) &= \sum_y p(z,e \mid y)p(y) \\ &= \sum_y p(z \mid y)p(e \mid y)p(y) \\ &= \sum_y p(z \mid y)p(e)p(y) \\ &= p(e)\sum_y p(z \mid y)p(y) \\ &= p(e)p(z), \end{aligned} \tag{10}$$

showing that Z and E are independent, i.e. $I(\phi(X); E) = 0$. □

A.3 Preserving Label Information

Proposition 3. Under the data-generating process defined in Equation equation 5, for any ϕ that satisfies $I(\phi(X);\ E \mid Y) = 0$, $I(\phi(X);\ Y) = I(X_C;\ Y)$ if and only if $H(Y \mid \phi(X),\ E) = 0$.

Proof. Let $Z = \phi(X)$. The model assumptions give $Y \perp E$, and Proposition 2 gives $I(Z; E) = 0$. Applying the chain rule of mutual information,

$$I(Z;Y) + I(Z;E \mid Y) = I(Z;E) + I(Z;Y \mid E). \tag{11}$$

Since $I(Z; E \mid Y) = I(Z; E) = 0$, it follows that $I(Z; Y) = I(Z; Y \mid E)$. For the discrete class label Y,

$$\begin{aligned} I(Z;Y) &= H(Y \mid E) - H(Y \mid Z, E) \\ &= H(Y) - H(Y \mid Z, E) \\ &= I(X_C;Y) - H(Y \mid Z, E), \end{aligned} \tag{12}$$

where the last equality follows from $Y = \psi_Y(X_C)$, so $H(Y \mid X_C) = 0$ and $I(X_C; Y) = H(Y)$. Therefore, $I(\phi(X); Y) = I(X_C; Y)$ if and only if $H(Y \mid \phi(X), E) = 0$. □

B Details of Performance Study

B.1 Model Used for Adaptation

For FMow, iWildCam and Waterbirds, we use the model provided by Wilds (Koh et al., 2021) and studied in (Sagawa* et al., 2020). The model was trained on the training domain of the two dataset using empirical risk minimization.

- FMoW model can be downloaded from `https://worksheets.codalab.org/rest/bundles/0x20182ee424504e4a916fe88c91afd5a2/contents/blob/fmow_seed%3A0_epoch%3Abest_model.pth`.
- iWildCam model can be downloaded from `https://worksheets.codalab.org/rest/bundles/0xc006392d35404899bf248d8f3dc8a8f2/contents/blob/best_model.pth`.
- Waterbirds model can be downloaded from `https://worksheets.codalab.org/rest/bundles/0x677545cb487b4c98831e70b16ff836e7/contents/blob/logs/last_model.pth`.

To our knowledge, there is no publicly available pretrained model for HuffPost. We train a DistilBERT model using empirical risk minimization. We split a subset of the training data as the out-of-distribution validation set for model selection, following the protocol suggested in Wilds (Koh et al., 2021). Unlike Wild-Time (Yao et al., 2022) that uses data from 2012 to 2015 as the training set, data from 2012 to 2014 are taken as the training set. Data from 2015 is taken as the out-of-distribution

validation set. We follow Wild-Time that takes data from 2016 to 2018 as the out-of-distributon test set. The training procedure follows that specified in Wild-Time. The model is trained with batch size 32 for 6000 iterations. AdamW (Loshchilov & Hutter, 2019) is used as the optimiser with learning rate 2e-5 and weight decay 1e-2. Model are selected using the average accuracy on the out-of-distribution validation set.

B.2 Hyperparameter Search Space

We perform a grid search to find the best hyperparameters for the baseline methods we compared with. For all methods that require a learning rate, we search the learning rate in $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, same as what we searched for our method. The adaptation is performed in a non-episodic way. At each test batch, one step of model update is performed. The hyperparameter used is presented below.

Active TTA baselines:

- SimATTA: Following (Gui et al., 2024), we search the number of steps for the cold start phase in {0, 100}, and the low entropy threshold in $\{10^{-3}, 10^{-4}\}$. The high entropy threshold is set to 1e-2. SGD is used as the optimizer with momentum 0.9. All parameters of the model is updated.
- EATTA: Following (Wang & Ding, 2025), the margin E_0 is set to $0.4\times \ln C$, where C is the number of classes. SGD is used as the optimizer with momentum 0.9. The affine parameters of normalization layers are updated.
- BiTTA: Following (Lee et al., 2025), we search dropout rate in {0.1, 0.3}, and set the number of dropout to 4. The weight for binary-feedback guided adaptation is 2.0, and the weight for agreement-based adaptation is 1.0. SGD is used as the optimizer with momentum 0.9. All parameters of the model is updated.

Non-Active TTA baselines:

- T3A: Following (Iwasawa & Matsuo, 2021), M, the number of representations stored to compute the centroid of each class is searched in {1,5,20,50,100, N/A}, where N/A means storing all representations.
- Tent: Following (Wang et al., 2021), SGD is used as the optimizer with momentum 0.9. The affine parameters of normalization layers are updated during adaptation.
- SAR: Following (Niu et al., 2023), the margin E_0 is set to $0.4\times \ln C$, where C is the number of classes. To recover the model, the moving average factor is set to 0.9, and the reset constant is set to 0.2. SGD is used as the base optimizer with sharpness-aware minimization (SAM). The momentum for SGD is set to 0.9. ρ in SAM is set to 0.05. The affine parameters of shallow normalization layers are updated. Normalization layers in the 9^{th}-11^{th} block in the feature extractor are frozen during adaptation.
- CAFA: Following (Jung et al., 2023), SGD is used as the optimizer with momentum 0.9. The affine parameters of normalization layers are updated during adaptation.
- AdaReAlign: Following (Zhang et al., 2024), we set the number of base learners to be the ceiling integer of the log of the length of the test loader. The margin E_0 is set to $0.4\times \ln C$, where C is the number of classes. eta_base is set to 0.01, and eta_meta is set to 0.9. The weight for KL divergence is set to 0.005. SGD is used as the optimizer with momentum 0.9. The affine parameters of normalization layers are updated during adaptation.
- GOLD: Following (Lai et al., 2026), we search the AGOP EMA coefficient in {0.1, 0.02}, and golden subspace rank in {64, 256}. The confidence threshold for AGOP is set to 0.8, and the eigen-decomposition is updated for every 50 batches. The weight of self-training loss is set to 1.0, and the weight of contrastive loss is searched in {1.0, 2.0}. The momentum for the teacher's moving average update is set to 0.999. Temperature is contrastive learning is set to 0.1, The projection dimension is set to 128, and the source prototypes are computed by using a maximum of 50 training batches. Adam is used as the optimizer to update the affine parameters in the normalization layer, projector, and adapter. The adapter uses the same learning rate as others.

- DeYO: Following (Lee et al., 2024), we search over the three augmentations {patch shuffling, pixel shuffling, occlusion} to destory causal features. The patch size in patch shuffling is set to 4. For occlusion, the occlusion size is set to $(H/2) \times (W/2)$, where H and W stand for the height and width of the image. The occulsion starts from $(H/4)^{th}$ row and $(W/4)^{th}$ column. The DeYO margin is set to $0.5\times \ln C$, and the margin E_0 is set to $0.4\times \ln C$, where C is the number of classes. The PLPD threshold is searched in $\{0.2, 0.3, 0.5\}$. SGD is used as the optimizer with momentum 0.9. The affine parameters of shallow normalization layers are updated. Normalization layers in the 9^{th}-11^{th} block in the feature extractor are frozen during adaptation.
- TACT: Following (Liu et al., 2025), we search the number of augmentations in $\{2^1, 2^2, \ldots, 2^8\}$ and the number of removed principled component in $\{1, 2, 3, \cdots, 16\}$. We apply RandomAugment to image datasets.

B.3 Hardware and Software Used

We perform experiments on the NVIDIA H100 GPU with 80GB of memory. The memory of 1 GPU is sufficient to perform test-time adaptation using ACTTA as well the baseline methods when the batch size is set to 64. PyTorch 2.1.2 is used for all implementations.

B.4 Effects of Hyperparameters

Figure 3 shows the TTA performance of ACTTA under different hyperparameters. Datasets with complex distribution shift and larger scale, such as FMoW and HuffPost, benefit from a larger inner dimension of the residual adapter. On the other hand, when the distribution shift is simpler, such as in iWildCam where camera traps have a fixed background, and when the number of test samples is small, a small inner dimension works better.

The distance percentile has a strong influence on the performance. Waterbirds request that the diversity be considered more in sample selection, as the performance degrades significantly when a small number of samples of high deviation are selected. In contrast, iWildCam requests that the deviation from training representations be considered more in sample selection. The performance is best when samples of the highest deviation are considered in subsequent diversity-based selection. FMoW falls in between. The adaptation performance is the best when the distance percentile is 0.7. HuffPost is the least affected by the change in distance percentile.

In general, the adaptation benefits from including more confident samples, as Waterbirds, HuffPost, and iWildCam have better performance when the entropy percentile is high. However, FMoW does not benefit from including more confident samples. Including a moderate number of samples shows the best performance.

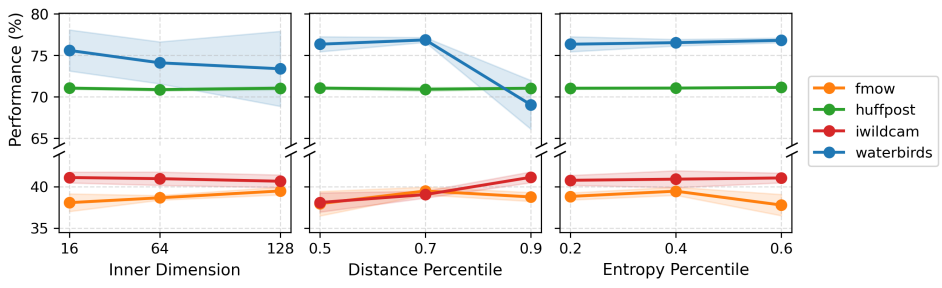


Figure 3: Effect of Hyperparameters on TTA Performance.

B.5 Effect of the Contrastive Objective.

Table 6 compares no adaptation, ACTTA without $\mathcal{L}_{\text{contrast}}$, and full ACTTA in terms of class-conditional maximum mean discrepancy (MMD) and target performance. Full ACTTA has the lowest mean MMD on all four datasets, indicating closer same-class source and target representation

Table 6: Effect of the contrastive objective.

Method	FMoW	iWildCam	HuffPost	Waterbirds
(a) Class-conditional MMD ↓				
No TTA	0.04531	0.4621	0.04290	0.03759
ACTTA w/o contrastive objective	0.04539 ± 0.00002	0.4617 ± 0.0003	0.04212 ± 0.00027	0.03767 ± 0.00089
ACTTA	**0.04521 ± 0.00002**	**0.4456 ± 0.0018**	**0.04187 ± 0.00002**	**0.03702 ± 0.00048**
Test-time adaptation performance (%) ↑				
No TTA	31.16	32.29	70.35	60.28
ACTTA w/o contrastive objective	36.05 ± 1.85	40.29 ± 1.51	71.08 ± 0.21	**76.86 ± 0.29**
ACTTA	**39.49 ± 0.38**	**40.31 ± 1.15**	**71.14 ± 0.07**	**76.86 ± 0.24**

Table 7: Query ranking criteria with diversity-based selection and ACTTA adaptation(%).

Query ranking criterion	FMoW	iWildCam	HuffPost	Waterbirds
Mahalanobis (ACTTA)	**39.49 ± 0.38**	**40.31 ± 1.15**	**71.14 ± 0.07**	**76.86 ± 0.24**
Predictive entropy	35.52 ± 0.68	38.20 ± 0.50	70.72 ± 0.10	76.06 ± 1.46
Low target kNN density	35.97 ± 3.63	37.84 ± 0.93	70.86 ± 0.14	72.79 ± 2.88

distributions. Adding the contrastive objective improves mean target performance by 3.44 percentage points on FMoW, with little or no change on the other datasets. These results support the contribution of contrastive alignment to predictive performance.

B.6 Query ranking criterion.

Table 7 compares source-relative Mahalanobis ranking with predictive entropy and low target kNN density. All variants use the same ACTTA adaptation configurations and only differ in the query ranking score to a pool of candidates for subsequent diversity-based selection. We compare ACTTA's Mahalanobis distance against entropy and test representation's nearest neighbour density. Samples of high entropy (uncertainty) and low nearest neighbour density measured by the highest Euclidean distance to nearest neighbours are selected respectively. The table below shows the results. Across four datasets, Mahalanobis distance performs the best.

B.7 Robustness to Noisy Annotations

To evaluate robustness to imperfect supervision, we assign each queried label a random label with probability 10% before using it for adaptation. Table 8 compares this setting with the original annotation setting, using the same dataset-specific metrics as Table 1. The row "Best non-active TTA" reports the best non-active baseline for each dataset from Table 1. The performances for active TTA methods do reduce when the label is noisy. However, active TTA methods still perform better than non-active TTA methods. ACTTA still performs the best on iWildCam, HuffPost, and Waterbirds, and performs the second best on FMoW.

B.8 Robustness to Delayed Annotations

To simulate delayed feedback of the labels, we evaluate ACTTA and other active TTA methods in a setting where the queried labels are only available 5 adaptation steps after the samples are selected for query. As a result, these samples can only be included in the update 5 steps after they are seen. The table below shows the result. Most of the time, the performance for ACTTA methods decreases when the feedback is delayed. However, active TTA methods still perform better than non-active TTA methods. ACTTA still performs the best on iWildCam, HuffPost, and Waterbirds, and performs the second best on FMoW, only around 0.5% lower than the best-performing active TTA method.

Table 8: Test-time adaptation performance (%) with noisy annotations. Each queried label is assigned a random label with probability 10%. Best active-method results within each annotation setting are in bold.

Method	FMoW	iWildCam	HuffPost	Waterbirds
no TTA	31.16	32.29	70.35	60.28
Best non-active TTA	32.65 ± 0.20	33.31 ± 0.69	70.69 ± 0.02	71.50 ± 0.76
Active TTA: no label noise				
SimATTA	38.28 ± 0.48	37.58 ± 0.48	70.77 ± 0.18	66.20 ± 1.90
EATTA	34.12 ± 1.01	34.57 ± 0.12	70.72 ± 0.11	59.40 ± 0.53
BiTTA	32.46 ± 1.23	35.26 ± 0.06	70.05 ± 0.17	75.08 ± 1.21
ACTTA	**39.49 ± 0.38**	**40.31 ± 1.15**	**71.14 ± 0.07**	**76.86 ± 0.24**
Active TTA: 10% label noise				
SimATTA	**37.48 ± 0.62**	34.95 ± 0.74	70.55 ± 0.03	65.42 ± 0.71
EATTA	33.15 ± 1.64	34.19 ± 0.35	70.72 ± 0.11	58.15 ± 0.26
BiTTA	30.04 ± 1.20	34.52 ± 0.90	70.05 ± 0.32	70.82 ± 1.91
ACTTA	36.28 ± 0.52	**36.52 ± 0.44**	**70.91 ± 0.09**	**73.59 ± 1.99**

Table 9: Test-time adaptation performance (%) with immediate feedback and with labels received five adaptation steps after querying. Best active-method results within each feedback setting are in bold.

Method	FMoW	iWildCam	HuffPost	Waterbirds
no TTA	31.16	32.29	70.35	60.28
Best non-active TTA	32.65 ± 0.20	33.31 ± 0.69	70.69 ± 0.02	71.50 ± 0.76
Active TTA: no label feedback delay				
SimATTA	38.28 ± 0.48	37.58 ± 0.48	70.77 ± 0.18	66.20 ± 1.90
EATTA	34.12 ± 1.01	34.57 ± 0.12	70.72 ± 0.11	59.40 ± 0.53
BiTTA	32.46 ± 1.23	35.26 ± 0.06	70.05 ± 0.17	75.08 ± 1.21
ACTTA	**39.49 ± 0.38**	**40.31 ± 1.15**	**71.14 ± 0.07**	**76.86 ± 0.24**
Active TTA: labels are received after 5 steps				
SimATTA	**38.05 ± 1.25**	36.09 ± 0.89	70.33 ± 0.28	61.63 ± 1.00
EATTA	33.26 ± 2.76	32.82 ± 0.97	70.70 ± 0.04	60.90 ± 1.71
BiTTA	27.52 ± 3.25	34.64 ± 2.18	69.31 ± 0.19	73.83 ± 1.49
ACTTA	37.51 ± 0.56	**40.39 ± 0.69**	**70.93 ± 0.20**	**75.34 ± 0.37**