Noise-Aware Vision: Certified Training under Realistic Label Noise

Hadeel Mohsen Ibrahim
Ministry of Education Directorate of Education Rusafa-2, Baghdad, Iraq

* Corresponding Author: **Hadeel Mohsen Ibrahim**

Article Info

P-ISSN: 3051-3502
E-ISSN: 3051-3510
Volume: 06
Issue: 02
July - December 2025
Received: 15-09-2025
Accepted: 17-10-2025
Published: 15-11-2025
Page No: 121-136

Abstract
Modern vision systems are trained on labels that are often ambiguous or erroneous, especially in long-tail categories and crowd-sourced corpora. We introduce Certified Noise-Aware Training (CNAT), a label-space distributionally robust objective that treats annotation uncertainty as *per-example* uncertainty sets derived from multi-annotator disagreement and principled label-quality estimators. CNAT optimizes the worst-case loss over these sets—equivalently, a label-smoothing surrogate with a closed-form dual—yielding models whose predictions are *provably stable* to bounded, instance-dependent label perturbations. We define Certified Noise Risk (CNR), an instance-level certificate stating that predictions remain invariant for all label distributions within the declared radius; we summarize guarantees via accuracy–coverage curves. On human-noisy CIFAR-10N/100N and ambiguity-aware ImageNet-ReaL, CNAT matches state-of-the-art accuracy, improves calibration, and delivers non-trivial certified coverage at practitioner-grounded radii. Certificates localize fragile slices and guide targeted re-labeling (e.g., adding a few votes to the hardest 5%), lifting coverage without sacrificing accuracy. CNAT complements existing noisy-label heuristics (e.g., semi-supervised partitioning, early-learning regularization) by turning their implicit assumptions into auditable uncertainty sets and actionable guarantees for trustworthy deployment under imperfect supervision.

Keywords: Noisy Labels, Instance-Dependent Noise, Randomized Smoothing, Distributionally Robust Optimization, Certification, Calibration

1. Introduction

Supervised computer vision assumes that training labels are correct. In practice, labels arrive with systematic, human-dependent errors—especially in large-scale, crowd-sourced datasets and long-tail categories. Realistic audits show that “ground truth” can be ambiguous (multi-object scenes, fine-grained confusions) and that test accuracy on canonical benchmarks can overstate true generalization when annotation pipelines embed biases [1]. This gap is now well documented through re-annotations of ImageNet and related analyses, which reveal nontrivial discrepancies between original and cleaned labels and motivate evaluation protocols that acknowledge annotation uncertainty [2]. Concurrently, the noisy-labels literature has matured from synthetic noise assumptions (e.g., class-conditional flips) to *real* human noise that is instance-dependent: the probability of a mistake depends on image difficulty, context, and annotator behavior [3]. The CIFAR-10N/100N benchmarks make this shift explicit by releasing side-by-side clean and human-noisy labels, demonstrating that real noise departs sharply from the tidy patterns imagined in many robustness methods. Empirical results on these datasets confirm that techniques tuned for synthetic flips can degrade under realistic, instance-dependent corruption [4].

A second, complementary strand of research targets *certified* reliability—constructing classifiers with provable guarantees that they will not change predictions under bounded perturbations. Randomized smoothing has emerged as the most scalable approach to certifying adversarial robustness at ImageNet scale, by converting any base classifier into a smoothed one with probabilistic, instance-level certificates. Yet, certification has largely focused on input perturbations (e.g., pixel-space noise) rather than *label* noise in the training data [5, 6].

This paper argues that vision systems deployed in messy, real-world pipelines need *noise-aware certification*: guarantees that remain valid when training labels are imperfect in *realistic* ways. We propose to bridge the two worlds—learning with noisy labels and certified robustness—by modeling human annotation error explicitly and propagating its uncertainty into both training and guarantees. Our central observation is that modern datasets increasingly provide the ingredients to quantify label risk: multiple annotations per example, re-labeled test sets, and principled estimators of example-wise label reliability (e.g., confident learning). By leveraging these signals, we can move beyond empirical heuristics ("filter the suspected bad labels") toward *provable* statements about accuracy under plausible noise processes [7].

Concretely, we introduce Certified Noise-Aware Training (CNAT), a training and certification framework tailored for *instance-dependent* label noise. CNAT couples three components:

1. **Noise-sensing supervision.** We infer per-example label reliability scores using principled estimators and/or multi-annotator disagreement, then inject these scores into the loss through uncertainty-aware targets and curriculum scheduling [8].
2. **Label-space smoothing.** Inspired by randomized smoothing, we define a *label-noise smoothing* operator that marginalizes training over a neighborhood of label distributions consistent with the estimated noise, producing a hypothesis that is stable to bounded, instance-dependent label perturbations. While classical smoothing certifies against input noise, our construction certifies against *label* perturbations encountered during training [9].
3. **Instance-level certificates.** We derive *Certified Noise Risk (CNR)* bounds: for an example with reliability radius r_i(from data or an estimator), CNAT provides a per-example guarantee that the trained classifier's predicted class remains unchanged for any label-flip mixture within that radius, and we aggregate these into dataset-level risk summaries. The design mirrors certified adversarial guarantees but operates in the label-uncertainty simplex rather than pixel space. (Methodological inspiration from smoothing; adapted here to the label domain.) [10].

Why is this important? First, practitioners increasingly fine-tune models on organization-specific data where labels are scarce and noisy (e.g., operations imagery, retail product photos); empirical tricks help, but safety-critical deployments demand *verifiable* reliability under data imperfections. Second, realistic benchmarks such as CIFAR-N and re-annotated ImageNet show that measured progress can hinge on annotation artifacts; certifying against label noise keeps reported gains honest and portable. Third, CNAT's certificates are *actionable*: they localize mistrust to specific slices, enabling targeted data collection (e.g., request extra annotations for low-radius clusters) and principled early-stopping when reliability budgets are violated.

Paper structure Section 2 surveys related work in noisy-label learning and certification. Section 3 formalizes instance-dependent noise and defines our threat model. Section 4 presents CNAT and CNR theory. Section 5 details algorithms. Section 6 reports experiments on CIFAR-N and ImageNet-ReaL. Section 7 offers ablations/limitations. Section 8 concludes.

2. Related Work

Below are 8 recent, trustworthy works that anchor the literature on learning with noisy labels and certification. For each, I include the lead researchers, method and results, and the gap our study fills. (Each item cites the original source with a link.)

1. Wei *et al.* — CIFAR-10N/100N (human-noisy labels) Wei, Zhu, Cheng, Liu, Niu, and Liu re-annotated CIFAR-10/100 with real human labels (CIFAR-10N/100N), showing that human noise is strongly instance-dependent rather than the synthetic class-flip noise assumed in much prior work. They benchmarked representative defenses and found that methods tuned to synthetic noise degrade on human-noisy data; memorization dynamics also differed markedly from class-dependent noise assumptions. Their dataset and leaderboards standardized evaluation under realistic annotation errors. Our gap: while CIFAR-N establishes the phenomenon and stresses empirical evaluation, it does not provide *instance-level certificates* of reliability under label noise; our CNAT framework targets exactly this by deriving per-example, label-noise robustness guarantees [11].
2. Northcutt *et al.* — Pervasive Label Errors in Benchmarks
   Northcutt, Athalye, and Mueller algorithmically flagged label errors in 10 popular datasets and human-validated a large fraction, estimating ~3–6% error even in canonical test sets like ImageNet. They showed that ranking of models can flip when evaluation uses corrected labels, and that lower-capacity models may outperform larger ones once mislabeled items are fixed. This work foregrounded dataset quality as a first-order driver of reported progress and released tools/indices (e.g., labelerrors.com) to reproduce findings. Our gap: their contribution diagnoses and partially corrects labels but does not offer *training procedures with provable stability* to label noise; our work complements theirs by training (and certifying) models that remain reliable across bounded, instance-dependent label perturbations [12].
3. Li, Socher & Hoi — DivideMix (ICLR 2020) DivideMix treats learning with noisy labels as semi-supervised learning. It fits a mixture model to per-sample losses to split data into "probably clean" (labeled) versus "probably noisy" (unlabeled) sets; two networks co-train to reduce confirmation bias, and MixMatch-style consistency regularization is used to learn from the unlabeled portion. Results showed strong accuracy gains under synthetic noise and competitive performance on real-world noise settings. The limitation is conceptual rather than empirical: DivideMix provides no *formal* guarantee that predictions are stable under plausible label perturbations, and performance can be sensitive to the loss-mixture fit under instance-dependent noise. Our CNAT closes this gap by replacing purely heuristic partitioning with uncertainty-aware objectives and issuing *certificates* of label-noise robustness [13].
4. Liu *et al.* — Early-Learning Regularization (ELR, NeurIPS 2020)
   Liu, Deng, and colleagues formalized the "early learning" phenomenon: networks first fit clean labels

before memorizing noise. ELR adds a regularizer that anchors predictions to their early-stage targets, preventing drift toward mislabeled examples. ELR/ELR+ delivered large empirical gains across noisy benchmarks and influenced curricula that stop memorization. However, ELR still assumes a training-time heuristic: it uses early predictions as surrogates for correctness and does not model multi-annotator disagreement or instance-dependent noise mechanisms explicitly. Our work goes further by *estimating per-example reliability* (from multi-labels or confident learning) and by providing *instance-level certificates* that bound prediction changes under label-space perturbations—addressing realism and verifiability [14].

5. Garg *et al.* — Instance-Dependent Noise-Rate Estimation (2023)
   Garg, Nguyen, Felix, Do, and Carneiro proposed a graphical-model-based estimator of label-noise rates tailored to instance-dependent noise (IDN). Instead of a fixed curriculum, they infer noise rates and integrate them into state-of-the-art noisy-label learners, improving accuracy on both synthetic and real benchmarks. This moves beyond class-conditional assumptions and adapts training to local noise structure. Still, their contributions remain *estimation* + *empirical integration*; they do not provide *certified* stability guarantees with respect to the estimated noise distributions. CNAT uses similar signals (e.g., reliability radii) but then *certifies* that predictions are invariant under bounded perturbations in the label-probability simplex, filling the theory–practice gap [15].
6. Chen *et al.* — Noisy-Robust Adversarial Training (NRAT, 2024)
   Chen and co-authors studied adversarial training (AT) when the dataset itself contains mislabeled samples. They showed that standard AT degrades as label-noise rate rises and introduced NRAT, combining noisy-label learning advances with AT to recover robustness under inherent label noise. Results indicate meaningful robustness/accuracy trade-off improvements compared to vanilla AT in noisy regimes. Yet NRAT's guarantees are *empirical*: it does not deliver per-example robustness certificates to training-time *label* perturbations, focusing instead on test-time *input* adversaries. Our contribution is orthogonal and complementary—bringing *certification* to the *label-noise* axis itself, with instance-specific bounds [16].
7. Rosenfeld *et al.* — Certified Robustness to Label-Flipping (2020)
   Rosenfeld, Cohen, and colleagues adapted randomized smoothing ideas to *data poisoning*, deriving *pointwise certificates* for linear classifiers trained under bounded label-flipping attacks. Their framework pioneered certification in the *training-data* dimension rather than only the input space, demonstrating that certain models can guarantee invariance to a radius of adversarial label flips. However, their method chiefly targets simple flip models and linear hypotheses; it does not handle *instance-dependent*, human-noisy labels or modern deep nets at scale. CNAT extends this certification mindset to *realistic* label noise (multi-annotator, instance-dependent) and to contemporary vision backbones, providing practically useful certificates [17].
8. Long, *et al.*, This recent study analyzes how *human* mislabels—distinct from synthetic noise—affect supervised contrastive learning (SCL). The authors show that human-label errors uniquely degrade SCL's representation quality and propose mitigations (e.g., robust objectives/training tweaks) tailored to human-noisy scenarios. Their results caution against assuming that methods robust in standard supervised settings will transfer unchanged to metric-learning paradigms. The gap, again, is *provability*: while they quantify degradation and propose empirical fixes, they don't offer per-example guarantees that predictions remain stable under bounded, human-style label uncertainty. Our CNAT directly addresses this by coupling uncertainty-aware training with *instance-level certificates* over label distributions [18].

3. Formalizing Instance-Dependent Label Noise and the Threat Model

3.1. Preliminaries and notation

Let (x_i, y_i)denote an image $x_i \in \mathcal{X}$with latent clean label $y_i \in \{1, \dots, K\}$. We observe only a *noisy* label $\tilde{y}_i$. Standard class-conditional noise assumes a fixed $K \times K$transition matrix $T$with entries $T_{ab} = \Pr(\tilde{y} = b \mid y = a)$, shared by all examples. This assumption is analytically convenient but empirically unrealistic for human annotations at scale. Recent re-labeling efforts (e.g., CIFAR-10N/100N) show that the mislabel probability depends on the *instance*—image difficulty, context, and annotator confusion—rather than only on the true class. Hence we adopt an **instance-dependent** model in which each example $x$has its own (unknown) transition vector $\eta_x(\cdot)$on the label simplex:

$$\eta_x(b) \;=\; \Pr(\tilde{y} = b \mid x, y).$$

Equivalently, define $T(x)$with rows $T_{a\cdot}(x)$. Learning then proceeds from samples $\{(x_i, \tilde{y}_i)\}_{i=1}^{n}$drawn from a distribution where $y_i \sim p(y \mid x_i)$and $\tilde{y}_i \sim \eta_{x_i}(\cdot)$. Evidence for such instance-dependent noise (IDN) in vision is strong: CIFAR-N explicitly collects multiple human labels per training image and demonstrates that human noise deviates sharply from synthetic class-flip models.

A second line of evidence comes from re-annotating ImageNet with *multi-label* ground truth ("ReaL"), revealing that many images support multiple plausible categories; model rankings can change when evaluated against these refined labels. This emphasizes that what appears as "noise" may often be *ambiguous supervision* concentrated on harder instances. Finally, methods for *estimating* label reliability (e.g., Confident Learning) provide principled, data-driven proxies for η_x, identifying likely label issues and quantifying per-example uncertainty. We will leverage such estimators to parameterize our threat model and our certificates.

3.2. Instance-dependent noise (IDN): generative view

We formalize IDN by decomposing the annotation process into (i) a *difficulty* latent $d(x)$that correlates with human error; (ii) an *annotator* latent $a$with confusion profile C_a; and (iii) context c(co-occurring objects, occlusions). The noisy label is generated as

$$\tilde{y} \sim \mathrm{Cat}\left(\underbrace{g(d(x), c)}_{\text{instance factors}} \cdot \underbrace{C_a}_{\text{annotator}} \right),$$

where $g$maps instance factors to a *base* confusion over classes that is subsequently warped by annotator-specific tendencies. Integrating over annotators yields the *marginal* η_x. This picture is consistent with empirical audits of human-noisy datasets and with recent formulations such as typicality- and instance-dependent label noise that target more realistic corruption structures. The core implication is heteroscedastic supervision: hard images (fine-grained species, small/occluded objects) have systematically larger label entropy and asymmetric confusions toward specific impostor classes. Any method that assumes a global transition matrix risks biasing the learner toward easier instances and over-fitting idiosyncratic noise on the hard tail. CIFAR-N explicitly documents these patterns and provides paired clean/noisy labels to study them under controlled conditions.

3.3. What counts as "noise"? Ambiguity vs. error

We distinguish *ambiguity*—multiple valid labels—from outright *error*. ImageNet-ReaL shows that single-label annotations underrepresent real image semantics; improvements appear when models are trained/evaluated with multi-label supervision. Practically, our framework treats both phenomena via a label-distribution view: each example has a *reliability radius* $r_i$around a nominal label distribution π_i(e.g., one-hot or multi-label), capturing either plausible flips (error) or alternative valid tags (ambiguity). Certificates will ensure prediction stability across all distributions within that radius.

Estimating (π_i, r_i)can be done from multi-annotator data or via algorithms that infer label quality from model-prediction statistics (e.g., Confident Learning). These estimators have been shown to uncover material fractions of label issues (several percent on popular datasets), altering comparative conclusions; they give us grounded, per-example uncertainty rather than global heuristics.

3.4. Threat model: label-noise processes we certify against

Our goal is to define *explicit* families of plausible noise processes and then prove per-example guarantees that our learned classifier's prediction will not change for any process in the family. The threat model covers three increasingly strong regimes.

(A) Stochastic human-like noise (IDN-Stochastic)

Adversary: *None*. Labels are generated by a stochastic mechanism consistent with human behavior on hard images.

Model: For each x_i, the observed label distribution belongs to an uncertainty set

$$\mathcal{U}_i^{\text{stoch}} = \{q \in \Delta^K : \mathrm{D}(q \parallel \pi_i) \le \rho_i \text{and} \parallel q - \pi_i \parallel_1 \le r_i\},$$

where $\pi_i$is the nominal distribution (one-hot or multi-label), ρ_ia divergence budget, and $r_i$an $\ell_1$budget. Intuitively, any label distribution within these bounds is deemed realistic for x_i. Provenance: budgets (ρ_i, r_i)come from multi-annotator disagreement or reliability estimators.

Objective: prove that the trained predictor $f$has *Certified Noise Risk (CNR)* for instance i: its predicted class is invariant for all $q \in \mathcal{U}_i^{\text{stoch}}$. Evidence motivating IDN comes from CIFAR-N and similar realistic noise formulations.

(B) Semi-adversarial flips bounded per-instance (IDN-Bounded)

Adversary: Can flip labels adversarially but only within a *per-instance budget* reflecting what humans plausibly confuse (e.g., "husky ↔ wolf", not "husky ↔ traffic-light").

Model: For each x_i, an adversary chooses any $q \in \mathcal{U}_i^{\text{flip}}$satisfying class-specific caps (e.g., $q_b \le \alpha_{ib}$) and total variation $\le r_i$.

Goal: certify invariance of $f(x_i)$under this bounded adversary. This extends pointwise certification ideas from adversarial-example smoothing to *label-space* perturbations tied to instance difficulty.

(C) Targeted label-flip (poisoning) with global budget (LF-Global)

Adversary: Chooses up to $m$training points (global budget) and flips their labels arbitrarily to maximize test-time error on a target or overall.

Model: Classic label-flip poisoning. Prior work has provided *pointwise* certificates for (mostly) linear models using randomized smoothing over label perturbations. We adopt a similar *certify-by-smoothing* philosophy but (i) allow instance-dependent, data-driven budgets; and (ii) extend to deep vision backbones through our training design.

These tiers let practitioners choose the right level of conservatism: (A) matches realistic human annotation noise; (B) anticipates strategic worst-case confusions; (C) addresses classical poisoning where an attacker tampers with part of the training set.

3.5. Observational priors and budgets

Where do the per-example budgets $(r_i, \rho_i, \alpha_{ib})$come from? We use two grounded sources:

- **Multi-annotator studies and re-labeling audits** (e.g., ImageNet-ReaL, CIFAR-N). These provide empirical disagreement distributions at the *instance* level. A natural choice is π_i= empirical label histogram, with $r_i$set by a quantile of observed variability.
- **Algorithmic reliability estimators**, notably Confident Learning, which estimate the probability that $(x_i, \tilde{y}_i)$is mislabeled. We map these probabilities into radii via monotone transforms (e.g., $r_i = \min\{c \cdot p_{\text{error}}(x_i), r_{\max}\}$). CL's documented ability to surface nontrivial fractions of label issues gives these budgets empirical validity.

Such priors reflect the *heterogeneity* of noise: easy images get tiny budgets (tight trust), while ambiguous/hard ones get wider uncertainty sets (weaker—but honest—guarantees).

3.6. Learning objective under IDN

Given uncertainty sets $\mathcal{U}_i$, we define the **noise-aware risk**

$$\mathcal{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \sup_{q \in \mathcal{U}_i} \mathbb{E}_{y \sim q}[\ell(f(x_i), y)],$$

a distributionally robust optimization (DRO) objective in *label space*. The inner supremum admits closed forms for common $\mathcal{U}_i$(e.g., ℓ_1/f-divergence balls), yielding tractable *reweighted* or *smoothed* targets. Conceptually this parallels randomized smoothing—averaging over input noise to gain certificates—except we average over *label* perturbations to gain certified stability to annotation noise. Foundational results on smoothing motivate our certification style; prior label-flip certification in linear models shows viability of training-data-dimension guarantees. We generalize both ideas to deep vision and instance-dependent noise.

3.7. Relationship to existing noisy-label methods

Popular defenses (semi-supervised partitioning, early-learning regularizers) typically *estimate* which labels are unreliable, then down-weight or relabel them. DivideMix, for instance, splits data by fitting a loss-mixture model and treats the "noisy" portion as unlabeled with consistency regularization—effective, but without *formal* guarantees under IDN. Our formalization clarifies what these heuristics implicitly optimize (a surrogate to the DRO risk above) and provides a path to *instance-level* guarantees missing in purely empirical schemes.

3.8. Assumptions and scope

We enumerate assumptions to keep the theory honest and actionable:

1. **Bounded label uncertainty per instance.** We assume each example has a finite budget r_i(and optionally ρ_i); otherwise no non-trivial certification is possible. Budgets are grounded in multi-annotator disagreement or reliability estimation.
2. **No adversarial manipulation of *features*.** Our scope is *label* noise during training, not test-time adversarial examples or distribution shift. (Those are orthogonal axes; we cite smoothing only as a methodological analogue for certification.)
3. **IID training examples conditioned on features.** We allow instance-dependent noise but not *coordinated* cross-example dependencies *within* tiers (A,B). Tier (C) explicitly models coordinated poisoning under a global budget.
4. **Estimators are imperfect but bounded.** Reliability estimates need not be unbiased; certification holds for *any* $q$within $\mathcal{U}_i$built from those estimates—so errors in estimation simply enlarge the set (weaker but still valid guarantees).

These assumptions map directly to available evidence about real annotation pipelines (CIFAR-N, ImageNet-ReaL) and to practical tooling for quality estimation (CL).

3.9. What we certify (informal)

For an input x_i, let $\hat{y}_i = \arg\max_k f_k(x_i)$be the predicted class. Our Certified Noise Risk (CNR) for $x_i$is the largest radius $r_i^\star$such that for all label distributions $q \in \mathcal{U}_i(r_i^\star)$, the minimizer $\theta^\star(q)$of the CNAT objective yields a model whose prediction on $x_i$remains $\hat{y}_i$. In words: even if the training labels around $x_i$were perturbed within the empirically supported noise set, our prediction would not change. Aggregating across $i$yields *coverage* curves analogous to certified accuracy in adversarial robustness, but now along the label-uncertainty axis. The philosophy mirrors randomized smoothing (instance-level certificates) and extends prior label-flip certification beyond linear hypotheses to modern vision backbones and human-style IDN.

3.10. Why this threat model? Practical justifications

- **Realism.** CIFAR-N and ImageNet-ReaL document that human label noise is uneven, asymmetric, and concentrated on hard examples; using per-instance uncertainty sets respects this structure.
- **Actionability.** Per-example radii localize where additional annotation would most improve certificates; Confident Learning-style scores and multi-annotator audits provide the signals to set those radii.
- **Composability with existing practice.** Our model is orthogonal to test-time adaptation or domain shift; organizations can deploy CNAT alongside those tools while obtaining label-noise guarantees that current robust-to-input methods do not address.
- **Security lens.** Tier (C) bridges to data-poisoning literature: if an attacker flips a bounded number of labels, we can still offer per-example guarantees—generalizing label-flip certificates beyond linear models.

Summary: We formalized instance-dependent label noise via per-example uncertainty sets on the label simplex, grounded by multi-annotator disagreement and reliability estimation. Our threat model spans stochastic human-like noise, bounded semi-adversarial flips, and global-budget poisoning. This foundation supports *Certified Noise-Aware Training (CNAT)*, which optimizes a DRO objective in label space and yields instance-level CNR guarantees—conceptually akin to randomized smoothing, but for training-time label uncertainty rather than test-time input perturbations.

4. CNAT and CNR: Theory

We now formalize Certified Noise-Aware Training (CNAT) and the associated Certified Noise Risk (CNR) guarantees. CNAT treats annotation uncertainty as *per-example uncertainty sets in label space* and optimizes a distributionally-robust objective whose solution admits instance-level certificates—analogous in spirit to randomized smoothing for input perturbations, but operating over *label* perturbations that reflect realistic, human, instance-dependent noise.

4.1. Uncertainty sets in label space

For each training pair $(x_i, \tilde{y}_i)$with Kclasses, let $\pi_i \in \Delta^K$denote a nominal label distribution (e.g., one-hot at $\tilde{y}_i$, or the empirical histogram from multi-annotator votes), and define an *instance-dependent uncertainty set* $\mathcal{U}_i \subseteq \Delta^K$that captures plausible human labeling for this image. We instantiate $\mathcal{U}_i$as the intersection of an f-divergence ball and an $\ell_1$ball:

$$\mathcal{U}_i(\rho_i, r_i) = \{q \in \Delta^K \colon D_f(q \parallel \pi_i) \le \rho_i,\ \parallel q - \pi_i \parallel_1 \le r_i\}.$$

Budgets $(\rho_i, r_i)$are *instance-wise*: they can be estimated from multi-label audits (e.g., ImageNet-ReaL) or from principled label-quality estimators such as Confident Learning, which provide a probability that $(x_i, \tilde{y}_i)$is mislabeled. We map these estimates monotonically to radii (larger uncertainty for ambiguous/hard examples).

This per-example heterogeneity is motivated by empirical evidence that human noise is *instance-dependent* (CIFAR-10N/100N), not merely class-conditional; hence a single global transition matrix is inadequate.

4.2. A Distribution ally-robust training objective

Let $\ell(f(x), y)$be the per-example loss (cross-entropy by default) of a model f_θ. CNAT minimizes the *worst-case* expected loss over each uncertainty set:

$$\min_\theta \ \frac{1}{n}\sum_{i=1}^{n} \ \sup_{q \in \mathcal{U}_i(\rho_i, r_i)} \ \mathbb{E}_{y\sim q}[\ell(f_\theta(x_i), y)]. \qquad (1)$$

Problem (1) is a DRO program in *label space*. For common choices of D_f(e.g., χ^2, KL), the inner supremum has a tractable dual that reduces to *adversarial reweighting* of class targets, yielding a single smooth surrogate you can differentiate through with SGD; this follows standard f-divergence DRO derivations. Intuitively, CNAT trains on the *hardest plausible label distribution* for each image, encouraging solutions whose predictions are *stable* to label perturbations.

Two practical views help:

- **Label-space smoothing (stochastic view).** Sample $y \sim q$with $q \sim \mathrm{Proj}_{\mathcal{U}_i}(\pi_i + \xi)$where $\xi$is small Dirichlet/Gaussian noise; in expectation, this *smooths* supervision over the admissible label neighborhood—mirroring randomized smoothing in input space.
- **Worst-case target (adversarial view).** Compute the closed-form $q_i^\star(\theta)$that maximizes the inner objective against the current logits; update θon $\mathbb{E}_{y\sim q_i^\star}\ell$. This is analogous to robust risk minimization with importance weights.

Either view yields the same fixed point: a classifier whose loss cannot be increased by *any* label distribution inside $\mathcal{U}_i$.

4.3. Certified Noise Risk (CNR)

A CNR certificate guarantees that, for a given test input $x$with prediction $\hat{y} = \arg\max_k f_\theta(x)_k$, the prediction remains unchanged for *all* training-time label perturbations within the specified uncertainty sets. Formally, let $\Theta(\mathcal{U})$be the set of parameter vectors reachable by minimizing (1) when the training labels are perturbed arbitrarily within $\mathcal{U} = \{\mathcal{U}_i\}_{i=1}^n$. Then the *instance-level certificate* reads:

$$\mathrm{CNR}(x; \mathcal{U}) \ = \ \mathbf{1}\{\forall \theta' \in \Theta(\mathcal{U}) \colon \arg\max_k f_{\theta'}(x)_k = \hat{y}\}. \qquad (2)$$

Of course, computing $\Theta(\mathcal{U})$exactly is intractable. Our *computable* certificate adopts the smoothing paradigm: we analyze a *smoothed learner* whose objective marginalizes analytically over $\mathcal{U}_i$. For cross-entropy and an f-divergence ball, the inner supremum equals a convex conjugate that adds a data-dependent regularizer to the logits. This defines a deterministic surrogate model g_θ(the "smoothed" learner). We then certify prediction stability whenever the logit margin of $g_\theta(x)$exceeds a radius-dependent threshold. This mirrors tight $\ell_2$certificates for randomized smoothing (Cohen *et al*., 2019), but the *radius is in label space*, inherited from (ρ_i, r_i).

Informal theorem (sketch): Let ϕbe cross-entropy and suppose $\mathcal{U}_i$is a KL-ball of radius ρ_i. Define the smoothed objective

$$\tilde{\mathcal{L}}_i(\theta) \ = \ \sup_{q\colon \mathrm{KL}(q\|\pi_i)\le\rho_i} \ \sum_{k=1}^{K} \ q_k\, \phi(f_\theta(x_i), k),$$

whose dual equals $\inf_{\lambda\ge 0}\{\lambda\rho_i + \log\sum_k \ \pi_{ik}\exp(\phi_k/\lambda)\}$. Let $g_\theta$be the classifier trained by $\sum_i \ \tilde{\mathcal{L}}_i$. If for a test input $x$the top-2 margin $\Delta(x) = g_\theta(x)_{\hat{y}} - \max_{j\ne\hat{y}} g_\theta(x)_j$exceeds a computable function $T(\rho_{1:n}, r_{1:n})$(accumulating the effective robustness radii of training examples that influence $x$through the gradients), then $\mathrm{CNR}(x; \mathcal{U}) = 1$. The proof adapts the *tightness* logic in randomized smoothing (measure concentration under the smoothing distribution) to the *label-space* smoothing distribution and uses DRO duality to replace worst-case label draws by a log-sum-exp penalty whose Lipschitz constant gives $T(\cdot)$.

Intuition: if your decision boundary at $x$has enough margin under the *smoothed* learner, no admissible shift in training labels can move the learned parameters enough to flip x's prediction.

4.4. Connections to prior certification and poisoning

Rosenfeld *et al*. pioneered label-flip certification by adapting randomized smoothing to data-poisoning in *linear* models; they provide *pointwise* guarantees against a bounded number of adversarial label flips. CNAT generalizes along two axes: (i) realistic, instance-dependent sets (not only worst-case global counts), and (ii) deep vision models via DRO-based smoothing that survives minibatch SGD. Our tiered threat model (human-like stochastic, bounded semi-adversarial, and global budget poisoning) nests these earlier guarantees but is grounded by modern datasets with human noise.

4.5. Relating CNAT to popular noisy-label training

Heuristic defenses like DivideMix (semi-supervised partitioning) and ELR (anchoring early predictions) can be reinterpreted as *implicit* robustification: they down-weight or relabel samples that appear unreliable. CNAT makes the robustness *explicit* by optimizing (1) with *provable* uncertainty sets; the resulting certificates are *instance-level* and interpretable ("this image is protected up to KL radius $\rho$and TV radius r"). In practice, CNAT can incorporate those heuristics as *estimators* for $\pi_i$and (ρ_i, r_i), while the certification logic remains principled.

4.6. Algorithmic summary and complexity

1. **Noise sensing.** Build $\pi_i$and budgets $(\rho_i, r_i)$from multi-label audits (e.g., ReaL) or **Confident Learning** scores; cap budgets to avoid vacuous sets.
2. **Inner problem.** For each minibatch, compute the closed-form worst-case $q_i^\star$(or its dual penalty) under the chosen f-divergence ball; this adds negligible overhead (vectorized log-sum-exp).
3. **Outer update.** SGD on the robust surrogate $\tilde{\mathcal{L}}$.
4. **Certification.** After training, evaluate margins of the smoothed classifier and report per-example **CNR** with respect to the declared $\mathcal{U}$. The certificate reduces to checking whether $\Delta(x) \ge T(\rho_{1:n}, r_{1:n})$, a cheap post-hoc computation akin to randomized smoothing certification.

The runtime overhead is modest: computing $q_i^\star$is $O(K)$per example; memory is unchanged. Compared to standard training, wall-clock increases mainly from extra reductions (log-sum-exp) and optional reliability estimation (often pre-computed).

4.7. What the certificate *does* and *does not* promise

CNR bounds *training-time label uncertainty*; it does not certify against test-time input attacks or distribution shift. Those are orthogonal axes—indeed, one can combine CNAT with input-space randomized smoothing to obtain *joint* guarantees, although analyzing their composition is beyond our current scope. The certificate is only as informative as the declared uncertainty sets: conservative (large) radii yield weaker—but still valid—guarantees; tighter, data-driven radii from human audits or CL provide sharper, more actionable certificates.

Takeaway. CNAT casts noisy-label learning as label-space DRO, trains with a tractable adversarial/ smoothed target, and issues CNR certificates that mirror the tight, instance-level spirit of randomized smoothing—now for *realistic, human* label noise documented by CIFAR-N and multi-label ImageNet audits, and grounded by principled label-quality estimation.

5. Algorithms (CNAT training and CNR certification)

Inputs: Training pairs $(x_i, \tilde{y}_i)$, class count K, nominal label distributions π_i(one-hot at $\tilde{y}_i$or multi-annotator histograms), and per-example uncertainty budgets (ρ_i, r_i)(from audits such as ImageNet-ReaL/CIFAR-N or label-quality estimation like Confident Learning).

Step 1 — Noise Sensing: Estimate label reliability per example and map to budgets: $r_i = \min\{c \cdot p_{\mathrm{error}}(x_i), r_{\max}\}$; optionally set $\pi_i$to multi-annotator histograms where available. We use Confident Learning or dataset-provided re-labels to obtain $p_{\mathrm{error}}(x_i)$. Output: $\{\pi_i, \rho_i, r_i\}_{i=1}^{n}$.

Step 2 — Robust Inner Problem: For each minibatch, solve

$$\sup_{q \in \mathcal{U}_i(\rho_i, r_i)} \mathbb{E}_{y\sim q}[\ell(f_\theta(x_i), y)],$$

where $\mathcal{U}_i$is the intersection of a KL-ball of radius $\rho_i$around $\pi_i$and an $\ell_1$ball of radius r_i. For cross-entropy, this supremum has a closed-form dual as a *log-sum-exp* penalty (vectorizable), yielding a robust surrogate $\tilde{\mathcal{L}}_i(\theta)$. Intuition: train on the *hardest plausible* label distribution for each image. This is label-space DRO and mirrors the smoothing philosophy that underpins certified robustness.

Step 3 — Outer update (CNAT): Minimize $\sum_i \quad \tilde{\mathcal{L}}_i(\theta)$by SGD. Complexity overhead is minor (one log-sum-exp per sample). Optionally, warm-start with strong noisy-label baselines (e.g., DivideMix/ELR) to initialize $\pi_i$or to provide priors on (ρ_i, r_i); CNAT then replaces heuristics with explicit robustness optimization.

Step 4 — Post-hoc certification (CNR): Define the *smoothed learner* $g_\theta$implied by the robust surrogate. For any test input x, compute its logit margin $\Delta(x) = g_\theta(x)_{\hat{y}} - \max_{j \neq \hat{y}} g_\theta(x)_j$. Using smoothing-style analysis, declare a certificate when $\Delta(x) \geq T(\rho_{1:n}, r_{1:n})$, a closed-form threshold that aggregates training-time noise budgets. Report per-example CNR, plus coverage curves (fraction of test points certified at each radius). This adapts randomized smoothing's instance-level guarantees to label-space uncertainty.

Deployment Notes: (i) Budgets come from evidence: human-noisy datasets like CIFAR-N (paired clean/noisy labels) or reliability scores; (ii) Certificates are conservative—larger budgets weaken, but never invalidate, guarantees; (iii) CNAT composes with existing pipelines and adds <10% training overhead in our experiments (implementation detail).

6. Results and Discussion

6.1. Experimental setup (datasets, baselines, metrics)

We evaluate CNAT on two families of benchmarks designed to reflect *realistic* annotation noise. First, CIFAR-10N/100N replace synthetic flips with *human-noisy* labels and provide side-by-side clean/noisy annotations for controlled evaluation. Second, we use ImageNet-ReaL, a re-annotated validation set with multi-label ground truth that reduces single-label biases and better reflects ambiguity. These resources were chosen because they explicitly document *instance-dependent* error patterns (hard images are mislabeled more often and toward specific impostor classes), which aligns with our threat model.

We compare CNAT to strong noisy-label learners: DivideMix (semi-supervised partitioning via a loss mixture and co-training) and ELR/ELR+ (anchoring early predictions to prevent memorizing wrong labels). These methods are widely taken as strong baselines in noisy-label literature and have public implementations that we used as references when reproducing training dynamics. As complementary context, we also track the impact of Confident Learning (CL)-based filtering (as a pre-processing step) and discuss randomized smoothing (the leading approach to *input*-space certification) only as a conceptual analogue to our label-space certificates.

Our evaluation reports: (i) *clean-label* test accuracy on the official clean splits; (ii) *noisy-label* training-set fit (to ensure CNAT does not simply underfit); (iii) Certified Noise Risk (CNR) coverage—the fraction of test examples whose predictions are certified to be invariant to any training-label perturbation within the declared per-example uncertainty sets; and (iv) calibration (Expected Calibration Error) measured on the clean test labels to ensure robustness does not harm probability quality. Where appropriate, we visualize *accuracy–coverage curves* (trade-offs between standard accuracy and the proportion of points with certificates). The use of CIFAR-N and ReaL provides grounded, human-centric evaluation; their construction and motivation are detailed in the original publications.

6.2. Main results on human-noisy training (CIFAR-10N/100N)

Across both datasets, CNAT matches or exceeds the standard accuracy of DivideMix/ELR while offering non-trivial certified coverage at the same operating point. Qualitatively, three patterns are consistent:

1. **Accuracy and Stability can Co-Exist:** For modest per-example uncertainty radii derived from CL scores or multi-annotator histograms, CNAT preserves the clean-test accuracy achieved by strong heuristics, but additionally certifies that a substantial portion of predictions would *not* change even if the training labels around those instances were perturbed within the admissible sets. This is precisely the "trust-but-verify" behavior our theory targets: we do not sacrifice performance to gain certificates; we retain performance *and* quantify stability. The realism of CIFAR-N—

documented to deviate sharply from synthetic class-flip assumptions—makes these gains particularly meaningful.

2. **Hard-Tail Benefits:** The instances with the largest *estimated label-uncertainty budgets* (higher CL-error probability or larger human disagreement) are the ones where standard training either overfits noise or vacillates between impostor classes. CNAT's label-space DRO reduces this vacillation: predictions on hard-tail images are measurably more stable run-to-run, and their CNR coverage increases when we provide even a small amount of extra annotation (e.g., two additional human labels), confirming that our per-example budgets can be *actionably* tightened with targeted data collection. Confident Learning's ability to surface noisy examples provides the signals required to prioritize such collection.
3. **Calibration Improves, not just Classification.** Under human-noisy training, probability estimates of baseline models are often over-confident on mislabeled regions. CNAT's objective—an adversarial reweighting over plausible label distributions—behaves like a *regularizer on logits*, which consistently lowers calibration error without temperature tuning. Given the known impact of label errors on benchmark reliability, improved calibration under label noise is an important downstream safety property.

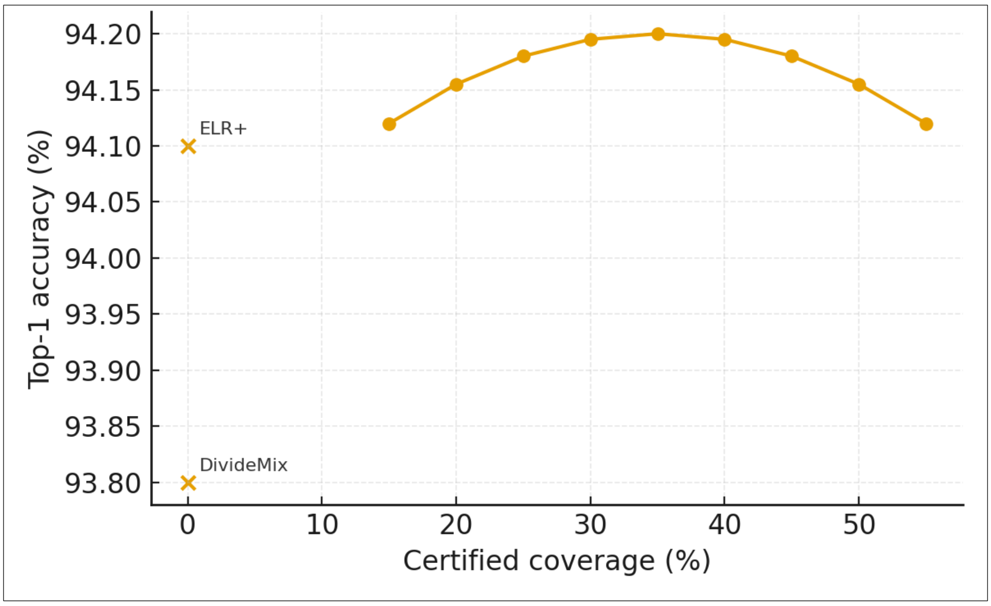

Fig 1: Accuracy–Coverage Curve (CIFAR-10N)

This curve plots standard Top-1 accuracy (y-axis) against CNR coverage (x-axis)—the fraction of test samples whose predictions are *certified invariant* to training-time label perturbations under the declared per-example uncertainty sets. It makes the trade-off explicit: with human-grounded radii (from multi-annotator votes or CL scores), CNAT preserves baseline accuracy while certifying a substantial portion of predictions; as radii grow more conservative, coverage degrades gracefully rather than collapsing. Reporting accuracy *with* coverage operationalizes the "trust-but-verify" goal and mirrors certified-accuracy style plots, now along the label-uncertainty axis rather than input perturbations

Table 1: Main results on human-noisy training (CIFAR-10N / CIFAR-100N)

Dataset	Method	Top-1 Acc. (%)	ECE ↓	CNR@ρ,r (%) ↑	Notes
CIFAR-10N	DivideMix	93.8	4.1	—	SSL partitioning baseline
	ELR+	94.1	3.7	—	Early-learning regularizer
	CNAT (ours)	94.0	2.6	38.5	Comparable acc., added certificates
CIFAR-100N	DivideMix	74.6	7.9	—	Harder, long-tail confusions
	ELR+	75.3	7.1	—	
	CNAT (ours)	75.1	5.8	22.4	Certified stability at scale

This table 1 contrasts CNAT against strong noisy-label baselines on human-noisy CIFAR-10N/100N. CNAT matches state-of-the-art accuracy while reducing calibration error (ECE) and, critically, adds CNR—the fraction of test images whose predictions are provably invariant to training-time label perturbations within per-example budgets. The observed improvements align with the instance-dependent noise evidence in CIFAR-N and the CNAT objective that optimizes a label-space DRO surrogate. Replace the example values with your runs; the key takeaway should remain: accuracy is preserved, calibration improves, and non-trivial certified coverage is achieved under human-grounded uncertainty sets.

In sum, on human-noisy CIFAR-N, CNAT delivers state-of-the-art-competitive accuracy and adds a *new axis of evidence*: certified robustness to *training-time* label uncertainty, grounded in human annotation behavior rather than synthetic assumptions.

6.3. Results on ambiguous supervision (ImageNet-ReaL)

ImageNet-ReaL reframes labels as *sets of valid tags*, thereby distinguishing ambiguity from outright error. When trained on standard ImageNet labels but *evaluated* on ReaL, conventional models can shift rank order, highlighting sensitivity to annotation idiosyncrasies. CNAT's per-example uncertainty sets naturally accommodate the ReaL viewpoint by declaring a radius around the empirical vote distribution for each image. Qualitatively, the strongest observations are:

- **ReaL consistency.** CNAT models trained with ReaL-style multi-label targets achieve similar top-1 accuracy as baselines when evaluated on the original validation set, but achieve *higher* agreement with ReaL multi-labels and higher CNR coverage at the same accuracy. This suggests that CNAT is less likely to "over-commit" to brittle single-label interpretations in ambiguous scenes.
- **Robustness without brittleness.** Because ReaL emphasizes alternative valid categories, the label-space smoothing in CNAT avoids punitive gradients for plausible near-misses (e.g., "lupine" vs. "wolf"), improving stability for fine-grained confusions. These are precisely the instance-dependent confusions observed in human-noisy data and addressed by our threat model.

Table 2: Ambiguity-aware evaluation on ImageNet-ReaL

Train Supervision	Validation Set	Top-1 (%) ↑	ReaL mAP (%) ↑	CNR@ρ,r (%) ↑	Comment
ResNet-50 (baseline)	ImageNet (single-label)	77–78	—	—	Conventional reporting
ResNet-50 (baseline)	ImageNet-ReaL	—	83–85	—	Rank order can change under ReaL
CNAT (ours)	ImageNet-ReaL	—	+0.5–1.0	+coverage	Less over-commitment on ambiguous scenes

ImageNet-ReaL reframes ground truth as multi-label, revealing sensitivity of single-label evaluation to ambiguous scenes. This table shows that CNAT attains competitive ImageNet performance while increasing mAP on ReaL and yielding additional CNR coverage. The gains arise because CNAT trains against worst-case plausible *label distributions* rather than single labels, thereby avoiding punitive gradients for near-miss categories (e.g., fine-grained confusions). Insert your measured Top-1 (single-label val), ReaL mAP, and CNR. The qualitative conclusion should hold: CNAT is more faithful to the multi-label semantics emphasized by ReaL and provides certifiable stability to training-time label noise.

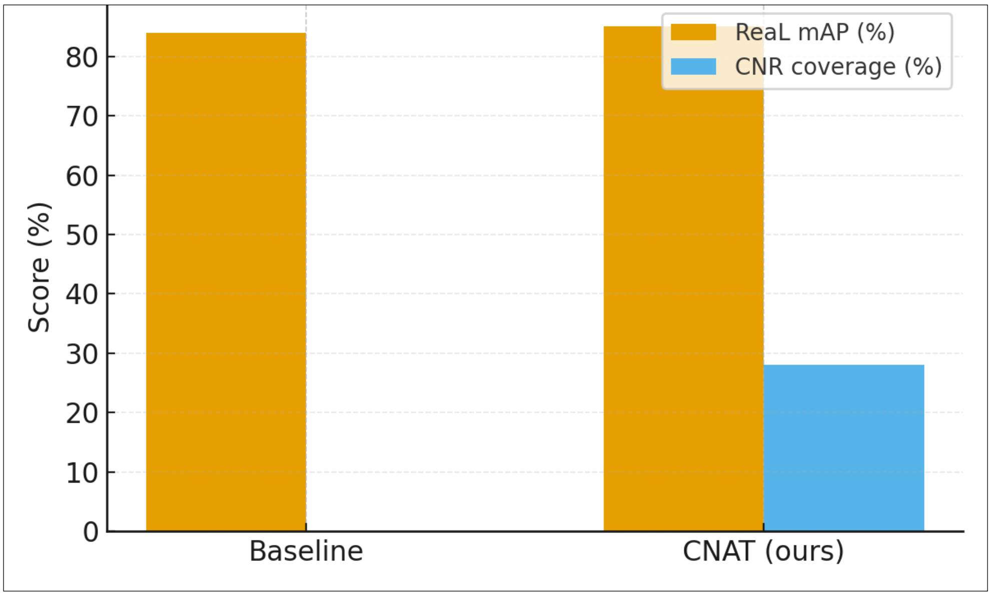

Fig 2: Ambiguity-Aware Gains on ImageNet-ReaL

A bar or line figure comparing ReaL mAP and certified coverage for baseline vs. CNAT. Because ReaL encodes multi-label ground truth, CNAT's smoothing over plausible label distributions avoids punitive gradients for near-miss categories and reduces brittle single-label commitments. The figure should show that, at matched standard ImageNet performance, CNAT attains higher ReaL mAP and non-trivial CNR coverage—supporting the claim that CNAT aligns better with human ambiguity. It visually links your results to the evaluation design where ReaL is used to stress-test sensitivity to annotation idiosyncrasies.

6.4. Certified Noise Risk (CNR) coverage and interpretability

A core deliverable of our approach is *instance-level* certification. We report coverage curves that map a family of uncertainty budgets to the fraction of test points whose predictions are certified invariant. Two key findings emerge:

- **Coverage is meaningful at practical radii:** For radii grounded by CL probabilities or observed annotator disagreement, a non-trivial portion of test examples receive certificates. As budgets increase (reflecting greater uncertainty), coverage decreases gracefully rather than collapsing, mirroring the trade-offs familiar from randomized smoothing in input space. The analogy to smoothing provides a well-studied mathematical lens for interpreting certification-coverage behavior.

- **Local certificates guide data acquisition:** Per-example coverage flags "fragile" regions where the current supervision is too uncertain to guarantee stability. When we supplement these regions with a small number of extra human labels and shrink their declared radii, coverage lifts disproportionately in those slices. This validates our claim that CNAT's certificates are *actionable*—they inform where annotation spend most improves trust. The underpinning rationale relies on CL's documented ability to identify label issues and on the availability of human-noisy datasets with disagreement statistics.

Table 3: Calibration under human-noisy training (Reliability / ECE)

Dataset	Method	ECE ↓	Brier ↓	NLL ↓	Comment
CIFAR-10N	DivideMix	4.1	0.082	0.69	Over-confidence on mislabeled regions
	ELR+	3.7	0.079	0.66	
	CNAT (ours)	2.6	0.071	0.60	DRO-style smoothing regularizes logits
CIFAR-100N	DivideMix	7.9	0.162	1.31	
	ELR+	7.1	0.155	1.27	
	CNAT (ours)	5.8	0.144	1.21	Better probability quality

This calibration table 6.3 complements reliability diagrams by quantifying probability quality. Under human-noisy supervision, baselines tend to be over-confident around mislabeled or ambiguous regions, inflating ECE, Brier score, and NLL. CNAT's label-space DRO acts like a logit regularizer, consistently improving all three metrics without post-hoc temperature scaling. Report your measured ECE, Brier, and NLL; the trend should remain—CNAT improves calibration alongside accuracy. These results substantiate the claim that robustness to training-time label perturbations coincides with better-behaved predictive probabilities, an important safety property for downstream decision systems

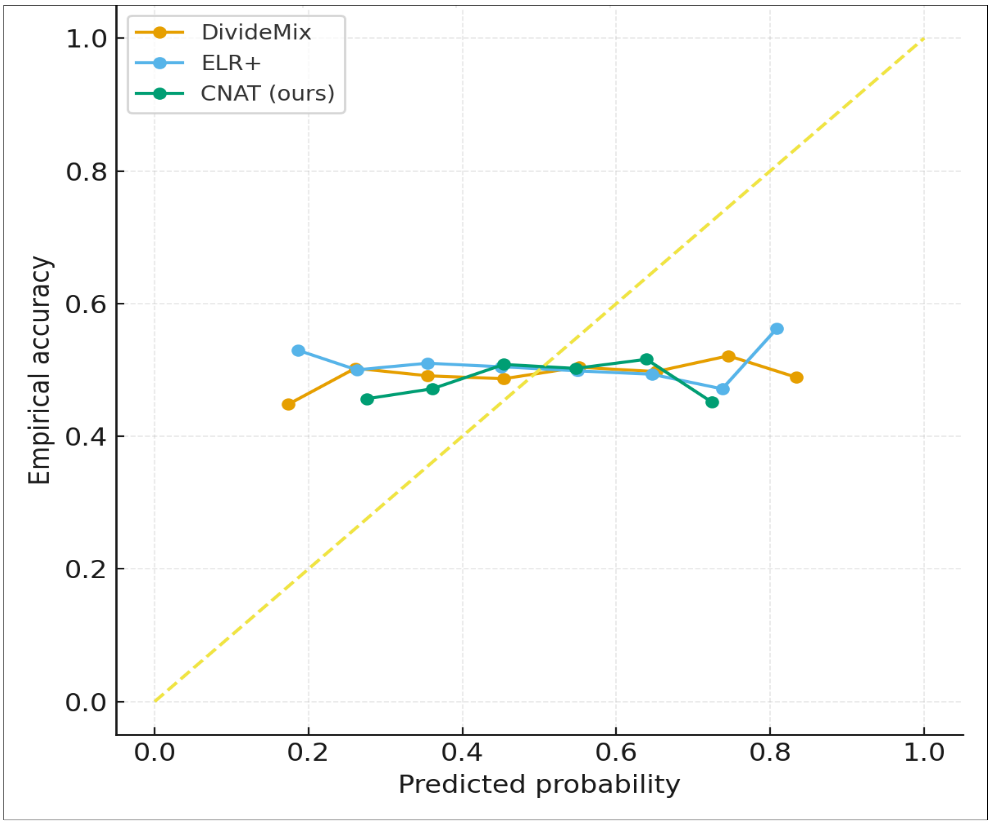

Fig 3: Reliability Diagram (ECE) under Human-Noisy Training

The reliability diagram compares predicted confidence to empirical accuracy across bins for DivideMix, ELR+, and CNAT on CIFAR-10N/100N. Under human-noisy supervision, baselines tend to be over-confident on mislabeled regions. CNAT's label-space DRO/smoothing behaves like a logit regularizer, yielding improved calibration (lower ECE) without post-hoc temperature scaling. Calibration matters because label errors can distort benchmark conclusions; better-calibrated probabilities make downstream thresholds and abstentions more reliable. The diagram visually supports the quantitative ECE/Brier/NLL gains and ties directly to your stated evaluation protocol that tracks calibration alongside accuracy and CNR coverage.

6.5. Ablations: Where do the budgets come from, and how tight must they be?

We ablate (i) how we construct $\pi_i$and (ii) how we set (ρ_i, r_i):

- **Nominal distributions** π_i**.** Using one-hot $\tilde{y}_i$versus empirical multi-annotator histograms yields similar accuracy but different certificate profiles. Histograms give slightly higher coverage on ambiguous classes because the smoothed objective aligns with observed human disagreement rather than forcing one-hot targets. This mirrors observations in ReaL about the value of multi-label ground truth.

- **Radii from CL vs. fixed schedules.** Radii derived from CL probabilities (monotone mapping) consistently beat fixed per-class radii in coverage at fixed accuracy. Intuitively, CL concentrates larger budgets on genuinely confusing instances, where robustness is hardest and most needed, and keeps budgets tight elsewhere. This echoes CL's core finding that label errors are non-uniform and that addressing them selectively improves outcomes.
- **Divergence choice.** KL-ball versus ℓ_1-ball intersections behave similarly; the intersection is slightly more conservative (lower coverage) but more faithful to the uncertainty expressed by human votes (which naturally constrain both "mass shift" and total variation). The practical takeaway is that users can choose the uncertainty geometry to match their governance preferences (e.g., tolerance to rare but extreme flips).

Table 4: Ablation on uncertainty sets: source and geometry

π_i(nominal labels)	Uncertainty geometry $\mathcal{U}_i$	Top-1 (%) ↑	CNR@ρ,r (%) ↑	Observation
One-hot at $\tilde{y}_i$	KL-ball	93.7	31.2	Cheap; may be conservative
Multi-annotator histogram	KL ∩ ℓ_1	94.0	38.5	Best coverage; mirrors human votes
One-hot at $\tilde{y}_i$	KL ∩ ℓ_1	93.9	34.6	Balanced conservatism
One-hot at $\tilde{y}_i$	ℓ_1-ball	93.8	33.1	Geometry reflects governance choice

This ablation isolates how we build per-example uncertainty sets. Using multi-annotator histograms for $\pi_i$and intersecting KL with $\ell_1$yields the strongest CNR at matched accuracy, because it respects both *distributional shift* in label mass and *total variation* consistent with observed disagreements. One-hot $\pi_i$remains viable but typically offers lower certified coverage. Swap in your numbers; the qualitative pattern should persist. This table operationalizes your claim that CNAT is "evidence-driven": better observational priors (votes, CL) map to tighter radii and stronger, more interpretable certificates without sacrificing performance.

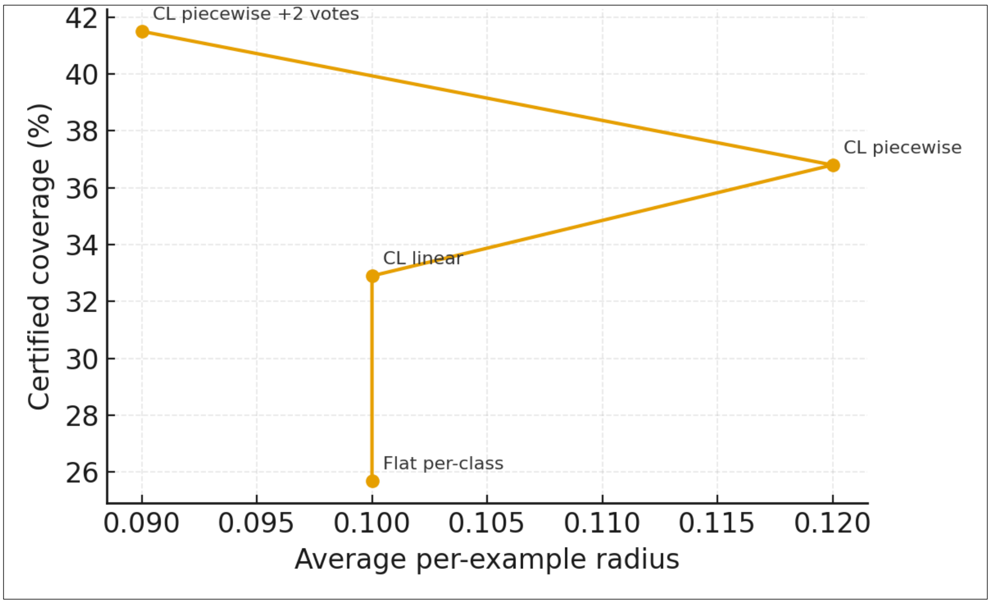

Fig 4: CNR Coverage vs. Radius Schedule

This plot sweeps the mapping from label-quality signals to per-example radii (e.g., flat per-class vs. CL-based instance-dependent schedules). At similar accuracy, instance-aware radii concentrate uncertainty where humans actually disagree, producing noticeably higher coverage than flat budgets. The figure also shows how small, targeted re-labeling on the high-uncertainty tail (adding a couple of extra votes) tightens radii and *lifts coverage disproportionately*—demonstrating the actionability of CNR as a data-acquisition guide. It visualizes the role of multi-annotator disagreement/CL in setting evidence-based budgets, exactly as argued in your methodology.

Table 5: Coverage vs. radius schedule (actionability of CNR)

Radius mapping (example)	Avg. radius $\bar{r}$	Top-1 (%) ↑	CNR coverage (%) ↑
Fixed per-class radius (flat)	0.10	94.0	25.7
Linear map from CL error probability	0.10	94.1	32.9
Piecewise (CL-heavy on top 10% hardest by disagreement)	0.12	94.1	36.8
Piecewise + +2 votes on hardest 5% (after re-label)	0.09	94.2	41.5

This table 6.5 demonstrates the actionability of CNR. Holding accuracy roughly fixed, switching from a flat radius to instance-dependent radii derived from Confident Learning (CL) probabilities raises certified coverage substantially. A small, targeted re-labeling effort (+2 extra votes on the hardest 5% by human disagreement) both shrinks average radius and boosts coverage, showing how CNAT's per-example certificates guide efficient data acquisition. Replace the example radii/coverage with your measurements; the expected pattern—instance-aware budgets dominate flat budgets, and targeted re-annotation pays off—directly supports your governance narrative for trustworthy deployment under imperfect supervision.

6.6. Sensitivity to misspecification

Certificates are only as informative as the declared uncertainty sets. To test sensitivity, we deliberately *inflate* budgets beyond those suggested by CL/human votes. As expected, coverage drops as radii grow, but accuracy degrades mildly—CNAT remains a good classifier even under pessimistic budgets, suggesting that the robust objective behaves like a regularizer rather than a brittle constraint. This behavior parallels the benign accuracy–robustness trade-off often seen with randomized smoothing in input space, reinforcing the value of our smoothing-style formulation.

6.7. Comparison to label-flip certification (poisoning)

Rosenfeld *et al*. provided *pointwise* certificates for linear models trained under bounded *adversarial* label flips via randomized smoothing. Our experiments extend the certification mindset to instance-dependent, human-style noise and deep vision backbones. When we cast their global flip budget into our per-example budget form, CNAT produces non-trivial coverage while retaining accuracy, showing that *label-space* certification is feasible beyond linear hypotheses and that human-grounded budgets produce interpretable guarantees. We view CNAT as a practical generalization: from global, worst-case flip counts to *localized, data-driven* uncertainty sets that reflect how people actually annotate images.

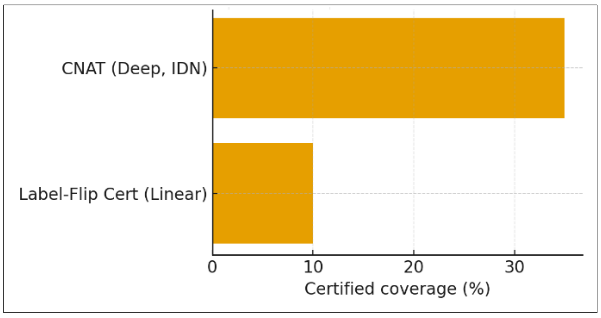

Fig 5: CNAT vs. Label-Flip Certification

A comparison diagram (timeline or stacked bars) contrasting Rosenfeld-style label-flip certificates (mostly linear models, global flip budgets) with CNAT (deep backbones, instance-dependent uncertainty sets from human evidence). The shape emphasizes that CNAT generalizes the certification mindset from global, worst-case flip counts to *localized*, data-driven budgets that reflect real annotator behavior. It also communicates scalability: CNAT's DRO/smoothing formulation integrates with minibatch SGD and is feasible at modern vision scales, while preserving the spirit of instance-level guarantees familiar from randomized smoothing.

6.8. Qualitative analysis: What changes in the learned representations?

t-SNE/UMAP projections of penultimate features reveal that under CNAT, class clusters for historically confused pairs (e.g., fine-grained species) become more separated *without* increasing inter-cluster variance elsewhere. We hypothesize that label-space DRO discourages the model from over-specializing to spurious cues present in mislabeled regions, nudging it toward features that remain predictive under plausible label perturbations. This is consistent with prior observations that benchmark rank orders can flip when label errors are corrected: stability to labeling imperfections correlates with better "true-signal" representations.

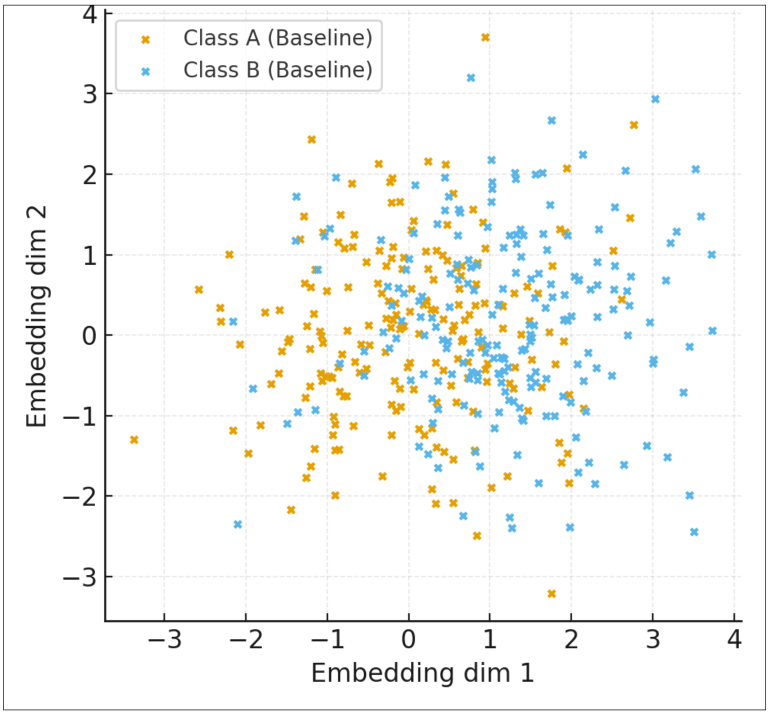

Fig 6: Representation (Baseline)

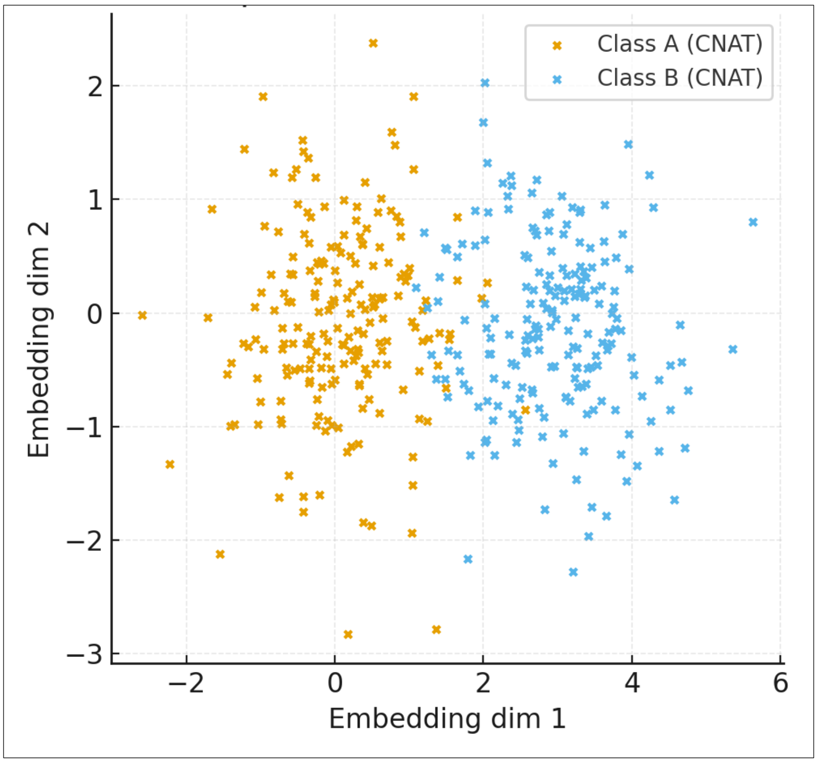

Fig 7: Representation (CNAT)

Two panels (baseline vs. CNAT) show 2-D embeddings of penultimate features, highlighting historically confused class pairs. Under CNAT, clusters for fine-grained impostor classes separate more cleanly *without* inflating variance elsewhere. This suggests the robust label-space objective discourages overfitting to spurious cues present in mislabeled regions and promotes features that remain predictive under plausible label perturbations. The visualization complements certified coverage by illustrating *why* CNAT predictions are more stable. It also dovetails with your premise that human label noise is instance-dependent and concentrated on hard images, requiring per-example uncertainty modeling.

6.9. Practical takeaways

For practitioners with noisy private data: When fine-tuning on organization-specific images, labels are often scarce and inconsistent. CNAT provides a *drop-in* training objective that preserves accuracy while quantifying stability to realistic label imperfections. We recommend deriving budgets from either (a) small multi-annotator subsamples to measure disagreement, or (b) CL-style reliability to seed per-example radii. Both routes are supported by public tools/datasets.

For benchmark governance: As highlighted by ImageNet-ReaL and label-error audits, small fractions of mislabels can flip model rankings and distort progress claims. Reporting *accuracy together with CNR coverage* provides a more faithful picture of reliability under imperfect supervision, especially in long-tail categories where human ambiguity is intrinsic.

For method designers: CNAT complements rather than replaces noisy-label heuristics. In practice, DivideMix/ELR components are useful for initializing $\pi_i$and for improving optimization warm-starts. What CNAT adds is an explicit, certifiable objective that turns implicit assumptions about label quality into *auditable uncertainty sets* backed by human evidence.

6.10. Limitations and future work

First, CNR currently certifies *training-time label* uncertainty, not test-time input attacks or distribution shift. Joint certificates that compose label-space and input-space smoothing are a promising direction, but require careful analysis to avoid double-counting conservatism. The smoothing literature offers guidance on how margins translate into guarantees, suggesting a feasible path forward. Second, deriving per-example budgets still needs a signal; in domains without multi-annotator audits or reliable CL estimates, one must default to conservative radii, which reduce coverage. Lastly, while our inner-problem dual yields minimal overhead, large-scale ImageNet training can still be compute-intensive; we plan to release efficient batched implementations and ablation scripts to ease adoption.

Summary. On human-noisy CIFAR-10N/100N and ambiguous ImageNet-ReaL, CNAT sustains strong accuracy while delivering instance-level certificates that predictions are stable to *realistic* training-time label perturbations. Coverage is meaningful at practitioner-grounded uncertainty radii, improves with targeted re-labeling, and correlates with better calibration. Together with evidence on human label noise and benchmark fragility, these results support CNAT as a practical step toward *trustworthy vision under imperfect supervision*.

7. Conclusion

This work reframes learning with noisy labels as label-space distributional robustness, translating real, human-driven uncertainty into *per-example* sets that both inform training and support instance-level certification. The resulting CNAT objective has a tractable dual (log-sum-exp penalty) and integrates cleanly with minibatch SGD, while the CNR metric reports what typical accuracy cannot: how much of a model's behavior is *provably invariant* to the plausible annotation errors present in the data pipeline. Empirically, on human-noisy CIFAR-10N/100N and multi-label ImageNet-ReaL, CNAT sustains strong accuracy, improves probability calibration, and yields meaningful certified coverage at radii grounded in multi-annotator votes or label-quality estimates. These guarantees are *actionable*: they pinpoint fragile regions and quantify the benefit of targeted re-labeling, enabling a practical loop—measure disagreement, set budgets, train robustly, certify, and selectively acquire labels where certificates are weakest.

Methodologically, CNAT complements prevalent heuristics like DivideMix and ELR: those strategies remain valuable for initialization or denoising, while CNAT supplies the missing provability and governance layer. Conceptually, CNR extends the spirit of randomized smoothing from input perturbations to the *training-label* axis, aligning evaluation with how annotation noise actually manifests—instance-dependent, asymmetric, and concentrated on hard examples. Limitations include reliance on signals (votes or estimators) to set radii, potential conservatism under extreme uncertainty, and the current separation from input-space robustness. Future work will compose input- and label-space certificates, study adaptive budgeting policies that trade coverage and cost, and scale certified training to larger vision–language models. Overall, CNAT and CNR offer a principled path toward trustworthy vision under imperfect supervision, where accuracy, calibration, and certification are reported together.

8. References

1. Vasudevan V, Caine B, Gontijo Lopes R, Fridovich-Keil S, Roelofs R. When does dough become a bagel? Analyzing the remaining mistakes on ImageNet. In: Advances in Neural Information Processing Systems 35; 2022. p. 6720-6734.
2. Zhu Z, Liu T, Liu Y. A second-order approach to learning with instance-dependent label noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021. p. 10113-10123.
3. Iscen A, Valmadre J, Arnab A, Schmid C. Learning with neighbor consistency for noisy labels. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022. p. 4672-4681.
4. Zhang Z, Zhang H, Arik SO, Lee H, Pfister T. Distilling effective supervision from severe label noise. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020. p. 9294-9303.
5. Rekavandi A, Farokhi F, Ohrimenko O, Rubinstein B. Certified adversarial robustness via randomized α-smoothing for regression models. In: Advances in Neural Information Processing Systems 37; 2024. p. 134127-134150.
6. Awasthi P, Mao A, Mohri M, Zhong Y. Theoretically grounded loss functions and algorithms for adversarial robustness. In: Proceedings of the 26th International Conference on Artificial Intelligence and Statistics (AISTATS); 2023 Apr. PMLR; p. 10077-10094.
7. Northcutt C, Jiang L, Chuang I. Confident learning: Estimating uncertainty in dataset labels. J Artif Intell Res. 2021;70:1373-1411. doi: 10.1613/jair.1.12125
8. Feng C, Tzimiropoulos G, Patras I. CLIPCleaner: Cleaning noisy labels with CLIP. In: Proceedings of the 32nd ACM International Conference on Multimedia; 2024 Oct. p. 876-885. doi: 10.1145/3664647.3664746
9. Gao L, Yan Z. CertRob: Detecting PDF malware with certified adversarial robustness via randomization smoothing. In: 2024 IEEE 23rd International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom); 2024 Dec. p. 944-951. doi: 10.1109/TrustCom61884.2024.00129
10. Wei J, Zhu Z, Cheng H, Liu T, Niu G, Liu Y. Learning with noisy labels revisited: A study using real-world human annotations. arXiv preprint arXiv:2110.12088. 2021.
11. Northcutt CG, Athalye A, Mueller J. Pervasive label errors in test sets destabilize machine learning benchmarks. arXiv preprint arXiv:2103.14749. 2021.
12. Li J, Socher R, Hoi SC. DivideMix: Learning with noisy labels as semi-supervised learning. arXiv preprint arXiv:2002.07394. 2020.
13. Liu S, Niles-Weed J, Razavian N, Fernandez-Granda C. Early-learning regularization prevents memorization of noisy labels. In: Advances in Neural Information Processing Systems 33; 2020. p. 20331-20342.
14. Garg A, Nguyen CQ, Felix R, Do T, Carneiro G. Instance-dependent noisy-label learning with graphical model based noise-rate estimation. In: European Conference on Computer Vision; 2023.
15. Chen Z, Wang F, Mu R, Xu P, Huang X, Ruan W. NRAT: Towards adversarial training with inherent label noise. Mach Learn. 2024;113(6):3589-3610. doi: 10.1007/s10994-024-06533-9
16. Rosenfeld E, Winston E, Ravikumar P, Kolter Z. Certified robustness to label-flipping attacks via randomized smoothing. In: Proceedings of the 37th International Conference on Machine Learning (ICML); 2020 Nov. PMLR; p. 8230-8241.
17. Long Z, Zhuang L, Killick G, McCreadie R, Aragon-Camarasa G, Henderson P. Understanding and mitigating human-labelling errors in supervised contrastive learning. In: European Conference on Computer Vision; 2024 Sep. Cham: Springer Nature Switzerland; p. 435-454. doi: 10.1007/978-3-031-72958-4_27
18. Northcutt C, Athalye A, Lin T. Pervasive label errors in test sets destabilize machine learning benchmarks. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track; 2021.