Test-Time Adaptation by Causal Trimming

Yingnan Liu[1,2] **Rui Qiao**[1,3] **Mong Li Lee**[1,2] **Wynne Hsu**[1,2]
[1]School of Computing, National University of Singapore
[2]Institute of Data Science, National University of Singapore
[3]Singapore-MIT Alliance for Research and Technology
{liu.yingnan, rui.qiao}@u.nus.edu, {dcsleeml, dcshsuw}@nus.edu.sg

Abstract

Test-time adaptation aims to improve model robustness under distribution shifts by adapting models with access to unlabeled target samples. A primary cause of performance degradation under such shifts is the model's reliance on features that lack a direct causal relationship with the prediction target. We introduce Test-time Adaptation by Causal Trimming (TACT), a method that identifies and removes non-causal components from representations for test distributions. TACT applies data augmentations that preserve causal features while varying non-causal ones. By analyzing the changes in the representations using Principal Component Analysis, TACT identifies the highest variance directions associated with non-causal features. It trims the representations by removing their projections on the identified directions, and uses the trimmed representations for the predictions. During adaptation, TACT continuously tracks and refines these directions to get a better estimate of non-causal features. We theoretically analyze the effectiveness of this approach and empirically validate TACT on real-world out-of-distribution benchmarks. TACT consistently outperforms state-of-the-art methods by a significant margin. Our code is available at https://github.com/NancyQuris/TACT.

1 Introduction

Machine learning models often exhibit significant performance degradation when evaluated on data drawn from a distribution that differs from their training data distribution [13]. To address this challenge, test-time adaptation (TTA) has emerged as a promising approach. TTA methods adapt a pretrained model to the test distribution dynamically, using the incoming test data to enhance predictive performance without requiring access to the original training data [19, 51, 56]. Despite recent advances, many existing TTA methods rely heavily on predicted labels generated by the model itself to guide the adaptation process [12, 37, 38, 51]. However, the effectiveness of these methods hinges critically on the quality of the predictions. When the model's predictions are influenced by non-causal features that do not have a direct causal relationship with the prediction target [26, 54], the predicted label may be unreliable, leading to sub-optimal adaptation outcomes [29, 47].

Unlike causal features that have stable associations with the semantic structure of the prediction task [27], non-causal features exhibit inconsistent or spurious correlations with the prediction target across training and test distributions [55]. Over-reliance on non-causal features is a key factor in model performance degradation under distribution shift. While DeYO [29] recognizes this issue, it does not explicitly mitigate reliance on non-causal features. Instead, it updates the model using predictions that leverage causal features only, relying on gradual adaptation to reinforce causal features over time. Consequently, early predictions may still be influenced by non-causal signals, requiring many adaptation steps to suppress their effects.

Given the above limitations, we propose to actively reduce non-causal features. Prior studies have shown that feature representations learned through standard training encode a mixture of causal and

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

non-causal features and that the causal part is often learned sufficiently well for accurate prediction [20, 27]. Motivated by this, we propose a Test-time Adaptation by Causal Trimming (TACT) framework that seeks to improve adaptation performance by isolating and removing non-causal components from the representations of samples from test distributions. Our framework aims to achieve more reliable predictions in the presence of distribution shift by reducing the model's dependence on unstable, non-causal features. To identify non-causal features in representations, we analyze how these representations change when we apply targeted perturbations to the input data. Specifically, we perform input augmentations that preserve the underlying causal contents while introducing variability in other, non-causal aspects of the input [9, 11, 18, 31, 32]. These augmentations produce multiple test-time samples that share the same causal semantics but differ in spurious or incidental attributes. By examining how the feature representations of these samples vary, we can disentangle causal and non-causal components.

We operationalize this by applying Principal Component Analysis (PCA) to the set of augmented representations and identify the direction of greatest variance. We interpret this dominant direction as being aligned with the non-causal features, under the assumption that causal content remains stable across augmentations, while non-causal attributes vary. This approach is inspired by prior work showing that high-level semantic factors are often linearly encoded in the learned representation space [1, 39, 46]. Building on the insight that linear manipulations in representation space can produce meaningful changes in semantic content [40, 50], we propose to reduce the influence of non-causal features by subtracting the projection of a test sample's representation along the identified non-causal direction. Since the prototypes used for prediction, defined as template representations corresponding to the weights for each class in the linear classifier, are influenced by non-causal features, we apply the same operation to them using the identified non-causal direction. During adaptation, we maintain a moving average of the updated prototypes to mitigate noise effects. Compared to DeYO, TACT can immediately produce predictions that are less affected by non-causal features, eliminating the need for iterative updates to achieve reliable results (see Figure 1). We provide a theoretical analysis to establish the conditions under which TACT can improve prediction accuracy under distribution shift. Empirically, we evaluate TACT on five real-world out-of-distribution datasets, demonstrating its effectiveness and superiority over state-of-the-art TTA methods.

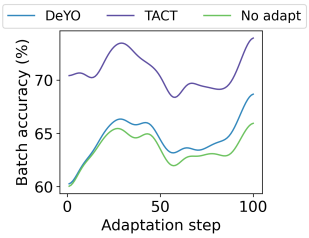


Figure 1: Batch accuracy on Camelyon17 dataset of the first 100 adaptation steps.

2 Related Work

Existing TTA methods can be broadly categorized into backpropagation-free and backpropagation-based methods. Backpropagation-free methods modify model outputs or intermediate representations without gradient-based optimization. These include modifiable prompts [36], re-normalized representations [44], updated prototypes [19, 57], and maximum likelihood estimation [4]. Backpropagation-based methods update the model with the gradient of objective functions such as entropy minimization [12, 37, 38, 51] and self-training with pseudo-labels [17, 25, 47, 52]. Entropy Minimization encourages more confident predictions by reducing the entropy of model predictions during adaptation. Self-training employs cross entropy [5, 17, 30, 47] and knowledge distillation [25, 52, 53] using model predictions as pseudo-labels. Regularization measures such as information maximization [30], representation statistics alignment [23, 58], and consistency regularization [35, 56] for invariant prediction under augmentations have been proposed to regularize the adaptation.

A key challenge in test-time adaptation is obtaining reliable pseudo-labels to guide model updates. One line of work assumes that correct predictions tend to exhibit low entropy, and update the model using only high-confidence samples with low-entropy predictions [19, 37, 38, 57]. However, DeYO [29] shows that spurious correlations can also result in low entropy predictions and proposes a causal intervention technique to identify predictions that are more likely based on causal features, using them selectively for model updates. Another line of work refines pseudo-labels by incorporating updated prototype and neighborhood information [5, 17, 21, 47, 53]. AdaContrast [5] uses soft voting among nearest neighbors. TSD [53] relies on updated prototypes and spatial local clustering. TAST [21] employs neighbourhood information in self-training. PROGRAM [47] considers both prototype

and neighbour-based pseudo-labels to enhance label quality. PASLE [17] progressively refines the pseudo-labels of uncertain predictions using updated prototypes.

All the above methods, except for DeYO, do not consider the effect of non-causal features on model prediction. Although DeYO finds that non-causal features would make entropy an unreliable metric to reflect prediction correctness, it does not adjust model predictions. TACT adjusts model predictions by reducing non-causal features, and our adjusted prediction can be used as a more reliable pseudo-label.

3 Preliminaries

We consider the problem of adapting a well-trained model to test-time distributions that differ from the training distribution. Our goal is to improve the model's performance on these shifted distributions with unlabeled test samples. Following prior work [18, 48], we model the distribution shift using a structural causal model that captures the underlying data-generating process, as illustrated in Figure 2.

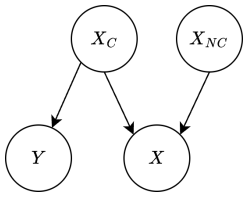


Figure 2: Structural causal model of the data-generating process.

We model the observed sample X and its label Y as being generated from causal factors X_C and non-causal factors X_{NC}. Only X_C is causally related to Y, while X is related to both X_C and X_{NC}. The correlation between X_C and Y is stable, i.e., the conditional distribution $P(Y|X_C)$ remains unchanged at test time. We also assume that the distribution of causal factors $P(X_C)$ remains invariant across the training and test datasets, whereas distribution shifts arise from changes in $P(X_{NC})$. The model would have stable performance across distributions if the prediction is based solely on features related to the causal factors X_C. In contrast, reliance on non-causal factors X_{NC} can lead to unreliable performance under distribution shifts.

We consider a c-class classification task, where the model $f := h \circ g$ used for adaptation is composed of a feature extractor g and a classifier h. The feature extractor g maps an input sample to a d-dimensional vector $z \in \mathbb{R}^d$ as the representation. The classifier h maintains a set of class prototypes $\{q_1, \ldots, q_c\} \in \mathbb{R}^d$, where each prototype q_i serves as a template representation for class i. Predictions are made by computing the similarity between the input representation z and each class prototype using the dot product $z \cdot q_i$, referred to as the logit of class i. A softmax function is then applied to the logits to obtain the probability distribution over the classes.

4 Proposed Method

The works in [20, 27] show that models are often capable of learning causal features, even when their predictions are predominantly driven by non-causal features with spurious correlations. However, the predictive influence of these causal features is frequently obscured or suppressed due to the heavily weighted non-causal components in the learned representations. Based on this observation, we propose TACT, a method that identifies non-causal features and reduces their influence by causally trimming the learned representations. We hypothesize that non-causal features are embedded in representations along a specific direction. Such direction is of the maximum variance when the non-causal features change. To suppress their influence, we subtract the projection of both the input representation and class prototypes onto this identified direction. This operation attenuates the non-causal information present in both elements. Since class prototypes serve as canonical representations for each class, and the non-causal direction estimated from a single test sample may be noisy, we maintain a moving average of the updated prototypes throughout test-time adaptation. At inference, predictions are made by measuring the similarity between the adapted representation and the moving average of the updated prototypes, thereby reducing influence of non-causal features.

4.1 Non-Causal Feature Identification

Given a sample x at test time, if we have access to additional samples generated with the same causal factors but different non-causal factors, we can compare their representations to infer the influence of non-causal factors. Changes in the representations across these samples can be attributed to variations

in these non-causal factors. By systematically analyzing these representational differences, we can isolate and identify the components of the representation that correspond to non-causal features.

To simulate variations in the data-generating process, we apply data augmentation to target non-causal features [9, 11, 18, 31, 32]. For a test sample x, we generate n augmented samples $\{\tilde{x}_i\}_{i=1}^n$ that preserve the causal feature while varying the non-causal factors. We collect the representations of these samples in a matrix $\mathbf{Z} = [z, \tilde{z}_1, ..., \tilde{z}_n]^\top$, where z is the representation of the original sample and $\tilde{z}_i$ are those of the augmented samples.

We interpret non-causal features as corresponding to specific, disentangled directions in the representation space, consistent with prior work that indicates high-level semantic concepts are linearly encoded as vector directions in learned representations [1, 39, 46]. For instance, the vector difference between "woman" and "man" would resemble that between "queen" and "king" [33], both aligning to the direction representing gender. Along this direction, specific instances of the gender concept, such as "male" and "female", take different magnitudes.

Given representations of samples that differ only in their non-causal factors, the direction along which the representations change the most is expected to capture the non-causal features. This dominant direction can be identified via Principal Component Analysis (PCA) which analyzes the covariance matrix of the representations to extract the principal components. Principal components are vectors along which the representations' projections exhibit maximum variances. We first compute the mean of the representations as: $\bar{z} = \frac{1}{n+1} z + \frac{n}{n+1} \sum_{i=1}^n \tilde{z}_i$. Using this mean, we construct a matrix $\bar{\mathbf{Z}} = [\bar{z}, \bar{z}, .., \bar{z}]^\top$ that has the same size as the representation matrix $\mathbf{Z}$. The covariance matrix of the representations is then given by $\mathbf{\Sigma_Z} = (\mathbf{Z} - \bar{\mathbf{Z}})^\top (\mathbf{Z} - \bar{\mathbf{Z}})$. The eigenvectors of $\mathbf{\Sigma_Z}$ correspond to the principal components, and their eigenvalues quantify the variance along these components [22]. Since $\mathbf{\Sigma_Z}$ is a real symmetric matrix, its eigenvectors form an orthonormal basis in $\mathbb{R}^d$ [16]. Using spectral decomposition, we express the covariance matrix as: $\mathbf{\Sigma_Z} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, where $\mathbf{Q} = [e_1, e_2, ..., e_d]$ is an orthogonal matrix whose columns are the orthonormal eigenvectors, and $\mathbf{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\mathbf{\Sigma_Z}$. Here, e_i denotes the direction along which the variance of the projected representations is the i^{th} largest.

4.2 Causal Trimming to Reduce Non-Causal Feature

Prior work has demonstrated that applying linear transformations to representations can manipulate the semantics they encode [40, 50]. Since the principal components $\{e_i\}_{i=1}^d$ form an orthonormal basis in $\mathbb{R}^d$, any representation z can be expressed as a linear combination of these components. To reduce the influence of non-causal features, we propose to trim the representation by removing its components along the top-m principal components:

$$\hat{z} = z - \sum_{i=1}^{m} (z \cdot e_i) e_i \tag{1}$$

Since each e_i is a vector of unit length, the term $(z \cdot e_i)e_i$ is the projection of z onto e_i whose magnitude is given by the dot product $(z \cdot e_i)$. By subtracting these terms, we obtain an updated representation $\hat{z}$ which is composed of components only formed by $\{e_i\}_{i=m+1}^d$. If causal features are invariant under data augmentations and their corresponding semantic directions are orthogonal to those of the removed directions, causal information present in z is preserved in the trimmed representation $\hat{z}$.

4.3 Model Adaptation

In a prototype-based classifier, each class prototype q_j serves as a template representation learned by the classifier h, summarizing the representations of samples belonging to class j. However, if the learned representations encode non-causal features, the prototypes will be influenced by these features. To mitigate this issue, we apply the same causal trimming to the class prototypes. Specifically, let q_j be the prototype of class j. Given the top-m principal components $\{e_i\}_{i=1}^m$ that are used to trim the test sample representation z, we obtain the trimmed prototype $\hat{q}_j$ for each class $j \in \{1, 2, \ldots, c\}$ as:

$$\hat{q}_j = q_j - \sum_{i=1}^{m} (q_j \cdot e_i) e_i \tag{2}$$

Since the identified non-causal directions may vary across samples due to noise or context-specific factors, we compute a batch-wise average of trimmed prototypes to obtain a more stable estimate. To track the estimate across batches during adaptation, we maintain a moving average of the trimmed prototypes during test-time adaptation. Suppose we obtain trimmed prototypes $\hat{q}_j^{(i)}$ for each class j at batch i, the moving-average $\bar{\hat{q}}_j$ is updated by $\bar{\hat{q}}_j = \frac{i-1}{i}\bar{\hat{q}}_j + \frac{1}{i}\hat{q}_j^{(i)}$. This moving average serves as a more robust estimate of the causally refined prototypes, effectively smoothing out sample-specific variance. The prediction made by the average of the trimmed prototypes is the same as that of an ensemble over the logits produced by individual trimmed prototypes, resulting in more stable predictions. At test time, for a given input sample x, we compute the causally trimmed representation $\hat{z}$ and compare it to the moving-averaged trimmed prototypes $\bar{\hat{q}}_j$. The logit for class j is given by the dot product $\hat{z} \cdot \bar{\hat{q}}_j$, and the final predicted label y is given by:

$$y = \arg\max_j \frac{\exp\left(\hat{z} \cdot \bar{\hat{q}}_j\right)}{\sum_{i=1}^{c} \exp\left(\hat{z} \cdot \bar{\hat{q}}_i\right)} \tag{3}$$

5 Theoretical Analysis

We present the conditions under which TACT would correct a wrong prediction and maintain a correct prediction. We consider binary classification $Y \in \{+1, -1\}$. The two prototypes learned by the binary classifier h are represented as $\{q_{+1}, q_{-1}\}$. We drop the bias term for simplicity. Meanwhile, we assume the existence of causal prototypes $\{p_{+1}, p_{-1}\}$, which always make correct predictions on the learned representations and do not leverage non-causal features. To simplify the analysis, we consider the decision boundary vectors $\Delta q = q_{+1} - q_{-1}$ and $\Delta p = p_{+1} - p_{-1}$. We analyze the representation z of an instance with label y. Given the principal components $\{e_i\}_{i=1}^{d}$ computed from z and its augmented variants, we write z as $\sum_{i=1}^{d} \alpha_i e_i$, where α_i is the magnitude of z's projection on e_i. Similarly, we define the learned decision boundary Δq's projection magnitude as $\{\gamma_i\}_{i=1}^{d}$. We write the projection magnitude of causal decision boundary Δp as $\{\eta_i \gamma_i\}_{i=1}^{d}$, to view Δp as a transformation from Δq by a projection magnitude η_i on the direction of each principal component.

We can obtain $\hat{z}$ by trimming the top-m principal components (PCs) for z. Proposition 1 shows the conditions under which TACT can correct a wrong prediction.

Proposition 1. *For any z that is misclassified by the learned decision boundary Δq, the misclassification can be corrected by using the representation obtained after removing the top-m principal components, if both of the following two conditions are satisfied:*

$$y \sum_{i=1}^{m} \alpha_i \gamma_i < 0 \quad \textit{and} \quad y \sum_{i=m+1}^{d} \alpha_i \gamma_i > 0 \tag{4}$$

$$\left| \sum_{i=1}^{m} \alpha_i \gamma_i \right| > \left| \sum_{i=m+1}^{d} \alpha_i \gamma_i \right| \tag{5}$$

Appendix A.1 provides the formal proof. Equation (4) captures the case in which a prediction based solely on the top-m PCs leads to an incorrect outcome, whereas a prediction based on the remaining PCs yields the correct result. Equation (5) requires the absolute value of the prediction score derived from the top-m PCs must be greater than that from the remaining PCs. Together, these conditions in Proposition 1 suggests that a wrong prediction can be corrected by TACT when the top-m PCs are solely responsible for the wrong prediction, and the prediction made by the top-m PCs weighs more than the prediction made by the remaining PCs.

In Proposition 2, we establish the conditions under which the trimmed representation $\hat{z}$ retains sufficient causal information to preserve the correct prediction by the causal decision boundary Δp.

Proposition 2 (Causal Preservation)**.** *For any original representation z, the trimmed representation $\hat{z}$ preserves the correct prediction under the causal decision boundary Δp if any one of the following*

conditions holds:

$$\begin{cases} y \sum_{i=1}^{m} \eta_i \alpha_i \gamma_i = 0 \\ y \sum_{i=1}^{m} \eta_i \alpha_i \gamma_i < 0 \\ 0 < y \sum_{i=1}^{m} \eta_i \alpha_i \gamma_i < y \sum_{i=1}^{d} \eta_i \alpha_i \gamma_i \end{cases} \tag{6}$$

The proof is provided in Appendix A.2. Equation (6) characterizes three cases: (a) the top-m PCs have no contribution to the causal prediction; (b) the top-m PCs has a negative influence on the causal prediction and thus their removal is beneficial; (c) the top-m PCs has a positive contribution, but the representation forms by all PCs contribute even more strongly. When the top-m PCs have no contribution to the causal predictions, they are considered non-causal features. In other words, the removed component $z - \hat{z}$ does not contain causal information. When the top-m PCs contain causal information, m should be selected such that the top-m PCs contribute less to the prediction compared to all the PCs, ensuring that the trimmed representation $\hat{z}$ remains causally informative. In other words, sufficient causal features need to be preserved after causal trimming.

Finally, in Proposition 3, we identify the conditions under which causal trimming would have no negative impact on the prediction of samples that are already correctly classified.

Proposition 3. *Suppose z is correctly classified by the learned decision boundary Δq. The trimmed representation $\hat{z}$ obtained via TACT will still be classified correctly if either of the conditions holds:*

1. $y(z - \hat{z})\Delta q \leq 0$*, or*
2. $y(z - \hat{z})\Delta q > 0$*, and Equation* (7) *holds, assuming $\hat{z}$ already satisfies the Causal Preservation condition (Proposition 2).*

$$\operatorname{sign}\left(\sum_{i=m+1}^{d} \eta_i \alpha_i \gamma_i\right) = \operatorname{sign}\left(\sum_{i=m+1}^{d} \alpha_i \gamma_i\right) \tag{7}$$

The proof can be found in Appendix A.3. Equation (7) indicates that when classification relies only on the representations formed by the remaining PCs, the learned decision boundary makes the same prediction as the causal decision boundary. Proposition 3 also shows that if a correct prediction is made by the learned decision boundary, TACT will preserve this correctness as long as the removed part $z - \hat{z}$ contributes negatively or does not contribute to the prediction. On the other hand, when the trimmed representation $\hat{z}$ contains sufficient causal information as established in Proposition 2, the learned decision boundary is required to align directionally with the causal decision boundary defined by the remaining PCs.

6 Performance Study

We study the test-time adaptation performance under real-world distribution shifts, using datasets from multiple modalities, including image, audio, and text. Compared to prior works that primarily benchmark on image data, our comprehensive experiments offer broader insights into the generalizability of TACT and other TTA methods.

Datasets. We summarize the datasets used in our experiments below:

- Birdcalls [15, 24, 34], curated by [9], is an audio classification dataset to identify bird species from clips recorded in diverse environments. Each clip is converted into a Mel spectrogram for classification. Distribution shifts stem from variations in microphone gain settings, habitat acoustics (e.g. other animal sounds), and bird population. The test set includes 724 audio clips.
- Camelyon17 [2], sourced from from the Wilds benchmark [28], is a medical imaging dataset for binary classification of tumor versus normal tissue images. The distribution shift arises from variations in slide staining protocols, patient demographics, and scanner equipment. The test set consists of 85,054 images.
- CivilComments [3], from the Wilds benchmark [28], is a natural language dataset comprising user-submitted text comments. The task is to classify whether a comment is toxic or non-toxic.

The toxicity is spuriously associated with the mention of certain demographics in the training data. The test set contains 133,782 comments.
- ImageNet-R [14] contains 30,000 images of objects from 200 ImageNet [42] classes. The images consist of various renditions, resulting in visual domain shifts from the original dataset.
- ImageNet-V2 [41] is collected years after the original ImageNet using the same methodology, and includes 10,000 images across 1,000 original classes. It represents a natural temporal shift.

Non-causal feature identification for TACT. We applied the following data augmentations to identify non-causal features in each dataset: For Birdcalls, we follow [9] that investigates augmentations that randomize features independent of labels but dependent on distributions. Here, random color jitter is applied to the Mel spectrograms to simulate changes in microphone gain settings. For Camelyon17, we use stain color jitter [49] as suggested in [9] to mimics variations in histopathological slide staining. For CivilComments, we randomly prepend or append short demographic-referencing sentences to the original text. The full list of sentences is provided in Appendix B. For ImageNet-R and ImageNet-V2, where the sources of distribution shift are unknown, we experiment with general-purpose image augmentations. Specifically, we apply AutoAugment [6] with ImageNet policy and RandomAugment [7]. Both methods apply a series of transformations to the images. A detailed discussion on augmentation design and selection in practice is presented in Appendix C.

Baselines. Since TACT is a backpropagation-free approach, we compare TACT with the following state-of-the-art (SOTA) TTA backpropagation-free algorithms:

- T3A [19] adapts the classifier by updating class prototypes using confident test-time representations.
- LAME [4] adjusts model output probabilities via Laplacian-adjusted maximum likelihood estimation.
- FOA [36] introduces an adaptable prompt at model input to match the representation statistics of test and train data.

We also implement a variant called TACT-adapt, where predictions from TACT are used to guide gradient-based model updates with cross entropy loss $\mathcal{L}_{CE}$. We employ the information maximization loss $\mathcal{L}_{IM}$ proposed in SHOT [30] as regularization. We optimize the model using the objective: $\mathcal{L} = \mathcal{L}_{CE}\left(\hat{y}, y_{\text{TACT}}\right) + \lambda \mathcal{L}_{IM}\left(\hat{y}\right)$. $\hat{y}$ is the model's prediction, and y_{TACT} is TACT's prediction. λ is the hyperparameter balancing the two terms.

We compare TACT-adapt with the following SOTA backpropagation-based methods:

- SHOT [30] adapts the feature extractor using information maximization and cross entropy loss on confident prediction.
- Tent [51] performs entropy minimization to update the affine parameters of normalization layers at test time.
- SAR [38] builds upon Tent by incorporating sharpness-aware minimization and model reset to mitigate overfitting to noisy samples.
- DeYO [29] identify confident samples that leverage causal features only by image augmentations that destroy shapes and using confidence-reweighted entropy minimization to update the affine parameters.
- TAST [21] adapts a trainable module on top of the trained feature extractor via self-training with nearest neighbor information.
- TSD [53] enhances feature representations through self-distillation and local clustering, ensuring alignment and uniformity while filtering noisy labels.
- PASLE [17] refines uncertain pseudo-labels progressively using selective label enhancement with candidate label sets and classifier-consistent loss.

Model architecture. We study TACT on transformer-based architectures, which are increasingly used in practice but remain relatively underexplored in TTA. Specifically, we use ViT-B/32 [8] as the backbone for Birdcalls, Camelyon17, ImageNet-R, and ImageNet-V2, and DistilBERT [43] for CivilComments. Appendix D.1 provides more details on model studied.

Hyperparameters and model selection. We use a test batch size of 64 [29, 36]. There are two hyperparameters in TACT, the number of augmentation n and the number of removed principal components m. We search $n \in \{2^1, 2^2, \ldots, 2^8\}$, $m \in [1, 16]$ and m is an integer. For TACT-adapt, we search $\lambda \in \{1, 5\} \times \{0.1, 1, 10, 100\}$. The rest hyperparameters follow the search space

Table 1: Test-time adaptation performance (%). We group the methods into backpropagation-free (BP-free) and backpropagation-based (BP-based). The best performance of each dataset is in bold.

	Method	Birdcalls	Camelyon17	CivilComments	ImageNet-R	ImageNet-V2
	No TTA	22.74	62.31	55.38	41.83	62.97
BP-free	T3A	26.16±1.33	69.96±1.98	56.43±0.00	41.78±0.12	62.93±0.02
	LAME	23.66±1.01	62.38±0.03	56.24±0.10	41.77±0.01	63.00±0.02
	FOA	26.95±1.81	58.36±0.77	-	41.46±0.16	62.76±0.08
	TACT	31.14±1.69	70.17±0.05	71.80±0.35	43.59±0.02	63.33±0.10
BP-based	SHOT	26.82±5.14	80.28±5.61	13.93±0.97	48.79±0.08	63.32±0.09
	Tent	23.16±0.42	62.29±0.01	55.38±0.00	42.08±0.05	63.09±0.03
	SAR	23.16±0.42	62.30±0.00	55.38±0.00	42.58±0.11	62.97±0.01
	DeYO	23.29±0.39	69.64±1.47	-	46.87±0.08	62.96±0.01
	TAST	26.08±1.11	83.01±1.42	56.56±0.20	41.09±0.08	62.84±0.07
	TSD	27.33±1.75	67.33±4.74	55.38±0.00	41.76±0.01	62.98±0.01
	PASLE	27.35±1.79	60.66±0.04	55.77±0.15	46.08±0.09	63.15±0.04
	TACT-adapt	**31.25±3.59**	**83.70±1.10**	**71.98±0.19**	**48.81±0.05**	**63.44±0.07**

of SHOT. For all baseline methods, we perform hyperparameter tuning within the search spaces specified in their respective papers. The detailed configurations and search procedures are provided in Appendix D.2. Following the protocol recommended in [59], we employ oracle selection to choose the best-performing hyperparameters, ensuring a fair and consistent evaluation across all methods.

6.1 Test-time Adaptation Performance

Following the evaluation protocol of each dataset, we use macro F1 for Birdcalls, accuracy for Camelyon17, ImageNet-R and ImageNet-V2, and worst-group accuracy for CivilComment, whose data are grouped by demographic attributes and toxicity. Due to the high variability observed in Birdcalls, each experiment is repeated ten times, whereas experiments on the remaining datasets are conducted three times. The mean and standard deviation are summarized in Table 1.

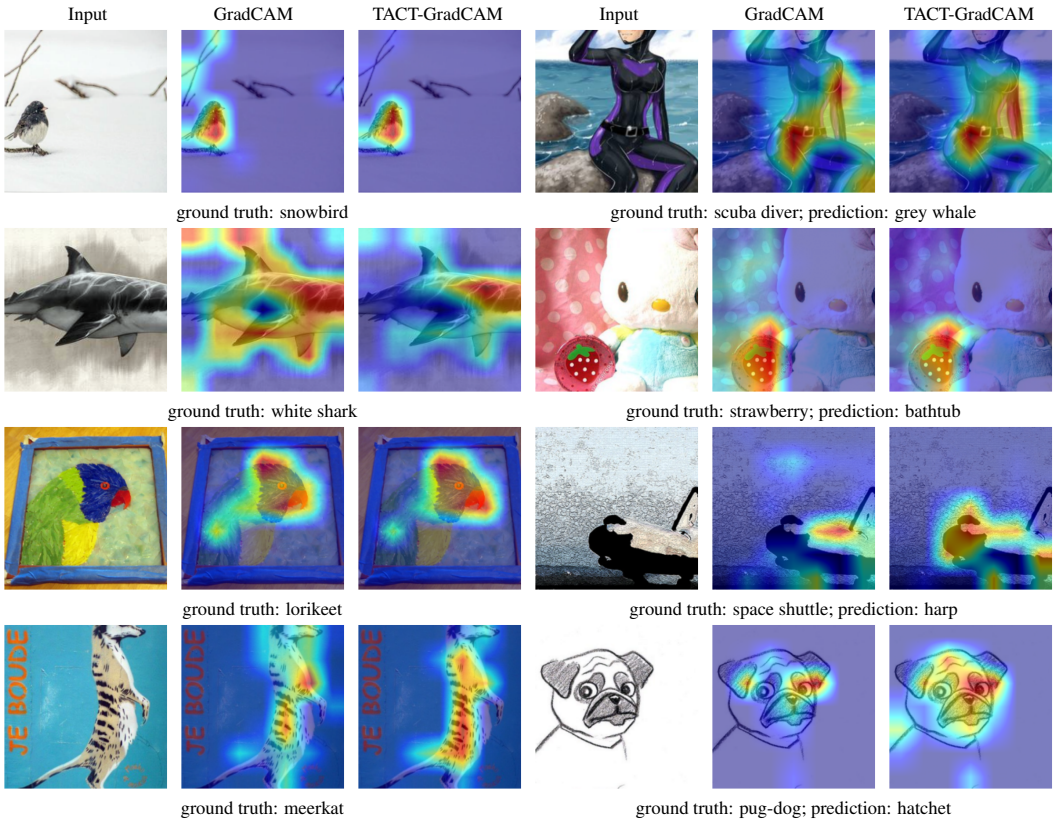
We see that TACT consistently outperforms existing backpropagation-free methods on all the datasets, with substantial gains of 4% on Birdcalls, 15% on CivilComments, and 1.7% on ImageNet-R. Further, TACT-adapt achieves the best overall performance across all datasets, outperforming both backpropagation-free and backpropagation-based baselines. These results suggest that non-causal features are a major source of performance degradation under distribution shift, and that removing them improves predictive reliability. It also confirms the value of TACT not only as a standalone method but also as a reliable supervisory signal for test-time learning.

We note that TACT performs well when causal features are approximately invariant under augmentation. For ImageNet-R and ImageNet-V2, AutoAugment [6] and RandomAugment [7] maintain the key causal features, which are object structure and shape [10, 29]. Other causal features that could be helpful in inferring objects, such as color when inferring strawberries, are altered. In addition, the models we perform adaptation on do not have their representation space explicitly constrained such that causal and non-causal features are linearly encoded, disentangled, or orthogonal. Yet, the approximate separation of causal and non-causal features by PCA yields consistent performance gains, suggesting the robustness of TACT.

6.2 Visualization of Predictions after Causal Trimming

To gain insight into the predictions made after causal trimming, we employ GradCAM [45] to visualize the focus of the original predictions and those made by TACT on samples from ImageNet-R. GradCAM identifies which parts of an input image contribute most to a prediction by computing the gradients of the predicted class score with respect to the embeddings of the image patches. The resulting heatmaps are overlaid on the input images, where brighter regions indicate higher importance for the prediction.

The visualization results are presented in Figure 3. Compared to the original predictions, TACT places less emphasis on non-causal information, such as background elements. For instance, in the



(a) correct predictions (b) wrong predictions corrected by TACT

Figure 3: GradCAM visualizations of the original predictions and TACT's predictions.

snowbird sample, TACT disregards irrelevant features like the surrounding branches. Similarly, in the white shark example, TACT restricts the focus to the object itself, unlike the original prediction that diffuses significant attention across the background. The sea background in the scuba diver example, and the dot texture in the background of the strawberry example, are likely to be spuriously correlated with certain prediction classes. These features are de-emphasized by TACT, contributing to a more accurate prediction.

Furthermore, TACT enhances attention on core causal features, leading to a sharper focus on an object's defining characteristics. This is clearly demonstrated in the lorikeet example, where the beak becomes the key focus, and the meerkat example, where attention is concentrated on the banded pattern and body. Moreover, in cases where the original prediction neglects causal features, as shown in the space shuttle and pug-dog example, TACT can redirect the emphasis to the actual salient features, such as the nose cone of the space shuttle and the face of the pug-dog, resulting in improved prediction performance.

6.3 Effect of Hyperparameters

The performance of TACT depends on two key hyperparameters: the number of augmentations n and the number of removed principal components m. These parameters govern the accuracy of non-causal direction estimation and the extent of causal trimming, respectively. Figure 4 shows the performance under different numbers of augmentations and removed principal components for the Camelyon17, CivilComments and ImageNet-R datasets.

Since representations from augmented samples are used to compute the covariance matrix from which the directions of maximum variances are identified, a larger number of augmentations n generally leads to more stable and accurate identification of non-causal directions. Empirically, we find that values of $n \in \{128, 256, 512\}$ provides sufficient performance, while small values of n often fail to

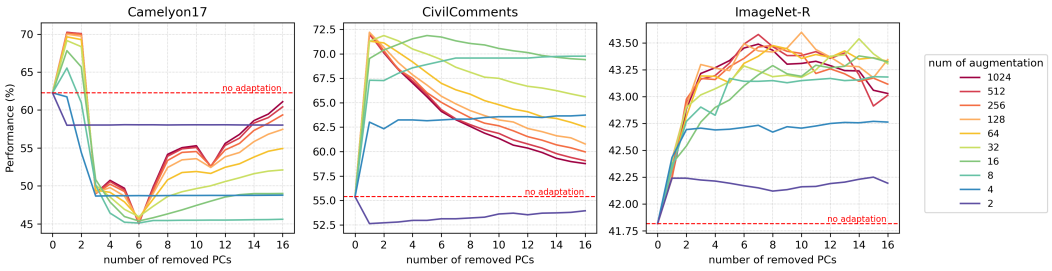


Figure 4: Performance across number of augmentation and number of removed principal components.

adequately capture the variance needed for accurate principal component estimation. The number of removed principal components m, should be carefully selected to ensure effective reduction of non-causal features while retaining sufficient causal features as suggested by our theoretical analysis. In practice, removing the top principal component, which typically captures the dominant non-causal variation, often suffices. However, for datasets with more complex or layered distribution shifts, such as ImageNet-R, removing more principal components would further boost the performance.

6.4 Ablation Study

We conduct two sets of ablation experiments. In the first experiment, we isolate the impact of representation trimming, without prototype averaging. In the second experiment, we assess whether prototype averaging alone can sufficiently filter out non-causal features.

Table 2 shows the results. We observe that trimming the representation z yields better performance than no adaptation, confirming that removing components aligned with non-causal directions in representations is beneficial. While using only the averaged trimmed prototypes $\hat{q}$ also improves performance over no adaptation, the gains are generally less significant than when trimmed representations are employed. This suggests that relying solely on the averaged trimmed prototypes is insufficient for effectively reducing non-causal features. The best performance is achieved when both the trimmed representation and the averaged trimmed prototypes are used in conjunction, indicating that mitigating non-causal features in both representations and prototypes is crucial.

Table 2: Results of ablation study.

trim z	trim q	average $\hat{q}$	Birdcalls	Camelyon17	CivilComments	ImageNet-R	ImageNet-V2
No TTA			22.74	62.31	55.38	41.83	62.97
✓			25.91±1.67	69.43±0.01	67.84±0.37	43.21±0.03	63.24±0.10
	✓	✓	27.36±0.23	64.74±0.05	62.41±0.08	42.24±0.00	63.03±0.01
✓	✓	✓	**31.14±1.69**	**70.17±0.05**	**71.80±0.35**	**43.59±0.02**	**63.33±0.10**

7 Conclusion and Future Work

We present TACT, a test-time adaptation method that reduces model reliance on non-causal features for test representations. TACT identifies non-causal components in the representation space by analyzing samples with identical causal features but varying non-causal features. The directions of maximum variance among the representations are treated as the non-causal directions. To adapt the model, we subtract the projection of the representation and class prototypes onto this non-causal direction. We keep track of the identified directions and utilize the average of the trimmed class prototypes for improved prediction. We analyze the theoretical conditions for TACT to enhance predictive performance. Extensive experiments on five real-world out-of-distribution datasets demonstrate the effectiveness and generalizability of our approach. While TACT demonstrates strong performance, it requires prior knowledge of the data to select augmentations that vary non-causal features without altering causal ones. Future work should explore identifying non-causal features when such knowledge is unavailable, and better methods to find non-causal features beyond PCA's orthogonality constraint.

Acknowledgement

We thank Dr Fusheng Liu for the helpful discussions. We appreciate the anonymous reviewers and AC for the constructive and valuable feedback.

References

[1] Sanjeev Arora, Yuanzhi Li, Yingyu Liang, Tengyu Ma, and Andrej Risteski. A latent variable model approach to pmi-based word embeddings. *Transactions of the Association for Computational Linguistics*, 4:385–399, 2016.

[2] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE Transactions on Medical Imaging*, 2018.

[3] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, 2019.

[4] Malik Boudiaf, Romain Mueller, Ismail Ben Ayed, and Luca Bertinetto. Parameter-free online test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8344–8353, 2022.

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

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019.

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

[8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[9] Irena Gao, Shiori Sagawa, Pang Wei Koh, Tatsunori Hashimoto, and Percy Liang. Out-of-domain robustness via targeted augmentations. In *International Conference on Machine Learning*, pages 10800–10834. PMLR, 2023.

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, 2019.

[11] Karan Goel, Albert Gu, Yixuan Li, and Christopher Re. Model patching: Closing the subgroup performance gap with data augmentation. In *International Conference on Learning Representations*, 2021.

[12] 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.

[13] Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *International Conference on Learning Representations*, 2021.

[14] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021.

[15] W. Alexander Hopping, Stefan Kahl, and Holger Klinck. A collection of fully-annotated soundscape recordings from the southwestern amazon basin. URL https://zenodo.org/records/7079124, 2022.

[16] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

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

[18] 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.

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

[20] 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.

[21] Minguk Jang, Sae-Young Chung, and Hye Won Chung. Test-time adaptation via self-training with nearest neighbor information. In *International Conference on Learning Representations*, 2023.

[22] I.T. Jolliffe. *Principal Component Analysis*. Springer Series in Statistics. Springer, 2002.

[23] 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*, pages 19060–19071, 2023.

[24] Stefan Kahl, Russell Charif, and Holger Klinck. A collection of fully-annotated soundscape recordings from the northeastern united states. URL https://zenodo.org/records/7018484, 2022.

[25] 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*, pages 15737–15752. PMLR, 2023.

[26] 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.

[27] 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.

[28] 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*, pages 5637–5664. PMLR, 2021.

[29] 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.

[30] Jian Liang, Dapeng Hu, and Jiashi Feng. Do we really need to access the source data? Source hypothesis transfer for unsupervised domain adaptation. In *International Conference on Machine Learning*, pages 6028–6039. PMLR, 2020.

[31] 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*, pages 8046–8056, 2022.

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

[33] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[34] Amanda Navine, Stefan Kahl, Ann Tanimoto-Johnson, Holger Klinck, and Patrick Hart. A collection of fully-annotated soundscape recordings from the island of hawai'i. URL https://doi.org/10.5281/zenodo.7078499, 2022.

[35] 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*, pages 24162–24171, June 2023.

[36] 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.

[37] 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*, pages 16888–16905. PMLR, 2022.

[38] 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.

[39] Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *International Conference on Machine Learning*, pages 39643–39666. PMLR, 2024.

[40] 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 *International Conference on Machine Learning*, pages 8748–8763. PmLR, 2021.

[41] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR, 2019.

[42] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[43] 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.

[44] Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems*, volume 33, pages 11539–11551, 2020.

[45] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[46] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020.

[47] 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.

[48] 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, pages 16846–16859, 2021.

[49] David Tellez, Maschenka Balkenhol, Irene Otte-Höller, Rob Van De Loo, Rob Vogels, Peter Bult, Carla Wauters, Willem Vreuls, Suzanne Mol, Nico Karssemeijer, et al. Whole-slide mitosis detection in h&e breast histology using phh3 as a reference to train distilled stain-invariant convolutional networks. *IEEE transactions on medical imaging*, 37(9):2126–2136, 2018.

[50] Paul Upchurch, Jacob Gardner, Geoff Pleiss, Robert Pless, Noah Snavely, Kavita Bala, and Kilian Weinberger. Deep feature interpolation for image content changes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7064–7073, 2017.

[51] 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.

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

[53] 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*, pages 20050–20060, 2023.

[54] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre-Alvise Rebuffi, Ira Ktena, Krishnamurthy Dj Dvijotham, and Ali Taylan Cemgil. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*, 2022.

[55] Nanyang Ye, Kaican Li, Haoyue Bai, Runpeng Yu, Lanqing Hong, Fengwei Zhou, Zhenguo Li, and Jun Zhu. Ood-bench: Quantifying and understanding two dimensions of out-of-distribution generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7947–7958, 2022.

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

[57] 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*, pages 41647–41676. PMLR, 2023.

[58] 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*, volume 37, pages 94607–94632, 2024.

[59] Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *International Conference on Machine Learning*, pages 42058–42080. PMLR, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly state the claims and contributions in the abstract and introduction, which are further justified in Sections 4, 5, and 6.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: we discuss the limitations in Section 7.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We discuss the assumptions of the Propositions in Section 5, and provide the proof in Appendix A.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided experimental configurations in Section 6 and Appendix D for reproducibility.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide the code and the used data in the supplemental materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have provided experimental configurations in Section 6 and Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have included error bars for both the main experiments and the ablation study in Section 6.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the computer resources used in the Appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper uses existing public datasets and develops a method trained on these datasets.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have cited the datasets in Section 6 and baseline related works in Section 2 and Section 6.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide the code and datasets used in the supplemental materials.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components, except that we use ChatGPT to generate the augmentations for the textual data in CivilComments dataset, as described in Appendix B.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Details of Theoretical Analysis

A.1 Conditions for TACT to correct a wrong prediction

We first restate Proposition 1 as follows:

Proposition 1. *For any z that is misclassified by the learned decision boundary Δq, the misclassification can be corrected by using the representation obtained after removing the top-m principal components, if both of the following two conditions are satisfied:*

$$y\sum_{i=1}^{m}\alpha_i\gamma_i < 0 \quad \textit{and} \quad y\sum_{i=m+1}^{d}\alpha_i\gamma_i > 0 \tag{4}$$

$$\left|\sum_{i=1}^{m}\alpha_i\gamma_i\right| > \left|\sum_{i=m+1}^{d}\alpha_i\gamma_i\right| \tag{5}$$

Proof. As the learned decision boundary Δq cannot classify z correctly, we have:

$$\begin{aligned} yz\cdot\Delta q &< 0 \\ y\sum_{i=1}^{d}\alpha_i e_i\cdot\sum_{i=1}^{d}\gamma_i e_i &< 0 \\ y\sum_{i=1}^{d}\alpha_i\gamma_i(e_i\cdot e_i) &< 0 \\ y\sum_{i=1}^{d}\alpha_i\gamma_i &< 0 \\ y\sum_{i=1}^{m}\alpha_i\gamma_i + y\sum_{i=m+1}^{d}\alpha_i\gamma_i &< 0 \end{aligned} \tag{8}$$

TACT updates z to $\hat{z}$ and q to $\hat{q}$ via causal trimming, and the resulting prediction is correct if and only if $y\hat{z}\cdot\Delta\hat{q} > 0$, which leads to:

$$\begin{aligned} y\hat{z}\cdot\Delta\hat{q} &> 0 \\ y\sum_{i=m+1}^{d}\alpha_i e_i\cdot\sum_{i=m+1}^{d}\gamma_i e_i &> 0 \\ y\sum_{i=m+1}^{d}\alpha_i\gamma_i(e_i\cdot e_i) &> 0 \\ y\sum_{i=m+1}^{d}\alpha_i\gamma_i &> 0 \end{aligned} \tag{9}$$

By combining Equation (8) and (9), we can derive:

$$y\sum_{i=1}^{m}\alpha_i\gamma_i < -y\sum_{i=m+1}^{d}\alpha_i\gamma_i < 0 \tag{10}$$

In addition:

$$\left|\sum_{i=1}^{m}\alpha_i\gamma_i\right| > \left|\sum_{i=m+1}^{d}\alpha_i\gamma_i\right| \tag{11}$$

□

A.2 Conditions for trimmed representations to preserve causal features

Proposition 2 (Causal Preservation)**.** *For any original representation z, the trimmed representation $\hat{z}$ preserves the correct prediction under the causal decision boundary Δp if any one of the following conditions holds:*

$$\begin{cases} y\sum_{i=1}^{m}\eta_i\alpha_i\gamma_i = 0 \\ y\sum_{i=1}^{m}\eta_i\alpha_i\gamma_i < 0 \\ 0 < y\sum_{i=1}^{m}\eta_i\alpha_i\gamma_i < y\sum_{i=1}^{d}\eta_i\alpha_i\gamma_i \end{cases} \tag{6}$$

Equation (6) characterizes three cases: (a) the top-m PCs have no contribution to the causal prediction; (b) the top-m PCs has a negative influence on the causal prediction and thus their removal is beneficial; (c) the top-m PCs has a positive contribution, but the representation forms by all PCs contribute even more strongly. When the top-m PCs have no contribution to the causal predictions, they are considered non-causal features. In other words, the removed component $z - \hat{z}$ does not contain causal information. When the top-m PCs contain causal information, m should be selected such that the causal information in the top-m PCs contributes less to the prediction compared to all the PCs, ensuring that the trimmed representation $\hat{z}$ remains causally informative.

The proof provided here corresponds to this corrected version.

Proof. As the causal decision boundary Δp can classify z correctly, we have:

$$\begin{aligned} yz \cdot \Delta p &> 0 \\ y\sum_{i=1}^{d}\alpha_i e_i \cdot \sum_{i=1}^{d}\eta_i\gamma_i e_i &> 0 \\ y\sum_{i=1}^{d}\eta_i\alpha_i\gamma_i(e_i \cdot e_i) &> 0 \\ y\sum_{i=1}^{d}\eta_i\alpha_i\gamma_i &> 0 \\ y\sum_{i=1}^{m}\eta_i\alpha_i\gamma_i + y\sum_{i=m+1}^{d}\eta_i\alpha_i\gamma_i &> 0 \end{aligned} \tag{12}$$

By rearranging Equation (12), we can derive:

$$y\sum_{i=m+1}^{d}\eta_i\alpha_i\gamma_i > -y\sum_{i=1}^{m}\eta_i\alpha_i\gamma_i \tag{13}$$

Using causal decision boundary to predict $\hat{z}$, the prediction is correct if and only if $y\hat{z} \cdot \Delta p > 0$, which leads to:

$$\begin{aligned} y\hat{z} \cdot \Delta p &> 0 \\ y\sum_{i=m+1}^{d}\alpha_i e_i \cdot \sum_{i=m+1}^{d}\eta_i\gamma_i e_i &> 0 \\ y\sum_{i=m+1}^{d}\eta_i\alpha_i\gamma_i(e_i \cdot e_i) &> 0 \\ y\sum_{i=m+1}^{d}\eta_i\alpha_i\gamma_i &> 0 \end{aligned} \tag{14}$$

Given Equation (13), Equation (14) is satisfied if any one of the following conditions holds:

$$\begin{cases} y \sum_{i=m+1}^{d} \eta_i \alpha_i \gamma_i > -y \sum_{i=1}^{m} \eta_i \alpha_i \gamma_i \geq 0 & (15) \\ y \sum_{i=m+1}^{d} \eta_i \alpha_i \gamma_i > 0 > -y \sum_{i=1}^{m} \eta_i \alpha_i \gamma_i & (16) \end{cases}$$

Equation (15) leads to:

$$y \sum_{i=1}^{m} \eta_i \alpha_i \gamma_i \leq 0 \tag{17}$$

By adding $y \sum_{i=1}^{m} \eta_i \alpha_i \gamma_i$ to Equation (16), we can derive:

$$y \sum_{i=1}^{d} \eta_i \alpha_i \gamma_i > y \sum_{i=1}^{m} \eta_i \alpha_i \gamma_i > 0 \tag{18}$$

□

A.3 Conditions for TACT to preserve a correct prediction

Proposition 3. *Suppose z is correctly classified by the learned decision boundary Δq. The trimmed representation $\hat{z}$ obtained via TACT will still be classified correctly if either of the conditions holds:*

1. *$y(z - \hat{z})\Delta q \leq 0$, or*
2. *$y(z - \hat{z})\Delta q > 0$, and Equation (7) holds, assuming $\hat{z}$ already satisfies the Causal Preservation condition (Proposition 2).*

$$\operatorname{sign}\left(\sum_{i=m+1}^{d} \eta_i \alpha_i \gamma_i\right) = \operatorname{sign}\left(\sum_{i=m+1}^{d} \alpha_i \gamma_i\right) \tag{7}$$

Equation (7) requires that when classification relies only on the representations formed by the remaining PCs, the learned decision boundary makes the same prediction as the causal decision boundary. Proposition 3 also shows that if a correct prediction is made by the learned decision boundary, TACT will preserve this correctness as long as the removed part $z - \hat{z}$ contributes negatively or does not contribute to the prediction. On the other hand, when the trimmed representation $\hat{z}$ contains sufficient causal information as established in Proposition 2, the learned decision boundary is required to align directionally with the causal decision boundary defined by the remaining PCs.

The proof provided here corresponds to this corrected version.

Proof. As the learned decision boundary Δq classify z correctly, we have:

$$\begin{aligned} yz \cdot \Delta q &> 0 \\ y(z - \hat{z}) \cdot \Delta q + y\hat{z} \cdot \Delta q &> 0 \end{aligned} \tag{19}$$

We can rewrite $y\hat{z} \cdot \Delta q$ as:

$$\begin{aligned} y\hat{z} \cdot \Delta q &= y \sum_{i=m+1}^{d} \alpha_i e_i \cdot \sum_{i=1}^{d} \gamma_i e_i \\ &= y \sum_{i=m+1}^{d} \alpha_i e_i \cdot \left(\sum_{i=1}^{m} \gamma_i e_i + \sum_{i=m+1}^{d} \gamma_i e_i\right) \\ &= y \sum_{i=m+1}^{d} \alpha_i e_i \cdot \sum_{i=1}^{m} \gamma_i e_i + y \sum_{i=m+1}^{d} \alpha_i e_i \cdot \sum_{i=m+1}^{d} \gamma_i e_i \\ &= 0 + y \sum_{i=m+1}^{d} \alpha_i e_i \cdot \sum_{i=m+1}^{d} \gamma_i e_i \\ &= y\hat{z} \cdot \Delta\hat{q} \end{aligned} \tag{20}$$

By combining Equation (19) and Equation (20), we can derive:

$$y(z - \hat{z}) \cdot \Delta q + y\hat{z} \cdot \Delta\hat{q} > 0 \tag{21}$$

The updated prediction by TACT is correct if and only if $y\hat{z} \cdot \Delta\hat{q} > 0$. Equation (21) shows that the value of $y(z - \hat{z}) \cdot \Delta q$ needs to be considered to derive the conditions under which $y\hat{z} \cdot \Delta\hat{q} > 0$.

1. When $y(z - \hat{z}) \cdot \Delta q \leq 0$, the removed part does not positively contribute to the prediction using the learned decision boundary, together with Equation (21), we can derive:

   $$y\hat{z} \cdot \Delta\hat{q} > -y(z - \hat{z}) \cdot \Delta q \geq 0 \tag{22}$$

   Equation (22) suggests that $y\hat{z} \cdot \Delta\hat{q} > 0$ is always true when $y(z - \hat{z}) \cdot \Delta q \leq 0$.

2. When $y(z - \hat{z}) \cdot \Delta q > 0$, the removed part positively contributes to the prediction using the learned decision boundary. We wish to connect with the causal decision boundary to understand the conditions. Therefore, we additionally assume $\hat{z}$ satisfies the Causal Preservation condition (Proposition 2), which suggests $y\hat{z} \cdot \Delta p > 0$.

   The updated prediction is correct, i.e. $y\hat{z} \cdot \Delta\hat{q} > 0$ if:

$$\begin{aligned}
\operatorname{sign}\left(y\hat{z} \cdot \Delta p\right) &= \operatorname{sign}\left(y\hat{z} \cdot \Delta\hat{q}\right) \\
\operatorname{sign}\left(y \sum_{i=m+1}^{d} \alpha_i e_i \cdot \sum_{i=1}^{d} \eta_i \gamma_i e_i\right) &= \operatorname{sign}\left(y \sum_{i=m+1}^{d} \alpha_i e_i \cdot \sum_{i=m+1}^{d} \gamma_i e_i\right) \\
\operatorname{sign}\left(y \sum_{i=m+1}^{d} \alpha_i \eta_i \gamma_i (e_i \cdot e_i)\right) &= \operatorname{sign}\left(y \sum_{i=m+1}^{d} \alpha_i \gamma_i (e_i \cdot e_i)\right) \\
\operatorname{sign}\left(y \sum_{i=m+1}^{d} \alpha_i \eta_i \gamma_i\right) &= \operatorname{sign}\left(y \sum_{i=m+1}^{d} \alpha_i \gamma_i\right) \\
\operatorname{sign}\left(\sum_{i=m+1}^{d} \eta_i \alpha_i \gamma_i\right) &= \operatorname{sign}\left(\sum_{i=m+1}^{d} \alpha_i \gamma_i\right)
\end{aligned} \tag{23}$$

□

B Data Augmentation for CivilComments

CivilComments considers the following demographics mentioned in a comment: male, female, LGBTQ, Christian, Muslim, other religions, Black, White. We group the demographics into gender (male/female), sexuality (LGBTQ), religion (Christian/Muslim/other religions), and race (Black/White). We notice that the comments tend to mention only one of the demographics in each group, and some comments mention more than one group. To vary demographics, we propose to introduce new demographics to the comments.

We propose to randomly insert a sentence before or after the comment. The sentences being inserted are randomly drawn from a set of sentences. Each sentence in the set mentions all demographics in one of the groups. The sentences are not toxic, so they would not affect the toxicity rating of the comment. Toxic comments remain toxic, and non-toxic comments remain non-toxic when the sentence is added. We ask ChatGPT via the web interface (`https://chatgpt.com`) to generate 20 sentences for each demographic group. We list the sentences below. Sentences from all groups make up the set from which we randomly sample for augmentation.

Gender(male/female)

- "This is a post about females and males."
- "The discussion focuses on women and men."
- "Females and males are the central topic here."
- "Women and men both contribute to this conversation."

- "This explores perspectives of females and males."
- "The post highlights contributions of women and men."
- "Both females and males are part of the narrative."
- "Women and men play essential roles in this story."
- "Females and males are equally represented here."
- "This covers aspects of both women and men."
- "This is a post about women and men."
- "The discussion centers on ladies and gentlemen."
- "Females and males are the key focus here."
- "Girls and boys both play significant roles."
- "Both genders are part of this discussion."
- "This highlights contributions from men and women."
- "Ladies and gentlemen are represented here equally."
- "The focus is on both sexes and their roles."
- "Womenfolk and menfolk shape this narrative."
- "Both males and females are included in this topic."

Sexuality (LGBTQ)

- "This is a post about LGBTQ+ and heterosexual individuals."
- "The discussion focuses on sexual minorities and heterosexual communities."
- "This highlights experiences of both LGBTQ+ and cisgender people."
- "The post compares queer and non-queer perspectives."
- "This covers topics relevant to both LGBTQ+ and straight groups."
- "Gender-diverse and cisgender voices are included in this conversation."
- "The focus is on LGBTQ+ and heterosexual rights and issues."
- "Both sexual minorities and heterosexual people's experiences are addressed here."
- "This post examines the lives of gender-nonconforming and cisgender individuals."
- "The post explores the intersection of queer and non-queer identities."
- "LGBTQ+ and heterosexual people both contribute to this topic."
- "This content engages with both gender-diverse and cisgender communities."
- "The article offers insights into the experiences of LGBTQ+ and non-LGBTQ+ individuals."
- "This is a post about LGBTQ+ and heterosexual experiences in society."
- "Both sexual minorities and heterosexual groups have a place in this discussion."
- "This conversation includes both LGBTQ+ and cisgender perspectives."
- "We explore issues affecting both sexual minorities and heterosexual individuals."
- "This is about the relationships between LGBTQ+ and heterosexual people."
- "The focus is on creating unity between LGBTQ+ and cisgender communities."
- "This post discusses challenges faced by both gender-diverse and cisgender people."

Religion (Christian/Muslim/other religions)

- "This is a post about Christians, Muslims, and followers of other faiths."
- "The discussion focuses on Christians, Muslims, and practitioners of different religions."
- "This highlights the experiences of Christians, Muslims, and believers from various traditions."
- "The post compares Christian, Muslim, and other spiritual practices."
- "This covers topics relevant to Christians, Muslims, and people of other religious backgrounds."
- "The voices of Christians, Muslims, and adherents of different faiths are included in this conversation."

- "The focus is on Christian, Muslim, and interfaith perspectives."
- "Both Christians, Muslims, and people of other beliefs contribute to this discussion."
- "This post examines the lives of Christians, Muslims, and followers of other religions."
- "The post explores the intersection of Christianity, Islam, and other spiritual practices."
- "Christians, Muslims, and people from diverse faiths share common values of compassion."
- "This content engages with Christians, Muslims, and those from various religious traditions."
- "The article offers insights into the teachings of Christians, Muslims, and other faith communities."
- "This is a post about Christians, Muslims, and adherents of various world religions."
- "Both Christians, Muslims, and individuals from different belief systems are included in this conversation."
- "The focus is on how Christians, Muslims, and people of other religions practice faith."
- "This conversation includes insights from Christians, Muslims, and followers of other spiritual paths."
- "We'll explore issues affecting Christians, Muslims, and people from various religious backgrounds."
- "This is about the relationships between Christians, Muslims, and those of other beliefs."
- "The post discusses shared values between Christians, Muslims, and adherents of other religions."

Race (Black/White)

- "This is a post about Black and White communities."
- "The discussion focuses on African American and Caucasian experiences."
- "This highlights the perspectives of Black and White individuals."
- "The post compares the lives of Black and White people."
- "This covers topics relevant to both Black and White races."
- "The voices of African Americans and Caucasians are included in this conversation."
- "The focus is on Black and White racial dynamics."
- "Both Black and White communities contribute to this discussion."
- "This post examines the experiences of Black and White individuals."
- "The post explores the intersection of African American and European American identities."
- "Black and White people play vital roles in shaping society."
- "This content engages with the experiences of Black and White groups."
- "The article offers insights into the lives of Black and White people in different settings."
- "This is a post about African American and White American experiences."
- "Both Black and White cultures have unique contributions to the world."
- "The focus is on both Black and White perspectives in social issues."
- "This conversation includes both Black and White voices."
- "We'll explore the relationship between Black and White individuals."
- "This is about the interactions between African Americans and Caucasians."
- "The post discusses challenges faced by both Black and White communities."

C Augmentation Design and Selection

Data augmentation requires careful consideration in order to achieve strong performance. It should heuristically maximize variations along non-causal directions and minimize variations along causal directions, so that the directions corresponding to non-causal features are well identified by Principal Component Analysis.

In practice, the augmentation can be treated as a hyperparameter to search over. The data collection process that raises variation and features that affect the prediction target should be analyzed to propose a set of augmentations that are semantically invariant with respect to the prediction target, yet introduce variability in other, non-causal aspects.

For example, for the commonly studied image classification task, we recommend searching over general image augmentations, such as AutoAugment [6] and RandomAugment [7]. These augmentations preserve the critical causal features, particularly the shape information of objects [10], while simultaneously injecting variability into less essential aspects. Our experiments examine the effect of different augmentation strategies on datasets where images serve as the predictive input. As shown in Table 3, augmentation affects model performance, but AutoAugment and RandomAugment could provide consistent improvements over no adaptation.

The most effective way to select the augmentation is to test on a small subset of labeled test data.

Table 3: Performance of TACT with different augmentation strategies.

Augmentation	Birdcalls	Camelyon17[1]	ImageNet-R	ImageNet-V2
no TTA	22.74	62.31	41.83	62.97
Stain color jitter/color jitter	31.14±1.69	70.17±0.05	41.78±0.01	61.88±0.11
AutoAugment	27.61±2.25	72.04±0.12	43.29±0.07	63.33±0.10
RandomAugment	32.19±1.26	79.71±0.07	43.59±0.02	62.99±0.10

D Details of Test-Time Adaptation Experiment

D.1 Model Used for Adaptation

For Birdcalls and Camelyon, to our knowledge, there were no publicly available ViT-B/32 models trained on the datasets. Therefore, we train a model using the standard empirical risk minimization. The training scripts and models can be found at our code repository `https://github.com/NancyQuris/TACT`. The details of the training are:

- Birdcalls uses a batch size of 16 and is trained for 100 epochs. AdamW is employed as the optimizer, with a learning rate of 5e-5 and weight decay of 0.001. As specified in [9], the training starts from a weight pretrained on ImageNet, and the best model is selected by macro F1 on the in-distribution validation split.
- Camelyon17 uses a batch size of 32 and is trained for 30 epochs. SGD is employed as the optimizer, with a learning rate of 5e-5 and momentum 0.9. As instructed in [28], the training starts from a randomly initialized weight, and the best model is selected by the average classification accuracy on the validation domain.

For CivilComments, we use the model provided by Wilds [28]. The model was trained on the training domain of CivilComments using empirical risk minimization. The model can be found in `https://worksheets.codalab.org/rest/bundles/0x17807ae09e364ec3b2680d71ca3d9623/contents/blob/best_model.pth`.

For ImageNet-R and ImageNet-V2, we use the model published by torchvision. The model was trained on ImageNet using empirical risk minimization. The pretrained weight `ViT_B_32_Weights.IMAGENET1K_V1` is loaded to the model for test-time adaptation.

D.2 Hyperparameter Search Space

We perform a grid search to find the best hyperparameters for the baseline methods we compared with. For backpropagation-free methods, here list the details of the hyperparameters searched:

[1]The performance of AutoAugment and RandomAugment on Camelyon17 is under the removal of principal components beginning with the 2nd. We observe that removing the first principal component only results in performance degradation. We hypothesize that important causal features might be present in the first principal component.

- T3A: Following [19], 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.
- LAME: Following [4], the k used in k-nearest neighbours is searched in {1,3,5}, and the kernel to compute distance is searched in {kNN, linear, rbf}.
- FOA: Following [36], we use 3 prompts. The population size is set to $4 + 3 \times \log(\text{prompt dim})$. The λ to balance entropy and representation distance is searched in {0.2, 0.4}.

For all backpropagation-based methods, we search the learning rate in {1e-3, 1e-4, 1e-5, 1e-6}. The adaptation is performed in a non-episodic way. For other hyperparameters used in each method, the details are listed below:

- SHOT: The method was originally proposed for source-free domain adaptation [30]. Following [19] that adapts it as a TTA strategy, β, the hyperparameter to balance information maximization and cross entropy, is set to 0.1. The hyperparameter to filter confident pseudo-labels is set to 0.9. Adam is used as the optimizer. The feature extractor is updated during adaptation. The adaptation step is set to 1.
- Tent: Following [51], SGD is used as the optimizer with momentum 0.9. The affine parameters of normalization layers are updated during adaptation. The adaptation step is set to 1.
- SAR: Following [38], 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. The adaptation step is set to 1.
- DeYO: Following [29], 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. The adaptation step is set to 1.
- TAST: Following [21], we search the number of nearby support examples N_s in {1, 2, 4, 8}. M, the number of support examples per class is searched in {1,5,20,50,100, N/A}, where N/A means storing all representations. The number of adaptation modules N_e is set to 20. Adam is used as the optimizer. The trainable module added on top of the feature extractor is adapted. The adaptation step is searched in {1, 3}.
- TSD: Following [53], the hyperparameter for feature filter M is searched in {1, 5, 20, 50, 100, N/A}, where N/A denotes no entropy filter. The tradeoff parameter λ to balance TSD loss and MSLC loss is set to 0.1. Adam is used as the optimizer. Adapting {affine parameters, classifier, feature extractor, all parameters} is searched. The adaptation step is set to 1.
- PASLE: Following [17], we search the the threshold in {0.2, 0.4, 0.6, 0.8}. The threshold gap is set to 0.1. The τ_{des} is searched in {1e-3, 1e-4}. The buffer size is set to 16, 1/4 of the batch size we used. Adam is used as the optimizer. Adapting {affine parameters, classifier, feature extractor, all parameters} is searched. The adaptation step is set to 1.

D.3 Hardware and Software Used

We perform experiments on the NVIDIA V100 GPU with 32GB memory. When the batch size is set to 64, the memory of 1 GPU is sufficient to perform test-time adaptation using TACT as well as all the baseline methods.

We implement TACT using PyTorch 2.1.2. Singular vector decomposition implemented by `torch.linalg.svd()` is used to compute the principal components, as it is computationally more stable than spectral decomposition. Since the covariance matrix is a symmetric positive semi-definite matrix, the singular vectors are the same as the eigenvectors.

E Additional Performance Study

E.1 TTA Performance on Larger Models

We examine TACT's effectiveness on larger models, specifically ViT-B/16 for images and BERT for texts. The experiment setup is consistent with that described in Section 6. Table 4 presents the performance of TACT and other state-of-the-art backpropagation-free methods on the larger architectures. Across all datasets except ImageNet-R, TACT achieves the best performance, ranking second on ImageNet-R. These results demonstrate the scalability of TACT to larger models.

The models for Birdcalls and Camelyon are trained under the same setting as that for ViT-B/32 stated in Appendix D.1. We follow the guidance of CivilComments' publisher to train BERT. The models we trained are included in our code repository. ViT-B/16 backbone for ImageNet-R and ImageNet-V2 is published by torchvision.

Table 4: Test-time adaptation performance of backpropagation-free methods on larger models. The best performance of each dataset is in bold.

Method	Birdcalls	Camelyon17	CivilComments	ImageNet-R	ImageNet-V2
No TTA	27.10	65.37	67.62	44.06	69.57
T3A	28.32±1.60	72.72±0.73	67.46±0.00	43.99±0.08	69.67±0.04
LAME	27.48±1.44	68.50±0.11	67.65±0.04	44.04±0.04	69.59±0.01
FOA	27.89±0.54	67.15±0.67	-	**47.53±2.73**	69.68±0.04
TACT	**33.65±2.11**	**72.85±0.02**	**69.76±0.44**	45.59±0.01	**69.71±0.02**

E.2 Synergy with Training-time Augmentation

The "no TTA" baselines of BirdCalls, Camelyon17, and CivilComments are trained without the augmentations used by TACT to identify and reduce non-causal features. To assess TACT's synergy with training-time augmentation, we trained models using the same augmentations as those applied by TACT and then performed test-time adaptation. For ImageNet-R and ImageNet-V2, the "no TTA" baseline provided by torchvision was trained with AutoAugment using the ImageNet policy.

Table 5 shows the test-time adaptation performance of TACT on models trained with the same augmentation strategy. The results show that, even when models are trained with these augmentations, TACT further improves test-time performance. This highlights TACT's ability to synergize with training-time augmentation and provides strong evidence of its effectiveness and generalizability.

Table 5: Test-time adaptation performance of TACT with training-time augmentation models.

	Birdcalls	Camelyon17	CivilComments	ImageNet-R	ImageNet-V2
no TTA (train time aug)	29.86	74.09	64.60	41.83	62.97
+ TACT	30.57±0.96	77.27±0.03	68.84±0.20	43.29±0.07	63.33±0.10

E.3 TTA Performance under Different Batch Size

We study the test-time adaptation performance of TACT on ImageNet-R when the test batch size varies. Table 6 shows the result when the test batch size is set to 1, 4, 16, 64 and 128, respectively. The performance remains stable across different batch sizes. Even with a batch size of 1, the performance only decreases by 0.06% compared to a batch size of 64. Moreover, TACT still improves performance by 1.7% over the no-adaptation baseline when only one sample is available per batch during adaptation. The result suggests that TACT is robust to variations in batch size, maintaining high performance even when batch sizes are small. This makes it well-suited for situations where the number of test samples per batch is constrained.

Table 6: Test-time adaptation performance (%) of TACT on ImageNet-R under different batch sizes.

no TTA	batch size = 1	batch size = 4	batch size = 16	batch size = 64	batch size = 128
41.83	43.53±0.02	43.51±0.03	43.55±0.06	43.59±0.02	43.56±0.03

E.4 Computational Cost

We compared the computational requirements of TACT with those of other backpropagation-free methods on the Birdcalls dataset using a ViT-b/32 backbone. As shown in Table 7, TACT incurs higher time and GPU memory consumption relative to alternative approaches. Nevertheless, this additional computational cost results in substantial performance gains (Table 1), which justifies the trade-off. Future work may explore optimization strategies, such as more efficient eigendecomposition techniques for PCA, to reduce the overhead.

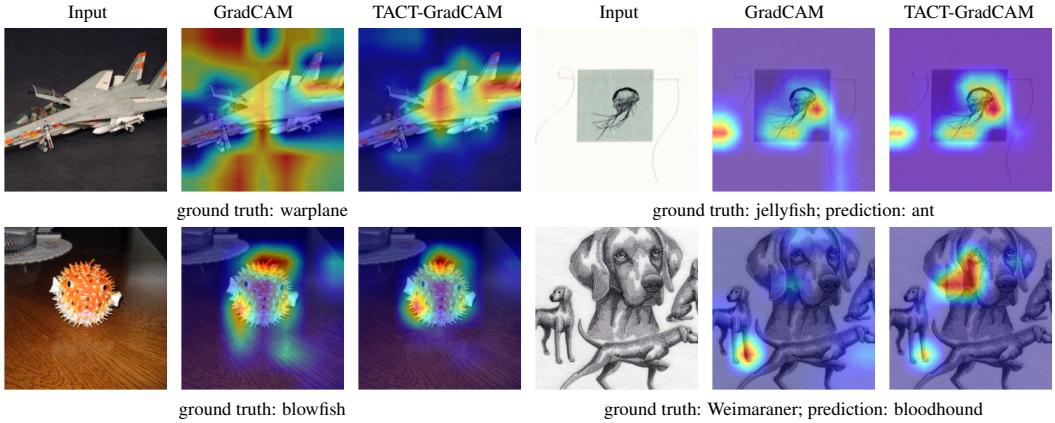
Table 7: Time and GPU memory required by backpropagation-free methods on Bridcalls.

	time (second)	GPU memory (MB)
T3A	7.67	667.42
LAME	7.34	667.42
FOA	16.83	667.42
TACT (num aug=128)	112.22	1750.21
TACT (num aug=256)	170.00	2966.21
TACT (num aug=512)	323.62	5398.21

E.5 Additional Visualization of Predictions after Causal Trimming

We provide more GradCAM visualization of the original predictions and the predictions made by TACT on samples from ImageNet-R. Figure 5 shows the visualizations.

Compared to original predictions, predictions made by TACT focus less on non-causal information. For example, TACT pays less attention to the background of the warplane example, and the blowfish example. The focus on the information that is semantically correlated with the class is retained in predictions made by TACT in the above examples. When the causal information is not important to the original prediction, prediction made by TACT leverages the causal information and thus turn the wrong prediction correct, as shown in the example of jellyfish and bloodhound.



(a) correct predictions (b) wrong predictions corrected by TACT

Figure 5: Additional GradCAM visualizations of the original predictions and TACT's predictions.

F Alternative Design of TACT

F.1 ICA to Find Non-Causal Directions

We experiment using an alternative direction finding method, Independent Component Analysis (ICA) with TACT. We rank the independent components by the variance of the scalars of features on the components. We remove the top independent components that have maximum variance. Table 8 shows the result on the Birdcalls dataset. ICA performs inferior to Principal Component Analysis (PCA), but better than no adaptation. Although ICA overcomes the orthogonality constraints of PCA, it only looks for statistically independent components and assumes each component follows a non-Gaussian distribution. Causal and non-causal features might not follow the non-Gaussian distribution assumption under augmentations that vary non-causal features.

Table 8: Performance of TACT with ICA to find non-causal directions.

no TTA	TACT w/ PCA	TACT w/ ICA
22.74	31.14±1.69	25.53±1.06

F.2 Causal Trimming Based on a Threshold

We consider using the variance that the top principal components (PC) account for as a threshold to decide whether causal trimming is conducted or not. When the augmentation only changes non-causal features and causal features remain unchanged, datapoints that are invariant to augmentations should have smaller variance of the top PCs. Thus, if the variance is smaller than a threshold, causal trimmings will not be conducted on the data. As the range of variance is not known and it could change significantly, setting a numerical threshold might not be feasible. We consider normalized variance, where we divide the variance of top PCs by the sum of variances of all PCs. Table 9 shows the result on the Birdcalls dataset. Removing components based on a threshold does not outperform using no threshold.

Table 9: Performance of TACT when causal trimming is performed based on a threshold τ.

no TTA	TACT	TACT (τ=0.1)	TACT (τ=0.2)	TACT (τ=0.3)
22.74	31.14±1.69	30.99±2.18	31.03±2.19	28.03±3.12