STOP REPLACING NOISE WITH NOISE: TWO-SOURCE RELIABILITY ASSESSMENT FOR LABEL CORRECTION AND SAMPLE REWEIGHTING IN LABEL-NOISE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Refurbishment-based noisy-label learning mixes an observed label with a model-derived pseudo target, typically using one sample-wise cleanliness score to control both branches. This creates a hidden coupling: reducing trust in the observed label automatically increases trust in the pseudo target. We show that this complementarity can replace one unreliable signal with another because a pseudo target learned from corrupted supervision may reproduce the noise it is meant to correct. Our representation diagnostics provide a consistent account of this mismatch: noisy supervision redirects deeper layers more strongly, whereas shallower relations remain comparatively stable and provide information beyond the loss posterior. We therefore propose **TRACE**, a **T**wo-Source **R**eliability **A**ssessment framework for Label **C**orrection and Sample R**e**weighting. TRACE assesses the observed label using loss fit, shallow relation stability, and prediction agreement, while separately assessing the pseudo target using model confidence. Its source-specific scores control target correction and supervision strength without assuming complementary reliability. Across synthetic and real-world noisy benchmarks, TRACE improves representative refurbishment baselines and yields more reliable pseudo supervision.

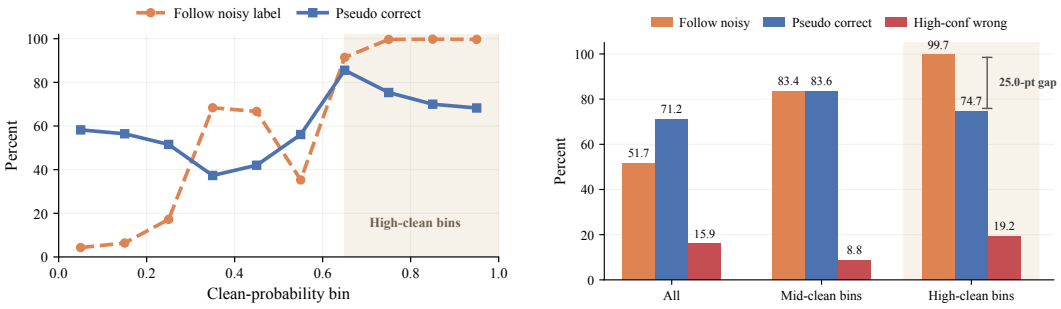
1 INTRODUCTION

Learning with noisy labels remains a central challenge in supervised learning (Natarajan et al., 2013; Papyan et al., 2020). Deep networks can recover semantic structure from corrupted data but also absorb erroneous supervision. The challenge is both to identify corrupted labels and to decide which supervision signal should guide each sample during training. Existing approaches include robust losses, sample selection, and label correction (Ghosh et al., 2017; Han et al., 2018; Zheng et al., 2020). Refurbishment-based methods such as DivideMix (Li et al., 2020), SELFIE (Song et al., 2019), SELC (Lu & He, 2022), and RoLR (Chen et al., 2023) address this decision by mixing a rectified or pseudo target with the observed noisy label.

Despite their differences, many refurbishment-based methods share one control interface: a single sample-wise cleanliness score interpolates between the observed label and a model-derived pseudo target Song et al. (2019); Li et al. (2020); Lu & He (2022); Chen et al. (2023), as formalized in Eq. (6). This collapses two questions into one scalar: how reliable is the observed label, and how reliable is the pseudo target? Because the pseudo branch receives the complementary weight, distrusting the observed label automatically increases the pseudo target's influence.

This coupling risks *replacing noise with noise*. A pseudo target comes from a model trained on the same corrupted supervision, not an external oracle. Having absorbed noisy bias, the model can reproduce the error it is meant to correct, making both supervision sources unreliable. On CIFAR-100N Wei et al. (2022), Fig. 1 shows that pseudo targets in high-clean bins follow the noisy observed label 99.7% of the time, yet are correct on only 74.7% of samples and make 19.2% high-confidence errors. Fig. 8 shows the same qualitative mismatch on a randomly sampled Clothing1M subset. The two sources therefore require separate reliability assessments.

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.



(a) Bin-wise mismatch (b) Across cleanliness regimes

Figure 1: Pseudo-target mismatch on CIFAR-100N. Correctness is evaluated against clean ground truth.

Our representation analysis supports this separation without introducing another learning objective. Consistent with recent studies Wongso et al. (2023); Maini et al. (2023), noisy supervision perturbs deeper representations more strongly, pulls mislabeled samples toward noisy prototypes, and reduces the direct usability of deep features. This pattern helps explain why a pseudo target generated from a noise-trained deep state requires its own reliability assessment. Shallower relations remain comparatively stable, and Fig. 3 shows that relation drift informs observed-label correctness beyond the loss posterior. Shallow structure can therefore refine observed-label reliability without certifying the pseudo target.

Motivated by this two-source view, we propose **TRACE**, a **T**wo-Source **R**eliability **A**ssessment framework for Label **C**orrection and Sample **Re**weighting. Without changing the base pseudo-target generator, TRACE evaluates the two sources separately: the observed-label score combines loss fit, shallow relation stability, and prediction agreement, while the pseudo-target score uses a separate confidence gate. Their composition enables need-aware pseudo supervision, with source-specific weights constructing corrected targets and adjusting each sample's contribution.

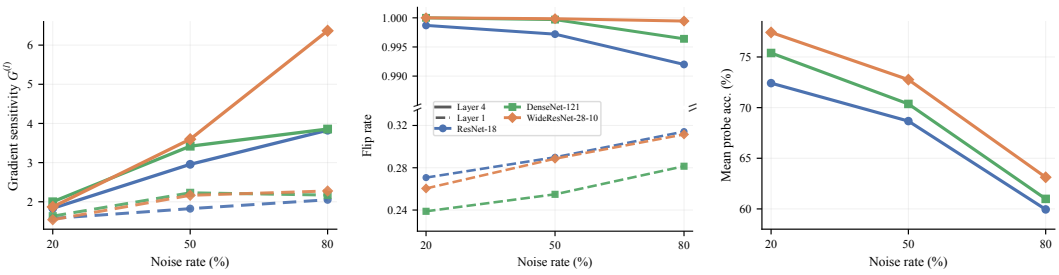
Contributions. Our contributions are threefold: we identify the *replacing-noise-with-noise* risk caused by using one cleanliness score to control both supervision sources; we propose TRACE, a plug-in two-source reliability framework for label correction and sample reweighting; and we show that TRACE improves refurbishment methods across synthetic, human-noise, and large-scale real-noise benchmarks.

2 Related Work

Sample Selection and Label Refurbishment. Noisy-label learning often uses robust losses or sample selection, especially the small-loss effect, to separate clean from corrupted samples (Ghosh et al., 2017; Han et al., 2018; Zheng et al., 2020). Later methods enrich this decision with meta learning, dynamic correction, clean-noisy splitting, neighborhood/eigenvector cues, or confidence tracking (Li et al., 2019; Kim et al., 2023; Li et al., 2023; Kim et al., 2025; Cordeiro & Carneiro, 2025; Pan et al., 2025). A related line performs label correction/refurbishment by mixing observed labels with pseudo targets (Li et al., 2020; Song et al., 2019; Wu et al., 2021; Tu et al., 2023; Lu & He, 2022; Chen et al., 2023). TRACE is complementary: rather than another clean-sample detector or pseudo-target generator, it asks whether one scalar should control trust in both branches.

Structure-Aware Reliability Estimation. Structure-aware methods use neighborhoods, contrastive relations, prototypes, and representation geometry to move beyond scalar losses (Iscen et al., 2022; Yi et al., 2022; Li et al., 2022; Karim et al., 2022; Kim et al., 2021; Wongso et al., 2023; Maini et al., 2023). They show that label noise appears in sample relations and layer-wise representations, not only in losses. TRACE uses shallow structural stability specifically to refine observed-label reliability, while a separate signal assesses the reliability of pseudo supervision from deeper states.

Pseudo-Target Reliability Beyond Sample Cleanliness. Adjacent areas also recognize that pseudo labels or model-provided supervision can be unreliable and may require filtering or rectification (Wang et al., 2024; Chen et al., 2024; Chelouche et al., 2024). These works support the broader intuition that pseudo supervision should not be treated as uniformly trustworthy. TRACE brings this



(i) Gradient sensitivity (ii) Prototype flipping (iii) Linear-probe degradation

Figure 2: Depth-dependent noise effects on CIFAR-10 (20%, 50%, and 80% symmetric noise) across ResNet-18, DenseNet-121, and WRN-28-10. The shared legend is in panel (ii).

distinction into standard refurbishment-based noisy-label classification: it separates observed-label reliability from pseudo-target reliability and ties the separation to depth-dependent noisy bias.

3 Preliminary Analysis

The two-source view requires separate evidence for observed-label and pseudo-target reliability. We use the following analyses for two supporting roles: the population-level shallow–deep contrast diagnoses why a noise-trained pseudo target cannot inherit reliability from the rejection of an observed label, while the sample-level relation-drift analysis tests whether shallow structure can refine observed-label reliability beyond the loss posterior. Clean labels and clean prototypes are used only for these diagnostics and are not required by TRACE during training.

3.1 Deeper Layers Are More Easily Redirected by Noisy Supervision

We first quantify how label corruption perturbs training dynamics across depth. Let g_l^{clean} and g_l^{noisy} denote the average cross-entropy gradients of layer l under clean and noisy supervision on the same batch, and define

$$G^{(l)} = \frac{\left\| g_l^{\text{noisy}} - g_l^{\text{clean}} \right\|_2}{\left\| g_l^{\text{clean}} \right\|_2 + \varepsilon} \tag{1}$$

Larger $G^{(l)}$ indicates stronger noise-induced redirection. On ResNet-18 with CIFAR-10 symmetric noise, $G^{(l)}$ rises from 1.5749 to 2.0510 at layer1 as the noise rate increases from 20% to 80%, but from 1.8334 to 3.8271 at layer4. Fig. 2 shows the same shallow–deep contrast for DenseNet-121 and WRN-28-10, indicating that deeper representations drift more strongly from comparatively stable shallow reference states.

3.2 Mislabeled Samples Flip Toward Noisy Prototypes in Deep Representations

We next examine the geometric consequence through prototype attraction. For a mislabeled sample x, let $h_l(x)$ be its layer-l feature and $\mu_l^{\text{clean}}(c)$ and $\mu_l^{\text{obs}}(c)$ be class-c prototypes under clean and observed labels. We compare

$$\begin{aligned} d_l^{\text{clean}}(x) &= \left\| h_l(x) - \mu_l^{\text{clean}}(y_{\text{clean}}) \right\|_2 \\ d_l^{\text{obs}}(x) &= \left\| h_l(x) - \mu_l^{\text{obs}}(y_{\text{obs}}) \right\|_2 \end{aligned} \tag{2}$$

and summarize the tendency to align with the observed-label prototype by

$$\text{flip_rate}_l = \mathbb{E}\left[\mathbf{1}\left[d_l^{\text{obs}}(x) < d_l^{\text{clean}}(x) \right] \right] \tag{3}$$

On CIFAR-10 with 50% symmetric noise, the flip rate rises from 29.0% at layer1 to 99.7% at layer4; on CIFAR-100, it rises from 28.3% to 98.9%. Fig. 2 shows the same late-stage flipping for DenseNet-121 and WRN-28-10. Thus, noise actively reorganizes deep features around incorrect class anchors, while the much lower shallow flip rate supports using shallow geometry to assess deep representation drift.

3.3 Prototype Flipping Reduces the Direct Usability of Deep Representations

Finally, we test whether prototype flipping reduces the direct usability of learned features. We freeze $h_l(x)$ at each depth, fit a linear classifier W_l, and measure

$$\mathrm{ProbeAcc}_l = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{(x,y)\in\mathcal{D}_{\text{test}}} \mathbf{1}[\arg\max W_l h_l(x) = y] \tag{4}$$

The layer1–4 mean probe accuracy drops from 72.42% to 59.95% on CIFAR-10 and from 42.79% to 28.48% on CIFAR-100 as symmetric noise increases from 20% to 80%. Fig. 2 shows the same monotonic decline for DenseNet-121 and WRN-28-10. Hence, rejecting an observed label does not justify trusting a pseudo target produced from deep representations that may no longer align with clean semantics.

3.4 Relation Drift Provides Information Beyond Loss Confidence

The preceding analyses establish a population-level shallow–deep contrast; TRACE, however, needs a sample-level signal. We therefore compare low- and high-drift samples within the same c_i^{loss} strata on CIFAR-100 with 50% symmetric noise. Fig. 3 shows that low-drift samples more often retain the clean label, especially when loss-based evidence is weak, with the gap narrowing as c_i^{loss} approaches one. Relation drift therefore provides conditional information beyond the loss posterior and is used as a complementary modifier rather than a standalone correctness certificate.

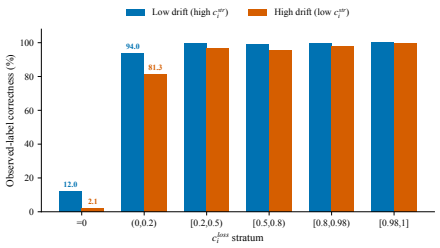


Figure 3: Observed-label correctness by relation drift within loss-confidence strata on CIFAR-100 (50% symmetric noise).

Together, Figs. 1 to 3 motivate two distinct reliability assessments. Observed-label unreliability does not establish pseudo-target reliability, and the depth-dependent diagnostics provide a consistent account of why the pseudo branch needs independent evidence. Relation drift, in contrast, supplies information beyond the loss posterior for refining observed-label reliability. TRACE therefore uses shallow-anchored stability only in the observed-label score and filters pseudo supervision through a separate reliability signal.

4 Method

TRACE reformulates refurbishment as separate reliability assessments for its two supervision sources: the observed label and the model-derived pseudo target. As shown in Fig. 4, the observed-label score combines the loss prior, shallow-to-deep relation stability, and dual-network agreement, while the pseudo-target score separately assesses the confidence of the proposed replacement. Their composition preserves the base pseudo-target generator and makes pseudo supervision depend on both the need for correction and the confidence of the replacement. The complete procedure is given in Algorithm 1.

Setup and Two-Source Reliability. Let $\mathcal{D} = \{(x_i, \hat{\mathbf{y}}_i)\}_{i=1}^N$, where $\hat{\mathbf{y}}_i \in \{0,1\}^C$ is the one-hot observed label, and let $p_i = f_\theta(x_i) \in \Delta^{C-1}$ be the model prediction. A refurbishment learner also constructs a pseudo target $q_i \in \Delta^{C-1}$; in our RoLR-style instantiation,

$$q_i = \mathrm{Sharpen}_T\left(\frac{p_i^{(1)} + p_i^{(2)}}{2}\right) \tag{5}$$

where $p_i^{(1)}$ and $p_i^{(2)}$ are predictions from two networks. Standard methods mix the observed and pseudo-target branches through

$$\mathbf{y}_i^{\text{rec}} = \lambda_i \hat{\mathbf{y}}_i + (1-\lambda_i) q_i, \qquad \lambda_i \in [0,1] \tag{6}$$

Following DivideMix and RoLR (Li et al., 2020; Chen et al., 2023), the observed-label coefficient λ_i is commonly derived from the loss-based clean posterior

$$c_i^{\text{loss}} = P(z_i = \text{clean} \mid \bar{\ell}_i) \tag{7}$$

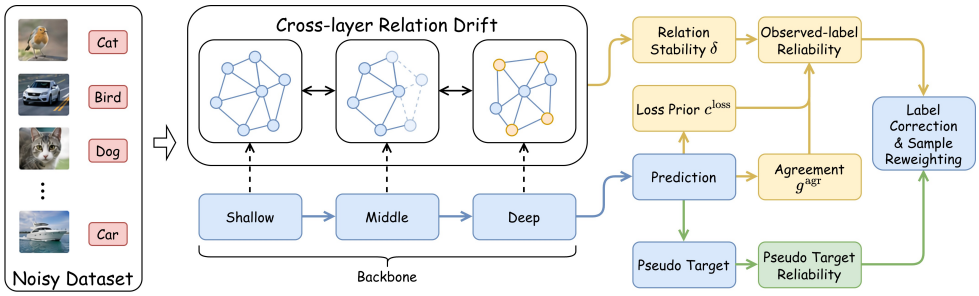


Figure 4: **TRACE overview.** Source-specific reliability scores guide label correction and sample reweighting.

where $z_i \in \{\text{clean}, \text{noisy}\}$ is the latent clean state and $\bar{\ell}_i$ is the normalized cross-entropy loss. Because $1-\lambda_i$ simultaneously weights the pseudo branch, the same signal couples the two reliability decisions and treats reduced observed-label trust as increased pseudo-target trust.

TRACE instead introduces s_i^{obs} for observed-label reliability and s_i^{pseudo} for pseudo-target reliability, and composes them into the branch weights

$$a_i = s_i^{\text{obs}}, \qquad b_i = \big(1 - s_i^{\text{obs}}\big) s_i^{\text{pseudo}}. \tag{8}$$

The observed branch is weighted directly by its reliability score. The pseudo branch is decomposed into two factors: $1 - s_i^{\text{obs}}$ expresses the need for correction, whereas s_i^{pseudo} measures the reliability of the proposed correction. The pseudo-target score thus modulates the correction weight made available by reduced observed-label trust.

Observed-Label Reliability. We first estimate s_i^{obs}, the reliability of the observed label. Fig. 3 shows that relation drift separates samples with different observed-label correctness within the same loss-confidence strata. We therefore combine c_i^{loss}, which captures the small-loss prior, with c_i^{str}, which measures shallow-to-deep relational stability.

For mini-batch features $\{h_l(x_i)\}_{i=1}^B$ at layer l, we compute the cosine relation matrix

$$A_{ij}^{(l)} = \left\langle \frac{h_l(x_i)}{\|h_l(x_i)\|_2}, \frac{h_l(x_j)}{\|h_l(x_j)\|_2} \right\rangle \tag{9}$$

To retain dominant local relations, let $\mathcal{N}_k(i)$ contain sample i and its k largest off-diagonal affinities, and set $M_{ij}^{(l)} = \mathbf{1}[j \in \mathcal{N}_k(i)]$. The symmetric local relation matrix $R^{(l)} = \mathcal{S}_k(A^{(l)})$ is

$$\mathcal{S}_k(A^{(l)}) = \frac{1}{2}\left(A^{(l)} \odot M^{(l)} + \left(A^{(l)} \odot M^{(l)}\right)^\top\right)$$

Relations provide a common reference for comparing representations across depth. Although feature vectors from different layers may have different dimensions and coordinate systems, each relation row describes affinities to the same mini-batch samples. Cross-layer comparison therefore measures how a sample's neighborhood changes as its representation becomes more task-specific. Retaining the strongest local affinities focuses this comparison on nearby samples. The resulting drift score summarizes changes in local relational structure and complements the label-fit information captured by the loss posterior.

The row $R_i^{(l)}$ describes sample i's local position at depth l. Its cross-layer change defines relation drift and structure confidence:

$$\begin{aligned} \delta_i &= \sum_{l=1}^{L-1} \frac{\left\| R_i^{(l+1)} - R_i^{(l)} \right\|_2}{\sqrt{B}} \\ \tilde{\delta}_i &= \text{Norm}(\delta_i) \\ c_i^{\text{str}} &= \text{Norm}\big(\exp(-\gamma \tilde{\delta}_i)\big) \end{aligned} \tag{10}$$

Here L is the number of analyzed layers, B is the batch size, $\gamma > 0$ maps drift to confidence, and $\text{Norm}(\cdot)$ is min-max normalization over the current scoring pass. A larger c_i^{str} indicates greater relational stability. For two networks, we use $\bar{c}_i^{\text{str}} = \frac{1}{2}(c_{i,1}^{\text{str}} + c_{i,2}^{\text{str}})$.

Table 1: Test accuracy (%) on CIFAR-10/100 with synthetic noise. Other methods' results are from the cited papers. Bold/underline: best/second-best. Avg. Δ: mean TRACE gain over its base learner across ten settings.

Dataset	**CIFAR-10**					**CIFAR-100**					**Avg. Δ**
Noise Type	**Sym**			**Pair**	**Ins**	**Sym**			**Pair**	**Ins**	
Method / Noise Rate	**20%**	**50%**	**80%**	**40%**	**40%**	**20%**	**50%**	**80%**	**40%**	**40%**	
JoCoR Wei et al. (2020)	89.4	53.3	25.8	56.1	60.9	55.4	32.7	6.6	34.1	34.9	–
RankMatch Zhang et al. (2023)	96.4	95.4	94.2	94.4	93.8	79.3	77.6	67.2	75.8	76.5	–
CrossSplit Kim et al. (2023)	96.9	96.3	95.4	96.0	95.8	79.9	75.7	64.6	76.8	79.2	–
CCL Fan & Li (2025)	97.0	<u>96.5</u>	94.6	96.1	96.2	79.5	77.4	70.3	77.2	<u>80.0</u>	–
NegScale Fan & Li (2026)	<u>97.2</u>	**96.6**	<u>95.6</u>	<u>96.3</u>	<u>96.5</u>	<u>80.9</u>	<u>78.7</u>	70.8	<u>77.8</u>	**80.4**	–
DivideMix Li et al. (2020)	95.7	94.4	92.9	92.1	95.1	76.9	74.2	59.6	52.3	76.1	–
+ TRACE	96.1	94.9	93.5	92.6	95.5	78.0	75.1	61.0	53.8	77.0	**+0.82**
RoLR Chen et al. (2023)	96.4	95.7	94.2	92.8	93.7	78.6	74.6	66.2	76.1	77.2	–
+ TRACE	97.0	96.2	94.7	93.9	94.1	80.4	76.2	68.0	**78.1**	78.5	**+1.16**
DISC Li et al. (2023)	96.3	95.4	92.9	94.6	96.0	78.6	76.3	59.3	75.1	78.4	–
+ TRACE	96.7	95.8	93.4	95.0	96.3	79.4	77.0	60.4	76.0	79.1	**+0.62**
ANNE Cordeiro & Carneiro (2025)	96.9	96.2	95.3	95.7	96.2	80.4	78.1	<u>73.0</u>	66.4	78.4	–
+ TRACE	**97.4**	**96.6**	**95.9**	**96.4**	**96.6**	**81.2**	**78.8**	**74.2**	67.2	79.3	**+0.70**

As depicted in Fig. 4, prediction agreement provides a lightweight consistency gate:

$$g_i^{\text{agr}} = \begin{cases} 1, & \arg\max p_i^{(1)} = \arg\max p_i^{(2)}, \\ \lambda_{\text{dis}}, & \text{otherwise} \end{cases} \tag{11}$$

where $\lambda_{\text{dis}} \in (0, 1]$ penalizes disagreement. The observed-label score combines this gate with the loss prior and Eq. (10):

$$\begin{aligned} \alpha_t &= 1 - \beta_t(1 - \alpha) \\ s_i^{\text{obs}} &= \text{clip}\big(\big(\alpha_t c_i^{\text{loss}} + (1 - \alpha_t)\bar{c}_i^{\text{str}}\big)\big((1 - \beta_t) + \beta_t g_i^{\text{agr}}\big), 0, 1\big) \end{aligned} \tag{12}$$

where $\alpha \in [0, 1]$ sets the final loss–structure balance and β_t gradually activates structure and agreement. The three components serve different roles: the loss posterior measures compatibility with the observed label, structural confidence measures relational stability across depth, and prediction agreement attenuates the score when the two networks disagree. When $\beta_t = 0$, the score reduces to the loss posterior; as β_t increases, structure and agreement increasingly influence the assessment. This gradual fusion is useful because early representations are still evolving. It also allows a sample with substantial relation drift to retain observed-label support when its loss-based evidence remains strong, accommodating the natural relational changes that can occur for hard-but-clean examples (Appendix B).

Pseudo-Target Reliability. We next estimate s_i^{pseudo}, the reliability of q_i as an alternative supervision signal. This assessment requires its own evidence because low s_i^{obs} indicates a need for correction but does not establish that the available pseudo target is reliable (Fig. 1). We score whether q_i is sufficiently certain to provide alternative supervision by

$$s_i^{\text{pseudo}} = \Big(\max_c q_{i,c}\Big)^{\rho} \tag{13}$$

where $\rho \geq 1$ controls the suppression of low-confidence targets. The gate is activated after a short warm-up. For a fixed observed-label score, decreasing pseudo-target confidence continuously reduces b_i, while increasing ρ strengthens this attenuation for a fixed confidence value. When $s_i^{\text{pseudo}} = 1$, the branch receives the full complementary weight $1 - s_i^{\text{obs}}$; lower scores reduce that allocation. Thus, the gate adjusts the influence of the existing pseudo target without changing its construction. The two reliability scores address different supervision decisions while sharing the underlying models.

Label Correction and Sample Reweighting. Using the source-specific weights in Eq. (8), the rightmost block in Fig. 4 constructs the corrected target

$$\tilde{\mathbf{y}}_i = \frac{a_i \hat{\mathbf{y}}_i + b_i q_i}{a_i + b_i + \varepsilon} \tag{14}$$

Table 2: Test accuracy (%) on CIFAR-N, from Wei et al. (2023) and our replication. Bold/underline: best/second-best.

Dataset		CIFAR-10N					CIFAR-100N	Avg. Δ
Method	Noise Type Noise Rate	Aggre 9.0%	Rand1 17.2%	Rand2 18.12%	Rand3 17.64%	Worst 40.2%	Fine 40.2%	
Cross-Entropy		88.8	83.8	83.5	83.9	67.7	47.8	
JoCoR Wei et al. (2020)		90.6	88.8	88.5	88.1	66.7	40.1	–
Co-learning Tan et al. (2021)		92.4	91.3	91.2	91.4	81.0	47.9	
RankMatch Zhang et al. (2023)		95.6	94.8	95.1	95.3	92.8	65.2	–
CCL Fan & Li (2025)		96.4	<u>96.0</u>	95.8	96.1	<u>93.1</u>	65.5	–
NegScale Fan & Li (2026)		<u>96.6</u>	**96.2**	<u>96.0</u>	<u>96.4</u>	**93.5**	<u>66.3</u>	–
DivideMix Li et al. (2020)		93.2	92.8	92.6	93.1	89.2	55.2	–
+ TRACE		93.6	93.2	93.1	93.5	89.8	56.0	**+0.52**
RoLR Chen et al. (2023)		95.4	94.9	94.7	95.2	92.3	62.3	–
+ TRACE		95.8	95.3	95.2	95.6	92.8	63.1	**+0.50**
ANNE Cordeiro & Carneiro (2025)		96.2	95.7	95.5	95.9	93.0	66.0	–
+ TRACE		**96.8**	**96.2**	**96.1**	**96.6**	**93.5**	**66.6**	**+0.58**

where $\varepsilon > 0$ ensures numerical stability. The supervision strength and training loss are

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} w_i \, \ell(p_i, \tilde{\mathbf{y}}_i) \, , w_i = \max(a_i + b_i, w_{\min}) \tag{15}$$

where $w_{\min} \geq 0$ is a lower bound and $\ell(\cdot, \cdot)$ is cross-entropy. In each mini-batch, w_i is normalized by its mean and the base prior regularizer is retained.

The corrected target specifies the relative contributions of the two supervision sources, while $a_i + b_i$ controls their total strength. Retaining this total as a sample weight prevents target normalization from restoring full supervision strength when both sources are unreliable.

Training Procedure. TRACE reuses RoLR's GMM loss posterior and pseudo-target generator, deriving additional reliability signals from existing features and predictions. After warm-up, structure is gradually introduced and pseudo-target gating is enabled. Corrected targets and sample weights are applied symmetrically to both networks, retaining the original prior regularizer.

5 Experiments

Experimental setup. Following Chen et al. (2023); Fan & Li (2026),We evaluate TRACE on synthetic noise with CIFAR datasets and real-world CIFAR-10N/CIFAR-100N, WebVision, Food-101N, and Clothing1M (Krizhevsky et al., 2009; Wei et al., 2022; Li et al., 2017; Lee et al., 2018). CIFAR experiments use ResNet-18, and the large-scale real-noise experiments use ResNet-50. For each run, test accuracy is averaged over the final 10 epochs, and the reported results are further averaged across three random seeds (0, 42, and 1027). TRACE preserves each base learner's pseudo-target generator, training schedule, and data pipeline, so each paired comparison changes only the reliability mechanism. Appendix A provides the full protocols and hyperparameters.

Table 3: Large-scale real-noise test accuracy (%), averaged over the last 10 epochs.

Method	WebVision		ILSVRC12		Food-101N	Avg. Δ
	Top1	Top5	Top1	Top5	Acc.	
CORES	71.7	89.0	68.4	88.3	84.4	–
+ CT	72.1	90.2	68.5	90.0	84.4	+0.68
+ TRACE	72.3	90.2	68.6	90.1	84.7	**+0.82**
DivideMix	77.4	91.9	74.7	92.1	86.5	–
+ CT	78.1	92.3	75.2	92.2	86.8	+0.40
+ TRACE	78.2	92.3	75.3	92.3	86.8	**+0.46**
f-DivideMix	78.4	92.5	75.2	92.2	86.8	–
+ CT	78.8	92.9	75.7	93.2	87.0	+0.50
+ TRACE	78.9	**93.2**	75.8	**93.3**	87.1	**+0.64**
DISC	80.1	92.4	77.4	92.4	87.3	–
+ CT	80.1	92.6	**78.3**	92.4	87.5	+0.26
+ TRACE	**80.4**	92.6	**78.3**	92.5	**87.8**	**+0.48**

CIFAR synthetic noise. Tab. 1 covers symmetric, pair/asymmetric, and instance-dependent noise on CIFAR-10/100. TRACE improves every paired base learner in all reported settings, with generally larger gains on CIFAR-100. This pattern is consistent with source-specific reliability assessment becoming more useful when the larger label space makes pseudo su-

pervision harder. ANNE+TRACE gives the strongest overall results, while RoLR+TRACE improves every RoLR result.

Real-world noisy datasets. Tabs. 2 and 3 show that TRACE improves all paired results on CIFAR-N and improves or matches every reported WebVision and Food-101N metric. The gains span four base learners with different pseudo-target generators, rather than depending on one refurbishment rule. On Clothing1M (Tab. 4), DivideMix+TRACE reaches 77.6%, compared with 74.8% for DivideMix.

Comparison with plug-in methods. Tab. 3 compares TRACE with CT under matched base-learner protocols, and Tab. 4 compares DivideMix-based plug-ins. Because the latter methods use different evaluation protocols, we report their displayed values as contextual comparisons rather than treating them as a strictly controlled leaderboard. The unmodified DivideMix row is the common reference, whereas the CT comparison keeps the base learner and evaluation protocol matched. TRACE exceeds CT's average gain for every matched base learner, improves DivideMix in all CIFAR columns, and gives the highest reported Clothing1M accuracy.

Table 4: Plug-in comparison: test accuracy (%).

Plug-in	CIFAR-10-Sym			CIFAR-100-Sym			Clothing1M
	20%	50%	80%	20%	50%	80%	
DivideMix	95.7	94.4	92.9	76.9	74.2	59.6	74.8
+ SNSCL	**96.1**	**95.2**	91.7	77.3	74.7	**64.3**	75.3
+ L2B	95.9	95.1	93.2	77.3	74.9	60.2	76.1
+ TRACE	**96.1**	94.9	**93.5**	**78.0**	**75.1**	61.0	**77.6**

Case study. Fig. 5 shows the same behavior at sample level. In the top row, a noisy label gives small a_i, while a correct, confident pseudo target gives large b_i. In the bottom row, TRACE keeps b_i small because the replacement is wrong or uncertain. Thus, the need for correction does not override pseudo-target reliability. The cases operationalize the two decisions: a_i asks whether to retain the observed label, while b_i asks whether the proposed replacement is usable.

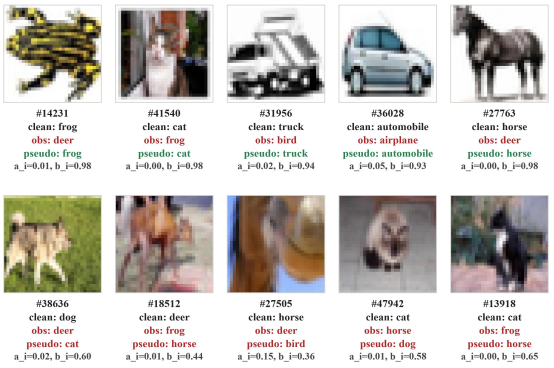


Figure 5: CIFAR-10 examples (50% sym. noise). Top: pseudo-target correction; bottom: unreliable targets down-weighted. a_i, b_i: observed-label and pseudo-target weights (Eq. (8)).

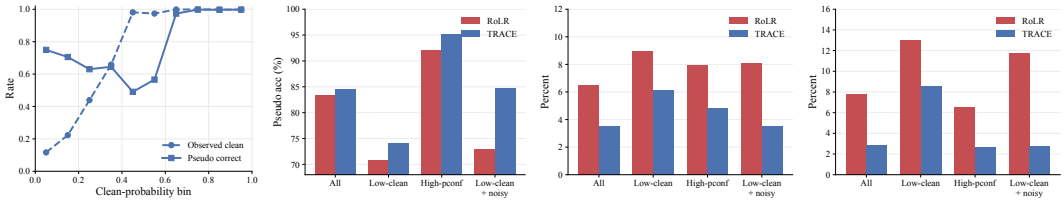
Ablation on two-source reliability. Tab. 5 examines structural confidence, prediction agreement, pseudo-target gating, and sample reweighting on CIFAR-100 with 50% symmetric noise. The coupled-null variant uses refined observed-label reliability but retains the complementary pseudo weight, improving best/last accuracy from RoLR's 75.68%/74.71% to 76.58%/75.86%. Adding the pseudo-target gate changes b_i from $1 - s_i^{\text{obs}}$ to $(1 - s_i^{\text{obs}})s_i^{\text{pseudo}}$, raising accuracy to 76.89%/76.10% under the same unweighted soft-target cross-entropy. Removing structural confidence or prediction agreement reduces performance relative to this decoupled variant, supporting both observed-label components. Finally, adding sample reweighting yields the full TRACE result of 76.95%/76.22%. These comparisons support complementary contributions from observed-label refinement and pseudo-target gating, with a smaller additional gain from sample reweighting.

Table 5: Component ablation on CIFAR-100 (50% symmetric noise): test accuracy (%).

Variant	Best	Last
RoLR (base)	75.68	74.71
Decoupled w/o c_i^{str}	76.51	75.78
Decoupled w/o g_i^{agr}	76.43	75.70
Coupled null ($b_i = 1 - s_i^{\text{obs}}$)	76.58	75.86
Decoupled reliability ($b_i = (1 - s_i^{\text{obs}})s_i^{\text{pseudo}}$)	76.89	76.10
TRACE	**76.95**	**76.22**

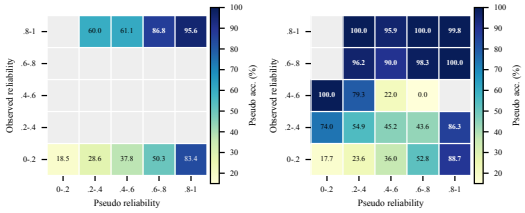
Table 6: Pseudo-target accuracy (%) on low-clean+noisy samples.

Setting	RoLR	TRACE	Gain
CIFAR100-Sym20	76.51	93.02	+16.51
CIFAR100-Sym50	72.92	84.70	+11.78
CIFAR100-Sym80	42.41	54.09	+11.68
CIFAR-100N	59.68	60.38	+0.70
CIFAR10-Sym50	96.42	97.52	+1.10
CIFAR-10N-Agg	93.02	94.07	+1.05
CIFAR-10N-Worst	93.83	95.78	+1.95



(a) Reliability by bin (b) Pseudo-target acc. (c) High-conf. errors (d) Calibration error

Figure 6: Pseudo-target reliability on CIFAR-100 with 50% symmetric noise.



(a) RoLR (b) TRACE

Figure 7: Pseudo-target accuracy (%) on CIFAR-100 (50% symmetric noise). Blank cells: fewer than 0.01% of samples.

Table 7: Sensitivity to α, ρ on CIFAR-100 (50% symmetric noise). HC Wrong: incorrect targets among high-confidence pseudo targets (%; lower is better).

Factor	Value	Best	Last	Pseudo Acc.	HC Wrong
α	0.5	76.72	75.98	84.31	6.42
	0.7	**76.95**	**76.22**	**84.70**	**6.16**
	0.9	76.83	76.07	84.48	6.28
ρ	0.5	76.78	76.01	84.18	6.55
	1.0	**76.95**	**76.22**	**84.70**	6.16
	1.5	76.87	76.14	84.56	**6.05**

Pseudo-target reliability. Tab. 6 evaluates low-clean+noisy samples, whose observed labels are corrupted and assigned low reliability. Subsets are identified separately by each method and may differ in membership because RoLR and TRACE use different observed-label scores. TRACE achieves higher pseudo-target accuracy in all reported settings, with an average gain of 13.32 percentage points across the three CIFAR-100 symmetric-noise rates. Tab. 17 provides the full results.

Error quality on CIFAR-100. Under 50% symmetric noise, TRACE raises pseudo-target accuracy from 70.75% to 74.15% on low-clean samples and from 72.92% to 84.70% on low-clean+noisy samples (Figs. 6a and 6b). It also reduces high-confidence errors from 8.93% to 6.16% and ECE from 13.03% to 8.52% (Figs. 6c and 6d), indicating improvements in both correctness and calibration. These diagnostics characterize supervision quality during training, complementing test accuracy.

Reliability landscape. Fig. 7 jointly examines observed-label and pseudo-target reliability; the observed-label axis uses loss-based cleanliness for RoLR and refined reliability for TRACE. Pseudo-target correctness varies across pseudo-reliability bins even when observed-label reliability is low, supporting modulation of $1 - s_i^{\text{obs}}$ by s_i^{pseudo}. In their respective low-observed/high-pseudo regions, RoLR and TRACE attain accuracies of 83.4% and 88.7%, with high-confidence error rates of 12.3% and 8.3%.

Sensitivity to hyperparameters. Tab. 7 varies the loss–structure balance α and pseudo-confidence power ρ individually on CIFAR-100 with 50% symmetric noise. Among the tested values, $\alpha = 0.7$ performs best across all metrics. We use $\rho = 1.0$ for its best classification and pseudo-target accuracy; $\rho = 1.5$ slightly reduces high-confidence errors at a small accuracy cost. Tab. 16 reports additional sensitivity results.

6 Conclusion

This work identifies the risk of replacing noise with noise in noisy-label refurbishment: low trust in an observed label does not establish that its pseudo target is reliable. TRACE addresses this mismatch through separate reliability assessment for the observed label and pseudo target. It combines loss, shallow-anchored relation stability, and prediction agreement for the former, while an independent confidence signal assesses the latter; the resulting source-specific weights control label correction and sample reweighting. Experiments across synthetic, human-annotated, and real-world noise show that TRACE improves refurbishment learners and pseudo-target reliability. Findings support source-specific reliability assessment as a safer interface for noisy-label learning.

AI Assistance Disclosure

Generative AI tools assisted with drafting and editing text, formulating mathematical claims, identifying relevant literature, code debugging, and figure drafting. As described in Appendix D.5, GPT-generated annotations serve as surrogate reference labels for the Clothing1M diagnostic. These labels are used only for this auxiliary analysis, not for model training or official benchmark evaluation, and are not treated as human-verified ground truth. The authors reviewed and revised AI-assisted text and checked suggested references against the original sources. The authors take full responsibility for the final content of this paper, including AI-assisted claims and artifacts.

References

Keryan Chelouche, Marie Lachaize, Marine Bernard, Louise Olgiati, and Rémi Cuingnet. When the small-loss trick is not enough: Multi-label image classification with noisy labels applied to CCTV sewer inspections. *CoRR*, abs/2410.07689, 2024. doi: 10.48550/ARXIV.2410.07689. URL https://doi.org/10.48550/arXiv.2410.07689.

Hao Chen, Jindong Wang, Zihan Wang, Ran Tao, Hongxin Wei, Xing Xie, Masashi Sugiyama, and Bhiksha Raj. Learning with noisy foundation models. *CoRR*, abs/2403.06869, 2024. doi: 10.48550/ARXIV.2403.06869. URL https://doi.org/10.48550/arXiv.2403.06869.

Mingcai Chen, Hao Cheng, Yuntao Du, Ming Xu, Wenyu Jiang, and Chongjun Wang. Two wrongs don't make a right: Combating confirmation bias in learning with label noise. In Brian Williams, Yiling Chen, and Jennifer Neville (eds.), *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pp. 14765–14773. AAAI Press, 2023. doi: 10.1609/AAAI.V37I12.26725. URL https://doi.org/10.1609/aaai.v37i12.26725.

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. In *AAAI Conference on Artificial Intelligence*, 2020. URL https://api.semanticscholar.org/CorpusID:228083495.

Filipe R. Cordeiro and Gustavo Carneiro. ANNE: adaptive nearest neighbours and eigenvector-based sample selection for robust learning with noisy labels. *Pattern Recognit.*, 159:111132, 2025. doi: 10.1016/J.PATCOG.2024.111132. URL https://doi.org/10.1016/j.patcog.2024.111132.

Wenxiao Fan and Kan Li. Combating semantic contamination in learning with label noise. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 2870–2878. AAAI Press, 2025. doi: 10.1609/AAAI.V39I3.32293. URL https://doi.org/10.1609/aaai.v39i3.32293.

Wenxiao Fan and Kan Li. Leveraging dissimilarity invariance as a robust anchor for learning with noisy labels. In Sven Koenig, Chad Jenkins, and Matthew E. Taylor (eds.), *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pp. 3804–3812. AAAI Press, 2026. doi: 10.1609/AAAI.V40I5.37381. URL https://doi.org/10.1609/aaai.v40i5.37381.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 1919–1925. AAAI Press, 2017. URL http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14759.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS):8527–8537, 2018. ISSN 10495258.

Ahmet Iscen, Jack Valmadre, Anurag Arnab, and Cordelia Schmid. Learning with neighbor consistency for noisy labels. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4662–4671, 2022. URL https://api.semanticscholar.org/CorpusID:246607828.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. *35th International Conference on Machine Learning, ICML 2018*, 5(6):3601–3620, 2018.

Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. UNICON: combating label noise through uniform selection and contrastive learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 9666–9676. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00945. URL https://doi.org/10.1109/CVPR52688.2022.00945.

Daehwan Kim, Kwangrok Ryoo, Hansang Cho, and Seungryong Kim. Splitnet: Learnable clean-noisy label splitting for learning with noisy labels. *Int. J. Comput. Vis.*, 133(2):549–566, 2025. doi: 10.1007/S11263-024-02187-4. URL https://doi.org/10.1007/s11263-024-02187-4.

Jihye Kim, Aristide Baratin, Yan Zhang, and Simon Lacoste-Julien. Crosssplit: Mitigating label noise memorization through data splitting. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16377–16392. PMLR, 2023. URL https://proceedings.mlr.press/v202/kim23a.html.

Taehyeon Kim, Jongwoo Ko, Sangwook Cho, Jinhwan Choi, and Se-Young Yun. FINE samples for learning with noisy labels. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 24137–24149, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/ca91c5464e73d3066825362c3093a45f-Abstract.html.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Kuang-Huei Lee, Xiaodong He, Lei Zhang, and Linjun Yang. Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5447–5456, 2018.

Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Learning to learn from noisy labeled data. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 5051–5059. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00519. URL http://openaccess.thecvf.com/content_CVPR_2019/html/Li_Learning_to_Learn_From_Noisy_Labeled_Data_CVPR_2019_paper.html.

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.

Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 316–325. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00041. URL https://doi.org/10.1109/CVPR52688.2022.00041.

Wen Li, Limin Wang, Wei Li, Eirikur Agustsson, and Luc Van Gool. Webvision database: Visual learning and understanding from web data. *CoRR*, abs/1708.02862, 2017. URL http://arxiv.org/abs/1708.02862.

Yifan Li, Hu Han, Shiguang Shan, and Xilin Chen. DISC: learning from noisy labels via dynamic instance-specific selection and correction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 24070–24079. IEEE, 2023. doi: 10.1109/CVPR52729.2023.02305. URL https://doi.org/10.1109/CVPR52729.2023.02305.

Yangdi Lu and Wenbo He. SELC: self-ensemble label correction improves learning with noisy labels. In Luc De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pp. 3278–3284. ijcai.org, 2022. doi: 10.24963/IJCAI.2022/455. URL https://doi.org/10.24963/ijcai.2022/455.

Pratyush Maini, Michael Curtis Mozer, Hanie Sedghi, Zachary Chase Lipton, J. Zico Kolter, and Chiyuan Zhang. Can neural network memorization be localized? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, Proceedings of Machine Learning Research, pp. 23536–23557. PMLR, 2023. URL https://proceedings.mlr.press/v202/maini23a.html.

Eran Malach and Shai Shalev-Shwartz. Decoupling "when to update" from "how to update". *Advances in Neural Information Processing Systems*, 2017-Decem:961–971, jun 2017. ISSN 10495258. URL http://arxiv.org/abs/1706.02613.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In Christopher J. C. Burges, Léon Bottou, Zoubin Ghahramani, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 1196–1204, 2013. URL https://proceedings.neurips.cc/paper/2013/hash/3871bd64012152bfb53fdf04b401193f-Abstract.html.

Weiran Pan, Wei Wei, Feida Zhu, and Yong Deng. Enhanced sample selection with confidence tracking: Identifying correctly labeled yet hard-to-learn samples in noisy data. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 19795–19803. AAAI Press, 2025. doi: 10.1609/AAAI.V39I19.34180. URL https://doi.org/10.1609/aaai.v39i19.34180.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *CoRR*, abs/2008.08186, 2020. URL https://arxiv.org/abs/2008.08186.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 2233–2241. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.240. URL https://doi.org/10.1109/CVPR.2017.240.

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. SELFIE: refurbishing unclean samples for robust deep learning. volume 97 of *Proceedings of Machine Learning Research*, pp. 5907–5915. PMLR, 2019. URL http://proceedings.mlr.press/v97/song19b.html.

Cheng Tan, Jun Xia, Lirong Wu, and Stan Z. Li. Co-learning: Learning from noisy labels with self-supervision. In Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo Cesar, Florian Metze, and Balakrishnan Prabhakaran (eds.), *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, pp. 1405–1413. ACM, 2021. doi: 10.1145/3474085.3475622. URL https://doi.org/10.1145/3474085.3475622.

Yuanpeng Tu, Boshen Zhang, Yuxi Li, Liang Liu, Jian Li, Yabiao Wang, Chengjie Wang, and Cairong Zhao. Learning from noisy labels with decoupled meta label purifier. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 19934–19943. IEEE, 2023. doi: 10.1109/CVPR52729.2023.01909. URL https://doi.org/10.1109/CVPR52729.2023.01909.

Haochen Wang, Yuchao Wang, Yujun Shen, Junsong Fan, Yuxi Wang, and Zhaoxiang Zhang. Using unreliable pseudo-labels for label-efficient semantic segmentation. *Int. J. Comput. Vis.*, October 2024.

Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu Tao Xia. Iterative Learning with Open-set Noisy Labels. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 8688–8696, 2018. ISSN 10636919. doi: 10.1109/CVPR.2018.00906.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 13723–13732. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.01374. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Wei_Combating_Noisy_Labels_by_Agreement_A_Joint_Training_Method_with_CVPR_2020_paper.html.

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 *ICML 2023*, 2023. URL https://api.semanticscholar.org/CorpusID:259138834.

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 *ICLR 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=TBWA6PLJZQm.

Shelvia Wongso, Rohan Ghosh, and Mehul Motani. Using sliced mutual information to study memorization and generalization in deep neural networks. In Francisco J. R. Ruiz, Jennifer G. Dy, and Jan-Willem van de Meent (eds.), *International Conference on Artificial Intelligence and Statistics, 25-27 April 2023, Palau de Congressos, Valencia, Spain*, Proceedings of Machine Learning Research, pp. 11608–11629. PMLR, 2023. URL https://proceedings.mlr.press/v206/wongso23a.html.

Yichen Wu, Jun Shu, Qi Xie, Qian Zhao, and Deyu Meng. Learning to purify noisy labels via meta soft label corrector. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021*, pp. 10388–10396. AAAI Press, 2021. URL https://ojs.aaai.org/index.php/AAAI/article/view/17244.

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 2691–2699. IEEE Computer Society, 2015. doi: 10.1109/CVPR.2015.7298885. URL https://doi.org/10.1109/CVPR.2015.7298885.

Linya Yi, Sheng Liu, Qi She, Alex McLeod, and Boyu Wang. On learning contrastive representations for learning with noisy labels. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16661–16670, 2022. URL https://api.semanticscholar.org/CorpusID:247223119.

Ziyi Zhang, Weikai Chen, Chaowei Fang, Zhen Li, Lechao Chen, Liang Lin, and Guanbin Li. Rankmatch: Fostering confidence and consistency in learning with noisy labels. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 1644–1654. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00158. URL https://doi.org/10.1109/ICCV51070.2023.00158.

Songzhu Zheng, Pengxiang Wu, Aman Goswami, Mayank Goswami, Dimitris N. Metaxas, and Chao Chen. Error-bounded correction of noisy labels. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11447–11457. PMLR, 2020. URL http://proceedings.mlr.press/v119/zheng20c.html.

Appendix Overview.

- Appendix A reports default hyperparameters, compute resources, and asset information.
- Appendix B develops the two-source reliability and shallow-anchor analyses, clarifies the treatment of hard-but-clean samples, and provides extended representation diagnostics.
- Appendix C specifies the experimental protocols, ablation variants, and synthetic-noise construction.
- Appendix D reports additional noise settings, sensitivity analyses, reliability statistics, and real-noise diagnostics.
- Appendix E explains the construction and interpretation of the diagnostic figures.
- Appendix F presents TRACE pseudocode, plug-in implementation details, and computational overhead.
- Appendix G discusses scope limitations and the extension to open-set label noise.

A Default Hyperparameters and Resources

Tab. 8 reports the default hyperparameters used for CIFAR-style TRACE experiments with the RoLR instantiation. The dataset, noise type, noise rate, and base learner are changed according to the experimental setting; all other values are kept fixed unless an ablation explicitly states otherwise. For large-scale real-noise experiments, we follow the training protocol of the corresponding base learner and use the same reliability defaults when TRACE is attached.

Table 8: RoLR+TRACE defaults for CIFAR-style experiments.

Category	**Hyperparameter**	**Default value**
Architecture	Backbone	ResNet-18
Training	Batch size	64
Training	Epochs / warm-up	500 / 15
Training	Optimizer	SGD, momentum 0.9, weight decay 5×10^{-4}
Training	Initial learning rate	0.02, divided by 10 in the last 100 epochs
Training	Supervised / prior penalty weights	$\lambda_l = 1,\ \lambda_p = 1$
Pseudo target	Sharpening temperature	$T = 1$
Loss posterior	GMM components	2
Loss posterior	GMM max iterations / tolerance / regularizer	$10/10^{-2}/5 \times 10^{-4}$
Observed label	Confidence mode	loss+structure
Observed label	Loss-structure fusion weight	$\alpha = 0.7$
Structure confidence	Transport drift temperature	$\gamma = 5.0$
Structure confidence	Sparse relation neighbors	$k = 50$
Structure confidence	Start epoch / ramp-up length	$30/20$
Agreement gate	Disagreement multiplier	$\lambda_{\text{dis}} = 0.5$
Pseudo-target reliability	Start epoch / confidence power	$30/1.0$
Supervision strength	Minimum sample weight	$w_{\min} = 0.2$

Compute resources. All experiments were run on a local workstation with 8 NVIDIA GeForce RTX 4090 GPUs, each with 24GB memory.

Existing assets. We use public benchmark datasets and previously published noisy-label baselines, and cite their original sources in the main paper. We follow the standard access and usage terms provided by the corresponding dataset and code maintainers.

Table 9: Additional backbone diagnostics. $\Delta G = G^{(4)} - G^{(1)}$; Probe: mean linear-probe accuracy over layers 1–4.

Backbone	ΔG		**Probe**	
	CIFAR10 Sym50	**CIFAR-10N Agg**	**CIFAR10 Sym50**	**CIFAR-10N Agg**
ResNet-34	0.523	0.042	68.29	74.53
ResNet-50	0.945	-0.034	68.77	74.93
MobileNetV2	1.066	0.269	64.55	70.86

Table 10: Additional CIFAR-10 backbone diagnostics. ΔP_9: layer4–layer1 noisy-gradient projection-ratio difference on mislabeled samples; Flip4/Gap4: layer4 prototype-flip rate/attraction gap; ΔRepair: Suffix34h minus Prefix12 recovery (positive favors late-stage repair).

Backbone	**Setting**	ΔP_9	**Flip4**	**Gap4**	**ΔRepair**
ResNet-34	CIFAR10-Sym50	0.049	0.998	3.841	0.41
ResNet-34	CIFAR-10N-Agg	0.005	0.999	5.559	0.99
ResNet-34	CIFAR-10N-Worst	0.058	0.998	3.931	-3.00
ResNet-50	CIFAR10-Sym50	0.001	0.997	3.363	0.71
ResNet-50	CIFAR-10N-Agg	0.006	1.000	5.009	2.30
ResNet-50	CIFAR-10N-Worst	0.026	0.997	3.550	-0.01
WRN-28-10	CIFAR10-Sym50	0.045	1.000	3.917	1.32
WRN-28-10	CIFAR-10N-Agg	0.037	1.000	5.124	3.13
WRN-28-10	CIFAR-10N-Worst	0.031	1.000	3.879	4.46
MobileNetV2	CIFAR10-Sym50	0.080	0.209	-0.738	3.56
MobileNetV2	CIFAR-10N-Agg	0.063	0.584	0.360	1.49
MobileNetV2	CIFAR-10N-Worst	0.077	0.396	-0.297	-0.24
DenseNet-121	CIFAR10-Sym50	-0.013	1.000	2.846	-0.54
DenseNet-121	CIFAR-10N-Agg	-0.110	1.000	4.052	10.99

B Reliability Analysis and Motivation

This section provides the supplementary analysis behind TRACE's decoupled reliability design, including the theoretical view, extended preliminary evidence, and representation-level diagnostics.

B.1 Reliability View of TRACE

This subsection gives a simple theoretical view of why TRACE separates observed-label reliability from pseudo-target reliability and why shallow relations are useful as reliability anchors. The goal is not to prove a generalization guarantee for deep networks under arbitrary noise, but to make explicit the assumptions hidden by a single cleanliness coefficient.

From label refurbishment to model structure. Label refurbishment first exposes a branch-coupling problem: the same sample-wise interpolation coefficient reduces observed-label weight and increases pseudo-target weight. This is reasonable only if low observed-label trust is evidence for high pseudo-target trust. However, the pseudo target is not an external supervision source; it is produced from predictions, teachers, or temporal ensembles built on the model's own representations. When noisy labels have already redirected these representations, observed-label unreliability and pseudo-target unreliability can coincide. This is why the main analysis moves from the refurbishment rule to model structure: it asks how noisy supervision changes representations across depth, and whether any part of the network remains stable enough to diagnose reliability. The observed shallow–deep contrast then motivates TRACE's two decisions: estimate observed-label reliability with a shallow structural anchor, and filter pseudo-target reliability with a separate confidence signal.

Dual reliability instead of one interpolation score. Consider one training sample x_i with clean label y_i, observed noisy label $\hat{y}_i$, and pseudo-target class $y_i^{\text{pseudo}} = \arg\max_c q_{i,c}$. Let

$$r_i^{\text{obs}} = P(\hat{y}_i = y_i \mid x_i), \qquad r_i^{\text{pseudo}} = P(y_i^{\text{pseudo}} = y_i \mid x_i)$$

denote the conditional reliability of the two supervision sources. A refurbishment target combines the observed-label branch and pseudo-target branch with nonnegative effective weights a_i and b_i. Ignoring class-dependent loss curvature and focusing on whether each branch points to the clean class, the expected amount of correct supervision is

$$S_i(a_i, b_i) = a_i r_i^{\text{obs}} + b_i r_i^{\text{pseudo}},$$

whereas the expected amount of incorrect supervision is

$$N_i(a_i, b_i) = a_i(1 - r_i^{\text{obs}}) + b_i(1 - r_i^{\text{pseudo}}).$$

The clean signal-to-noise ratio of the mixed target can therefore be summarized as

$$\text{SNR}_i(a_i, b_i) = \frac{a_i r_i^{\text{obs}} + b_i r_i^{\text{pseudo}}}{a_i(1 - r_i^{\text{obs}}) + b_i(1 - r_i^{\text{pseudo}}) + \varepsilon}. \tag{16}$$

For fixed observed-label weight a_i, increasing the pseudo-target weight in the idealized ratio is beneficial only when the pseudo target is more reliable than the observed-label branch. The stabilizer ε is only a numerical device; omitting it for the analytic comparison gives

$$\frac{\partial \text{SNR}_i}{\partial b_i} = \frac{a_i\big(r_i^{\text{pseudo}} - r_i^{\text{obs}}\big)}{\big[a_i(1 - r_i^{\text{obs}}) + b_i(1 - r_i^{\text{pseudo}})\big]^2}. \tag{17}$$

With the stabilizer retained, the numerator additionally contains the small numerical term $\varepsilon r_i^{\text{pseudo}}$, which should not be interpreted as statistical evidence for the pseudo branch. Thus, when $a_i > 0$, lowering trust in the observed label does not by itself justify increasing pseudo supervision; the pseudo branch improves the idealized mixture only if its own reliability is sufficiently high relative to the observed-label branch. This observation formalizes the failure mode discussed in the main text: a single cleanliness coefficient estimates r_i^{obs}, but then also controls b_i, implicitly treating $1 - r_i^{\text{obs}}$ as evidence for pseudo-target reliability.

The common interpolation rule $a_i = \lambda_i$, $b_i = 1 - \lambda_i$, with λ_i driven by a loss-based clean posterior, has exactly this coupling. If λ_i is small because the observed label looks unreliable, the pseudo branch becomes large even when r_i^{pseudo} is low. In contrast, TRACE uses

$$a_i = s_i^{\text{obs}}, \qquad b_i = (1 - s_i^{\text{obs}}) s_i^{\text{pseudo}},$$

so the pseudo branch opens only when two conditions hold: observed-label reliability is low enough to need an alternative, and pseudo-target reliability is high enough to provide one. The total supervision strength

$$w_i^{(0)} = a_i + b_i = s_i^{\text{obs}} + (1 - s_i^{\text{obs}}) s_i^{\text{pseudo}}$$

also decreases when both branches are unreliable. The implemented loss uses $w_i = \max(w_i^{(0)}, w_{\min})$, as in Eq. (15), only to impose a small optimization lower bound. This is important because the right response to an unreliable observed label and an unreliable pseudo target is not to force a hard replacement target, but to reduce the sample's effective contribution. Under the view in Eq. (16), TRACE is therefore a branch-wise reliability filter rather than merely a different interpolation schedule.

This analysis is intentionally local and model-agnostic. It does not require the observed label and pseudo target to be independent: if both branches share a bias, then r_i^{pseudo} can be low exactly when r_i^{obs} is low, making the single-score coupling even less justified. It also matches the noisy-label literature's broader lesson that loss correction or sample selection is helpful only when the selected supervision remains aligned with the clean target (Natarajan et al., 2013; Patrini et al., 2017; Han et al., 2018). TRACE keeps the useful small-loss prior for the observed-label branch, but prevents that prior from automatically becoming a pseudo-target trust score.

B.2 Why Shallow Relations Can Serve as Anchors

TRACE's observed-label reliability further asks whether a sample preserves its local relation pattern as representations become deeper. This design is motivated by the preliminary evidence that noisy supervision perturbs deeper layers more strongly while shallower relations remain comparatively stable. Here we give a small perturbation argument that connects this empirical pattern to the structural confidence in Eq. (10).

Let $u_i^{(l)}$ be the unit-normalized clean feature of sample i at layer l, and let the noisy-trained feature be

$$\tilde{u}_i^{(l)} = \frac{u_i^{(l)} + \eta_i^{(l)}}{\|u_i^{(l)} + \eta_i^{(l)}\|_2}, \qquad \|\eta_i^{(l)}\|_2 \le \epsilon_l.$$

For two samples i, j, the cosine relation is $A_{ij}^{(l)} = \langle u_i^{(l)}, u_j^{(l)} \rangle$. Since the inner product of unit vectors is Lipschitz in each argument,

$$\left| \left\langle \tilde{u}_i^{(l)}, \tilde{u}_j^{(l)} \right\rangle - \left\langle u_i^{(l)}, u_j^{(l)} \right\rangle \right| \le \left\| \tilde{u}_i^{(l)} - u_i^{(l)} \right\|_2 + \left\| \tilde{u}_j^{(l)} - u_j^{(l)} \right\|_2 . \tag{18}$$

When $\epsilon_l < 1$, normalization gives

$$\left\| \tilde{u}_i^{(l)} - u_i^{(l)} \right\|_2 \le \frac{2\epsilon_l}{1 - \epsilon_l},$$

and therefore

$$\left| \tilde{A}_{ij}^{(l)} - A_{ij}^{(l)} \right| \le \frac{4\epsilon_l}{1 - \epsilon_l}. \tag{19}$$

Thus, if shallow features have smaller perturbation ϵ_l, their pairwise relations are also more stable. For the sparse local relation matrix $R^{(l)} = \mathcal{S}_k(A^{(l)})$, this statement requires a local margin condition because top-k off-diagonal selection is discontinuous at neighbor-ordering ties. Let $\tau_l = 4\epsilon_l/(1 - \epsilon_l)$, and let $m_i^{(l)}$ be the gap between the k-th retained off-diagonal affinity of sample i and the largest unretained affinity. If $m_i^{(l)} > 2\tau_l$, the top-k off-diagonal neighborhood of sample i is unchanged under the perturbation bound above; in this region, sparsification and symmetrization are fixed linear operations, so the row-wise sparse relation perturbation is controlled by the dense relation perturbation up to constants. When this margin condition fails, we do not claim a deterministic

continuity bound; such samples correspond to locally ambiguous neighborhoods, for which relation drift is a useful warning signal rather than a certified perturbation measure.

Now decompose the measured cross-layer relation drift as

$$\begin{aligned}\left\|\tilde{R}_i^{(l+1)} - \tilde{R}_i^{(l)}\right\|_2 \le &\left\|R_i^{(l+1)} - R_i^{(l)}\right\|_2 \\ &+ \left\|\tilde{R}_i^{(l+1)} - R_i^{(l+1)}\right\|_2 \\ &+ \left\|\tilde{R}_i^{(l)} - R_i^{(l)}\right\|_2 . \end{aligned} \tag{20}$$

The first term is the clean model's natural relation evolution across depth; the last two terms are noise-induced relation perturbations. This inequality should be read as a diagnostic decomposition rather than a lower bound: by itself it does not prove that the measured drift must increase whenever the deeper-layer perturbation increases. Under the additional empirical regime where the clean relation pattern changes smoothly and the perturbation terms do not cancel the deeper noisy redirection, larger deeper-layer perturbations explain larger measured shallow-to-deep drift. The main paper's gradient sensitivity and prototype-flipping analyses provide evidence that TRACE operates in this regime.

The structural confidence

$$c_i^{\mathrm{str}} = \mathrm{Norm}\Big(\exp(-\gamma\tilde{\delta}_i)\Big), \qquad \delta_i = \sum_{l=1}^{L-1} \frac{\left\|\tilde{R}_i^{(l+1)} - \tilde{R}_i^{(l)}\right\|_2}{\sqrt{B}},$$

can therefore be interpreted as a monotone proxy for whether sample i's local geometry remains stable as supervision propagates through the network. A high value does not mean the shallow feature is perfectly clean, nor does it turn shallow neighbors into labels. It only says that the sample has not undergone strong shallow-to-deep relational drift. This conservative use of shallow structure is consistent with structure-aware noisy-label methods (Iscen et al., 2022; Yi et al., 2022; Wongso et al., 2023; Maini et al., 2023), while differing from nearest-neighbor correction because TRACE uses the anchor to estimate reliability rather than to rewrite the pseudo target.

Combining the two arguments, TRACE can be read as a reliability-preserving control rule. The SNR view explains why the two supervision branches need separate scores, and the relation-stability view explains why shallow anchored drift is a useful signal for refining observed-label trust. Together they support the main design choice: low observed-label reliability should create an opportunity for pseudo supervision, but only a reliable pseudo target should be allowed to fill it.

B.3 Hard-but-Clean Samples and the Scope of Relation Drift

Hard-but-clean samples clarify the intended scope of relation drift. A clean observed label does not imply that the corresponding sample must preserve an invariant neighborhood across depth. Samples near class boundaries, samples sharing visual attributes with other classes, and atypical instances within a class can undergo natural relational reorganization as the network builds increasingly task-specific features. Such drift may reflect sample difficulty rather than corruption. Consequently, c_i^{str} in Eq. (10) should be interpreted as a soft risk signal about the stability of the current representation path, not as a standalone clean/noisy decision.

TRACE incorporates this distinction directly into observed-label reliability. The structural confidence is blended with the loss posterior and modulated by prediction agreement, while β_t gradually activates the structure-aware terms. Thus, high relation drift alone is insufficient to determine s_i^{obs}: a hard-but-clean sample can still receive observed-label support through its loss-based evidence, and the agreement gate introduces no additional penalty when the two networks agree. Conversely, when the available signals jointly indicate uncertainty, the resulting reduction in s_i^{obs} expresses lower confidence in the observed label rather than a declaration that it is incorrect.

Lower observed-label reliability also does not force a correction. By Eq. (8), the pseudo-branch weight is $b_i = (1 - s_i^{\mathrm{obs}})s_i^{\mathrm{pseudo}}$, so relation drift can create a need for alternative supervision but cannot make the pseudo target influential without separate pseudo-target evidence. If both the observed-label and pseudo-target signals are weak, Eq. (15) reduces the sample's relative supervision strength subject to $w_{\min}$, rather than committing to either source. TRACE therefore treats

hard-but-clean ambiguity through continuous evidence fusion, independent pseudo-target filtering, and soft attenuation; relation drift modifies the level of trust assigned to a sample without being equated with label corruption.

B.4 Extended Preliminary Analysis

This subsection provides the extended version of the preliminary analysis summarized in the main paper, including additional backbone results and representation-level follow-up metrics under both synthetic and real-noise settings.

B.4.1 Details for Fig. 2

Fig. 2 summarizes the preliminary representation diagnostics on CIFAR-10 with 20%, 50%, and 80% symmetric noise. The three panels use the same representative backbones: ResNet-18, DenseNet-121, and WRN-28-10. The purpose is to separate three related effects of noisy supervision: how strongly gradients are perturbed, whether mislabeled samples move toward noisy prototypes, and whether the resulting features remain linearly useful.

Panel (i): layerwise gradient sensitivity. This panel reports the gradient-sensitivity ratio $G^{(l)}$ defined in the main text. Larger values mean that the layer's cross-entropy gradient under noisy supervision deviates more from the clean-supervision gradient. The key comparison is across depth and noise rate: deeper layers show larger increases as the corruption rate grows, indicating stronger noisy redirection in late representations.

Panel (ii): layerwise prototype flipping. This panel measures the fraction of mislabeled samples that are closer to their observed-label prototype than to their clean-label prototype. The contrast between layer1 and layer4 shows whether noisy labels merely perturb features or actively pull deep representations toward incorrect class anchors. The near-saturated layer4 flip rates for the representative backbones indicate that deep features can become strongly organized around noisy labels.

Panel (iii): linear-probe degradation. This panel reports the layer1–4 mean linear-probe accuracy after freezing the learned representations. Lower probe accuracy at higher noise rates means that the representation has become less directly aligned with clean semantics. Together with panels (i) and (ii), this supports the use of shallow relations as anchors and motivates separating observed-label reliability from pseudo-target reliability.

B.4.2 Details for Fig. 3

Fig. 3 is a conditional reliability diagnostic on CIFAR-100 with 50% symmetric label noise. Clean labels are used only to evaluate whether the observed label is correct and are not available to the training procedure. For every training sample, we obtain the loss posterior c_i^{loss} and the shallow-to-deep relation drift δ_i from the same model-scoring pass. The displayed loss-confidence strata are $\{0\}$, $(0, 0.2)$, $[0.2, 0.5)$, $[0.5, 0.8)$, $[0.8, 0.98)$, and $[0.98, 1]$.

Within-stratum drift comparison. Within each c_i^{loss} stratum, samples are ranked by δ_i and split at the stratum-wise median. The lower and upper halves are denoted low drift and high drift, respectively; because c_i^{str} is a monotone decreasing transformation of drift before dual-network fusion, these groups correspond to relatively high and low structural confidence. Each bar reports

$$\frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{G}} \mathbf{1}\left[y_i^{\mathrm{obs}} = y_i^{\star}\right],$$

where $\mathcal{G}$ is the corresponding loss-stratum/drift group and $y_i^{\star}$ is the clean reference label used only for analysis.

Interpretation and scope. The comparison controls loss confidence at the resolution of the displayed strata rather than claiming exact sample matching. Its purpose is to test whether relation drift provides conditional information beyond c_i^{loss}. The separation is strongest where loss-based evidence is weak and naturally contracts in high-confidence strata because observed-label correctness approaches its ceiling. Accordingly, Fig. 3 supports using c_i^{str} as a complementary modifier of the loss posterior, not as a standalone label-correctness certificate.

Table 11: Additional CIFAR-100 backbone diagnostics, with metrics defined in Tab. 10.

Backbone	Setting	ΔP_9	Flip4	Gap4	ΔRepair
ResNet-34	CIFAR100-Sym50	-0.014	0.999	3.546	-0.25
ResNet-50	CIFAR100-Sym50	-0.019	0.999	4.195	0.19
WRN-28-10	CIFAR100-Sym50	0.064	1.000	7.098	0.10
WRN-28-10	CIFAR-100N	0.123	1.000	7.255	0.46
MobileNetV2	CIFAR100-Sym50	0.013	0.826	0.789	0.53
MobileNetV2	CIFAR-100N	0.001	0.884	1.243	0.19
DenseNet-121	CIFAR100-Sym50	-0.017	0.999	3.396	0.79
DenseNet-121	CIFAR-100N	-0.025	0.999	4.832	0.87

Table 12: ResNet-18 representation diagnostics, averaged over layers 1–4. CKA: clean–noisy similarity; FDR: class separability.

Setting	CKA	FDR
CIFAR10-Sym20	0.7223	0.5780
CIFAR10-Sym50	0.6544	0.4339
CIFAR10-Sym80	0.5275	0.2877
CIFAR-10N-Agg	0.7606	0.7186
CIFAR-10N-Worst	0.6805	0.4372
CIFAR100-Sym20	0.6975	0.3849
CIFAR100-Sym50	0.6340	0.3305
CIFAR100-Sym80	0.5026	0.2828
CIFAR-100N	0.7186	0.4186

B.4.3 Additional Backbone Results Beyond Fig. 2

Tab. 9 clarifies why the main text focuses on ResNet-18, DenseNet-121, and WRN-28-10. ResNet-34 and ResNet-50 broadly follow the same trends as the main-text backbones, although the CIFAR-10N-Agg depth gap is weak for ResNet-34 and slightly negative for ResNet-50. MobileNetV2 provides a more substantial deviation in the prototype-flipping analysis, which is reported in Tab. 10: its layer4 flip rate is only 0.209 on CIFAR-10 with 50% symmetric noise and 0.584 on CIFAR-10N-Agg, far below the near-saturated behavior observed for the three representative backbones in the main paper. We therefore use the main text to emphasize the stable pattern shared by the three selected architectures, while retaining the remaining architectures here to document the broader empirical boundary.

For compact table captions, we use three protocol labels for auxiliary mechanism diagnostics: P9 denotes the noisy-gradient projection ratio on mislabeled samples, P10 denotes prototype-attraction diagnostics, and P11 denotes checkpoint-surgery repair.

The additional mechanism tables are meant to define the empirical boundary rather than to strengthen the claim into a universal depth law. P9 generally supports deeper-layer gradient redirection for ResNet/WRN/MobileNet-style backbones, but DenseNet-121 is often weaker or negative in the layer4–layer1 contrast. P10 is the most stable mechanism signal for ResNet, WRN, and DenseNet, whereas MobileNetV2 is a clear CIFAR-10 exception and becomes only moderately supportive on CIFAR-100. P11 remains an auxiliary intervention check because checkpoint surgery is harsh; its sign is useful mainly as relative evidence for whether deeper-stage repair is preferable to early-stage repair.

B.5 Additional Representation Analysis

Class separability degradation. We ask whether the diagnosed depth-dependent degradation is also visible through class separability. For layer feature T_l, we use the Fisher discriminant ratio

$$\mathrm{FDR}_l = \frac{\mathrm{tr}(S_b)}{\mathrm{tr}(S_w) + \varepsilon}, \tag{21}$$

Table 13: Clean–noisy CKA and FDR for additional backbones, averaged over layers 1–4.

Backbone	Setting	CKA	FDR
ResNet-34	CIFAR10-Sym50	0.6125	0.4029
ResNet-34	CIFAR-10N-Agg	0.7786	0.8044
ResNet-34	CIFAR-10N-Worst	0.6675	0.4304
ResNet-34	CIFAR100-Sym50	0.5552	0.2969
ResNet-50	CIFAR10-Sym50	0.7000	0.4127
ResNet-50	CIFAR-10N-Agg	0.7783	0.6502
ResNet-50	CIFAR-10N-Worst	0.7402	0.4342
WRN-28-10	CIFAR10-Sym50	0.6397	0.5356
WRN-28-10	CIFAR-10N-Agg	0.7904	1.3144
WRN-28-10	CIFAR-10N-Worst	0.6590	0.5685
WRN-28-10	CIFAR100-Sym50	0.5839	0.4041
WRN-28-10	CIFAR-100N	0.7019	0.5870
MobileNetV2	CIFAR10-Sym50	0.6702	0.7133
MobileNetV2	CIFAR-10N-Agg	0.7567	0.8522
MobileNetV2	CIFAR-10N-Worst	0.6990	0.7032
MobileNetV2	CIFAR100-Sym50	0.6460	0.4252
MobileNetV2	CIFAR-100N	0.7228	0.4902
DenseNet-121	CIFAR10-Sym50	0.6368	0.3335
DenseNet-121	CIFAR-10N-Agg	0.8299	0.6660
DenseNet-121	CIFAR100-Sym50	0.6392	0.3288
DenseNet-121	CIFAR-100N	0.7499	0.4327

Table 14: Checkpoint-surgery recovery relative to the noisy baseline. Prefix12 replaces layers 1–2 with clean counterparts; Suffix34h replaces layers 3–4 and the classifier.

Setting	Layer1	Layer4	Prefix12	Suffix34h
CIFAR10-Sym50	-48.80	-44.61	-49.01	-46.14
CIFAR-10N-Rand1	-71.84	-67.63	-73.86	-62.08
CIFAR100-Sym50	-39.86	-39.53	-40.01	-39.69
CIFAR-100N	-51.85	-51.94	-52.47	-51.54

where S_b and S_w are the between-class and within-class scatter matrices. Larger values indicate that classes remain well separated. As summarized in Tab. 12, the layer1–4 mean FDR on ResNet-18 drops from 0.5780 to 0.4339 to 0.2877 as CIFAR-10 symmetric noise increases from 20% to 80%, and from 0.3849 to 0.3305 to 0.2828 on CIFAR-100. Real-noise results show the same trend: CIFAR-10N-Worst is substantially below CIFAR-10N-Agg (0.4372 versus 0.7186). Tab. 13 reports the same CKA/FDR diagnostics for the remaining backbones wherever the corresponding representation suite is available.

Layerwise repair preference. Starting from a noisy checkpoint, we transplant selected stages from the corresponding clean model and measure recovery relative to the noisy baseline:

$$\mathrm{recover}(v) = \mathrm{Acc}(\text{repair variant } v) - \mathrm{Acc}(\text{noisy base}). \tag{22}$$

The checkpoint-surgery protocol is intentionally simple and therefore harsh, so the absolute numbers in Tab. 14 should be interpreted only directionally. The variants named in the table replace the corresponding clean stages into the noisy checkpoint: Layer1 and Layer4 replace a single residual stage, Prefix12 replaces layer1–2, and Suffix34h replaces layer3–4 plus the classifier head. Even so, the relative ordering is informative. On CIFAR-10 with 50% symmetric noise, repairing layer4 is less harmful than repairing layer1 (-44.61 versus -48.80), and repairing Suffix34h is better than repairing Prefix12 (-46.14 versus -49.01). The same ordering becomes clearer on CIFAR-10N-Rand1, where Suffix34h repair outperforms Prefix12 repair by more than 11 points.

Clean–noisy representation mismatch. Finally, we compare noisy representations with a clean reference model through linear CKA:

$$\mathrm{CKA}(X, Y) = \frac{\|X_c^\top Y_c\|_F^2}{\sqrt{\|X_c^\top X_c\|_F^2 \cdot \|Y_c^\top Y_c\|_F^2}}, \tag{23}$$

where X_c and Y_c are centered activations from the clean and noisy models. Lower CKA indicates that noisy representations have drifted farther from their clean counterparts. Tab. 12 shows the expected progression: on CIFAR-10 symmetric noise, the layer1–4 mean CKA declines from 0.7223 to 0.6544 to 0.5275 as the noise rate increases; on CIFAR-100, it declines from 0.6975 to 0.6340 to 0.5026. Real noise again tells the same story, with CIFAR-10N-Agg remaining closer to the clean reference than CIFAR-10N-Worst (0.7606 versus 0.6805).

C Experimental Protocols

C.1 Additional Experimental Details

The main evaluation metric is classification accuracy. Unless otherwise stated, CIFAR-style results report the average test accuracy over the last 10 epochs. In ablation and sensitivity tables with **Best** and **Last** columns, Best denotes the highest test accuracy reached by any checkpoint during training, while Last denotes the average test accuracy over the final 10 epochs. Large-scale real-noise experiments follow the reporting protocol of the corresponding base learner.

For synthetic noise, we use CIFAR-10/100 (Krizhevsky et al., 2009) with 20%, 50%, and 80% symmetric noise (Song et al., 2019), 40% pair/asymmetric noise (Song et al., 2019), and 40% instance-dependent noise (Chen et al., 2020). For real-world noisy datasets, we use CIFAR-10N/CIFAR-100N (Wei et al., 2022), including CIFAR-10N-Agg, CIFAR-10N-Rand1–3, CIFAR-10N-Worst, and CIFAR-100N; we also evaluate WebVision (Li et al., 2017), Food-101N (Lee et al., 2018), and Clothing1M under the reporting protocols used by the corresponding base methods.

TRACE is attached to representative refurbishment-based learners without changing their pseudo-target generators or training pipelines. For the RoLR instantiation, we reuse the GMM loss-clean posterior, add shallow structural confidence and dual-network agreement for observed-label reliability, and apply a separate confidence score to the pseudo branch. Augmentations and reproduced plug-in variants follow the corresponding base protocols. Experiments vary only the dataset, noise setting, base learner, or explicitly named ablation factors unless otherwise specified.

C.1.1 Component Ablation Variants

Tab. 5 compares variants with the same RoLR pseudo-target generator, data pipeline, backbone, and training schedule. Coupled null replaces RoLR's loss-based cleanliness coefficient with the refined observed-label score s_i^{obs}, but retains $b_i = 1 - s_i^{\mathrm{obs}}$ and uses unweighted soft-target cross-entropy. Decoupled reliability adds the pseudo-target gate, setting $b_i = (1 - s_i^{\mathrm{obs}})s_i^{\mathrm{pseudo}}$, while retaining the same loss. Decoupled w/o c_i^{str} removes structural confidence from this variant, leaving the loss posterior and agreement gate to determine observed-label reliability. Decoupled w/o g_i^{agr} instead disables the agreement gate by setting its multiplier to one. The final TRACE row adds the sample-wise weight w_i in Eq. (15) to the decoupled variant. Thus, the comparisons vary the reliability components and sample weighting while preserving the base pseudo-target construction and training protocol.

C.1.2 Synthetic Noise Construction

For CIFAR-style synthetic-noise experiments, we first construct a fixed noisy-label file for each dataset, noise type, noise rate, and random seed, and all compared methods are trained with the same observed labels. Let $y_i \in \{1, \ldots, C\}$ be the clean class of sample x_i, and let $\hat{y}_i$ be the label actually used for training.

Symmetric noise corrupts labels independently of the image content and treats all wrong classes uniformly (Jiang et al., 2018; Malach & Shalev-Shwartz, 2017; Song et al., 2019). At noise rate ξ, a sample keeps its clean label with probability $1 - \xi$; otherwise its label is replaced by one of the other $C - 1$ classes uniformly at random. Equivalently, the class-transition matrix has diagonal probability $1 - \xi$ and off-diagonal probability $\xi/(C - 1)$. This setting models unstructured random annotation errors.

Asymmetric (pair) noise is still class-conditional, but the wrong label is not chosen uniformly (Jiang et al., 2018; Malach & Shalev-Shwartz, 2017; Song et al., 2019). Instead, each corrupted source

Table 15: RoLR/TRACE test accuracy (%) under 20% pair and instance-dependent noise.

Method	CIFAR-10		CIFAR-100	
	Pair-20%	**Ins-20%**	**Pair-20%**	**Ins-20%**
RoLR	95.0	95.4	78.6	79.1
+ TRACE	**95.7**	**95.8**	**80.1**	**80.3**

class is mapped to a semantically related target class, so the transition matrix places most corrupted mass on a single paired class. For CIFAR-10, this follows the common benchmark mapping used by refurbishment baselines, such as bird $\rightarrow$ airplane, cat $\leftrightarrow$ dog, deer $\rightarrow$ horse, and truck $\rightarrow$ automobile. For CIFAR-100, we use the corresponding benchmark pair-noise labels from the same evaluation protocol. This noise type is harder than symmetric noise because the wrong labels are visually plausible rather than arbitrary.

Instance-dependent noise allows the corruption distribution to depend on the input x_i, following the standard synthetic protocol for instance-dependent label noise (Chen et al., 2020). Instead of using a single class-transition matrix shared by every sample in a class, each example receives its own transition vector T_i. The clean class keeps the remaining probability mass, while candidate wrong classes receive probabilities determined from sample-dependent scores. We then sample $\hat{y}_i$ from this per-example distribution. Thus, visually ambiguous or feature-specific samples can have different corruption tendencies even when they share the same clean class, which makes the setting closer to realistic annotation mistakes than class-conditional noise.

To test reliability directly, we also measure pseudo-target accuracy on all samples, low-clean samples, and low-clean+noisy samples. Low-clean samples are those assigned low observed-label reliability by the corresponding method, and low-clean+noisy samples further restrict this subset to examples whose observed label is actually corrupted. When ground-truth clean labels are available for analysis, pseudo-target accuracy is computed by comparing the selected pseudo target with the clean label. Follow-noisy is computed only on observed-noisy samples and measures how often the pseudo target agrees with the noisy observed label. HC Wrong denotes high-confidence pseudo targets that are nevertheless incorrect, so lower values indicate safer pseudo supervision. Representation-level follow-up metrics are reported to connect the final method back to the preliminary analysis.

The qualitative examples in Fig. 5 are shown from CIFAR-10's native 32×32 images; their limited visual sharpness reflects the dataset resolution rather than a rendering artifact.

D Additional Experimental Results

This section collects supplementary experimental results beyond the main paper.

D.1 Results under 20% Pair and Instance-dependent Noise

To complement the 40% pair and instance-dependent noise results in Tab. 1, we additionally evaluate RoLR and TRACE under the milder 20% settings on CIFAR-10 and CIFAR-100. We keep the backbone, noisy-label realization, training schedule, and reporting protocol matched within each pair, changing only the reliability mechanism.

D.2 Additional Hyperparameter Sensitivity

Tab. 16 reports the remaining one-factor-at-a-time sensitivity settings on CIFAR-100 with 50% symmetric noise. These factors control the structural anchor, scheduling, agreement attenuation, and minimum supervision strength. They are kept in the appendix because they are implementation-facing checks rather than the two branch-defining hyperparameters emphasized in the main text.

These additional sweeps support the default choices used in the main experiments. A moderate neighborhood size, temperature, pseudo-score start time, agreement penalty, and minimum sample weight provide the best or most balanced accuracy–reliability trade-off. More conservative pseudo

Table 16: One-factor sensitivity on CIFAR-100 (50% symmetric noise); other defaults are fixed. HC Wrong: high-confidence pseudo-target error rate (lower is better).

Factor	Value	Best	Last	Pseudo Acc.	HC Wrong
Relation neighbors k	20	76.84	76.10	84.52	6.24
	50	**76.95**	**76.22**	**84.70**	**6.16**
	100	76.76	76.03	84.35	6.31
Drift temperature γ	2.5	76.88	76.12	84.61	6.22
	5.0	**76.95**	**76.22**	**84.70**	**6.16**
	10.0	76.71	75.96	84.29	6.38
Pseudo-score start epoch	20	76.86	76.11	84.46	6.32
	30	**76.95**	**76.22**	**84.70**	6.16
	40	76.69	75.91	84.28	**6.09**
Agreement penalty λ_{dis}	0.25	76.74	75.99	84.33	6.41
	0.5	**76.95**	**76.22**	**84.70**	6.16
	0.75	76.82	76.08	84.55	**6.04**
Minimum sample weight $w_{\min}$	0.0	76.63	75.87	84.26	6.47
	0.2	**76.95**	**76.22**	**84.70**	6.16
	0.4	76.79	76.04	84.49	**6.02**

filtering can slightly reduce high-confidence errors, but it also lowers best/last accuracy and pseudo-target accuracy, so we keep the middle settings as the default configuration. Across the main and appendix sweeps, performance changes remain small and smooth, indicating that TRACE's two-source reliability interface is not tied to a fragile hyperparameter setting.

D.3 Additional Reliability Statistics

Tab. 17 extends the main-text reliability summary with three complementary views of pseudo-target quality. All Pseudo Acc. measures global pseudo-target accuracy, Low-clean Pseudo Acc. focuses on samples assigned low observed-label reliability by the corresponding method, and Low-clean+Noisy Pseudo Acc. further restricts that subset to samples whose observed label is actually corrupted. The last view most directly tests the supervision used when refurbishment needs a replacement target. Because the low-clean subset is selected separately by RoLR and TRACE, its two entries are diagnostic rather than a matched-sample comparison.

TRACE improves Low-clean+Noisy Pseudo Acc. in every reported setting. The gains are largest across the three CIFAR-100 symmetric-noise rates, where the larger label space makes reliable replacement harder, and remain positive under CIFAR-10 and human-annotation noise. Global pseudo-target accuracy also generally improves, although the small decreases on CIFAR100-Sym20 and CIFAR-100N mark the boundary of this effect. Together, the results show that TRACE's most consistent benefit occurs on actually corrupted samples that depend on the pseudo branch, rather than uniformly improving every method-dependent subset.

D.4 CIFAR-100N Hard-Regime Diagnostics

Tab. 18 provides subset-level diagnostics for the hardest real-noise regime. CIFAR-100N is where the main-text reliability gain is positive but small, and the subset statistics show that pseudo targets can still follow noisy observed labels. Follow-noisy measures pseudo-target agreement with the noisy observed label, and is computed only where the observed label is noisy. On observed-noisy samples, pseudo-target accuracy drops to 30.93%, while follow-noisy remains 51.72%, reinforcing that pseudo-target reliability is related to sample cleanliness but cannot be reduced to it.

Table 17: Pseudo-target accuracy (%) across datasets and noise settings. Low-clean+noisy: low estimated cleanliness with a corrupted observed label.

Setting	All Pseudo Acc. RoLR / TRACE	Low-clean Pseudo Acc. RoLR / TRACE	Low-clean+Noisy Pseudo Acc. RoLR / TRACE
Synthetic Noise			
CIFAR10-Sym20	98.38 / 98.58	96.05 / 93.20	96.89 / 98.84
CIFAR10-Sym50	96.34 / 97.16	95.93 / 94.51	96.42 / 97.52
CIFAR10-Sym80	93.63 / 94.68	94.82 / 93.55	95.17 / 95.21
CIFAR100-Sym20	93.51 / 93.05	73.14 / 74.85	76.51 / 93.02
CIFAR100-Sym50	83.47 / 84.56	70.75 / 74.15	72.92 / 84.70
CIFAR100-Sym80	44.11 / 53.66	38.20 / 50.20	42.41 / 54.09
Real Noise			
CIFAR-10N-Agg	96.58 / 97.42	90.96 / 89.44	93.02 / 94.07
CIFAR-10N-Rand1	95.80 / 96.89	93.51 / 93.07	95.48 / 96.16
CIFAR-10N-Rand2	95.70 / 96.82	93.98 / 92.73	95.42 / 95.70
CIFAR-10N-Rand3	94.76 / 96.81	94.92 / 92.93	95.81 / 96.02
CIFAR-10N-Worst	91.23 / 94.39	92.87 / 94.36	93.83 / 95.78
CIFAR-100N	71.86 / 71.22	57.43 / 57.82	59.68 / 60.38

Table 18: TRACE reliability on CIFAR-100N subsets.

Subset	Count	Pseudo Acc.	Follow-noisy
All	50000	71.22	51.72
Low-clean	10466	57.82	4.91
Observed-noisy	20100	30.93	51.72
Low-clean+noisy	9993	60.38	4.91

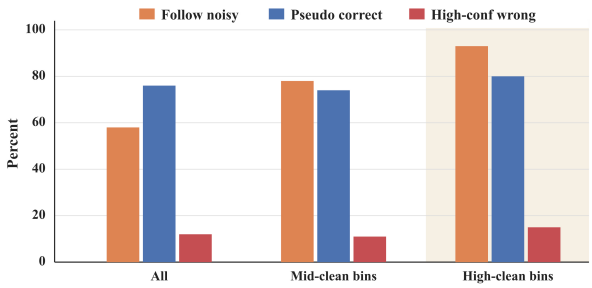


Figure 8: Pseudo-target mismatch on 800 randomly sampled Clothing1M noisy-training images, using GPT-5.6 API annotations as surrogate reference labels.

D.5 Clothing1M Cross-Dataset Diagnostic

We further examine whether the pseudo-target mismatch in Fig. 1 also appears under large-scale real-world label noise. We randomly sample 800 images from the noisy training split of Clothing1M. Its training images were collected from online shopping websites, and their noisy labels were derived from surrounding text and label keywords rather than image-level manual verification Xiao et al. (2015). The benchmark also contains a clean subset formed by manually refining 72,409 image labels, whereas the 800 images analyzed here come from the million-image noisy split. We therefore query the GPT-5.6 API to assign one of the 14 Clothing1M categories to each sampled image and use these annotations as surrogate reference labels for this diagnostic. Following the analysis in Fig. 1, we report how often the pseudo target follows the noisy observed label, matches the surrogate reference label, or remains wrong at high confidence across estimated-cleanliness strata.

As shown in Fig. 8, the high-clean bins exhibit the clearest mismatch: pseudo targets follow the noisy observed labels on 93% of samples, but match the surrogate reference labels on only 80%, while 15% remain high-confidence errors. A similar gap appears in the mid-clean bins. These results qualitatively reproduce the phenomenon in Fig. 1 under a distinct, real-world noise source:

a high cleanliness estimate does not by itself certify the pseudo target. Because the reference labels are model-generated and the analysis covers only 800 sampled images, this experiment is intended as evidence of recurrence rather than a dataset-wide prevalence estimate or a substitute for human-verified ground truth.

E FIGURE DETAILS

E.1 DETAILS FOR FIG. 1

Fig. 1 is constructed from the CIFAR-100N real-noise setting under the noisy condition. The statistics are exported from the same pseudo-target reliability analysis pipeline used throughout the paper. Samples are first grouped by the estimated cleanliness score into ten equal-width bins over $[0, 1]$, and we then summarize how pseudo-target behavior changes as samples move from low-clean to high-clean regimes.

Panel (a): bin-wise pseudo-target mismatch. The horizontal axis is the midpoint of each cleanliness bin, and the vertical axis reports percentages. The blue curve ("Pseudo correct") is the fraction of samples in that bin whose pseudo target matches the ground-truth class. The orange dashed curve ("Follow noisy label") is computed only on observed-noisy samples and measures how often the pseudo target instead agrees with the noisy observed label. The shaded region marks the high-clean regime with bin midpoint at least 0.65. This panel is intended to show that once samples move into apparently clean regions, pseudo targets can still inherit the noisy label direction rather than becoming uniformly trustworthy replacements.

Panel (b): aggregate failure across cleanliness regimes. The horizontal axis groups samples into three regimes: all samples, mid-clean bins $[0.35, 0.65)$, and high-clean bins $[0.65, 1.0]$. The vertical axis again reports percentages. The orange bars ("Follow noisy") aggregate, over observed-noisy samples only, the rate at which the pseudo target agrees with the noisy observed label. The blue bars ("Pseudo correct") aggregate the overall pseudo-target accuracy in that regime. The red bars ("High-conf wrong") report the fraction of samples whose pseudo target is wrong despite satisfying the high-confidence criterion used by the reliability-analysis pipeline. The bracket in the high-clean regime highlights the gap between noisy-label following and actual pseudo-target correctness.

Main takeaway. The central message of Fig. 1 is not merely that noisy samples are hard; it is that high estimated sample cleanliness does not automatically imply high pseudo-target reliability. In the high-clean regime, the pseudo target follows the noisy observed label 99.7% of the time, while being correct on only 74.7% of samples and still making 19.2% high-confidence errors. This mismatch motivates the main paper's decision to treat observed-label reliability and pseudo-target reliability as related but distinct quantities.

E.2 DETAILS FOR FIG. 6

Fig. 6 reports the pseudo-target reliability diagnostics on CIFAR-100 with 50% symmetric noise. The same trained baseline and TRACE-enhanced model are evaluated against the clean ground-truth labels, which are used only for analysis. Low-clean samples are those assigned low observed-label reliability, high-pconf samples are those whose pseudo targets pass the high-confidence criterion, and low-clean+noisy samples further restrict low-clean examples to cases whose observed label is actually corrupted.

Panel (a): reliability by bin. Samples are grouped by clean-probability bins. The observed-clean curve measures how often the observed label is correct in each bin, while the pseudo-correct curve measures how often the selected pseudo target matches the clean class. The panel shows that pseudo-target correctness does not simply mirror observed-label cleanliness, motivating a separate pseudo-target reliability estimate.

Panel (b): pseudo-target accuracy. This panel compares RoLR and TRACE across all samples, low-clean samples, high-pconf samples, and low-clean+noisy samples. The low-clean+noisy subset is the most diagnostic regime because the observed label is both unreliable and actually wrong, so refurbishment depends heavily on whether the pseudo branch is trustworthy.

Panels (c) and (d): error quality. Panel (c) reports high-confidence pseudo-target errors, where lower values mean that fewer confident pseudo targets are wrong. Panel (d) reports pseudo-target calibration error. Together, these panels show whether the method only increases pseudo-target usage or also makes pseudo supervision safer.

Main takeaway. Fig. 6 supports the main text's error-quality claim: TRACE improves pseudo-target accuracy most strongly in the low-clean+noisy regime, while also reducing high-confidence wrong pseudo targets and lowering calibration error. The gain is therefore not only a quantity increase in pseudo supervision, but an improvement in the reliability of the supervision that replaces corrupted labels.

E.3 Details for Fig. 7

Fig. 7 visualizes pseudo-target accuracy over a two-dimensional reliability grid on CIFAR-100 with 50% symmetric noise. The horizontal axis bins observed-label reliability into five intervals, and the vertical axis bins pseudo-target reliability into the same intervals. Each nonempty cell reports pseudo-target accuracy within that reliability pair; cells covering less than 0.01% of samples are omitted to avoid over-interpreting extremely sparse regions.

Panel (a): RoLR landscape. For RoLR, the observed axis is the original loss-based cleanliness score, while the pseudo axis reflects the pseudo-target confidence used for diagnostic binning. The landscape shows that low observed-label reliability alone does not guarantee reliable pseudo supervision: pseudo-target accuracy remains low in low-pseudo cells and becomes useful mainly when pseudo reliability is high.

Panel (b): TRACE landscape. For TRACE, the observed axis uses the refined observed-label reliability score and the pseudo axis uses the separate pseudo-target reliability score. Compared with RoLR, TRACE improves the key low-observed / high-pseudo region, where pseudo supervision is most needed because the observed label is unreliable but a confident replacement is available.

Main takeaway. The landscape explains why TRACE uses two gates rather than a single cleanliness coefficient. A sample should not move automatically from the observed-label branch to the pseudo branch just because observed-label reliability is low; it should receive strong pseudo supervision only in cells where pseudo-target reliability is also high.

F Plug-in Implementation and Computational Overhead

F.1 TRACE Pseudocode

F.2 Plug-in Implementation Notes

TRACE is implemented without changing the base method's pseudo-target generator or data pipeline. Given the base cleanliness score, it computes the shallow-anchored structural confidence, applies the dual-network agreement gate to refine observed-label reliability, filters pseudo targets with a separate confidence score, and combines the two branches through the unified supervision weight described in the main method section.

In the compact loss in Eq. (15), p_i denotes the prediction of the network currently being optimized. In two-network refurbishment pipelines, the same target correction and sample reweighting are applied symmetrically to the paired networks.

In the RoLR instantiation, the original GMM loss-clean posterior is reused as the loss-side input to observed-label reliability. The structural branch is activated after warm-up, computed from sparse local relations, and ramped in gradually so that early noisy features do not dominate the reliability estimate. The pseudo branch is activated with the same start epoch as the structural branch and uses only pseudo-target confidence, leaving pseudo-target construction itself unchanged. Thus, all ablation factors in the main paper modify only the reliability interface, not the underlying pseudo-label generator.

Algorithm 1 TRACE reliability interface for a mini-batch

Require: mini-batch $\{x_i, \hat{\mathbf{y}}_i\}_{i=1}^{B}$, network predictions $p_i^{(1)}, p_i^{(2)}$, layer features $\{h_l^{(1)}(x_i), h_l^{(2)}(x_i)\}_{l=1}^{L}$, loss-clean posterior c_i^{loss}
Schedules/hyperparameters: $\beta_t, T, k, \gamma, \alpha, \lambda_{\text{dis}}, \rho, \varepsilon, w_{\min}$, pseudo-score start epoch
Ensure: corrected target $\tilde{\mathbf{y}}_i$, sample weight w_i, weighted training loss $\mathcal{L}$

$q_i \leftarrow \text{Sharpen}_T\left(\frac{p_i^{(1)}+p_i^{(2)}}{2}\right)$ for all i
$\alpha_t \leftarrow 1 - \beta_t(1-\alpha)$
for $m \in \{1, 2\}$ **do**
  for $l = 1, \ldots, L$ **do**
    form cosine relation matrix $A^{(l,m)}$ from normalized features $\{h_l^{(m)}(x_i)\}_{i=1}^{B}$
    keep self and top-k off-diagonal local affinities: $R^{(l,m)} \leftarrow \mathcal{S}_k(A^{(l,m)})$
  end for
  for each sample i **do**
    $\delta_{i,m} \leftarrow \sum_{l=1}^{L-1} \|R_i^{(l+1,m)} - R_i^{(l,m)}\|_2 / \sqrt{B}$
    $c_{i,m}^{\text{str}} \leftarrow \text{Norm}(\exp(-\gamma\, \text{Norm}(\delta_{i,m})))$
  end for
end for
$\bar{c}_i^{\text{str}} \leftarrow \frac{1}{2}(c_{i,1}^{\text{str}} + c_{i,2}^{\text{str}})$ for all i
$g_i^{\text{agr}} \leftarrow 1$ if $\arg\max p_i^{(1)} = \arg\max p_i^{(2)}$, else λ_{dis}
$s_i^{\text{obs}} \leftarrow \text{clip}\big((\alpha_t c_i^{\text{loss}} + (1-\alpha_t)\bar{c}_i^{\text{str}})\big((1-\beta_t) + \beta_t g_i^{\text{agr}}\big), 0, 1\big)$
$s_i^{\text{pseudo}} \leftarrow (\max_c q_{i,c})^{\rho}$ after pseudo-score warm-up
$a_i \leftarrow s_i^{\text{obs}}, \quad b_i \leftarrow (1 - s_i^{\text{obs}}) s_i^{\text{pseudo}}$
$\tilde{\mathbf{y}}_i \leftarrow \frac{a_i \hat{\mathbf{y}}_i + b_i q_i}{a_i + b_i + \varepsilon}$
$w_i \leftarrow \max(a_i + b_i, w_{\min})$
normalize $\{w_i\}$ by the batch mean and optimize the base soft-target loss with the original prior regularizer
return $\tilde{\mathbf{y}}_i$, w_i, and $\mathcal{L}$

F.3 Computational Overhead

TRACE adds one batchwise relation computation for the structural confidence and a lightweight agreement check between the paired networks. The structural term is computed from the intermediate features already produced by the forward pass: for each analyzed layer, we normalize the B mini-batch features, form a $B \times B$ cosine relation matrix, keep the self-affinity and top-k off-diagonal local affinities per sample, and compare sparse relation rows across depth. This requires no dataset-level neighbor graph, memory bank, or clean validation pass. Its temporary dense memory is $O(LB^2)$ for L analyzed layers and batch size B, and the retained sparse relations are $O(LB(k+1)) = O(LBk)$. With the default CIFAR-style setting $B = 64$ and $k = 50$, this overhead is small relative to storing backbone activations. The dual-network agreement gate reuses the predictions already available in two-network refurbishment pipelines, so it does not introduce an additional model or extra forward pass.

G Limitations and Future Extensions

TRACE is designed for refurbishment-based pipelines with explicit pseudo-target construction and is evaluated on closed-set image classification, so its applicability to other tasks remains unverified. Its model-derived reliability scores can still inherit noisy bias and cannot rule out confidently incorrect pseudo targets. The hardest real-noise setting, CIFAR-100N, also shows that reliability decoupling improves the difficult subset only modestly. Future extensions could test stronger backbones, additional refurbishment pipelines, and broader reliability signals beyond the shallow structural anchor used here.

Open-set label noise is a fundamentally different research setting from the closed-set noise considered in this work: some corrupted samples belong to classes outside the predefined label space, so the system must identify class-external samples rather than merely decide whether an in-distribution label or replacement target is reliable Wang et al. (2018). Extending TRACE to this setting therefore requires an explicit rejection mechanism and open-set sample identification, not a direct reuse

of the current scores. We plan to pursue this direction by integrating open-set detection with the two-source reliability framework.