Robust Multiview Learning Under Noisy Correspondence

Shuxian Li
College of Computer Science
Sichuan University
Chengdu, China
lishuxian.gm@gmail.com

Changhao He
College of Computer Science
Sichuan University
Chengdu, China
hechanghao.gm@gmail.com

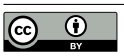
Xi Peng
School of Artificial Intelligence
Sichuan University
Chengdu, China
pengx.gm@gmail.com

Peng Hu*
College of Computer Science
Sichuan University
Chengdu, China
penghu.ml@gmail.com

Abstract

Multiview learning aims to exploit cross-view consistency and complementarity to improve prediction performance. In practice, however, asynchronous acquisition, transmission errors, and sensor malfunctions may introduce *noisy correspondence*, where some views within a sample are misaligned, inevitably breaking the cross-view alignment assumption underlying standard multiview learning and corrupting both training and inference. Existing methods typically learn to downweight mismatched views at inference time, while assuming correctly aligned views during training. As a result, they overlook the biased learning induced by mismatched training views, which act as view-level noisy supervision, thus leading to inaccurate reliability estimation and performance degradation. To address this issue, we propose **CARF**, a **C**onsistency-**A**ware **R**obust **F**usion framework for robust multiview learning under noisy correspondence, which integrates two key components. Specifically, we introduce a robust Hellinger-based objective that aligns each view with the fused prediction (*i.e.*, consensus) while mitigating the adverse impact of mismatched views, thereby promoting cross-view consistency among correctly matched views and stabilizing multiview fusion. To further suppress the influence of noisy views, we estimate view reliability by jointly modeling *view consistency* and *label consistency*, which measure decision-level agreement across different views and fidelity to the assigned label, respectively. The resulting reliability scores are then used to reweight the view-specific optimization during training and to perform reliability-aware fusion during both training and inference, thus yielding more robust consensus learning and more reliable final prediction. Extensive experiments on nine datasets demonstrate that **CARF** consistently outperforms state-of-the-art methods on both clean and noisy data. Code is available at https://github.com/XLearning-SCU/2026-ACM-MM-CARF.

*Corresponding author.



This work is licensed under a Creative Commons Attribution 4.0 International License.
MM '26, Rio de Janeiro, Brazil
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2213-4/2026/11
https://doi.org/10.1145/3767308.3836492

CCS Concepts

• **Computing methodologies** → **Supervised learning by classification**; *Learning latent representations.*

Keywords

Multiview Learning, Noisy Correspondence, Multimodal Learning

ACM Reference Format:
Shuxian Li, Changhao He, Xi Peng, and Peng Hu. 2026. Robust Multiview Learning Under Noisy Correspondence. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26), November 10–14, 2026, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3767308.3836492

1 Introduction

With the increasing availability of multiview data in autonomous driving [6], medical diagnosis [1, 33], and embodied intelligence [27], multiview learning has attracted growing attention for its ability to exploit both cross-view consistency and complementarity. In practice, however, existing multiview data collection pipelines (*e.g.*, web crawling and crowdsourcing), as well as real-time asynchronous acquisition and sensor malfunctions, may incorrectly associate data from different views, which unavoidably introduce *noisy correspondence* into both training and inference, thus degrading the prediction performance.

Existing trusted multiview learning methods [28, 31, 44], which estimate uncertainty based on Dempster-Shafer theory [8, 13, 45] or fuzzy set theory [10, 41, 52], may be intuitively regarded as partial remedies for noisy correspondence. However, they are primarily designed to handle conflicting views at test time and generally assume that views within each training sample are correctly aligned during training. By contrast, noisy correspondence disrupts cross-view consistency in both phases. During training, it induces a view-level noisy-label effect, leading to overfitting and biased uncertainty estimation. During inference, this biased learning, together with the cross-view conflicts introduced by noisy correspondence, further results in unreliable fusion and potentially incorrect predictions. This coupled training-inference failure mode makes noisy correspondence a particularly challenging problem in multiview learning.

To address this issue, we propose **CARF**, a **C**onsistency-**A**ware **R**obust **F**usion framework for learning from noisy multiview data

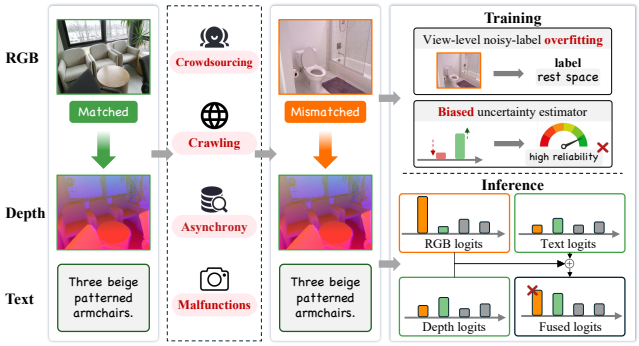


Figure 1: *Noisy correspondence in multiview learning*. Due to transmission disturbances, asynchronous acquisition, or sensor malfunctions, a multiview sample may contain partially mismatched views. In the example, the RGB view is replaced by an image from another instance, while the depth and text views remain correctly aligned and retain the original label. During training, such noisy correspondence corrupts cross-view consistency and induces a view-level noisy-label effect, which can lead to overfitting and biased reliability estimation. During inference, the mismatched inputs can further worsen the confusion in the learned biased model, leading to unreliable fusion and incorrect predictions.

and producing reliable predictions. CARF consists of two synergistic components: **Consensus Consistency Learning** (CCL) and **View Reliability Estimation** (VRE). Specifically: i) In CCL, we introduce a Hellinger-based objective that aligns each view with the fused predictive consensus. By using the relatively stable fused prediction as a soft consensus target and leveraging the robustness of the minimum-Hellinger objective, CCL improves cross-view consistency and stabilizes multiview fusion. ii) In VRE, we estimate the reliability of each view to identify mismatched correspondences through two complementary criteria: *view consistency*, which measures decision-level agreement across views, and *label consistency*, which measures fidelity to the assigned label. These two signals are jointly integrated into calibrated reliability weights with two roles: 1) reweighting view-specific objectives (*i.e.*, consistency learning and single-view classification) during training to suppress the influence of unreliable views; and 2) constructing reliability-aware fused predictions from view logits during both training and inference, thereby providing a more reliable consensus for consistency learning and more trustworthy final predictions.

Our contributions are summarized as follows:

(1) We identify and formalize *noisy correspondence* in multiview learning, a practical yet largely overlooked problem in which mismatched views corrupt cross-view consistency during both training and inference.
(2) We propose **CARF**, a unified framework for robust learning and reliable inference under noisy correspondence. CARF consists of two tightly coupled components: i) *Consensus Consistency Learning*, which promotes consistency among matched views; and ii) *View Reliability Estimation*, which quantifies the reliability of each view by jointly considering cross-view agreement and label fidelity.
(3) Extensive experiments on nine datasets demonstrate that CARF consistently outperforms state-of-the-art methods on both clean and noisy data.

2 Related Work

2.1 Multiview Learning

Multiview learning aims to exploit the consistency and complementarity across multiple views to improve prediction performance. Depending on whether predictive uncertainty is explicitly modeled, existing methods can be broadly categorized into *conventional multiview learning* and *trusted multiview learning*. i) Conventional multiview learning mainly improves predictive performance through better feature fusion, representation alignment, or consistency regularization. For example, MV-HFMD [4] introduces a hybrid fusion strategy and mutual distillation, while MAMC [29] alleviates feature heterogeneity and boundary ambiguity through multi-scale alignment and an expanded boundary. ii) Trusted multiview learning [14, 36, 44, 45] explicitly models uncertainty and has emerged as an important line of research for reliable multiview prediction. For instance, TMC [13] estimates view-specific evidence via Evidential Deep Learning [35] and performs uncertainty-aware fusion under Dempster-Shafer theory [8]. Subsequent methods further extend this framework from different perspectives. CCML [31] improves trusted fusion by dynamically decoupling consistent and complementary evidence, and TMCEK [28] enhances trusted multiview learning through expert knowledge constraints and a distribution-aware subjective opinion mechanism. More recently, FUML [10] departs from the conventional evidential framework by introducing Fuzzy Set Theory [52] to better characterize ambiguity and inter-view conflict. Despite their ability to handle contaminated or conflicting views at test time, trusted multiview learning methods typically assume that the views within each training sample are correctly aligned. Under noisy correspondence, however, mismatched views may occur not only at test time but also during training, where they corrupt cross-view consistency and introduce noisy supervision into view-specific learning. Consequently, these methods may overfit to spurious cross-view associations, which can in turn bias uncertainty estimation and undermine prediction reliability. In contrast, our CARF explicitly addresses noisy correspondence in both training and inference through consensus consistency learning and reliability estimation, thereby improving robustness.

2.2 Learning with Noisy Correspondence

Noisy correspondence generally refers to alignment errors in paired data, where mismatched samples are incorrectly treated as corresponding pairs, thereby causing spurious relation modeling and overfitting that degrade model performance. Existing methods improve robustness mainly through sample selection, robust objective design, and correspondence rectification. These ideas have been studied in cross-modal retrieval and matching [19, 20, 26, 34, 53], visible-infrared person re-identification [48], graph matching [30], vision-language pre-training [21], and clustering [16, 39, 54]. However, many existing noisy-correspondence methods are developed

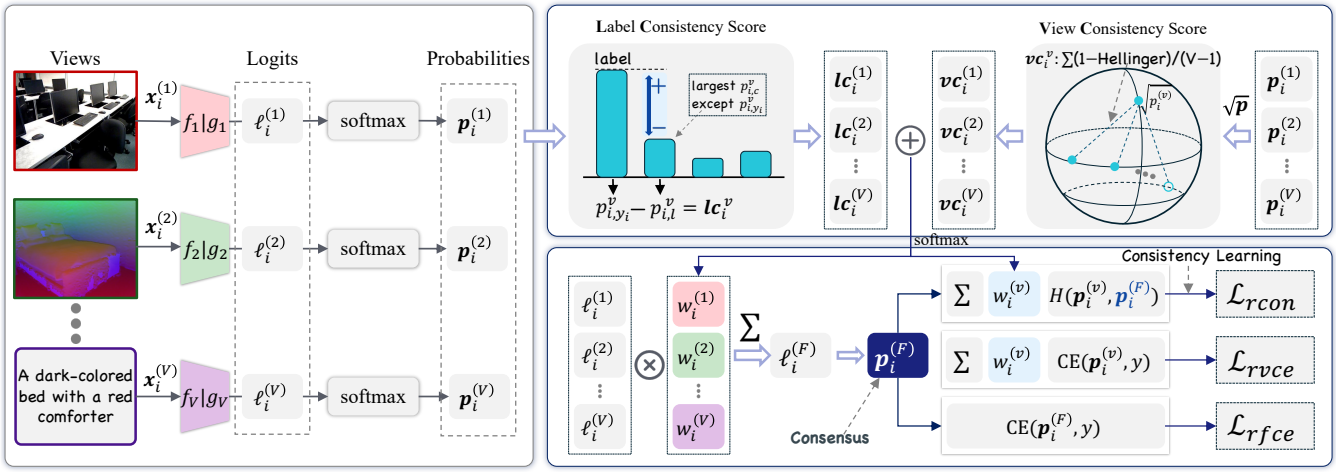


Figure 2: *Pipeline of CARF.* Given a multiview sample $X_i = (x_i^{(1)}, \cdots, x_i^{(V)}, y_i)$, view-specific encoders and classifier heads map each view to logits, which are converted to predictive distributions by softmax. CARF first introduces a Hellinger-based consistency objective $\mathcal{L}_{rcon}$ for view-to-fusion learning, promoting cross-view consistency and stable fusion. It then estimates reliability weights w_i from two complementary cues: *view consistency*, which measures decision-level agreement via Hellinger distance, and *label consistency*, which measures fidelity to the assigned label by the probability margin. The weights reweight view-specific objectives during training and enable reliability-aware weighted fusion in both training and inference.

for paired-data settings. This treatment is not well-suited to multiview learning with noisy correspondence, where corruption is often partial rather than holistic. Within a corrupted multiview sample, only some views may be mismatched, while the remaining views can still provide useful discriminative information. Therefore, simply downweighting or discarding an entire sample may remove informative views together with corrupted ones, leading to suboptimal classification performance. Furthermore, unlike conventional noisy-correspondence settings that primarily focus on training-time corruption, noisy correspondence in multiview learning can also arise at test time, where mismatched views continue to disrupt cross-view fusion and prediction. These characteristics make noisy correspondence in multiview learning a challenging problem that requires robust handling of view-level corruption during both training and inference. Adjacent studies on MLLMs also improve multimodal reliability through vMF-based trustworthiness estimation for hallucination detection and fuzzy semantic self-feedback for hallucination mitigation [5, 15]; however, they target hallucination rather than view-level correspondence noise.

3 Method

3.1 Problem Formulation

We consider a multiview classification task on a dataset $\mathcal{S} = \mathcal{S}^{tr} \cup \mathcal{S}^{te}$, where $\mathcal{S}^{tr}$ and $\mathcal{S}^{te}$ denote the training and test sets, respectively. Each sample is represented as

$$\boldsymbol{X}_i = (\boldsymbol{x}_i^{(1)}, \boldsymbol{x}_i^{(2)}, \cdots, \boldsymbol{x}_i^{(V)}, y_i), \quad y_i \in \{1, \ldots, C\}, \tag{1}$$

where $\boldsymbol{x}_i^{(v)}$ denotes the instance of the v-th view, V is the number of views, and y_i is the class label among C categories. Given the v-th view, a view-specific encoder f_v and classifier head g_v produce its latent representation and logits:

$$\boldsymbol{z}_i^{(v)} = f_v(\boldsymbol{x}_i^{(v)}), \quad \boldsymbol{\ell}_i^{(v)} = g_v(\boldsymbol{z}_i^{(v)}), \tag{2}$$

where $\boldsymbol{z}_i^{(v)} \in \mathbb{R}^D$ and $\boldsymbol{\ell}_i^{(v)} \in \mathbb{R}^C$. The fused logits are obtained by averaging the view-specific logits: $\boldsymbol{\ell}_i^{(F)} = \frac{1}{V}\sum_{v=1}^{V} \boldsymbol{\ell}_i^{(v)}$. Applying softmax yields the fused and view-specific class probabilities: $\boldsymbol{p}_i^{(\cdot)} = \text{softmax}(\boldsymbol{\ell}_i^{(\cdot)}) \in \mathbb{R}^C$. A standard multiview classifier is trained with cross-entropy losses on both the fused prediction and the view-specific predictions:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{fce} + \mathcal{L}_{vce} \\ &= -\frac{1}{B}\sum_{i=1}^{B} \log p_{i,y_i}^{(F)} - \frac{1}{B}\sum_{i=1}^{B}\frac{1}{V}\sum_{v=1}^{V} \log p_{i,y_i}^{(v)}, \end{aligned} \tag{3}$$

where B is the batch size, and $p_{i,y_i}^{(\cdot)}$ denotes the predicted probability of the ground-truth class y_i.

However, the above formulation assumes that the views within each sample are correctly aligned, an assumption that is often violated in real-world applications due to imperfect data construction pipelines or unstable real-time acquisition. To characterize this problem, we denote the underlying clean multiview sample, in which all views are correctly aligned by

$$\tilde{\boldsymbol{X}}_i = (\tilde{\boldsymbol{x}}_i^{(1)}, \tilde{\boldsymbol{x}}_i^{(2)}, \cdots, \tilde{\boldsymbol{x}}_i^{(V)}, y_i). \tag{4}$$

For each view, we define a binary correspondence indicator $m_i^{(v)} \in \{0, 1\}$, where $m_i^{(v)} = 1$ if $\boldsymbol{x}_i^{(v)} = \tilde{\boldsymbol{x}}_i^{(v)}$, and $m_i^{(v)} = 0$ otherwise. Here, $m_i^{(v)} = 0$ indicates that the observed $\boldsymbol{x}_i^{(v)}$ is a mismatched view originating from another instance. To exclude degenerate cases, we assume

$$\sum_{v=1}^{V}\left(1 - m_i^{(v)}\right) \le \left\lfloor \frac{V}{2} \right\rfloor, \tag{5}$$

for both training and test samples, so that each sample remains semantically meaningful.

Noisy correspondence introduces two major challenges. During training, mismatched views contaminate the fused prediction and simultaneously inject noisy supervision into the view-specific objective $\mathcal{L}_{vce}$, thereby degrading representation learning. During inference, the model can no longer rely on full agreement across views; instead, it must identify reliable views among mismatched ones, or aggregate the dominant evidence from the majority of consistent views. To address these challenges, we propose CARF, which tackles noisy correspondence through consensus consistency learning and view reliability estimation.

3.2 Consensus Consistency Learning

As discussed in the problem formulation, noisy correspondence disrupts cross-view consistency and makes naive view-wise alignment unreliable. Still, consistency across correctly matched views remains beneficial in multiview learning [46, 50, 51], as it regularizes view-specific predictions through multiview consensus, suppresses incidental view-specific cues, thereby improving fused-prediction stability and generalization [4]. However, when some views within a training sample are mismatched, directly enforcing agreement across all views may introduce additional noisy supervision and further distort the learned representation space.

Nevertheless, noisy correspondence does not entirely destroy the semantic information within a multiview instance. In many cases, some views remain correctly matched and still convey consistent class evidence, whereas mismatched views tend to be less coherent with the majority of views. From this perspective, the fused prediction can serve as a soft consensus target, since it is more likely to preserve the class evidence repeatedly supported across views. Based on this intuition, we introduce a robust Hellinger-based objective for view-to-fusion consensus consistency learning:

$$\mathcal{L}_{con} = \frac{1}{B}\sum_{i=1}^{B}\frac{1}{V}\sum_{v=1}^{V} H\left(\boldsymbol{p}_i^{(v)}, \boldsymbol{p}_i^{(F)}\right), \tag{6}$$

where

$$H(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{\sqrt{2}}\left\|\sqrt{\boldsymbol{p}} - \sqrt{\boldsymbol{q}}\right\|_2. \tag{7}$$

This objective encourages each view's prediction to align with the fused prediction (*i.e.*, consensus), thereby improving cross-view consistency.

Although the consensus may provide a biased target for mismatched views, this Hellinger-based objective avoids assigning excessively large penalties to such outliers, which is also supported by robust statistics, where minimum-Hellinger-based procedures are known to remain stable under contamination [2, 3, 18]. To further illustrate this point, we analyze the robustness of this Hellinger-based objective as follows.

i) Bounded risk. By definition, $H(\boldsymbol{p}, \boldsymbol{q}) = \frac{1}{\sqrt{2}}\|\sqrt{\boldsymbol{p}} - \sqrt{\boldsymbol{q}}\|_2$, and thus $H^2(\boldsymbol{p}, \boldsymbol{q}) = 1 - \langle\sqrt{\boldsymbol{p}}, \sqrt{\boldsymbol{q}}\rangle = 1 - \mathrm{BC}(\boldsymbol{p}, \boldsymbol{q})$, where

$$\mathrm{BC}(\boldsymbol{p}, \boldsymbol{q}) = \sum_{c=1}^{C}\sqrt{p_c q_c} \in [0, 1] \tag{8}$$

is the Bhattacharyya coefficient. Therefore, $0 \le H(\boldsymbol{p}, \boldsymbol{q}) \le 1$. Consequently, if m out of V views are corrupted by noisy correspondence, then the total contribution of these corrupted views to the per-sample consistency loss is bounded by

$$\frac{1}{V}\sum_{v\in\mathcal{M}} H\left(\boldsymbol{p}_i^{(v)}, \boldsymbol{p}_i^{(F)}\right) \le \frac{m}{V}, \tag{9}$$

where $\mathcal{M}$ denotes the set of mismatched views.

ii) Gradient attenuation on mismatched views. Treating the fused prediction $\boldsymbol{p}_i^{(F)}$ as a locally fixed target, consider the partial gradient of $H(\boldsymbol{p}_i^{(v)}, \boldsymbol{p}_i^{(F)})$ with respect to the logit $\ell_{i,y_i}^{(v)}$ corresponding to class y_i:

$$\frac{\partial H}{\partial \ell_{i,y_i}^{(v)}} = \frac{\mathrm{BC}(\boldsymbol{p}_i^{(v)}, \boldsymbol{p}_i^{(F)})\, p_{i,y_i}^{(v)} - \sqrt{p_{i,y_i}^{(v)} p_{i,y_i}^{(F)}}}{4H(\boldsymbol{p}_i^{(v)}, \boldsymbol{p}_i^{(F)})}. \tag{10}$$

Since $\mathrm{BC}(\boldsymbol{p}_i^{(v)}, \boldsymbol{p}_i^{(F)}) \in [0, 1]$, the condition $p_{i,y_i}^{(v)} < p_{i,y_i}^{(F)}$ is sufficient for the gradient to be negative. In this case, gradient descent increases $\ell_{i,y_i}^{(v)}$ and pushes the view prediction toward the fused consensus on class y_i. Although this introduces noisy supervision when $\boldsymbol{x}_i^{(v)}$ is mismatched, the magnitude of the induced gradient is attenuated. Specifically,

$$\left|\frac{\partial H}{\partial \ell_{i,y_i}^{(v)}}\right| \le \frac{p_{i,y_i}^{(v)} + \sqrt{p_{i,y_i}^{(v)} p_{i,y_i}^{(F)}}}{4H(\boldsymbol{p}_i^{(v)}, \boldsymbol{p}_i^{(F)})} \le \frac{\sqrt{p_{i,y_i}^{(v)}}}{2H(\boldsymbol{p}_i^{(v)}, \boldsymbol{p}_i^{(F)})}. \tag{11}$$

Therefore, for a severely mismatched view, which typically assigns vanishing probability to class y_i and remains far from the consensus, i.e., $p_{i,y_i}^{(v)} \to 0$ and $H(\boldsymbol{p}_i^{(v)}, \boldsymbol{p}_i^{(F)})$ is bounded away from zero, we have

$$\frac{\partial H}{\partial \ell_{i,y_i}^{(v)}} = O\left(\sqrt{p_{i,y_i}^{(v)}}\right) \to 0. \tag{12}$$

This shows that severely mismatched views are automatically assigned attenuated gradients.

iii) Geometric interpretation. Under the square-root mapping $\boldsymbol{p} \mapsto \sqrt{\boldsymbol{p}}$, probability vectors lie on the positive orthant of the unit sphere, and minimizing the Hellinger distance is equivalent to maximizing the Bhattacharyya overlap:

$$H^2(\boldsymbol{p}, \boldsymbol{q}) = 1 - \sum_{c=1}^{C}\sqrt{p_c q_c}. \tag{13}$$

Therefore, the proposed objective promotes consistency by encouraging alignment at the distribution level, rather than aggressively penalizing coordinate-wise discrepancies, making it less sensitive to isolated mismatched views.

3.3 View Reliability Estimation

Although $\mathcal{L}_{con}$ promotes consistency among matched views and improves fusion robustness, it still relies on the fused consensus being reasonably reliable. When noisy correspondence is severe, mismatched views may still distort the consensus and propagate noisy supervision through view-to-consensus alignment. Moreover, since CCL does not explicitly regulate the contribution of unreliable views to view-specific learning or final fusion, the model may still overfit noisy supervision during training and remain vulnerable to conflicting evidence at inference time. Therefore, we

further estimate the reliability of each view based on two complementary criteria, namely view consistency and label consistency, to determine whether a view is likely to be correctly matched.

View Consistency Score. The view consistency score $vc_i^{(v)}$ for the v-th view of the i-th sample is defined as

$$vc_i^{(v)} = \frac{1}{V-1} \sum_{k=1, k\neq v}^{V} \left(1 - H\left(\boldsymbol{p}_i^{(v)}, \boldsymbol{p}_i^{(k)}\right)\right). \tag{14}$$

Since the Hellinger distance measures the discrepancy between predictive distributions, the view consistency score quantifies the extent to which a view is supported by the remaining views in the same sample. Matched views typically provide more coherent class evidence and thus receive higher consistency scores, whereas mismatched views tend to receive less support from the others.

Label Consistency Score. We further define the label consistency score $lc_i^{(v)}$ for the v-th view of the i-th sample as

$$lc_i^{(v)} = p_{i,y_i}^{(v)} - p_{i,lar_i^{(v)}}^{(v)}, \tag{15}$$

where $p_{i,y_i}^{(v)}$ denotes the predicted probability of the ground-truth label y_i, and $lar_i^{(v)}$ is the index of the largest non-ground-truth probability:

$$lar_i^{(v)} = \arg\max_{c,\, c\neq y_i} p_{i,c}^{(v)}. \tag{16}$$

The label consistency score measures how well the prediction of $\boldsymbol{x}_i^{(v)}$ agrees with the assigned label, as well as its confidence margin over the most competitive incorrect class. Under noisy correspondence, reliable matched views tend to assign higher confidence to y_i and maintain a larger margin, whereas mismatched views often yield weaker or conflicting evidence. Therefore, this score provides an effective indicator of view reliability from the perspective of label supervision.

3.4 Reliability-Aware Training and Inference

The view consistency score and label consistency score provide complementary evidence for assessing whether a view is correctly matched: the former reflects cross-view agreement at the decision level, while the latter evaluates the fidelity of a view prediction to the assigned label.

During training, based on these two scores, we compute an overall reliability score for each view and normalize it with a softmax function to obtain the view weights $\boldsymbol{w}_i$ for the i-th instance:

$$\boldsymbol{w}_i = \mathrm{softmax}(\boldsymbol{vc}_i + \boldsymbol{lc}_i). \tag{17}$$

These weights are then used to construct a reliability-aware fused logit

$$\boldsymbol{\ell}_i^{(F)} = \sum_{v=1}^{V} w_i^{(v)} \boldsymbol{\ell}_i^{(v)}, \tag{18}$$

which provides a more reliable consensus $\boldsymbol{p}_i^{(F)} = \mathrm{softmax}(\boldsymbol{\ell}_i^{(F)})$ for both training and prediction. With this reliability-aware fusion, we further reduce the influence of mismatched views during optimization by reweighting both the single-view classification loss and the consensus consistency objective:

$$\begin{aligned} \mathcal{L}_{rfce} &= -\frac{1}{B}\sum_{i=1}^{B} \log(p_{i,y_i}^{(F)}), \\ \mathcal{L}_{rvce} &= -\frac{1}{B}\sum_{i=1}^{B}\sum_{v=1}^{V} w_i^{(v)} \log p_{i,y_i}^{(v)}, \\ \mathcal{L}_{rcon} &= \frac{1}{B}\sum_{i=1}^{B}\sum_{v=1}^{V} w_i^{(v)} H\left(\boldsymbol{p}_i^{(v)}, \boldsymbol{p}_i^{(F)}\right). \end{aligned} \tag{19}$$

Accordingly, the final training objective is

$$\mathcal{L} = \mathcal{L}_{rfce} + \mathcal{L}_{rvce} + \lambda \mathcal{L}_{rcon}. \tag{20}$$

During inference, only the view consistency score is available since the ground-truth label is unknown. We therefore estimate the view weights as:

$$\boldsymbol{w}_i' = \mathrm{softmax}(\boldsymbol{vc}_i), \tag{21}$$

and use them to form the fused logit for the final prediction. The complete training and inference procedures are summarized in algorithm 1.

Algorithm 1 Implementation of CARF

// Training Phase

Input: Training set $\mathcal{S}^{tr} = \{(\boldsymbol{x}_i^{(1)}, \cdots, \boldsymbol{x}_i^{(V)}, y_i)\}_{i=1}^{N}$, view encoders $\{f_v\}_{v=1}^{V}$, classifier heads $\{g_v\}_{v=1}^{V}$, warmup epoch E_w, total epochs E, hyperparameter λ.

Initialize epoch $e \leftarrow 0$.
while $e < E_w$ **do**
  Compute the view logits $\{\boldsymbol{\ell}_i^{(v)}\}$ and prediction probabilities $\{\boldsymbol{p}_i^{(v)}\}$ via eq. (2) and softmax.
  Set the view weights as $w_i^{(v)} = 1/V$ for all i and v.
  Update $\{f_v\}_{v=1}^{V}$ and $\{g_v\}_{v=1}^{V}$ via eq. (20)
  $e \leftarrow e + 1$.
end while
while $e < E$ **do**
  Compute the view consistency scores $\boldsymbol{vc}_i$ and label consistency scores $\boldsymbol{lc}_i$ via eqs. (14) and (15).
  Compute the reliability-aware weights $\boldsymbol{w}_i$ via eq. (17).
  Update $\{f_v\}_{v=1}^{V}$ and $\{g_v\}_{v=1}^{V}$ via eq. (20).
  $e \leftarrow e + 1$.
end while

Output: Trained encoders $\{f_v\}_{v=1}^{V}$ and classifier heads $\{g_v\}_{v=1}^{V}$.

// Inference Phase

Input: Trained encoders $\{f_v\}_{v=1}^{V}$, classifier heads $\{g_v\}_{v=1}^{V}$, test sample $\boldsymbol{x}_t = (\boldsymbol{x}_t^{(1)}, \cdots, \boldsymbol{x}_t^{(V)})$.

Compute the view consistency scores $\boldsymbol{vc}_t$ via eq. (14).
Set the view weights as $\boldsymbol{w}_t' = \mathrm{softmax}(\boldsymbol{vc}_t)$.
Compute fused logits as $\boldsymbol{\ell}_t^{(F)} = \sum_{v=1}^{V} w_t'^{(v)} \boldsymbol{\ell}_t^{(v)}$.
Output the predicted class $k = \arg\max_c \ell_{t,c}^{(F)}$.

Table 1: Classification Performance (Acc±Std) under varying noise ratios on eight feature-vector datasets over ten random seeds. The best and second-best results are highlighted in bold and underlined, respectively.

Noise	Method	Dataset							
		Caltech	Leaves	LandUse	Scene	CCV	Fashion	NUS-OBJ	AWA
0%	TMC (ICLR' 21)	95.96±1.12	95.59±0.85	46.14±1.82	72.47±0.91	44.58±0.81	96.07±0.31	39.72±0.51	36.87±0.44
	UIMC (CVPR' 23)	97.29±0.61	93.94±1.50	49.90±1.66	73.61±1.15	45.48±1.29	98.25±0.22	41.27±0.68	35.95±0.42
	ECML (AAAI' 24)	96.25±0.79	93.47±1.30	44.64±1.85	70.45±1.00	42.59±1.09	95.43±0.38	35.34±0.49	32.93±0.56
	CCML (ACM MM'24)	95.79±1.09	96.81±0.69	41.38±2.11	71.56±1.39	43.62±0.67	95.66±0.52	38.04±0.58	31.46±0.92
	MAMC (ICLR' 25)	97.82±0.72	89.88±1.55	71.67±1.60	80.71±1.28	54.66±1.26	98.53±0.28	45.84±0.57	32.15±0.83
	TMCEK (ICML' 25)	96.18±0.83	93.53±1.28	45.10±2.29	71.04±1.22	42.90±0.97	94.53±0.33	35.40±0.59	33.70±0.47
	FUML (ICML' 25)	95.29±0.93	99.28±0.63	74.95±1.71	78.92±1.68	47.38±2.43	98.09±0.33	47.14±0.56	38.30±0.43
	GTMC-HOA (TMM' 2026)	97.46±0.82	97.94±0.40	64.10±1.50	77.92±1.05	50.20±1.16	97.67±0.37	42.53±0.56	36.13±0.61
	CARF	**97.93±0.55**	**99.81±0.15**	**80.31±1.21**	**82.37±1.29**	**57.94±0.92**	**98.72±0.16**	**51.06±0.30**	**45.80±0.51**
50%	TMC (ICLR' 21)	83.54±1.87	78.47±2.14	39.38±2.05	62.22±1.31	38.47±0.53	83.28±0.70	33.89±0.27	25.96±0.65
	UIMC (CVPR' 23)	85.57±1.56	76.00±1.42	40.45±1.36	63.76±1.60	39.23±1.05	85.16±0.69	34.91±0.55	24.61±0.72
	ECML (AAAI' 24)	84.00±2.23	76.66±1.80	37.19±1.96	60.46±1.61	36.41±0.62	82.74±0.52	30.49±0.51	23.92±0.66
	CCML (ACM MM'24)	82.89±2.19	75.25±1.62	33.64±1.78	59.60±1.11	35.50±0.79	84.83±0.50	31.30±0.46	19.66±0.47
	MAMC (ICLR' 25)	94.93±1.21	66.81±1.73	59.02±1.83	70.02±2.04	45.54±1.30	94.08±0.33	37.76±0.62	16.10±0.73
	TMCEK (ICML' 25)	84.07±2.35	76.47±1.70	38.21±1.75	61.05±1.41	36.52±0.57	82.89±0.51	30.53±0.65	24.50±0.49
	FUML (ICML' 25)	93.11±1.21	91.69±0.63	64.55±1.98	70.94±2.12	35.37±13.27	92.10±0.58	40.39±0.58	20.13±8.56
	GTMC-HOA (TMM' 2026)	86.50±2.45	74.56±1.99	49.10±1.22	66.51±1.67	40.58±1.13	89.76±0.98	34.80±0.59	23.89±0.51
	CARF	**95.50±1.10**	**94.47±1.10**	**71.29±1.92**	**76.70±1.58**	**48.75±0.87**	**95.48±0.38**	**43.70±0.51**	**32.59±0.42**
100%	TMC (ICLR' 21)	68.32±2.70	60.06±2.22	31.86±1.32	52.07±1.17	32.10±1.48	67.91±1.14	28.79±0.30	15.58±0.36
	UIMC (CVPR' 23)	65.57±2.34	58.06±2.77	32.79±1.49	53.60±0.99	32.97±0.81	67.96±0.99	28.63±0.26	14.40±0.45
	ECML (AAAI' 24)	67.36±1.82	57.81±2.42	30.50±1.34	50.51±0.72	29.79±1.13	68.99±1.25	25.57±0.47	15.20±0.27
	CCML (ACM MM'24)	66.36±3.36	54.44±2.36	27.83±1.07	46.70±1.06	27.33±0.91	73.17±0.99	24.57±0.57	11.79±0.38
	MAMC (ICLR' 25)	86.11±1.15	49.13±2.32	45.76±2.43	61.58±2.02	37.14±0.94	90.40±0.55	31.43±0.36	5.08±0.25
	TMCEK (ICML' 25)	67.79±1.97	57.25±2.49	30.64±1.95	51.15±0.65	29.98±1.09	69.95±1.09	25.54±0.53	15.41±0.40
	FUML (ICML' 25)	84.29±3.55	79.31±2.94	49.76±11.02	64.80±1.43	25.79±13.46	80.94±0.74	34.13±0.32	1.93±0.21
	GTMC-HOA (TMM' 2026)	77.32±3.14	53.94±2.11	38.57±1.74	57.93±1.23	32.75±0.86	83.90±1.30	29.25±0.59	13.86±0.33
	CARF	**89.82±1.12**	**87.97±2.47**	**63.07±1.33**	**70.07±0.97**	**40.63±0.75**	**91.32±0.45**	**36.79±0.42**	**18.09±0.38**

Table 2: Classification Performance (Acc±Std) under varying noise ratios on SUN R-D-T over five random seeds. The best and second-best results are highlighted in bold and underlined, respectively.

Method	SUN R-D-T		
	0%	50%	100%
MAMC	62.91±0.86	52.33±1.02	37.20±1.73
TMCEK	63.55±0.62	53.12±0.30	41.47±0.95
FUML	64.01±0.60	55.61±0.50	42.92±0.37
GTMC-HOA	64.05±0.52	50.43±1.49	37.90±1.24
CARF	**66.81±0.23**	**59.61±0.30**	**48.68±0.58**

4 Experiments

4.1 Experimental Setup

Datasets and Metric. We evaluate CARF on nine datasets: Caltech [11], Leaves [40], LandUse [49], Scene [12], CCV [23], Fashion [42], NUS-OBJ [7], AWA [25] and SUN R-D-T [14]. SUN R-D-T is a three-view raw dataset constructed in [14], consisting of visual, depth, and textual views. Specifically, the visual and depth views are derived from RGB-D data [22, 37, 38, 43], while the textual view is a single-sentence description generated for each RGB image using Qwen3-VL-32B-Instruct [47]. For the eight feature-vector datasets, we adopt stratified 8:2 train-test splits. For SUN R-D-T, we follow the original train-test split of the underlying RGB-D benchmark. For all datasets, we report classification accuracy at the final epoch.

Noise Simulation. To evaluate the robustness of CARF, we simulate noisy correspondence in both the training and test sets with noise ratio σ. Specifically, for each split, we randomly sample a subset $\mathcal{S}_{sub}$ containing $\lfloor \sigma N \rfloor$ instances. For each instance in $\mathcal{S}_{sub}$, we randomly select $\lfloor V/2 \rfloor$ views and shuffle them across instances. As a result, each corrupted instance retains at least $\lceil V/2 \rceil$ views that remain correctly matched to the ground-truth label, corresponding to a practically meaningful contamination regime.

Implementation Details. We implement CARF in PyTorch 2.10.0 and conduct all experiments on a single NVIDIA RTX 4090 GPU with 24GB memory. The experimental settings are configured according to the dataset type. For feature-vector datasets, each encoder f_v consists of four linear layers with hidden dimension 1024 and three ReLU activations. CARF is trained for 200 epochs using Adam [24] with a batch size of 2048. The learning rate is scheduled by cosine decay with a peak value of 2×10^{-3}. For SUN R-D-T, we adopt an ImageNet-pretrained ResNet18 [17] for the visual and depth views, and BERT [9] for the textual view. Each classifier head is implemented as a linear layer. Input images are resized to

Table 3: Ablation study of the components in CARF under $\sigma = 50\%$. The best and second-best results are highlighted in bold and underlined, respectively.

No.	Combinations			Datasets					
	$\boldsymbol{vc}$	$\boldsymbol{lc}$	$\mathcal{L}_{rcon}$	Caltech	Leaves	Scene	CCV	Fashion	NUS-OBJ
#1				92.21±1.54	86.38±2.04	71.57±1.31	46.50±1.08	89.19±0.50	38.34±0.54
#2			✓	94.04±1.21	92.75±0.85	74.55±1.51	47.79±1.08	92.85±0.24	39.91±0.56
#3	✓	✓		93.54±0.92	91.00±1.12	<u>76.25±1.26</u>	48.39±0.69	93.48±0.55	<u>43.41±0.25</u>
#4	✓		✓	<u>94.82±1.33</u>	93.91±1.26	76.14±1.59	48.27±0.73	<u>94.88±0.58</u>	43.31±0.36
#5		✓	✓	94.54±1.11	<u>94.03±0.77</u>	76.01±1.70	<u>48.44±1.22</u>	93.16±0.29	41.32±0.44
#6	✓	✓	✓	**95.50±1.10**	**94.47±1.10**	**76.70±1.58**	**48.75±0.87**	**95.48±0.38**	**43.70±0.51**

256×256 and randomly cropped to 224×224. We train CARF for 100 epochs using AdamW [32] with a cosine annealing scheduler, weight decay of 0.01, initial learning rates of 10^{-4} for ResNet18 and 2×10^{-5} for BERT, and a batch size of 64. The hyperparameter λ is fixed to 0.2, and the warmup epoch $E_w = 5$ across all datasets.

4.2 Comparison with State-of-the-Art

To demonstrate the superiority of CARF, we compare it with representative multiview learning methods under varying noise ratios on the feature-vector datasets in table 1. The baselines include: (i) *trusted multiview learning methods*, which exploit uncertainty estimation to derive fusion weights, including TMC [13], UIMC [44], ECML [45], CCML [31], TMCEK [28], FUML [10], and GTMC-HOA [36]; and (ii) a representative *conventional multiview learning method*, MAMC [29], which improves representation learning without uncertainty-aware fusion. Furthermore, we compare CARF with MAMC [29], TMCEK [28], FUML [10], and GTMC-HOA [36] on the multimodal dataset SUN R-D-T in table 2. Since these methods were originally evaluated on feature-vector datasets, we adapt them to SUN R-D-T using the same view encoders as CARF (*i.e.*, ResNet18 and BERT). CARF achieves the best performance across all evaluated noise levels on all datasets. We highlight two observations below:

- On the feature-vector datasets, CARF consistently outperforms existing state-of-the-art methods on all datasets, even without synthetic noisy correspondence (*i.e.*, $\sigma = 0\%$). For example, CARF surpasses FUML by 5.36% on LandUse and by 7.50% on AWA. This suggests that the proposed consensus consistency learning and reliability estimation also help CARF better cope with intrinsic noise in multiview data.
- On SUN R-D-T, CARF maintains clear superiority. As the noise ratio increases, CARF exhibits stronger robustness, and the performance gap between CARF and competing methods becomes larger. For example, when σ increases from 0% to 50%, the strongest baseline, GTMC-HOA, drops from 64.05 to 50.43 (−13.62), whereas CARF decreases only from 66.81 to 59.61 (−7.20). Moreover, the performance gap between CARF and FUML further widens from 2.80 to 4.00. These results indicate that, on SUN R-D-T, CARF achieves superior robustness under noisy correspondence by combining a robust Hellinger-based objective for consistency learning with reliability estimation to suppress unreliable views in both representation learning and prediction fusion.

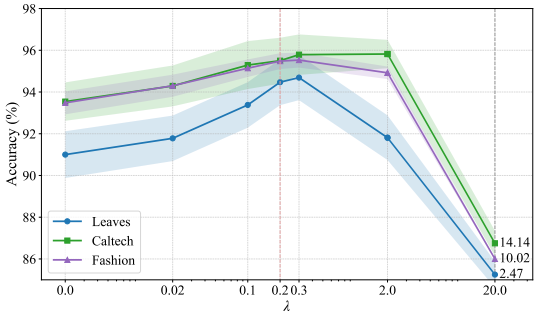


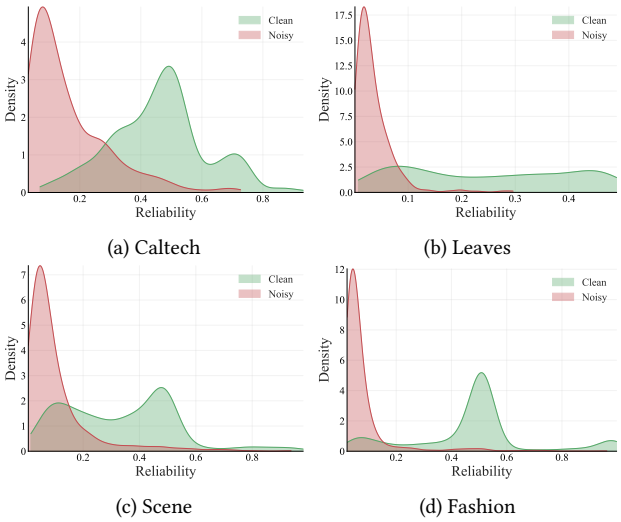
Figure 3: Performance variation with respect to λ on the Caltech, Leaves, and Fashion datasets. The default value is indicated by a red dashed line. The accuracies at $\lambda = 20.0$ are annotated in the figure because they fall outside the displayed range.

4.3 Ablation

We conduct ablation studies on Caltech, Leaves, Scene, CCV, Fashion, and NUS-OBJ under noisy correspondence with $\sigma = 50\%$ to quantify the contribution of each component in CARF. Specifically, we evaluate the effects of: (i) the consensus consistency learning objective $\mathcal{L}_{rcon}$; and (ii) the two reliability scores, *i.e.*, the view consistency score $\boldsymbol{vc}$ and the label consistency score $\boldsymbol{lc}$. The results are reported in table 3. Overall, each component consistently improves performance across datasets. The observations are summarized below.

Effect of $\mathcal{L}_{rcon}$. The objective $\mathcal{L}_{rcon}$ consistently improves performance, both with and without reliability estimation. For example, comparing #1 with #2, introducing $\mathcal{L}_{rcon}$ improves accuracy on Leaves from 86.38 to 92.75 (+6.37). Comparing #3 with #6, it further improves accuracy from 93.54 to 95.50 (+1.96) on Caltech. These results indicate that $\mathcal{L}_{rcon}$ effectively promotes robust view consistency under noisy correspondence, thereby improving fusion stability and generalization.

Effect of $\boldsymbol{vc}$. Comparing #5 with #6, introducing $\boldsymbol{vc}$ improves accuracy from 76.01 to 76.70 (+0.69) on Scene and from 93.16 to 95.48 (+2.32) on Fashion. This suggests that $\boldsymbol{vc}$ effectively measures view reliability from cross-view predictive agreement.



(a) Caltech (b) Leaves

(c) Scene (d) Fashion

Figure 4: Density plots of the reliability scores on the Caltech, Leaves, Scene, and Fashion test sets.

Effect of $\boldsymbol{lc}$. Comparing #4 with #6, introducing $\boldsymbol{lc}$ improves accuracy from 43.31 to 43.70 (+0.39) on NUS-OBJ and from 48.27 to 48.75 (+0.48) on CCV. This indicates that $\boldsymbol{lc}$, which reflects the margin to the most competitive incorrect class, provides useful reliability information under noisy correspondence.

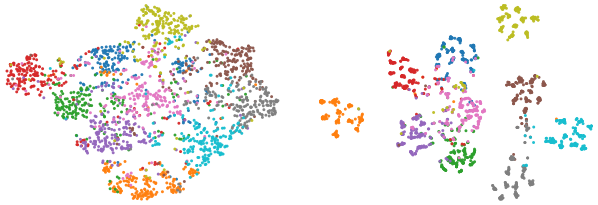
Joint effect of $\boldsymbol{vc}$ *and* $\boldsymbol{lc}$. The two scores are complementary and yield further gains when used together. For example, comparing #1 with #3, jointly introducing $\boldsymbol{vc}$ and $\boldsymbol{lc}$ improves accuracy on Fashion by 4.29. Moreover, when one score is already used, adding the other still brings additional improvement. This shows that $\boldsymbol{vc}$ and $\boldsymbol{lc}$ capture different yet complementary cues for estimating view reliability.

4.4 Parameter Analysis

To evaluate the robustness of CARF to the optimization configuration, we analyze its sensitivity to the hyperparameter λ, which controls the strength of consensus consistency learning. The results on Caltech, Leaves, and Fashion under noisy correspondence ($\sigma = 50\%$) are shown in fig. 3. Across all datasets, the performance exhibits an inverted-U trend as λ varies, indicating that a moderate value effectively promotes view consistency and improves classification performance. CARF remains relatively stable over a broad range of $\lambda \in [0.02, 2.0]$. However, when λ becomes excessively large ($\lambda = 20.0$), the consistency objective dominates label supervision, leading to a sharp performance drop.

4.5 Visualization

Figure 4 shows that clean and noisy views form clearly separated reliability-score distributions, indicating that the proposed reliability estimation can identify mismatched views. Figure 5 further shows that reliability estimation yields more compact and separable fused logits, and fig. 6 confirms that mismatched views receive lower reliability scores in representative SUN R-D-T examples.



(a) W/O $\boldsymbol{w}_i$ (Acc=78.50%) (b) W/ $\boldsymbol{w}_i$ (Acc=91.10%)

Figure 5: t-SNE visualization of fused logits on the Fashion test set under noisy correspondence ($\sigma = 100\%$), comparing models with and without view reliability estimation.

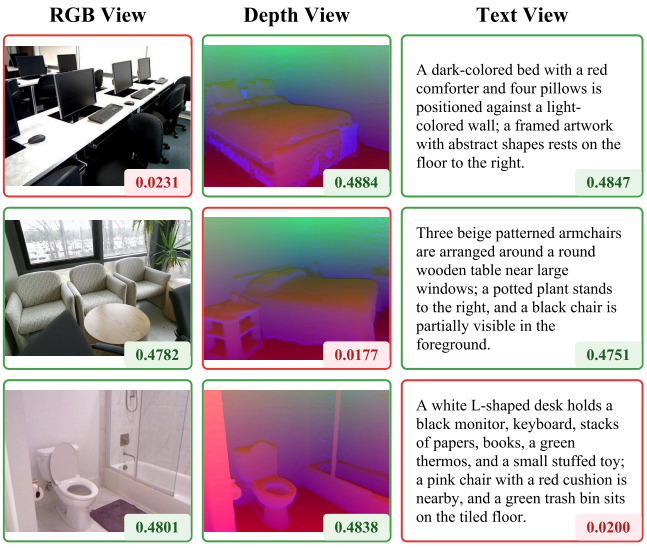


Figure 6: Case study of view reliability estimation on the SUN R-D-T test set under noisy correspondence ($\sigma = 100\%$). The red box indicates the mismatched view, and the estimated reliability score of each view is shown in the bottom-right corner.

5 Conclusion

In this work, we revisit and study the challenging problem of noisy correspondence in multiview learning, which violates the common assumption in existing methods that views are correctly aligned in both training and inference. To address this issue, we propose **CARF**, a consistency-aware robust fusion framework for learning and inferring under noisy correspondence. By combining consensus consistency learning with view reliability estimation, CARF effectively suppresses the influence of mismatched views and yields more reliable multiview predictions. Extensive experiments on eight feature-vector datasets and one raw dataset demonstrate that **CARF** consistently achieves superior performance and stronger robustness on both clean and noisy data.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2024YFB4710604; in part by the National Natural Science Foundation of China under Grants U25B6003, U25A20534, and 62472295; in part by the Fundamental Research Funds for the Central Universities under Grants CJ202403 and CJ202303; in part by Sichuan Science and Technology Planning Project under Grant 24NSFTD0130; in part by the Luzhou City School-Local-Enterprise-Academy Science and Technology Cooperation Project under Grant 2024XDY200; in part by the National Social Science Foundation under Grant 24BGL131; in part by the Sichuan Science and Technology Program under Grant 2024NSFSC0521; and in part by the Chengdu Science and Technology Bureau Project under Grant 2025-YF05-00391-SN.

References

[1] Joshua P Barrios, Minhaj U Ansari, Jeffrey E Olgin, Sean Abreau, Jacques Delfrate, Elodie L Langlais, Robert Avram, and Geoffrey H Tison. 2026. Multiview deep learning improves detection of major cardiac conditions from echocardiography. *Nature Cardiovascular Research* 5, 3 (2026), 234–245.

[2] Ayanendranath Basu and Bruce G Lindsay. 1994. Minimum disparity estimation for continuous models: efficiency, distributions and robustness. *Annals of the Institute of Statistical Mathematics* 46, 4 (1994), 683–705.

[3] Rudolf Beran. 1977. Minimum Hellinger distance estimates for parametric models. *The annals of Statistics* (1977), 445–463.

[4] Samuel Black and Richard Souvenir. 2024. Multi-view classification using hybrid fusion and mutual distillation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 270–280.

[5] Kaiqi Chen, Yang Qin, Changhao He, Xi Peng, and Peng Hu. 2026. DOUBT: Decoupled Object-level Understanding and Bridging via vMF-based Trustworthiness for Hallucination Detection in MLLMs. In *Forty-third International Conference on Machine Learning*.

[6] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. 2017. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 1907–1915.

[7] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. 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*. 1–9.

[8] Arthur P Dempster. 2008. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions*. Springer, 57–72.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[10] Siyuan Duan, Yuan Sun, Dezhong Peng, Guiduo Duan, Xi Peng, and Peng Hu. 2025. Deep fuzzy multi-view learning for reliable classification. In *Forty-second International Conference on Machine Learning*.

[11] Li Fei-Fei, Rob Fergus, and Pietro Perona. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*. IEEE, 178–178.

[12] Li Fei-Fei and Pietro Perona. 2005. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, Vol. 2. IEEE, 524–531.

[13] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2021. Trusted Multi-View Classification. In *International Conference on Learning Representations*.

[14] Changhao He, Di Xue, Shuxian Li, Yanji Hao, Xi Peng, and Peng Hu. 2026. Bootstrapping Multi-view Learning for Test-time Noisy Correspondence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*.

[15] Changhao He, Shuhao Yan, Shuxian Li, Xi Peng, and Peng Hu. 2026. RLSF-V: Mitigating Hallucinations in MLLMs via Fuzzy Semantic Self-Feedback. In *Forty-third International Conference on Machine Learning*.

[16] Changhao He, Hongyuan Zhu, Peng Hu, and Xi Peng. 2024. Robust Variational Contrastive Learning for Partially View-unaligned Clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 4167–4176.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[18] Giles Hooker and Anand N Vidyashankar. 2014. Bayesian model robustness via disparities. *Test* 23, 3 (2014), 556–584.

[19] Zhenyu Huang, Peng Hu, Guocheng Niu, Xinyan Xiao, Jiancheng Lv, and Xi Peng. 2024. Learning with Noisy Correspondence. *International Journal of Computer Vision* (2024), 1–22.

[20] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. 2021. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems* 34 (2021), 29406–29419.

[21] Zhenyu Huang, Mouxing Yang, Xinyan Xiao, Peng Hu, and Xi Peng. 2024. Noise-robust Vision-language Pre-training with Positive-negative Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2024), 1–13. doi:10.1109/TPAMI.2024.3462996

[22] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T Barron, Mario Fritz, Kate Saenko, and Trevor Darrell. 2011. A category-level 3-d object dataset: Putting the kinect to work. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. IEEE, 1168–1174.

[23] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C Loui. 2011. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proceedings of the 1st ACM international conference on multimedia retrieval*. 1–8.

[24] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[25] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. 2013. Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence* 36, 3 (2013), 453–465.

[26] Shuxian Li, Changhao He, Xiting Liu, Joey Tianyi Zhou, Xi Peng, and Peng Hu. 2025. Learning with noisy triplet correspondence for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 19628–19637.

[27] Sizhe Lester Li, Annan Zhang, Boyuan Chen, Hanna Matusik, Chao Liu, Daniela Rus, and Vincent Sitzmann. 2025. Controlling diverse robots by inferring Jacobian fields with deep networks. *Nature* 643, 8070 (2025), 89–95.

[28] Xinyan Liang, Shijie Wang, Yuhua Qian, Qian Guo, Liang Du, Bingbing Jiang, Tingjin Luo, and Feijiang Li. 2025. Trusted multi-view classification with expert knowledge constraints. In *Forty-second International Conference on Machine Learning*.

[29] Yuena Lin, Yiyuan Wang, Gengyu Lyu, Yongjian Deng, Haichun Cai, Huibin Lin, Haobo Wang, and Zhen Yang. 2025. Enhance multi-view classification through multi-scale alignment and expanded boundary. In *The Thirteenth International Conference on Learning Representations*.

[30] Yijie Lin, Mouxing Yang, Jun Yu, Peng Hu, Changqing Zhang, and Xi Peng. 2023. Graph matching with bi-level noisy correspondence. In *Proceedings of the IEEE/CVF international conference on computer vision*. 23362–23371.

[31] Ying Liu, Lihong Liu, Cai Xu, Xiangyu Song, Ziyu Guan, and Wei Zhao. 2024. Dynamic evidence decoupling for trusted multi-view learning. In *Proceedings of the 32nd ACM international conference on multimedia*. 7269–7277.

[32] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[33] Xuejun Qian, Jing Pei, Hui Zheng, Xinxin Xie, Lin Yan, Hao Zhang, Chunguang Han, Xiang Gao, Hanqi Zhang, Weiwei Zheng, et al. 2021. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. *Nature biomedical engineering* 5, 6 (2021), 522–532.

[34] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. 2024. Noisy-correspondence learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 27197–27206.

[35] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems* 31 (2018).

[36] Long Shi, Chuanqing Tang, Huangyi Deng, Cai Xu, Lei Xing, and Badong Chen. 2026. Generalized trusted multi-view classification framework with hierarchical opinion aggregation. *IEEE Transactions on Multimedia* (2026).

[37] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*. Springer, 746–760.

[38] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 567–576.

[39] Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. 2024. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering* (2024).

[40] Hao Wang, Linlin Zong, Bing Liu, Yan Yang, and Wei Zhou. 2019. Spectral perturbation meets incomplete multi-view data. *arXiv preprint arXiv:1906.00098* (2019).

[41] Jinbo Wang, Weihua Xu, Qinghua Zhang, Yuhua Qian, and Weiping Ding. 2025. Dynamic Multi-View Classification and Knowledge Fusion: A Fuzzy Concept-Cognitive Learning Perspective. *IEEE Transactions on Fuzzy Systems* (2025).

[42] Han Xiao, Kashif Rasul, and Roland Vollgraf. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* (2017).

[43] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. 2013. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision*. 1625–1632.

[44] Mengyao Xie, Zongbo Han, Changqing Zhang, Yichen Bai, and Qinghua Hu. 2023. Exploring and exploiting uncertainty for incomplete multi-view classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 19873–19882.

[45] Cai Xu, Jiajun Si, Ziyu Guan, Wei Zhao, Yue Wu, and Xiyue Gao. 2024. Reliable conflictive multi-view learning. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 16129–16137.

[46] Chang Xu, Dacheng Tao, and Chao Xu. 2013. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634* (2013).

[47] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388* (2025).

[48] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. 2022. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14308–14317.

[49] Yi Yang and Shawn Newsam. 2010. 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*. 270–279.

[50] Yang Yang, Fengqiang Wan, Qing-Yuan Jiang, and Yi Xu. 2024. Facilitating multimodal classification via dynamically learning modality gap. *Advances in Neural Information Processing Systems* 37 (2024), 62108–62122.

[51] Zhiwen Yu, Ziyang Dong, Chenchen Yu, Kaixiang Yang, Ziwei Fan, and CL Philip Chen. 2025. A review on multi-view learning. *Frontiers of Computer Science* 19, 7 (2025), 197334.

[52] Lotfi A Zadeh. 1965. Fuzzy sets. *Information and control* 8, 3 (1965), 338–353.

[53] Xu Zhang, Hao Li, and Mang Ye. 2024. Negative Pre-aware for Noisy Cross-Modal Matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 7341–7349.

[54] Haochen Zhou, Guofeng Ding, Mouxing Yang, Peng Hu, Yijie Lin, and Xi Peng. 2026. Uncover Underlying Correspondence for Robust Multi-view Clustering. In *The Fourteenth International Conference on Learning Representations*.