RELOF: RELIABILITY-GUIDED LOGARITHMIC FUSION FOR ROBUST MULTI-VIEW CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-view learning integrates heterogeneous observations of the same object to improve classification modeling. Despite numerous advances, robustness remains a persistent challenge due to the presence of low-quality or conflicting views. Effectively modeling view reliability in the fused space is therefore nontrivial for robust multi-view classification. This paper proposes Reliability-Guided Logarithmic Fusion (ReLoF), which unifies instance-wise reliability estimation and the logarithmic opinion pool (LogOP) within a categorical probability-opinion framework. In a nutshell, ReLoF first forms a categorical probability opinion for each view, and then estimates its reliability using normalized-entropy confidence and Jensen–Shannon-based inter-view consistency relative to a leave-one-out consensus. Their normalized values serve as LogOP exponents for aggregating class evidence. In addition, ReLoF uses Maximum-Logit (MaxLog) regularization to moderate excessively sharp single-view predictions and mitigate the effect of near-zero true-class probabilities assigned by erroneous, overconfident views. Experimental results on seven benchmark datasets show that ReLoF outperforms baselines remarkably under both normal and conflictive conditions. Ablation studies further show that LogOP improves over conventional weighted averaging and that MaxLog yields consistent numerical improvements under view conflict.

1 INTRODUCTION

Multi-view learning exploits heterogeneous descriptions of the same object and uses their consensus and complementary information for representation learning, clustering, and classification (Zhao et al., 2017; Li et al., 2019). These views may arise from different sensors, modalities, feature descriptors, or feature extractors. Within this paradigm, multi-view classification integrates view-specific representations or decisions into a unified label prediction (Wang et al., 2023; Han et al., 2021; 2023). It supports applications such as medical diagnosis, multi-sensor perception, remote-sensing recognition, and image understanding.

However, complementarity alone does not guarantee equal reliability. Noise, occlusion, acquisition failure, limited semantic content, or cross-view mismatch may degrade a small subset of views while most others retain mutually consistent, label-relevant evidence. An abnormal view may then produce either a nearly uniform, uninformative opinion or a highly confident but incorrect one. Without sample-wise control, such opinions can dilute reliable evidence or mislead the final prediction. Robust multi-view classification must therefore identify trustworthy views and determine how their class evidence should interact.

To handle unreliable views, trusted multi-view learning and quality-aware fusion estimate uncertainty, conflict, or sample-wise quality. Trusted Multi-View Classification (TMC) models predictive uncertainty through Dirichlet evidence and subjective opinions (Han et al., 2021); Evidential Conflictive Multi-view Learning (ECML) estimates reliability for conflictive samples (Xu et al., 2024a); Deep Fuzzy Multi-view Learning (FUML) combines within-view uncertainty with inter-view conflict (Duan et al., 2025); and Quality-aware Multimodal Fusion (QMF) dynamically weights modalities by estimated quality (Zhang et al., 2023). These methods demonstrate the value of sample-wise reliability, but mainly determine how much each view contributes. The pooling rule remains an independent choice. Given the same opinions and weights, the arithmetic opinion pool (AOP) computes $p_{\mathrm{A}}(k) = \sum_{v=1}^{V} w_v p_v(k)$, whereas the logarithmic opinion pool (LogOP) com-

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.

putes $p_{\mathrm{L}}(k) \propto \prod_{v=1}^{V} p_v(k)^{w_v}$ (Genest & Zidek, 1986; Heskes, 1997; Hinton, 2002). AOP allows strong support from one view to compensate for weak support from others; LogOP instead aggregates weighted log-evidence and favors classes supported across multiple reliable views. Thus, fusion geometry can change the decision even when the opinions and reliability weights are fixed.

This observation creates two coupled challenges. First, LogOP exponents must reflect reliability rather than confidence alone: a uniform opinion is uninformative, whereas a concentrated opinion can still be confidently wrong. Reliability assessment should therefore combine intra-view confidence with consistency against prevailing cross-view evidence. Second, LogOP is sensitive to near-zero probabilities. An erroneous overconfident view that assigns negligible probability to the true class can create a soft veto through its large negative log-evidence, even when its weight is small (Tax et al., 2000; Alkoot & Kittler, 2002). Robust logarithmic fusion consequently requires both reliability auditing and overconfidence control.

We address these challenges with Reliability-Guided Logarithmic Fusion (ReLoF). A private encoder and classifier map each view to a categorical opinion in the shared label space. ReLoF combines normalized-entropy confidence with Jensen–Shannon consistency against a leave-one-out consensus, then normalizes the resulting reliability scores into LogOP exponents. With unit-sum weights, it implements fusion stably as a softmax over weighted logits. ReLoF further applies Maximum-Logit (MaxLog) regularization to moderate overly sharp opinions and mitigate their soft-veto risk. Reliability weighting limits the influence of abnormal views, while MaxLog controls their residual logit dominance.

Our main contributions are summarized as follows:

- We propose ReLoF, a robust multi-view probability-opinion fusion framework that uses sample-wise reliability scores as LogOP exponents to govern cross-view class-evidence aggregation in log space.
- We further support this fusion using both a closed-form reliability score that combines intra-view confidence with leave-one-out inter-view consistency and MaxLog regularization that mitigates LogOP's soft-veto effect caused by erroneous, overconfident views.
- We conduct systematic evaluations and controlled fusion comparisons under normal and conflictive conditions, showing that LogOP improves over conventional weighted averaging and MaxLog yields consistent numerical gains under view conflict.

2 Related Work

Multi-view learning and classification. Multi-view learning exploits shared and complementary structure across heterogeneous observations for representation learning, clustering, and supervised prediction (Zhao et al., 2017; Li et al., 2019). Multi-view classification adapts this paradigm to unified label prediction through representation-level or decision-level integration. MetaViewer learns a unified representation through bi-level optimization (Wang et al., 2023), uncertainty-aware methods extend classification to incomplete views (Xie et al., 2023), and evidential opinions with view-specific noise-correlation modeling support learning under noisy labels (Xu et al., 2024b). These studies broaden the applicability of multi-view classification, while robust prediction under sample-specific view degradation further requires explicit reliability assessment.

Trusted multi-view fusion. Trusted methods seek robust decisions under view uncertainty, disagreement, and sample-wise quality variation. TMC combines Dirichlet opinions through evidential fusion (Han et al., 2021). Enhanced TMC (ETMC) introduces dynamic evidential fusion, while QMF weights modalities by estimated quality (Han et al., 2023; Zhang et al., 2023). ECML jointly models decision evidence and reliability for conflictive samples (Xu et al., 2024a), whereas Consistent and Complementary-aware trusted Multi-view Learning (CCML) separates consistent and complementary evidence to address semantic vagueness (Liu et al., 2024). More recently, FUML combines within-view uncertainty and inter-view conflict through fuzzy credibility (Duan et al., 2025). These approaches primarily determine each view's contribution. Inspired by FUML's joint consideration of uncertainty and conflict, ReLoF adopts a simple confidence–consistency score to instantiate sample-wise reliability, while focusing on how the resulting categorical opinions are pooled.

Fusion geometry and confidence control. After reliability weights are obtained, the pooling rule determines how class evidence interacts. AOP combines probabilities linearly, whereas LogOP uses a weighted geometric mean and aggregates relative evidence in log space (Genest & Zidek, 1986; Heskes, 1997; Abbas, 2009). Linear and logarithmic pools have been studied with reliability or consensus weights (Benediktsson & Swain, 1992; Altınçay & Demirekler, 2000), and generalized product-of-experts models use input-dependent exponents to control expert influence (Hinton, 2002; Cao & Fleet, 2014). Logarithmic fusion, however, is vulnerable to overconfident predictions: an expert may assign a numerically zero or near-zero probability to the correct class, creating a veto effect (Tax et al., 2000; Alkoot & Kittler, 2002). Confidence-control methods therefore seek to limit excessive prediction sharpness, and MaxSup penalizes the maximum logit to curb overly dominant outputs (Zhou et al., 2025). Motivated by this gap, we propose ReLoF, which uses sample-wise reliability weights as LogOP exponents and applies maximum-logit regularization to private and fused predictions; controlled AOP–LogOP comparisons isolate the pooling effect.

3 Problem Formulation

Consider a complete multi-view classification dataset $\mathcal{D} = \{(\mathcal{X}_i, y_i)\}_{i=1}^{N}$ containing N labeled samples, V complete views, and K classes, where $\mathcal{X}_i = \{x_i^{(v)}\}_{v=1}^{V}$. Here $i \in \{1, \ldots, N\}$ indexes samples, $v \in \{1, \ldots, V\}$ indexes views, $x_i^{(v)} \in \mathbb{R}^{D_v}$ is the vth view of the ith sample, D_v is its input dimensionality, and $y_i \in \{1, \ldots, K\}$ is the ground-truth label. The views describe different attributes or observation sources and may differ in dimensionality, information density, and noise mechanisms. We seek a classifier $f : \mathbb{R}^{D_1} \times \cdots \times \mathbb{R}^{D_V} \to \{1, \ldots, K\}$ that exploits complementary information, dynamically controls each view's contribution, and forms a robust joint prediction. We focus on complete-view classification rather than missing-view recovery.

Conflictive samples can arise from localized view degradation: a small subset of views becomes uninformative or misleading, while most other views preserve a shared label-consistent signal. This prevailing signal makes cross-view agreement informative for reliability assessment without requiring the abnormal views to be identified in advance.

For clarity, we fix one sample and omit its sample index hereafter. Thus, $x_i^{(v)}$, y_i, $p_{i,v}$, and $w_{i,v}$ are written as x_v, y, p_v, and w_v, respectively. Let $\mathcal{V} = \{1, \ldots, V\}$ denote the view-index set. When unambiguous, u is a dummy view index, and k and t are class indices.

4 Reliability-Guided Logarithmic Fusion

Figure 1 shows the framework of the proposed ReLoF. Given heterogeneous views of the same instance, ReLoF first maps each view to a probability opinion in a shared class space, estimates sample-wise reliability from intra-view confidence and inter-view consistency, and then forms a joint prediction through reliability-weighted LogOP. Because LogOP is sensitive to near-zero probabilities, we penalize maximum-logit dominance during training to mitigate the soft-veto risk caused by erroneous but overconfident single-view opinions.

4.1 Categorical Probability Opinions

The private encoder E_v and classifier C_v associated with the vth view first produce a representation $h_v = E_v(x_v)$ and logits $z_v = C_v(h_v) \in \mathbb{R}^K$. The logits are then converted to a probability opinion $p_v = \mathrm{softmax}(z_v)$ in the shared label space, where $p_v(k) > 0$ and $\sum_{k=1}^{K} p_v(k) = 1$. Here $p_v(k)$ denotes the probability assigned by the vth view to class k. These opinions place heterogeneous views on the same K-class probability simplex, allowing within-view uncertainty, inter-view class relations, and subsequent fusion to be defined in one semantic space.

4.2 Instance-Wise Reliability Score

Let w_v denote the eventual fusion weight assigned to the vth view. In LogOP, this weight exponentiates the probability opinion, or equivalently scales the log-evidence, that the view supplies for

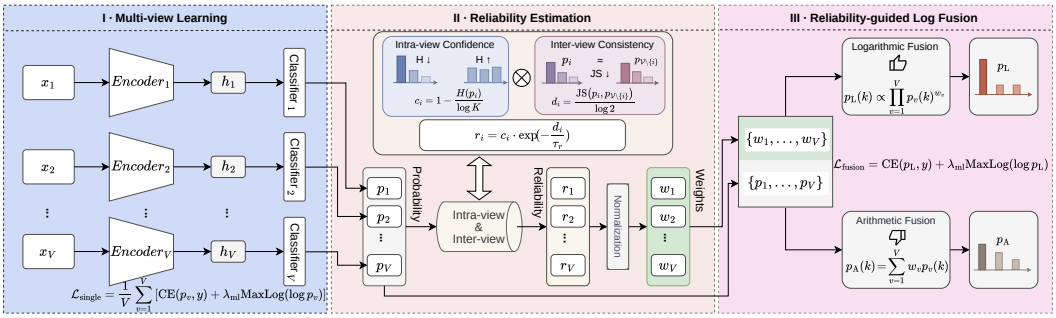


Figure 1: Overview of ReLoF. Private encoders and classifiers produce categorical opinions. Each view is audited using normalized-entropy confidence and Jensen–Shannon-based inter-view consistency with a leave-one-out consensus. The normalized reliability scores become LogOP exponents, and AOP is shown as the controlled counterpart. The MaxLog penalty is displayed in its equivalent probability-space form, $\mathrm{MaxLog}(\log p) = \mathrm{MaxLog}(z)$, for private and fused supervision.

every class. A high-weight view should therefore exhibit both high intra-view confidence and strong inter-view consistency. ReLoF constructs sample-wise reliability from these two criteria.

We first measure the intra-view confidence of the vth opinion using the complement of normalized entropy:

$$c_v = 1 - \frac{H(p_v)}{\log K}, \qquad H(p_v) = -\sum_{k=1}^{K} p_v(k) \log p_v(k). \tag{1}$$

The entropy $H(p_v)$ is categorical entropy (Shannon, 1948), and $c_v \in [0, 1]$ measures departure from the uniform distribution. Nearly uniform opinions usually lack stable class evidence and should receive a low fusion weight, whereas concentrated distributions indicate higher intra-view confidence.

Intra-view confidence alone cannot determine whether that confidence is warranted: a view can be highly confident while conflicting sharply with the others. In such localized conflicts, the evidence shared by the other views provides a natural reference for auditing an individual opinion. We therefore construct a leave-one-out consensus that excludes the current view:

$$p_{\mathcal{V}\setminus\{v\}} = \frac{1}{V-1} \sum_{u \in \mathcal{V}\setminus\{v\}} p_u. \tag{2}$$

Here u indexes the remaining views. The leave-one-out construction prevents the target opinion from mixing with its own reference. We assess inter-view consistency by quantifying the deviation of the vth view from this consensus using normalized Jensen–Shannon (JS) divergence,

$$d_v = \frac{\mathrm{JS}(p_v, p_{\mathcal{V}\setminus\{v\}})}{\log 2}. \tag{3}$$

The divergence $\mathrm{JS}(\cdot, \cdot)$ is symmetric and bounded (Lin, 1991). A larger $d_v \in [0, 1]$ indicates greater deviation and therefore lower inter-view consistency with the other views.

ReLoF maps inter-view deviation to a continuous consistency factor and combines it multiplicatively with intra-view confidence:

$$r_v = c_v \exp\left(-\frac{d_v}{\tau_r}\right), \tag{4}$$

where $\tau_r > 0$ is a reliability temperature that controls the decay induced by inter-view deviation. This multiplicative form acts as a soft AND gate: a view is highly eligible for fusion only when it is both confident and consistent with the other views. Accordingly, the confidence factor suppresses nearly uniform, uninformative opinions, whereas the consistency factor suppresses sharply concentrated opinions that contradict the prevailing cross-view evidence.

In the nondegenerate case $\sum_{u=1}^{V} r_u > 0$, we normalize reliability scores into relative within-sample view weights:

$$w_v = \frac{r_v}{\sum_{u=1}^{V} r_u}, \qquad w_v \geq 0, \qquad \sum_{v=1}^{V} w_v = 1. \tag{5}$$

The weights express relative contributions within the current sample rather than fixed global view importance. During training, the reliability score r_v is detached from the computational graph before the fusion weights are formed, so the fusion loss does not backpropagate through the reliability-score computation. When no view provides a valid reliability signal, the weights can fall back to a uniform distribution. The score r_v is not a calibrated probability that the vth view is correct; it is an eligibility score for reliable fusion.

4.3 Reliability-Guided LogOP Fusion

Given normalized weights, ReLoF uses LogOP to aggregate probability opinions. For class k, the fused probability is

$$p_{\mathrm{L}}(k) = \frac{\prod_{v=1}^{V} p_v(k)^{w_v}}{\sum_{t=1}^{K} \prod_{v=1}^{V} p_v(t)^{w_v}}. \tag{6}$$

Equivalently,

$$p_{\mathrm{L}} = \mathrm{softmax}\left(\sum_{v=1}^{V} w_v \log p_v \right). \tag{7}$$

Since $p_v = \mathrm{softmax}(z_v)$ and $\sum_{v=1}^{V} w_v = 1$, implementation can equivalently use weighted logits:

$$z_{\mathrm{L}} = \sum_{v=1}^{V} w_v z_v, \qquad p_{\mathrm{L}} = \mathrm{softmax}(z_{\mathrm{L}}). \tag{8}$$

This form avoids numerical underflow from explicit probability products. In contrast to AOP, LogOP aggregates evidence in log-evidence space and emphasizes whether a class receives support from multiple reliable views. As a controlled counterpart, we compute AOP using exactly the same view opinions and reliability weights:

$$p_{\mathrm{A}}(k) = \sum_{v=1}^{V} w_v p_v(k). \tag{9}$$

This comparison replaces only the final fusion rule and therefore isolates the behavioral difference between LogOP and AOP under identical inputs.

4.4 Maximum-Logit Regularization and Training Objective

The multiplicative structure of LogOP introduces an important risk. For the true class y, the joint log-evidence contains

$$\sum_{v=1}^{V} w_v \log p_v(y). \tag{10}$$

If an erroneous overconfident view drives $p_v(y)$ close to zero, then $\log p_v(y)$ makes a large negative contribution. Even when its reliability weight is small, the view can create a soft veto. Reliability weighting reduces this influence but cannot completely remove the near-zero-probability problem.

We therefore use Maximum-Logit (MaxLog) regularization. For any logits vector $z \in \mathbb{R}^K$, the maximum-logit penalty is

$$\mathrm{MaxLog}(z) = \max_{1 \leq k \leq K} z(k) - \frac{1}{K} \sum_{k=1}^{K} z(k). \tag{11}$$

It penalizes the gap between the dominant logit and the class-average logit, reducing excessively sharp single-view predictions. Since $p = \mathrm{softmax}(z)$ implies that $\log p$ differs from z only by a class-independent normalization term, this term cancels in the maximum-minus-mean difference and

yields $\mathrm{MaxLog}(\log p) = \mathrm{MaxLog}(z)$. Let $\mathrm{CE}(z, y)$ denote the cross-entropy (CE) loss computed from logits z and label y. The resulting supervised loss is

$$\ell_{\mathrm{sup}}(z, y) = \mathrm{CE}(z, y) + \lambda_{\mathrm{ml}} \mathrm{MaxLog}(z), \lambda_{\mathrm{ml}} \geq 0, \tag{12}$$

where λ_{ml} controls the regularization strength. When $\lambda_{\mathrm{ml}} = 0$, the loss reduces to standard cross-entropy. When the ground-truth class is also the class with the maximum logit, the loss is equivalent to cross-entropy with a uniformly smoothed target $q(y) = 1 - \lambda_{\mathrm{ml}} + \lambda_{\mathrm{ml}}/K$ and $q(k) = \lambda_{\mathrm{ml}}/K$ for $k \neq y$. Thus, the regularizer softens the one-hot target and preserves probability mass for non-target classes. This equivalence is conditional: when an incorrect class has the maximum logit, MaxLog directly penalizes that dominant class rather than applying a fixed label-smoothed target regardless of which class has the maximum logit.

ReLoF jointly supervises the fused and private predictions:

$$\begin{aligned} \mathcal{L}_{\mathrm{fusion}} &= \ell_{\mathrm{sup}}(z_{\mathrm{L}}, y), \\ \mathcal{L}_{\mathrm{single}} &= \frac{1}{V} \sum_{v=1}^{V} \ell_{\mathrm{sup}}(z_v, y), \\ \mathcal{L} &= \mathcal{L}_{\mathrm{fusion}} + \mathcal{L}_{\mathrm{single}}. \end{aligned} \tag{13}$$

The fused loss trains the joint prediction, while the single-view loss preserves the independent discriminative ability of each private classifier. MaxLog regularization does not change the LogOP rule itself; instead, it reduces excessive confidence when probability opinions are generated. Thus, it helps prevent an erroneous confident view from collapsing the true-class probability, while its reliability weight further attenuates its negative log-evidence contribution.

5 Experiments

We evaluate ReLoF through overall performance comparisons, controlled ablations of LogOP and MaxLog, and parameter-sensitivity analyses. The overall comparison reports results on all seven benchmark datasets. Owing to space constraints, the main paper presents the ablation studies on a representative subset of the datasets and includes only the most informative parameter-sensitivity analysis; results for the remaining datasets, additional sensitivity analyses, and further experimental details are provided in the appendix. The ablations first use counterfactual fusion-geometry diagnostics to isolate the test-time effect of replacing AOP with LogOP while holding the view-level opinions and reliability weights fixed. They then separate the effects of the training-time pooling rule and compare CE with MaxLog under a fixed LogOP fusion scheme.

5.1 Experimental Setup

Datasets. We use seven multi-view classification datasets: Handwritten (HW) (Duin, 1998), LandUse (Yang & Newsam, 2010; Dai & Van Gool, 2013), NUS-WIDE-OBJ (NUSOBJ) (Chua et al., 2009; Kang et al., 2020), Scene15 (Scene) (Fei-Fei & Perona, 2005; Dai & Van Gool, 2013), Leaves100 (Leaves) (Cope et al., 2013), MSRC-v1 (MSRC) (Winn & Jojic, 2005; Zhuge et al., 2017), and Caltech-5V (Caltech) (Fei-Fei et al., 2004; Xu et al., 2022). Each dataset is split using an 8:2 training-to-test ratio. Table 1 summarizes their statistics.

Table 1: Dataset statistics.

Dataset	Size	Classes	Dimensions
HW	2,000	10	240; 76; 216; 47; 64; 6
LandUse	2,100	21	20; 59; 40
NUSOBJ	30,000	31	65; 226; 145; 74; 129
Scene	4,485	15	20; 59; 40
Leaves	1,600	100	64; 64; 64
MSRC	210	7	1302; 48; 512; 100; 256; 210
Caltech	1,400	7	40; 254; 1984; 512; 928

To further assess robustness, we follow the conflictive test setting of prior work (Xu et al., 2024a). Gaussian noise with varying standard deviations is added to designated views of selected test instances to simulate low-quality observations. Separately, a randomly selected subset of views is altered to conflict with the true label. In both cases, only the selected views are corrupted, leaving the remaining observations unchanged and preserving the prevailing cross-view signal.

Table 2: Accuracy (%) on normal test sets. ReLoF denotes the reliability-weighted LogOP model trained with MaxLog regularization. Entries report means with standard deviations in parentheses. Throughout, bold marks the highest mean accuracy and underlining the second highest. Δ is ReLoF's gain over the strongest competing mean.

Method	HW	LandUse	NUSOBJ	Scene	Leaves	MSRC	Caltech
TMC	98.25(0.59)	46.71(2.27)	39.56(0.35)	72.85(0.95)	92.84(1.24)	95.24(2.61)	94.18(1.61)
ETMC	98.70(0.42)	62.52(1.72)	46.88(0.34)	79.30(0.83)	94.53(1.58)	96.67(3.05)	97.25(0.89)
CCML	98.48(0.36)	55.33(2.15)	45.90(0.36)	75.47(1.24)	97.84(0.68)	96.67(2.86)	97.29(0.64)
TMNR	98.72(0.39)	48.07(2.95)	42.58(0.53)	64.60(1.16)	86.69(3.70)	96.43(2.66)	97.75(0.77)
UIMC	98.60(0.37)	64.48(2.51)	41.14(1.94)	75.77(1.13)	95.53(1.30)	96.19(3.23)	98.04(0.29)
QMF	98.02(0.49)	44.98(1.72)	40.22(0.56)	71.58(0.95)	99.28(0.71)	97.38(2.25)	97.36(0.74)
ECML	98.07(0.32)	55.76(1.37)	41.78(0.43)	75.46(0.94)	94.75(1.52)	93.33(3.96)	97.46(0.72)
FUML	98.90(0.32)	76.62(1.23)	47.69(0.56)	80.48(0.86)	99.81(0.21)	98.57(1.90)	96.68(1.68)
ReLoF	**99.15**(0.41)	**79.45**(1.98)	**49.60**(0.55)	**81.58**(0.86)	**99.91**(0.20)	**99.05**(1.58)	**98.11**(0.58)
Δ	+0.25	+2.83	+1.91	+1.10	+0.10	+0.48	+0.07

Table 3: Accuracy (%) on conflictive test sets.

Method	HW	LandUse	NUSOBJ	Scene	Leaves	MSRC	Caltech
TMC	96.80(0.50)	39.29(2.96)	36.34(0.31)	63.15(1.03)	75.84(2.48)	92.14(3.85)	90.64(1.56)
ETMC	96.42(0.59)	51.12(1.92)	42.55(0.30)	68.61(1.63)	77.59(2.68)	93.10(3.44)	94.57(1.47)
CCML	93.25(1.01)	46.38(2.65)	42.13(0.39)	66.32(1.96)	80.78(2.05)	93.57(3.85)	90.79(0.89)
TMNR	98.10(0.48)	40.95(2.13)	39.57(0.45)	57.59(1.30)	67.19(2.46)	94.05(3.57)	96.93(0.74)
UIMC	98.07(0.56)	53.17(3.13)	37.76(0.41)	65.96(1.39)	78.66(1.56)	93.33(4.10)	97.39(0.48)
QMF	92.62(0.94)	37.14(1.99)	37.39(0.46)	60.38(1.38)	79.84(1.61)	93.81(3.40)	91.54(1.06)
ECML	94.35(0.88)	48.29(1.90)	39.31(0.44)	66.74(1.00)	78.09(2.43)	91.19(3.85)	91.86(1.43)
FUML	98.62(0.28)	69.90(1.55)	45.74(0.30)	74.04(1.17)	93.88(0.93)	96.90(3.02)	92.89(4.41)
ReLoF	**98.92**(0.46)	**71.95**(2.68)	**47.26**(0.33)	**75.25**(1.09)	**94.97**(0.81)	**98.57**(1.58)	**97.57**(0.52)
Δ	+0.30	+2.05	+1.52	+1.21	+1.09	+1.67	+0.18

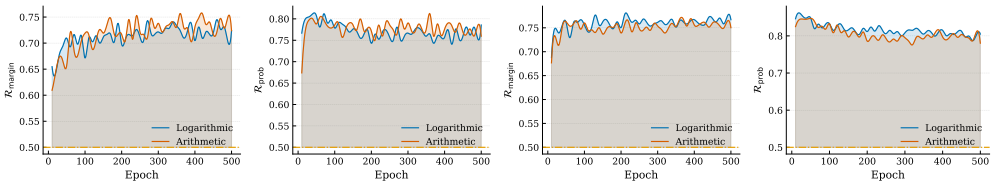


Figure 2: Fusion-geometry diagnostics on LandUse (left) and Scene (right). Each pair shows $\mathcal{R}_{\text{margin}}$ followed by $\mathcal{R}_{\text{prob}}$; curves compare LogOP and AOP with fixed opinions and reliability weights, and the legend identifies the training pool.

Baselines. We compare ReLoF with TMC (Han et al., 2021), ETMC (Han et al., 2023), CCML (Liu et al., 2024), Trusted Multi-view Noise Refining (TMNR) (Xu et al., 2024b), Uncertainty-induced Incomplete Multi-View Data Classification (UIMC) (Xie et al., 2023), QMF (Zhang et al., 2023), ECML (Xu et al., 2024a), and FUML (Duan et al., 2025). All methods use the same data partitions and random-seed protocol. In the main results, ReLoF denotes the final model trained and evaluated with reliability-weighted LogOP and MaxLog regularization; the CE-only variant is reported solely in the ablations. Unless stated otherwise, we set the MaxLog weight to $\lambda_{\text{ml}} = 0.1$ and the reliability temperature to $\tau_r = 0.25$ for all datasets.

Evaluation metrics. We report accuracy as the mean and standard deviation over 10 independent random seeds. For two counterfactual diagnostics, let the test set contain N_{te} samples with labels y_i. From a fixed trained model, we compute the LogOP and AOP distributions $p_{\text{L},i}$ and $p_{\text{A},i}$ from

exactly the same view opinions $\{p_{i,v}\}_{v=1}^{V}$ and reliability weights $\{w_{i,v}\}_{v=1}^{V}$, so only the final fusion rule changes. Define $R_{\mathrm{prob},i} = p_{\mathrm{L},i}(y_i)/p_{\mathrm{A},i}(y_i)$ and $\mathcal{R}_{\mathrm{prob}} = N_{\mathrm{te}}^{-1} \sum_{i=1}^{N_{\mathrm{te}}} \mathbb{I}[R_{\mathrm{prob},i} > 1]$, the fraction of samples for which LogOP increases the true-class probability. For $X \in \{\mathrm{A}, \mathrm{L}\}$, define the log-margin $m_{X,i} = \log p_{X,i}(y_i) - \max_{k \neq y_i} \log p_{X,i}(k)$ and $\mathcal{R}_{\mathrm{margin}} = N_{\mathrm{te}}^{-1} \sum_{i=1}^{N_{\mathrm{te}}} \mathbb{I}[m_{\mathrm{L},i} > m_{\mathrm{A},i}]$, the fraction for which LogOP enlarges this margin. Values above 0.5 indicate improvement on a majority of samples; unlike accuracy, the margin diagnostic remains sensitive when the predicted label is unchanged.

5.2 Comparison with State-of-the-Art Methods

We evaluate all methods on seven datasets under normal and conflictive test conditions using identical data splits and ten random seeds. Tables 2 and 3 show the resulting mean accuracy and standard deviation. We can see that our ReLoF achieves the highest mean accuracy on every dataset in both settings. On the normal test sets, the largest gains over the strongest counterpart are 2.83, 1.91, and 1.10 percentage points on datasets LandUse, NUSOBJ, and Scene, respectively. Under localized view degradation, the corresponding margins reach 2.05, 1.67, 1.52, 1.21 points on datasets LandUse, MSRC, NUSOBJ, and Scene, respectively. These results may be attributed to the complementary roles of reliability auditing as they suppress abnormal views. The results also show that the proposed LogOP reinforces label-consistent evidence shared across unaffected views, and MaxLog limits the residual dominance of erroneous overconfident opinions. We next examine these mechanisms through three ablations: *fusion-geometry evaluation*, *training–test pool comparison*, and a *CE-versus-MaxLog comparison*.

5.3 Ablation Studies

Fusion-geometry evaluation. Overall accuracy does not reveal whether changing only the final pool improves the true-class score for individual samples. Figure 2 therefore tracks $\mathcal{R}_{\mathrm{margin}}$ and $\mathcal{R}_{\mathrm{prob}}$ over training on the LandUse and Scene test sets. The first two panels report the diagnostics on LandUse, while the last two report their counterparts on Scene. At each checkpoint, p_{L} and p_{A} are computed from exactly the same view opinions and reliability weights, so the curves isolate the immediate effect of replacing AOP with LogOP without changing the learned representations or model parameters.

Across both datasets, $\mathcal{R}_{\mathrm{margin}}$ and $\mathcal{R}_{\mathrm{prob}}$ remain well above 0.5 for both LogOP- and AOP-trained models throughout the recorded checkpoints. Thus, replacing AOP with LogOP increases the true-class probability and enlarges the true-class log-margin on a majority of test samples, regardless of which pool produced the learned opinions. The consistent patterns on LandUse and Scene show that this immediate test-time geometry effect is not confined to a single dataset. These sample-level results reveal LogOP's potential and motivate the two controlled ablations that follow.

Training–test pool comparison. Guided by counterfactual evaluation, we fix the supervision loss to CE and vary only the opinion pool used during training and testing. Table 4 organizes each dataset into two training rows, one for LogOP and one for AOP, while the last two columns report predictions obtained with AOP and LogOP at test time. Within a fixed training row, the two test columns share exactly the same view opinions and reliability weights and differ only in the final pool; their difference therefore isolates test-time fusion geometry. Within a fixed test column, comparing the two training rows reveals how the training operator shapes the learned view opinions. In the discussion below, $X \to Y$ denotes training with pool X and testing with pool Y.

Table 4: CE training–test pool comparison on four conflictive datasets (accuracy in %).

Dataset	Train	Test AOP	Test LogOP
LandUse	LogOP	69.74 ± 1.95	$\underline{71.69 \pm 2.17}$
	AOP	70.55 ± 1.89	$\mathbf{72.52 \pm 1.88}$
NUSOBJ	LogOP	44.90 ± 0.45	$\mathbf{47.20 \pm 0.40}$
	AOP	44.79 ± 0.30	$\underline{46.34 \pm 0.20}$
Scene	LogOP	73.34 ± 1.63	$\mathbf{74.77 \pm 1.37}$
	AOP	73.17 ± 1.16	$\underline{74.47 \pm 1.00}$
Leaves	LogOP	91.59 ± 0.82	$\mathbf{94.38 \pm 1.03}$
	AOP	91.56 ± 0.48	$\underline{94.31 \pm 0.90}$

Across all four datasets, using LogOP rather than AOP at test time consistently improves accuracy under either training pool. With LogOP training, the gains reach 2.30 percentage points on NUSOBJ and 2.79 points on Leaves; with AOP training, they reach 1.97 points on LandUse and 2.75 points on Leaves. The matched LogOP $\rightarrow$ LogOP configuration also outperforms AOP $\rightarrow$ AOP on every dataset. These gains at fixed view opinions and reliability weights demonstrate an immediate test-time fusion-geometry benefit, while the differences between training rows indicate that the training pool additionally shapes the learned opinions.

MaxLog versus CE under LogOP. We fix LogOP for both training and testing and vary only the supervision loss. Table 5 shows the mean and standard deviation of conflictive-test accuracy across ten random seeds on four datasets. MaxLog improves the mean on every dataset, with gains ranging from 0.06 percentage points on NUSOBJ to 0.59 on Leaves; the gains on LandUse and Scene are 0.26 and 0.48 points, respectively. The standard deviation is also lower with MaxLog on NUSOBJ, Scene, and Leaves, although it increases on LandUse. Because the reliability-weight computation and LogOP rule are fixed, these modest but consistent mean gains isolate the contribution of the supervision objective and support MaxLog's intended role in confidence control.

Table 5: Conflictive-test accuracy (%) of CE and MaxLog under LogOP, where Δ denotes MaxLog gain.

Dataset	CE	MaxLog	Δ
LandUse	71.69(2.17)	**71.95**(2.68)	+0.26
NUSOBJ	47.20(0.40)	**47.26**(0.33)	+0.06
Scene	74.77(1.37)	**75.25**(1.09)	+0.48
Leaves	94.38(1.03)	**94.97**(0.81)	+0.59

5.4 Parameter Sensitivity

The MaxLog weight λ_{ml} controls the strength of overconfidence suppression: a value that is too small may be insufficient to prevent near-zero probabilities, whereas an excessively large value can weaken class separation and reduce classification accuracy. We therefore vary λ_{ml} over $\{0.05, 0.10, 0.15, 0.20, 0.30\}$ on the conflictive LandUse and Scene test sets while fixing LogOP for training and testing and holding all other settings constant.

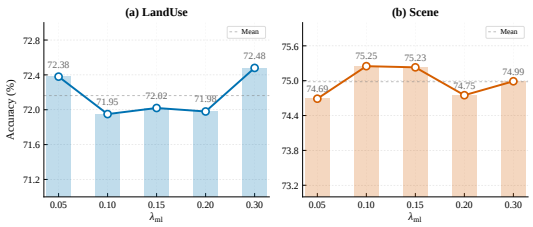


Figure 3: Sensitivity of conflictive-test accuracy to the MaxLog weight λ_{ml} on LandUse and Scene, with LogOP used for both training and test.

The results in Figure 3 show that performance remains stable across the tested range on both datasets, without systematic degradation as the weight changes. Although the best setting varies by dataset, the limited variation indicates that the model is not sensitive to moderate changes in λ_{ml}. We therefore use $\lambda_{\mathrm{ml}} = 0.10$ as a shared regularization weight throughout, avoiding dataset-specific tuning.

6 Conclusion

We presented ReLoF for robust multi-view classification, combining entropy-based confidence and leave-one-out consistency into reliability-weighted LogOP fusion and applying MaxLog regularization to control overconfidence. Across seven benchmarks, ReLoF achieves the highest mean accuracy under both normal and conflictive conditions. Controlled ablations distinguish LogOP's immediate test-time geometric benefit from its influence during training and show that MaxLog provides additional gains under fixed reliability weighting and LogOP fusion, supporting the complementary roles of the three components. Like existing methods, ReLoF may nevertheless degrade when erroneous views form a majority, reliable evidence is confined to few views, or true-class probabilities approach zero. Future work will investigate majority-bias detection, class-conditional reliability, missing views, and adaptive LogOP exponents.

REPRODUCIBILITY STATEMENT

The main paper specifies the model, reliability computation, fusion rule, training objective, datasets, and evaluation metrics. The supplementary material further documents the network architecture, preprocessing and data splits, optimization settings, random seeds, software and hardware environment, conflictive-test construction, and complete experimental results.

AI USE STATEMENT

A generative AI tool was used solely for language polishing, grammar correction, and formatting assistance. It was not used to generate scientific ideas, methods, experimental results, analyses, or conclusions. The authors manually reviewed all AI-assisted edits for technical accuracy and consistency with the original manuscript and take full responsibility for the final content of this work.

REFERENCES

Ali E. Abbas. A kullback–leibler view of linear and log-linear pools. *Decision Analysis*, 6(1):25–37, 2009. doi: 10.1287/deca.1080.0133.

Fuad M. Alkoot and Josef Kittler. Modified product fusion. *Pattern Recognition Letters*, 23(8): 957–965, 2002. doi: 10.1016/S0167-8655(02)00025-9.

Hakan Altınçay and Mübeccel Demirekler. An information theoretic framework for weight estimation in the combination of probabilistic classifiers for speaker identification. *Speech Communication*, 30(4):255–272, 2000. doi: 10.1016/S0167-6393(99)00054-0.

Jón Atli Benediktsson and Philip H. Swain. Consensus theoretic classification methods. *IEEE Transactions on Systems, Man, and Cybernetics*, 22(4):688–704, 1992. doi: 10.1109/21.156582.

Yanshuai Cao and David J. Fleet. Generalized product of experts for automatic and principled fusion of gaussian process predictions, 2014. URL https://arxiv.org/abs/1410.7827.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. NUS-WIDE: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 1–9, 2009. doi: 10.1145/1646396.1646452.

James Cope, Thibaut Beghin, Paolo Remagnino, and Sarah Barman. One-hundred plant species leaves data set. UCI Machine Learning Repository, 2013. Dataset.

Dengxin Dai and Luc Van Gool. Ensemble projection for semi-supervised image classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2072–2079, 2013. doi: 10.1109/ICCV.2013.259.

Siyuan Duan, Yuan Sun, Dezhong Peng, Guiduo Duan, Xi Peng, and Peng Hu. Deep fuzzy multi-view learning for reliable classification. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14735–14758. PMLR, 2025. URL https://proceedings.mlr.press/v267/duan25a.html.

Robert P. W. Duin. Multiple features. UCI Machine Learning Repository, 1998. Dataset.

Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 524–531, 2005. doi: 10.1109/CVPR.2005.16.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 178, 2004. doi: 10.1109/CVPR.2004.383.

Christian Genest and James V. Zidek. Combining probability distributions: A critique and an annotated bibliography. *Statistical Science*, 1(1):114–135, 1986. doi: 10.1214/ss/1177013825.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=OOsR8BzCnl5.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2551–2566, 2023. doi: 10.1109/TPAMI.2022.3171983.

Tom Heskes. Selecting weighting factors in logarithmic opinion pools. In *Advances in Neural Information Processing Systems 10*, pp. 266–272, 1997. URL https://proceedings.neurips.cc/paper_files/paper/1997/hash/59f51fd6937412b7e56ded1ea2470c25-Abstract.html.

Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800, 2002. doi: 10.1162/089976602760128018.

Zhao Kang, Wangtao Zhou, Zhitong Zhao, Junming Shao, Meng Han, and Zenglin Xu. Large-scale multi-view subspace clustering in linear time. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4412–4419, 2020. doi: 10.1609/aaai.v34i04.5867.

Yingming Li, Ming Yang, and Zhongfei Zhang. A survey of multi-view representation learning. *IEEE Transactions on Knowledge and Data Engineering*, 31(10):1863–1883, 2019. doi: 10.1109/TKDE.2018.2872063.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Ying Liu, Lihong Liu, Cai Xu, Xiangyu Song, Ziyu Guan, and Wei Zhao. Dynamic evidence decoupling for trusted multi-view learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7269–7277, 2024. doi: 10.1145/3664647.3681404.

Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3–4):379–423, 623–656, 1948.

David M. J. Tax, Martijn van Breukelen, Robert P. W. Duin, and Josef Kittler. Combining multiple classifiers by averaging or by multiplying? *Pattern Recognition*, 33(9):1475–1485, 2000. doi: 10.1016/S0031-3203(99)00138-7.

Ren Wang, Haoliang Sun, Yuling Ma, Xiaoming Xi, and Yilong Yin. MetaViewer: Towards a unified multi-view representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11590–11599, 2023. doi: 10.1109/CVPR52729.2023.01115.

John Winn and Nebojsa Jojic. LOCUS: Learning object classes with unsupervised segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 1, pp. 756–763, 2005. doi: 10.1109/ICCV.2005.148.

Mengyao Xie, Zongbo Han, Changqing Zhang, Yichen Bai, and Qinghua Hu. Exploring and exploiting uncertainty for incomplete multi-view classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19873–19882, 2023. doi: 10.1109/CVPR52729.2023.01903.

Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. Reliable conflictive multi-view learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 16129–16137, 2024a. doi: 10.1609/aaai.v38i14.29546.

Cai Xu, Yilin Zhang, Ziyu Guan, and Wei Zhao. Trusted multi-view learning with label noise. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pp. 5263–5271, 2024b. doi: 10.24963/ijcai.2024/582. URL https://www.ijcai.org/proceedings/2024/0582.

Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16051–16060, 2022. doi: 10.1109/CVPR52688.2022.01558. URL https://openaccess.thecvf.com/content/CVPR2022/html/Xu_Multi-Level_Feature_Learning_for_Contrastive_Multi-View_Clustering_CVPR_2022_paper.html.

Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pp. 270–279, 2010. doi: 10.1145/1869790.1869829.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41753–41769. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhang23ar.html.

Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017. doi: 10.1016/j.inffus.2017.02.007.

Yuxuan Zhou, Heng Li, Zhi-Qi Cheng, Xudong Yan, Yifei Dong, Mario Fritz, and Margret Keuper. MaxSup: Overcoming representation collapse in label smoothing. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://neurips.cc/virtual/2025/poster/116896.

Wenzhang Zhuge, Chenping Hou, Yuanyuan Jiao, Jia Yue, Hong Tao, and Dongyun Yi. Robust auto-weighted multi-view subspace clustering with common subspace representation matrix. *PLOS ONE*, 12(5):e0176769, 2017. doi: 10.1371/journal.pone.0176769.

SUPPLEMENTARY MATERIAL: THEORETICAL AND EXPERIMENTAL DETAILS FOR RELOF

SUPPLEMENT OVERVIEW

This supplement provides the theoretical foundations and complete experimental details for ReLoF. It first relates the probability-score interface of ReLoF to Dirichlet evidential opinions and fuzzy-set membership opinions, derives AOP and LogOP as KL barycenters, contrasts their weighting properties, and establishes the softmax-specific equivalence between LogOP and reliability-weighted logits. The experimental part then documents the complete configuration and evaluation protocol, and reports the full training–test pool comparisons, fusion-geometry diagnostics, and MaxLog sensitivity results.

A THEORETICAL FOUNDATIONS AND SCOPE OF RELOF

The characterizations in this section are standard results included to clarify the scope and implementation of ReLoF; they are not claimed as novel theoretical contributions. Different trusted multi-view methods encode a view-specific decision using different mathematical objects. This section distinguishes two representative uncertainty-aware constructions from the first-order categorical opinions used by ReLoF. Let

$$\Delta_K = \left\{ q \in \mathbb{R}^K_{\geq 0} : \sum_{k=1}^{K} q(k) = 1 \right\}$$

denote the K-class probability simplex.

A.1 OPINION REPRESENTATIONS PRIOR TO FUSION

Dirichlet Evidential Opinions. Trusted Multi-View Classification (TMC) (Han et al., 2021) represents the uncertainty of view v by a Dirichlet distribution over a latent categorical probability vector $\pi_v \in \Delta_K$. Given nonnegative evidence $e_v \in \mathbb{R}^K_{\geq 0}$, its concentration parameters and total strength are

$$\alpha_v(k) = e_v(k) + 1, \qquad S_v = \sum_{k=1}^{K} \alpha_v(k).$$

The corresponding density is

$$\mathrm{Dir}(\pi_v \mid \alpha_v) = \frac{\Gamma(S_v)}{\prod_{k=1}^{K} \Gamma(\alpha_v(k))} \prod_{k=1}^{K} \pi_v(k)^{\alpha_v(k)-1}. \tag{14}$$

Thus, a Dirichlet opinion is a distribution over points on the categorical simplex rather than a single categorical distribution.

Under the uniform-base-rate subjective-logic mapping used by TMC, the class belief masses and residual uncertainty mass are

$$b_v(k) = \frac{e_v(k)}{S_v} = \frac{\alpha_v(k) - 1}{S_v}, \qquad u_v = \frac{K}{S_v}, \tag{15}$$

and satisfy $\sum_k b_v(k) + u_v = 1$. The predictive categorical probability is the Dirichlet mean,

$$p_v^{\mathrm{D}}(k) := \mathbb{E}[\pi_v(k)] = \frac{\alpha_v(k)}{S_v} = b_v(k) + \frac{u_v}{K}. \tag{16}$$

Equation (16) is the appropriate first-order probability interface; the uncertainty mass u_v is not itself an additional class.

TMC natively combines the mass opinions with Dempster's rule. For two opinions, let

$$C_{12} = \sum_{k \neq t} b_1(k) b_2(t).$$

Table 6: Distinction between representative uncertainty-aware opinion representations and ReLoF's first-order categorical opinion.

Aspect	Dirichlet evidential opinion	Fuzzy-set opinion	ReLoF opinion
Primary object	Dirichlet density over $\pi_v \in \Delta_K$	Membership vector $m_v \in [0,1]^K$	Categorical distribution $p_v \in \Delta_K^\circ$
Class-level quantity	Predictive mean $\alpha_v(k)/S_v$	Membership $m_v(k)$ and credibility $c_v^{\mathrm{F}}(k)$	Probability score $s_v(k) = p_v(k)$
Uncertainty	Residual mass $u_v = K/S_v$	Mean binary entropy of category credibility	Normalized categorical entropy, used in reliability estimation
Native combination	Dempster's combination of belief and uncertainty masses	Reliability-weighted arithmetic membership aggregation	Reliability-weighted LogOP; AOP is the controlled alternative
Probability interface	Use the Dirichlet mean p_v^{D}	Normalize a selected membership or credibility score	Already normalized; no conversion is required

When $C_{12} < 1$, their joint masses are

$$b_{1\oplus 2}(k) = \frac{b_1(k)b_2(k) + b_1(k)u_2 + b_2(k)u_1}{1 - C_{12}}, \qquad u_{1\oplus 2} = \frac{u_1 u_2}{1 - C_{12}}. \tag{17}$$

This rule operates on belief and uncertainty masses and is therefore distinct from pooling only the predictive means in Equation (16).

Fuzzy-Set Membership Opinions. Deep Fuzzy Multi-view Learning (FUML) (Duan et al., 2025) instead interprets bounded classifier outputs as fuzzy-set memberships. If $a_v \in \mathbb{R}^K$ is a nonzero view-specific network output, FUML constructs

$$m_v(k) = \mathrm{ReLU}\left(\frac{a_v(k)}{\|a_v\|_\rho}\right) \in [0,1], \tag{18}$$

where $\rho \geq 1$ and $\|\cdot\|_\rho$ is the normalization used by the membership layer. A membership vector m_v is generally not constrained to sum to one: $m_v(k)$ is a possibility-like degree of membership, not necessarily a categorical probability.

FUML complements membership with the necessity-like quantity

$$n_v(k) = 1 - \max_{t \neq k} m_v(t)$$

and defines category credibility by

$$c_v^{\mathrm{F}}(k) = \frac{m_v(k) + n_v(k)}{2}. \tag{19}$$

Writing $h_2(x) = -x\log x - (1-x)\log(1-x)$ with $0\log 0 := 0$, its view-specific uncertainty is

$$U_v^{\mathrm{F}} = \frac{1}{K\log 2}\sum_{k=1}^{K} h_2(c_v^{\mathrm{F}}(k)). \tag{20}$$

For nonzero membership vectors, FUML further measures the conflict of view v against the remaining views through the mean cosine distance

$$o_v = \frac{1}{V-1}\sum_{u \neq v}\left(1 - \frac{m_v^\top m_u}{\|m_v\|_2\|m_u\|_2}\right). \tag{21}$$

Its native Dual-reliable Multi-view Fusion assigns

$$w_v^{\mathrm{F}} = \frac{\exp((1-U_v^{\mathrm{F}})(1-o_v))}{\sum_{u=1}^{V}\exp((1-U_u^{\mathrm{F}})(1-o_u))} \tag{22}$$

and aggregates memberships arithmetically as

$$m_{\mathrm{agg}} = \sum_{v=1}^{V} w_v^{\mathrm{F}} m_v.$$

Hence, this construction retains fuzzy membership, necessity, uncertainty, and conflict as separate semantic quantities.

ReLoF's Probability-Score Interface. ReLoF uses a simpler first-order representation. Each view directly produces a categorical distribution

$$p_v = \mathrm{softmax}(z_v) \in \Delta_K^{\circ},$$

which is also used as its class-score vector, i.e., $s_v(k) := p_v(k)$. Consequently, no second-order density, residual uncertainty mass, or non-normalized membership vector is introduced. Uncertainty and cross-view inconsistency are assessed from these categorical opinions only to construct the reliability weights; the opinions themselves remain points on the shared class simplex.

The pooling layer can also accept any nonnegative class-score vector s_v with $\sum_k s_v(k) > 0$ after the probability conversion

$$q_v(k) = \frac{s_v(k)}{\sum_{t=1}^{K} s_v(t)}. \tag{23}$$

The two reliability-weighted pools are then

$$\mathcal{F}_{\mathrm{A}}(\{q_v\})(k) = \sum_{v=1}^{V} w_v q_v(k), \tag{24}$$

$$\mathcal{F}_{\mathrm{L}}(\{q_v\})(k) = \frac{\prod_{v=1}^{V} q_v(k)^{w_v}}{\sum_{t=1}^{K}\prod_{v=1}^{V} q_v(t)^{w_v}}. \tag{25}$$

For LogOP, the positively weighted opinions have a nonempty common support. Thus, ReLoF's softmax opinions satisfy the stronger condition $q_v(k) > 0$.

This interface permits the same pooling geometry to be applied to other opinion representations. For example, choosing $s_v = \alpha_v$ in a Dirichlet model gives $q_v = p_v^{\mathrm{D}}$, while a fuzzy model may choose $s_v = m_v$ or $s_v = c_v^{\mathrm{F}}$ and normalize it with Equation (23). These conversions replace the representations' native fusion operators; they do not preserve, or claim equivalence to, Dempster's mass combination or FUML's membership-space DRF. Moreover, the KL-barycenter results below apply to the normalized probabilities q_v, not to arbitrary unnormalized scores.

Table 6 summarizes the semantic distinctions among the three opinion representations.

A.2 KL Barycenter of AOP and LogOP

The KL-barycentric interpretations below are standard properties of linear and logarithmic opinion pools (Genest & Zidek, 1986; Heskes, 1997; Abbas, 2009).

Setup and notation. We fix one sample and suppress its sample index. Let $p_v \in \Delta_K^{\circ}$ be the categorical opinion of view $v \in \{1, \ldots, V\}$, where

$$\Delta_K^{\circ} = \left\{ q \in \mathbb{R}_{>0}^{K} : \sum_{k=1}^{K} q(k) = 1 \right\}.$$

The strict positivity follows from the softmax parameterization used by ReLoF. Let $r_v \geq 0$ denote the reliability score of view v. In the nondegenerate case $\sum_{u=1}^{V} r_u > 0$, ReLoF uses the normalized weights

$$w_v = \frac{r_v}{\sum_{u=1}^{V} r_u}, \qquad w_v \geq 0, \qquad \sum_{v=1}^{V} w_v = 1.$$

Table 7: Properties of arithmetic and logarithmic opinion pooling under the same nonnegative unit-sum weights.

Property	AOP	LogOP
KL barycenter	$\arg\min_q \sum_v w_v \, \mathrm{KL}(p_v \| q)$	$\arg\min_q \sum_v w_v \, \mathrm{KL}(q \| p_v)$
Aggregation geometry	Weighted arithmetic mean in probability space; lies in the convex hull	Normalized weighted geometric mean; linear aggregation in log-evidence space
Role of w_v	Scales the probability contribution $p_v(k)$	Scales every class log-odds contribution from view v
Support with exact zeros	Union of positively weighted supports	Intersection of positively weighted supports, if nonempty
Conflict behavior	Compensatory; isolated low probabilities can be offset by other views	Consensus-seeking; near-zero probabilities can exert a soft-veto effect
Shared update	Does not commute in general	Commutes when the weights sum to one

If all reliability scores are zero, the weights are set to $w_v = 1/V$. Thus, the fusion weights always lie on the probability simplex. For $p, q \in \Delta_K^\circ$, define

$$\mathrm{KL}(p\|q) = \sum_{k=1}^{K} p(k) \log \frac{p(k)}{q(k)}.$$

Thus, the two fusion rules arise from the same KL divergence with its arguments placed in opposite orders.

AOP from Expert-to-Pool KL. Consider the weighted expert-to-pool objective

$$J_{\mathrm{A}}(q) = \sum_{v=1}^{V} w_v \, \mathrm{KL}(p_v\|q), \qquad q \in \Delta_K^\circ.$$

Let $\bar{p} = \sum_{v=1}^{V} w_v p_v$. Terms independent of q can be collected into constants C_{A} and C'_{A}, giving

$$J_{\mathrm{A}}(q) = C_{\mathrm{A}} - \sum_{k=1}^{K} \bar{p}(k) \log q(k) \tag{26}$$

$$= C'_{\mathrm{A}} + \mathrm{KL}(\bar{p}\|q). \tag{27}$$

Nonnegativity of KL divergence, with equality only when its two arguments coincide, therefore gives the unique minimizer

$$p_{\mathrm{A}} = \arg\min_{q \in \Delta_K^\circ} J_{\mathrm{A}}(q) = \sum_{v=1}^{V} w_v p_v. \tag{28}$$

Hence, AOP is the KL barycenter when each view opinion occupies the first argument and the fused distribution occupies the second. The weights act linearly on probabilities.

LogOP from Pool-to-Expert KL. Reverse the KL arguments and consider

$$J_{\mathrm{L}}(q) = \sum_{v=1}^{V} w_v \, \mathrm{KL}(q\|p_v).$$

Define the unnormalized weighted geometric mean and its normalizer as

$$g(k) = \exp\left(\sum_{v=1}^{V} w_v \log p_v(k)\right) \tag{29}$$

$$= \prod_{v=1}^{V} p_v(k)^{w_v}, \tag{30}$$

$$Z_{\mathrm{L}} = \sum_{t=1}^{K} g(t). \tag{31}$$

Because every $p_v(k)$ is positive, $Z_{\mathrm{L}} > 0$. Let $p_{\mathrm{L}}(k) = g(k)/Z_{\mathrm{L}}$. Then

$$J_{\mathrm{L}}(q) = \sum_{k=1}^{K} q(k) \log \frac{q(k)}{g(k)} \tag{32}$$

$$= \mathrm{KL}(q \| p_{\mathrm{L}}) - \log Z_{\mathrm{L}}. \tag{33}$$

The second term is independent of q, so the unique minimizer is

$$p_{\mathrm{L}} = \arg\min_{q \in \Delta_K^\circ} J_{\mathrm{L}}(q), \qquad p_{\mathrm{L}}(k) = \frac{\prod_{v=1}^{V} p_v(k)^{w_v}}{\sum_{t=1}^{K} \prod_{v=1}^{V} p_v(t)^{w_v}}. \tag{34}$$

LogOP is therefore the KL barycenter when the fused distribution occupies the first argument. Its weights act linearly on log probabilities, or equivalently as exponents on probabilities.

A.3 Properties of Weighted AOP and LogOP

Both pools are normalized, permutation-equivariant with respect to matched opinion-weight pairs, and idempotent: if $p_v = p$ for every positively weighted view, then $p_{\mathrm{A}} = p_{\mathrm{L}} = p$. Their different KL orientations, however, produce different fusion geometries.

Arithmetic weighting. AOP lies in the convex hull of the view opinions. In particular,

$$\min_{v: w_v > 0} p_v(k) \le p_{\mathrm{A}}(k) \le \max_{v: w_v > 0} p_v(k)$$

for every class k. Evidence is compensatory: a low probability from one view can be offset by a high probability from another. If boundary distributions with zeros are admitted, AOP has the union of the positively weighted supports and therefore does not impose a veto merely because one view assigns zero probability.

Logarithmic weighting. LogOP averages class evidence in log space. For any classes k and t,

$$\log \frac{p_{\mathrm{L}}(k)}{p_{\mathrm{L}}(t)} = \sum_{v=1}^{V} w_v \log \frac{p_v(k)}{p_v(t)}. \tag{35}$$

Thus, each weight directly scales a view's pairwise log-odds. LogOP favors classes that receive compatible support across reliable views. If exact zeros are admitted, its support is the intersection of the positively weighted supports, provided that this intersection is nonempty. ReLoF's softmax opinions have full support, so exact vetoes do not occur; nevertheless, a near-zero probability can still make a large negative log-evidence contribution, which explains the soft-veto behavior addressed by MaxLog.

Shared evidence. LogOP also commutes with a common Bayesian update (Genest & Zidek, 1986). For a positive likelihood $L(k)$, define

$$\mathcal{U}_L[p](k) = \frac{L(k)p(k)}{\sum_{t=1}^{K} L(t)p(t)}.$$

Table 8: Dataset and training configurations. The train/test counts follow the seeded 8:2 split used in all experiments.

Dataset	Samples	Views	Classes	Train/Test	Batch	Learning rate
HW	2,000	6	10	1,600/400	100	1×10^{-3}
LandUse	2,100	3	21	1,680/420	100	1×10^{-3}
NUSOBJ	30,000	5	31	24,000/6,000	400	2×10^{-4}
Scene	4,485	3	15	3,588/897	100	1×10^{-3}
Leaves	1,600	3	100	1,280/320	100	1×10^{-3}
MSRC	210	6	7	168/42	100	1×10^{-3}
Caltech	1,400	5	7	1,120/280	100	1×10^{-3}

Table 9: Complete CE-only training–test pool comparison under normal and conflictive testing. Each entry reports the mean accuracy (%) with the population standard deviation over ten seeds in parentheses.

		Training pool → test pool			
Dataset	Setting	LogOP → LogOP	LogOP → AOP	AOP → AOP	AOP → LogOP
HW	Normal	99.13(0.36)	98.95(0.38)	99.00(0.42)	99.15(0.37)
	Conflictive	98.90(0.42)	98.60(0.45)	98.45(0.58)	98.92(0.45)
LandUse	Normal	79.76(1.70)	76.83(1.70)	77.14(1.17)	79.45(1.30)
	Conflictive	71.69(2.17)	69.74(1.95)	70.55(1.89)	72.52(1.88)
NUSOBJ	Normal	49.64(0.39)	47.03(0.45)	47.00(0.43)	48.69(0.47)
	Conflictive	47.20(0.40)	44.90(0.45)	44.79(0.30)	46.34(0.20)
Scene	Normal	81.78(0.95)	79.40(1.01)	79.25(0.70)	81.23(0.76)
	Conflictive	74.77(1.37)	73.34(1.63)	73.17(1.16)	74.47(1.00)
Leaves	Normal	99.94(0.13)	98.41(0.55)	98.38(0.65)	99.94(0.13)
	Conflictive	94.38(1.03)	91.59(0.82)	91.56(0.48)	94.31(0.90)
MSRC	Normal	99.29(1.52)	99.29(2.14)	99.76(0.71)	99.52(1.43)
	Conflictive	99.05(1.17)	99.29(1.09)	99.05(1.17)	99.05(1.17)
Caltech	Normal	98.25(0.59)	98.21(0.62)	98.11(0.58)	98.25(0.59)
	Conflictive	97.54(0.54)	97.50(0.45)	97.54(0.56)	97.64(0.48)

Because $\sum_v w_v = 1$,

$$\begin{aligned} &\mathrm{LogOP}_w\left(\{\mathcal{U}_L[p_v]\}_{v=1}^V\right) \\ &\quad = \mathcal{U}_L\left[\mathrm{LogOP}_w\left(\{p_v\}_{v=1}^V\right)\right]. \end{aligned} \tag{36}$$

Indeed, the common factor becomes $\prod_v L(k)^{w_v} = L(k)$ and the view-specific normalizers cancel during renormalization. AOP generally does not satisfy this identity because the normalizer of $\mathcal{U}_L[p_v]$ differs across views.

Table 7 summarizes the resulting contrast between the two pooling geometries.

A.4 Softmax-Specific Weighted-Logit Equivalence

For ReLoF, view v produces logits $z_v \in \mathbb{R}^K$ and the categorical opinion

$$p_v(k) = \mathrm{softmax}(z_v)_k = \frac{\exp(z_v(k))}{Z_v}, \tag{37}$$

$$Z_v = \sum_{t=1}^{K} \exp(z_v(t)). \tag{38}$$

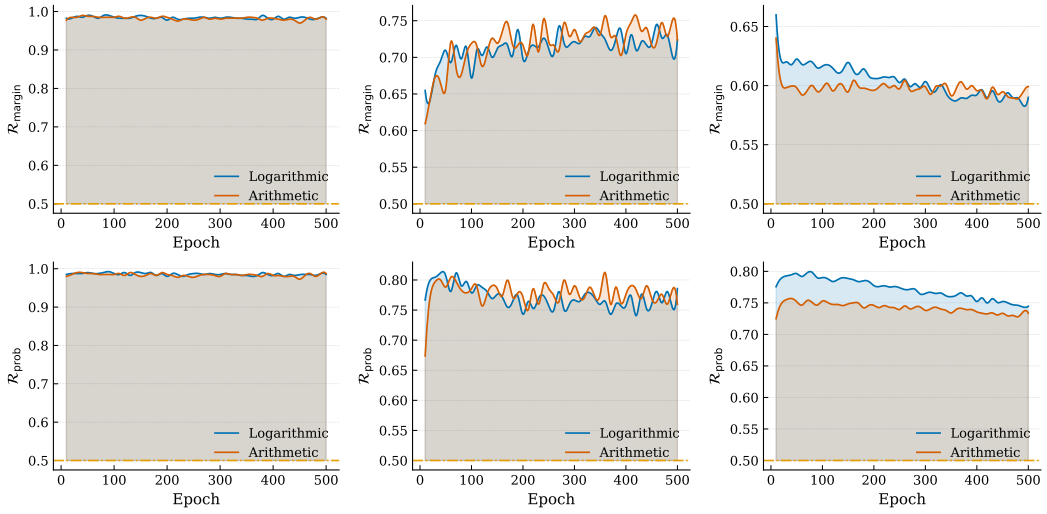


Figure 4: Fusion-geometry diagnostics on HW, LandUse, and NUSOBJ. Within each dataset, the upper and lower panels report $\mathcal{R}_{\text{margin}}$ and $\mathcal{R}_{\text{prob}}$, respectively. The legend labels "Logarithmic" and "Arithmetic" denote models trained with LogOP and AOP, respectively; both diagnostics compare the two test-time pools using fixed opinions and reliability weights.

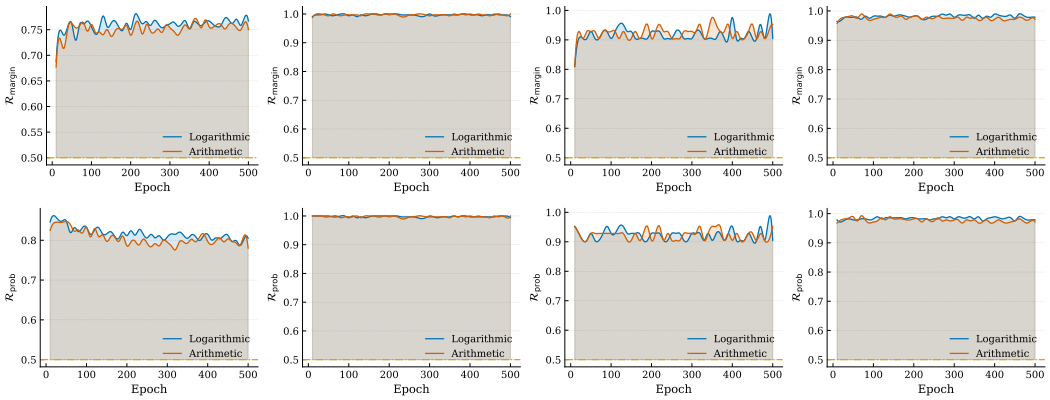


Figure 5: Fusion-geometry diagnostics on Scene, Leaves, MSRC, and Caltech. Within each dataset, the upper and lower panels report $\mathcal{R}_{\text{margin}}$ and $\mathcal{R}_{\text{prob}}$, respectively. The legend labels "Logarithmic" and "Arithmetic" denote models trained with LogOP and AOP, respectively; both diagnostics compare the two test-time pools using fixed opinions and reliability weights.

Substitution into the unnormalized LogOP score yields

$$\prod_{v=1}^{V} p_v(k)^{w_v} = \prod_{v=1}^{V} \left(\frac{\exp(z_v(k))}{Z_v} \right)^{w_v} \tag{39}$$

$$= \exp\left(\sum_{v=1}^{V} w_v z_v(k) \right) \prod_{v=1}^{V} Z_v^{-w_v}. \tag{40}$$

The factor $\prod_v Z_v^{-w_v}$ depends on the sample and views but not on the class index k. Following the main paper, define the fused logits as $z_{\text{L}} = \sum_{v=1}^{V} w_v z_v$. The class-independent factor therefore

cancels between the numerator and denominator of LogOP:

$$p_{\mathrm{L}}(k) = \frac{\exp\left(z_{\mathrm{L}}(k)\right)}{\sum_{t=1}^{K} \exp\left(z_{\mathrm{L}}(t)\right)} \tag{41}$$

$$= \operatorname{softmax}\left(z_{\mathrm{L}}\right)_k . \tag{42}$$

Thus, probability-space LogOP is exactly equivalent to applying softmax to the reliability-weighted logits; it is not an approximation. Unit-sum weights preserve idempotence and the common-update property above. If a shared opinion temperature $T > 0$ is used, the same derivation gives $p_{\mathrm{L}} = \operatorname{softmax}(z_{\mathrm{L}}/T)$. ReLoF uses $T = 1$.

Centered log-evidence form. ReLoF forms centered log evidence for numerical clarity:

$$\ell_v^{\mathrm{c}}(k) = \log p_v(k) - \frac{1}{K}\sum_{t=1}^{K} \log p_v(t) \tag{43}$$

$$= z_v(k) - \frac{1}{K}\sum_{t=1}^{K} z_v(t). \tag{44}$$

Consequently, $\sum_v w_v \ell_v^{\mathrm{c}}$ differs from z_{L} only by a scalar shared by all classes. Since softmax is invariant to addition of a class-independent scalar,

$$\operatorname{softmax}\left(\sum_v w_v \ell_v^{\mathrm{c}}\right) = \operatorname{softmax}\left(z_{\mathrm{L}}\right) = p_{\mathrm{L}}. \tag{45}$$

Therefore, centered log evidence, weighted logits, and normalized probability products compute exactly the same fused distribution.

Scope of the equivalence. The "logits" in Equation (42) are specifically the pre-softmax network outputs that generate the ReLoF opinions. For generic positive probability opinions, one may always define canonical log scores $\ell_v = \log q_v$ and write

$$\mathcal{F}_{\mathrm{L}}(\{q_v\}) = \operatorname{softmax}\left(\sum_v w_v \ell_v\right).$$

This identity is a log-probability representation, not an equivalence to the original outputs of an arbitrary uncertainty model. In particular, applying LogOP to Dirichlet means yields weighted log concentrations $\sum_v w_v \log \alpha_v$ up to class-independent constants, and applying it to positive fuzzy scores yields weighted log memberships or log credibilities. Neither expression is generally a weighted average of the originating network's logits. Exact fusion of the original logits is obtained only when each pooled opinion is generated by the corresponding softmax parameterization (with a common temperature when a single shared logit scale is asserted).

B Experimental Settings and Full Results

B.1 Datasets and Optimization

We evaluate ReLoF on the seven complete multi-view classification benchmarks used in the main paper. Table 8 shows dataset sizes and optimization settings. Every feature dimension in each view is min–max normalized to $[0, 1]$ before the random split. All models are trained for 500 epochs with Adam and a weight decay of 10^{-5}. We use a batch size of 400 and a learning rate of 2×10^{-4} for NUSOBJ; all other datasets use a batch size of 100 and a learning rate of 10^{-3}. The model is evaluated every 10 epochs, and the best observed test accuracy is retained for each seed. This checkpoint-selection rule is applied identically to ReLoF and all baselines. Accuracy is used as the common multiclass metric to maintain direct comparability with prior trusted multi-view classification evaluations; the probability-geometry diagnostics separately assess changes that may not alter the predicted label.

B.2 Fixed ReLoF Configuration

Each private encoder is an MLP with hidden dimensions $(128, 64)$ and a 32-dimensional output, using ReLU activations, layer normalization at the output, and dropout 0.1. The opinion temperature is 1.0, the reliability temperature is $\tau_r = 0.25$. The fused and mean private-view losses have equal unit weights. Unless an ablation explicitly states otherwise, both training and testing use LogOP and the MaxLog weight is $\lambda_{\mathrm{ml}} = 0.10$.

B.3 Hardware and Software Environment

All experiments were conducted on a single NVIDIA GeForce RTX 5070 Ti GPU with 16,303 MiB of memory under Linux 6.14.0-37-generic (x86_64, glibc 2.39), using NVIDIA driver 570.133.07. The software environment comprised Python 3.9.25, PyTorch 2.8.0+cu129, CUDA 12.9, cuDNN 9.10.2, NumPy 2.0.1, and scikit-learn 1.6.1. For deterministic execution, PyTorch deterministic algorithms and cuDNN deterministic mode were enabled, cuDNN benchmarking was disabled, and `CUBLAS_WORKSPACE_CONFIG` was set to `:4096:8`.

B.4 Random Seeds and Data Partitioning

Each experimental configuration is run with the ten integer seeds $\mathcal{S} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. At the beginning of each run, the same seed is assigned to Python's `random` module, NumPy, the PyTorch CPU generator, and all CUDA generators. We enable deterministic PyTorch algorithms, set cuDNN to deterministic mode, disable cuDNN benchmarking, and use the cuBLAS workspace configuration specified above.

After seeding, all sample indices are randomly shuffled and split once into 80% training and 20% test data. This is a non-stratified index split: the first $\lfloor 0.8N \rfloor$ shuffled indices form the training set and the remainder form the test set. No separate validation partition is used, and no split file is loaded or saved. Instead, each seed regenerates its own split. The same seed also controls model initialization, batch shuffling, and the conflictive-test sampling described below. The training loader is reshuffled at every epoch using the seeded PyTorch random stream, whereas the test loader preserves the split order. For each configuration, we report the arithmetic mean and population standard deviation (`ddof=0`) of the ten retained test accuracies.

B.5 Uniform Comparison Protocol

For every seed, ReLoF and all baselines use the same training–test partition, the same Gaussian-noise and cross-class-conflict realizations, and the same best-test-accuracy checkpoint-selection rule. Thus, the reported comparisons hold the data partition, corruption pattern, and selection criterion fixed across methods.

B.6 Conflictive-Test Construction

Only the test split is modified in the conflictive setting; the training data remain unchanged. Starting from a seeded test split, we apply the following two corruptions in order.

Gaussian-noise corruption. We sample without replacement $\lfloor 0.10 N_{\mathrm{te}} \rfloor$ test instances. For each selected instance i, we draw the number of affected views M_i uniformly from $\{0, \ldots, V-1\}$ and then sample M_i distinct views. For every selected view v, the normalized feature vector is replaced by

$$\widetilde{x}_i^{(v)} = x_i^{(v)} + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, 0.5^2 I_{D_v}).$$

The same absolute standard deviation is applied independently to every feature coordinate in every dataset and view after normalization.

Cross-class conflict corruption. Independently, we sample without replacement $\lfloor 0.40 N_{\mathrm{te}} \rfloor$ test instances. For each selected instance, one view is sampled uniformly from $\{1, \ldots, V\}$ and replaced by the corresponding view of a reference instance from the next class:

$$\widetilde{x}_i^{(v)} = x_j^{(v)}, \qquad y_j = 1 + (y_i \bmod K).$$

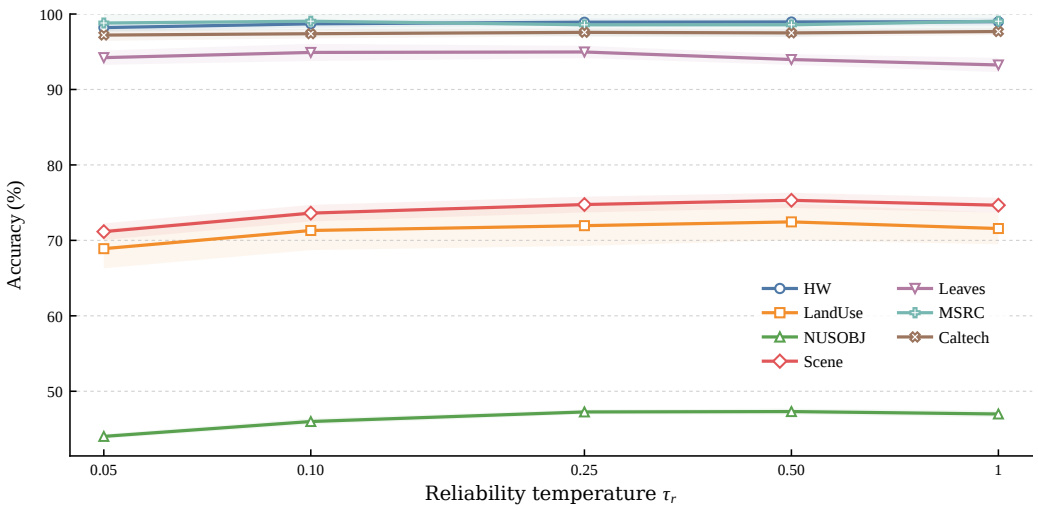


Figure 6: Sensitivity of conflictive-test accuracy to the reliability temperature τ_r on all seven datasets. All other settings are fixed to the default configuration.

This expression follows the main paper's label convention $y_i \in \{1, \ldots, K\}$ and cyclically selects the next class. The reference j is the first stored sample of that class in the complete dataset. Reference views are collected after the Gaussian-noise stage. The noise and conflict subsets are sampled separately and may overlap. Because conflict replacement is performed after Gaussian corruption, the chosen conflicting view overwrites any earlier noise applied to that same view, while other corrupted views remain unchanged. Both corruptions are resampled for every seed. The normal-test setting uses the corresponding seeded split without either modification.

B.7 Complete Training–Test Pool Comparisons

Table 9 shows CE-only accuracy for all four training–test pool combinations under normal and conflictive testing on the seven benchmarks. Within a fixed training pool, both test pools reuse the same learned opinions and reliability weights.

B.8 Complete Fusion-Geometry Diagnostics

Figures 4 and 5 extend fusion-geometry evaluation to all seven benchmarks. At each checkpoint, the LogOP and AOP test-time predictions are computed from the same view opinions and reliability weights. The two curves distinguish the pool used to train those shared opinions, while the horizontal reference at 0.5 marks equal proportions of improved and non-improved test instances.

B.9 Reliability-Temperature Sensitivity

We further evaluate the sensitivity of ReLoF to the reliability temperature τ_r, which controls the decay induced by inter-view disagreement. We vary τ_r over $\{0.05, 0.10, 0.25, 0.50, 1.00\}$ on all seven datasets while keeping all other settings fixed.

As shown in Figure 6, very small values of τ_r produce a sharper decay of reliability under inter-view disagreement and reduce accuracy on several datasets. Moderate values provide a better balance between suppressing inconsistent views and retaining complementary evidence, whereas larger values offer no consistent improvement across all datasets. We therefore use $\tau_r = 0.25$ as a shared setting throughout the experiments, without dataset-specific tuning.

B.10 MaxLog Search Range

Following the maximum-logit suppression principle of MaxSup (Zhou et al., 2025), we use the same MaxLog operator and supervised loss as in the main paper. For logits $z \in \mathbb{R}^K$, they are

$$\mathrm{MaxLog}(z) = \max_{1 \le k \le K} z(k) - \frac{1}{K} \sum_{k=1}^{K} z(k), \tag{46}$$

$$\ell_{\mathrm{sup}}(z, y) = \mathrm{CE}(z, y) + \lambda_{\mathrm{ml}} \, \mathrm{MaxLog}(z), \qquad \lambda_{\mathrm{ml}} \ge 0. \tag{47}$$

We treat $\lambda_{\mathrm{ml}} = 0$ as the CE reference and evaluate the nonzero sensitivity grid $\lambda_{\mathrm{ml}} \in \{0.05, 0.10, 0.15, 0.20, 0.30\}$ on the conflictive LandUse and Scene test sets, with LogOP fixed for both training and testing and all other settings unchanged.

The range spans weak to strong confidence suppression around 0.10. Small weights may not control the soft-veto effect of an erroneous view, whereas large weights may over-flatten logits and weaken class separation. Conflictive-test accuracy varies by less than 0.6 percentage points across the grid on both datasets. We therefore use $\lambda_{\mathrm{ml}} = 0.10$ as one dataset-independent value rather than tuning it separately for each benchmark.