# Bootstrapping Multi-view Learning for Test-time Noisy Correspondence

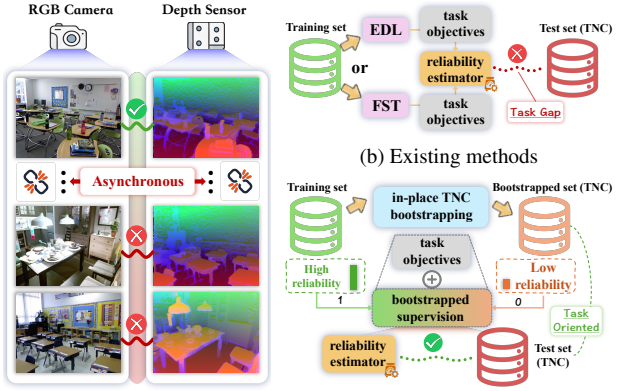
Changhao He[1], Di Xue[2], Shuxian Li[1], Yanji Hao[2], Xi Peng[1,3], Peng Hu[1]*
[1]Sichuan University, China [2]AVIC Chengdu Aircraft Design & Research Institute, China
[3]National Key Laboratory of Fundamental Algorithms and Models for Engineering Simulation, China
https://github.com/XLearning-SCU/2026-CVPR-BML

## Abstract

*Multi-view learning fuses complementary views to improve perception, but real-world deployments often suffer from Test-time Noisy Correspondence (TNC) — cross-view misalignment caused by asynchronous sampling, transient network congestion, or other disturbances. Such misalignment introduces semantic inconsistency and significantly degrades performance. Existing remedies typically estimate view-specific reliability from clean, well-aligned training data and then extrapolate to noisy fusion at inference, resulting in a train-test task gap and reduced robustness against TNC. To bridge this gap, we propose **B**ootstrapping **M**ulti-view **L**earning (BML) — a plug-and-play framework that explicitly learns to fuse under TNC. Specifically, BML performs in-place TNC bootstrapping to construct a controllable noise-augmented training set that simulates realistic correspondence distortion, thereby eliminating the task gap without external data. Unlike prior uncertainty-based approaches that model reliability in an unsupervised manner, BML presents a reveal-supervised paradigm, wherein a lightweight estimator jointly models intra-view predictive uncertainty (view quality) and inter-view prediction discrepancy (correspondence consistency) to produce calibrated reliability weights guided by both task objectives and bootstrapped supervision. Once deployed, these reliability weights directly modulate fusion, suppressing corrupted views while preserving informative ones. Across 11 benchmarks spanning diverse noise ratios, BML consistently outperforms state-of-the-art baselines and maintains robustness against TNC.*

## 1. Introduction

Multi-view learning [4, 32, 72] integrates complementary information across views/modalities (*e.g.*, images, text, video, *etc.*) to improve perception, robustness, and decision-making in various applications, such as autonomous driving [5, 6], embodied intelligence [21, 42], and medical diagnostics [3]. In practice, however, real-world deployments frequently encounter Test-time Noisy Correspondence (TNC), where asynchronous sampling [41], transient network congestion [59], and other disturbances introduce cross-view misalignment during inference as illustrated in Figure 1(a). This misalignment inevitably yields semantic inconsistencies across misaligned views, thereby significantly degrading model performance.




(a) TNC scenarios (b) Existing methods (c) Our BML

Figure 1. (a) In real-world deployments, sensors are prone to suffer from asynchronous sampling, transient network congestion, and other disturbances, resulting in **Test-time Noisy Correspondence (TNC)**. (b) Existing methods (*e.g.*, Evidential Deep Learning (EDL) or Fuzzy Set Theory (FST)) typically infer reliability from a clean training set under task objectives but lack explicit supervision, leading to blind estimation. (c) BML addresses TNC from both data and model perspectives, where data is augmented via in-place TNC bootstrapping and model is optimized with both task objectives and bootstrapped supervision.

To mitigate this problem, many prior works model predictive uncertainty via evidential deep learning [33, 44]. For example, evidential learning-based methods such as TMC [18] aggregate evidence across views to parameterize class distributions [45], and ECML [64] adopts a tailored robust aggregation strategy to down-weight unreliable opinions from individual views during fusion. More recently, several works [11, 55] step outside the evidential framework and employ fuzzy set

*Corresponding author: Peng Hu (penghu.ml@gmail.com).

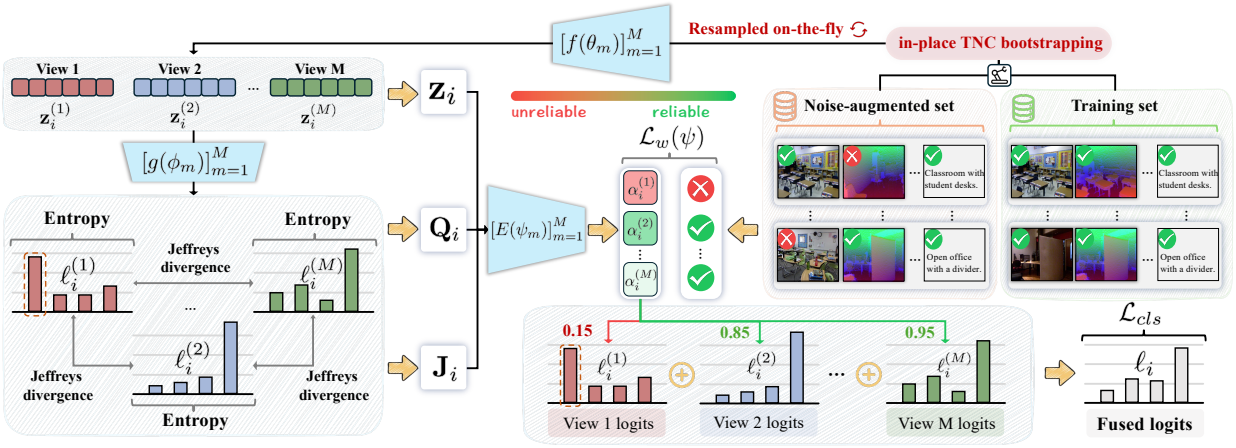



Figure 2. Pipeline of **Bootstrapping Multi-view Learning (BML)**. Starting from a clean, off-the-shelf training set, BML generates noise-augmented data via in-place TNC bootstrapping and interleaves it with the clean data during training. With known noise locations, BML learns a lightweight estimator $E_\psi$ that integrates view-specific features $\mathbf{z}_i$, inter-view predictive discrepancy $\mathbf{J}_i$, and intra-view predictive uncertainty $\mathbf{Q}_i$ to perform robust test-time fusion.

theory [71] to reduce the sensitivity of uncertainty modeling to the total amount of evidence and dataset-specific characteristics. Despite their effectiveness, these methods suffer from two key limitations: First, they learn reliability weights from clean, well-aligned training data, yet are applied to noisy, misaligned inputs at test time, thereby leading to a *train-test task gap* that hampers robustness under TNC. Second, they typically infer reliability indirectly without explicit supervision, often resulting in overconfident or miscalibrated weights under TNC.

To overcome the aforementioned limits, we propose a plug-and-play Bootstrapping Multi-view Learning (BML) framework, which adopts a reveal-supervised paradigm and bridges the task gap from both the data and model perspectives. Specifically, on the data side, BML performs in-place TNC bootstrapping to construct a controllable noise-augmented set that explicitly simulates downstream cross-view misalignment, which is interleaved with the original clean data during training. On the model side, BML trains a lightweight reliability estimator under task constraints as well as bootstrapped supervision, which ingests both intra-view predictive uncertainty (view quality) and inter-view prediction discrepancy (correspondence consistency) to produce calibrated reliability weights. At test time, the estimated weights directly modulate multi-view fusion to suppress misalignment and enhance robustness against TNC. Our main contributions can be summarized as follows:

- We identify and formalize **Test-time Noisy Correspondence (TNC)** as a practical yet overlooked challenge in multi-view learning, where cross-view misalignment arises during inference rather than training, revealing a critical train-test task gap in existing fusion frameworks.
- We propose **in-place TNC bootstrapping** to inject controllable cross-view misalignment into clean samples and record the corrupted views. This produces noise-augmented data that effectively aligns training objectives with test-time conditions, bridging the task gap without requiring extra data or annotations.
- We present a **reveal-supervised paradigm** with a lightweight reliability estimator that unifies intra-view predictive uncertainty and inter-view prediction discrepancy into calibrated reliability weights. Unlike prior unsupervised uncertainty-driven weighting, our estimator is directly optimized with both task objectives and bootstrapped supervision, yielding stable and interpretable fusion under TNC.
- Extensive experiments across 11 benchmarks under varying noise ratios demonstrate that BML consistently improves robustness against TNC while preserving performance on clean data.

## 2. Related Work

### 2.1. Multi-view Learning

Multi-view learning has garnered significant attention due to its ability to integrate complementary information from different views [2, 17, 49, 56, 67]. Depending on whether per-view uncertainty is explicitly estimated, existing methods typically fall into two categories: i) deterministic multi-view learning methods [19, 36, 72], which exploit complementarity by improving representation quality or cross-view alignment. For example, MAMC [37] mitigates feature heterogeneity and reduces information redundancy to enhance generalization, while RML [66] proposes a perturbation-simulated multi-view contrastive framework that achieves both representation alignment and

fusion. However, these methods often assume clean cross-view correspondences at test time, which renders them brittle when noisy correspondence occurs after deployment [38, 64]. ii) trustworthy multi-view learning methods [18], which enable uncertainty-aware weighting at test time [35]. Here, uncertainty is commonly modeled via Evidential Deep Learning (EDL) [44] and aggregated using Dempster-Shafer theory [9] or related fusion rules [38, 40, 65]. More recently, FUML [11] replaces EDL with Fuzzy Set Theory (FST) [71] to reduce the sensitivity of uncertainty estimation to the number of classes. Despite these advances, most trustworthy methods infer uncertainty relying solely on clean, aligned training data, lacking explicit exposure to mismatched samples, which inevitably introduces a *task gap* and undermines reliability. In contrast, we propose a reveal-supervised paradigm that uses a noise-augmented set and noisy correspondences as supervision, thereby bridging the gap between training and test sets and improving robustness in the presence of TNC.

### 2.2. Noisy Correspondence Problem

The Noisy Correspondence (NC) problem [22] has emerged as a distinct paradigm for handling imperfect pairings across views or modalities in real-world scenarios, which has been investigated across a wide range of tasks, including person re-identification (ReID) [43, 46, 63, 69], image-text Matching [8, 12, 13, 51, 57], vision-language pre-training [23, 30], composed image retrieval [31, 47], *etc*. Typically, most existing methods treat correspondence noise primarily as a *training-time phenomenon*, focusing on robust objectives [52, 53], sample reweighting [73], or correspondence rectification [22] during the training phase. However, in practice, factors such as sensor asynchrony [41], calibration drift [1], domain shift [26], *etc*., often make NC inevitable at test time, which has rarely been explored before. This common misalignment undermines cross-view alignment during inference, degrading decision quality and reliability [38, 64]. To address this gap, we introduce a simple yet effective bootstrapped reliability estimation framework for test-time NC, which complements prior NC work that operates only at training time, thus extending robustness to the deployment setting.

## 3. Method

### 3.1. Problem Formulation

Considering an $M$-view classification task with label space $\mathcal{Y} = \{1, \dots, C\}$, given a sample index $i$, let $\mathcal{X}_i = \{\mathbf{x}_i^{(m)}\}_{m=1}^{M}$ denote the set of view-specific inputs and $y_i \in \mathcal{Y}$ be the ground-truth label. Each view $m$ is processed by an encoder–classifier pair $[f(\cdot;\theta_m),\, g(\cdot;\phi_m)]$ that produces a feature representation $\mathbf{z}_i^{(m)} = f(\mathbf{x}_i^{(m)};\theta_m) \in \mathbb{R}^{d_m}$, class logits $\boldsymbol{\ell}_i^{(m)} = g(\mathbf{z}_i^{(m)};\phi_m) \in \mathbb{R}^C$, and the corresponding predictive distribution $\mathbf{p}_i^{(m)} = \mathrm{softmax}(\boldsymbol{\ell}_i^{(m)})$. However, at test time, sensor asynchrony or link blockage is prone to induce noisy correspondence, where one or more views in $\mathcal{X}_i$ no longer correspond to $y_i$. Formally, we define the test-time noisy correspondence scenario as follows:

**Definition 1 (Test-time Noisy Correspondence (TNC)).** *An instance $\mathcal{X}_i$ is said to suffer from TNC if a nonempty subset of its views is misaligned at test time. Mathematically, let*

$$k_i = \sum_{m=1}^{M} \mathbb{I}\big[\mathbf{x}_i^{(m)} \not\leftrightarrow y_i\big] \tag{1}$$

*denote the number of misaligned views, where $\mathbb{I}[\cdot]$ is the indicator function and $\mathbf{x}_i^{(m)} \not\leftrightarrow y_i$ indicates that view $m$ does not correspond to the ground-truth label $y_i$. We focus on the regime:*

$$1 \le k_i \le \big\lfloor \tfrac{M}{2} \big\rfloor, \tag{2}$$

*so that a majority of views remain correctly aligned.*

To achieve robust inference under TNC, we perform late fusion using nonnegative, sample- and view-specific coefficients. Specifically, the fused logits and final prediction are given by:

$$\bar{\boldsymbol{\ell}}_i = \sum_{m=1}^{M} \alpha_i^{(m)}\, \boldsymbol{\ell}_i^{(m)}, \qquad \bar{y}_i = \arg\max_{c \in \mathcal{Y}} \bar{\boldsymbol{\ell}}_{i,c}, \tag{3}$$

where $\alpha_i^{(m)}$ is the view reliability produced by a view-specific estimator $E(\cdot;\psi_m)$. This generator is trained via a reveal-supervised paradigm on an in-place augmented set containing simulated misalignment, allowing it to learn how to detect and down-weight noisy views. We will elaborate on this learning mechanism in the following subsections.

### 3.2. Bootstrapped Reveal-Supervised Learning

To obtain view reliability $\alpha_i^{(m)}$ for robust fusion, heuristic weighting or purely uncertainty-based solutions often require manual tuning or take a task gap between training and test sets, yielding suboptimal decision quality. We address this by learning reliability with a controllable in-place TNC-augmented set, where the noise introduced by TNC is naturally revealed to the estimator as supervision. To prevent overfitting to fixed corruption patterns and improve coverage of plausible misalignments, the augmentation is resampled on-the-fly in a bootstrapped manner.

**Bootstrapped in-place augmentation.** To start with, at the beginning of each training epoch, given the clean training set:

$$\mathcal{D} = \{(\mathcal{X}_i, y_i)\}_{i=1}^{N}, \quad \mathcal{X}_i = \{\mathbf{x}_i^{(m)}\}_{m=1}^{M}, \tag{4}$$

we first sample an index subset $\widetilde{S} \subseteq \{1, \ldots, N\}$ with $|\widetilde{S}| = |\rho N|$, where $\rho \in (0, 1]$ denotes the augmented rate. To bridge the train-test gap, for each $i \in \widetilde{S}$, we draw a view-level mask $\mathbf{s}_i = (s_i^{(1)}, \ldots, s_i^{(M)}) \in \{0, 1\}^M$ and constrain it by the TNC regime:

$$1 \leq \sum_{m=1}^{M} s_i^{(m)} \leq \lfloor \tfrac{M}{2} \rfloor, \tag{5}$$

which ensures that at least one and at most half of the views are corrupted, thereby preserving identifiability through remaining clean views. After that, we corrupt view $m$ whenever $s_i^{(m)} = 1$ by shuffling its input within the bootstrapped pool $\widetilde{S}$:

$$\breve{\mathbf{x}}_i^{(m)} \leftarrow \begin{cases} \mathbf{x}_j^{(m)}, & \text{if } s_i^{(m)} = 1, \ \ j \in \widetilde{S}, \\ \mathbf{x}_i^{(m)}, & \text{if } s_i^{(m)} = 0. \end{cases} \tag{6}$$

On the other hand, for samples in $\mathcal{D}$ but not in $\widetilde{S}$, we keep $\breve{\mathcal{X}}_i = \mathcal{X}_i$ and define their corrupted mask $\mathbf{s}_i \equiv \mathbf{0}$. At last, collecting all tuples yields the reveal-supervised training set:

$$\breve{\mathcal{D}} \ = \ \big\{(\breve{\mathcal{X}}_i, y_i, \mathbf{s}_i)\big\}_{i=1}^{N}. \tag{7}$$

In practice, both $\widetilde{S}$ and the masks $\mathbf{s}_i$ are resampled every epoch to diversify mismatches and reduce corrupted memorization.

**View reliability estimator.** After obtaining the bootstrapped set for the reveal-supervised paradigm, we instantiate a lightweight MLP $E(\cdot; \psi_m)$ that maps per-view evidence to a nonnegative reliability score. Specifically, given an input (clean or corrupted) from $\breve{\mathcal{D}}$, the reliability for view $m$ is computed as:

$$\alpha_i^{(m)} = \sigma\big[E(\mathbf{u}_i^{(m)}; \psi_m)\big] \in (0, 1), \tag{8}$$

where $\mathbf{u}_i^{(m)} = \breve{\mathbf{z}}_i^{(m)} = f\left(\breve{\mathbf{x}}_i^{(m)}; \theta_m\right) \in \mathbb{R}^d$ is the latent representation from the $m$-th encoder and $\sigma[\cdot]$ is the sigmoid function. After that, as the reveal-supervised signal $\mathbf{s}_i$ is known on $\breve{\mathcal{D}}$, the view reliability estimators $E(\cdot; \psi)$ can be optimized by aligning $\alpha_i^{(m)}$ with the clean indicator $1 - s_i^{(m)}$ via a binary cross-entropy as follows:

$$\mathcal{L}_w(\psi) = -\frac{1}{NM} \sum_{i=1}^{N} \sum_{m=1}^{M} \Big[\big(1 - s_i^{(m)}\big) \log \alpha_i^{(m)} + s_i^{(m)} \log\big(1 - \alpha_i^{(m)}\big)\Big]. \tag{9}$$

Intuitively, this objective encourages $\alpha_i^{(m)} \to 1$ for clean views and $\alpha_i^{(m)} \to 0$ for corrupted ones, yielding reliability estimates that are directly useful for robust fusion. In the next subsection, we further refine $\alpha_i^{(m)}$ with another two complementary signals derived from the view predictions.

### 3.3. Dual Prediction-derived Refinements

Learning reliabilities from features $\breve{\mathbf{z}}_i^{(m)}$ alone may be insufficient to detect mismatched views, as features themselves do not quantify per-view noisiness directly. To tackle this, we further refine the input of the generator with two prediction-derived signals that expose *cross-view disagreement* and *within-view quality* as follows:

**Proposition 1** (**Inter-view prediction discrepancy**). *Given prediction distribution $\mathbf{p}_i^{(m)}$ for view $m$, we quantify disagreement with other views via the averaged Jeffreys divergence [25] as follows:*

$$J_i^{(m)} = \frac{1}{M-1} \sum_{n=1, n\neq m}^{M} \Big[D_{KL}(\mathbf{p}_i^{(m)} \,\|\, \mathbf{p}_i^{(n)}) + D_{KL}(\mathbf{p}_i^{(n)} \,\|\, \mathbf{p}_i^{(m)})\Big], \tag{10}$$

*where $D_{KL}$ denotes the Kullback–Leibler (KL) divergence operator and a larger $J_i^{(m)}$ indicates stronger cross-view disagreement and potential mismatch.*

**Proposition 2** (**Intra-view prediction uncertainty**). *Given logits $\boldsymbol{\ell}_i^{(m)}$ for view $m$, we simply use Shannon entropy to measure the view quality as follows:*

$$\begin{aligned} Q_i^{(m)} &= -\log\left[1 - \frac{H_i^{(m)}}{\log C}\right] \\ &= -\log\left[1 + \frac{\sum_{c=1}^{C} p_{i,c}^{(m)} \log p_{i,c}^{(m)}}{\log C}\right], \end{aligned} \tag{11}$$

*which is small for confident predictions and large for ambiguous ones.*

These two quantities provide complementary cues for assessing view reliability. Specifically, the inter-view discrepancy $J_i^{(m)}$ measures how far the prediction of view $m$ deviates from the consensus of the other views, so a larger value indicates stronger inconsistency and a higher risk that this view is misaligned. On the other hand, the intra-view uncertainty $Q_i^{(m)}$ reflects the intrinsic confidence of the prediction from view $m$, where corrupted or low-quality views typically yield high-entropy, ambiguous predictions and thus large $Q_i^{(m)}$. Motivated by this, we refine the per-view input of the reliability estimator by concatenation:

$$\mathbf{u}_i^{(m)} = \big[\, \mathbf{z}_i^{(m)} \parallel J_i^{(m)} \parallel Q_i^{(m)} \big] \in \mathbb{R}^{d+2}. \tag{12}$$

Table 1. Classification Performance (Acc±Std) under varying noise ratios and ten different random seeds. The best is indicated in **red**, while the second is indicated in blue.

| Noise | Method | Caltech | Leaves | HW | LandUse | Scene | AVG. |
|---|---|---|---|---|---|---|---|
| 0% | TMC [ICLR'21] | 95.96±1.12 | 95.59±0.85 | 97.55±0.56 | 46.14±1.82 | 72.47±0.91 | 81.54 |
| | UIMC [CVPR'23] | 97.29±0.61 | 93.94±1.50 | 98.10±0.56 | 49.90±1.66 | 73.61±1.15 | 82.57 |
| | ECML [AAAI'24] | 96.25±0.79 | 93.47±1.30 | 97.22±0.56 | 44.64±1.85 | 70.45±1.00 | 80.41 |
| | CCML [ACM MM'24] | 95.79±1.09 | 96.81±0.69 | 97.38±0.81 | 41.38±2.11 | 71.56±1.39 | 80.58 |
| | MAMC [ICLR'25] | 97.82±0.72 | 89.88±1.55 | 98.90±0.58 | 71.67±1.60 | 80.71±1.28 | 87.80 |
| | ETF [ICML'25] | 96.47±0.91 | 98.44±0.31 | 97.40±0.60 | 46.09±1.71 | 75.27±0.99 | 82.73 |
| | TMCEK [ICML'25] | 96.18±0.83 | 93.53±1.28 | 97.30±0.63 | 45.10±2.29 | 71.04±1.22 | 80.63 |
| | FUML [ICML'25] | 95.29±0.93 | 99.28±0.63 | 98.45±0.81 | 74.95±1.71 | 78.92±1.68 | 89.38 |
| | RML [ICCV'25] | 95.54±1.58 | 97.03±1.06 | 98.20±0.82 | 75.40±2.43 | 78.06±1.13 | 88.85 |
| | **BML** | **98.11±0.75** | **99.75±0.27** | **99.03±0.59** | **81.43±1.06** | **83.92±1.39** | **92.45** |
| 50% | TMC [ICLR'21] | 87.07±1.08 | 79.69±1.63 | 83.42±1.34 | 39.62±2.07 | 61.83±1.54 | 70.33 |
| | UIMC [CVPR'23] | 92.96±1.21 | 78.09±1.45 | 88.75±1.29 | 42.43±1.18 | 63.55±1.27 | 73.16 |
| | ECML [AAAI'24] | 83.25±2.16 | 77.69±1.94 | 81.20±1.48 | 38.69±1.92 | 60.33±1.65 | 68.23 |
| | CCML [ACM MM'24] | 82.14±2.23 | 78.88±1.54 | 79.12±1.38 | 34.88±1.97 | 60.80±1.15 | 67.16 |
| | MAMC [ICLR'25] | 96.36±0.75 | 73.69±2.30 | 96.47±0.90 | 62.93±1.19 | 74.06±1.49 | 80.70 |
| | ETF [ICML'25] | 86.46±2.72 | 78.72±1.58 | 81.57±1.63 | 38.98±1.51 | 63.23±1.29 | 69.79 |
| | TMCEK [ICML'25] | 83.54±1.92 | 77.62±2.17 | 79.53±1.93 | 38.86±2.48 | 60.70±1.60 | 68.05 |
| | FUML [ICML'25] | 85.46±5.15 | 94.25±1.31 | 95.35±1.52 | 68.29±2.34 | 74.10±1.35 | 83.49 |
| | RML [ICCV'25] | 86.00±1.84 | 86.16±1.02 | 87.97±1.26 | 66.26±2.46 | 69.25±1.16 | 79.13 |
| | **BML** | **96.79±0.83** | **96.16±0.84** | **97.28±0.66** | **75.62±1.29** | **79.25±1.53** | **89.02** |
| 100% | TMC [ICLR'21] | 77.43±3.50 | 61.78±3.10 | 67.35±2.08 | 33.50±1.51 | 51.71±1.11 | 58.35 |
| | UIMC [CVPR'23] | 88.71±1.47 | 60.78±3.49 | 78.05±1.55 | 34.90±1.64 | 53.60±1.35 | 63.21 |
| | ECML [AAAI'24] | 68.11±2.23 | 59.91±3.60 | 64.28±1.73 | 32.48±1.92 | 51.01±0.95 | 55.16 |
| | CCML [ACM MM'24] | 67.71±2.47 | 60.09±3.21 | 59.45±2.10 | 29.52±1.89 | 50.08±1.08 | 53.37 |
| | MAMC [ICLR'25] | 95.21±1.50 | 58.81±2.57 | 93.28±0.49 | 53.33±2.35 | 67.92±1.30 | 73.71 |
| | ETF [ICML'25] | 76.14±4.71 | 59.53±2.27 | 64.33±1.92 | 31.52±1.05 | 51.80±0.90 | 56.67 |
| | TMCEK [ICML'25] | 69.11±2.78 | 59.91±3.71 | 61.80±2.26 | 33.02±1.50 | 51.65±0.84 | 55.10 |
| | FUML [ICML'25] | 75.79±9.43 | 89.88±1.93 | 92.10±3.13 | 63.74±2.49 | 69.05±1.22 | 78.11 |
| | RML [ICCV'25] | 75.29±3.36 | 76.81±2.89 | 77.05±1.06 | 55.43±1.99 | 60.45±1.01 | 69.01 |
| | **BML** | **95.36±1.13** | **93.56±0.73** | **94.97±1.02** | **69.02±1.50** | **74.04±1.10** | **85.39** |

By jointly considering discrepancy and uncertainty, the reliability estimator $E(\cdot;\psi)$ can better distinguish trustworthy evidence (low $J_i^{(m)}$, low $Q_i^{(m)}$) from noisy views (high $J_i^{(m)}$ and/or high $Q_i^{(m)}$), yielding reliability scores that are directly useful for robust late fusion.

### 3.4. Optimization

**Training.** To optimize BML, we minimize the cross-entropy over mini-batches:

$$\mathcal{L}_{cls} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}} \log \hat{p}_{i,y_i}, \tag{13}$$

where $\mathcal{B}$ denotes the mini-batch sampled from $\breve{\mathcal{D}}$ and the whole framework is trained end-to-end with the joint objective:

$$\mathcal{L} = \mathcal{L}_{cls} + \lambda\,\mathcal{L}_w, \tag{14}$$

where $\lambda > 0$ balances the task objectives and bootstrapped supervision.

**Inference.** At test time, given a potential noisy input $\hat{\mathcal{X}}_i$, we first compute $\{\hat{\mathbf{z}}_i^{(m)}, \hat{\boldsymbol{\ell}}_i^{(m)}, \hat{\mathbf{p}}_i^{(m)}\}_{m=1}^M$ to form $\{\hat{J}_i^{(m)}\}_{m=1}^M$ and $\{\hat{Q}_i^{(m)}\}_{m=1}^M$. After that, we obtain the reliabilities by $\{\alpha_i^{(m)} = \sigma\big[E(\hat{\mathbf{u}}_i^{(m)};\psi_m)\big]\}_{m=1}^M$ and produce the final fused prediction as follows:

$$\hat{y}_i = \arg\max_{c\in\{1,\ldots,C\}} \Big[\sum_{m=1}^{M} \alpha_i^{(m)}\,\hat{\boldsymbol{\ell}}_i^{(m)}\Big]_c. \tag{15}$$

Since $E(\cdot;\psi_m)$ is trained on a TNC-bootstrapped set under the reveal-supervised paradigm, it encourages aligned views and suppresses inconsistent ones, yielding robustness under TNC scenarios without requiring any explicit noise indicator at test time.

## 4. Experiment

In this section, we conduct a series of experiments to answer the following three questions:

- Q1: Is BML effective and robust across various levels of TNC scenarios?

Table 2. Classification Performance (Acc±Std) under varying noise ratios and ten different random seeds. The best is indicated in **red**, while the second is indicated in blue.

| Noise | Method | CCV | Fashion | NUS-OBJ | AWA | YouTubeFace | AVG. |
|---|---|---|---|---|---|---|---|
| 0% | TMC [ICLR'21] | 44.58±0.81 | 96.07±0.31 | 39.72±0.51 | 36.87±0.44 | 75.44±0.36 | 58.54 |
| | UIMC [CVPR'23] | 45.48±1.29 | 98.25±0.22 | 41.27±0.68 | 35.95±0.42 | 81.95±0.24 | 60.58 |
| | ECML [AAAI'24] | 42.59±1.09 | 95.43±0.38 | 35.34±0.49 | 32.93±0.56 | 52.52±1.22 | 51.76 |
| | CCML [ACM MM'24] | 43.62±0.67 | 95.66±0.52 | 38.04±0.58 | 31.46±0.92 | 40.36±0.63 | 49.83 |
| | MAMC [ICLR'25] | 54.66±1.26 | 98.53±0.28 | 45.84±0.57 | 32.15±0.83 | 87.18±0.39 | 63.67 |
| | ETF [ICML'25] | 48.40±0.79 | 96.78±0.26 | 45.98±0.41 | 48.38±0.35 | 83.02±0.25 | 64.51 |
| | TMCEK [ICML'25] | 42.90±0.97 | 94.53±0.33 | 35.40±0.59 | 33.70±0.47 | 53.69±1.26 | 52.04 |
| | FUML [ICML'25] | 47.38±2.43 | 98.09±0.33 | 47.14±0.56 | 38.30±0.43 | 83.10±0.43 | 62.80 |
| | RML [ICCV'25] | 47.50±1.03 | 98.66±0.21 | 44.00±0.52 | 41.58±0.38 | 88.72±0.13 | 64.09 |
| | **BML** | **59.56±0.99** | **98.85±0.16** | **51.91±0.40** | **48.90±0.27** | **89.72±0.22** | **69.79** |
| 50% | TMC [ICLR'21] | 38.92±0.79 | 84.11±0.61 | 34.51±0.29 | 28.98±0.42 | 63.48±0.28 | 50.00 |
| | UIMC [CVPR'23] | 39.68±1.02 | 91.43±0.49 | 35.62±0.71 | 28.68±0.53 | 68.72±0.24 | 52.83 |
| | ECML [AAAI'24] | 37.62±1.00 | 82.34±0.50 | 31.26±0.55 | 26.06±0.51 | 45.79±0.91 | 44.61 |
| | CCML [ACM MM'24] | 37.23±1.05 | 83.07±0.52 | 32.45±0.56 | 24.14±0.68 | 35.81±0.51 | 42.54 |
| | MAMC [ICLR'25] | 46.87±1.10 | 95.66±0.75 | 38.58±0.58 | 21.84±0.70 | 76.39±0.69 | 55.87 |
| | ETF [ICML'25] | 41.67±1.04 | 88.28±0.45 | 38.26±0.45 | 36.94±0.36 | 68.97±0.24 | 54.82 |
| | TMCEK [ICML'25] | 37.70±0.72 | 82.03±0.32 | 31.27±0.59 | 26.63±0.56 | 46.57±0.98 | 44.84 |
| | FUML [ICML'25] | 41.97±2.42 | 95.25±0.48 | 42.72±0.33 | 31.28±0.42 | 73.98±0.50 | 57.04 |
| | RML [ICCV'25] | 41.47±1.23 | 90.87±0.69 | 37.48±0.59 | 32.68±0.58 | 76.06±0.17 | 55.71 |
| | **BML** | **51.73±1.10** | **96.24±0.29** | **45.99±0.44** | **40.52±0.31** | **84.12±0.32** | **63.72** |
| 100% | TMC [ICLR'21] | 33.12±1.42 | 72.00±0.94 | 29.56±0.45 | 21.39±0.55 | 51.64±0.33 | 41.54 |
| | UIMC [CVPR'23] | 34.04±1.22 | 84.99±0.85 | 29.86±0.47 | 21.53±0.27 | 55.61±0.41 | 45.21 |
| | ECML [AAAI'24] | 31.99±0.94 | 68.60±0.90 | 27.44±0.41 | 19.17±0.41 | 38.99±0.63 | 37.24 |
| | CCML [ACM MM'24] | 30.67±1.08 | 69.91±1.05 | 27.23±0.36 | 16.65±0.67 | 31.43±0.33 | 35.18 |
| | MAMC [ICLR'25] | 38.09±1.54 | 93.47±0.51 | 31.28±0.48 | 11.28±0.59 | 65.89±1.02 | 48.00 |
| | ETF [ICML'25] | 33.30±1.45 | 80.55±0.59 | 31.56±0.29 | 25.92±0.51 | 55.57±0.45 | 45.38 |
| | TMCEK [ICML'25] | 32.18±0.88 | 69.21±1.06 | 27.40±0.39 | 19.40±0.33 | 39.46±0.65 | 37.53 |
| | FUML [ICML'25] | 36.66±2.29 | 92.75±0.52 | 37.96±0.52 | 24.34±0.41 | 64.87±0.50 | 51.32 |
| | RML [ICCV'25] | 35.09±1.04 | 83.31±1.71 | 31.17±0.48 | 24.08±0.59 | 63.45±0.34 | 47.42 |
| | **BML** | **44.16±0.97** | **94.37±0.35** | **39.98±0.53** | **32.22±0.51** | **78.42±0.26** | **57.83** |

- Q2: What underlying mechanisms drive BML's behavior and performance?
- Q3: Are the view reliabilities estimated by BML in TNC settings reasonable, *i.e.*, do they appropriately down-weight noisy views?

### 4.1. Datasets

To ensure a comprehensive evaluation, we employ eleven benchmark datasets spanning diverse scales and data types. Specifically, the widely used feature-vector datasets include Caltech [16], Leaves [54], HW [14], LandUse [70], Scene [15], CCV [27], Fashion [60], NUS-OBJ [7], AWA [29], and YouTubeFace [58], which are partitioned into stratified 8:2 train-test splits. In addition, to assess BML on raw data inputs, we construct a three-view dataset, where the visual views are the RGB and depth images sourced from SUN RGB-D [24, 48, 50, 61], while the textual view is a single-sentence content description generated for each RGB image using Qwen3-VL-32B-Instruct [68]. We refer to this constructed benchmark as SUN R-D-T, which adopts the original split from SUN RGB-D. Due to space constraints, detailed descriptions of all datasets and the construction pipeline are provided in the supplementary material.

### 4.2. Implementation Details

We implement BML in PyTorch v2.1.2 and conduct all experiments on a machine running Ubuntu 20.04 with a single NVIDIA RTX 3090 GPU. To assess robustness under varying noise levels in downstream test sets, we report results at noise ratios $\eta \in \{0\%, 50\%, 100\%\}$* in the main paper, where $\eta$ denotes the proportion of test samples that suffer the scenario defined in Definition 1. Our experiments are organized by dataset type. Specifically, for feature-vector datasets, to avoid per-dataset tuning, we employ the same fully connected feature extractor and a unified hyperparameter configuration across all datasets. Models are trained with Adam [28] for 200 epochs, using an initial learning rate

*Due to space constraints, comprehensive results covering noise ratios $\eta = 0\% - 100\%$ are provided in the supplementary material.

Table 3. Ablation study of the components in our BML framework under 50% TNC noise, where the best is indicated in **red**.

| Type | Caltech | Leaves | HW | LandUse | Scene | NUS-OBJ | SUN R-D-T | AVG. |
|---|---|---|---|---|---|---|---|---|
| W/O $\mathcal{L}_w$ | 91.50±1.48 | 80.81±1.77 | 91.40±0.96 | 65.60±1.86 | 73.78±1.09 | 40.82±0.50 | 58.50±0.93 | 71.77 |
| W/O $J$ | 96.32±0.68 | 95.12±0.81 | 96.58±0.68 | 74.50±1.98 | 78.65±1.53 | 45.55±0.53 | 61.76±0.22 | 78.35 |
| W/O $Q$ | 96.68±0.62 | 96.06±0.91 | 97.17±0.54 | 75.10±1.89 | 78.81±1.49 | 45.93±0.57 | 64.20±0.55 | 79.14 |
| W/O on-the-fly | 94.00±1.13 | 86.66±1.73 | 94.85±1.34 | 69.48±1.65 | 76.43±1.43 | 42.03±0.46 | 56.27±0.93 | 74.25 |
| **FULL** | **96.79±0.83** | **96.16±0.84** | **97.28±0.66** | **75.62±1.29** | **79.25±1.53** | **45.99±0.44** | **64.54±0.59** | **79.38** |

of $2 \times 10^{-3}$ and a batch size of 2048. We set the in-place augmentation ratio to $\rho = 0.5$, and balance $\mathcal{L}_w$ and $\mathcal{L}_{cls}$ with $\lambda = 1.0$. For the SUN R-D-T dataset, we adopt BERT [10] for text and a ResNet-18 [20] backbone pretrained on ImageNet for images. Input images are resized to $256 \times 256$ pixels and randomly cropped to $224 \times 224$. Models are trained with AdamW [39] for 100 epochs, using weight decay 0.01 and an initial learning rate of $1 \times 10^{-4}$ for ResNet and $2 \times 10^{-5}$ for BERT, respectively. We set $\lambda = 50.0$ to balance $\mathcal{L}_w$ and $\mathcal{L}_{cls}$, and all other settings follow those used for the feature-vector datasets.

Table 4. Classification Performance (Acc±Std) on SUN R-D-T dataset under varying noise ratios and five different random seeds. Indicators are the same as Table 1 and Table 2.

| Method | 0% | 50% | 100% |
|---|---|---|---|
| MAMC [ICLR'25] | 62.54±1.66 | 58.58±1.77 | 54.43±2.13 |
| TMCEK [ICML'25] | 63.30±0.53 | 55.47±0.59 | 47.70±0.70 |
| FUML [ICML'25] | 63.77±0.89 | 59.68±0.96 | 55.55±0.81 |
| RML [ICCV'25] | 58.23±1.15 | 50.15±1.13 | 42.73±1.32 |
| **BML** | **68.15±0.28** | **64.54±0.59** | **60.97±0.60** |

### 4.3. Comparison with State-of-the-arts (Q1)

**Results on feature-vector datasets.** To demonstrate the effectiveness of BML, we compare its classification performance against nine state-of-the-art baseline methods, which comprise: i) trustworthy methods that leverage uncertainty estimation to derive fusion weights for multi-view learning—TMC [18], UIMC [62], ECML [64], CCML [38], ETF [40], TMCEK [34], and FUML [11], and ii) deterministic methods that improve either representation quality (*e.g.*, MAMC [37]) or cross-view alignment (*e.g.*, RML [66]). We report the last epoch mean accuracy (*Acc*) and standard deviation (*Std*), and average all results over 10 different seeds to mitigate randomness. The results of ten benchmarks (sorted by data size) under both clean and TNC scenarios are summarized in Table 1 and 2, from which we make the following observations:

- In the absence of noisy correspondence at test time (*i.e.*, W/O TNC), BML still outperforms existing SOTA methods on all datasets—most notably surpassing RML [66] by 6.03% on LandUse and FUML [11] by 4.77% on NUS-OBJ, suggesting that the reveal-supervised learning paradigm enables BML to infer and leverage per-view quality even when correspondences are clean.
- Under TNC scenarios (50% and 100% noise), BML exhibits strong and consistent dominance, with gains particularly pronounced on large-scale datasets (*e.g.*, AWA, YouTubeFace). We attribute these margins to BML's explicit bridging of the train-test task gap induced by TNC, where bootstrapped supervision accurately identifies mismatched views and down-weights their contributions during fusion, yielding more robust predictions.

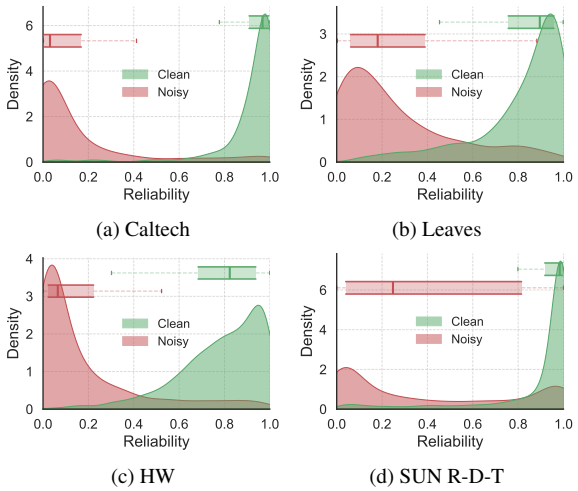
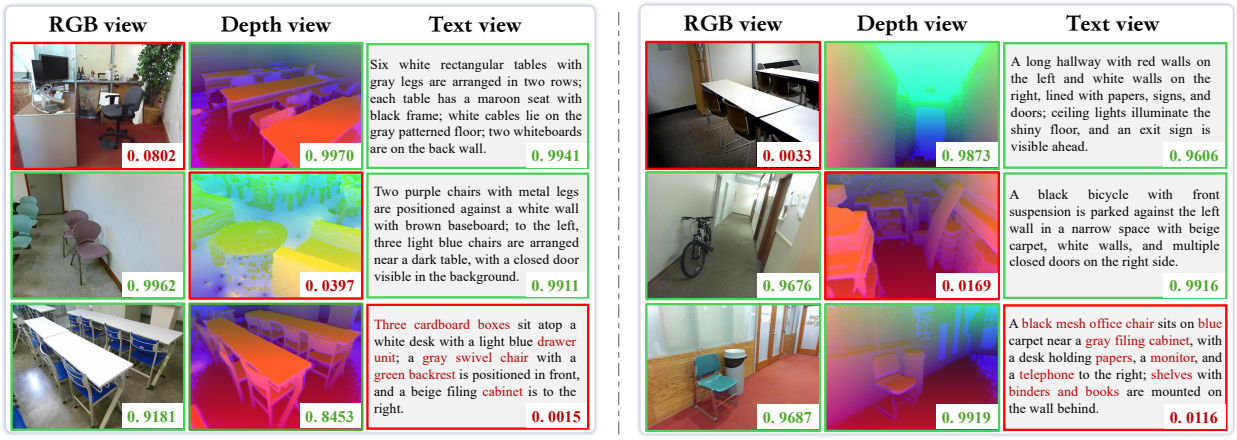



Figure 3. Density plots and box plots (top half of each subplot) of the estimated view reliability $\alpha_i^{(m)}$ on the test set of four datasets.

**Results on RGB-D Scene Recognition.** To emphasize the generalizability of BML to raw data, we further evaluate it on our proposed SUN R-D-T scene recognition benchmark. We compare BML with two state-of-the-art trusted baselines (TMCEK [34], FUML [11]) and two strong deterministic baselines (MAMC [37], RML [66]), reporting results in Table 4 across five distinct random seeds. The results show that under TNC scenarios, BML significantly outperforms all baselines, while on clean test sets it also remains highly competitive, highlighting its promise for deployment in practical multi-view scenarios.




(a) Ground Truth: *Classroom* (b) Ground Truth: *Corridor*

Figure 4. Case study of the view-specific reliability on the SUN R-D-T test set under the TNC scenario, where the red box represents the TNC sample, and the estimated reliability weight for each view is indicated in the bottom-right corner of its respective panel.

(a) W/O $\alpha_i^{(m)}$ ($Acc = 90.75$) (b) W/ $\alpha_i^{(m)}$ ($Acc = 94.15$)

Figure 5. t-SNE visualization comparison with and without our view reliability on the Fashion test set.

### 4.4. Ablation Study (Q2)

We dissect BML's working mechanism from a component-wise perspective. First, we report the performance of a naive classifier trained solely with cross-entropy $\mathcal{L}_{cls}$, *i.e.*, W/O the reveal-supervised loss ($\mathcal{L}_w$), which is substantially worse than all BML variants. Next, we ablate the two prediction-derived signals from the input to the view reliability estimator one at a time: removing $J$ (inter-view predictive discrepancy) reduces performance by $1.03\%$, while removing $Q$ (intra-view predictive uncertainty) yields a $0.24\%$ drop, both relative to the full model. Finally, disabling the on-the-fly, bootstrapped construction of the noise-augmented set causes the reveal-supervised paradigm to overfit the injected noise patterns, leading to a pronounced decline of $5.13\%$.

### 4.5. Visualization Analysis (Q3)

**Reliability Visualization.** To build intuition about the estimated reliabilities, Figure 3 shows the test-time reliability density predicted by the estimator $E(\cdot;\psi)$ (with boxplot overlays at the top of each subfigure). Obviously, BML successfully captures the TNC patterns learned via the reveal-supervised paradigm, clearly peeling off noisy correspondence samples from the clean cluster and assigning them lower fusion weights.

**Case Study.** Figure 4 visualizes the reliability weights estimated by BML for the "classroom" and "corridor" categories from the SUN R-D-T test set under TNC scenarios. The results show that BML is highly sensitive to misalignment, where views involved in TNC are consistently assigned very low weights (mostly less than $0.1$), whereas aligned views retain high reliability. This provides strong evidence of BML's ability to resist TNC during deployment.

**t-SNE Visualization.** Given the reliabilities that capture TNC, we visualize fused logits with and without applying them for fusion in Figure 5. When the test set includes noisy correspondence and all views are fused with equal weights (Figure 5(a)), class separability deteriorates. After training with $\mathcal{L}_w$ and applying the estimated reliabilities (Figure 5(b)), the clusters become markedly more discriminative.

## 5. Conclusion

In this paper, we identify and formalize a pervasive deployment failure mode for multi-view systems, *i.e.*, test-time noisy correspondence arising from asynchronous view acquisition. To address it, we introduce a Bootstrapping Multi-view Learning framework that narrows the train–test gap by constructing an in-place, TNC-bootstrapped noise-augmented set and optimizing it with a reveal-supervised paradigm. An important next step is to extend this framework to a unified treatment of TNC and test-time view missingness, enabling joint handling of misalignment and incompleteness to further strengthen real-world reliability.

## 6. Acknowledgments

This work was supported in part by the National Key R&D Program of China 2024YFB4710604; in part by NSFC under Grant 62472295, U25A201523, and U25B6003; in part by the Fundamental Research Funds for the Central Universities under Grant CJ202303 and CJ202403; in part by Sichuan Science and Technology Planning Project under Grant 24NSFTD0130; and in part by System of Systems and Artificial Intelligence Laboratory pioneer fund grant under Grant HLJGGG20240327517-15. In terms of author contributions, Peng Hu conceived and coordinated the research, designed the BML algorithm, and critically revised the manuscript. Xi Peng co-designed the BML algorithm, provided high-level methodological guidance, and contributed to manuscript revision. Changhao He co-designed and implemented the BML algorithm, conducted the majority of the experiments, analyzed the data, and drafted the initial version of the manuscript. Shuxian Li conducted the baseline evaluations and assisted with revising the manuscript. Di Xue and Yanji Hao contributed to the problem formulation, provided domain expertise, and assisted with the analysis and interpretation of the results. All authors read and approved the final manuscript.

## References

[1] Pei An, Junfeng Ding, Siwen Quan, Jiaqi Yang, You Yang, Qiong Liu, and Jie Ma. Survey of extrinsic calibration on lidar-camera system for intelligent vehicle: Challenges, approaches, and trends. *IEEE Transactions on Intelligent Transportation Systems*, 25(11):15342–15366, 2024. 3

[2] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013. 2

[3] Sicong Che, Zhaoming Kong, Hao Peng, Lichao Sun, Alex Leow, Yong Chen, and Lifang He. Federated multi-view learning for private medical data integration and analysis. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 13(4):1–23, 2022. 1

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 1

[5] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017. 1

[6] Sauhaarda Chowdhuri, Tushar Pankaj, and Karl Zipser. Multinet: Multi-modal multi-task learning for autonomous driving. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1496–1504. IEEE, 2019. 1

[7] 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*, pages 1–9, 2009. 6

[8] Zhuohang Dang, Minnan Luo, Jihong Wang, Chengyou Jia, Haochen Han, Herun Wan, Guang Dai, Xiaojun Chang, and Jingdong Wang. Disentangled noisy correspondence learning. *IEEE Transactions on Image Processing*, 34:2602–2615, 2025. 3

[9] Arthur P Dempster. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions*, pages 57–72. Springer, 2008. 3

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, 2019. 7

[11] Siyuan Duan, Yuan Sun, Dezhong Peng, Guiduo Duan, Xi Peng, and Peng Hu. Deep fuzzy multi-view learning for reliable classification. In *Forty-second International Conference on Machine Learning*. 1, 3, 7

[12] Siyuan Duan, Yuan Sun, Dezhong Peng, Zheng Liu, Xiaomin Song, and Peng Hu. Fuzzy multimodal learning for trusted cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20747–20756, 2025. 3

[13] Yue Duan, Zhangxuan Gu, Zhenzhe Ying, Lei Qi, Changhua Meng, and Yinghuan Shi. Pc2: Pseudo-classification based pseudo-captioning for noisy correspondence learning in cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9397–9406, 2024. 3

[14] Robert Duin. Multiple Features. UCI Machine Learning Repository, 1998. 6

[15] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pages 524–531. IEEE, 2005. 6

[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 *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 6

[17] Kai Guo, Jiedong Wang, Xi Peng, Peng Hu, and Hao Wang. Disentangling multi-view representations via curriculum learning with learnable prior. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 5262–5270, 2025. 2

[18] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In

*International Conference on Learning Representations*, 2021. 1, 3, 7

[19] Changhao He, Hongyuan Zhu, Peng Hu, and Xi Peng. Robust variational contrastive learning for partially view-unaligned clustering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 4167–4176, 2024. 2

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[21] Yining Hong, Chunru Lin, Yilun Du, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 3d concept learning and reasoning from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9202–9212, 2023. 1

[22] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34: 29406–29419, 2021. 3

[23] Zhenyu Huang, Mouxing Yang, Xinyan Xiao, Peng Hu, and Xi Peng. Noise-robust vision-language pre-training with positive-negative learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):338–350, 2025. 3

[24] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T Barron, Mario Fritz, Kate Saenko, and Trevor Darrell. A category-level 3-d object dataset: Putting the kinect to work. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1168–1174. IEEE, 2011. 6

[25] Harold Jeffreys. *The theory of probability*. OuP Oxford, 1998. 4

[26] Liang Jiang, Zhenyu Huang, Jia Liu, Zujie Wen, and Xi Peng. Robust domain adaptation for machine reading comprehension. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8060–8069, 2023. 3

[27] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C Loui. 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*, pages 1–8, 2011. 6

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[29] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014. 6

[30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 3

[31] Shuxian Li, Changhao He, Xiting Liu, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Learning with noisy triplet correspondence for composed image retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19628–19637, 2025. 3

[32] 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, 2018. 1

[33] Yuan Li, Liangli Zhen, Yuan Sun, Dezhong Peng, Xi Peng, and Peng Hu. Deep evidential hashing for trustworthy cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 18566–18574, 2025. 1

[34] Xinyan Liang, Shijie Wang, Yuhua Qian, Qian Guo, Liang Du, Bingbing Jiang, Tingjin Luo, and Feijiang Li. Trusted multi-view classification with expert knowledge constraints. In *Forty-second International Conference on Machine Learning*. 7

[35] Xinyan Liang, Pinhan Fu, Yuhua Qian, Qian Guo, and Guoqing Liu. Trusted multi-view classification via evolutionary multi-view fusion. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[36] Yijie Lin, Yuanbiao Gou, Xiaotian Liu, Jinfeng Bai, Jiancheng Lv, and Xi Peng. Dual contrastive prediction for incomplete multi-view representation learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4447–4461, 2022. 2

[37] Yuena Lin, Yiyuan Wang, Gengyu Lyu, Yongjian Deng, Haichun Cai, Huibin Lin, Haobo Wang, and Zhen Yang. Enhance multi-view classification through multi-scale alignment and expanded boundary. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 7

[38] 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*, pages 7269–7277, 2024. 3, 7

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[40] Jueqing Lu, Wray Buntine, YUANYUAN QI, Joanna Dipnall, Belinda Gabbe, and Lan Du. Navigating conflicting views: Harnessing trust for learning. In *Forty-second International Conference on Machine Learning*. 3, 7

[41] Kuldeep Mahato, Tamoghna Saha, Shichao Ding, Samar S Sandhu, An-Yi Chang, and Joseph Wang. Hybrid multimodal wearable sensors for comprehensive health monitoring. *Nature Electronics*, 7(9):735–750, 2024. 1, 3

[42] Chang Nie, Guangming Wang, Zhe Lie, and Hesheng Wang. Ermv: Editing 4d robotic multi-view images to enhance embodied agents. *arXiv preprint arXiv:2507.17462*, 2025. 1

[43] Yang Qin, Yingke Chen, Dezhong Peng, Xi Peng, Joey Tianyi Zhou, and Peng Hu. Noisy-correspondence

learning for text-to-image person re-identification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27197–27206, 2024. 3

[44] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31, 2018. 1, 3

[45] Glenn Shafer. Dempster-shafer theory. *Encyclopedia of artificial intelligence*, 1:330–331, 1992. 1

[46] Jiangming Shi, Xiangbo Yin, Yachao Zhang, Yuan Xie, Yanyun Qu, et al. Learning commonality, divergence and variety for unsupervised visible-infrared person re-identification. *Advances in Neural Information Processing Systems*, 37:99715–99734, 2024. 3

[47] Jiangming Shi, Xiangbo Yin, Yeyun Chen, Yachao Zhang, Zhizhong Zhang, Yuan Xie, and Yanyun Qu. Multi-schema proximity network for composed image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19999–20008, 2025. 3

[48] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 6

[49] Qihong Song, Hongyuan Zhu, Joey Tianyi Zhou, Xi Peng, Peng Hu, et al. Deep unsupervised hashing via external guidance. In *Forty-second International Conference on Machine Learning*, 2025. 2

[50] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 6

[51] Yuan Sun, Kaiming Liu, Yongxiang Li, Zhenwen Ren, Jian Dai, and Dezhong Peng. Distribution consistency guided hashing for cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5623–5632, 2024. 3

[52] Yuan Sun, Yang Qin, Yongxiang Li, Dezhong Peng, Xi Peng, and Peng Hu. Robust multi-view clustering with noisy correspondence. *IEEE Transactions on Knowledge and Data Engineering*, 36(12):9150–9162, 2024. 3

[53] Yuan Sun, Yongxiang Li, Zhenwen Ren, Guiduo Duan, Dezhong Peng, and Peng Hu. Roll: Robust noisy pseudo-label learning for multi-view clustering with noisy correspondence. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30732–30741, 2025. 3

[54] Hao Wang, Linlin Zong, Bing Liu, Yan Yang, and Wei Zhou. Spectral perturbation meets incomplete multi-view data. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 3677–3683, 2019. 6

[55] Jinbo Wang, Weihua Xu, Qinghua Zhang, Yuhua Qian, and Weiping Ding. Dynamic multiview classification and knowledge fusion: A fuzzy concept-cognitive learning perspective. *IEEE Transactions on Fuzzy Systems*, 33 (10):3476–3490, 2025. 1

[56] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR, 2015. 2

[57] Yunbo Wang, YuJie Wu, Zhien Dai, Can Tian, Jun Long, and Jianhai Chen. Noisy correspondence rectification via asymmetric similarity learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21384–21392, 2025. 3

[58] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE, 2011. 6

[59] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024. 1

[60] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 6

[61] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision*, pages 1625–1632, 2013. 6

[62] 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*, pages 19873–19882, 2023. 7

[63] Zequn Xie, Haoming Ji, and Lingwei Meng. Dynamic uncertainty learning with noisy correspondence for text-based person search. *arXiv preprint arXiv:2505.06566*, 2025. 3

[64] 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*, pages 16129–16137, 2024. 1, 3, 7

[65] 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*, pages 5263–5271, 2024. 3

[66] Jie Xu, Na Zhao, Gang Niu, Masashi Sugiyama, and Xiaofeng Zhu. Robust multi-view learning via representation fusion of sample-level attention and alignment of simulated perturbation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4232–4241, 2025. 2, 7

[67] Shilin Xu, Yuan Sun, Xingfeng Li, Siyuan Duan, Zhenwen Ren, Zheng Liu, and Dezhong Peng. Noisy label calibration for multi-view classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 21797–21805, 2025. 2

[68] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 6

[69] Mouxing Yang, Zhenyu Huang, Peng Hu, Taihao Li, Jiancheng Lv, and Xi Peng. Learning with twin noisy labels for visible-infrared person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14308–14317, 2022. 3

[70] Yi Yang and Shawn Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010. 6

[71] Lotfi A Zadeh. Fuzzy sets. *Information and control*, 8 (3):338–353, 1965. 2, 3

[72] Changqing Zhang, Zongbo Han, Huazhu Fu, Joey Tianyi Zhou, Qinghua Hu, et al. Cpm-nets: Cross partial multi-view networks. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 2

[73] Xu Zhang, Hao Li, and Mang Ye. Negative pre-aware for noisy cross-modal matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7341–7349, 2024. 3