# A COMMITTEE OF ONE: DETECTING AND CORRECTING LABEL NOISE WITH TEST-TIME AUGMENTATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Label noise can substantially degrade the stability and generalization of deep classifiers. Rather than designing a noise-robust training procedure, we address the problem at the data level by identifying and correcting corrupted labels before training the final model. We propose a data-cleaning framework based on a Siamese network and test-time augmentation. The network is trained once on the noisy dataset using a joint contrastive and classification objective and is then frozen. For each training sample, we generate views from a fixed set of twelve augmentation families and measure how often their predictions disagree with the recorded label. Our key insight is that mislabeled samples are more sensitive to small input perturbations, particularly near decision boundaries, whereas clean samples with larger margins remain more stable. We use this augmentation-induced disagreement to identify suspicious samples, relabel them to the most frequently predicted class when predictions are sufficiently consistent, and remove them otherwise. We further show that the gap in expected disagreement splits into a margin term, non-negative whenever clean samples dominate in sensitivity-normalized margin, and an excess we measure to be positive. The detector is trained once and never retrained; only threshold selection trains a downstream classifier. Across CIFAR-10, CIFAR-100, and Fashion-MNIST with synthetic instance-dependent noise, our approach achieves detection AUROC values ranging from $0.907$ to $0.994$. On real annotation noise, a standard classifier trained on our cleaned labels reaches $93.90 \pm 0.11$ accuracy on CIFAR-10N, compared with $92.06 \pm 0.12$ for the strongest of nine single-model robust-training baselines evaluated together in a published common-codebase study, while matching its performance on CIFAR-100N. On Animal-10N, it achieves $87.11 \pm 0.44$, compared with $84.1$ for the strongest of the baselines we compare against.

## 1 INTRODUCTION

Supervised classification typically assumes that every recorded training label is correct. In practice, however, labels are often corrupted by annotator mistakes or unreliable automated pipelines, especially in datasets collected through crowdsourcing or web scraping. Early studies modeled this *label noise* as uniformly random (Angluin & Laird, 1988; Natarajan et al., 2013), but real annotation errors are rarely uniform. They tend to occur on difficult or ambiguous examples, or on examples that require expertise the annotator does not have; even experienced domain experts can disagree on such cases (Schmarje et al., 2022; Lloyd et al., 2004). In this setting, known as *instance-dependent label noise* (IDN), the examples most likely to be mislabeled are often also the most informative. Detecting and correcting these errors is therefore essential (Xia et al., 2020; Cheng et al., 2021; Song et al., 2023).

Most methods for noisy labels either train a model that resists them or flag mislabeled samples using ensembles, cross-validation, or prediction histories (Section 2). Our method needs neither an ensemble nor a stored training history. It trains a Siamese classifier once on the noisy dataset and then freezes it. For each training sample, the model predicts the classes of several augmented views and counts how many disagree with the recorded label. Clean samples well inside their class regions tend to remain stable, whereas the conflict between a mislabeled sample's visual content and observed label can place it near a decision boundary, making its predictions more sensitive to small perturbations. Samples with high disagreement are relabeled when the views support a consistent



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

alternative class and discarded otherwise. The output is a cleaned dataset rather than a model, so any downstream learner can use it.

Our contributions are as follows:

- We clean labels by counting how many augmented views of one frozen model disagree with each recorded label, then relabel or discard the flagged samples. At a matched cleaning budget this keeps more correct labels and less residual noise than cleaning with the same model's single-pass confidence, and detection improves with each added view (Appendix F).
- We split the expected disagreement gap exactly into a margin term, which stochastic dominance of the sensitivity-normalized margins makes non-negative, and a measured excess; on four benchmark cells the margin term carries $0.79$ to $0.92$ of the gap.
- Across nine synthetic-IDN and two human-noise cells, the disagreement ranking reaches AUROC $0.840$–$0.994$, and cleaning increases the number of correct labels in every evaluable cell. A standard learner on the cleaned data outperforms the strongest reported common-codebase baseline on CIFAR-10N, matches it on CIFAR-100N, and reaches $87.11 \pm 0.44$ on Animal-10N. A label-free threshold rule has a worst-case gap of $0.026$ F1 to the per-cell oracle ceiling.

## 2 Related Work

Learning with noisy labels has been studied through robust training and explicit label-error detection (Song et al., 2023; George et al., 2024). Robust training methods select likely clean samples, as in Co-teaching and DivideMix (Han et al., 2018; Li et al., 2020), or correct unreliable labels during training (Song et al., 2019). UniCon selects clean samples uniformly across classes and adds contrastive learning (Karim et al., 2022). These approaches become harder under instance-dependent noise, where corrupted labels are tied to sample difficulty and resemble genuinely hard examples (Xia et al., 2020; Cheng et al., 2021). This setting better reflects human annotation errors than simple class-conditional noise models (Wei et al., 2022; Chen et al., 2021b). Our goal is different from robust training because we explicitly identify and correct suspicious annotations, producing a cleaned dataset for an arbitrary downstream learner.

Existing detectors obtain evidence against an observed label from several sources. Committee-based methods use disagreement across independently trained models (Brodley & Friedl, 1999), while Confident Learning uses out-of-sample predictions to identify likely label errors (Northcutt et al., 2021). Other methods exploit training dynamics through margins or forgetting events (Pleiss et al., 2020; Toneva et al., 2019). Representation-based approaches instead compare a sample with its neighborhood (Ortego et al., 2021; Wang et al., 2018; Zhu et al., 2022). Among them, SimiFeat also aggregates over augmented views after training, but it votes on the noisy labels of nearby samples in a frozen pretrained feature space (Zhu et al., 2022). Augmented-view consistency has also been used during noisy-label training, including Jo-SRC and DISC (Yao et al., 2021; Li et al., 2023). Our method instead reads the predictions of a classifier trained on the noisy labels themselves. It queries one fixed classifier and uses disagreement across augmented views directly, requiring neither a retrained committee nor per-example training histories. This is related to test-time augmentation, where different views can expose prediction instability (Shanmugam et al., 2021), but here that instability is deliberately retained as a label-error signal rather than suppressed by aggregation.

A limitation of any such post-hoc probe is memorization because deep networks can eventually fit corrupted labels (Zhang et al., 2017; Arpit et al., 2017). Early-learning regularization and bounded predictions have been proposed to reduce this effect (Liu et al., 2020; Wei et al., 2023), while recent work shows that some mislabeled examples can be learned particularly early (Yuan et al., 2025). We mitigate this failure mode using an $\ell_2$-normalized classifier embedding together with the peak-validation checkpoint before scoring augmented views.

## 3 Methodology

We are given a training set whose labels cannot be trusted. Write it as $\mathcal{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$, where $x_i \in \mathcal{X} \subseteq \mathbb{R}^n$ is an input, here an image, and $\tilde{y}_i \in \mathcal{Y} = \{1, \ldots, C\}$ is the label recorded for it in

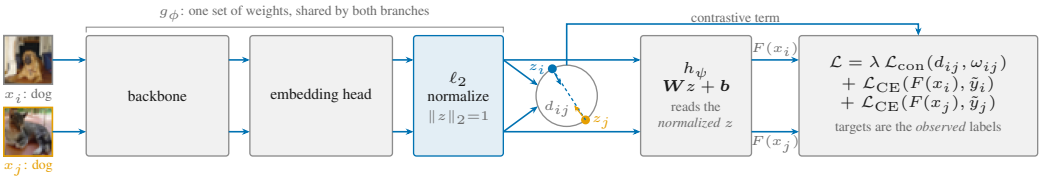



Figure 1: Training the Siamese network on a positive pair, a dog and a cat recorded as a dog. Both observed labels read "dog", so the contrastive term pulls the cat toward the dogs.

the dataset. We call $\tilde{y}_i$ the *observed* label. Behind it stands the true class $y_i$. For most samples the two agree; for some they do not, and those are the ones we want to find. Write $\mathcal{N} = \{i : \tilde{y}_i \neq y_i\}$ for their indices and $\mathcal{C}$ for its complement; neither is known during training, and $y_i$ enters only at evaluation.

The task has two parts. *Detection* flags training samples whose observed labels are likely incorrect, and *cleaning* either corrects or discards the flagged samples. We perform both with a single Siamese network, trained once on the noisy training set under a joint contrastive–classification objective and then frozen. The network scores each sample by classifying twelve augmented views in one batched forward pass and counting how many predictions contradict the observed label. A flagged sample is relabeled when the views agree on an alternative class and discarded when they scatter, yielding a cleaned dataset for any downstream learner.

Section 3.1 describes the network architecture, Section 3.1.1 its training on the observed labels alone, Section 3.2 detection and cleaning after training, and Section 4 analyzes why the disagreement signal separates mislabeled from clean samples.

### 3.1 The Siamese Network

The Siamese architecture processes a pair of inputs through two branches that share all parameters and therefore compute the same function. We first describe the path of a single input $x$ through one branch and then return to the paired formulation.

As shown from left to right in Figure 1, each branch maps $x$ to a unit-norm embedding in three stages:

1. A *backbone encoder*, a pretrained ResNet, maps the image to a feature vector.
2. An *embedding head*, implemented as a small multilayer perceptron, maps the feature vector to an embedding $u \in \mathbb{R}^d$.
3. An *$\ell_2$-normalization layer* rescales the embedding as $z = u/\|u\|_2$, placing it on the unit sphere in $\mathbb{R}^d$.

Write $g_\phi$ for the composition of the three stages, with $\phi$ collecting their parameters, so $z = g_\phi(x)$. A *linear classifier* $h_\psi(z) = \boldsymbol{W}z + \boldsymbol{b}$, with weights $\boldsymbol{W} \in \mathbb{R}^{C\times d}$ and bias $\boldsymbol{b} \in \mathbb{R}^C$, reads that embedding and outputs one score per class. Those scores are the logits, so the full map from image to logits is $F(x) = h_\psi(g_\phi(x)) \in \mathbb{R}^C$, and the predicted class is the one with the largest logit, $\hat{y}(x) = \arg\max_c F_c(x)$.

The classifier reads the *normalized* embedding rather than the raw backbone features, which closes one route to memorization. A network that has fit a wrong label reproduces it under every view, and the detector of Section 3.2 sees nothing. The $c$-th logit is $F_c(x) = \boldsymbol{W}_c z + b_c$, with $\boldsymbol{W}_c$ the $c$-th row of $\boldsymbol{W}$; since $\|z\|_2 = 1$ it can never exceed $\|\boldsymbol{W}_c\|_2 + |b_c|$, and weight decay keeps $\|\boldsymbol{W}_c\|_2$ small. A network can fit a wrong label by inflating the norm of the feature vector, which lets the logit of the wrong class grow without bound (Zhang et al., 2017; Liu et al., 2020; Wei et al., 2023). On the unit sphere that route is bounded. What remains is to move $z$ across the sphere into the region the classifier assigns it, and Section 3.1.1 explains why training on the clean majority resists that move. Appendix F measures what removing the bound costs.

Training operates on pairs. A pair $(x_i, x_j)$ is two samples drawn from $\mathcal{D}$ and sent through the two tied-weight branches, with $\omega_{ij} = \mathbf{1}[\tilde{y}_i = \tilde{y}_j]$ recording whether their observed labels agree.

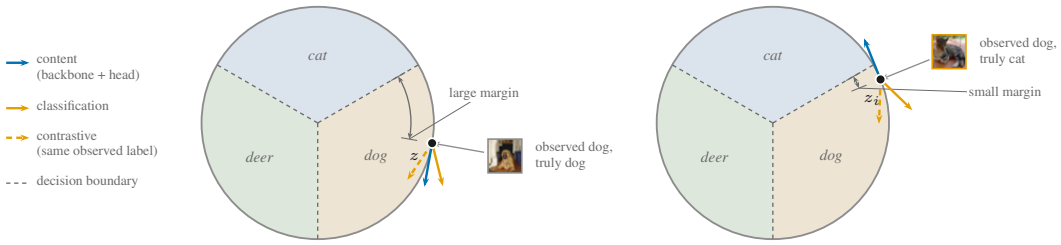



Figure 2: The pair of Figure 1 on the unit circle of a $d = 2$ embedding. On the left the three demands agree; on the right content pulls toward the cats and both label terms toward the dogs, leaving a small margin.

Contrastive learning calls a pair with $\omega_{ij} = 1$ a *positive pair* and one with $\omega_{ij} = 0$ a *negative pair*. The observed labels decide this, so the dog and the cat of Figure 1, both recorded "dog", count as a positive pair.

### 3.1.1 Contrastive Training Objective

Each pair $(x_i, x_j)$ gives two embeddings $z_i = g_\phi(x_i)$ and $z_j = g_\phi(x_j)$ and two logit vectors $F(x_i)$ and $F(x_j)$, from the shared network of Section 3.1. Write $d_{ij} = \|z_i - z_j\|_2$ for the distance between the embeddings; both lie on the unit sphere, so $d_{ij} \in [0, 2]$. Two loss terms act on these.

1. A *contrastive* term acts on the distance. For a positive pair it is $d_{ij}^2$, so minimizing it pulls the two embeddings together. For a negative pair it is $\max(0, \gamma - d_{ij})^2$, which pushes them apart until they are at least $\gamma$ apart. We set $\gamma = 2$, the largest distance two points on the sphere can reach, so the term stays active on essentially every negative pair and spreads the classes over the sphere throughout training.
2. A *classification* term acts on the logits. It is the cross-entropy $\mathcal{L}_{\text{CE}}$ of each branch's logits against that branch's observed label.

The loss of the pair adds the two terms with a balance $\lambda > 0$,

$$\mathcal{L} \;=\; \lambda \underbrace{\big[\, \omega_{ij}\, d_{ij}^2 + (1 - \omega_{ij}) \max(0,\, \gamma - d_{ij})^2 \,\big]}_{\mathcal{L}_{\text{con}}} \;+\; \underbrace{\mathcal{L}_{\text{CE}}(F(x_i),\, \tilde{y}_i) + \mathcal{L}_{\text{CE}}(F(x_j),\, \tilde{y}_j)}_{\mathcal{L}_{\text{cls}}}. \qquad (1)$$

The objective itself is established for learning representations under label noise (Ortego et al., 2021; Wang et al., 2018). We adopt it for what it does to a mislabeled sample. Take a cat image $x_i$ recorded as "dog" and follow the three demands the loss places on it, drawn in Figure 2. The classification term asks the classifier to output "dog" for $x_i$. Every positive pair containing $x_i$ pairs it with another sample recorded as "dog", and since most of those are real dogs, the contrastive term pulls $z_i$ toward the dog region of the sphere. The backbone and embedding head, however, are shared with every other sample and are shaped mostly by the clean majority. They learn to place cat-like content in the cat region, so they keep placing $z_i$ near the cats. For a correctly labeled sample the three demands agree. For the mislabeled cat they conflict, because its label pulls $z_i$ toward the dogs while its content holds $z_i$ near the cats. The compromise leaves $z_i$ between the two regions, close to the boundary the classifier draws between "cat" and "dog". A sample that sits there is the kind whose perturbed copies most often land on the wrong side of that boundary relative to its observed label, which is what the detector of Section 3.2 counts. Section 4 turns the resulting margin deficit into a measurable quantity. Appendix F.2 tests the contrastive term by removing it.

To train the network we first split the noisy training set, with stratification, into a $90\%$ training part and a $10\%$ validation part. From the training part we draw $P$ pairs once, half of them positive and half negative, and apply strong augmentation to both inputs of every pair. The network minimizes Equation 1 over these pairs with AdamW under cosine annealing. During training we track the accuracy on the validation part, measured against its observed and therefore noisy labels, and at the end we restore the checkpoint at which that accuracy peaked. Empirically, validation accuracy rises

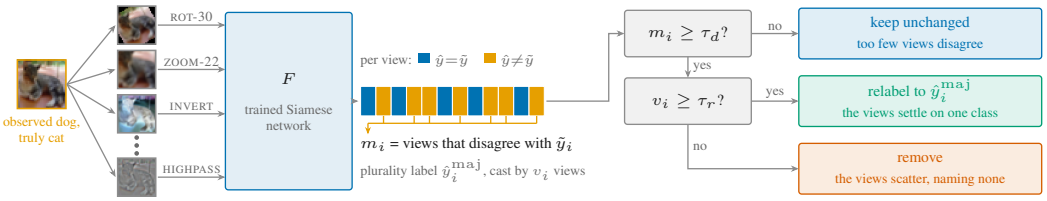



Figure 3: We freeze the network of Figure 1 to score each sample. It predicts a class for each of the $K = 12$ views and lets both thresholds apply post hoc (Equation 2).

while the network fits the clean majority, which generalizes, and decays once the network starts to memorize mislabeled samples, so the restored checkpoint precedes the strongest memorization. Section 5.3 reports the memorization the restored checkpoint still carries. Appendix G gives this procedure as pseudocode.

### 3.2 TTACLEANER: SINGLE-MODEL DETECTION AND CLEANING

With the network $F$ of Section 3.1.1 frozen, each training sample needs one decision, whether to keep its label, replace it, or drop the sample. TTACLEANER makes that decision in three steps, drawn in Figure 3 and applied to each $(x_i, \tilde{y}_i)$ on its own.

1. A fixed set of $K = 12$ label-preserving view families, one transform $t_1, \ldots, t_K$ from each, listed with their parameters in Appendix E, turns $x_i$ into $K$ views of itself, and one batched forward pass through $F$ predicts a class for each view. Scoring covers all of $\mathcal{D}$, including the 10% validation part the network never trained on.
2. The *TTA mistakes count* $m_i = \sum_{k=1}^{K} \mathbf{1}[\hat{y}(t_k(x_i)) \neq \tilde{y}_i]$ counts the views that disagree with the observed label. The *plurality prediction* $\hat{y}_i^{\text{maj}} = \arg\max_c \#\{k : \hat{y}(t_k(x_i)) = c\}$ is the class the views name most often, and $v_i$ is how many of them name it.
3. Two integer thresholds read those numbers. The first, $\tau_d$, sets how much disagreement makes a sample suspect; the second, $\tau_r$, sets how much agreement a replacement label needs. Writing $\emptyset$ for the sample leaving the training set,

$$(x_i, \tilde{y}_i) \longmapsto \begin{cases} (x_i,\ \tilde{y}_i) & m_i < \tau_d & \textit{keep} \\ (x_i,\ \hat{y}_i^{\text{maj}}) & m_i \geq \tau_d,\ v_i \geq \tau_r & \textit{relabel} \\ \emptyset & m_i \geq \tau_d,\ v_i < \tau_r & \textit{discard} \end{cases} \tag{2}$$

The keep row is the whole detector, and $m_i$ is deliberately its sole signal (Section 4 analyzes why it separates mislabeled samples from clean ones). The other two rows are what makes this cleaning and not sample selection or relabeling alone. When a flagged sample's views scatter, the model names nothing to put in place of $\tilde{y}_i$, so we drop the sample instead of guessing. Appendix G gives the pseudocode, Section 5.1 the grid search that sets both thresholds, and Appendix J a cheaper label-free rule.

## 4 ANALYSIS

Sections 3.1–3.2 give the full procedure; this section analyzes the disagreement signal it relies on. We split the expected disagreement gap exactly into a margin term, which dominance of the normalized margins makes non-negative, and a measured excess.

**Setup.** Let $F : \mathcal{X} \to \mathbb{R}^C$ be the trained logit map of Section 3.1.1, with prediction $\hat{y}$ as defined in Section 3.1. Write $M_{\tilde{y}}(x) = F_{\tilde{y}}(x) - \max_{c \neq \tilde{y}} F_c(x)$ for the margin at the observed label. It is positive exactly when the model agrees with $\tilde{y}$ and negative when another class has a larger logit. The margin alone does not say how stable that decision is, because the same margin can sit in regions of very different sensitivity. Linearizing the margin at $x$ gives

$$M_{\tilde{y}}(x + \delta) \approx M_{\tilde{y}}(x) + \nabla_x M_{\tilde{y}}(x)^\top \delta,$$

so the smallest $\delta$ that drives the linearized margin to zero has norm $|M_{\tilde{y}}(x)|/\|\nabla_x M_{\tilde{y}}(x)\|$. Keeping the sign gives the *sensitivity-normalized margin*

$$s(x) \;=\; \frac{M_{\tilde{y}}(x)}{\|\nabla_x M_{\tilde{y}}(x)\|}, \tag{3}$$

the signed first-order distance from $x$ to the surface $M_{\tilde{y}} = 0$ (Moosavi-Dezfooli et al., 2016), read at the observed label rather than at the predicted class. A large positive $s$ puts the observed-label decision well clear of the boundary; a negative one means the model already prefers another class. The ratio also cancels the logit scale, since multiplying every logit by a positive constant scales numerator and denominator alike. Equation 3 measures stability and not just margin size. If the exact gradient norm is replaced by a Lipschitz bound on the margin, the same ratio becomes a certified radius, a distance inside which no perturbation can change the prediction (Hein & Andriushchenko, 2017; Tsuzuku et al., 2018). That bound has to be estimated and is conservative, so we keep the exact gradient norm and read $s$ as a first-order distance rather than a guarantee; Appendix B traces the construction and gives the relation exactly. The numerator is a forward pass and the denominator one backward pass, so $s$ is computed exactly rather than estimated. It is defined wherever $F$ is differentiable at $x$, the strongest competing class is unique, and the gradient does not vanish; Lemma 1 gives the gradient there. We treat these as regularity conditions on the trained network. Test-time augmentation applies one transform $t_k$ from each of the $K = 12$ view families of Appendix E, eight fixed and four drawn at random, and $m(x)$ counts the views whose prediction differs from $\tilde{y}$, the mistakes count $m_i$ of Section 3.2. Its mean rate over the random draws, $D(x) = \mathbb{E}[m(x)]/K$, is the disagreement rate under the uniform mixture $\mathcal{T}$ of the families. We write $Q_{\text{clean}}$ and $Q_{\text{noisy}}$ for the CDFs of $s(x)$ over $\mathcal{C}$ and $\mathcal{N}$.

**Assumption 1** (Dominance of normalized margins)**.** *The clean normalized margins first-order stochastically dominate the noisy ones: $Q_{\text{clean}}(t) \le Q_{\text{noisy}}(t)$ for every $t$. The two distributions may overlap; dominance implies $\mu_{\text{clean}} \ge \mu_{\text{noisy}}$ for the means but is not implied by it.*

**Assumption 2** (Monotone clean response)**.** *There is a non-increasing $m_{\mathcal{C}} : \mathbb{R} \to [0, 1]$ with $\mathbb{E}[D(x) \mid s(x),\, x \in \mathcal{C}] = m_{\mathcal{C}}(s(x))$.*

Assumption 1 is where label noise enters, and $s$ is its most direct form. Mislabeled samples carry smaller margins at their observed labels (Pleiss et al., 2020; Lin & Bradic, 2021) and sit in regions of elevated perturbation sensitivity (Arpit et al., 2017), shrinking the numerator of Equation 3 and inflating its denominator. Neither assumption uses the linearization behind $s$, so the views may be as large as invert or rot-30.

**Theorem 1** (Gap decomposition)**.** *The gap $\Delta D = \mathbb{E}_{x\in\mathcal{N}}[D(x)] - \mathbb{E}_{x\in\mathcal{C}}[D(x)]$ splits exactly, for any extension of $m_{\mathcal{C}}$ beyond the clean margins, as $\Delta D = G + X$ into the margin term $G = \mathbb{E}_{x\in\mathcal{N}}[m_{\mathcal{C}}(s(x))] - \mathbb{E}_{x\in\mathcal{C}}[m_{\mathcal{C}}(s(x))]$ and the excess $X = \mathbb{E}_{x\in\mathcal{N}}[D(x) - m_{\mathcal{C}}(s(x))]$. Under Assumptions 1 and 2, $G \ge 0$, and for every $a < b$ the window bound $G \ge \big(m_{\mathcal{C}}(a) - m_{\mathcal{C}}(b)\big)\big(Q_{\text{noisy}}(a) - Q_{\text{clean}}(b^-)\big)$ holds. Without Assumption 2, both statements hold with $m_{\mathcal{C}}$ replaced by its closest non-increasing fit (Lemma 4), so Assumption 1 alone suffices. Conversely, if $Q_{\text{clean}}(t) > Q_{\text{noisy}}(t)$ for some $t$, the response $R(x) = \mathbf{1}[s(x) \le t]$ satisfies Assumption 2 with $R$ in place of $D$ and has a negative gap.*

In words, $G$ is the part of the gap that the margins alone explain. It is the gap we would see if every mislabeled sample flipped exactly as often as a clean sample with the same $s$. The excess $X$ is the rest, the extra fragility of mislabeled samples at the same margin. The split itself holds by construction. What Theorem 1 adds is that $G$ cannot be negative, and that dominance is exactly what guarantees this. The assumptions do not sign $X$, so we measure it. The window bound says how large $G$ must be. If the clean response falls from $m_{\mathcal{C}}(a)$ to $m_{\mathcal{C}}(b)$ between two margins $a < b$, then $G$ is at least that fall times the share of noisy samples with $s \le a$ minus the share of clean samples with $s < b$. The bound is largest around $s = 0$, where the model switches from disagreeing to agreeing with the observed label; Lemma 2 ties the two shares there to the clean training accuracy and to the fraction of mislabeled samples the model fits to their wrong label. The same split holds for the detector itself. For the flag $\mathbf{1}[m(x) \ge \tau]$ the gap is the true-positive rate minus the false-positive rate at $\tau_d = \tau$, and $\text{AUROC}(m) - \frac{1}{2}$ is a weighted sum of these gaps with non-negative weights (Lemma 5). Appendix C gives the proofs.

We measure every term on 2,000-sample subsets of the training sets of four cells, one with human noise (Appendix D, Tables 5 and 6). Dominance holds exactly on every subset. The measured clean

response is not perfectly monotone, so we replace it by its closest non-increasing fit, computed by isotonic regression. Lemma 4 shows that the split stays exact after this replacement and that $G$ stays non-negative. On the four cells the margin term explains $0.79$ to $0.92$ of the measured gap, the window bound alone guarantees $0.31$ to $0.63$ of it, and the excess is positive. For the detector, the true-positive rate exceeds the false-positive rate at all twelve thresholds on every cell.

# 5 EXPERIMENTS

Each result below tests one link in a chain. Augmentation disagreement separates mislabeled samples from clean ones, acting on that separation yields a dataset with more correct labels than it started with, and a plain learner trained on that dataset matches the single-model robust-training baselines of a common-codebase study.

## 5.1 EVALUATION PROTOCOL

Noisy is the positive class throughout, so we report precision, recall, and F1 rather than accuracy or macro F1, since realistic noise rates leave the classes imbalanced and those averages would flatter the result. We also report AUROC (Fawcett, 2006), a threshold-free, prevalence-independent measure of the same separation, the probability that a random noisy sample outranks a random clean one. Cleaning is scored by the *residual noise ratio*, the fraction of kept samples whose final label still differs from ground truth. The *memorization ratio* (Lin & Bradic, 2021) is the fraction of truly-noisy trained-on samples that the scoring checkpoint still classifies as their observed wrong label; it separates architectural variants.

Both thresholds are chosen per cell by one grid search over the pairs at or above the floor $\lceil K/2 \rceil = 6$. We clean at each candidate, train a downstream classifier for a few epochs, and keep the pair with the best accuracy on the noisy validation split, so no clean label enters selection. Every $(\tau_d, \tau_r)$ we report is that pair. The criterion is downstream accuracy, not detection F1, so the pair is usually not the F1 maximum; Table 12 gives that maximum beside it. Appendix J gives a cheaper rule reading the disagreement histogram alone, exact on five cells and with a worst-case gap of $0.026$ to the F1 ceiling elsewhere. Only Table 2 uses that rule, and it says so.

## 5.2 SETUP

These metrics require knowing each sample's true label, which a naturally noisy dataset cannot supply. We therefore inject synthetic instance-dependent noise (IDN) into CIFAR-10, CIFAR-100 (Krizhevsky, 2009) and Fashion-MNIST (Xiao et al., 2017) at nominal rates of 20%, 30% and 40%, following Xia et al. (2020). Each sample draws its own flip probability from a truncated normal (s.d. $0.1$), and a per-class random projection picks the wrong class. Appendix H.3 adds symmetric noise.

We complement this with real noise. CIFAR-10N and CIFAR-100N (Wei et al., 2022) re-annotate CIFAR-10/100 by crowdsourcing while keeping the original clean label, giving $9.0\%$ and $40.2\%$ noise. Animal-10N (Song et al., 2019) has no clean-label reference for its training set, so it is excluded from Table 1; its clean test split still supports the downstream accuracy of Section 5.4, and Appendix J uses the cell for label-free threshold selection.

Since our output is a cleaned dataset, not a trained model, we compare against methods that instead train robustly on noisy data, taking those numbers from Englesson & Azizpour's (2024) common-codebase evaluation of nine robust-learning methods and cross-entropy on CIFAR-10N/100N. For detection, Table 2 compares against five published detectors, and Figure 6 in Appendix H sweeps every threshold.

## 5.3 DETECTION AND CLEANING

**Detection.** A single scoring pass reaches F1 $0.939$–$0.956$ on all six ten-class synthetic cells (peaking on Fashion-MNIST) from twelve augmented views. Against five published detectors on the four cells whose noise generator and label set match ours, our label-free operating point leads the strongest, DynaCor, by $2.7$, $1.7$ and $1.5$ F1 points on CIFAR-10 IDN 40%, CIFAR-100 IDN 40% and CIFAR-10N, and trails it by $7.3$ on CIFAR-100N (Table 2).

Table 1: Detection and cleaning at the operating point of Section 5.1. F1 is measured there, not at the per-cell maximum (Appendix H gives both); $\Delta$ correct is the change in retained correct labels against no cleaning.

| Benchmark | Type | $\tau_d/\tau_r$ | F1 | Noise (%) | $\Delta$ correct | Kept (%) |
|---|---|---|---|---|---|---|
| CIFAR-10 IDN 20% | Synth. | 10/6 | 0.939 | $20.5 \to 2.93$ | +8,498 | 99.4 |
| CIFAR-10 IDN 30% | Synth. | 10/6 | 0.949 | $29.9 \to 4.11$ | +12,501 | 99.2 |
| CIFAR-10 IDN 40% | Synth. | 9/6 | 0.942 | $40.0 \to 5.77$ | +16,263 | 98.2 |
| CIFAR-100 IDN 20% | Synth. | 12/6 | 0.789 | $20.8 \to 8.63$ | +4,287 | 96.1 |
| CIFAR-100 IDN 30% | Synth. | 12/6 | 0.789 | $30.2 \to 13.10$ | +5,785 | 93.6 |
| CIFAR-100 IDN 40% | Synth. | 11/6 | 0.801 | $40.2 \to 17.02$ | +6,646 | 88.0 |
| Fashion-MNIST IDN 20% | Synth. | 10/6 | 0.956 | $20.6 \to 2.46$ | +10,748 | 99.7 |
| Fashion-MNIST IDN 30% | Synth. | 9/6 | 0.956 | $30.0 \to 3.33$ | +15,629 | 99.4 |
| Fashion-MNIST IDN 40% | Synth. | 9/6 | 0.940 | $40.0 \to 6.56$ | +19,541 | 99.1 |
| CIFAR-10N | Real | 9/6 | 0.724 | $9.0 \to 4.60$ | +1,964 | 99.5 |
| CIFAR-100N | Real | 10/6 | 0.676 | $40.2 \to 28.42$ | +1,121 | 86.7 |

Table 2: Detection F1 (%) against published detectors, noisy positive, on the four cells whose noise generator and label set match ours. Our row uses the label-free $\tau_d$ of Appendix J; † from Kim et al. (2024), and Appendix H gives the other sources. **Bold**: best per column.

| Method | CIFAR-10 IDN 40% | CIFAR-100 IDN 40% | CIFAR-10N (Aggr.) | CIFAR-100N |
|---|---|---|---|---|
| CORES (Cheng et al., 2021) | 49.2 | 39.2 | 63.6 | 31.9 |
| AUM (Pleiss et al., 2020) | 57.7 | 57.8 | 16.7 | 58.0 |
| Confident learning (Northcutt et al., 2021) | 82.5 | 67.3 | 57.0 | 65.2 |
| SimiFeat-V (Zhu et al., 2022) | 81.1 | 74.3 | 69.4† | 73.2 |
| SimiFeat-R (Zhu et al., 2022) | 82.3 | 73.5 | 68.8† | 74.7 |
| DynaCor (Kim et al., 2024) | 91.5 | 79.5 | 72.6 | **77.3** |
| Ours (label-free $\tau_d$) | **94.2** | **81.2** | **74.1** | 70.0 |

Class count matters more than noise level. CIFAR-100 spans only 0.789–0.801 against 0.939–0.956 on the ten-class benchmarks, a recall gap rather than a precision one (Table 12). Clean samples sit closer to a boundary when a hundred classes divide the space, so they disagree more often too, and holding precision at 0.84 to 0.87 costs a $\tau_d$ of 11 or 12, which caps recall at 0.72 to 0.76.

Real human noise is harder. Read at each cell's best F1, the score falls from 0.950 on CIFAR-10 IDN 30% to 0.741 on CIFAR-10N. Synthetic IDN flips a label to an arbitrary target; CIFAR-10N's human errors are plausible confusions (truck $\to$ automobile, cat $\to$ dog) that the model predicts anyway, so views agree with the wrong label instead of disagreeing. The memorization ratio supports this, at 0.376/0.381 on the two real benchmarks against 0.014–0.163 synthetically, an eightfold rise over the CIFAR-10 IDN 30% cell (0.046).

**Cleaning.** Every cell in Table 1 ends with more correct labels than it began with, discarding at most 13.3% of the data; Fashion-MNIST IDN 40% gains the most. Discarding every flagged sample instead would lose correct labels on every cell, so the gain comes from relabeling (Appendix F.5). At the same flagged and relabeled counts, cleaning with the model's single-pass confidence keeps fewer correct labels and more noise on the four replayed cells (Appendix F.6). Figure 4 shows the composition before and after cleaning.

### 5.4 Downstream accuracy on human-annotated noise

On CIFAR-10N the cleaned set reaches $93.90 \pm 0.11$, 1.84 points above the strongest published entry ($92.06 \pm 0.12$) and clearly separable from it. On CIFAR-100N it reaches $60.52 \pm 0.38$ against $60.36 \pm 0.71$, a 0.16-point difference indistinguishable from noise (Appendix I gives both significance tests). Animal-10N carries no clean training label, so the downstream number is all that is measurable there. At $\tau_d/\tau_r = 7/10$, which keeps 44,477 of its 50,000 images, the cleaned set reaches $87.11 \pm 0.44$

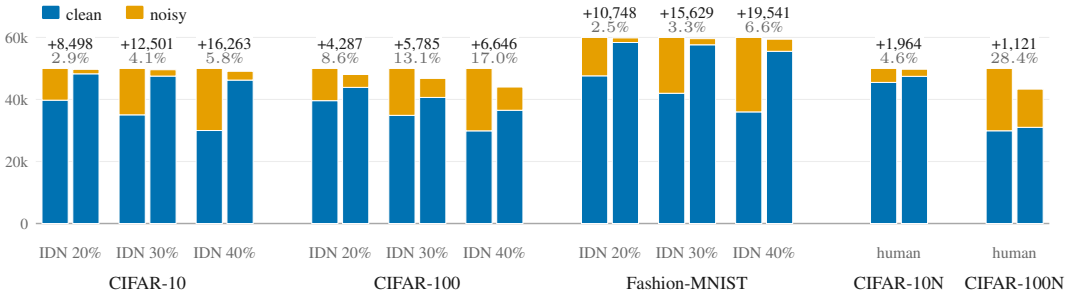



Figure 4: Training set clean-noisy distribution before and after the cleaning process. Each pair of bars is one cell, counted in training images, with before cleaning on the left and after on the right. Printed above each pair are the change in correctly labeled images and the residual noise rate.

Table 3: Test accuracy (%) on CIFAR-N. **Bold**: best mean per column and anything not separable from it (Welch, $p > 0.05$). Shown are CE and the five strongest methods on CIFAR-100N; the other four are in Appendix I.

| **Method** | **Intervention** | **CIFAR-10N (Aggr.)** | **CIFAR-100N** |
|---|---|---|---|
| *Reference* (Englesson & Azizpour, 2024) | | | |
| CE | — | $83.59 \pm 0.98$ | $42.75 \pm 0.93$ |
| LN (Englesson et al., 2023) | noise model | $85.35 \pm 1.33$ | $50.37 \pm 0.50$ |
| ELR (Liu et al., 2020) | regularization | $89.61 \pm 0.12$ | $54.91 \pm 1.11$ |
| SOP (Liu et al., 2022) | label correction | $89.54 \pm 0.57$ | $50.20 \pm 0.84$ |
| NAL (Lu et al., 2022) | loss reweighting | $91.30 \pm 0.16$ | $56.33 \pm 1.21$ |
| SGN (Englesson & Azizpour, 2024) | reweight + correct | $92.06 \pm 0.12$ | $\mathbf{60.36 \pm 0.71}$ |
| **Ours** (clean, then train) | dataset | $\mathbf{93.90 \pm 0.11}$ | $\mathbf{60.52 \pm 0.38}$ |

over three seeds, against $84.1 \pm 0.1$ for Nested Co-teaching (Chen et al., 2021c) and $83.08 \pm 0.39$ for AdaptCDRP (Guo et al., 2024). This comparison and the CIFAR-N one above place the cleaned dataset in a published field without attributing the gap to cleaning; the one arm that does is the uncleaned CIFAR-10 IDN 40% run of Appendix F.4.

## 6 Discussion

One frozen classifier, read under a small set of label-preserving views, gives enough evidence to find and fix label errors. The disagreement count ranks mislabeled samples above clean ones with AUROC 0.951 on CIFAR-10N and 0.840 on CIFAR-100N, where 40% of the labels are wrong. Acting on that ranking leaves every cell with more correct labels than it started with, and most of the data is kept (Table 1). A standard learner trained on the cleaned set then matches or exceeds the single-model robust-training baselines of a common-codebase study (Table 3).

**Limitations.** The probe relies on transformations that preserve the label. Images offer many of them, while text or tabular data would need a perturbation family of their own. The disagreement count is also an integer in $[0, K]$, so with $K = 12$ views the threshold moves in coarse steps.

**Future work.** We see two natural extensions. The first is multi-label classification, where an image can carry several correct labels and an error can be a missing label as well as a wrong one. The disagreement count carries over label by label, counting the views whose prediction for that label contradicts its recorded presence or absence. The correction rule is what has to change, since it must add or remove single labels instead of replacing one class with another. The second is a per-instance threshold. One $\tau_d$ now serves every sample in a cell, yet ambiguous clean samples disagree more often than easy ones, and on CIFAR-100 this forces a high $\tau_d$ that limits recall (Section 5.3). A threshold that adapts to each sample, for example through its margin $s$ or its predicted class, could tolerate the extra disagreement of hard clean samples and still flag mislabeled ones.

### AI USE STATEMENT

We used generative AI to help draft and edit the manuscript text and format the LaTeX source and references, but not to formulate mathematical claims, write proofs, design the methodology or experiments, or generate, clean, or analyze data. All AI-assisted text was reviewed. The authors verified mathematical claims and proofs, and checked citations against their original sources. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

This work uses public benchmarks only. CIFAR-10, CIFAR-100 and Fashion-MNIST are standard released datasets, and the CIFAR-10N/100N and Animal-10N noisy labels are human re-annotations released by Wei et al. (2022) and Song et al. (2019). We collected no new human-subject data, and none of these datasets carries personally identifying information.

We are not aware of any ethical concerns raised by this work.

### REPRODUCIBILITY STATEMENT

The experimental protocol — datasets, noise-injection procedure, evaluation metrics, and threshold selection — is described in Sections 5.1 and 5.2, and the network architecture and its backbone in Section 3.1 and Appendix K, with a label-free thresholding alternative in Appendix J. Notation, the full set of assumptions, and complete proofs of all theoretical claims are in Appendices A and C. Implementation details — the $K = 12$-transform augmentation policy, per-stage training configuration, and per-cell hardware and wall-clock — are in Appendices E, K, and L. The code is available in an anonymized repository at `https://anonymous.4open.science/r/SiameseTTA`.

## REFERENCES

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, Apr 1988. ISSN 1573-0565. doi: 10.1023/A:1022873112823. URL https://doi.org/10.1023/A:1022873112823.

Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 233–242. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arpit17a.html.

Richard E. Barlow, David J. Bartholomew, J. M. Bremner, and H. Daniel Brunk. *Statistical Inference under Order Restrictions: The Theory and Application of Isotonic Regression*. Wiley, London, 1972.

C. E. Brodley and M. A. Friedl. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167, August 1999. ISSN 1076-9757. doi: 10.1613/jair.606. URL http://dx.doi.org/10.1613/jair.606.

Peng Cao, Yilun Xu, Yuqing Kong, and Yizhou Wang. Max-MIG: an information theoretic approach for joint learning from crowds. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=BJg9DoR9t7.

Pengfei Chen, Guangyong Chen, Junjie Ye, Jingwei Zhao, and Pheng-Ann Heng. Noise against noise: stochastic label noise helps combat inherent label noise. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021a. URL https://openreview.net/forum?id=80FMcTSZ6J0.

Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(13):11442–11450, May 2021b. ISSN 2159-5399. doi: 10.1609/aaai.v35i13.17363. URL http://dx.doi.org/10.1609/aaai.v35i13.17363.

Yingyi Chen, Xi Shen, Shell Xu Hu, and Johan A. K. Suykens. Boosting co-teaching with compression regularization for label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2688–2692, June 2021c.

Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=2VXyy9mIyU3.

Zhendong Chu, Jing Ma, and Hongning Wang. Learning from crowds by modeling common confusions. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):5832–5840, May 2021. ISSN 2159-5399. doi: 10.1609/aaai.v35i7.16730. URL http://dx.doi.org/10.1609/aaai.v35i7.16730.

Mark Collier, Basil Mustafa, Efi Kokiopoulou, Rodolphe Jenatton, and Jesse Berent. Correlated input-dependent label noise in large-scale image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1551–1560, June 2021.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, 28(1):20–28, 1979. ISSN 0035-9254. doi: 10.2307/2346806. URL http://dx.doi.org/10.2307/2346806.

Gamaleldin Elsayed, Dilip Krishnan, Hossein Mobahi, Kevin Regan, and Samy Bengio. Large margin deep networks for classification. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/42998cf32d552343bc8e460416382dca-Paper.pdf.

Erik Englesson and Hossein Azizpour. Robust classification via regression for learning with noisy labels. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 7143–7163, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/1cb5b3d64bdf3c6642c8d9a8fbecd019-Paper-Conference.pdf.

Erik Englesson, Amir Mehrpanah, and Hossein Azizpour. Logistic-normal likelihoods for heteroscedastic label noise. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=7wA65zL3B3.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006. ISSN 0167-8655. doi: 10.1016/j.patrec.2005.10.010. URL http://dx.doi.org/10.1016/j.patrec.2005.10.010.

Thomas George, Pierre Nodet, Alexis Bondu, and Vincent Lemaire. Mislabeled examples detection viewed as probing machine learning models: concepts, survey and extensive benchmark. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=3YlOr7BHkx.

Melody Guan, Varun Gulshan, Andrew Dai, and Geoffrey Hinton. Who said what: Modeling individual labelers improves classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2159-5399. doi: 10.1609/aaai.v32i1.11756. URL http://dx.doi.org/10.1609/aaai.v32i1.11756.

Hui Guo, Grace Y. Yi, and Boyu Wang. Learning from noisy labels via conditional distributionally robust optimization. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing*

*Systems*, volume 37, pp. 82627–82672. Curran Associates, Inc., 2024. doi: 10.52202/079017-2627. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/96af549bb29d01b1713e88e5c3899ceb-Paper-Conference.pdf.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/a19744e268754fb0148b017647355b7b-Paper.pdf.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/e077e1a544eec4f0307cf5c3c721d944-Paper.pdf.

Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9676–9686, June 2022.

Ashish Khetan, Zachary C. Lipton, and Animashree Anandkumar. Learning from noisy singly-labeled data. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL https://openreview.net/forum?id=H1sUHgb0Z.

Suyeon Kim, Dongha Lee, SeongKu Kang, Sukang Chae, Sanghwan Jang, and Hwanjo Yu. Learning discriminative dynamics with label corruption for noisy label detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22477–22487, June 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Junnan Li, Richard Socher, and Steven C. H. Hoi. DivideMix: Learning with noisy labels as semi-supervised learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HJgExaVtwr.

Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. DISC: Learning from noisy labels via dynamic instance-specific selection and correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24070–24079, June 2023.

Jason Z. Lin and Jelena Bradic. Learning to combat noisy labels via classification margins, 2021. URL https://arxiv.org/abs/2102.00751.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 20331–20342. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/ea89621bee7c88b2c5be6681c8ef4906-Paper.pdf.

Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 14153–14172. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/liu22w.html.

Ricardo V Lloyd, Lori A Erickson, Mary B Casey, King Y Lam, Christine M Lohse, Sylvia L Asa, John K. C Chan, Ronald A DeLellis, H Ruben Harach, Kennichi Kakudo, Virginia A LiVolsi, Juan Rosai, Thomas J Sebo, Manuel Sobrinho-Simoes, Bruce M Wenig, and Marick E Lae. Observer variation in the diagnosis of follicular variant of papillary thyroid carcinoma. *The American Journal of Surgical Pathology*, 28(10):1336–1340, October 2004. ISSN 0147-5185. doi: 10.1097/01.pas.0000135519.34847.f6. URL http://dx.doi.org/10.1097/01.pas.0000135519.34847.f6.

Yangdi Lu, Yang Bo, and Wenbo He. Noise attention learning: Enhancing noise robustness by gradient scaling. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 23164–23177. Curran Associates, Inc., 2022. doi: 10.52202/068431-1683. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/92864e1191ed272deb0914b3bb50f97c-Paper-Conference.pdf.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/3871bd64012152bfb53fdf04b401193f-Paper.pdf.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411, April 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12125. URL http://dx.doi.org/10.1613/jair.1.12125.

Diego Ortego, Eric Arazo, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Multi-objective interpolation training for robustness to label noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6606–6615, June 2021.

Geoff Pleiss, Tianyi Zhang, Ethan Elenberg, and Kilian Weinberger. Identifying mislabeled data using the area under the margin ranking. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 17044–17056. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/c6102b3727b2a7d8b1bb6981147081ef-Paper.pdf.

Filipe Rodrigues and Francisco Pereira. Deep learning from crowds. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. ISSN 2159-5399. doi: 10.1609/aaai.v32i1.11506. URL http://dx.doi.org/10.1609/aaai.v32i1.11506.

Lars Schmarje, Vasco Grossmann, Claudius Zelenka, Sabine Dippel, Rainer Kiko, Mariusz Oszust, Matti Pastell, Jenny Stracke, Anna Valros, Nina Volkmann, and Reinhard Koch. Is one annotation enough? - A data-centric image classification benchmark for noisy and ambiguous label estimation. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 33215–33232. Curran Associates, Inc., 2022. doi: 10.52202/068431-2407. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/d6c03035b8bc551f474f040fe8607cab-Paper-Datasets_and_Benchmarks.pdf.

Moshe Shaked and J. George Shanthikumar. *Stochastic Orders*. Springer Series in Statistics. Springer, New York, 2007.

Divya Shanmugam, Davis Blalock, Guha Balakrishnan, and John Guttag. Better aggregation in test-time augmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1214–1223, October 2021.

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: Refurbishing unclean samples for robust deep learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th*

*International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5907–5915. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/song19b.html.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, November 2023. ISSN 2162-2388. doi: 10.1109/tnnls.2022.3152527. URL http://dx.doi.org/10.1109/TNNLS.2022.3152527.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Ryutaro Tanno, Ardavan Saeedi, Swami Sankaranarayanan, Daniel C. Alexander, and Nathan Silberman. Learning from noisy labels by regularized estimation of annotator confusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. URL https://openreview.net/forum?id=BJlxm30cKm.

Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/485843481a7edacbfce101ecb1e4d2a8-Paper.pdf.

Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Hongxin Wei, Huiping Zhuang, Renchunzi Xie, Lei Feng, Gang Niu, Bo An, and Yixuan Li. Mitigating memorization of noisy labels by clipping the model prediction. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 36868–36886. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/wei23e.html.

Jiaheng Wei, Zhaowei Zhu, Hao Cheng, Tongliang Liu, Gang Niu, and Yang Liu. Learning with noisy labels revisited: A study using real-world human annotations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=TBWA6PLJZQm.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 7597–7610. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/5607fe8879e4fd269e88387e8cb30b7e-Paper.pdf.

Xiaobo Xia, Bo Han, Yibing Zhan, Jun Yu, Mingming Gong, Chen Gong, and Tongliang Liu. Combating noisy labels with sample selection by mining high-discrepancy examples. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1833–1843, October 2023.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms, 2017. URL https://arxiv.org/abs/1708.07747.

Shuo Yang, Erkun Yang, Bo Han, Yang Liu, Min Xu, Gang Niu, and Tongliang Liu. Estimating instance-dependent Bayes-label transition matrix using a deep neural network. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25302–25312. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/yang22p.html.

Yazhou Yao, Zeren Sun, Chuanyi Zhang, Fumin Shen, Qi Wu, Jian Zhang, and Zhenmin Tang. Jo-SRC: A contrastive approach for combating noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5192–5201, June 2021.

Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. How does disagreement help generalization against label corruption? In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7164–7173. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/yu19b.html.

Suqin Yuan, Lei Feng, Bo Han, and Tongliang Liu. Enhancing sample selection against label noise by cutting mislabeled easy examples. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 43942–43970. Curran Associates, Inc., 2025. doi: 10.52202/085713-1466. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3e826f682178d9830a3b704141b4989e-Paper-Conference.pdf.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Sy8gdB9xx.

Hansong Zhang, Shikun Li, Dan Zeng, Chenggang Yan, and Shiming Ge. Coupled confusion correction: Learning from crowds with sparse annotations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(15):16732–16740, March 2024. ISSN 2159-5399. doi: 10.1609/aaai.v38i15.29613. URL http://dx.doi.org/10.1609/aaai.v38i15.29613.

Yikai Zhang, Songzhu Zheng, Pengxiang Wu, Mayank Goswami, and Chao Chen. Learning with feature-dependent label noise: A progressive approach. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=ZPa2SyGcbwh.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/f2925f97bc13ad2852a7a551802feea0-Paper.pdf.

Zhaowei Zhu, Zihao Dong, and Yang Liu. Detecting corrupted labels without training a model to predict. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27412–27427. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhu22a.html.

## A Table of Notation

Table 4 lists the symbols used in more than one section, in two blocks. The first covers the data, the model and the training objective; the second covers the theory, the detector and threshold selection.

Table 4: Notation.

| Symbol | Meaning |
|---|---|
| *Data, model and training objective* | |
| $\mathcal{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$ | noisy training set, $N$ samples |
| $x_i \in \mathcal{X} \subseteq \mathbb{R}^n$ | an input, of dimension $n$ |
| $y_i$ / $\tilde{y}_i \in \mathcal{Y} = \{1, \dots, C\}$ | latent ground-truth label (evaluation only) / observed label |
| $C$ / $c$ / $c^*$ | number of classes / class index / strongest competing class at $x$ |
| $\mathcal{N}$ / $\mathcal{C}$ | index sets of mislabeled / correctly labeled samples |
| $\hat{y}(x) = \arg\max_c F_c(x)$ / $\hat{y}_i^{\mathrm{maj}}$ | prediction at $x$ / plurality prediction over the $K$ views |
| $F : \mathcal{X} \to \mathbb{R}^C$, $F_c(x)$ | trained logit map $h_\psi \circ g_\phi$, and its $c$-th logit |
| $g_\phi$ / $h_\psi$ | encoder (backbone, embedding head, $\ell_2$ normalization) / linear classifier $h_\psi(z) = \boldsymbol{W}z + \boldsymbol{b}$ |
| $u_i \in \mathbb{R}^d$, $z_i = u_i / \|u_i\|_2$ | embedding before and after $\ell_2$ normalization |
| $\omega_{ij} = \mathbf{1}[\tilde{y}_i = \tilde{y}_j]$ | whether a pair shares its observed label |
| $d_{ij} = \|z_i - z_j\|_2$ | distance between normalized embeddings, in $[0, 2]$ |
| $\gamma$ / $\lambda$ / $P$ | contrastive margin / loss balance / pairs drawn |
| $\mathcal{L}$, $\mathcal{L}_{\mathrm{con}}$, $\mathcal{L}_{\mathrm{cls}}$ | objective of Equation 1 and its two terms |
| *Theory, detection and threshold selection* | |
| $M_{\tilde{y}}(x)$ | margin at the observed label |
| $s(x) = M_{\tilde{y}}(x) / \|\nabla_x M_{\tilde{y}}(x)\|$ | sensitivity-normalized margin, Equation 3 |
| $\mathcal{T}_k$, $\mathcal{T}$, $t_k$ | law of view family $k$, their uniform mixture, and the draw from family $k$ |
| $K = 12$ | number of test-time-augmented views |
| $D(x)$ / $\hat{D}(x)$ | expected disagreement rate over the view draws / its $K$-view estimate $m/K$ |
| $\Delta D$ | expected gap $\mathbb{E}_{x \in \mathcal{N}}[D] - \mathbb{E}_{x \in \mathcal{C}}[D]$ |
| $m_{\mathcal{C}}$ | clean response $\mathbb{E}[D \mid s, \mathcal{C}]$ (Assumption 2) |
| $G$, $X$ | margin term and excess of Theorem 1, $\Delta D = G + X$ |
| $J(\tau)$ | true-positive minus false-positive rate of the flag $\mathbf{1}[m \geq \tau]$ |
| $a_{\mathcal{C}}$, $a_{\mathcal{N}}$ | clean training accuracy and memorization ratio (Lemma 2) |
| $Q_{\mathrm{clean}}$, $Q_{\mathrm{noisy}}$ | CDFs of $s(x)$ over $\mathcal{C}$ and over $\mathcal{N}$ |
| $\mu_{\mathrm{clean}}$, $\mu_{\mathrm{noisy}}$ | the corresponding means (Assumption 1) |
| $r(x)$, $L_x$, $\kappa(x)$ | certified radius, local constant, $2L_x / \|\nabla_x M_{\tilde{y}}\|$ (Appendix B) |
| $m_i = K\hat{D}(x_i)$ / $v_i$ | TTA mistakes count in $[0, K]$ / views casting the plurality label |
| $\tau_d$ / $\tau_r$ | detection / relabel threshold, both floored at $\lceil K/2 \rceil = 6$ |
| $\pi$ / $\hat{\pi}$ | noise rate / the mixture's estimate of it |
| $p_{\mathrm{clean}}$, $p_{\mathrm{noisy}}$ | the two binomial rates of Equation 4 |
| $h[m]$, $\rho$ | disagreement histogram and its bimodality ratio |

It is worth noting that a few letters carry more than one meaning, told apart by typeface, subscript or context. $C$ counts classes, $c$ indexes one, and $\mathcal{C}$ is the correctly labeled index set. The letter $d$ is the embedding dimension and $d_{ij}$ is a distance. $m_i$ and $m(x)$ count disagreeing views, while $m_{\mathcal{C}}$ is the clean response.

## B Where $s$ Comes From

The ratio in Equation 3 predates this paper and has been read two ways. Moosavi-Dezfooli et al. (2016) read it as the first-order minimal perturbation that carries a sample across the decision boundary, and Elsayed et al. (2018) impose it during training as a margin at the input or at a hidden layer, on tasks that include corrupted labels. Both measure from the predicted class, so the quantity is non-negative by construction, and both read the gradient exactly.

The second reading replaces the exact gradient norm by a Lipschitz bound and buys a guarantee with it. Let $L_x$ bound the local Lipschitz constant of $F$ on a ball $B$ around $x$. The observed-label margin is one coordinate of $F$ minus the largest of the others, each $L_x$-Lipschitz on $B$, so it is $2L_x$-Lipschitz on $B$ and $\|\nabla_x M_{\tilde{y}}(x)\| \leq 2L_x$ wherever the gradient exists. Write $r(x) = M_{\tilde{y}}(x)/(2L_x)$. When $M_{\tilde{y}}(x) > 0$, no perturbation that stays in $B$ and has norm below $r(x)$ moves the prediction off $\tilde{y}$ (Hein & Andriushchenko, 2017; Tsuzuku et al., 2018). Equation 3 is that radius with the bound replaced by the exact gradient norm. Writing $\kappa(x) = 2L_x / \|\nabla_x M_{\tilde{y}}(x)\| \geq 1$ for how conservative the bound is, the two satisfy $\kappa(x)\, r(x) = s(x)$ identically, so $r$ carries the same information as $s$ up to a rescaling that differs from sample to sample.

Ours differs on three points. First, we read the margin at the *observed* label, which may be wrong, and we keep its sign; the two readings above measure from the predicted class and never see the negative branch. That branch is where mislabeled samples live. At the memorization ratios of Section 5.3, $M_{\tilde{y}} < 0$ on most of $\mathcal{N}$ on the synthetic cells. Second, $s$ is a statistic we measure, not a training objective and not a certificate; nothing is optimized for it and Theorem 1 claims no guaranteed radius. Third, we keep the exact gradient norm, so $s$ needs no estimate of $L_x$. Because $\kappa$ varies from sample to sample, neither dominance nor a monotone response transfers from $s$ to $r$. The ordering $r(x) \le s(x)$ also holds only where $M_{\tilde{y}}(x) > 0$. Where the model disagrees with the observed label, dividing a negative margin by the larger denominator $2L_x$ returns a less negative number, so $r(x) \ge s(x)$ there. Assumptions 1 and 2 are therefore stated for $s$, which is what Appendices D.1 and D.2 measure.

# C PROOFS

Throughout, $F : \mathcal{X} \to \mathbb{R}^C$ is the trained logit map, held fixed, and expectations run over labeled training samples $(x, \tilde{y})$ drawn from the training distribution and over the random draws of the four stochastic view families. $\mathcal{C}$ and $\mathcal{N}$ are the events that $\tilde{y}$ is correct and wrong, both of positive probability, and $\mathbb{E}_{\mathcal{C}}$, $\mathbb{E}_{\mathcal{N}}$ denote expectations conditioned on them. $M_{\tilde{y}}$ is the observed-label margin of Section 4 and $s(x) = M_{\tilde{y}}(x)/\|\nabla_x M_{\tilde{y}}(x)\|$ the sensitivity-normalized margin of Equation 3, defined off a null set by the regularity conditions of Section 4. We write $S_{\mathcal{C}} \sim Q_{\text{clean}}$ and $S_{\mathcal{N}} \sim Q_{\text{noisy}}$ for $s(x)$ under $\mathcal{C}$ and $\mathcal{N}$, and $Q(t^-) = \lim_{u \uparrow t} Q(u)$ for a left limit. Nothing below needs $Q_{\text{clean}}$ or $Q_{\text{noisy}}$ to be continuous, so ties and atoms in $s$ are allowed, and neither support need contain the other.

**The disagreement rate.** View family $k$ is a law $\mathcal{T}_k$ on label-preserving transforms, a point mass for the eight deterministic families of Appendix E. The detector draws $t_k \sim \mathcal{T}_k$ once per family, independently across families and of $(x, \tilde{y})$, and counts $m(x) = \sum_{k=1}^{K} \mathbf{1}[\hat{y}(t_k(x)) \ne \tilde{y}]$. Writing $p_k(x) = \Pr_{t \sim \mathcal{T}_k}[\hat{y}(t(x)) \ne \tilde{y}]$,

$$D(x) \;=\; \frac{\mathbb{E}[m(x) \mid x, \tilde{y}]}{K} \;=\; \frac{1}{K}\sum_{k=1}^{K} p_k(x) \;=\; \Pr_{t \sim \mathcal{T}}\big[\hat{y}(t(x)) \ne \tilde{y}\big], \qquad \mathcal{T} = \frac{1}{K}\sum_{k=1}^{K} \mathcal{T}_k,$$

so $\hat{D} = m/K$ is a stratified sample of the mixture $\mathcal{T}$ with one draw per stratum. For a deterministic family $p_k(x) \in \{0, 1\}$ and its indicator equals $p_k(x)$ exactly. Given $x$ the indicators are independent, so $\mathrm{Var}[m(x) \mid x, \tilde{y}] = \sum_{k \text{ random}} p_k(x)(1 - p_k(x)) \le 4 \cdot \frac{1}{4} = 1$. Theorem 1 and Lemma 4 use only that the response takes values in $[0, 1]$. They therefore hold for $D$, for the realized rate $\hat{D}$, and for any flag $\mathbf{1}[m(x) \ge \tau]$, and because $s$ is a function of $(x, \tilde{y})$, $\mathbb{E}[\hat{D} \mid s, \mathcal{C}] = \mathbb{E}[D \mid s, \mathcal{C}]$.

## C.1 DIFFERENTIABILITY AND SIGN OF THE MARGIN

The logit map $F$ composes affine layers, ReLU nonlinearities and one $\ell_2$ normalization, so it is continuous and piecewise smooth on $\mathcal{X}$. It is differentiable off the boundaries between regions and off the set where the pre-normalization embedding vanishes. That does not by itself give differentiability of the observed-label margin $M_{\tilde{y}}$, which is a maximum over classes and so need not be differentiable even where all of its arguments are.

**Lemma 1** (Margin differentiability)**.** *Let $F$ be differentiable at $x$ with a unique strongest competing class $c^* = \arg\max_{c \ne \tilde{y}} F_c(x)$. Then $M_{\tilde{y}}$ is differentiable at $x$ with*

$$\nabla_x M_{\tilde{y}}(x) \;=\; \nabla_x F_{\tilde{y}}(x) - \nabla_x F_{c^*}(x).$$

*Proof.* $F$ is continuous, so the strict inequality $F_{c^*}(x) > F_c(x)$ for every $c \notin \{\tilde{y}, c^*\}$ persists on some open neighborhood of $x$; on that neighborhood $c^*$ remains the strongest competing class and

$$M_{\tilde{y}}(x') \;=\; F_{\tilde{y}}(x') - F_{c^*}(x')$$

identically, a difference of two functions each differentiable at $x$. Hence $M_{\tilde{y}}$ is differentiable at $x$ with gradient $\nabla_x F_{\tilde{y}}(x) - \nabla_x F_{c^*}(x)$. □

**Lemma 2** (Sign of the normalized margin)**.** *Wherever $s(x)$ is defined, $s(x) > 0$ if and only if $\tilde{y}$ is the unique largest logit of $F(x)$. Hence $Q_{\mathrm{clean}}(0) = 1 - a_{\mathcal{C}}$ and $Q_{\mathrm{noisy}}(0) = 1 - a_{\mathcal{N}}$, where $a_{\mathcal{C}} = \Pr_{\mathcal{C}}[\hat{y}(x) = \tilde{y}]$ is the training accuracy on clean samples and $a_{\mathcal{N}} = \Pr_{\mathcal{N}}[\hat{y}(x) = \tilde{y}]$ the memorization ratio, with $\hat{y}$ read as $\tilde{y}$ only when $\tilde{y}$ is the unique maximizer. Assumption 1 at $t = 0$ is the inequality $a_{\mathcal{C}} \geq a_{\mathcal{N}}$.*

*Proof.* The denominator of $s$ is positive where $s$ is defined, so $s(x)$ has the sign of $M_{\tilde{y}}(x) = F_{\tilde{y}}(x) - \max_{c \neq \tilde{y}} F_c(x)$, which is positive exactly when $F_{\tilde{y}}(x)$ exceeds every other logit. Therefore $Q(0) = \Pr[s \leq 0] = 1 - \Pr[s > 0]$ under either population. □

The unperturbed input alone gives the response $R_0(x) = \mathbf{1}[s(x) \leq 0]$, the one-view detector that flags a sample when the model disagrees with its label. It is the response of the converse in Theorem 1 at $t = 0$, and its gap is $Q_{\mathrm{noisy}}(0) - Q_{\mathrm{clean}}(0) = a_{\mathcal{C}} - a_{\mathcal{N}}$.

### C.2 Proof of Theorem 1

The proof rests on one identity for non-increasing functions.

**Lemma 3** (Level sets)**.** *Let $h : \mathbb{R} \to [0, 1]$ be non-increasing and $A_\eta = \{\sigma : h(\sigma) > \eta\}$. Each $A_\eta$ is empty, all of $\mathbb{R}$, or a half-line $(-\infty, \beta)$ or $(-\infty, \beta]$, and*

$$\mathbb{E}[h(S_{\mathcal{N}})] - \mathbb{E}[h(S_{\mathcal{C}})] \;=\; \int_0^1 \Big( \Pr[S_{\mathcal{N}} \in A_\eta] - \Pr[S_{\mathcal{C}} \in A_\eta] \Big)\, d\eta.$$

*Proof.* If $\sigma \in A_\eta$ and $\sigma' < \sigma$ then $h(\sigma') \geq h(\sigma) > \eta$, so $A_\eta$ is closed downward, which gives the four shapes. A monotone function is Borel, and $h(\sigma) = \int_0^1 \mathbf{1}[h(\sigma) > \eta]\, d\eta$ for $h(\sigma) \in [0, 1]$. The integrand $(\sigma, \eta) \mapsto \mathbf{1}[h(\sigma) > \eta]$ is jointly measurable and non-negative, so by Tonelli $\mathbb{E}[h(S)] = \int_0^1 \Pr[S \in A_\eta]\, d\eta$ for either population, and the two integrals are finite. Subtracting gives the identity. □

*Proof of Theorem 1. Decomposition.* By Assumption 2 and the tower property, $\mathbb{E}_{\mathcal{C}}[m_{\mathcal{C}}(s(x))] = \mathbb{E}_{\mathcal{C}}[\mathbb{E}[D(x) \mid s(x), \mathcal{C}]] = \mathbb{E}_{\mathcal{C}}[D(x)]$. Hence

$$\Delta D \;=\; \Big( \mathbb{E}_{\mathcal{N}}[m_{\mathcal{C}}(s)] - \mathbb{E}_{\mathcal{C}}[m_{\mathcal{C}}(s)] \Big) + \Big( \mathbb{E}_{\mathcal{N}}[D] - \mathbb{E}_{\mathcal{N}}[m_{\mathcal{C}}(s)] \Big) \;=\; G + X.$$

Assumption 2 fixes $m_{\mathcal{C}}$ only up to a $Q_{\mathrm{clean}}$-null set, which may carry $Q_{\mathrm{noisy}}$ mass where the noisy margins fall outside the clean support. Any two admissible choices give the same $\Delta D$ and may split it differently between $G$ and $X$; everything below holds for each choice.

*Non-negativity.* Apply Lemma 3 with $h = m_{\mathcal{C}}$. For $A_\eta = (-\infty, \beta]$ the integrand is $Q_{\mathrm{noisy}}(\beta) - Q_{\mathrm{clean}}(\beta)$; for $A_\eta = (-\infty, \beta)$ it is $Q_{\mathrm{noisy}}(\beta^-) - Q_{\mathrm{clean}}(\beta^-)$; for $\emptyset$ and $\mathbb{R}$ it is zero. Assumption 1 makes the first non-negative and, on taking left limits, the second as well. Hence $G \geq 0$.

*Window bound.* Fix $a < b$. If $m_{\mathcal{C}}(a) \leq m_{\mathcal{C}}(b)$ the bound is at most zero and follows from $G \geq 0$. Otherwise take $\eta \in [m_{\mathcal{C}}(b), m_{\mathcal{C}}(a))$. Every $\sigma \leq a$ has $m_{\mathcal{C}}(\sigma) \geq m_{\mathcal{C}}(a) > \eta$, so $(-\infty, a] \subseteq A_\eta$, and every $\sigma \geq b$ has $m_{\mathcal{C}}(\sigma) \leq m_{\mathcal{C}}(b) \leq \eta$, so $A_\eta \subseteq (-\infty, b)$. The integrand of Lemma 3 is therefore at least $Q_{\mathrm{noisy}}(a) - Q_{\mathrm{clean}}(b^-)$ on an interval of $\eta$ of length $m_{\mathcal{C}}(a) - m_{\mathcal{C}}(b)$, and non-negative elsewhere by the previous step. Integrating gives $G \geq (m_{\mathcal{C}}(a) - m_{\mathcal{C}}(b))(Q_{\mathrm{noisy}}(a) - Q_{\mathrm{clean}}(b^-))$.

*Converse.* Suppose $Q_{\mathrm{clean}}(t) > Q_{\mathrm{noisy}}(t)$ and consider the response $R(x) = \mathbf{1}[s(x) \leq t]$ in place of $D$. It is a function of $s$, so its clean response is $m_{\mathcal{C}}(\sigma) = \mathbf{1}[\sigma \leq t]$, which is non-increasing with values in $[0, 1]$, and Assumption 2 holds for $R$. Since $R = m_{\mathcal{C}}(s)$ on both populations, its excess is zero and its gap is $G = Q_{\mathrm{noisy}}(t) - Q_{\mathrm{clean}}(t) < 0$. □

*Remark* 1 (What the theorem does and does not use). Non-negativity of $G$ for every non-increasing response is the standard characterization of first-order stochastic dominance (Shaked & Shanthikumar, 2007, Section 1.A.1), and the converse shows that, over all non-increasing responses, Assumption 1 is exactly the condition under which it holds. In particular the mean ordering $\mu_{\mathrm{clean}} \geq \mu_{\mathrm{noisy}}$ does not suffice, since any two laws that satisfy it and cross admit a $t$ as in the converse. The window bound needs no rate of decrease. Non-negativity also degrades gracefully. If $Q_{\mathrm{clean}}(t) - Q_{\mathrm{noisy}}(t) \leq$

$\varepsilon$ for every $t$, the same argument gives $G \geq -\varepsilon$. Chaining windows $a_0 < a_1 < \cdots < a_q$ over disjoint intervals of $\eta$ gives $G \geq \sum_{j=1}^{q} (m_{\mathcal{C}}(a_{j-1}) - m_{\mathcal{C}}(a_j))(Q_{\text{noisy}}(a_{j-1}) - Q_{\text{clean}}(a_j^-))$. The excess $X$ is not signed by either assumption; Appendix D.2 measures it.

### C.3 Removing Assumption 2

**Lemma 4** (Closest non-increasing response)**.** *Let $R$ be any response with values in $[0,1]$, $m_{\mathcal{C}}(\sigma) = \mathbb{E}[R \mid s = \sigma, \mathcal{C}]$ its clean response, and $\mathcal{K}$ the set of $g \in L^2(Q_{\text{clean}})$ with a non-increasing version. The minimizer $h^*$ of $\mathbb{E}_{\mathcal{C}}[(m_{\mathcal{C}}(s) - g(s))^2]$ over $g \in \mathcal{K}$ exists, is unique in $L^2(Q_{\text{clean}})$, has a non-increasing version with values in $[0,1]$, and satisfies $\mathbb{E}_{\mathcal{C}}[R - h^*(s)] = 0$. Theorem 1 therefore holds under Assumption 1 alone, with $h^*$ in place of $m_{\mathcal{C}}$ in $G$ and $X$. Under Assumption 2, $h^* = m_{\mathcal{C}}$.*

*Proof.* $\mathcal{K}$ is a convex cone. It is closed. If $g_n \to g$ in $L^2(Q_{\text{clean}})$, a subsequence converges on a set $B$ of full $Q_{\text{clean}}$ measure, where $g$ is then a pointwise limit of non-increasing functions and so non-increasing on $B$. The function $\tilde{g}(\sigma) = \sup\{g(u) : u \in B,\, u \geq \sigma\}$, set to $\inf_B g$ where the set is empty, is non-increasing on $\mathbb{R}$ and equals $g$ on $B$. The projection $h^*$ onto a non-empty closed convex subset of a Hilbert space exists, is unique, and satisfies $\mathbb{E}_{\mathcal{C}}[(m_{\mathcal{C}} - h^*)(g - h^*)] \leq 0$ for every $g \in \mathcal{K}$. Constants of either sign lie in $\mathcal{K}$ and $\mathcal{K} + c = \mathcal{K}$, so $g = h^* \pm 1$ gives $\mathbb{E}_{\mathcal{C}}[m_{\mathcal{C}}(s) - h^*(s)] = 0$, and the tower property gives $\mathbb{E}_{\mathcal{C}}[R - h^*(s)] = 0$. Clipping $h^*$ to $[0,1]$ keeps it in $\mathcal{K}$ and, since $m_{\mathcal{C}} \in [0,1]$, does not increase the distance to $m_{\mathcal{C}}$; by uniqueness $h^*$ already takes values in $[0,1]$ almost surely. The proof of Theorem 1 used $m_{\mathcal{C}}$ only through $\mathbb{E}_{\mathcal{C}}[R - m_{\mathcal{C}}(s)] = 0$, monotonicity and the range $[0,1]$, which $h^*$ has. If $m_{\mathcal{C}} \in \mathcal{K}$ it is its own projection. □

On a finite sample the empirical measure is itself a distribution, and $h^*$ is the isotonic regression of the observed $\hat{D}$ on $s$ over the clean points, computed exactly by pool-adjacent-violators (Barlow et al., 1972). The decomposition of Appendix D.2 is Theorem 1 applied to that measure. Where noisy margins fall outside the range of the clean ones, the fit is extended by the smallest and the largest non-increasing extension, $\sigma \mapsto h^*(\min\{u \in \mathcal{S}_{\mathcal{C}} : u \geq \sigma\})$ (zero past the largest clean margin) and $\sigma \mapsto h^*(\max\{u \in \mathcal{S}_{\mathcal{C}} : u \leq \sigma\})$ (one below the smallest), with $\mathcal{S}_{\mathcal{C}}$ the clean margins; every admissible extension gives a $G$ between the two.

### C.4 Thresholds and ranking

Theorem 1 applies to any response in $[0,1]$. For the flag $R_\tau = \mathbf{1}[m(x) \geq \tau]$ its gap is $J(\tau) = \Pr_{\mathcal{N}}[m \geq \tau] - \Pr_{\mathcal{C}}[m \geq \tau]$, the detector's true-positive minus false-positive rate at $\tau_d = \tau$, and it splits as $J(\tau) = G_\tau + X_\tau$ with $G_\tau \geq 0$ under Assumption 1, taking $h^*$ from Lemma 4 for $R_\tau$. With noise rate $\pi$, $J(\tau) \geq 0$ is equivalent to a flagged set whose precision $\pi\,\mathrm{TPR}/(\pi\,\mathrm{TPR} + (1-\pi)\,\mathrm{FPR})$ is at least $\pi$. The mean gap and the ranking are both non-negative weighted sums of $J$.

**Lemma 5** (Mean gap and AUROC from thresholds)**.** *Let $Z_{\mathcal{N}}$ and $Z_{\mathcal{C}}$ be independent counts drawn from the noisy and the clean population, $p_{\mathcal{N}}(k) = \Pr[Z_{\mathcal{N}} = k]$, and $\mathrm{AUROC}(m) = \Pr[Z_{\mathcal{N}} > Z_{\mathcal{C}}] + \frac{1}{2}\Pr[Z_{\mathcal{N}} = Z_{\mathcal{C}}]$. Then*

$$\Delta D \;=\; \frac{1}{K}\sum_{\tau=1}^{K} J(\tau), \qquad \mathrm{AUROC}(m) - \frac{1}{2} \;=\; \sum_{\tau=1}^{K} w_\tau\, J(\tau), \quad w_\tau = \frac{p_{\mathcal{N}}(\tau-1) + p_{\mathcal{N}}(\tau)}{2} \;\geq\; 0.$$

*If $J(\tau) \geq 0$ at every $\tau$ then $\mathrm{AUROC}(m) \geq \frac{1}{2}$.*

*Proof.* For a count in $\{0, \ldots, K\}$, $\mathbb{E}[m] = \sum_{\tau=1}^{K} \Pr[m \geq \tau]$; dividing the difference of the two populations by $K$ gives the first identity. For the second, write $S(k) = \Pr[m \geq k]$ for either population, so that $S(0) = 1$ and $S(K+1) = 0$ and $J(0) = J(K+1) = 0$. The mid-rank function $\Phi(k) = \Pr[m < k] + \frac{1}{2}\Pr[m = k] = 1 - \frac{1}{2}(S(k) + S(k+1))$ gives $\mathrm{AUROC}(m) = \mathbb{E}[\Phi_{\mathcal{C}}(Z_{\mathcal{N}})]$ by independence, and $\mathbb{E}[\Phi_{\mathcal{N}}(Z_{\mathcal{N}})] = \frac{1}{2}$ because $\sum_k p(k)(\Pr[m < k] + \frac{1}{2}p(k)) = \frac{1}{2}(\sum_k p(k))^2$. Hence

$$\mathrm{AUROC}(m) - \tfrac{1}{2} \;=\; \sum_{k=0}^{K} p_{\mathcal{N}}(k)\,\big(\Phi_{\mathcal{C}}(k) - \Phi_{\mathcal{N}}(k)\big) \;=\; \sum_{k=0}^{K} p_{\mathcal{N}}(k)\,\frac{J(k) + J(k+1)}{2},$$

and collecting the coefficient of each $J(\tau)$ gives $w_\tau$. □

# D EMPIRICAL CHECKS OF THE ASSUMPTIONS

## D.1 ASSUMPTION 1: DOMINANCE OF NORMALIZED MARGINS

Assumption 1 states that the clean normalized margins stochastically dominate the noisy ones. Because $s$ is computed exactly by automatic differentiation, it can be checked directly. We check it on four cells, each on a seeded 2,000-sample subset of its training split (Table 5, Figure 5).

On every measured cell the clean margins dominate the noisy ones pointwise, $Q_{\text{clean}}(t) \le Q_{\text{noisy}}(t)$ for every $t$, so Assumption 1 holds throughout, and a Mann-Whitney test rejects equality on each cell (Table 5). Cohen's effect size $d_{\text{Cohen}}$ runs from 0.66 on CIFAR-100N, where the distributions overlap visibly near $s = 0$ (Figure 5), to 1.78 on CIFAR-10 IDN 20%, where they barely do. Because $s$ is an input-space distance, its scale depends on the network and the data, so means and medians are not comparable between cells, while dominance, $d_{\text{Cohen}}$ and the ranking statistics are.

Table 5: Assumptions 1 and 2 per cell, on a seeded 2,000-sample subset of the training split. $n_{\mathcal{N}}$ is the number of mislabeled samples in the subset and $d_{\text{Cohen}}$ is Cohen's effect size between the clean and the noisy $s$. Dominance marks whether $Q_{\text{clean}} \le Q_{\text{noisy}}$ holds at every $t$, and $z$ is the tie-corrected Mann-Whitney statistic (all $p < 10^{-70}$). Steps counts the non-increasing steps of $\mathbb{E}[\hat{D} \mid s]$ over eighteen quantile bins. Share is the fraction of the measured gap that the margin term of Theorem 1 carries, with the clean response fitted by isotonic regression.

| Cell | $n_{\mathcal{N}}$ | $d_{\text{Cohen}}$ | dominance | MW $z$ | steps | share |
|---|---|---|---|---|---|---|
| CIFAR-10 IDN 20% | 434 | 1.78 | yes | 31.0 | 14/17 | 0.91–0.92 |
| CIFAR-10 IDN 40% | 839 | 1.58 | yes | 37.1 | 16/17 | 0.82–0.83 |
| Fashion-MNIST IDN 40% | 789 | 0.98 | yes | 34.6 | 16/17 | 0.79–0.85 |
| CIFAR-100N | 785 | 0.66 | yes | 26.1 | 17/17 | 0.84–0.86 |

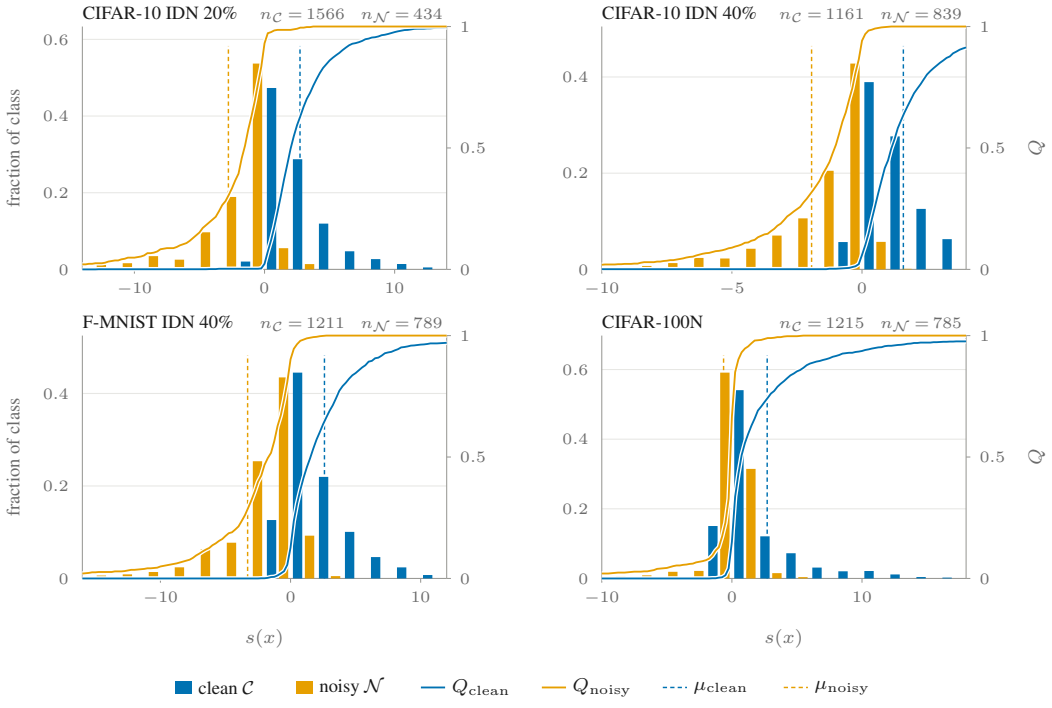



Figure 5: Distribution of the sensitivity-normalized margin $s(x)$ over clean ($\mathcal{C}$) and noisy ($\mathcal{N}$) samples, one panel per cell, with within-class bin fractions (left axis), empirical CDFs (right axis) and dotted class means. The clean CDF lies below the noisy one at every $s$ on every cell, which is Assumption 1. Each panel spans the 2nd percentile of noisy $s$ to the 90th of clean $s$.

### D.2 Assumption 2: monotone response

Assumption 2 is checked on the same cells and subsets, with $D$ estimated by the twelve-view rate $\hat{D}$. We cut $s$ into eighteen quantile bins and average $\hat{D}$ within each; Table 5 counts how many of the seventeen steps between adjacent bins are non-increasing. On every cell the response falls from above $0.85$ in the lowest bin to below $0.13$ in the highest and is non-increasing on at least 14 of the 17 steps. Lemma 4 makes the remaining small increases harmless, because the decomposition below uses the closest non-increasing fit and needs no binning.

Table 6 applies Theorem 1 to the empirical measure of each subset. Dominance holds exactly there, so $G \geq 0$ is guaranteed, and the table reports its size. The margin term carries $0.79$ to $0.92$ of the measured gap across the four cells (Table 5), the range within a cell coming from how the fit is extended below the smallest clean margin. The excess $X$ lies between $0.058$ and $0.145$ and is positive on every cell under either extension, so mislabeled samples flip more often than the clean response predicts at their margins. The window bound, maximized over $a < b$, alone guarantees $0.31$ to $0.63$ of the gap. Its maximizing window straddles $s = 0$ on every cell, where Lemma 2 ties $Q_{\text{noisy}}(0) - Q_{\text{clean}}(0)$ to $a_{\mathcal{C}} - a_{\mathcal{N}}$. The flag $\mathbf{1}[m \geq \tau]$ has positive $J(\tau)$ and positive $X_\tau$ at all twelve thresholds on every cell.

Table 6: Theorem 1 on the empirical measure of each $2{,}000$-sample subset, with $m_{\mathcal{C}}$ the isotonic fit of the clean response. $\Delta D$ is the measured gap, and $G$ and $X$ range over the smallest and largest non-increasing extensions; window is the bound of Theorem 1 maximized over $a < b$.

| Cell | $\Delta D$ | $G$ | $X$ | window |
|---|---|---|---|---|
| CIFAR-10 IDN 20% | 0.802 | 0.731–0.734 | 0.068–0.071 | 0.503 |
| CIFAR-10 IDN 40% | 0.758 | 0.620–0.628 | 0.130–0.138 | 0.344 |
| Fashion-MNIST IDN 40% | 0.681 | 0.537–0.581 | 0.100–0.145 | 0.306 |
| CIFAR-100N | 0.399 | 0.335–0.341 | 0.058–0.064 | 0.122 |

## E The Twelve Augmented Views

Table 7 lists the $K = 12$ views in the order the implementation applies them, with their parameters at the $32\times32$ working resolution of CIFAR-10, CIFAR-100 and Fashion-MNIST. The names are fixed but every spatial parameter is rescaled with the input, so Animal-10N at its native $64\times64$ (Appendix K) doubles the crop padding, the centre-crop extent and the two Gaussian $\sigma$. All views end with tensor conversion and per-channel normalization. Eight are deterministic; CROP, CROP+HFLIP, AFFINE and COMBO resample per call.

Table 7: The $K = 12$ TTA views.

| View | Transformation |
|---|---|
| 1 | HFLIP: horizontal flip, applied with $p = 1$ |
| 2 | CROP: random crop $32\times32$, padding 4 |
| 3 | CROP+HFLIP: random crop $32\times32$/pad 4 + horizontal flip |
| 4 | AFFINE: random affine, rotation $\pm15^\circ$, translation $10\%$, shear $10^\circ$ |
| 5 | COMBO: random crop $32\times32$/pad 4 + flip + jitter $(0.3/0.3/0.2)$ + affine ($\pm10^\circ$, translation $5\%$) |
| 6 | SHARPEN-2: sharpness adjustment, factor 2.0 |
| 7 | SHARPEN-8: sharpness adjustment, factor 8.0 |
| 8 | HIGHPASS-1.0: image minus its Gaussian blur, $\sigma = 1.0$, kernel 5 |
| 9 | HIGHPASS-2.0: image minus its Gaussian blur, $\sigma = 2.0$, kernel 9 |
| 10 | INVERT: photometric inversion |
| 11 | ROT-30: rotation by a fixed $30^\circ$ |
| 12 | ZOOM-22: centre crop $22\times22$, resized to $32\times32$ |

# F Ablation Studies

## F.1 Normalized embedding against backbone features

Section 3.1 places the classifier on the $\ell_2$-normalized contrastive embedding rather than on backbone features. The ablation attaches a second classifier head to the backbone features alongside the normalized embedding, matched on cell, seed, backbone and optimizer settings, and run with the unmodified ImageNet stem rather than the one of Appendix K. Detection survives at a cost of 1–5 points of F1, while the memorization ratio rises from $0.028$–$0.068$ to $0.093$–$0.168$, a factor of 2.5–3.3.

## F.2 Contrastive term

We retrained the detector on CIFAR-10 IDN 40% with $\lambda = 0$, keeping the noise, recipe, views and seed of the reported run, and cleaned both at $(\tau_d, \tau_r) = (9, 6)$. With the contrastive term the detector flags fewer clean samples, relabels more accurately, leaves less residual noise and ranks slightly better (Table 8). Only detection F1 favours $\lambda = 0$, by $0.005$, because that detector flags more samples overall.

Table 8: With and without the contrastive term on CIFAR-10 IDN 40%. Bold is better.

| | $\boldsymbol{\lambda = 1}$ **(reported)** | $\boldsymbol{\lambda = 0}$ |
|---|---|---|
| Detection F1 | $0.9421$ | $\mathbf{0.9467}$ |
| AUROC | $\mathbf{0.9873}$ | $0.9865$ |
| AP | $\mathbf{0.9756}$ | $0.9740$ |
| Clean samples flagged ($m \geq 9$) | $\mathbf{2.6}\%$ | $3.6\%$ |
| Accuracy of relabeled samples | $\mathbf{92.7}\%$ | $90.4\%$ |
| Clean labels relabeled to a wrong class | $\mathbf{563}$ | $859$ |
| Residual noise | $\mathbf{5.77}\%$ | $5.92\%$ |

## F.3 Number of views and single-pass confidence

The disagreement count pools twelve views. To see what the pooling adds, we score each view on its own and every one of the $4{,}095$ non-empty view subsets, using the per-view predictions the reported checkpoints already cached, so nothing is retrained. Table 9 averages the AUROC of the count over all subsets of each size $k$. On every cell AUROC rises with each added view and levels off after about eight, and the best single view, picked with the clean labels, stays $0.013$ to $0.109$ below the full count.

One view gives only a binary flag. The stronger single-pass baseline is the model's continuous confidence on the un-augmented image. We replayed four reported checkpoints on the un-augmented training images, reproducing their cached per-view predictions on at least $99.99\%$ of samples. The softmax probability of the observed label then reaches AUROC $0.971$, $0.959$, $0.978$ and $0.834$ on CIFAR-10 IDN 20%, CIFAR-10 IDN 40%, Fashion-MNIST IDN 30% and CIFAR-100N, against $0.991$, $0.987$, $0.991$ and $0.840$ for the count. Cross-entropy on the observed label ranks samples identically, and the observed-label logit margin reaches $0.968$, $0.961$, $0.977$ and $0.831$.

A continuous twelve-view score, the mean probability of the observed label over the views, also beats single-pass confidence on all four cells, at AUROC $0.990$, $0.974$, $0.983$ and $0.855$. The count ranks and cleans better on three of the four. On CIFAR-100N the mean ranks better ($0.855$ against $0.840$), and at the budget of Appendix F.6 it keeps $131$ more correct labels.

## F.4 Cleaning against no cleaning

On CIFAR-10 IDN 40%, we also train the downstream classifier on the uncleaned labels, with the same recipe, architecture and seeds as the cleaned arm. The uncleaned set gives $86.93 \pm 0.96$ against $91.95 \pm 0.46$ for the cleaned one, a gain on each of the three seeds (paired $t = 11.6$, 2 df, $p = 0.007$, 95% CI $[+3.15, +6.88]$).

Table 9: Detection AUROC of the disagreement count over $k$ of the twelve views, averaged over every subset of that size, and of the best single view chosen with the clean labels.

| Benchmark | $k = 1$ | 2 | 4 | 8 | 12 | Best view |
|---|---|---|---|---|---|---|
| CIFAR-10 IDN 20% | 0.904 | 0.960 | 0.982 | 0.989 | 0.991 | 0.956 |
| CIFAR-10 IDN 30% | 0.907 | 0.961 | 0.981 | 0.989 | 0.991 | 0.959 |
| CIFAR-10 IDN 40% | 0.880 | 0.945 | 0.974 | 0.984 | 0.987 | 0.939 |
| CIFAR-100 IDN 20% | 0.783 | 0.871 | 0.923 | 0.944 | 0.950 | 0.880 |
| CIFAR-100 IDN 30% | 0.770 | 0.857 | 0.909 | 0.932 | 0.939 | 0.870 |
| CIFAR-100 IDN 40% | 0.734 | 0.818 | 0.873 | 0.898 | 0.907 | 0.831 |
| Fashion-MNIST IDN 20% | 0.911 | 0.966 | 0.987 | 0.992 | 0.994 | 0.981 |
| Fashion-MNIST IDN 30% | 0.884 | 0.946 | 0.979 | 0.989 | 0.991 | 0.970 |
| Fashion-MNIST IDN 40% | 0.852 | 0.928 | 0.965 | 0.978 | 0.982 | 0.940 |
| CIFAR-10N | 0.809 | 0.873 | 0.916 | 0.942 | 0.951 | 0.842 |
| CIFAR-100N | 0.699 | 0.764 | 0.808 | 0.832 | 0.840 | 0.751 |

### F.5 Relabeling against discarding

Every flagged sample is either relabeled or discarded, and $\tau_r$ decides which. Table 10 fixes each cell's $\tau_d$ and compares the reported $\tau_r$ with its two extremes, computed from the cached per-sample records. Discarding every flagged sample leaves the least residual noise but loses correct labels on every cell, because the clean samples the detector flags are thrown away with the mislabeled ones. Relabeling every flagged sample recovers the most correct labels but also keeps the most wrong ones. The reported rule sits between the two. It ends every cell with more correct labels than it started with, at lower residual noise than relabeling everything.

Table 10: Cleaning at each cell's reported $\tau_d$ when every flagged sample is discarded, when every flagged sample is relabeled, and under the reported rule. $\Delta$ correct is the change in retained correct labels against no cleaning; residual is the noise rate of the kept set.

| | | Discard all | | Relabel all | | Reported rule | |
|---|---|---|---|---|---|---|---|
| Benchmark | $\tau_d/\tau_r$ | $\Delta$ correct | Residual (%) | $\Delta$ correct | Residual (%) | $\Delta$ correct | Residual (%) |
| CIFAR-10 IDN 20% | 10/6 | −296 | 2.26 | +8,604 | 3.26 | +8,498 | 2.93 |
| CIFAR-10 IDN 30% | 10/6 | −544 | 2.70 | +12,653 | 4.61 | +12,501 | 4.11 |
| CIFAR-10 IDN 40% | 9/6 | −787 | 4.85 | +16,596 | 6.83 | +16,263 | 5.77 |
| CIFAR-100 IDN 20% | 12/6 | −1,096 | 7.02 | +4,883 | 11.01 | +4,287 | 8.63 |
| CIFAR-100 IDN 30% | 12/6 | −1,568 | 11.33 | +6,708 | 16.83 | +5,785 | 13.10 |
| CIFAR-100 IDN 40% | 11/6 | −2,854 | 15.02 | +8,304 | 23.64 | +6,646 | 17.02 |
| Fashion-MNIST IDN 20% | 10/6 | −669 | 0.89 | +10,810 | 2.61 | +10,748 | 2.46 |
| Fashion-MNIST IDN 30% | 9/6 | −764 | 1.97 | +15,770 | 3.72 | +15,629 | 3.33 |
| Fashion-MNIST IDN 40% | 9/6 | −1,743 | 3.27 | +19,750 | 7.10 | +19,541 | 6.56 |
| CIFAR-10N | 9/6 | −397 | 3.68 | +2,039 | 4.93 | +1,964 | 4.60 |
| CIFAR-100N | 10/6 | −3,203 | 23.53 | +2,474 | 35.25 | +1,121 | 28.42 |

### F.6 Cleaning with single-pass confidence

Appendix F.3 compares rankings, and this subsection compares the cleaned datasets. On the four replayed checkpoints, a single-pass cleaner flags as many samples as the reported rule, taking those with the lowest probability of the observed label or, in a second variant, the lowest observed-label logit margin. It relabels as many flagged samples as the reported rule does, choosing the most confident ones and giving each its single-pass prediction, and discards the rest. Every cleaned set therefore keeps as many samples as ours. Table 11 shows that the twelve-view rule ends with more correct labels and less residual noise on all four cells.

## G Algorithms

Algorithm 1 trains the detector and Algorithm 2 scores and cleans. Neither reads a true label $y_i$.

Table 11: Cleaning at a matched budget, with the same numbers of flagged, relabeled and discarded samples as the reported rule and the flags and labels taken from the twelve views or from one forward pass. Δ correct is the change in retained correct labels against no cleaning; residual is the noise rate of the kept set.

| | Twelve views (ours) | | Single-pass probability | | Single-pass margin | |
|---|---|---|---|---|---|---|
| **Benchmark** | **Δ correct** | **Residual (%)** | **Δ correct** | **Residual (%)** | **Δ correct** | **Residual (%)** |
| CIFAR-10 IDN 20% | +8,498 | 2.93 | +7,987 | 3.96 | +8,047 | 3.84 |
| CIFAR-10 IDN 40% | +16,263 | 5.77 | +14,660 | 9.04 | +15,196 | 7.95 |
| Fashion-MNIST IDN 30% | +15,629 | 3.33 | +15,143 | 4.15 | +15,475 | 3.59 |
| CIFAR-100N | +1,121 | 28.42 | +564 | 29.71 | +680 | 29.44 |

**Algorithm 1** Training the Siamese detector (Sections 3.1 and 3.1.1). One shared network is applied to both inputs of every pair, and the kept checkpoint is the validation-accuracy peak.

**Input:** noisy training set $\mathcal{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$, split into a training part and a validation part; margin $\gamma$; balance $\lambda$; number of pairs $P$
**Output:** trained logit map $F = h_\psi \circ g_\phi$

initialize the encoder $g_\phi$ (backbone, head, $\ell_2$ normalization) and the classifier $h_\psi$
draw $P$ index pairs $(i, j)$ from the training part, half positive ($\tilde{y}_i = \tilde{y}_j$) and half negative, and set $\omega_{ij} \leftarrow \mathbf{1}[\tilde{y}_i = \tilde{y}_j]$
**for** each training epoch **do**
  **for** each minibatch of pairs **do**
    **for** each pair $(i, j)$ in the minibatch **do**
      apply a random augmentation to $x_i$ and to $x_j$
      $z_i \leftarrow g_\phi(x_i), \quad z_j \leftarrow g_\phi(x_j)$ ▷ same weights on both branches; $\|z_i\|_2 = \|z_j\|_2 = 1$
      $d_{ij} \leftarrow \|z_i - z_j\|_2$ ▷ lies in $[0, 2]$
      $\mathcal{L}_{\text{con}} \leftarrow \omega_{ij}\, d_{ij}^2 + (1 - \omega_{ij}) \max(0,\, \gamma - d_{ij})^2$ ▷ contrastive term
      $\mathcal{L}_{\text{cls}} \leftarrow \mathcal{L}_{\text{CE}}(h_\psi(z_i), \tilde{y}_i) + \mathcal{L}_{\text{CE}}(h_\psi(z_j), \tilde{y}_j)$ ▷ classification term; $h_\psi(z_i) = F(x_i)$
      $\mathcal{L}_{ij} \leftarrow \lambda\, \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{cls}}$ ▷ Equation 1
    **end for**
    update $(\phi, \psi)$ by one gradient step on the minibatch mean of $\mathcal{L}_{ij}$
  **end for**
  record the accuracy of $F = h_\psi \circ g_\phi$ on the validation part, against its observed labels
**end for**
**return** $F$ from the epoch with the highest validation accuracy

# H Additional Detection Results

## H.1 Sources for Table 2

Table 2 compares five published detectors, with SimiFeat's two variants as separate rows. All are evaluated on the instance-dependent generator of Xia et al. (2020) at $40\%$ and on the same two CIFAR-N label sets, so four of our cells admit a matched comparison. Fashion-MNIST has no published detection F1 under this generator. DISC (Li et al., 2023), the closest augmented-view method (Section 2), publishes none either, because its selection is used inside its own training loop and never emitted. SimiFeat's rows come from its own paper (Zhu et al., 2022, Table 1), except the two entries marked †. Those and all other rows come from Kim et al. (2024, Table 2), because Zhu et al. (2022) report CIFAR-10 human noise at $\eta \approx 0.16$ rather than the $9.01\%$ aggregate set. The best threshold on the same grid (Table 12) is at most 2.6 points above our row.

## H.2 The headline F1 decomposed

Table 12 decomposes the headline F1 of Table 1 into precision, recall and the operating threshold. The first block is Table 1's operating point, the grid-search pick of Section 5.1. Best F1 is the maximum over the same floored grid, usually at a different $\tau_d$ and so a different cleaning; it is a reference rather than a point we run at, since selection scores downstream accuracy. AUROC and AP

**Algorithm 2** TTACLEANER detection and cleaning (Section 3.2). The scoring loop is independent of both thresholds, so the cached records support any $(\tau_d, \tau_r)$ without a second pass over the data.

**Input:** noisy training set $\mathcal{D} = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$; trained logit map $F$ from Algorithm 1; transforms $t_1, \ldots, t_K$ (Appendix E); thresholds $\tau_d, \tau_r$
**Output:** cleaned training set $\mathcal{D}'$

**for** each sample $(x_i, \tilde{y}_i) \in \mathcal{D}$ **do** ▷ scoring pass: one batched forward call per sample
  **for** $k = 1$ to $K$ **do**
    $\hat{y}_{ik} \leftarrow \arg\max_c F_c(t_k(x_i))$ ▷ prediction on the $k$-th view
  **end for**
  $m_i \leftarrow \#\{k : \hat{y}_{ik} \neq \tilde{y}_i\}$ ▷ TTA mistakes count
  $\hat{y}_i^{\mathrm{maj}} \leftarrow \arg\max_c \#\{k : \hat{y}_{ik} = c\}, \quad v_i \leftarrow \#\{k : \hat{y}_{ik} = \hat{y}_i^{\mathrm{maj}}\}$
  cache the record $(m_i, \hat{y}_i^{\mathrm{maj}}, v_i)$
**end for**
$\mathcal{D}' \leftarrow \emptyset$
**for** each sample $(x_i, \tilde{y}_i) \in \mathcal{D}$ **do** ▷ cleaning pass: Equation 2, cached records only
  **if** $m_i < \tau_d$ **then** ▷ not flagged
    $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{(x_i, \tilde{y}_i)\}$ ▷ keep the observed label
  **else if** $v_i \geq \tau_r$ **then** ▷ flagged, and the views agree on a replacement
    $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{(x_i, \hat{y}_i^{\mathrm{maj}})\}$ ▷ relabel
  **end if** ▷ flagged with scattered views: discard $x_i$
**end for**
**return** $\mathcal{D}'$

describe the ranking the count induces, independent of any threshold, and AUROC is tie-corrected because the count is an integer in $[0, K]$.

Table 12: Detection quality from the per-sample records, noisy positive, at the reported operating point and at the best threshold on the same grid.

| Benchmark | At the reported operating point | | | | Best F1 | Threshold-free | |
|---|---|---|---|---|---|---|---|
| | $\tau_d$ | Precision | Recall | F1 | F1 ($\tau_d$) | AUROC | AP |
| *Synthetic instance-dependent noise* | | | | | | | |
| CIFAR-10 IDN 20% | 10 | 0.9692 | 0.9109 | 0.9391 | 0.9413 (9) | 0.9909 | 0.9693 |
| CIFAR-10 IDN 30% | 10 | 0.9626 | 0.9360 | 0.9491 | 0.9503 (9) | 0.9907 | 0.9741 |
| CIFAR-10 IDN 40% | 9 | 0.9592 | 0.9255 | 0.9421 | 0.9439 (8) | 0.9873 | 0.9756 |
| CIFAR-100 IDN 20% | 12 | 0.8722 | 0.7199 | 0.7888 | 0.8088 (11) | 0.9500 | 0.8115 |
| CIFAR-100 IDN 30% | 12 | 0.8739 | 0.7187 | 0.7888 | 0.8243 (10) | 0.9390 | 0.8304 |
| CIFAR-100 IDN 40% | 11 | 0.8432 | 0.7626 | 0.8009 | 0.8142 (9) | 0.9067 | 0.8330 |
| Fashion-MNIST IDN 20% | 10 | 0.9470 | 0.9660 | 0.9564 | 0.9564 (10) | 0.9939 | 0.9685 |
| Fashion-MNIST IDN 30% | 9 | 0.9574 | 0.9540 | 0.9557 | 0.9557 (9) | 0.9906 | 0.9714 |
| Fashion-MNIST IDN 40% | 9 | 0.9291 | 0.9517 | 0.9403 | 0.9403 (9) | 0.9816 | 0.9615 |
| *Real human annotation noise* | | | | | | | |
| CIFAR-10N | 9 | 0.8750 | 0.6171 | 0.7238 | 0.7412 (7) | 0.9512 | 0.7814 |
| CIFAR-100N | 10 | 0.7877 | 0.5913 | 0.6756 | 0.7267 (7) | 0.8401 | 0.7541 |

### H.3 SYMMETRIC NOISE

Every cell above carries instance-dependent or human noise. To check that the detector is not tied to either, we trained it once more on CIFAR-10 with symmetric noise, replacing exactly $50\%$ of the labels with a class drawn uniformly from the other nine. The recipe, views and seed are those of the CIFAR-10 IDN cells. This cell has no grid search, so both thresholds come from the label-free rule of Appendix J. The mixture estimates $\hat{\pi} = 0.513$ against a true rate of $0.500$ and sets $\tau_d = 9$, which is also the F1-maximizing threshold. There the detector reaches precision $0.9435$, recall $0.9498$ and F1 $0.9466$, with AUROC $0.9801$ and AP $0.9702$, in line with the CIFAR-10 IDN cells of Table 12. Cleaning at $\tau_r = 6$ relabels $23{,}326$ samples and discards $1{,}841$. The noise rate falls from $50.0\%$ to $9.4\%$, and the cleaned set holds $18{,}631$ more correct labels than the uncleaned one.

### H.4 Detection ROC across all twelve thresholds

The disagreement count is an integer in $[0, K]$, so it takes thirteen values and leaves exactly twelve non-trivial thresholds. Figure 6 plots all of them, one panel per dataset, each holding that dataset's synthetic cells and, where one exists, the human re-annotation of the same images. Every marker is an operating point the detector can reach and the segments between them are interpolation, so the number in parentheses in each legend is the area under the resulting fourteen-point trapezoid. Because the score is discrete, this is the whole curve rather than a sample of it, and no finer sweep can find a threshold between two adjacent markers. Every panel shows false-positive rates up to $0.4$ and true-positive rates from $0.4$, which holds every point at $\tau_d \geq 6$, and the ringed marker is the F1-maximizing threshold reported as Best F1 in Table 12.

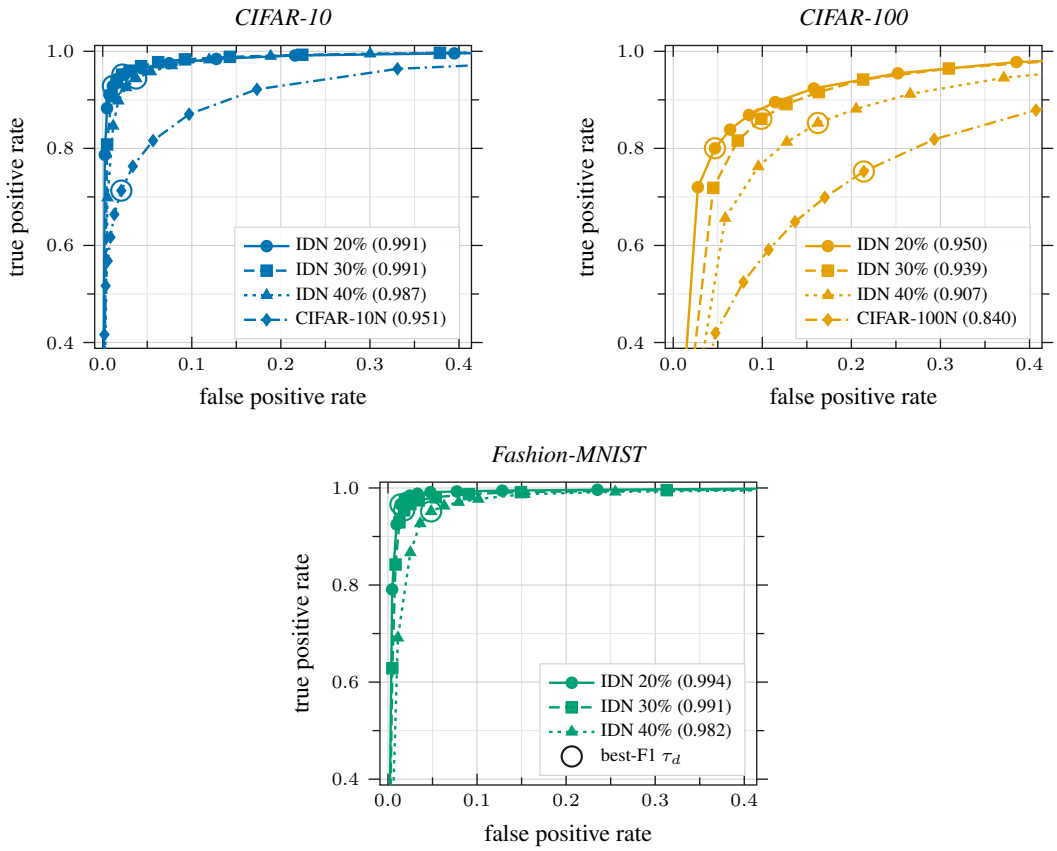



Figure 6: Detection ROC on the eleven cells with clean labels, noisy positive, over the score's twelve non-trivial thresholds. Colour is dataset; dash pattern and marker are noise level, and the ringed marker is the F1-maximizing threshold of Table 12.

## I Downstream Comparisons

Our $\pm$ is over the three downstream seeds. Welch tests against SGN (Englesson & Azizpour, 2024), the highest-scoring entry in both CIFAR-N columns (their $n = 5$), separate the cleaned dataset on CIFAR-10N ($t = 22.13$, $df = 4.68$, $p \approx 6.4 \times 10^{-6}$, 95% CI of the difference $[+1.62, +2.06]$) but not on CIFAR-100N ($t = 0.41$, $df = 6.00$, $p \approx 0.69$, 95% CI $[-0.78, +1.10]$). Table 13 gives the Animal-10N result against its reference rows.

Table 14 covers synthetic instance-dependent noise. Yang et al. (2022) report no CIFAR-100 results, so Table 15 gives our three CIFAR-100 cells on their own.

Table 13: Downstream accuracy (%) under real-world annotation noise. Top: the four Englesson & Azizpour (2024) methods not in Table 3, with our row repeated. Bottom: Animal-10N. Reference values come from the re-training study named in each block heading, which also states what $\pm$ denotes; our rows are mean $\pm$ std over three seeds. **Bold**: best mean per column and anything not separable from it (Welch, $p > 0.05$); a reported standard error enters the test as the standard error of the mean.

| **Method** | **CIFAR-10N (Aggr.)** | **CIFAR-100N** |
|---|---|---|
| *Re-trained by Englesson & Azizpour (2024); $\pm$ std* | | |
| GCE (Zhang & Sabuncu, 2018) | $86.66 \pm 0.68$ | $48.81 \pm 0.46$ |
| LS (Szegedy et al., 2016) | $85.08 \pm 0.54$ | $45.98 \pm 1.44$ |
| HET (Collier et al., 2021) | $84.45 \pm 0.57$ | $45.58 \pm 0.80$ |
| NAN (Chen et al., 2021a) | $85.53 \pm 0.83$ | $50.00 \pm 0.72$ |
| **Ours** (clean, then train) | $\mathbf{93.90 \pm 0.11}$ | $\mathbf{60.52 \pm 0.38}$ |

| **Method** | **Animal-10N** |
|---|---|
| *Re-trained by Guo et al. (2024); $\pm$ standard error* | |
| LogitClip (Wei et al., 2023) | $70.89 \pm 0.65$ |
| CoDis (Xia et al., 2023) | $73.08 \pm 0.35$ |
| MBEM (Khetan et al., 2018) | $76.96 \pm 3.17$ |
| CCC (Zhang et al., 2024) | $78.36 \pm 0.35$ |
| CrowdLayer (Rodrigues & Pereira, 2018) | $79.70 \pm 0.35$ |
| Max-MIG (Cao et al., 2019) | $79.78 \pm 0.80$ |
| CE (MV) | $79.88 \pm 0.38$ |
| DoctorNet (Guan et al., 2018) | $79.96 \pm 0.55$ |
| CE (EM) (Dawid & Skene, 1979) | $80.18 \pm 0.34$ |
| TraceReg (Tanno et al., 2019) | $80.34 \pm 0.66$ |
| CoNAL (Chu et al., 2021) | $80.45 \pm 0.49$ |
| Co-teaching+ (Yu et al., 2019) | $81.24 \pm 0.22$ |
| Co-teaching (Han et al., 2018) | $81.96 \pm 0.48$ |
| AdaptCDRP (Guo et al., 2024) | $83.08 \pm 0.39$ |
| *Re-trained by Chen et al. (2021c); $\pm$ std over three runs* | |
| CE (Chen et al., 2021c) | $79.4 \pm 0.1$ |
| Nested Dropout (Chen et al., 2021c) | $81.3 \pm 0.6$ |
| SELFIE (Song et al., 2019) | $81.8 \pm 0.1$ |
| PLC (Zhang et al., 2021) | $83.4 \pm 0.4$ |
| Nested Co-teaching (Chen et al., 2021c) | $84.1 \pm 0.1$ |
| **Ours** (clean, then train) | $\mathbf{87.11 \pm 0.44}$ |

Table 14: Downstream accuracy (%) under synthetic instance-dependent noise. Reference rows are as published by Yang et al. (2022), who use the same generator (Xia et al., 2020). **Bold**: best mean per column and anything not separable from it (Welch, $p > 0.05$). Our row is three seeds.

| | **Fashion-MNIST** | | | **CIFAR-10** | | |
|---|---|---|---|---|---|---|
| **Method** | **IDN-20%** | **IDN-30%** | **IDN-40%** | **IDN-20%** | **IDN-30%** | **IDN-40%** |
| CE | $88.31 \pm 0.37$ | $85.22 \pm 0.56$ | $76.56 \pm 2.50$ | $71.49 \pm 1.35$ | $67.52 \pm 1.68$ | $58.63 \pm 4.92$ |
| GCE | $88.59 \pm 0.26$ | $86.64 \pm 0.76$ | $76.93 \pm 1.64$ | $72.11 \pm 0.43$ | $69.31 \pm 0.18$ | $56.86 \pm 0.92$ |
| APL | $85.32 \pm 0.88$ | $85.59 \pm 0.85$ | $74.66 \pm 2.77$ | $68.89 \pm 0.27$ | $65.17 \pm 0.35$ | $53.22 \pm 2.21$ |
| Decoupling | $86.29 \pm 1.13$ | $86.01 \pm 1.01$ | $78.78 \pm 0.53$ | $74.23 \pm 1.18$ | $70.85 \pm 1.88$ | $54.73 \pm 1.02$ |
| MentorNet | $87.92 \pm 0.83$ | $87.24 \pm 0.99$ | $79.01 \pm 2.30$ | $73.59 \pm 1.29$ | $72.32 \pm 1.04$ | $57.85 \pm 1.88$ |
| Co-teaching | $88.77 \pm 0.41$ | $86.98 \pm 0.67$ | $78.92 \pm 1.36$ | $75.93 \pm 0.87$ | $74.86 \pm 0.42$ | $59.07 \pm 1.03$ |
| Co-teaching+ | $89.15 \pm 1.77$ | $86.15 \pm 1.04$ | $79.23 \pm 1.30$ | $75.14 \pm 0.61$ | $71.92 \pm 2.13$ | $59.15 \pm 0.87$ |
| Joint | $78.46 \pm 1.24$ | $72.73 \pm 2.44$ | $65.93 \pm 2.08$ | $76.45 \pm 0.45$ | $75.93 \pm 1.65$ | $63.22 \pm 5.37$ |
| DMI | $90.33 \pm 0.66$ | $85.96 \pm 1.52$ | $77.77 \pm 2.15$ | $73.49 \pm 0.88$ | $73.93 \pm 0.34$ | $60.22 \pm 3.47$ |
| Forward | $88.61 \pm 0.77$ | $85.01 \pm 0.43$ | $78.59 \pm 0.38$ | $70.98 \pm 0.32$ | $66.53 \pm 1.96$ | $58.63 \pm 1.25$ |
| Reweight | $88.81 \pm 0.44$ | $84.93 \pm 0.42$ | $76.07 \pm 1.93$ | $71.49 \pm 0.57$ | $68.76 \pm 0.37$ | $60.32 \pm 1.03$ |
| S2E | $89.93 \pm 1.08$ | $86.77 \pm 1.15$ | $76.12 \pm 1.21$ | $75.53 \pm 0.32$ | $71.21 \pm 2.51$ | $64.62 \pm 0.68$ |
| T-Revision | $90.24 \pm 1.01$ | $85.59 \pm 1.77$ | $78.24 \pm 1.12$ | $73.42 \pm 0.64$ | $71.15 \pm 0.43$ | $59.93 \pm 1.33$ |
| PTD | $90.78 \pm 0.64$ | $87.86 \pm 0.78$ | $79.46 \pm 1.58$ | $76.05 \pm 1.72$ | $75.42 \pm 1.33$ | $65.92 \pm 2.33$ |
| DivideMix | $\mathbf{95.92 \pm 0.73}$ | $90.37 \pm 0.83$ | $80.92 \pm 2.32$ | $81.42 \pm 0.28$ | $80.73 \pm 1.28$ | $70.29 \pm 1.97$ |
| BLTM | $94.97 \pm 0.33$ | $91.47 \pm 1.36$ | $82.88 \pm 2.72$ | $80.26 \pm 0.63$ | $77.69 \pm 1.37$ | $71.96 \pm 2.27$ |
| BLTM-V | $\mathbf{95.55 \pm 0.59}$ | $\mathbf{92.24 \pm 1.87}$ | $83.43 \pm 1.72$ | $80.37 \pm 1.98$ | $78.82 \pm 1.07$ | $72.93 \pm 4.00$ |
| **Ours** (clean, then train) | $94.80 \pm 0.06$ | $\mathbf{93.90 \pm 0.12}$ | $\mathbf{91.69 \pm 0.20}$ | $\mathbf{93.93 \pm 0.39}$ | $\mathbf{93.28 \pm 0.06}$ | $\mathbf{91.95 \pm 0.46}$ |

Table 15: Downstream accuracy (%) on CIFAR-100 under synthetic instance-dependent noise, at the operating points of Table 1. Three seeds on one recipe and one schedule.

| Benchmark | $\tau_d/\tau_r$ | Accuracy |
|---|---|---|
| CIFAR-100 IDN 20% | 12/6 | $73.23 \pm 0.19$ |
| CIFAR-100 IDN 30% | 12/6 | $70.39 \pm 0.21$ |
| CIFAR-100 IDN 40% | 11/6 | $67.15 \pm 0.32$ |

## J LABEL-FREE THRESHOLD SELECTION

Section 5.1 sets every reported $(\tau_d, \tau_r)$ by a grid search that retrains a downstream classifier at each grid point. The rule below needs no retraining. It reads only the histogram of disagreement counts, thirteen integers the scoring pass already produces, and fixes $\tau_r = 6$. Its robustness checks cover ten of the eleven labelled cells, all but CIFAR-100 IDN 40%.

**The rule.** Fit a two-component binomial mixture to that histogram by EM,

$$P(m) \;=\; (1-\pi)\,\mathrm{Bin}(m; K, p_{\mathrm{clean}}) \;+\; \pi\,\mathrm{Bin}(m; K, p_{\mathrm{noisy}}), \tag{4}$$

read off the mixing weight $\hat{\pi}$, and take the smallest $\tau_d$ whose flagged fraction does not exceed it, $\tau_d = \min\{\tau : N^{-1}\sum_i \mathbf{1}[m_i \geq \tau] \leq \hat{\pi}\}$, which is well defined because that fraction is non-increasing in $\tau$. The guard $\tau_d \geq 6$ applies and $\tau_r = 6$ throughout.

**Fitting the mixture.** Initialization is deterministic and identical on every cell, $(\pi, p_{\mathrm{clean}}, p_{\mathrm{noisy}}) = (0.30, 0.05, 0.95)$; EM stops at $|\ell_t - \ell_{t-1}| < \epsilon \max(1, |\ell_t|)$, $\epsilon = 10^{-13}$, or 2,000 iterations; probabilities are clipped to $[10^{-12}, 1 - 10^{-12}]$; the components take canonical order ($p_{\mathrm{noisy}} > p_{\mathrm{clean}}$) only after convergence; one fit is performed per cell.

**Why the mixing weight and not the posterior.** The twelve views are correlated perturbations of one image, so Equation 4 is not a calibrated per-example model; it estimates aggregate prevalence, not individual error probabilities. On all ten cells $\hat{\pi}$ is within 5.1 percentage points of the true noise rate. The empirical crossing is the first bin in which more than half the samples are truly noisy, and the rule's threshold lands within one bin of it on all ten, with mean absolute error 0.70 bins.

Table 16: The label-free rule (LF) against the reported operating point (Rep., Table 1). F1 is the rule's detection F1, and to ceiling is its gap to the maximum F1 over $\tau_d \in [1, K]$, which equals Table 12's Best F1 on every cell. $\Delta$ correct is the change in retained correct labels against no cleaning, and $\Delta$ vs rep. the same count against the reported point. CIFAR-100 IDN 40% (last row) has this comparison only; the robustness checks cover the other ten cells.

| Benchmark | Rep. $\tau_d$ | LF $\tau_d$ | F1 | to ceiling | $\Delta$ correct | $\Delta$ vs rep. |
|---|---|---|---|---|---|---|
| CIFAR-10 IDN 20% | 10 | 9 | 0.9413 | 0.0000 | +8,478 | −20 |
| CIFAR-10 IDN 30% | 10 | 9 | 0.9503 | 0.0000 | +12,448 | −53 |
| CIFAR-10 IDN 40% | 9 | 9 | 0.9421 | −0.0018 | +16,263 | 0 |
| CIFAR-100 IDN 20% | 12 | 10 | 0.8056 | −0.0032 | +3,520 | −767 |
| CIFAR-100 IDN 30% | 12 | 10 | 0.8243 | 0.0000 | +5,096 | −689 |
| Fashion-MNIST IDN 20% | 10 | 9 | 0.9546 | −0.0018 | +10,667 | −81 |
| Fashion-MNIST IDN 30% | 9 | 8 | 0.9546 | −0.0011 | +15,495 | −134 |
| Fashion-MNIST IDN 40% | 9 | 9 | 0.9403 | 0.0000 | +19,541 | 0 |
| CIFAR-10N | 9 | 7 | 0.7412 | 0.0000 | +1,608 | −356 |
| CIFAR-100N | 10 | 9 | 0.7005 | −0.0262 | +398 | −723 |
| CIFAR-100 IDN 40% | 11 | 10 | 0.8122 | −0.0021 | +6,124 | −522 |

**What it delivers.** The rule reaches the F1 ceiling exactly on five of the eleven cells (Table 16), and its largest gap is 0.026, on CIFAR-100N. Across 200 EM restarts the fit returns the same $\tau_d$ on all ten checked cells, and 200 bootstrap resamples return the modal $\tau_d$ in at least 97.0% of draws. On subsamples down to 5% of the rows, the chosen $\tau_d$ never moves by more than one bin.

**A validity condition on the histogram.** Fed only the clean rows of CIFAR-10N, the fit still returns a confident threshold. It estimates $\hat{\pi} \approx 10\%$, flags $3.4\%$ of samples, is fully stable across restarts (1.000) and fits the histogram *better* than on any labelled cell. Only a statistic of the histogram's shape detects that. With the two tallest local maxima at $m_{\mathrm{lo}} < m_{\mathrm{hi}}$ satisfying $m_{\mathrm{hi}} - m_{\mathrm{lo}} \geq 2$ and $m_*$ the argmin strictly between them, the bimodality ratio $\rho = h[m_*]/\min(h[m_{\mathrm{lo}}], h[m_{\mathrm{hi}}])$ must be defined and satisfy $\rho < 0.5$ before we select; otherwise we decline. The ten labelled cells sit at $\rho \leq 0.410$ and Animal-10N at $0.452$, so all pass. The clean-only input reads $0.736$ and a single-mode histogram leaves $\rho$ undefined, so the rule declines both.

On Animal-10N at its native $64 \times 64$, $\rho = 0.452$ passes the gate, $\hat{\pi} = 0.142$, and the rule selects $\tau_d = 8$.

## K Implementation Details

**Detector.** The backbone is a torchvision ResNet-34, ImageNet-pretrained, with the $7\times7$ stride-2 stem replaced by a $3\times3$ stride-1 convolution and the stride-2 max-pool replaced by the identity, so that `layer4` closes at $4\times4$ on a $32\times32$ input rather than $1\times1$. Global average pooling gives a 512-dimensional feature, which an MLP head maps to the embedding through $\mathrm{Linear}(512, 512)$, ReLU, dropout $0.3$, $\mathrm{Linear}(512, 128)$ and $\ell_2$ normalization, and the classifier is a single linear layer on that normalized embedding. Total parameters are $21.61$M, of which $21.28$M are the trunk.

Both images of a pair pass through the same weights, and each is augmented independently with a random crop to the working resolution with padding one eighth of it, a horizontal flip with probability $0.5$, RandAugment at its default two operations and magnitude $9$, normalization, and random erasing with probability $0.25$. Validation uses normalization alone, and scoring uses the battery of Appendix E. The objective is $\lambda\mathcal{L}_{\mathrm{contrastive}} + \mathrm{CE}(\mathrm{logits}_1, y_1) + \mathrm{CE}(\mathrm{logits}_2, y_2)$ with $\lambda = 1$, contrastive margin $\gamma = 2$ on Euclidean distances between normalized embeddings (the diameter of the unit sphere), and no label smoothing at this stage. Training uses AdamW at learning rate $3.969 \times 10^{-4}$ and weight decay $5.784 \times 10^{-4}$, cosine annealed to $2\%$ of the initial rate, effective batch $256$, on a stratified $90/10$ split of the noisy labels, with the scoring checkpoint taken at best validation accuracy and early stopping at patience $10$. These optimization settings are the transferred CIFAR-10 winner and are not tuned per cell. They come from an Optuna search scored by validation accuracy against the noisy labels, so no clean label enters it. What varies with dataset size alone is the epoch budget and the number of sampled pairs, drawn balanced one to one between same-class and different-class. The budgets are $20$ epochs and $120{,}000$ pairs on the CIFAR-10 cells and CIFAR-10N, $25$ and $150{,}000$ on the CIFAR-100 cells, $20$ and $144{,}000$ on Fashion-MNIST, $30$ and $200{,}000$ on CIFAR-100N and Animal-10N.

**Datasets as loaded.** Fashion-MNIST carries $60{,}000$ training images where the other cells carry $50{,}000$, and is bicubic-resized from $28\times28$ to $32\times32$ and its single channel replicated to three, which keeps the pretrained weights usable; Animal-10N is kept at its native $64\times64$. Both conversions happen in the dataset class at load time rather than in a transform, so the view battery is unaffected, and normalization constants are measured through that same load path. At $64\times64$ the spatial parameters of the battery are rescaled so that each view removes the same fraction of the frame, while angles, area fractions and sharpness factors are held fixed.

**Downstream classifier.** Training uses a $100$-epoch cosine schedule, on a stratified $10\%$ validation split of the cleaned set redrawn per seed, with three seeds (42, 43, 44) per arm. Nothing is pretrained. Augmentation is $\mathrm{RandomCrop}$ with padding $S/8$, horizontal flip, $\mathrm{RandAugment}(2, 9)$, normalization and $\mathrm{RandomErasing}(p{=}0.25)$; evaluation normalizes only. The architecture is a PreActResNet-34 on every cell except the three Fashion-MNIST cells, which use a PreActResNet-18. Recipes were searched with Optuna TPE, scored on *validation* accuracy of the cleaned set, with the winner re-run at the full schedule; the test split never influences checkpoint selection or recipe choice. Five cells carry a search of their own (CIFAR-10 IDN 20%, Fashion-MNIST IDN 20%, CIFAR-10N, CIFAR-100N and Animal-10N), and the rest transfer one. Most cells run Adam at learning rate $7.476 \times 10^{-4}$, weight decay $2.342 \times 10^{-4}$, label smoothing $0.05$ and three warmup epochs. Fashion-MNIST IDN 40% and Animal-10N run Adam at $1.596 \times 10^{-3}$ with weight decay $1.209 \times 10^{-6}$, no label smoothing and five warmup epochs. Fashion-MNIST IDN 20% is the one cell on SGD, at learning rate $0.1307$, weight decay $3.245 \times 10^{-4}$, momentum $0.9$ and label smooth-

ing 0.1; CIFAR-10N runs Adam at $4.395 \times 10^{-3}$ with weight decay $1.057 \times 10^{-6}$, label smoothing 0.05 and five warmup epochs.

**Seeding.** A single global seed of 42 drives noise realization, the detector's train/validation split, contrastive pair sampling, view augmentation and weight initialization, and the scoring pass reseeds to it before the first augmentation, so replaying a checkpoint reproduces its counts exactly. Downstream arms run seeds 42, 43 and 44. Cleaning is computed once from the cached per-sample record and the recipe search runs once, so the seeds differ only in weight initialization, batch order and the validation split. For synthetic noise each sample draws its own flip probability from a truncated normal centered at the nominal rate with standard deviation 0.1.

## L COMPUTE

Everything in Table 17 ran on a single NVIDIA Tesla T4 on hosted notebook instances. Animal-10N runs at $64 \times 64$, four times the pixels of a $32 \times 32$ cell, and takes 2.9 times as long as CIFAR-100N, which has the same epoch and pair budget.

Table 17: Detection cost on a single T4 for Siamese training, the $K = 12$ scoring pass, and their total.

| **Benchmark** | **Siamese** (min) | **TTA** (min) | **Total** (h) |
|---|---|---|---|
| CIFAR-10 IDN 20% | 90 | 10.2 | 1.7 |
| CIFAR-10 IDN 30% | 95 | 11.1 | 1.8 |
| CIFAR-10 IDN 40% | 83 | 11.3 | 1.6 |
| CIFAR-100 IDN 20% | 138 | 10.9 | 2.5 |
| CIFAR-100 IDN 30% | 123 | 9.8 | 2.2 |
| CIFAR-100 IDN 40% | 117 | 10.1 | 2.1 |
| Fashion-MNIST IDN 20% | 131 | 13 | 2.4 |
| Fashion-MNIST IDN 30% | 135 | 13 | 2.5 |
| Fashion-MNIST IDN 40% | 133 | 13 | 2.4 |
| CIFAR-10N | 91 | 10.0 | 1.7 |
| CIFAR-100N | 169 | 10.1 | 3.0 |
| Animal-10N | 507 | 24.1 | 8.8 |