# CS-TTA: Preserving Concept Sensitivity in Test-Time Adaptation

Junah Jung[1⋆], Yeon Gyu Han[2⋆], Chang Min Park[3,4⋆⋆], and Dongheon Lee[3,4,5⋆⋆]

[1] Interdisciplinary Program in Medical Informatics, Seoul National University College of Medicine, Seoul, Republic of Korea
[2] Department of Biomedical Engineering, Chungnam National University College of Medicine, Daejeon, Republic of Korea
[3] Department of Radiology, Seoul National University Hospital, Seoul National University College of Medicine, Seoul, Republic of Korea
[4] Institute of Medical and Biological Engineering, Seoul National University Medical Research Center, Seoul, Republic of Korea
[5] Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Republic of Korea

**Abstract.** Test-time adaptation (TTA) methods are typically evaluated by task accuracy, leaving open the question of whether adaptation also changes *which features* the model relies on. We examine this question by tracking concept sensitivity throughout adaptation using TCAV directional derivatives. Across five TTA methods and five benchmarks spanning natural image and medical imaging domains, we observe recurring shifts in concept sensitivity: adaptation can reduce sensitivity to task-relevant or pathology-related concepts while increasing sensitivity to nuisance concepts such as backgrounds or institutional artifacts—a phenomenon we term *Concept Sensitivity Drift*, even when accuracy improves. On Waterbirds, where the causal/spurious split is defined by construction, this drift coincides with degraded worst-group accuracy; on cross-hospital medical adaptation, it coincides with increased reliance on hospital-specific artifacts rather than pathology. Motivated by this observation, we propose CS-TTA, a source-concept-supervised but target-label-free plug-in regularizer that mitigates excessive concept-sensitivity drift during adaptation. CS-TTA requires no architectural changes and can be added to any existing TTA objective. Experiments show that CS-TTA provides consistent worst-group accuracy gains on spurious-correlation benchmarks, improves AUC by 1.4–1.8 points on cross-hospital medical adaptation, and gives consistent overall accuracy gains on CIFAR-10-C and ImageNet-C.1234

**Keywords:** Test-time adaptation · Concept sensitivity · Spurious correlations

⋆ Equal contribution.
⋆⋆ Corresponding authors.

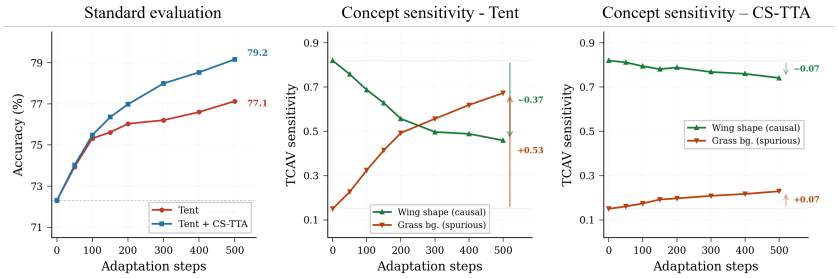


**Fig. 1: Concept sensitivity drifts during TTA despite accuracy gains.** (a) Tent improves accuracy from 72% to 77%, and CS-TTA further improves it to 79%. (b) However, Tent shifts sensitivity away from the causal concept (*wing shape*, −0.37) and toward the spurious concept (*grass background*, +0.53). (c) CS-TTA preserves concept sensitivity, limiting drift to −0.07 and +0.07.

## 1 Introduction

Test-time adaptation (TTA) has become a widely adopted approach for handling distribution shift: given a model pretrained on source data, TTA updates its parameters on unlabeled target samples to improve robustness [21, 28, 29, 32]. Methods based on entropy minimization [28], teacher–student consistency [29], and selective sample filtering [21] have demonstrated consistent accuracy gains across a wide range of corruption and domain shift benchmarks. Recent advances further address practical concerns such as small batch sizes [14], catastrophic forgetting [29], and temporally correlated test streams [22].

These methods, however, evaluate adaptation success through a single lens: task accuracy. A model that achieves higher accuracy after adaptation is deemed successful regardless of *how* it arrives at its predictions. This leaves a critical question unexamined: does adaptation preserve the model's reliance on task-relevant features, or does it increase dependence on nuisance or spurious correlations? This distinction has practical consequences. In domains where models are known to exploit spurious correlations—chest X-ray classifiers attending to hospital tags rather than pathology [4], or object recognizers relying on texture over shape [6]—an accuracy improvement driven by increased shortcut learning [6] would undermine the very trustworthiness that motivates deployment.

We provide a systematic investigation of this question. Using TCAV [15]—a well-established tool for measuring how sensitive a model's predictions are to human-interpretable concepts via directional derivatives in activation space—we track concept sensitivity *throughout* the adaptation process. Unlike pixel-level attribution methods such as GradCAM [25], which show *where* the model attends, TCAV quantifies *how much* each semantic concept contributes to the prediction, providing quantitative measurements at the concept level.

Figure 1 illustrates our key finding on the Waterbirds benchmark. When Tent [28] adapts a ResNet-50, accuracy improves from 72% to 77%—a result that would conventionally be reported as a success. However, TCAV analysis

reveals that the sensitivity to a causal concept (*e.g.*, *wing shape*) drops by 0.37, while the sensitivity to a spurious concept (*e.g.*, *grass background*) rises by 0.53. The model's prediction basis shifts toward spurious features despite the accuracy gain. This observation is not specific to Tent: across multiple TTA methods and domains, we observe analogous shifts from task-relevant or pathology-related concepts toward nuisance or artifact concepts.

We formalize this phenomenon as *Concept Sensitivity Drift* (CSD), defined as the change in TCAV scores between the source and adapted models for a given set of concepts. CSD provides a single, interpretable number that quantifies how much adaptation has altered the model's concept sensitivity. Across our experiments, we find that CSD is (i) consistently observed across all tested methods—every method exhibits non-trivial drift even when accuracy improves, and (ii) associated with later degradation—higher CSD at early adaptation steps strongly correlates with later accuracy degradation, suggesting its potential as an early diagnostic signal for adaptation failure in the evaluated settings.

Building on this diagnostic, we propose CS-TTA: a plug-in regularization term that penalizes deviations in concept sensitivity from the source model during adaptation. For concepts whose CSD exceeds a threshold, the regularizer encourages the adapted model's sensitivity profile to remain close to the source. This additional loss can be combined with *any* existing TTA objective without modifying the model architecture, the set of trainable parameters, or the adaptation procedure. In experiments, CS-TTA provides consistent worst-group accuracy gains on Waterbirds and CelebA, improves AUC by 1.4–1.8 points on cross-hospital chest X-ray classification, and provides consistent overall accuracy gains on corruption benchmarks.

Our contributions are as follows:

- We observe that popular TTA methods can shift concept sensitivity toward nuisance features and away from task-relevant ones in our evaluated settings (five methods and five benchmarks spanning general and medical domains), a pattern we term *Concept Sensitivity Drift* (CSD).
- CSD is associated with later degradation: higher early drift correlates with later accuracy degradation ($\rho = 0.76$, pooled across methods and datasets), suggesting its potential as an early diagnostic signal.
- CS-TTA, a source-concept-supervised but target-label-free plug-in regularizer that mitigates excessive concept-sensitivity drift, improves both overall and worst-group accuracy across all tested baselines without architectural changes.

## 2 Related Work

**Test-Time Adaptation.** Since Tent [28] demonstrated that entropy minimization over batch normalization (BN) parameters yields immediate robustness gains at test time, a rich line of work has expanded the TTA toolkit along several axes. EATA [21] improves reliability by filtering high-entropy samples before gradient updates and regularizing important parameters via Fisher information.

CoTTA [29] mitigates catastrophic forgetting in continual shift settings through mean-teacher consistency and stochastic weight restoration. SAR [22] introduces sharpness-aware minimization to stabilize the loss landscape during adaptation. MEMO [32] enforces consistency across augmented views of each test sample. EcoTTA [27] reduces memory consumption by freezing the backbone and adapting only a lightweight meta-network. More recently, DeYO [17] observed that entropy alone cannot distinguish genuine confidence from spurious-correlation-driven confidence, and introduced PLPD, a shape-awareness metric that filters samples based on the influence of object shape on predictions. MemBN [14] addressed the small-batch failure mode of BN-based TTA by aggregating normalization statistics across a memory queue. Despite these methodological advances, all of these approaches evaluate success exclusively through task accuracy or error rate. The question of whether the model's concept sensitivity—which features it relies on and why—is preserved or degraded during adaptation has not been addressed.

**Concept-Based Explanations.** Rather than attributing predictions to individual pixels [25], concept-based methods explain model behavior in terms of human-interpretable semantic units. TCAV [15] learns a Concept Activation Vector (CAV) via linear classification in activation space and measures the model's sensitivity to that concept through a directional derivative. Subsequent work has extended this framework to automatic concept discovery [7], concept bottleneck architectures [16, 31], efficient CAV computation [24], and cross-layer consistency [9]. A shared assumption is that the model under analysis is *static*: explanations are computed for a fixed, pretrained model. We study how concept sensitivity *changes* as TTA updates model parameters—a setting that remains unexplored. The most related efforts are CONDA [2], which adapts a concept bottleneck layer within the prediction pipeline, and PatchSAE [18], which analyzes how adaptation remaps sparse-autoencoder features. Unlike both, CS-TTA uses concept sensitivity as a frozen, external diagnostic without altering the model's architecture or prediction pipeline.

## 3 Methods

We begin with the necessary background on TTA and concept sensitivity (Section 3.1), then formalize the concept sensitivity drift phenomenon (Section 3.2), and finally present our sensitivity-preserving regularization (Section 3.3).

### 3.1 Preliminaries

**Test-Time Adaptation.** Consider a model $f_\theta$ pretrained on source data $\mathcal{D}_s$. At test time, the model encounters target samples $\{x_t\}$ drawn from a shifted distribution. TTA methods update a subset of model parameters $\theta$ (*e.g.*, BN affine parameters) by minimizing an unsupervised objective:

$$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}_{\text{TTA}}(\theta_t; x_t) \tag{1}$$

where $\mathcal{L}_{\text{TTA}}$ varies by method (*e.g.*, entropy for Tent [28], augmentation consistency for CoTTA [29]).

**Concept Activation Vectors.** Given a set of images labeled as positive or negative for a concept $C$ (*e.g.*, *striped texture* vs. random images), a Concept Activation Vector (CAV) is the normal vector $v_C$ of a linear classifier trained on the model's intermediate activations at a chosen layer $l$ [15]. This vector captures the direction in activation space that best separates the concept from its complement.

**TCAV Score.** The sensitivity of the model's prediction for class $k$ to concept $C$ at input $x$ is given by the directional derivative:

$$S_{C,k}(x) = \nabla_{a^{(l)}} f_k(a^{(l)}(x)) \cdot v_C \tag{2}$$

where $a^{(l)}(x)$ is the activation of layer $l$ and $f_k$ is the logit for class $k$. A positive $S_{C,k}$ means that moving the activation in the concept direction increases the predicted probability of class $k$. The TCAV score for concept $C$ and class $k$ is the proportion of inputs for which $S_{C,k}$ is positive:

$$\text{TCAV}_{C,k} = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathbb{1}[S_{C,k}(x) > 0] \tag{3}$$

TCAV score close to 1 indicates strong positive reliance of class $k$ on concept $C$.

### 3.2 Concept Sensitivity Drift

During TTA, the model parameters $\theta$ evolve according to Equation 1. This changes the activation function $a^{(l)}(\cdot; \theta_t)$ at each adaptation step, which in turn changes the directional derivative of Equation 2 even though the CAV $v_C$ remains frozen. Keeping $v_C$ fixed lets us measure changes in the model's sensitivity along a consistent direction; we examine the validity of this design in Section 4.2 (Adapted-feature CAV stability).

We define the *Concept Sensitivity Drift* for concept $C$ and class $k$ at adaptation step $t$ as:

$$\text{CSD}_{C,k}(t) = \text{TCAV}_{C,k}(\theta_t) - \text{TCAV}_{C,k}(\theta_0) \tag{4}$$

where $\theta_0$ denotes the source model and $\theta_t$ denotes the model after $t$ adaptation steps. In practice, for each input $x$ we compute CSD with respect to its predicted class $k(x)$ (denoted $\text{CSD}_{\text{cur}}$ when disambiguation is needed); we write $\text{CSD}_C$ when the class index is clear from context. To isolate class-target effects, we additionally consider three analysis-only diagnostics evaluated in Section 4.2: $\text{CSD}_{\text{src}}$ ($k$ fixed to the source prediction $k_0(x)$), $\text{CSD}_{y/\text{task}}$ ($k$ fixed to the ground-truth class, or a fixed task logit for binary medical findings), and $\text{CSD}_{\text{noflip}}$ (common subset where the adapted prediction matches the source prediction under both the base TTA method and +CS-TTA). The sign of $\text{CSD}_C$ is interpreted relative to the concept's role: a positive CSD for a nuisance concept signals increasing reliance, while a negative CSD for a task-relevant concept signals decreasing

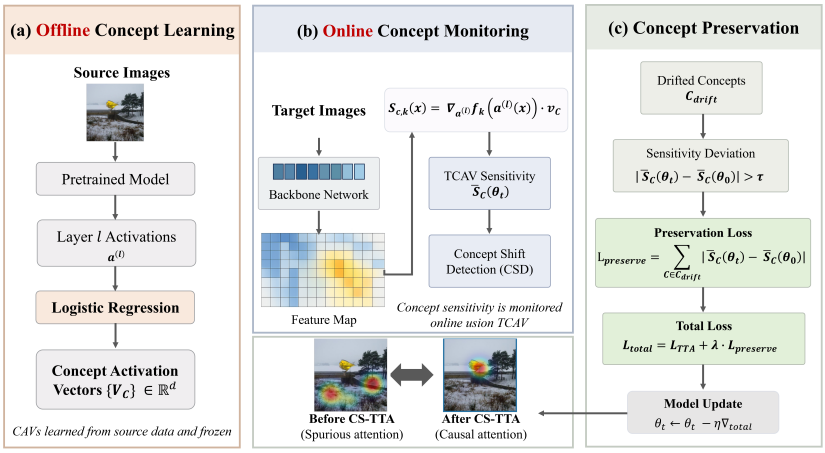


**Fig. 2: Overview of CS-TTA.** (a) CAVs are learned offline from source activations via logistic regression and frozen. (b) During adaptation, concept sensitivity is monitored online using TCAV. (c) Concepts with drift exceeding $\tau$ are regularized via $\mathcal{L}_{\text{preserve}}$, and the model is updated with $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{TTA}} + \lambda \mathcal{L}_{\text{preserve}}$.

reliance. Whether a concept is task-relevant or nuisance is determined by the benchmark's prior knowledge (*e.g.*, dataset annotations on Waterbirds/CelebA, radiologist-defined regions on chest X-rays); CSD itself is correlational, not a causal certificate.

To obtain an aggregate measure of drift across all monitored concepts, we define:

$$\text{CSD}(t) = \frac{1}{|\mathcal{C}|} \sum_{C \in \mathcal{C}} |\text{CSD}_C(t)| \tag{5}$$

where $\mathcal{C}$ is the set of monitored concepts. We show in Section 4.2 that this aggregate CSD is consistently positive across TTA methods (*i.e.*, all tested methods exhibit concept sensitivity drift) and correlates with later accuracy degradation, suggesting its potential as an early diagnostic signal for adaptation failure in our evaluated settings.

### 3.3 Sensitivity-Preserving Regularization

Given that concept sensitivity drift is measurable and associated with later degradation, a natural response is to constrain the adaptation process to preserve the source model's sensitivity profile. We propose a regularization term that penalizes large deviations in concept sensitivity for concepts that exhibit significant drift. The key insight is that preserving concept sensitivity does not prevent adaptation—it steers the model away from solutions that exploit spurious correlations.

**Batch-level sensitivity.** For a mini-batch $\mathcal{B}_t$ at adaptation step $t$, the sample-level sensitivity of concept $C$ for class $k$ is given by Equation 2. We define the

mean sensitivity of concept $C$ over the batch as:

$$\bar{S}_C(\theta_t) = \frac{1}{|\mathcal{B}_t|} \sum_{x \in \mathcal{B}_t} S_{C,k(x)}(x; \theta_t) \tag{6}$$

where $k(x)$ denotes the predicted class for input $x$. Unlike the TCAV score (Equation 3), which binarizes $S_{C,k}$ and is therefore not differentiable, $\bar{S}_C$ is a continuous quantity that is differentiable with respect to $\theta_t$ through the activation function and the logit output. This makes $\bar{S}_C$ suitable as an optimization target, while CSD (Equation 4) serves as the criterion for selecting *which* concepts to regularize.

**Sensitivity-preserving loss.** We selectively regularize concepts whose sensitivity has drifted beyond a threshold $\tau$, penalizing deviation from the source sensitivity:

$$\mathcal{L}_{\text{preserve}}(\theta_t) = \sum_{C:\, |\bar{S}_C(\theta_t) - \bar{S}_C(\theta_0)| > \tau} \big(\bar{S}_C(\theta_t) - \text{sg}\big(\bar{S}_C(\theta_0)\big)\big)^2 \tag{7}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator, treating the source sensitivity $\bar{S}_C(\theta_0)$ as a fixed target. The selection criterion $|\bar{S}_C(\theta_t) - \bar{S}_C(\theta_0)| > \tau$ is evaluated per batch, using the same batch-level sensitivity defined in Equation 6; no population-level statistics are needed during adaptation. The threshold $\tau$ serves a dual role: it avoids constraining concepts that are not meaningfully affected by adaptation, and it reduces computational cost by limiting the number of active regularization terms. The total adaptation objective becomes:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{TTA}} + \lambda \cdot \mathcal{L}_{\text{preserve}} \tag{8}$$

where $\lambda$ controls the regularization strength. Figure 2 provides an overview of the complete pipeline.

**Efficient computation.** The exact sensitivity $S_{C,k}$ is used in the second-order variant; the first-order default optimizes the CAV-projection surrogate below, whose online cost is dominated by a dot product between activations and each monitored CAV. For $|\mathcal{C}|$ concepts, batch size $B$, and dimension $d$, the overhead is $O(|\mathcal{C}| \cdot B \cdot d)$ multiplications—with typical values ($|\mathcal{C}| = 15$, $B = 64$, $d = 2048$), amounting to 1.1–2.8% additional wall-clock time per step (Supplementary Section F).

**First-order approximation.** The sensitivity $S_{C,k}$ depends on $\theta_t$ through both the activation $a^{(l)}$ and the gradient $\nabla_{a^{(l)}} f_k$; differentiating $\mathcal{L}_{\text{preserve}}$ thus involves second-order Hessian terms. Following first-order approximation strategies common in meta-learning [20], our default replaces $\bar{S}_C$ with the CAV-projection surrogate $P_C$ defined below, which preserves the projection of the activation $a^{(l)}(\theta_t)$ onto the CAV direction $v_C$ (the "concept component" of the activation). Our ablation (Section 4.4) confirms this approximation introduces no measurable accuracy difference while halving overhead.

**Diagnostic vs. optimized quantity.** The binary TCAV score (Equation 3) is non-differentiable and reported only as a diagnostic; the exact preservation loss

is defined on the continuous batch-level sensitivity $\bar{S}_C$ (Equation 6). In the first-order default, $\bar{S}_C$ in Equation 7 is replaced by the differentiable CAV-projection surrogate

$$P_C(\theta_t) = \frac{1}{|\mathcal{B}_t|} \sum_{x \in \mathcal{B}_t} a^{(l)}(x; \theta_t)^\top v_C \tag{9}$$

which uses the same source-defined direction $v_C$ but avoids the Hessian; the exact $\bar{S}_C$-based version (second-order ablation, Section 4.4) yields nearly identical performance.

Since the first-order default is implemented as an additive regularizer on internal activations and CAV directions, CS-TTA makes no assumptions about the form of $\mathcal{L}_{\text{TTA}}$ or which parameters are updated. It can be combined with any TTA method—entropy minimization (Tent), consistency regularization (CoTTA), sample selection (EATA), sharpness-aware minimization (SAR), or disagreement-based filtering (DeYO)—by adding $\lambda \cdot \mathcal{L}_{\text{preserve}}$ to the existing loss. The only prerequisite is a set of concept examples from the source domain to learn the CAVs, which are computed once offline via logistic regression ($< 1$ second per concept on CPU) and remain frozen throughout adaptation.

## 4 Experiments

We design experiments to answer three questions: (i) Does concept sensitivity drift occur across TTA methods, datasets, and domains? (Section 4.2) (ii) Does preserving concept sensitivity improve adaptation outcomes? (Section 4.3) (iii) How sensitive is CS-TTA to design choices? (Section 4.4)

### 4.1 Setup

**Datasets.** We evaluate across two domains to demonstrate generalizability.

*General domain.* (i) **CIFAR-10-C** [10]: 15 corruption types at severity level 5, applied to CIFAR-10 test images (10 classes). (ii) **ImageNet-C** [10]: the same 15 corruption types at severity level 5, on 1,000 ImageNet classes. (iii) **Waterbirds** [23]: a spurious-correlation benchmark where bird type (waterbird vs. landbird) is confounded with background (water vs. land). (iv) **CelebA** [19]: hair color prediction confounded with gender, following the standard worst-group evaluation.

*Medical domain.* (v) **CheXpert→MIMIC-CXR**: we train on CheXpert [12] and adapt at test time to MIMIC-CXR [13], evaluating three pathological findings (Pneumonia, Pneumothorax, Enlarged Cardiomediastinum). Cross-hospital chest X-ray classification is a natural setting for concept sensitivity drift: models are known to exploit hospital-specific markers (*e.g.*, text overlays, device shadows) rather than pathological features [4]. Full results across all four source datasets (CheXpert, NIH ChestX-ray14 [30], PadChest [1], RSNA Pneumonia [26]) are in the supplementary material.

**TTA baselines.** We apply CS-TTA on top of five TTA methods with publicly available code: Tent [28], EATA [21], CoTTA [29], SAR [22], and DeYO [17]. The

non-adaptive source model serves as a reference. All five baselines are evaluated on both general and medical domains.

**Backbones.** For CIFAR-10-C, we use WideResNet-28-10 following the standard TTA protocol [28]. For ImageNet-C, Waterbirds, and CelebA, we use ResNet-50 [8] pretrained on ImageNet; ViT-B/16 [5] results are in the supplementary. For the medical domain, we use DenseNet-121 [11] from TorchXRayVision [3], which provides pretrained weights for chest X-ray classification. Following standard practice [21, 28], TTA updates only batch normalization affine parameters ($\gamma$, $\beta$), keeping all other weights frozen.

**Concept definitions.** For CIFAR-10-C (15 concepts) and ImageNet-C (20 concepts), CLIP zero-shot classification produces task-relevant attributes (shape, texture, material; *e.g.*, *round*, *furry*, *metallic*) alongside nuisance scene/background attributes (*e.g.*, *outdoor*, *indoor*). For Waterbirds (12 concepts) and CelebA (15 concepts), dataset annotations [19, 23] provide causal/target-related attributes (*e.g.*, *wing shape*, *hair color*), spurious/group-correlated nuisance features (*e.g.*, *water background*, *male*), and neutral properties. For the medical domain (20 concepts), we use pathological findings (*e.g.*, *lung opacity*), anatomical structures (*e.g.*, *lung field*), and institutional artifacts (*e.g.*, *patient markers*, *device shadows*); medical CAV positives are obtained from radiologist-annotated regions in the source dataset. Complete concept lists and CAV quality validation are in the Supplementary material. CAVs are trained on source-domain activations via logistic regression at the penultimate residual block (WRN-28 / ResNet-50) or the final dense block (DenseNet-121).

**Concept supervision.** CS-TTA is source-concept-supervised but target-label-free: CAVs are trained offline from source-domain concept examples (200 positive + 200 random per concept; 4.8–8.0k instances per dataset) and frozen, with no target labels used during adaptation. CLIP is used only offline to obtain concept-positive labels for CIFAR-10-C, ImageNet-C, and the Waterbirds weak-label ablation; no CLIP or VLM inference is performed at test time.

**Metrics.** For general domain, we report overall accuracy (Acc) on all datasets, and worst-group accuracy (WG) on Waterbirds and CelebA. For the medical domain, we report AUC following standard practice in chest X-ray classification. CSD (Equation 5) is reported separately in Table 1 as a diagnostic measure.

**Implementation details.** All experiments use batch size 64 (general) or 16 (medical, due to higher resolution), learning rate $10^{-3}$ for BN parameters, and 1 epoch of adaptation. For CS-TTA, we set $\lambda = 0.1$ and $\tau = 0.05$ based on a small validation split (5% of source data).

### 4.2 Concept Sensitivity Drift Analysis

**CSD across methods.** Figure 3 plots the aggregate CSD and accuracy over adaptation steps for five TTA methods on Waterbirds with ResNet-50. Every tested method exhibits monotonically increasing CSD: as adaptation progresses, the model's concept sensitivity drifts further from the source. CoTTA shows the highest CSD (0.35 at step 500), consistent with its aggressive pseudo-label-driven updates that modify the model more substantially. SAR and DeYO exhibit

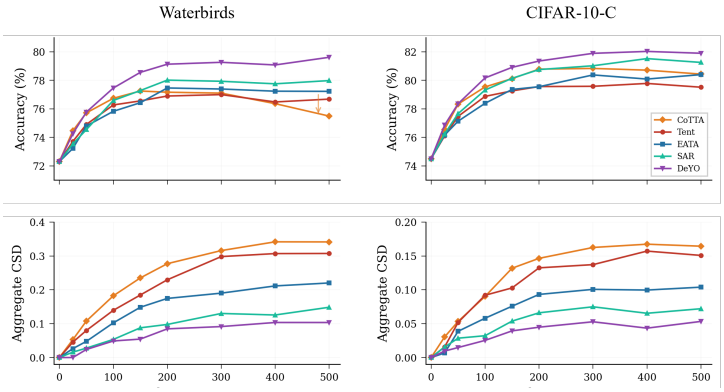


**Fig. 3: Accuracy and CSD over adaptation steps** (Waterbirds, ResNet-50). Left: without CS-TTA, accuracy peaks then degrades for aggressive methods (*e.g.*, CoTTA), while CSD increases monotonically. Right: with CS-TTA, accuracy is higher and more stable, and CSD is substantially reduced across all methods.

lower but non-trivial drift (0.15 and 0.11, respectively), reflecting their conservative update strategies—sharpness-aware regularization and disagreement filtering partially constrain parameter updates, but neither explicitly monitors concept sensitivity. Critically, CSD increases *even when accuracy is still improving*, revealing a cost of adaptation that accuracy alone does not capture. This observation is consistent with prior findings that entropy minimization can reduce prediction uncertainty by leveraging spurious patterns [17], and provides concept-level evidence for this failure mode.

**Signed per-concept CSD analysis.** Per-concept analysis reveals a consistent signed pattern: object-related concepts (*e.g.*, *wing shape*, *beak type*) exhibit negative CSD (sensitivity *decreases*), while background-related concepts (*e.g.*, *water background*, *grass background*) exhibit positive CSD (sensitivity *increases*). On Waterbirds, the mean CSD for the four object-related concepts under Tent is $-0.15$, while the mean CSD for the four background-related concepts is $+0.22$. A similar signed pattern appears in the CheXpert→MIMIC-CXR Pneumonia analysis: pathology-related concepts (*e.g.*, *lung opacity*) show negative CSD ($-0.13$ on average), while institutional-artifact concepts (*e.g.*, *patient markers*) show positive CSD ($+0.18$). A full per-concept CSD breakdown is provided in the Supplementary material (Section B.3).

**Correlation with accuracy.** Table 1 reports CSD values at convergence for each method and dataset. To assess whether early CSD is associated with later degradation, we compute the Spearman rank correlation between CSD at step 100 and the accuracy change over the subsequent 400 steps, pooling across all method–dataset pairs ($n = 25$). The correlation is $\rho = 0.76$ ($p < 0.001$): methods and datasets with higher early CSD tend to show larger subsequent accuracy drops. For instance, CoTTA on Waterbirds accumulates CSD of 0.18 by step 100 and exhibits the most pronounced later degradation, while DeYO's lower

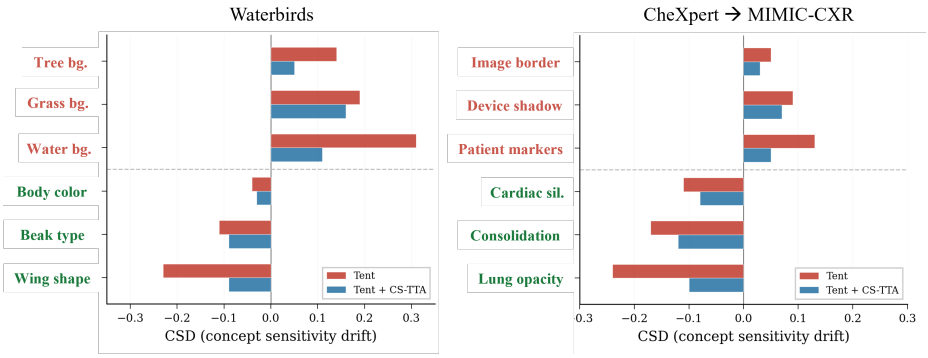


**Fig. 4: Per-concept CSD breakdown** for Tent vs. Tent+CS-TTA. (a) Waterbirds: object-related concepts (green, bottom) drift negatively while background concepts (red, top) drift positively. CS-TTA reduces drift in both directions. (b) CheXpert→MIMIC-CXR: the same asymmetric pattern holds for pathology-related vs. institutional-artifact concepts.

**Table 1: CSD after adaptation** (averaged over 3 runs). Lower values indicate less drift from the source model's concept sensitivity.

| Method | General Domain | | | | Medical |
|---|---|---|---|---|---|
| | C10-C | IN-C | Waterbirds | CelebA | Medical |
| Tent | 0.15 | 0.18 | 0.31 | 0.24 | 0.22 |
| EATA | 0.10 | 0.12 | 0.22 | 0.19 | 0.17 |
| CoTTA | 0.17 | 0.21 | 0.35 | 0.28 | 0.25 |
| SAR | 0.07 | 0.09 | 0.15 | 0.14 | 0.11 |
| DeYO | 0.05 | 0.07 | 0.11 | 0.10 | 0.08 |

early CSD (0.04) is associated with more stable long-term performance. This association suggests that CSD could serve as an online criterion for deciding when to stop or restart adaptation—a direction we leave for future work.

**Robustness to class-target and coordinate confounds.** Two potential confounds could explain the observed signed per-concept pattern without invoking semantic drift: (i) the predicted class $k(x)$ may change during adaptation, and (ii) the feature space may rotate as $\theta$ evolves, so frozen source CAVs might measure coordinate drift rather than semantic reliance. We address both with fixed-target and CAV-stability diagnostics. Across four fixed-target variants ($\mathrm{CSD}_{\mathrm{cur}}$, $\mathrm{CSD}_{\mathrm{src}}$, $\mathrm{CSD}_{y/\mathrm{task}}$, $\mathrm{CSD}_{\mathrm{noflip}}$; defined in Section 3.2), the task-relevant-down / nuisance-up signed pattern persists, and +CS-TTA reduces signed CSD toward zero in every case. Frozen source CAVs remain meaningful probes after adaptation: held-out source-concept accuracies on Tent-adapted activations drop by only 1.6–2.0 points across all five benchmarks, and CAVs re-fitted on adapted activations have cosine 0.83–0.87 with source CAVs (Waterbirds signed-group

**Table 2: Quantitative results** (mean±std over 3 runs). CS-TTA improves performance across all baselines; worst-group and medical AUC gains exceed seed-level standard deviations. Medical column reports mean AUC across Pneumonia, Pneumothorax, and Enl. Cardiomediastinum on CheXpert→MIMIC-CXR (DenseNet-121); per-finding results are in the Supplementary material.

| Method | Corruption | | Waterbirds | | CelebA | | Medical |
|---|---|---|---|---|---|---|---|
| | C10-C | IN-C | Acc | WG | Acc | WG | AUC |
| Source | 74.5±0.2 | 39.2±0.3 | 72.3±0.5 | 58.1±0.8 | 85.1±0.3 | 47.2±0.9 | 0.762±.004 |
| Tent | 79.8±0.3 | 44.8±0.4 | 76.8±0.5 | 56.2±1.0 | 88.3±0.3 | 45.8±1.1 | 0.778±.004 |
| +CS-TTA | **81.2**±0.3 | **45.9**±0.3 | **78.5**±0.4 | **65.8**±0.9 | **89.1**±0.3 | **54.3**±1.0 | **0.795**±.003 |
| EATA | 80.5±0.3 | 45.3±0.4 | 77.5±0.5 | 57.8±1.0 | 88.9±0.3 | 46.5±1.1 | 0.783±.004 |
| +CS-TTA | **81.6**±0.3 | **46.1**±0.3 | **79.2**±0.4 | **66.5**±0.9 | **89.5**±0.3 | **55.1**±1.0 | **0.798**±.003 |
| CoTTA | 80.9±0.4 | 45.8±0.4 | 77.2±0.6 | 55.1±1.2 | 89.1±0.3 | 44.7±1.3 | 0.774±.005 |
| +CS-TTA | **82.0**±0.3 | **46.5**±0.3 | **79.8**±0.5 | **65.2**±1.0 | **89.8**±0.3 | **53.8**±1.1 | **0.792**±.004 |
| SAR | 81.3±0.3 | 46.1±0.3 | 78.1±0.4 | 59.3±0.9 | 89.4±0.2 | 48.1±1.0 | 0.786±.003 |
| +CS-TTA | **82.1**±0.3 | **46.8**±0.3 | **79.5**±0.4 | **67.1**±0.8 | **89.9**±0.2 | **56.2**±0.9 | **0.800**±.003 |
| DeYO | 81.8±0.3 | 46.5±0.3 | 79.5±0.4 | 63.8±0.9 | 89.8±0.2 | 51.2±1.0 | 0.788±.003 |
| +CS-TTA | **82.4**±0.3 | **47.1**±0.3 | **80.8**±0.4 | **68.5**±0.8 | **90.3**±0.2 | **56.8**±0.9 | **0.802**±.003 |

CSD is $-0.14/+0.21$ vs. $-0.15/+0.22$ with frozen CAVs). The signed CSD pattern is therefore not explained by coordinate rotation alone; full per-diagnostic and per-benchmark breakdown is in Supplementary Section D. We emphasize that CSD remains a drift diagnostic, not a causal certificate—TCAV sensitivity is correlational [15].

### 4.3 Results of CS-TTA

Table 2 presents results across all five benchmarks; adding CS-TTA consistently improves all baselines, with the largest gains on spurious-correlation and cross-hospital settings (Waterbirds, CelebA, and CheXpert→MIMIC-CXR).

**Worst-group accuracy.** Tent, EATA, and CoTTA—which optimize entropy or pseudo-label consistency without considering feature quality—*decrease* worst-group accuracy relative to the source model: CoTTA drops WG from 58.1% to 55.1% on Waterbirds, and Tent drops it to 56.2%. SAR and DeYO fare better: SAR's sharpness-aware optimization provides modest WG gains (+1.2 points on Waterbirds), while DeYO's shape-awareness filtering achieves larger improvements (+5.7 points). However, even SAR and DeYO exhibit non-trivial CSD (Table 1), indicating that concept sensitivity drift is not limited to methods that degrade WG. Adding CS-TTA improves WG for *all* methods: Tent+CS-TTA raises WG from 56.2 to 65.8 (+9.6 points) and DeYO+CS-TTA reaches 68.5 (+4.7 points), well above the source baseline in every case. These gains are consistent with preserving object-related concept sensitivity: when the model

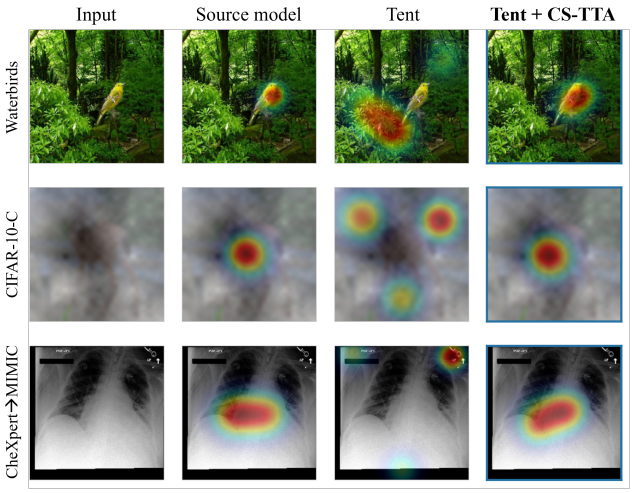


**Fig. 5: Qualitative visualization.** GradCAM attention maps for Source, Tent, and Tent+CS-TTA across three domains. Row 1 (Waterbirds): Tent shifts attention to background; CS-TTA refocuses on the bird. Row 2 (CIFAR-10-C): Tent disperses attention; CS-TTA maintains focus on the object. Row 3 (Chest X-ray): Tent attends to institutional markers; CS-TTA preserves focus on pathological regions. Blue borders indicate CS-TTA.

maintains reliance on *wing shape* rather than *background*, predictions for the minority group (waterbirds on land) improve the most.

**Medical domain.** Cross-hospital adaptation introduces spurious correlations between institutional markers and diagnoses [4]. All five TTA methods improve AUC over the source model, but CSD analysis (Figure 4) reveals that these improvements coincide with increased sensitivity to institutional artifacts. Adding CS-TTA further improves AUC by 1.4–1.8 points across all baselines. The improvement is consistent with preserving sensitivity to pathology-related concepts: when the model maintains reliance on *lung opacity* rather than *patient markers*, predictions generalize better to the target hospital. Full results across all four source datasets (12 source–target pairs) are in the Supplementary material and show consistent AUC gains; fixed-target CSD diagnostics are reported for the CheXpert→MIMIC-CXR Pneumonia setting (Supplementary Section D).

**Overall accuracy.** On CIFAR-10-C and ImageNet-C, CS-TTA provides consistent accuracy improvements of 0.6–1.4 points across all baselines. The gains are smaller than on spurious-correlation benchmarks, which is expected: corruption shifts affect all samples similarly, so concept sensitivity drift has a less concentrated effect on any specific subgroup. We note that CONDA [2] addresses a related but distinct setting, requiring a concept bottleneck architecture; CS-TTA operates as a plug-in on standard models, and the two approaches are complementary.

**Table 3: Ablation study** on Waterbirds (ResNet-50, Tent+CS-TTA).

| Component | Setting | Acc | WG | CSD |
|---|---|---|---|---|
| $\lambda$ | 0.01 | 77.2 | 62.3 | 0.25 |
| | 0.1 (default) | **78.5** | **65.8** | **0.12** |
| | 1.0 | 77.8 | 65.1 | 0.07 |
| | 10.0 | 74.5 | 63.5 | 0.03 |
| $\tau$ | 0.02 | 78.1 | 65.1 | 0.10 |
| | 0.05 (default) | **78.5** | **65.8** | 0.12 |
| | 0.10 | 77.8 | 64.2 | 0.16 |
| Layer | Early (layer2) | 77.2 | 62.8 | 0.18 |
| | Late (layer4, default) | **78.5** | **65.8** | **0.12** |
| | Multi-layer | 78.3 | 65.5 | 0.11 |
| Concept source | Annotation | **78.5** | **65.8** | **0.12** |
| | CLIP pseudo-label | 77.9 | 64.5 | 0.14 |
| Approximation | First-order (default) | **78.5** | 65.8 | 0.12 |
| | Second-order | 78.6 | **66.0** | **0.11** |
| Regularization | CS-TTA (default) | **78.5** | **65.8** | **0.12** |
| | Blind L2 (all activations) | 77.1 | 57.5 | 0.26 |
| | Random concepts | 77.3 | 58.2 | 0.24 |

### 4.4 Ablation Study and Analysis

We conduct ablations on Waterbirds with Tent+CS-TTA (ResNet-50) unless noted otherwise.

**Hyperparameters.** The default $\lambda$=0.1 and $\tau$=0.05 balance accuracy and sensitivity preservation; small $\lambda$ under-regularizes, large $\lambda$=10 over-constrains adaptation, and $\tau$ trades concept coverage for flexibility.

**Layer and concept source.** Late layers (layer4 of ResNet-50) provide the most semantically meaningful CAVs [15]; annotation-based concepts outperform CLIP pseudo-labels by 1.3 points WG, but the gap is modest, indicating CS-TTA is practical without manual annotations. For CheXpert→MIMIC-CXR, replacing radiologist-region positives with source-side CheXpert structured-report keywords yields mean AUC 0.793 vs. 0.795, confirming source-side structured text is a viable substitute when fine-grained annotations are unavailable.

**Approximation and backbone.** The first-order approximation closely matches the second-order variant (78.5/65.8 vs. 78.6/66.0 in Acc/WG) at halved overhead; with ViT-B/16, CS-TTA improves worst-group accuracy by 3–5 points across methods (Supplementary Section B.1).

**Regularization type.** We compare CS-TTA against two controls: (i) blind L2 that penalizes all activation changes without concept selection, and (ii) regularization using randomly selected concepts instead of object/background ones. Neither control improves worst-group accuracy meaningfully (WG: 57.5 and 58.2

vs. 56.2 for Tent alone), while CS-TTA raises it to 65.8, indicating that concept-guided selection—not generic regularization—drives the gain.

**Mechanism controls beyond Waterbirds+Tent.** We extend the controls in Table 3 to Tent on CheXpert→MIMIC-CXR (AUC) and EATA on Waterbirds (WG), and add source-logit KL—penalizing the KL divergence between source-model and adapted-model predictive distributions, a generic functional anchor concept-agnostic but stronger than blind L2. Generic anchoring (blind-L2: 0.781/58.1; random-CAV: 0.776/58.6; source-logit KL: 0.783/59.4 in AUC/WG) modestly stabilizes the adapted model relative to the baselines (0.778/57.8), while +CS-TTA gives the strongest gain in both settings (0.795/66.5); full breakdown is in Supplementary Section E. We phrase this as targeted anchoring beyond generic regularization, not as a universally essential ingredient.

## 5 Discussion

Using TCAV, we examine an aspect of TTA that standard accuracy evaluation does not capture: popular methods can shift sensitivity toward nuisance concepts and away from task-relevant ones in our evaluated settings—a pattern we term Concept Sensitivity Drift. In the medical domain, this manifests as increased reliance on institutional artifacts—the failure mode of DeGrave *et al.* [4] now observed during cross-hospital adaptation. CS-TTA mitigates excessive concept-sensitivity drift as a lightweight plug-in, improving overall and worst-group accuracy without architectural changes.

**Limitations and future work.** CS-TTA requires concept examples to train CAVs; CLIP pseudo-labels reduce this burden but affect effectiveness, and structured-report keywords offer a viable substitute. Not all sensitivity drift may be harmful—only drift above $\tau$ enters $\mathcal{L}_{\text{preserve}}$, and we do not claim recovery from a strongly biased source: when the source profile is sub-optimal, preserving it constrains improvement (over-constrained $\lambda = 10$ row in Table 3). Future work includes distinguishing beneficial drift, concept discovery via ACE [7], and continual TTA.

**Conclusion.** Concept sensitivity drift is a consistent and measurable side effect of TTA that standard evaluation overlooks; CS-TTA offers a simple, effective mitigation that existing TTA methods can adopt.

## Acknowledgements

This work was supported in part by Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University); and No. RS-2025-25442867), by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number: RS-2025-02307233), and by the "Advanced GPU Utilization Support Program" funded by the Government of the Republic of Korea (Ministry of Science and ICT).

## References


1. Bustos, A., Pertusa, A., Salinas, J.M., de la Iglesia-Vaya, M.: PadChest: A large chest x-ray image dataset with multi-label annotated reports. Medical Image Analysis **66**, 101797 (2020)
2. Choi, J., Raghuram, J., Li, Y., Jha, S.: CONDA: Adaptive concept bottleneck for foundation models under distribution shifts. In: Int. Conf. Learn. Represent. (2025)
3. Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M., Bertrand, H.: TorchXRayVision: A library of chest x-ray datasets and models. In: Proceedings of The 5th International Conference on Medical Imaging with Deep Learning. Proceedings of Machine Learning Research, vol. 172, pp. 231–249. PMLR (2022)
4. DeGrave, A.J., Janizek, J.D., Lee, S.I.: AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Mach. Intell. **3**, 610–619 (2021)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Int. Conf. Learn. Represent. (2021)
6. Geirhos, R., Jacobsen, J.H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Mach. Intell. **2**, 665–673 (2020)
7. Ghorbani, A., Wexler, J., Zou, J., Kim, B.: Towards automatic concept-based explanations. In: Adv. Neural Inform. Process. Syst. (2019)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016)
9. He, Z., Sinha, S., Xiong, G., Zhang, A.: GCAV: A global concept activation vector framework for cross-layer consistency in interpretability. In: Int. Conf. Comput. Vis. (2025)
10. Hendrycks, D., Dietterich, T.: Benchmarking neural network robustness to common corruptions and perturbations. In: Int. Conf. Learn. Represent. (2019)
11. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: IEEE Conf. Comput. Vis. Pattern Recog. (2017)
12. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: AAAI (2019)
13. Johnson, A.E.W., Pollard, T.J., Berkowitz, S.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Mark, R.G., Horng, S.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data **6**, 317 (2019)
14. Kang, J., Kim, N., Ok, J., Kwak, S.: Membn: Robust Test-Time adaptation via Batch Norm with statistics memory. In: Eur. Conf. Comput. Vis. (2024)
15. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: Int. Conf. Mach. Learn. (2018)
16. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: Int. Conf. Mach. Learn. (2020)
17. Lee, J., Jung, D., Lee, S., Park, J., Shin, J., Hwang, U., Yoon, S.: Entropy is not enough for test-time adaptation: From the perspective of disentangled factors. In: Int. Conf. Learn. Represent. (2024)
18. Lim, H., Choi, J., Choo, J., Schneider, S.: Sparse autoencoders reveal selective remapping of visual concepts during adaptation. In: Int. Conf. Learn. Represent. (2025)

19. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Int. Conf. Comput. Vis. (2015)
20. Nichol, A., Achiam, J., Schulman, J.: On first-order meta-learning algorithms. arXiv preprint arXiv:1803.02999 (2018)
21. Niu, S., Wu, J., Zhang, Y., Chen, Y., Zheng, S., Zhao, P., Tan, M.: Efficient test-time model adaptation without forgetting. In: Int. Conf. Mach. Learn. (2022)
22. Niu, S., Wu, J., Zhang, Y., Wen, Z., Chen, Y., Zhao, P., Tan, M.: Towards stable test-time adaptation in dynamic wild world. In: Int. Conf. Learn. Represent. (2023)
23. Sagawa, S., Koh, P.W., Hashimoto, T.B., Liang, P.: Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In: Int. Conf. Learn. Represent. (2020)
24. Schmalwasser, L., Penzel, N., Denzler, J., Niebling, J.: FastCAV: Efficient computation of concept activation vectors for explaining deep neural networks. In: Int. Conf. Mach. Learn. (2025)
25. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: Int. Conf. Comput. Vis. (2017)
26. Shih, G., Wu, C.C., Halabi, S.S., Kohli, M.D., Prevedello, L.M., Cook, T.S., Sharma, A., Amorosa, J.K., et al.: Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. Radiology: Artificial Intelligence **1**(1), e180041 (2019). https://doi.org/10.1148/ryai.2019180041
27. Song, J., Lee, J., Kweon, I.S., Choi, S.: Ecotta: Memory-efficient continual test-time adaptation via self-distilled regularization. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
28. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. In: Int. Conf. Learn. Represent. (2021)
29. Wang, Q., Fink, O., Van Gool, L., Dai, D.: Continual test-time domain adaptation. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
30. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: IEEE Conf. Comput. Vis. Pattern Recog. (2017)
31. Yüksekgonül, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. In: Int. Conf. Learn. Represent. (2023)
32. Zhang, M., Levine, S., Finn, C.: Memo: Test time robustness via adaptation and augmentation. In: Adv. Neural Inform. Process. Syst. (2022)