ACTIVATION-DISTRIBUTION PROTOTYPES FOR ROBUST LEARNING UNDER NOISY SUPERVISION

Anonymous authors
Paper under double-blind review

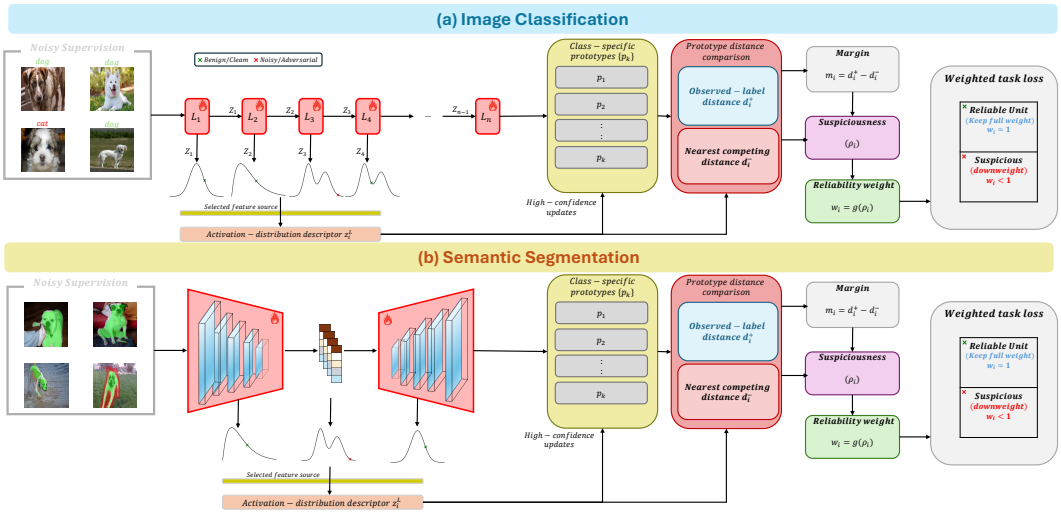

Figure 1: Overview of the proposed activation-distribution reliability framework for (a) image classification and (b) semantic segmentation. A selected intermediate feature source is summarized by an activation-distribution descriptor, which is compared against class-specific activation-distribution prototypes. The observed-label and nearest-competing prototype distances define the disagreement margin $m_i = d_i^+ - d_i^-$, which is converted into a suspiciousness score ρ_i and reliability weight $w_i \in [\alpha_{\min}, 1]$. High-confidence predictions update the prototypes, while supervision inconsistent with the class-specific activation distribution is adaptively downweighted during training.

ABSTRACT

Deep neural networks trained with noisy annotations may eventually memorize corrupted supervision because standard objectives treat all labels as equally reliable. We introduce a distribution-aware framework for estimating supervision reliability from intermediate network activations. Rather than representing a supervised unit only as a point in feature space, we summarize its activation behavior through sparsity, mean, standard deviation, skewness, and kurtosis. High-confidence predictions are used to construct class-specific activation-distribution prototypes. Reliability is then estimated by comparing the descriptor of each unit with the prototype of its observed label and the nearest competing prototype. When the activation distribution is more consistent with a competing class, the corresponding supervision is smoothly downweighted according to the resulting disagreement margin. This yields an online reliability signal complementary to prediction confidence, task loss, and point-wise feature representations. Experiments across image classification and semantic segmentation, synthetic and structured corruption, real-world label noise, and CNN and ViT architectures demonstrate consistent robustness gains. Direct reliability evaluation further shows that activation-distribution disagreement separates corrupted supervision more effectively than confidence-, loss-, feature-prototype-, and Mahalanobis-based alternatives. These results show that class-specific activation-distribution modeling provides a practical and informative signal for learning under imperfect supervision.

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.

1 INTRODUCTION

Deep neural networks achieve strong performance across visual recognition tasks, yet standard training typically assumes that supervision is reliable. In practice, annotations obtained from human labeling, web data, weak supervision, pseudo-labeling, or semi-automatic pipelines frequently contain errors. Under such supervision, standard objectives assign equal importance to reliable and corrupted targets, allowing incorrect labels to generate misleading gradients and eventually be memorized by the network. Robust learning under noisy supervision therefore depends critically on identifying which supervision should be trusted.

Existing approaches address this problem through robust objectives, sample selection, and reliability estimation. Bootstrapping combines observed labels with model predictions Reed et al. (2014), while generalized and symmetric cross-entropy reduce sensitivity to corrupted labels Zhang & Sabuncu (2018); Wang et al. (2019). MentorNet and Co-teaching identify reliable samples through curriculum learning or small-loss selection Jiang et al. (2018); Han et al. (2018), whereas DivideMix separates likely clean and noisy samples within a semi-supervised framework Li et al. (2020). Despite their effectiveness, these methods primarily estimate reliability from prediction confidence, loss values, training dynamics, or model agreement. We investigate a different source of information: the *distributional structure of intermediate activations*.

Intermediate representations contain more than a point location in feature space. They also encode how a network internally expresses an input through characteristic activation statistics. We hypothesize that these statistics exhibit class-specific structure, such that corrupted supervision can be revealed when the observed label disagrees with the activation behavior associated with that class. This leads to the central question of this work: **can class-specific activation distributions provide a reliable signal for identifying supervision that is inconsistent with the underlying data?**

We answer this question with a confidence-gated framework based on *class-specific activation-distribution prototypes*. Each supervised unit is represented by a compact descriptor summarizing activation sparsity, mean, standard deviation, skewness, and kurtosis. High-confidence model predictions are used to construct and update a prototype for each class, capturing its characteristic intermediate activation behavior. Reliability is then estimated directly in this distributional space. For a unit with observed target $\tilde{y}_i$, we compare its descriptor with the prototype of $\tilde{y}_i$ and with the nearest competing class prototype. If the activation distribution is more consistent with a competing class, the resulting prototype-disagreement margin signals potentially unreliable supervision, and the corresponding task-loss contribution is smoothly reduced.

The proposed signal is fundamentally different from conventional confidence-, loss-, and embedding-based criteria. Hard-example mining identifies difficult samples through large task losses Shrivastava et al. (2016), while metric-learning methods reason about distances between point-wise embeddings Schroff et al. (2015); Snell et al. (2017); Wen et al. (2016). In contrast, our method asks whether the *statistical behavior of intermediate activations* is compatible with the observed supervision. The weighting mechanism itself is deliberately simple; the main contribution is the reliability signal that makes this weighting informative. Empirically, we show that activation-distribution disagreement provides stronger corruption signals than confidence-, loss-, Mahalanobis-, and point-prototype alternatives, that clean and corrupted supervision become increasingly separated as the prototypes stabilize, and that the same formulation transfers across classification, segmentation, CNN, and ViT architectures.

Our contributions are summarized as follows:

We introduce **class-specific activation-distribution prototypes**, which model the statistical behavior of intermediate activations rather than only their point-wise feature location, providing a complementary signal for supervision reliability.

We formulate reliability through an **assigned-label versus nearest-competing prototype disagreement margin** in activation-distribution space, allowing suspicious supervision to be identified and smoothly downweighted according to the strength of the disagreement.

We demonstrate the applicability of the same reliability mechanism across image classification and semantic segmentation under image-level, pixel-level, structured, and real-world label noise, with consistent improvements over conventional loss-, confidence-, and prototype-based alternatives.

2 Related Work

Learning with Noisy Labels Learning with noisy labels reduces the influence of unreliable supervision through robust objectives, sample selection, and reliability estimation. Generalized and symmetric cross-entropy improve robustness to corrupted labels Zhang & Sabuncu (2018); Wang et al. (2019), while dynamics-aware losses use training behavior to suppress memorized noise Li et al. (2023b). MentorNet, Co-teaching, and DivideMix estimate reliable samples through curriculum learning, small-loss selection, or semi-supervised separation Jiang et al. (2018); Han et al. (2018); Li et al. (2020). Related approaches include label correction, sample refurbishment, probabilistic correction, meta-learned weighting, agreement regularization, early-learning regularization, and constrained reweighting Patrini et al. (2017); Song et al. (2019); Yi & Wu (2019); Shu et al. (2019); Wei et al. (2020); Liu et al. (2020); Kumar & Amid (2021). Recent work further studies prediction-consistency regularization, asymmetric robust losses, and debiased sample selection Sun et al. (2024); Wang et al. (2025); Pan et al. (2026). Most existing methods infer reliability from confidence, loss, training dynamics, corrected labels, or model agreement. Our method instead uses the class-conditional distributional behavior of intermediate activations as a complementary supervision-reliability signal.

Prototype-Based Representation Learning. Prototype-based methods summarize classes using representative feature vectors. Prototypical Networks perform classification according to distances from class prototypes Snell et al. (2017), center loss encourages features from the same class to concentrate around learned centers Wen et al. (2016), and proxy-based metric-learning methods replace individual samples with class-level representatives. More recent noisy-label approaches similarly exploit prototype structure, including multi-prototype modeling for separating reliable and corrupted samples Zhang et al. (2025), class-prototype modeling for robust representation learning Xia et al. (2025), and noise-aware prototype-assisted contrastive learning Ghazal (2026). These methods primarily construct prototypes from point-wise feature embeddings. Our formulation differs in the object represented by the prototype. Rather than encoding a location or mode in embedding space, our activation-distribution prototypes characterize class-specific statistical behavior of intermediate activations through sparsity, mean, standard deviation, skewness, and kurtosis. Reliability is then determined by comparing the observed-label prototype with the nearest competing activation-distribution prototype, providing a signal based on internal activation behavior rather than point-wise feature geometry.

Activation Statistics and Distributional Representations Statistics of intermediate activations are fundamental to deep neural networks. Batch normalization uses batch-level first- and second-order statistics Ioffe & Szegedy (2015), while layer and group normalization use alternative statistical groupings to stabilize representation learning Ba et al. (2016); Wu & He (2018). Intermediate representations have also been used for uncertainty and out-of-distribution analysis. Mahalanobis-based methods estimate class-conditional structure in feature space Lee et al. (2018), while deep k-nearest-neighbor approaches exploit representations from multiple layers for interpretable and robust prediction Papernot & McDaniel (2018). These works demonstrate that intermediate features contain information not fully captured by the final softmax output. Our objective is different: we use statistics of intermediate activations as a *training-time supervision reliability signal*. We construct class-specific activation-distribution prototypes online and measure whether the activation behavior of a supervised unit agrees with its observed target or is more consistent with a competing class. This signal is then used to control the influence of potentially corrupted supervision during optimization.

Noisy Dense Supervision Semantic segmentation introduces additional forms of annotation noise, including inaccurate boundaries, missing regions, spurious foreground, and class-confusion errors. Yi et al. (2021) studied noisy pixel-level supervision in semi-supervised segmentation and showed that noisy pseudo masks can remain useful when combined with limited clean supervision. Li et al. (2023a) introduced Diverse Learning Groups and evaluated robustness to pseudo-label noise, random dilation and erosion, and similar-class perturbations on PASCAL VOC 2012 Everingham et al. (2010). These approaches develop mechanisms specifically for segmentation under imperfect masks. Our formulation is not tied to a particular segmentation corruption model. Instead, it operates on a generic *supervised unit*, which may correspond to an image, region, proposal, mask, pixel group, or other prediction entity. The same activation-distribution reliability principle can

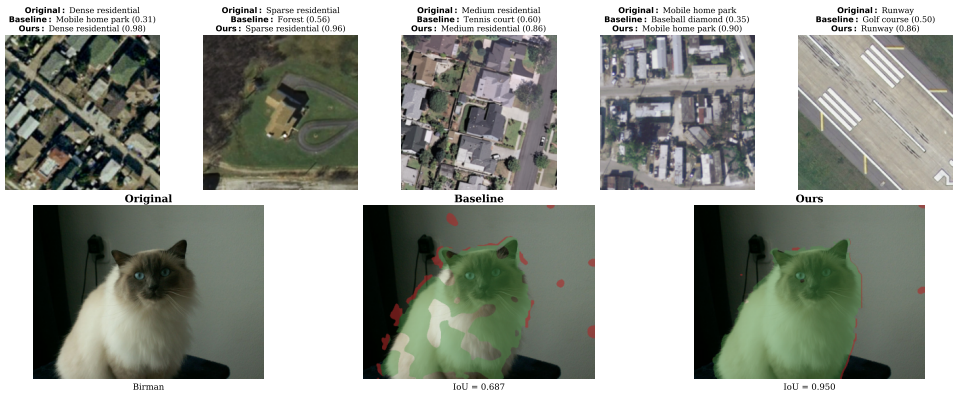

Figure 2: Qualitative comparison under noisy supervision. **Top:** classification examples with corrupted labels, where $\tilde{y}$ denotes the observed noisy label and y the clean label. The baseline predicts an incorrect class, whereas the proposed method recovers the clean class. **Bottom:** segmentation predictions under pixel-level label noise. Green denotes correctly predicted foreground pixels, $\hat{M} \cap M$, and red denotes false-positive pixels, $\hat{M} \setminus M$. The proposed method reduces prediction leakage outside the clean foreground region.

therefore be applied to image-level and dense supervision without changing the underlying criterion used to assess reliability. Unlike existing reliability criteria based on losses, confidence, training dynamics, or point-wise feature distances, we model class-specific statistical behavior of intermediate activations. Reliability is then estimated through the observed-label versus nearest-competing activation-distribution prototype margin.

3 Methodology

We estimate supervision reliability from class-specific activation-distribution prototypes. Let u_i denote a supervised unit with possibly corrupted target $\tilde{y}_i$, and let $A_i^\ell \in \mathbb{R}^{C_\ell \times H_\ell \times W_\ell}$ be its intermediate activation tensor at layer ℓ. The framework has three steps: intermediate activations are summarized by an activation-distribution descriptor, high-confidence predictions construct class-specific prototypes, and reliability is determined by comparing the observed-label prototype with the nearest competing prototype. Figure 1 summarizes the framework.

3.1 Activation-Distribution Descriptor

A conventional feature representation describes where a supervised unit lies in embedding space; our descriptor instead summarizes how its intermediate activations are distributed. For image-level supervision, let $A_i^\ell \in \mathbb{R}^{C_\ell \times H_\ell \times W_\ell}$ be the activation tensor at layer ℓ, $N_\ell = H_\ell W_\ell$, and $\{x_{i,q,n}^\ell\}_{n=1}^{N_\ell}$ the spatial activations of channel q. Each channel is summarized by five statistics:

$$
\begin{aligned}
r_{i,q}^\ell &= \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} \mathbb{I}\big[|x_{i,q,n}^\ell| \le \epsilon\big], & \mu_{i,q}^\ell &= \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} x_{i,q,n}^\ell, \\
\sigma_{i,q}^\ell &= \sqrt{\frac{1}{N_\ell} \sum_{n=1}^{N_\ell} \big(x_{i,q,n}^\ell - \mu_{i,q}^\ell\big)^2 + \varepsilon}, & \gamma_{i,q}^\ell &= \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} \left(\frac{x_{i,q,n}^\ell - \mu_{i,q}^\ell}{\sigma_{i,q}^\ell}\right)^3, \\
\kappa_{i,q}^\ell &= \frac{1}{N_\ell} \sum_{n=1}^{N_\ell} \left(\frac{x_{i,q,n}^\ell - \mu_{i,q}^\ell}{\sigma_{i,q}^\ell}\right)^4.
\end{aligned}
\tag{1}
$$

These statistics capture near-zero activation ratio, mean, standard deviation, skewness, and kurtosis, with $\varepsilon > 0$ for numerical stability. The channel descriptor is $z_{i,q}^\ell = [r_{i,q}^\ell, \mu_{i,q}^\ell, \sigma_{i,q}^\ell, \gamma_{i,q}^\ell, \kappa_{i,q}^\ell]$,

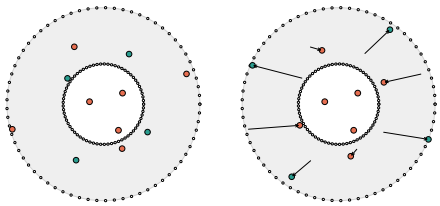

Figure 3: Reliability estimation in activation-distribution space. d_i^+ and d_i^- denote the observed-label and nearest-competing prototype distances, respectively. Their difference $m_i = d_i^+ - d_i^-$ determines the reliability assigned to the supervised unit.

and the image descriptor is $z_i^\ell = \mathrm{concat}_{q=1}^{C_\ell} z_{i,q}^\ell \in \mathbb{R}^{5C_\ell}$. For semantic segmentation, each spatial location s is treated as a supervised unit. Given $a_{i,s}^\ell \in \mathbb{R}^{C_\ell}$, the same five statistics are computed across channels, yielding $z_{i,s}^\ell = [r_{i,s}^\ell, \mu_{i,s}^\ell, \sigma_{i,s}^\ell, \gamma_{i,s}^\ell, \kappa_{i,s}^\ell] \in \mathbb{R}^5$. Thus, although the same statistical descriptor is used in both tasks, the distribution axis is chosen according to the supervised unit: classification summarizes spatial activations within each channel, whereas segmentation summarizes channel responses at each spatial prediction unit. The subsequent class-specific prototype construction, disagreement margin, and reliability weighting are unchanged across the two tasks.

3.2 Confidence-Gated Activation-Distribution Prototypes

Because the observed targets $\tilde{y}_i$ may be corrupted, prototypes are constructed from high-confidence model predictions rather than observed labels. For unit u_i, let $a_i = \mathrm{softmax}(g_\theta(u_i))$, $\hat{y}_i = \arg\max_k a_i(k)$, and $h_i = \max_k a_i(k)$. Given threshold δ, define $\mathcal{B}_k = \{i \in \mathcal{B} : h_i \geq \delta, \hat{y}_i = k\}$. Each class prototype p_k^ℓ is initialized from the mean descriptor of its first non-empty $\mathcal{B}_k$ and subsequently updated by

$$p_k^\ell \leftarrow \beta p_k^\ell + (1-\beta)\frac{1}{|\mathcal{B}_k|}\sum_{i\in\mathcal{B}_k} z_i^\ell.$$

Uninitialized prototypes are excluded from distance comparisons; if the observed-label prototype is unavailable, the corresponding unit retains full task-loss weight. Thus, p_k^ℓ captures class-specific activation-distribution behavior inferred from confident predictions. Confidence gating reduces, but does not assume the absence of, prototype contamination; its effect is analyzed in Sec. 3.6.

3.3 Prototype-Based Supervision Reliability

Reliability is estimated by comparing each descriptor with the class-specific activation-distribution prototypes. We normalize descriptors and prototypes as $\bar{z}_i^\ell = \mathrm{norm}(z_i^\ell)$ and $\bar{p}_k^\ell = \mathrm{norm}(p_k^\ell)$, and define $d_{i,k}^\ell = \|\bar{z}_i^\ell - \bar{p}_k^\ell\|_2$. To encourage discriminative prototype structure, distances are converted to prototype logits $s_{i,k}^\ell = -d_{i,k}^\ell/\tau$. For $s_i^\ell = [s_{i,1}^\ell, \ldots, s_{i,K}^\ell]$, the prototype-alignment loss is $\mathcal{L}_{\mathrm{proto}} = \frac{1}{|\mathcal{B}_{\mathrm{conf}}|}\sum_{i\in\mathcal{B}_{\mathrm{conf}}} \mathrm{CE}(s_i^\ell, \hat{y}_i)$. The central reliability signal is the relative agreement between the observed label and the closest competing class. We define $d_i^+ = d_{i,\tilde{y}_i}^\ell$ and $d_i^- = \min_{c\neq\tilde{y}_i} d_{i,c}^\ell$, yielding the activation-distribution disagreement margin $m_i = d_i^+ - d_i^-$. A negative margin indicates stronger consistency with the observed label, whereas $m_i > 0$ indicates that the activation distribution is closer to a competing class. Larger positive margins therefore provide stronger evidence that the observed supervision is unreliable.

3.4 Reliability-Weighted Training

The disagreement margin is mapped to a continuous suspiciousness score and reliability weight:

$$\rho_i = \mathbb{I}[h_i \geq \delta]\left[2\,\mathrm{sigm}\left(\frac{m_i}{\tau_w}\right) - 1\right]_+, \quad w_i = 1 - (1-\alpha_{\min})\rho_i. \tag{2}$$

Here, $\mathrm{sigm}(\cdot)$ is the sigmoid function, $[x]_+ = \max(x, 0)$, and $\alpha_{\min} \leq w_i \leq 1$ controls the maximum attenuation. Reliable units retain full task-loss weight, while increasingly inconsistent units

are smoothly downweighted rather than discarded. The final objective is

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} w_i \, \mathcal{L}_{\text{task}} \left(g_\theta(u_i), \tilde{y}_i \right) + \lambda \mathcal{L}_{\text{proto}}. \tag{3}$$

The adaptive weighting itself is deliberately simple; the central contribution is the reliability signal derived from observed-label versus competing-class agreement in class-specific activation-distribution space.

3.5 When Does Activation-Distribution Disagreement Identify Corruption?

We analyze sufficient conditions under which the activation-distribution disagreement margin separates clean from corrupted supervision. Omitting the layer index for clarity, let z_i denote the normalized descriptor of unit u_i, p_k the learned prototype of class k, ν_k its population counterpart, y_i the clean class, and $\tilde{y}_i$ the observed label. We assume that descriptors concentrate around their class-specific population prototypes, learned prototypes remain close to these population counterparts, and distinct classes are sufficiently separated:

$$\Pr(\|z_i - \nu_{y_i}\|_2 \le r) \ge 1 - \eta_z, \tag{4}$$

$$\Pr\left(\max_k \|p_k - \nu_k\|_2 \le \epsilon_p \right) \ge 1 - \eta_p, \tag{5}$$

and

$$\Delta = \min_{k \neq k'} \|\nu_k - \nu_{k'}\|_2 > 2(r + \epsilon_p). \tag{6}$$

Theorem 1: Activation-distribution margin separation. Let

$$m_i = \|z_i - p_{\tilde{y}_i}\|_2 - \min_{k \neq \tilde{y}_i} \|z_i - p_k\|_2. \tag{7}$$

Under the conditions above, with probability at least $1 - \eta_z - \eta_p$,

$$\tilde{y}_i = y_i \Longrightarrow m_i \le -\zeta, \qquad \tilde{y}_i \neq y_i \Longrightarrow m_i \ge \zeta, \tag{8}$$

where $\zeta = \Delta - 2(r + \epsilon_p) > 0$. Thus, sufficiently separated class-specific activation distributions induce opposite margin signs for clean and corrupted supervision. The complete proof is provided in the appendix. Figure 4 empirically supports the analysis. The clean-corrupted margin gap grows from $\Delta_m = 0.068$ at prototype construction to 0.639 at convergence, with final mean margins of -0.325 for clean and 0.314 for corrupted supervision. Thus, separation emerges as the prototypes stabilize. Negative margins preserve full task-loss weight, whereas positive margins reduce it. Reliability weights are detached from the computational graph; the corresponding corollary and proof are given in the appendix.

3.6 Prototype Stability under Imperfect Confidence Gating

Theorem 1 depends on the prototype error ϵ_p. Since high-confidence predictions may still be incorrect, we additionally characterize prototype stability under contaminated updates. Let $\widehat{p}_{k,t}$ denote the mini-batch prototype estimate for class k and $\xi_{k,t}$ the fraction of admitted units whose clean class is not k. Assume $\|\widehat{p}_{k,t} - \nu_k\|_2 \le \epsilon_b + \xi_{k,t} D_{\max}$, where ϵ_b is finite-batch estimation error and $D_{\max}$ bounds the effect of an incorrect assignment.

Proposition 2: Finite-time prototype-error bound. If $\xi_{k,t} \le \xi_{\max}$ and

$$p_{k,t} = \beta p_{k,t-1} + (1 - \beta) \widehat{p}_{k,t}, \tag{9}$$

then

$$\|p_{k,t} - \nu_k\|_2 \le \beta^t \|p_{k,0} - \nu_k\|_2 + (1 - \beta^t) \left(\epsilon_b + \xi_{\max} D_{\max} \right). \tag{10}$$

Consequently, Theorem 1 remains applicable whenever

$$\Delta > 2 \left(r + \epsilon_p(t) \right), \tag{11}$$

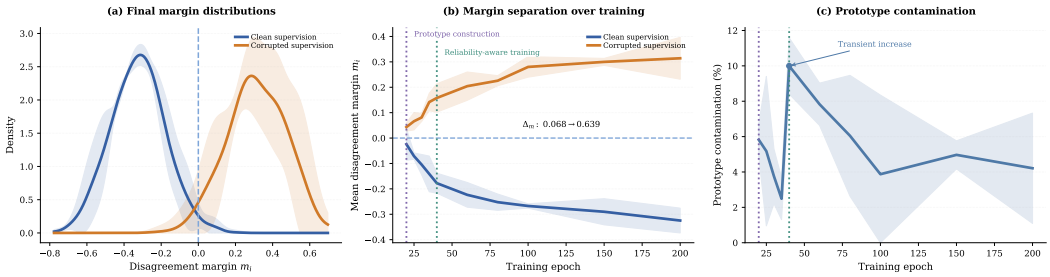

Figure 4: **Empirical validation on CIFAR-10 with 40% asymmetric label noise.** (a) Final disagreement-margin distributions for clean and corrupted supervision. (b) Mean margin evolution across training with 95% confidence intervals over five seeds; dotted lines mark prototype construction and reliability-aware training. (c) Contamination rate of confidence-gated prototype updates. The observed margin separation and bounded contamination are consistent with Theorem 1 and Proposition 2, respectively.

where the resulting prototype error converges to a neighborhood controlled by $\epsilon_b + \xi_{\max} D_{\max}$. The full derivation is provided in the appendix. Empirically, Fig. 4(c) shows that contamination remains bounded despite imperfect confidence gating. After the transition to reliability-aware training, contamination decreases from 10.00% to 4.21% by the final epoch, while the clean-corrupted margin gap continues to increase. This supports the proposition's central implication that prototype assignments need not be perfectly clean for the reliability signal to remain stable.

Training Procedure Training proceeds in three stages. First, the network is trained using only the task loss until T_{warm}. Next, high-confidence predictions update the class-specific activation-distribution prototypes by exponential moving average, without reliability weighting. From $T_{\text{proto}} \geq T_{\text{warm}}$, disagreement margins determine the reliability weights, training uses the weighted task loss together with the prototype-alignment loss. Algorithm 1 summarizes the procedure.

Algorithm 1 Training with Activation-Distribution Reliability

Require: $\mathcal{D} = \{(u_i, \tilde{y}_i)\}$, network f_θ, layer ℓ, thresholds $\delta, T_{\text{warm}}, T_{\text{proto}}$, and $\alpha_{\min}, \lambda, \beta, \tau, \tau_w$

```
Mark prototypes {p_k^ℓ}_{k=1}^K as uninitialized
for epoch = 1 to T do
    for mini-batch B do
        Compute logits, descriptors z_i^ℓ, predictions ŷ_i, and confidences h_i
        if epoch ≥ T_warm then
            Form B_k = {i : h_i ≥ δ, ŷ_i = k} and initialize p_k^ℓ from its first non-empty B_k; thereafter update by EMA
        end if
        if epoch ≥ T_proto then
            Compute prototype distances using initialized prototypes, L_proto, and m_i = d_i^+ − d_i^−
            Compute ρ_i and w_i = 1 − (1 − α_min)ρ_i; set w_i = 1 if the required prototypes are unavailable
        else
            Set w_i = 1 and L_proto = 0
        end if
        Update θ using Eq. 3
    end for
end for
```

4 Experimental Protocol and Results

We evaluate activation-distribution reliability across image classification and semantic segmentation under image-level and pixel-level label noise. For classification, we use CIFAR-10/100 and UC

Merced Krizhevsky et al. (2009); Yang & Newsam (2010) with 20% and 40% symmetric and asymmetric corruption, reporting clean-test accuracy. For segmentation, we use PASCAL VOC 2012 Everingham et al. (2010) under pseudo-label noise (PL), random dilation and erosion (RDE), and similar-class perturbation (SCP), as well as a separate $1/2$ labeled-data protocol. DeepLabV3+ and a ViT-L/8 encoder-decoder Chen et al. (2018); Dosovitskiy et al. (2020) are evaluated using mIoU. Full training and corruption details are provided in the appendix. Table 1 directly tests whether the proposed activation-distribution signal provides information beyond conventional reliability criteria. Under identical training conditions, the full method outperforms confidence-only weighting, prediction-label disagreement, point-wise feature prototypes, Mahalanobis reliability, small-loss selection, and the variant without prototype alignment. At 40% asymmetric CIFAR-10 noise, it reaches 93.51% accuracy, compared with 85.67% for feature prototypes and 83.92% for confidence-only weighting. Under 40% segmentation noise, it achieves 68.17, 69.24, and 70.31 mIoU for PL, RDE, and SCP, respectively, supporting the central claim that class-specific activation-distribution consistency provides a stronger supervision-reliability signal than conventional point-wise alternatives.

Table 1: **Controlled ablation of supervision-reliability signals.** Classification results use CIFAR-10 and report test accuracy (%). Segmentation results use DeepLabV3+ on PASCAL VOC 2012 and report mIoU (%). All variants within each task use identical training settings and corruption realizations; only the reliability mechanism is changed.

Component / Setting	Full Model	Confidence Only	Prediction-Label Disagreement	Feature Prototypes	Mahalanobis Reliability	Small-Loss Selection	w/o Prototype Alignment
Reliability formulation							
Activation-distribution descriptor	✓	✗	✗	✗	✗	✗	✓
Point-wise feature representation	✗	✗	✗	✓	✓	✗	✗
Confidence gating	✓	✓	✓	✓	✓	✗	✓
Prediction-label disagreement	✗	✗	✓	✗	✗	✗	✗
Prototype-based reliability	✓	✗	✗	✓	✗	✗	✓
Prototype-alignment loss	✓	✗	✗	✓	✗	✗	✗
Small-loss selection	✗	✗	✗	✗	✗	✓	✗
Reliability weighting	✓	✓	✓	✓	✓	✗	✓
Classification performance: CIFAR-10 accuracy (%)							
Clean	**95.83**	94.72	94.38	95.06	94.81	94.17	95.29
Symmetric 20%	**96.17**	89.74	89.21	90.36	89.92	88.67	90.58
Symmetric 40%	**92.64**	87.83	87.16	89.28	88.74	86.31	89.73
Asymmetric 20%	**94.05**	91.27	90.84	92.18	91.69	89.76	92.42
Asymmetric 40%	**93.51**	83.92	82.85	85.67	84.96	81.73	86.14
Semantic segmentation performance: PASCAL VOC 2012 mIoU (%)							
Clean	**81.34**	80.27	79.86	80.68	80.14	79.23	80.91
PL 20%	**75.42**	72.84	71.96	74.18	73.51	70.87	74.63
PL 40%	**68.17**	63.92	62.71	66.24	65.38	60.86	67.03
RDE 20%	**76.14**	73.63	72.87	74.92	74.26	71.48	75.31
RDE 40%	**69.24**	64.81	63.69	67.17	66.43	61.72	68.06
SCP 20%	**76.92**	74.37	73.58	75.46	74.82	72.16	76.08
SCP 40%	**70.31**	66.18	64.93	68.24	67.56	63.27	69.13

We next evaluate activation-distribution reliability under structured pixel-level supervision on PASCAL VOC 2012 using pseudo-label noise (PL), random dilation and erosion (RDE), and similar-class perturbation (SCP). With DeepLabV3+, ours outperforms cross-entropy, Bootstrapping, GCE, and SCE across all corruption types and noise rates. At 40% corruption, it achieves 68.17, 69.24, and 70.31 mIoU under PL, RDE, and SCP, respectively, about 5.8-6.0 mIoU above SCE. Using the same reliability mechanism with a ViT-L/8 encoder-decoder further improves performance to 70.63, 71.58, and 72.46 mIoU.

Table 3 compares the proposed method with representative noisy-label approaches across CIFAR-10/100, and UC Merced. The proposed activation-distribution reliability signal remains consistently effective across symmetric and asymmetric corruption, with the largest gains under severe asymmetric noise. At 40% asymmetric corruption, our method achieves 93.51% accuracy on CIFAR-10, slightly exceeding DivideMix at 93.4%, and reaches 72.15% on CIFAR-100 compared with 71.30% for TPCR. These results support the central premise that class-specific activation-distribution consistency provides reliability information beyond conventional loss-, confidence-, and feature-based criteria. Additional results under the matched $1/2$ labeled-data segmentation protocol and on the real-world noisy-label Clothing1M benchmark are provided in Appendices A.3 and A.4, respectively.

Table 2: Controlled comparison on PASCAL VOC 2012 under PL, RDE, and SCP. The corruption rate η denotes the fraction of corrupted training pixels; validation and test masks remain clean. All values are mIoU (%).

Method	Noise Type	Backbone	Noise Rate η				
			0%	10%	20%	30%	40%
Cross-Entropy	PL	DeepLabV3+	79.81	74.92	69.43	63.82	57.64
Bootstrapping		DeepLabV3+	79.68	75.71	70.82	65.37	59.86
GCE		DeepLabV3+	79.47	76.18	71.64	66.52	61.27
SCE		DeepLabV3+	79.76	76.63	72.29	67.48	62.36
Ours		**DeepLabV3+**	81.34	78.61	75.42	71.94	68.17
Ours		**ViT-L/8**	**82.08**	**79.74**	**77.13**	**74.26**	**70.63**
Cross-Entropy	RDE	DeepLabV3+	79.81	75.48	70.21	64.76	58.53
Bootstrapping		DeepLabV3+	79.68	76.32	71.57	66.23	60.71
GCE		DeepLabV3+	79.47	76.83	72.38	67.41	62.19
SCE		DeepLabV3+	79.76	77.29	73.16	68.34	63.27
Ours		**DeepLabV3+**	81.34	79.18	76.14	72.86	69.24
Ours		**ViT-L/8**	**82.08**	**80.26**	**77.84**	**75.13**	**71.58**
Cross-Entropy	SCP	DeepLabV3+	79.81	76.07	71.02	65.83	59.72
Bootstrapping		DeepLabV3+	79.68	76.93	72.45	67.19	61.87
GCE		DeepLabV3+	79.47	77.46	73.24	68.31	63.38
SCE		DeepLabV3+	79.76	77.92	74.08	69.26	64.41
Ours		**DeepLabV3+**	81.34	79.73	76.92	73.78	70.31
Ours		**ViT-L/8**	**82.08**	**80.81**	**78.61**	**76.04**	**72.46**

Table 3: Classification performance under symmetric and asymmetric training-label corruption. All values denote clean test accuracy (%) and are reported as mean ± STD when available. A dash indicates that a directly comparable result is not reported.

Method	Dataset	Symmetric Noise		Asymmetric Noise	
		Noise Rate η		Noise Rate η	
		0.2	0.4	0.2	0.4
Co-teaching Han et al. (2018)	CIFAR-10	82.32±0.07	–	–	–
	CIFAR-100	54.23±0.08	–	–	–
DivideMix Li et al. (2020)	CIFAR-10	96.1	–	–	93.4
	CIFAR-100	77.3	–	–	–
SCE / SL Wang et al. (2019)	CIFAR-10	87.63±0.06	85.34±0.07	88.24±0.05	80.64±0.10
	CIFAR-100	60.01±0.19	53.69±0.07	65.58±0.06	63.10±0.13
GCE / Trunc. L_q Zhang & Sabuncu (2018)	CIFAR-10	89.70±0.11	87.62±0.26	89.45±0.29	82.28±0.67
	CIFAR-100	67.61±0.18	62.64±0.33	66.59±0.23	47.66±0.69
DAL Li et al. (2023b)	CIFAR-10	90.95±0.09	88.73±0.10	90.08±0.10	73.34±0.18
	CIFAR-100	69.00±0.09	64.80±0.17	66.60±0.04	45.36±0.42
CICW / CICW-M Kumar & Amid (2021)	CIFAR-10	91.34±0.16	89.95±0.15	92.44±0.48	86.41±0.51
	CIFAR-100	68.23±0.23	63.64±0.45	70.08±0.40	65.93±0.27
TPCR Sun et al. (2024)	CIFAR-10	93.20±0.20	92.70±0.10	93.30±0.30	91.00±0.60
	CIFAR-100	74.80±0.30	68.70±0.50	77.30±0.20	71.30±0.60
JAL-CE Wang et al. (2025)	CIFAR-100	68.25±0.39	64.11±0.55	67.90±0.59	56.17±0.32
JAL-FL Wang et al. (2025)	CIFAR-100	68.33±0.34	64.55±0.61	67.63±0.50	56.26±0.05
Ours	CIFAR-10	96.17±0.57	92.64±0.39	94.05±0.41	93.51±0.12
	CIFAR-100	77.79±0.21	69.67±0.91	78.97±0.32	72.15±0.42
	UC Merced	97.87±0.12	97.11±0.37	97.14±0.21	96.69±0.51

5 Conclusion and Future Work

We introduced a supervision-reliability framework based on the distributional behavior of intermediate activations. The method constructs class-specific activation-distribution prototypes and measures reliability through disagreement between the observed-label prototype and the nearest competing prototype, providing a signal complementary to loss, confidence, and point-wise feature criteria. Experiments across classification and segmentation, multiple noise types, corruption levels, and architectures show consistent robustness gains, while the analysis characterizes conditions under which the proposed margin separates clean and corrupted supervision. Overall, the results demonstrate that class-specific activation statistics provide useful reliability information beyond model outputs alone. **Future work** will explore richer descriptors, multi-layer aggregation, and broader structured-prediction settings.

AI USE STATEMENT

In this work, generative AI tools were used for text polishing and assistance with code writing. All AI-assisted text was reviewed and edited by the authors, and all generated or modified code was manually inspected and tested for correctness. The authors take full responsibility for the final content, implementation, results, and claims presented in this work.

ETHICS STATEMENT

This work studies robustness to noisy supervision and does not involve human subjects or the collection of sensitive personal data. All experiments use standard publicly available benchmarks and synthetic or existing annotation corruption protocols. The proposed method is intended to improve the reliability of learning systems trained with imperfect annotations and does not introduce new data privacy or security risks beyond those associated with the underlying models and datasets.

REPRODUCIBILITY STATEMENT

We provide detailed information to support reproducibility of our results. The methodology specifies the activation-distribution descriptor, prototype construction, reliability estimation, and training objective, while Algorithm 1 summarizes the complete training procedure. The appendix reports dataset preparation, corruption procedures, model architectures, optimization settings, hyperparameters, and training schedules used in all experiments. It also provides complete proofs for the theoretical results and additional experimental analyses. All reported comparisons use the same data splits and corruption realizations within each evaluation protocol.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2613–2622, 2021.

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

Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2): 303–338, 2010.

Geoff French, Samuli Laine, Timo Aila, Michal Mackiewicz, and Graham Finlayson. Semi-supervised semantic segmentation needs strong, varied perturbations. *arXiv preprint arXiv:1906.01916*, 2019.

Mohammed Talal Ghazal. Noise-aware prototype-assisted contrastive learning for learning with noisy labels. *Expert Systems with Applications*, pp. 133007, 2026.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31, 2018.

Zhizhong Huang, Junping Zhang, and Hongming Shan. Twin contrastive learning with noisy labels. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11661–11670. IEEE, 2023.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pp. 448–456. pmlr, 2015.

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. In *International conference on machine learning*, pp. 2304–2313. PMLR, 2018.

Zhanghan Ke, Di Qiu, Kaican Li, Qiong Yan, and Rynson WH Lau. Guided collaborative training for pixel-wise semi-supervised learning. In *European conference on computer vision*, pp. 429–445. Springer, 2020.

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

Abhishek Kumar and Ehsan Amid. Constrained instance and class reweighting for robust learning under label noise. *arXiv preprint arXiv:2111.05428*, 2021.

Xin Lai, Zhuotao Tian, Li Jiang, Shu Liu, Hengshuang Zhao, Liwei Wang, and Jiaya Jia. Semi-supervised semantic segmentation with directional context-aware consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1205–1214, 2021.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020.

Peixia Li, Pulak Purkait, Thalaiyasingam Ajanthan, Majid Abdolshah, Ravi Garg, Hisham Husain, Chenchen Xu, Stephen Gould, Wanli Ouyang, and Anton Van Den Hengel. Semi-supervised semantic segmentation under label noise via diverse learning groups. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1229–1238, 2023a.

Xiu-Chuan Li, Xiaobo Xia, Fei Zhu, Tongliang Liu, Xu-Yao Zhang, and Cheng-Lin Liu. Dynamics-aware loss for learning with label noise. *Pattern Recognition*, 144:109835, 2023b.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33:20331–20342, 2020.

Yuyuan Liu, Yu Tian, Yuanhong Chen, Fengbei Liu, Vasileios Belagiannis, and Gustavo Carneiro. Perturbed and strict mean teachers for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4258–4267, 2022.

Robert Mendel, Luis Antonio de Souza Jr, David Rauber, Joao Paulo Papa, and Christoph Palm. Semi-supervised segmentation based on error-correcting supervision. In *European conference on computer vision*, pp. 141–157. Springer, 2020.

Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12674–12684, 2020.

Weiran Pan, Wei Wei, and Wenfeng Xie. Debiased sample selection for learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32047–32057, 2026.

Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1944–1952, 2017.

Scott Reed, Honglak Lee, Dragomir Anguelov, Christian Szegedy, Dumitru Erhan, and Andrew Rabinovich. Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*, 2014.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 761–769, 2016.

Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Hwanjun Song, Minseok Kim, and Jae-Gil Lee. Selfie: Refurbishing unclean samples for robust deep learning. In *International conference on machine learning*, pp. 5907–5915. PMLR, 2019.

Xinkai Sun, Sanguo Zhang, and Shuangge Ma. Prediction consistency regularization for learning with noise labels based on contrastive clustering. *Entropy*, 26(4):308, 2024.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in neural information processing systems*, 30, 2017.

Jialiang Wang, Xianming Liu, Xiong Zhou, Gangfeng Hu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. Joint asymmetric loss for learning with noisy labels. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1947–1956. IEEE, 2025.

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 322–330, 2019.

Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 13726–13735, 2020.

Qi Wei, Haoliang Sun, Xiankai Lu, and Yilong Yin. Self-filtering: A noise-aware sample selection for label noise with confidence penalization. In *European Conference on Computer Vision*, pp. 516–532. Springer, 2022.

Qi Wei, Lei Feng, Haoliang Sun, Ren Wang, Chenhui Guo, and Yilong Yin. Fine-grained classification with noisy labels. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11651–11660. IEEE, 2023.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pp. 499–515. Springer, 2016.

Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Qiangqiang Xia, Feifei Lee, Lin Xie, Shuai Yang, Qing Bao, and Qiu Chen. Cpsl: A semi-supervised framework with class prototype-based modeling for combating noisy labels. *Pattern Recognition*, pp. 112318, 2025.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pp. 270–279, 2010.

Kun Yi and Jianxin Wu. Probabilistic end-to-end noise correction for learning with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7017–7025, 2019.

Rumeng Yi, Yaping Huang, Qingji Guan, Mengyang Pu, and Runsheng Zhang. Learning from pixel-level label noise: A new perspective for semi-supervised semantic segmentation. *IEEE Transactions on Image Processing*, 31:623–635, 2021.

Yue Zhang, Yiyi Chen, Chaowei Fang, Qian Wang, Jiayi Wu, and Jingmin Xin. Learning from open-set noisy labels based on multi-prototype modeling. *Pattern Recognition*, 157:110902, 2025.

Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.

A Appendix

A.1 Additional Qualitative Results

Figure 5 presents additional classification examples from UC Merced under corrupted training labels. Each example shows the underlying clean class, the baseline prediction, and the prediction produced by the proposed method. The baseline frequently predicts an incorrect class under corrupted supervision, whereas the proposed activation-distribution reliability mechanism more consistently recovers the underlying clean class.

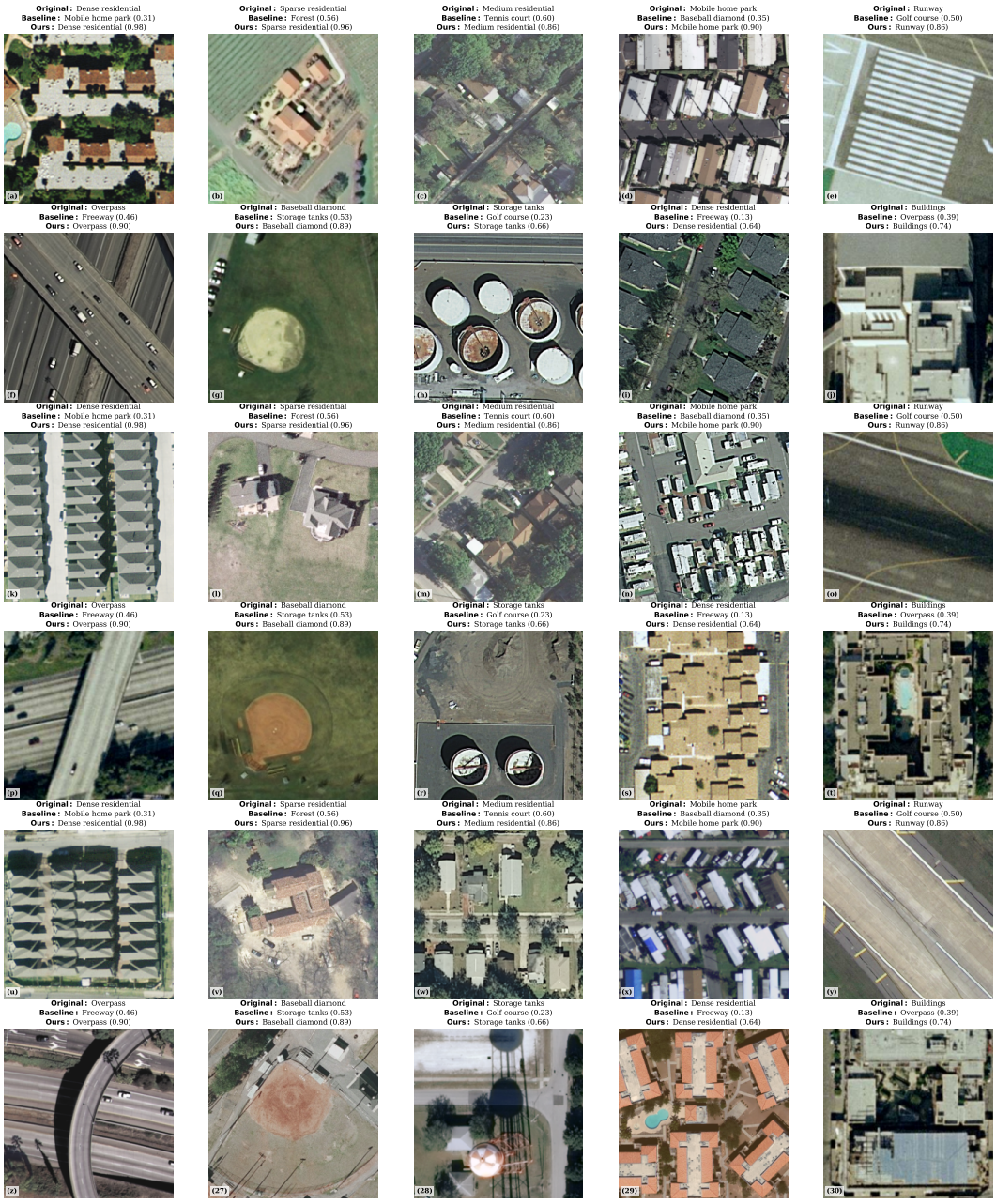

Figure 5: **Additional qualitative classification results on UC Merced.** For each image, we report the clean class, baseline prediction, and prediction of the proposed method, together with their confidence scores. These examples illustrate cases in which activation-distribution reliability enables recovery of the underlying clean target despite corrupted training supervision.

Figure 6 provides additional semantic segmentation comparisons under noisy pixel-level supervision. Green regions denote correctly predicted foreground pixels, while red regions denote false-positive predictions outside the clean ground-truth mask. Compared with the baseline, the proposed method produces predictions that more closely follow the target regions and substantially reduces leakage into the background.

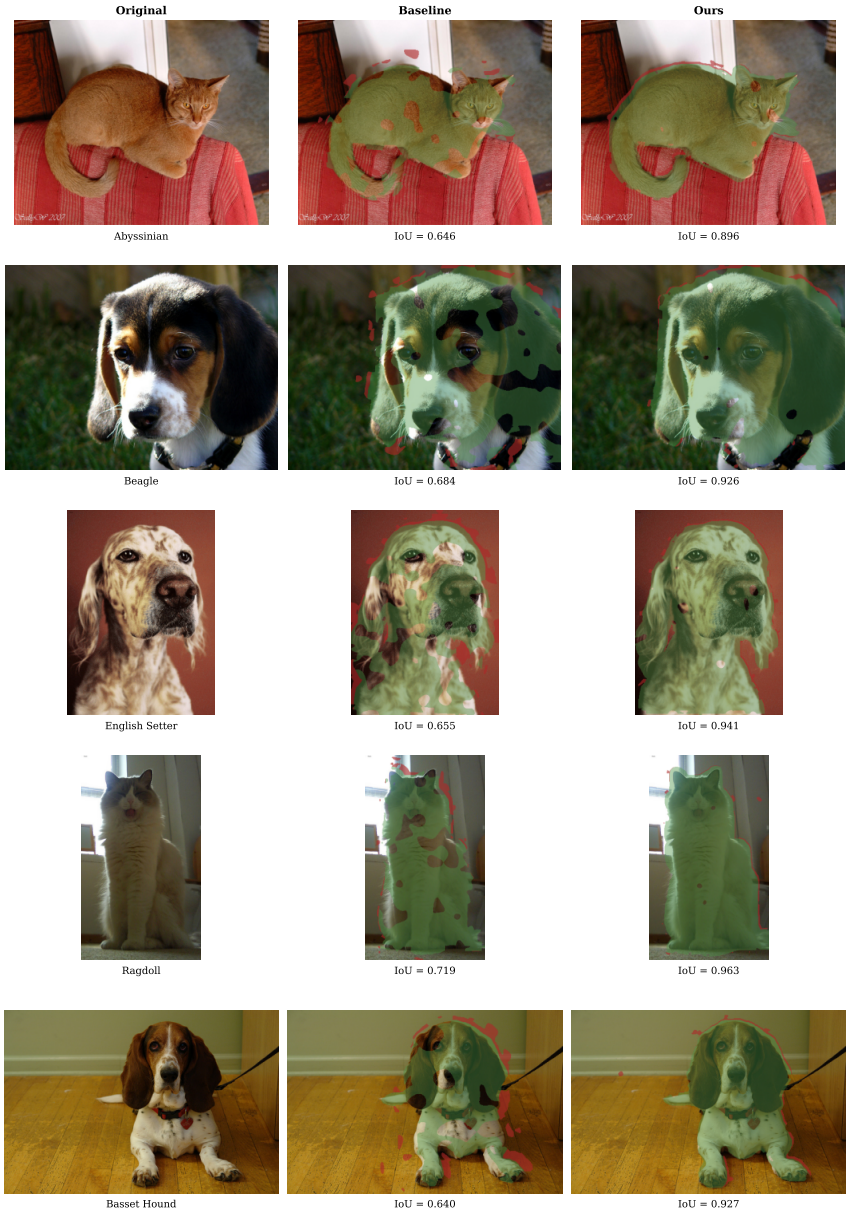

Figure 6: **Additional qualitative semantic segmentation results.** From left to right, each row shows the input image and clean ground-truth mask, the baseline prediction, and the prediction of the proposed method. Green denotes correctly predicted foreground pixels and red denotes false-positive regions. The activation-distribution reliability mechanism produces cleaner predictions and fewer background false positives under noisy supervision.

A.2 Disentangling Reliability Weighting and Prototype Alignment

The full framework combines two components: the proposed activation-distribution reliability weighting and the prototype-alignment objective $\mathcal{L}_{\text{proto}}$. To determine whether the observed ro-

bustness gains arise primarily from the auxiliary prototype objective or from the proposed reliability signal, we independently ablate these two components under the CIFAR-10 setting with 40% asymmetric label noise.

We compare five variants: standard cross-entropy training, activation-distribution reliability weighting without prototype alignment, prototype alignment without reliability weighting, the complete framework, and a point-wise feature-prototype variant using the same alignment and reliability mechanism. For the prototype-alignment-only variant, all task-loss weights are fixed to $w_i = 1$, while $\mathcal{L}_{\text{proto}}$ is retained. Conversely, the reliability-only variant uses the activation-distribution disagreement weights while setting $\lambda = 0$.

Table 4: Disentangling the effects of activation-distribution reliability weighting and prototype alignment on CIFAR-10 with 40% asymmetric label noise. All variants use identical training settings and corruption realizations.

Variant	Reliability Weighting	$\mathcal{L}_{\text{proto}}$	Accuracy (%)
Cross-Entropy Baseline	×	×	81.68
Distribution Reliability Only	✓	×	86.14
Prototype Alignment Only	×	✓	85.23
Full Model	✓	✓	**93.51**
Feature Prototype + Alignment	✓	✓	85.67

As shown in Table 4, prototype alignment alone achieves **85.23%**, while activation-distribution reliability weighting alone reaches 86.14%. Combining both components increases performance to 93.51%. These results indicate that prototype alignment alone does not account for the improvement of the full framework; instead, the proposed activation-distribution reliability signal provides complementary information that becomes substantially more effective when coupled with prototype alignment. The substantially lower performance of the point-wise feature-prototype variant further suggests that the gain is not explained by prototype regularization alone, but depends on modeling class-specific activation-distribution behavior.

A.3 Comparison under the $1/2$ Labeled-Data Protocol

Table 5 provides an additional comparison under the $1/2$ labeled-data protocol used in prior semi-supervised segmentation work. Following DLG Li et al. (2023a), the competing methods use DeepLabV3+, allowing direct comparison with our DeepLabV3+ implementation.

Under clean supervision, our DeepLabV3+ model reaches 79.97 mIoU compared with 79.54 for DLG. Under 9% corrupted pixels, the corresponding results for our method are 74.47, 74.67, and 76.06 mIoU under PL, RDE, and SCP, respectively. These results are comparable to or slightly above DLG, which obtains 74.44, 74.47, and 75.95 under the same settings.

The ViT-L/8 instantiation further reaches 75.12, 76.71, and 77.90 mIoU under PL, RDE, and SCP. These results should not be interpreted as a backbone-matched comparison with the DeepLabV3+-based competitors; rather, they show that the same activation-distribution reliability formulation can be applied to a substantially different feature representation.

A.4 Real-World Label Noise on Clothing1M

Table 6 evaluates the proposed method on Clothing1M, which contains naturally occurring label noise. All methods use a ResNet-50 backbone. Our method achieves 76.91% test accuracy, outperforming the strongest directly comparable baseline by 1.60 percentage points.

A.5 Direct Evaluation of Supervision Reliability

We directly evaluate whether the proposed activation-distribution disagreement margin identifies corrupted supervision. We formulate corruption identification as a binary detection problem, where corrupted labels are treated as the positive class. We compare our disagreement margin against commonly available sample-level reliability signals: prediction confidence, task loss, point-feature prototype distance, and Mahalanobis distance.

Table 5: Comparison on PASCAL VOC 2012 under clean and noisy supervision using the $1/2$ labeled-data ratio. Competing methods use DeepLabV3+. Each noisy setting contains 9% corrupted training pixels, while validation and test masks remain clean. Results are reported as mIoU (%).

Method	Noisy Pixels	Noise Type	mIoU ↑
MT Tarvainen & Valpola (2017)	0%	None	75.41
French French et al. (2019)	0%	None	74.49
CCT Ouali et al. (2020)	0%	None	74.75
GCT Ke et al. (2020)	0%	None	75.20
ECS Mendel et al. (2020)	0%	None	72.89
CPS Chen et al. (2021)	0%	None	76.15
CAC Lai et al. (2021)	0%	None	–
PS-MT Liu et al. (2022)	0%	None	77.88
DLG Li et al. (2023a)	0%	None	79.54
Ours (DeepLabV3+)	0%	None	79.97
Ours (ViT-L/8)	0%	None	**80.06**
PS-MT Liu et al. (2022)	9%	PL	66.78
DLG Li et al. (2023a)	9%	PL	74.44
Ours (DeepLabV3+)	9%	PL	74.47
Ours (ViT-L/8)	9%	PL	**75.12**
PS-MT Liu et al. (2022)	9%	RDE	63.46
DLG Li et al. (2023a)	9%	RDE	74.47
Ours (DeepLabV3+)	9%	RDE	74.67
Ours (ViT-L/8)	9%	RDE	**76.71**
PS-MT Liu et al. (2022)	9%	SCP	53.69
DLG Li et al. (2023a)	9%	SCP	75.95
Ours (DeepLabV3+)	9%	SCP	76.06
Ours (ViT-L/8)	9%	SCP	**77.90**

Table 6: **Comparison on Clothing1M.** Results are clean test accuracy (%).

Method	Test Acc. ↑
CE	69.21
DivideMix Li et al. (2020)	74.76
ELR Liu et al. (2020)	74.81
TCL Huang et al. (2023)	74.80
SFT+* Wei et al. (2022)	75.08
DivideMix + SNSCL Wei et al. (2023)	75.31
JAL-CE Wang et al. (2025)	70.31
JAL-FL Wang et al. (2025)	70.11
DSS+ Pan et al. (2026)	75.13
Ours	**76.91**

For confidence, we use $1 - \max_k a_{i,k}$ so that larger values indicate lower reliability. Task loss is the per-unit cross-entropy loss. Feature-prototype and Mahalanobis baselines use distance to the prototype or distribution associated with the observed label. For our method, larger activation-distribution disagreement margins indicate stronger disagreement between the observed-label prototype and the nearest competing class.

Table 7: **Direct evaluation of supervision-reliability signals.** Corrupted supervision is treated as the positive class. Results report AUROC and AUPRC under 40% asymmetric noise on CIFAR-10 and 40% SCP corruption on PASCAL VOC. Higher is better.

	CIFAR-10		PASCAL VOC	
Reliability Signal	AUROC	AUPRC	AUROC	AUPRC
Confidence	0.78	0.72	0.74	0.68
Task Loss	0.82	0.77	0.77	0.71
Feature-Prototype Distance	0.85	0.81	0.81	0.76
Mahalanobis Distance	0.84	0.80	0.80	0.75
Activation-Distribution Margin (Ours)	**0.92**	**0.89**	**0.89**	**0.85**

As shown in Table 7, the proposed activation-distribution disagreement margin provides the strongest separation between clean and corrupted supervision across both image-level and pixel-level noise. In particular, it improves AUROC from **0.85** to **0.92** on CIFAR-10 and from **0.81** to **0.89** on PASCAL VOC relative to the strongest competing reliability signal. Similar improvements are observed in AUPRC.

These results provide direct evidence that the benefit of the proposed method does not arise merely from applying another reweighting strategy. Instead, class-specific activation-distribution prototypes provide a more informative signal of supervision reliability than prediction confidence, instantaneous loss, conventional point-feature prototypes, or Mahalanobis distance. This directly supports the central motivation of our approach.

A.6 Implementation and Reproducibility Details

A.6.1 Activation-Distribution Descriptor Construction

The activation-distribution principle is instantiated according to the granularity of the supervised unit. For image classification, the supervised unit is an entire image, whereas for semantic segmentation the supervised unit corresponds to a spatial prediction location.

Image classification. Activation-distribution descriptors are extracted from the output of the third feature-extraction layer of the backbone. Given an activation tensor

$$A_i^\ell \in \mathbb{R}^{C_\ell \times H_\ell \times W_\ell}, \tag{12}$$

the near-zero activation ratio, mean, standard deviation, skewness, and kurtosis are computed independently for each channel over its $H_\ell W_\ell$ spatial entries. The resulting channel-level statistics are concatenated to form

$$z_i^\ell \in \mathbb{R}^{5C_\ell}. \tag{13}$$

Thus, each image is represented by a descriptor that summarizes the distributional behavior of all channels in the selected intermediate representation.

Semantic segmentation. For dense prediction, descriptors are extracted from the spatial latent representation immediately before the final prediction head. Each spatial location (h, w) is treated as an individual supervised unit. At that location, the distributional statistics are computed over the corresponding C_ℓ-dimensional activation vector, producing a five-dimensional local activation-distribution descriptor.

This dense formulation preserves the same underlying principle as the classification setting: supervision reliability is estimated from statistical activation behavior rather than directly from task loss, prediction confidence, or the raw point-wise feature vector.

The supervision mask is resized to the latent feature resolution using nearest-neighbor interpolation so that discrete class labels are preserved. Reliability values computed at the latent resolution are subsequently resized to the output-logit resolution using bilinear interpolation.

Activations satisfying

$$|A_{i,j}^\ell| < 10^{-5} \tag{14}$$

are treated as near zero. Only one feature source is used for each task, and therefore no multi-layer aggregation is required. Descriptors and prototypes are independently normalized to unit ℓ_2 norm before prototype-distance computation.

Pixel-Level Prototype Assignment and Reliability Weighting. For semantic segmentation, each latent spatial location contributes independently to reliability estimation. A location contributes to the activation-distribution prototype of class k only when its predicted class is k and its prediction confidence exceeds the threshold δ.

For each valid location, the distance to the prototype corresponding to its observed label and the distance to its nearest competing prototype are computed. Their difference defines the activation-distribution disagreement margin, which produces one reliability weight per latent spatial location.

No region-level or image-level aggregation is applied. Boundary locations are treated identically to other valid locations, without additional boundary weighting or exclusion. Ignore-label locations, when present, are excluded from both prototype updates and task-loss computation.

Only the final latent feature resolution is used for dense descriptor construction. The resulting reliability map is bilinearly upsampled to the output-logit resolution and applied element-wise to the pixel-level task loss.

A.6.2 Training Configuration

CIFAR-10 and CIFAR-100. We use a ResNet-34 initialized from scratch and train for 200 epochs using SGD with momentum 0.9, an initial learning rate of 0.1, weight decay 5×10^{-4}, and batch size 128. The learning rate follows a cosine decay schedule. Data augmentation consists of random cropping with four pixels of padding and random horizontal flipping.

Clothing1M. For Clothing1M, we follow the standard training protocol used in prior noisy-label learning studies. We use an ImageNet-pretrained ResNet-50 and the standard Clothing1M train, validation, and test splits. Training, augmentation, optimizer, learning-rate schedule, and model-selection settings follow the corresponding baseline protocol to ensure direct comparability. Checkpoints are selected using the clean validation set, and final performance is reported on the clean test set.

UC Merced. We use an ImageNet-pretrained ResNet-34 and resize images to 224×224. The model is trained for 100 epochs using SGD with momentum 0.9, an initial learning rate of 0.01, weight decay 5×10^{-4}, and batch size 32. The learning rate follows a cosine decay schedule. Augmentation consists of random resized cropping, horizontal flipping, and color jitter.

Semantic segmentation. The DeepLabV3+ model uses an ImageNet-pretrained ResNet-101 encoder and is optimized using SGD with momentum 0.9, learning rate 0.01, weight decay 10^{-4}, and batch size 16.

The ViT-L/8 model is initialized from ImageNet-pretrained weights and optimized using AdamW with learning rate 6×10^{-5}, weight decay 10^{-2}, and batch size 8. Both segmentation models are trained for 80 epochs.

Segmentation augmentation consists of random scaling in $[0.5, 2.0]$, random cropping to 513×513, and random horizontal flipping.

A.6.3 Reliability Hyperparameters

The same activation-distribution reliability configuration is used across datasets and corruption settings unless otherwise stated.

Table 8: **Hyperparameters of the activation-distribution reliability mechanism.**

Hyperparameter	Value
Near-zero threshold ϵ	10^{-5}
Confidence threshold δ	0.8
Prototype momentum β	0.9
Prototype temperature τ	0.1
Weighting temperature τ_w	0.5
Minimum reliability weight $\alpha_{\min}$	0.2
Prototype-alignment coefficient λ	0.1
Representation warm-up T_{warm}	$0.1T$
Reliability activation T_{proto}	$0.2T$

During the first $0.1T$ epochs, the network is optimized using only the original task objective. Confidence-gated activation-distribution prototype estimation begins after $0.1T$ epochs. Prototype

alignment and reliability weighting are activated after $0.2T$ epochs, where T denotes the total number of training epochs.

A.6.4 Noise Generation and Evaluation Protocol

Classification noise. For symmetric classification noise, a fraction η of the training samples is selected uniformly at random, and each selected label is replaced by a uniformly sampled incorrect class. For asymmetric noise, we follow the standard class-dependent corruption protocols used in prior noisy-label learning studies. For CIFAR-10, the fixed mappings are truck $\rightarrow$ automobile, bird $\rightarrow$ airplane, deer $\rightarrow$ horse, and cat $\leftrightarrow$ dog. For CIFAR-100, corruption is restricted within each of the 20 superclasses, each containing five fine-grained classes, by cyclically mapping each class to another class within the same superclass. For UC Merced, we follow the established remote-sensing asymmetric-noise protocol, in which labels are mapped between semantically or visually related scene categories. For all settings, the corresponding mapping is applied independently to eligible samples with probability η. Corrupted labels are generated once before training and remain fixed throughout optimization and across compared methods.

Segmentation noise. For semantic segmentation, we follow the noisy-supervision protocol based on pseudo-label noise (PL), random dilation and erosion (RDE), and similar-class perturbation (SCP). All corrupted masks are generated before training, while validation and test annotations remain clean.

For **pseudo-label noise (PL)**, noisy masks are obtained from SEAM predictions. Clean training annotations are replaced by the corresponding pseudo-label masks according to the target corruption setting.

For **random dilation and erosion (RDE)**, annotation noise is introduced by randomly applying morphological dilation or erosion to semantic regions, producing spatial boundary and shape perturbations while preserving the underlying semantic category.

For **similar-class perturbation (SCP)**, we follow the class-similarity procedure defined from SEAM predictions. For each ground-truth class, SEAM probability vectors over pixels belonging to that class are averaged to construct a class-similarity distribution. The probability of the original class is removed and the remaining probabilities are normalized. Corrupted labels are then sampled from this distribution, causing semantic classes to be preferentially replaced by visually similar classes while preserving the spatial extent of the original region.

For PL, RDE, and SCP, the corruption strength is adjusted to obtain the target pixel-level corruption rate η, defined as the fraction of training pixels whose corrupted annotation differs from the corresponding clean annotation. The same pre-generated corrupted masks are used for all methods within each experimental setting.

Evaluation. Within each controlled comparison, all methods use the same data split, corruption realization, backbone, augmentation, optimizer, and training schedule. Results are reported as the mean and standard deviation over three independent runs using seeds $\{1, 2, 3\}$. Classification checkpoints are selected according to validation accuracy, whereas semantic-segmentation checkpoints are selected according to validation mIoU. Test annotations are used only for final evaluation and never for model selection.

A.7 Proofs and Additional Theoretical Analysis

This subsection provides complete proofs for the theoretical results stated in the main paper, together with additional optimization analysis omitted from the main text for brevity. We use the same notation as in Sec. 3.5. In particular, z_i denotes the normalized activation-distribution descriptor of supervised unit u_i, p_k denotes the learned activation-distribution prototype of class k, and ν_k denotes the corresponding population prototype. The clean class of u_i is denoted by y_i, whereas $\tilde{y}_i$ denotes its observed, potentially corrupted, label.

A.7.1 Proof of Theorem 1: Activation-Distribution Margin Separation

Define the events

$$\mathcal{E}_z = \left\{ \|z_i - \nu_{y_i}\|_2 \le r \right\}, \tag{15}$$

and

$$\mathcal{E}_p = \left\{ \max_{k \in \{1,\dots,K\}} \|p_k - \nu_k\|_2 \le \epsilon_p \right\}. \tag{16}$$

By Assumptions 1 and 2,

$$\Pr(\mathcal{E}_z) \ge 1 - \eta_z, \qquad \Pr(\mathcal{E}_p) \ge 1 - \eta_p. \tag{17}$$

Applying the union bound gives

$$\Pr\left(\mathcal{E}_z \cap \mathcal{E}_p\right) \ge 1 - \eta_z - \eta_p. \tag{18}$$

We condition on the joint event $\mathcal{E}_z \cap \mathcal{E}_p$.

Correctly labeled supervision. Suppose that $\tilde{y}_i = y_i$. By the triangle inequality, the distance to the observed-label prototype satisfies

$$\begin{aligned} d_{i,y_i} &= \|z_i - p_{y_i}\|_2 \\ &\le \|z_i - \nu_{y_i}\|_2 + \|\nu_{y_i} - p_{y_i}\|_2 \\ &\le r + \epsilon_p. \end{aligned} \tag{19}$$

For any competing class $k \ne y_i$, the reverse triangle inequality yields

$$\begin{aligned} d_{i,k} &= \|z_i - p_k\|_2 \\ &\ge \|\nu_{y_i} - \nu_k\|_2 - \|z_i - \nu_{y_i}\|_2 - \|p_k - \nu_k\|_2 \\ &\ge \Delta - r - \epsilon_p. \end{aligned} \tag{20}$$

Consequently,

$$\begin{aligned} m_i &= d_{i,y_i} - \min_{k \ne y_i} d_{i,k} \\ &\le (r + \epsilon_p) - (\Delta - r - \epsilon_p) \\ &= -\zeta, \end{aligned} \tag{21}$$

where

$$\zeta = \Delta - 2(r + \epsilon_p) > 0. \tag{22}$$

Thus, under the stated assumptions, correctly labeled supervision produces a strictly negative activation-distribution disagreement margin.

Corrupted supervision. Now suppose that $\tilde{y}_i \ne y_i$. The distance from the descriptor to the prototype associated with the observed label satisfies

$$\begin{aligned} d_{i,\tilde{y}_i} &= \|z_i - p_{\tilde{y}_i}\|_2 \\ &\ge \|\nu_{y_i} - \nu_{\tilde{y}_i}\|_2 - \|z_i - \nu_{y_i}\|_2 - \|p_{\tilde{y}_i} - \nu_{\tilde{y}_i}\|_2 \\ &\ge \Delta - r - \epsilon_p. \end{aligned} \tag{23}$$

Because the clean class y_i belongs to the set of competing classes,

$$\begin{aligned} \min_{k \neq \tilde{y}_i} d_{i,k} &\leq d_{i,y_i} \\ &\leq r + \epsilon_p. \end{aligned} \tag{24}$$

Therefore,

$$\begin{aligned} m_i &= d_{i,\tilde{y}_i} - \min_{k \neq \tilde{y}_i} d_{i,k} \\ &\geq (\Delta - r - \epsilon_p) - (r + \epsilon_p) \\ &= \zeta. \end{aligned} \tag{25}$$

Hence, conditioned on $\mathcal{E}_z \cap \mathcal{E}_p$, correctly labeled units satisfy $m_i \leq -\zeta$, while corrupted units satisfy $m_i \geq \zeta$. Together with Eq. equation 18, this occurs with probability at least $1 - \eta_z - \eta_p$. □

A.7.2 Proof of Corollary 1: Reliability-Weight Separation

For a confident unit satisfying $h_i \geq \delta$, the suspiciousness score is

$$\rho_i = \left[2 \operatorname{sigm}\left(\frac{m_i}{\tau_w} \right) - 1 \right]_+, \qquad \tau_w > 0. \tag{26}$$

For correctly labeled supervision, Theorem 1 gives

$$m_i \leq -\zeta < 0. \tag{27}$$

Since the sigmoid function is monotonically increasing,

$$2 \operatorname{sigm}\left(\frac{m_i}{\tau_w} \right) - 1 \leq 2 \operatorname{sigm}\left(-\frac{\zeta}{\tau_w} \right) - 1 < 0. \tag{28}$$

The positive-part operator therefore yields

$$\rho_i = 0. \tag{29}$$

Consequently,

$$w_i = 1 - (1 - \alpha_{\min})\rho_i = 1. \tag{30}$$

For corrupted supervision, Theorem 1 gives

$$m_i \geq \zeta > 0. \tag{31}$$

Again using monotonicity of the sigmoid,

$$\rho_i \geq 2 \operatorname{sigm}\left(\frac{\zeta}{\tau_w} \right) - 1 > 0. \tag{32}$$

Since the reliability weight decreases monotonically with ρ_i,

$$w_i \leq 1 - (1 - \alpha_{\min}) \left[2 \operatorname{sigm}\left(\frac{\zeta}{\tau_w} \right) - 1 \right] < 1. \tag{33}$$

Thus, under the separation conditions of Theorem 1, confident correctly labeled units retain their complete task-loss contribution, while confident corrupted units are strictly downweighted. □

A.7.3 Proof of Proposition 2: Finite-Time Prototype-Error Bound

We analyze the effect of imperfect confidence gating on the learned activation-distribution prototypes.

Recall that $\widehat{p}_{k,t}$ denotes the mini-batch prototype estimate for class k at training step t, and let $\xi_{k,t}$ denote the fraction of units assigned to class k by confidence gating whose clean class differs from k. By assumption,

$$\|\widehat{p}_{k,t} - \nu_k\|_2 \leq \epsilon_b + \xi_{k,t} D_{\max}, \tag{34}$$

where ϵ_b represents finite-batch estimation error and $D_{\max}$ bounds the descriptor deviation introduced by an incorrectly assigned unit.

Suppose that

$$\xi_{k,t} \leq \xi_{\max} \tag{35}$$

for all classes and training steps. Define the prototype-estimation error

$$e_{k,t} = \|p_{k,t} - \nu_k\|_2 . \tag{36}$$

Using the EMA update,

$$p_{k,t} = \beta p_{k,t-1} + (1-\beta)\widehat{p}_{k,t}, \tag{37}$$

we obtain

$$\begin{aligned} e_{k,t} &= \|\beta\,(p_{k,t-1} - \nu_k) + (1-\beta)\,(\widehat{p}_{k,t} - \nu_k)\|_2 \\ &\leq \beta\,\|p_{k,t-1} - \nu_k\|_2 + (1-\beta)\,\|\widehat{p}_{k,t} - \nu_k\|_2 \\ &\leq \beta e_{k,t-1} + (1-\beta)\,(\epsilon_b + \xi_{\max} D_{\max}) . \end{aligned} \tag{38}$$

Unrolling the recursion gives

$$\begin{aligned} e_{k,t} &\leq \beta^t e_{k,0} + (1-\beta)\,(\epsilon_b + \xi_{\max} D_{\max}) \sum_{j=0}^{t-1} \beta^j \\ &= \beta^t e_{k,0} + (1-\beta^t)\,(\epsilon_b + \xi_{\max} D_{\max}) . \end{aligned} \tag{39}$$

Therefore,

$$\|p_{k,t} - \nu_k\|_2 \leq \beta^t\,\|p_{k,0} - \nu_k\|_2 + (1-\beta^t)\,(\epsilon_b + \xi_{\max} D_{\max}) . \tag{40}$$

Let

$$\epsilon_0 = \max_k \|p_{k,0} - \nu_k\|_2 . \tag{41}$$

Then a uniform time-dependent prototype-error bound is

$$\epsilon_p(t) = \beta^t \epsilon_0 + (1-\beta^t)\,(\epsilon_b + \xi_{\max} D_{\max}) . \tag{42}$$

Therefore, the separation condition required by Theorem 1 remains satisfied at training step t whenever

$$\Delta > 2\,(r + \epsilon_p(t)) . \tag{43}$$

Finally, because $\beta \in [0, 1)$,

$$\lim_{t\to\infty} \beta^t = 0. \tag{44}$$

Taking the limit superior of Eq. equation 40 yields

$$\limsup_{t\to\infty} \|p_{k,t} - \nu_k\|_2 \leq \epsilon_b + \xi_{\max} D_{\max}. \tag{45}$$

Thus, the effect of prototype initialization decays geometrically, while the asymptotic error is controlled by finite-batch estimation error and the rate of confident misassignment. This proves the finite-time and asymptotic prototype-error bounds. □

A.7.4 Proof of Proposition 3: Task-Gradient Non-Amplification

Let

$$\ell_i(\theta) = \mathcal{L}_{\text{task}}\left(g_\theta(u_i), \tilde{y}_i\right) \tag{46}$$

denote the task loss of supervised unit u_i.

The reliability weight satisfies

$$w_i = 1 - (1 - \alpha_{\min})\rho_i, \tag{47}$$

with

$$0 \leq \rho_i \leq 1. \tag{48}$$

Therefore,

$$\alpha_{\min} \leq w_i \leq 1. \tag{49}$$

When w_i is detached from the computational graph, it is treated as a constant during differentiation. Hence,

$$\nabla_\theta \left(w_i \ell_i(\theta)\right) = w_i \nabla_\theta \ell_i(\theta). \tag{50}$$

Taking the Euclidean norm gives

$$\begin{aligned} \|\nabla_\theta \left(w_i \ell_i(\theta)\right)\|_2 &= w_i \|\nabla_\theta \ell_i(\theta)\|_2 \\ &\leq \|\nabla_\theta \ell_i(\theta)\|_2 . \end{aligned} \tag{51}$$

Therefore, reliability weighting cannot amplify the task-gradient contribution of any supervised unit.

When the activation-distribution reliability signal indicates no disagreement, $\rho_i = 0$, and therefore

$$w_i = 1. \tag{52}$$

In this case,

$$\nabla_\theta \left(w_i \ell_i(\theta)\right) = \nabla_\theta \ell_i(\theta), \tag{53}$$

so the original task gradient is preserved exactly. □

A.7.5 Additional Result: Aggregate Clean-Gradient Preservation

The previous proposition establishes a per-unit property. We now characterize how false-positive reliability decisions affect the aggregate optimization signal contributed by correctly labeled units.

Let $\mathcal{C}$ denote the set of correctly labeled units in a mini-batch and define

$$G_{\mathcal{C}} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \nabla_\theta \ell_i, \tag{54}$$

and

$$G_{\mathcal{C}}^{w} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} w_i \nabla_\theta \ell_i. \tag{55}$$

Let $\mathcal{F} \subseteq \mathcal{C}$ denote the correctly labeled units that are incorrectly assigned a weight smaller than one. Assume

$$\frac{|\mathcal{F}|}{|\mathcal{C}|} \leq q_{\mathrm{fp}}, \tag{56}$$

and that individual clean task gradients satisfy

$$\|\nabla_\theta \ell_i\|_2 \leq G_{\max}. \tag{57}$$

For every $i \in \mathcal{C} \setminus \mathcal{F}$, we have $w_i = 1$. Therefore,

$$G_{\mathcal{C}}^{w} - G_{\mathcal{C}} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{F}} (w_i - 1) \nabla_\theta \ell_i. \tag{58}$$

Applying the triangle inequality,

$$\begin{aligned} \|G_{\mathcal{C}}^{w} - G_{\mathcal{C}}\|_2 &\leq \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{F}} (1 - w_i) \|\nabla_\theta \ell_i\|_2 \\ &\leq \frac{|\mathcal{F}|}{|\mathcal{C}|} (1 - \alpha_{\min}) G_{\max} \\ &\leq q_{\mathrm{fp}} (1 - \alpha_{\min}) G_{\max}. \end{aligned} \tag{59}$$

Hence,

$$\|G_{\mathcal{C}}^{w} - G_{\mathcal{C}}\|_2 \leq q_{\mathrm{fp}} (1 - \alpha_{\min}) G_{\max}. \tag{60}$$

In particular, if $q_{\mathrm{fp}} = 0$, then $\mathcal{F} = \varnothing$ and

$$G_{\mathcal{C}}^{w} = G_{\mathcal{C}}. \tag{61}$$

Thus, when no correctly labeled units are falsely downweighted, the aggregate clean optimization signal is preserved exactly.

We can further establish a sufficient condition for positive alignment between the weighted and unweighted clean gradients. Define

$$E = G_{\mathcal{C}}^{w} - G_{\mathcal{C}}. \tag{62}$$

Then,

$$\begin{aligned}\langle G_{\mathcal{C}}^{w}, G_{\mathcal{C}} \rangle &= \langle G_{\mathcal{C}} + E, G_{\mathcal{C}} \rangle \\ &= \|G_{\mathcal{C}}\|_2^2 + \langle E, G_{\mathcal{C}} \rangle .\end{aligned} \tag{63}$$

By the Cauchy–Schwarz inequality,

$$\langle E, G_{\mathcal{C}} \rangle \geq - \|E\|_2 \|G_{\mathcal{C}}\|_2 . \tag{64}$$

Therefore,

$$\langle G_{\mathcal{C}}^{w}, G_{\mathcal{C}} \rangle \geq \|G_{\mathcal{C}}\|_2 (\|G_{\mathcal{C}}\|_2 - \|E\|_2) . \tag{65}$$

If

$$q_{\mathrm{fp}}(1 - \alpha_{\min}) G_{\max} < \|G_{\mathcal{C}}\|_2 , \tag{66}$$

then Eq. equation 59 implies

$$\|E\|_2 < \|G_{\mathcal{C}}\|_2 . \tag{67}$$

Consequently,

$$\langle G_{\mathcal{C}}^{w}, G_{\mathcal{C}} \rangle > 0. \tag{68}$$

Thus, even when a fraction of correctly labeled units is mistakenly downweighted, the weighted clean gradient remains positively aligned with the original clean optimization direction whenever the distortion induced by these false-positive reliability decisions is sufficiently small. □

A.8 Computational Efficiency and Descriptor Complexity

Table 9 evaluates both the computational cost of the proposed activation-distribution reliability mechanism and the contribution of progressively richer descriptor statistics. For classification, we consider CIFAR-10 with 40% asymmetric label noise using ResNet-34. For semantic segmentation, we use PASCAL VOC 2012 with 40% SCP corruption and DeepLabV3+ with a ResNet-101 encoder.

The results show a consistent benefit from enriching the activation descriptor. On CIFAR-10, using only the activation mean increases accuracy from 81.68% to 88.90%. Adding standard deviation and the near-zero activation ratio further improves performance to 90.80% and 92.10%, respectively. The complete descriptor, which additionally includes skewness and kurtosis, achieves $93.51 \pm 0.12\%$. A similar trend is observed for semantic segmentation: performance increases from 59.72 mIoU with standard cross-entropy training to 64.80, 66.90, and 68.60 mIoU as progressively richer statistics are introduced, reaching 70.31 mIoU with the full descriptor. In particular, adding skewness and kurtosis to the first three statistics provides a further gain of 1.41 percentage points on CIFAR-10 and 1.71 mIoU on PASCAL VOC, indicating that higher-order activation statistics capture supervision-relevant information beyond simple location, dispersion, and sparsity.

These gains are obtained with modest additional training cost. For CIFAR-10, the full descriptor increases training time from 7.80 to 8.90 seconds per epoch, corresponding to approximately 14.1% overhead, while peak memory increases from 1.80 to 1.98 GB. For PASCAL VOC, training time increases from 118 to 136 seconds per epoch, approximately 15.3%, and peak memory rises from 11.20 to 12.10 GB. Prototype storage itself is negligible, requiring only 0.10 MB for CIFAR-10 and less than 0.001 MB for PASCAL VOC. The reduction in throughput therefore primarily reflects the computation of activation-distribution statistics rather than storage of class prototypes.

Importantly, this overhead is restricted to training. Descriptor construction, prototype updates, prototype alignment, and reliability weighting are removed at inference time, leaving the original prediction network unchanged. Consequently, the proposed method introduces no additional inference-time parameters or computation. Overall, the results indicate that the richer activation-distribution

representation provides substantial robustness gains while retaining a relatively small training-time and memory overhead.

Table 9: **Computational overhead and activation-distribution descriptor ablation.** Classification results use CIFAR-10 with 40% asymmetric label noise and a ResNet-34 backbone. Semantic segmentation results use PASCAL VOC 2012 with 40% SCP corruption and DeepLabV3+ with a ResNet-101 encoder. μ, σ, r, γ, and κ denote mean, standard deviation, near-zero activation ratio, skewness, and kurtosis, respectively. Training time, peak GPU memory, throughput, and predictive performance are measured under the same implementation, hardware, batch size, and training configuration for all variants within each task.

Task	Variant	μ	σ	r	γ	κ	Perf. ↑	Time/Epoch (s) ↓	Peak VRAM (GB) ↓	Throughput (img/s) ↑	Prototype Memory
CIFAR-10 ResNet-34	Cross-Entropy	✗	✗	✗	✗	✗	81.68	7.80	1.80	6410	0
	Mean only	✓	✗	✗	✗	✗	88.90	8.10	1.84	6170	0.02 MB
	Mean + Std.	✓	✓	✗	✗	✗	90.80	8.30	1.88	6020	0.04 MB
	Mean + Std. + Sparsity	✓	✓	✓	✗	✗	92.10	8.50	1.92	5880	0.06 MB
	Full Descriptor	✓	✓	✓	✓	✓	**93.51±0.12**	8.90	1.98	5620	0.10 MB
PASCAL VOC DeepLabV3+	Cross-Entropy	✗	✗	✗	✗	✗	59.72	118.0	11.20	89.7	0
	Mean only	✓	✗	✗	✗	✗	64.80	122.0	11.40	86.7	< 0.001 MB
	Mean + Std.	✓	✓	✗	✗	✗	66.90	126.0	11.60	84.0	< 0.001 MB
	Mean + Std. + Sparsity	✓	✓	✓	✗	✗	68.60	130.0	11.80	81.4	< 0.001 MB
	Full Descriptor	✓	✓	✓	✓	✓	**70.31**	136.0	12.10	77.8	< 0.001 MB

A.8.1 Computational and Memory Overhead

The proposed method introduces no additional trainable model parameters and does not modify the inference architecture. Activation-distribution descriptor construction, prototype updates, and reliability weighting are used only during training and are removed entirely at inference time.

For image classification, each class prototype has dimension $5C_\ell$, where C_ℓ is the number of channels in the selected feature layer. The complete prototype bank therefore contains

$$5KC_\ell \tag{69}$$

floating-point values for a K-class problem.

For dense prediction, the local descriptor contains activation-distribution statistics, so the corresponding prototype bank contains $5K$ values.

The additional training computation consists primarily of descriptor-statistic evaluation, prototype-distance computation, and inexpensive exponential moving-average updates. All experiments reported in this work were conducted using a single NVIDIA H100 GPU. Since the reliability mechanism is disabled at test time, inference-time parameter count, memory consumption, and computational cost are identical to those of the underlying backbone.