Supplementary Material for 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

In this supplementary material, we first provide detailed proofs for the propositions stated in the main paper (Section 1 and Section 2). We then describe the datasets used in our experiments, including the construction of our proposed three-view scene recognition benchmark, SUN R-D-T (Section 3), and present summary statistics for ten widely used feature-vector datasets (Section 4). Finally, in Section 5 we report additional experimental results, including performance comparisons under varying noise ratios, ablation studies, and a comprehensive sensitivity analysis of the two hyperparameters in the BML framework.

1. Proof of Proposition 1

Given a fixed sample i and view m, recall that the inter-view prediction discrepancy is defined as:

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

where D_{KL} denotes the Kullback–Leibler (KL) divergence operator between two distributions. By Gibbs' inequality, for any two distributions $\mathbf{p}$ and $\mathbf{q}$, we have:

$$D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}) \;\geq\; 0, \tag{2}$$

and equality holds iff $\mathbf{p} = \mathbf{q}$ almost everywhere. Therefore, each summand

$$D_{\mathrm{KL}}(\mathbf{p}_i^{(m)}\|\mathbf{p}_i^{(n)}) + D_{\mathrm{KL}}(\mathbf{p}_i^{(n)}\|\mathbf{p}_i^{(m)}) \tag{3}$$

is nonnegative, and it is strictly positive whenever $\mathbf{p}_i^{(m)} \neq \mathbf{p}_i^{(n)}$. Since $J_i^{(m)}$ is the average of these nonnegative terms, we have $J_i^{(m)} \geq 0$, and $J_i^{(m)} = 0$ iff $\mathbf{p}_i^{(m)} = \mathbf{p}_i^{(n)}$ for all $n \neq m$. Hence, a larger $J_i^{(m)}$ corresponds to stronger disagreement between the prediction of view m and those of the other views, which justifies using $J_i^{(m)}$ as a signal of inter-view prediction discrepancy.

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

2. Proof of Proposition 2

Given a fixed sample i and view m, let $\mathbf{p}_i^{(m)}$ be the predictive distribution obtained from the logits $\boldsymbol{\ell}_i^{(m)}$ via the softmax function. The Shannon entropy of $\mathbf{p}_i^{(m)}$ is:

$$H_i^{(m)} = -\sum_{c=1}^{C} p_{i,c}^{(m)} \log p_{i,c}^{(m)}. \tag{4}$$

It is well known that $H_i^{(m)} \in [0, \log C]$, with $H_i^{(m)} = 0$ iff the prediction is perfectly confident (one class has probability 1), and $H_i^{(m)}$ attaining its maximum $\log C$ when $\mathbf{p}_i^{(m)}$ is the uniform distribution. The proposed intra-view prediction uncertainty is:

$$Q_i^{(m)} = -\log\left(1 - \frac{H_i^{(m)}}{\log C}\right), \tag{5}$$

substituting the expression of $H_i^{(m)}$ gives:

$$Q_i^{(m)} = -\log\left[1 + \frac{\sum_{c=1}^{C} p_{i,c}^{(m)} \log p_{i,c}^{(m)}}{\log C}\right], \tag{6}$$

which is the second expression in the proposition. Define the normalized entropy $\tilde{H}_i^{(m)} = H_i^{(m)} / \log C \in [0, 1]$ and the scalar function

$$f(x) = -\log(1-x), \qquad x \in [0, 1). \tag{7}$$

Then $Q_i^{(m)} = f(\tilde{H}_i^{(m)})$. The derivative of f is:

$$f'(x) = \frac{1}{1-x} > 0 \quad \text{for } x \in [0, 1), \tag{8}$$

so f is strictly increasing on $[0, 1)$. Consequently, $Q_i^{(m)}$ is a strictly increasing function of $H_i^{(m)}$. When the prediction is perfectly confident, $H_i^{(m)} = 0$ and hence

$$Q_i^{(m)} = -\log(1-0) = 0, \tag{9}$$

which is small. As the prediction becomes more ambiguous, the entropy $H_i^{(m)}$ increases towards its maximum $\log C$, so $\tilde{H}_i^{(m)} \to 1$ and

$$Q_i^{(m)} = -\log\left(1 - \tilde{H}_i^{(m)}\right) \to +\infty. \tag{10}$$

Therefore, $Q_i^{(m)}$ remains small for confident predictions and grows large for highly uncertain, ambiguous predictions, which justifies its use as a signal of intra-view prediction uncertainty.

Remark 1 (Numerical stability in implementation)**.** *Theoretically, both the inter-view prediction discrepancy $J_i^{(m)}$ and the intra-view uncertainty $Q_i^{(m)}$ can become arbitrarily large: $J_i^{(m)}$ may diverge when some predictive probabilities vanish (leading to terms of the form $\log 0$ in the KL divergences), and $Q_i^{(m)}$ diverges as the entropy $H_i^{(m)}$ approaches its maximum $\log C$. In our implementation, we adopt standard numerical stabilization techniques to avoid infinities. Specifically, when computing the symmetric KL divergences, all predictive probabilities are clipped away from zero by a small constant $\varepsilon > 0$, i.e., $p_{i,c}^{(m)}$ is replaced by $\max\{p_{i,c}^{(m)}, \varepsilon\}$ before evaluating the KL terms. Likewise, for the intra-view uncertainty, we compute:*

$$Q_i^{(m)} = -\log\left(\max\{1 - H_i^{(m)}/\log C,\ \varepsilon\}\right),$$

so that the argument of the logarithm is always bounded below by ε. These modifications guarantee that $J_i^{(m)}$ and $Q_i^{(m)}$ remain finite in practice, while leaving their theoretical interpretations in Propositions 1 and 2 unchanged. For consistency, we use $\varepsilon = 1e-8$ throughout our implementation.

3. Details of the SUN R-D-T Dataset

This section details the construction of the SUN R-D-T benchmark, which extends the original SUN RGB-D dataset [12] by adding sentence-level content descriptions. We elaborate on the data generation process, statistical analysis, qualitative examples, and limitations.

3.1. Data Generation

The RGB-Depth image pairs in SUN R-D-T are sourced entirely from the SUN RGB-D dataset [12], which aggregates indoor RGB-D images from NYU Depth v2 [11], Berkeley B3DO [7] and SUN3D [16]. Following [1, 6, 20], we adopt the same train-test splits, resulting in a total of 9,504 samples. To simulate a realistic multi-view/modal perception scenario, we employ the Qwen3-VL-32B-Instruct model [18] to generate a descriptive caption for each RGB image, with an explicit focus on capturing the salient scene content. The model is conditioned on both the RGB image and a carefully designed structured prompt that specifies the TASK, STRICT PROHIBITIONS, STYLE AND LENGTH requirements, UNCERTAINTY POLICY and OUTPUT format, as illustrated in Figure 2. Inference is performed on a machine equipped with 8 NVIDIA 3090 GPUs using VLLM [9] for efficient batched decoding. We adopt a batch size of 64 and greedy decoding with `max_tokens=64`. Under this configuration, caption generation for the entire dataset requires approximately 12.8 GPU hours.

Table 1. Statistical summary of the constructed text descriptions.

Metric	chars	words
Avg text length	188.90	34.64
Median text length	179.00	33.00
Std dev of length	45.11	8.58
Vocabulary size	-	2664.00

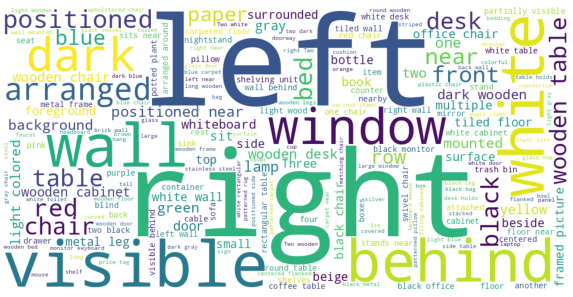


Figure 1. Visualization of the most frequent semantic concepts in the text descriptions.

3.2. Dataset Statistics

SUN R-D-T retains the original 19-class split of SUN RGB-D, comprising 4,845 training samples and 4,659 test samples. For the text view, we report corpus-level statistics in Table 1. The generated descriptions contain on average 188.90 characters and 34.64 words, with a median length of 179 characters and 33 words and a standard deviation of 45.11 characters and 8.58 words, respectively. The resulting vocabulary size is 2,664 unique word types, indicating that the captions are short yet lexically diverse.

To further characterize the semantics of the text, we visualize the token distribution in a word cloud in Figure 1. High-frequency terms predominantly correspond to object categories (e.g., *table*, *chair*, *window*), attributes (e.g., *wooden*, *dark*, *visible*), and spatial relations (e.g., *left*, *right*, *behind*, *near*). This focus is con-

```
You are generating strictly content-focused image descriptions for
research on multimodal classification.

TASK
Describe only visible objects, attributes, spatial relations
(left/right/near/behind/under), counts, and human actions if plainly
observable.  Base every token on visual evidence.

STRICT PROHIBITIONS (to prevent label leakage)
Do NOT name, imply, or hint at any place or scene type.  The following
words/phrases and their plurals, synonyms, or morphological variants are
FORBIDDEN and must not appear:  bathroom, bedroom, classroom, computer
room, conference room, corridor, dining area, dining room, discussion
area, furniture store, home office, kitchen, lab, laboratory, lecture
theatre, lecture theater, library, living room, office, rest space, study
space, interior, exterior, indoors, outdoors.

STYLE AND LENGTH
1.  No scene/type labels, no brand/model guesses, no value judgments, no
speculation.
2.  One sentence in English, less than 20 words.

UNCERTAINTY POLICY
If a detail is unclear, omit it rather than guessing.

OUTPUT
Return only the single sentence (no prefixes, no metadata).
```

Figure 2. The structured prompt used to elicit content-focused, scene-agnostic image descriptions for the SUN R-D-T dataset.

sistent with the prompt design in Figure 2, which enforces content-centric, single-sentence descriptions, explicitly forbids scene-type words to reduce label leakage, and encourages omitting uncertain details rather than hallucinating. Overall, the statistics and visualizations confirm that SUN R-D-T provides concise, structurally homogeneous, and semantically grounded text descriptions aligned with our intended multi-view setting.

3.3. Qualitative Examples

Figure 7 – Figure 25 present qualitative examples from SUN R-D-T across all 19 scene categories, including *bathroom*, *bedroom*, *classroom*, *computer room*, *conference room*, *corridor*, *dining area*, *dining room*, *discussion area*, *furniture store*, *home office*, *kitchen*, *lab*, *lecture theatre*, *library*, *living room*, *office*, *rest space*, and *study space*. Each example includes the original RGB image, its corresponding depth map, and the generated text description. These triplets illustrate how the three views complement one another, with the text capturing object types, attributes, and spatial relations that are consistent with the visual evidence in both RGB and depth.

3.4. Limitations

While SUN R-D-T provides a controlled testbed for studying modality robustness, it also has several limitations:

- First, the textual descriptions necessarily offer a compressed view of the RGB content and cannot fully capture fine-grained appearance cues or small objects.
- Second, all captions are generated by a single MLLM (Qwen3-VL-32B-Instruct) under a fixed prompt, which introduces systematic biases and occasional hallucinations or omissions despite the uncertainty policy and strict prohibitions in Figure 2. Large-scale human verification is not performed, and thus some descriptions may deviate from the exact scene content.
- Third, the model only partially adheres to the intended style and length. As shown in Table 1, the average sentence length noticeably exceeds the requested value, and the vocabulary statistics together with the word cloud in Figure 1 reveal a strong emphasis on frequent object nouns and spatial prepositions, while rarer concepts are underrepresented. This may bias downstream models toward dominant patterns in the

corpus.

- Finally, SUN R-D-T is restricted to English descriptions and the 19 indoor scene categories inherited from SUN RGB-D, which limits its coverage of outdoor environments, multilingual settings, and more diverse domains. Despite these constraints, SUN R-D-T still serves as a useful benchmark for studying multi-view test-time noisy correspondence (TNC), where models must remain robust under imperfect alignment among text, RGB, and Depth views.

4. Details of Feature-vector datasets

To evaluate the effectiveness of our method, we conduct extensive experiments on ten widely used feature-vector datasets containing diverse data types and scales. The details of these datasets are summarized below.

- **Caltech** [5] implies a collection of RGB images organized into multiple views. Following [17], we select 1,400 images representing 7 categories, and employ five distinct views for feature extraction including WM, CENTRIST, LBP, GIST, and HOG.
- **Leaves** [13] comprises 1,600 images distributed across 100 classes. Each image is characterized by three views consisting of texture histogram features, a shape descriptor, and a fine scale margin feature.
- **HW (Handwritten)** [3] consists of 2,000 samples evenly distributed among 10 digit classes with 200 samples per class. This dataset incorporates six feature types including 76 Fourier coefficients of the character shapes (FOU), 216 profile correlations (FAC), 64 Karhunen–Loeve coefficients (KAR), 240 pixel averages in 2×3 windows (PIX), 47 Zernike moments (ZER), and 6 morphological (MOR) features.
- **LandUse** [19] encompasses 2,100 satellite images categorized into 21 classes. We utilize GIST, PHOG, and LBP as the three feature sets for view generation.
- **Scene** [4] depicts indoor and outdoor environments across 15 categories with 4,485 images. The features extracted for three distinct views are GIST, PHOG, and LBP.
- **CCV (Columbia Consumer Video)** [8] contains 6,773 video samples divided into 20 categories. Each sample utilizes a Bag of Words representation derived from three views including STIP, SIFT, and MFCC.
- **Fashion** [15] includes 70,000 product images covering 10 fashion categories. For the test set of 10,000 images, we construct the first view from the set itself while generating the second and third views by randomly selecting samples from the same category.
- **NUS-OBJ (NUSWIDE-OBJ)** [2] targets object recognition tasks and contains 30,000 images across 31 categories. Five views are generated using color histogram, CM, CORR, edge direction histogram, and wavelet texture features.
- **AWA (Animals with Attributes)** [10] serves as a large-scale dataset with 30,475 samples spanning 50 animal categories. We utilize six features for each image including Color Histogram (CQ), Local Self Similarity (LSS), PyramidHOG (PHOG), SIFT, colorSIFT (RGSIFT), and SURF.
- **YouTubeFace** [14] features 101,499 face video samples across 31 categories. The dataset employs five different views extracted from video frames.

The statistical details of these datasets are provided in Table 2.

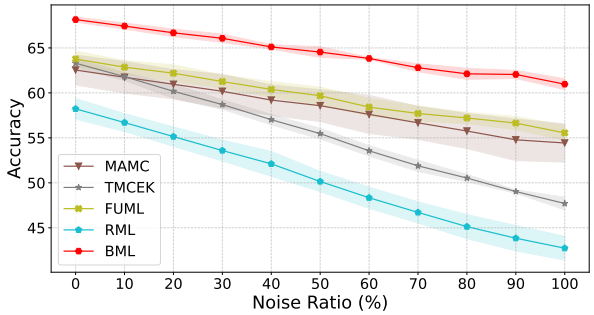


Figure 3. The classification performance comparison results on the SUN R-D-T dataset with noise ratios of 0%-100%, where the solid line represents the average result of five different random seeds, and the shaded area represents the standard deviation.

5. Additional Experiments

5.1. Algorithm

We illustrate the Bootstrapping Multi-view Learning (BML) algorithm for the TNC problem in Algorithm 1.

5.2. Results on varying noise ratios

Figure 3 and Figure 4 report the performance of BML and all baseline methods on the SUN R-D-T dataset and 10 feature-vector datasets, respectively, under varying levels of TNC noise. On most datasets, BML consistently and substantially outperforms all baselines, demonstrating the effectiveness of our method across a broad range of downstream TNC scenarios. Notably, on the large-scale AWA and YouTubeFace datasets, the performance gap between BML and all competing methods further widens as the noise ratio increases. This observation corroborates the task gap discussed in Figure 1 of the main paper, *i.e.*, when the weight estimator is trained only on a clean, well-aligned training set, extrapolating it to noisy test conditions is prone to failure, as the method cannot capture appropriate weights in the presence of TNC noise.

Table 2. Statistical details of the used feature-vector datasets.

Datasets	Categories	Samples	Views	dimensions
Caltech [5]	7	1,400	5	[40, 254, 1984, 512, 928]
Leaves [13]	100	1,600	3	[64, 64, 64]
HW [3]	10	2,000	6	[240, 76, 216, 47, 64, 6]
LandUse [19]	21	2,100	3	[20, 59, 40]
Scene [4]	15	4,485	3	[20, 59, 40]
CCV [8]	20	6,773	3	[20, 20, 20]
Fashion [15]	10	10,000	3	[1×28×28, 1×28×28, 1×28×28]
NUS-OBJ [2]	31	30,000	5	[65, 226, 145, 74, 129]
AWA [10]	50	30,475	6	[2688, 2000, 252, 2000, 2000, 2000]
YouTubeFace [14]	31	101,499	5	[64, 512, 64, 647, 838]

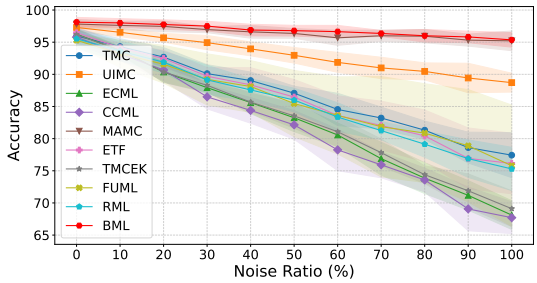
Algorithm 1 Bootstrapping Multi-view Learning (BML) procedure for TNC problem.

Require: Training set $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, encoders $\{f(\cdot;\theta_m)\}_{m=1}^{M}$, classifiers $\{g(\cdot;\phi_m)\}_{m=1}^{M}$, reliability estimators $\{E(\cdot;\psi_m)\}_{m=1}^{M}$, augmentation rate ρ, loss coefficient λ, number of epochs T.

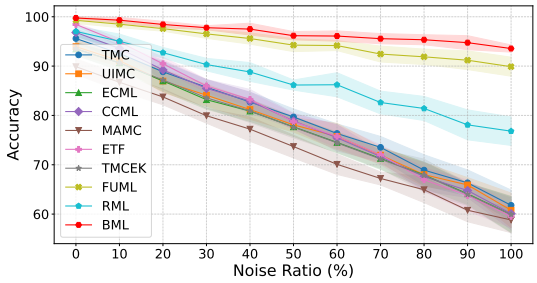
for $t = 1$ **to** T **do**
  // Constructing Bootstrapped TNC-augmented set
  Sample index subset $\tilde{S} \subseteq \{1, \ldots, N\}$ with $|\tilde{S}| = \rho N$.
  for each $i \in \tilde{S}$ **do**
    Sample view mask $\mathbf{s}_i$ constrained by the TNC regime in Eq. (5).
    Construct bootstrapped sample $\breve{\mathbf{x}}_i$ by Eq. (6).
  end for
  Collecting $\{\breve{\mathbf{x}}_i, y_i, \mathbf{s}_i\}_{i=1}^{N}$ and build reveal-supervised set $\breve{\mathcal{D}}$.
  // BML Training
  for mini-batch $\mathcal{B} \subseteq \breve{\mathcal{D}}$ **do**
    for each $(\breve{\mathbf{x}}_i, y_i, \mathbf{s}_i) \in \mathcal{B}$ and $m = 1, \ldots, M$ **do**
      Encode features $\breve{\mathbf{z}}_i^{(m)} = f(\breve{\mathbf{x}}_i; \theta_m)$ and obtain logits $\breve{\boldsymbol{\ell}}_i^{(m)} = g(\breve{\mathbf{z}}_i^{(m)}; \phi_m)$ via the per-view models.
      Compute inter-view discrepancy $\breve{J}_i^{(m)}$ and intra-view uncertainty $\breve{Q}_i^{(m)}$ by Eq. (10) and Eq. (11).
      Form view evidence $\breve{\mathbf{u}}_i^{(m)}$ by Eq. (12) and get the reliability score $\alpha_i^{(m)} = E(\breve{\mathbf{u}}_i^{(m)}; \psi_m)$.
      Update all parameters $\{\theta_m, \phi_m, \psi_m\}_{m=1}^{M}$ by the joint objective $\mathcal{L} = \mathcal{L}_{cls} + \lambda \mathcal{L}_w$.
    end for
  end for
end for
// Inference
for $m = 1$ **to** M **do**
  Encode features $\mathbf{z}_i^{(m)} = f(\mathbf{x}_i; \theta_m)$ and obtain logits $\boldsymbol{\ell}_i^{(m)} = g(\mathbf{z}_i^{(m)}; \phi_m)$ via the per-view models.
  Compute discrepancy $J_i^{(m)}$ and uncertainty $Q_i^{(m)}$ by Eq. (10) and Eq. (11).
  Form evidence $\mathbf{u}_i^{(m)}$ by Eq. (12) and get the reliability score $\alpha_i^{(m)} = E(\mathbf{u}_i^{(m)}; \psi_m)$.
end for
Fuse logits with the learned reliabilities and output final prediction $\hat{y}_i$ via Eq. (15).

5.3. Ablation results on SUN R-D-T

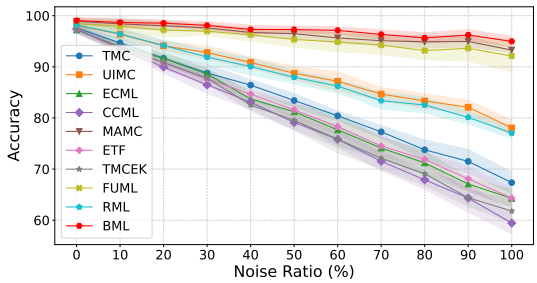
To provide an in-depth analysis of the components in BML, Table 3 presents an ablation study of several BML variants on the SUN R-D-T dataset under different noise ratios, as a complement to Table 3 in the main paper. Consistent with the observations in the main paper, BML exhibits the effectiveness of each of its components across all noise levels. When we do not on-the-fly resample the bootstrapped pool $\widetilde{S}$ at each training epoch, the task gap is not reduced; instead, the model quickly overfits the TNC noise patterns, leading to a substantial performance drop. Similarly, discarding the reliability-learning loss $\mathcal{L}_w$ and using a naïve classifier alone also results in a marked degradation in performance. For the refined input to $E(\cdot;\psi)$ that we propose, the inter-view



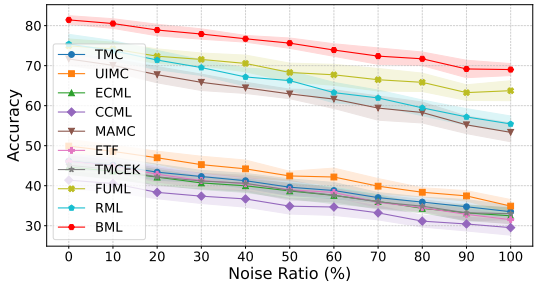
(a) Caltech



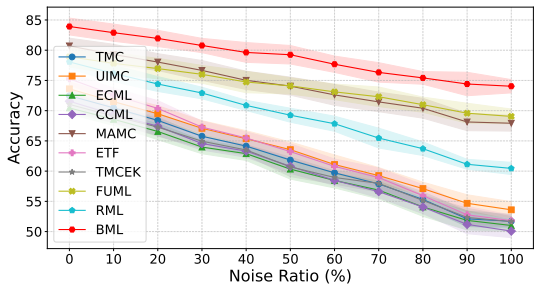
(b) Leaves



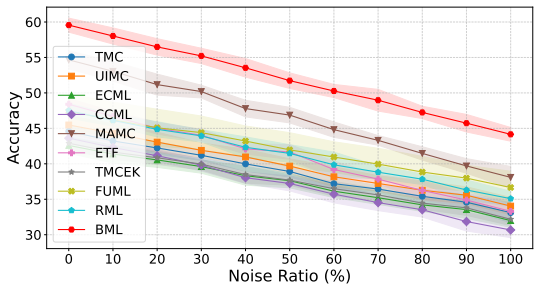
(c) HW



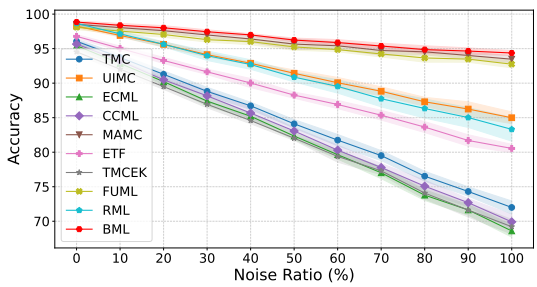
(d) LandUse



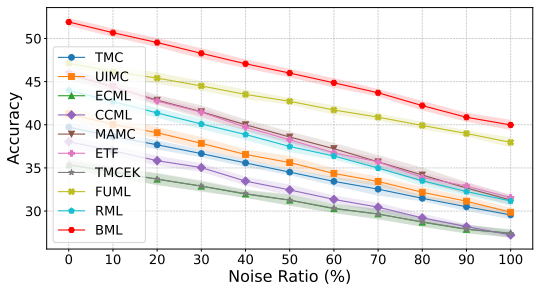
(e) Scene



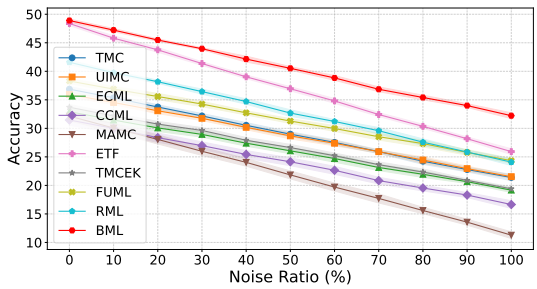
(f) CCV



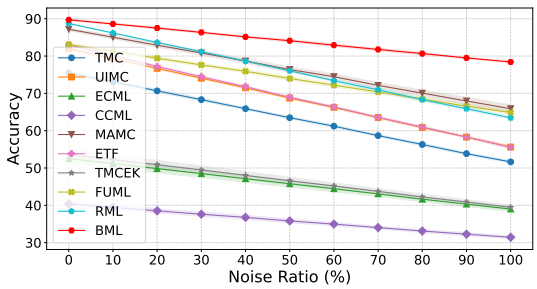
(g) Fashion



(h) NUS-OBJ



(a) AWA

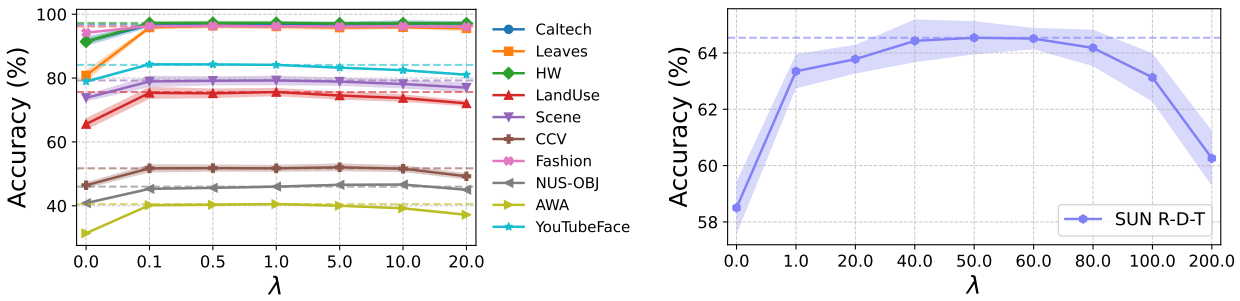


(b) YouTubeFace

Figure 4. The classification performance comparison results on the Feature-vector type dataset with noise ratios of 0%-100%, where the solid line represents the average result of ten different random seeds, and the shaded area represents the standard deviation.

Table 3. Ablation results of BML at various noise ratios on the SUN R-D-T dataset.

Noise	W/O $\mathcal{L}_w$	W/O J	W/O Q	W/O on-the-fly	FULL
0%	64.51±0.48	67.02±0.50	67.57±0.27	63.30±0.48	**68.15±0.28**
10%	63.28±0.74	66.01±0.53	66.79±0.28	61.97±0.54	**67.43±0.34**
20%	62.15±0.49	65.03±0.57	66.18±0.25	60.44±0.54	**66.68±0.42**
30%	60.72±0.69	63.82±0.30	65.39±0.29	58.79±0.42	**66.07±0.46**
40%	59.65±0.42	62.72±0.58	64.53±0.31	57.63±0.36	**65.11±0.25**
50%	58.50±0.93	61.76±0.22	64.20±0.55	56.27±0.93	**64.54±0.59**
60%	56.93±0.63	60.46±0.44	63.10±0.51	54.90±0.34	**63.83±0.11**
70%	55.14±0.78	59.14±0.18	62.26±0.31	53.00±0.69	**62.80±0.45**
80%	54.43±0.97	58.22±0.77	61.78±0.47	51.75±0.58	**62.12±0.61**
90%	53.17±0.72	57.45±0.42	60.96±0.54	50.25±0.58	**62.05±0.45**
100%	52.07±1.00	56.80±0.46	60.78±0.61	48.95±0.94	**60.97±0.60**



(a) Feature-vector datasets (b) SUN R-D-T

Figure 5. Parameter analysis on the loss coefficient λ, where the dashed line represents the performance of the values used in BML.

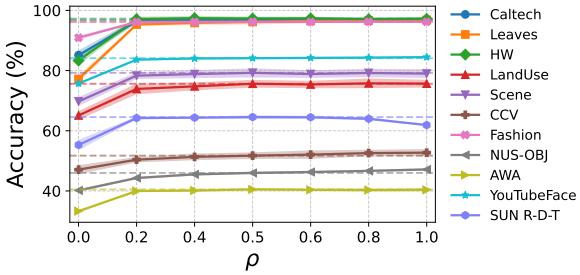


Figure 6. Parameter analysis on the noise-augmented proportion ρ, where the dashed line represents the performance of the values used in BML.

prediction discrepancy signal J is more critical than the intra-view prediction uncertainty signal Q. This is expected: J is specifically designed to capture TNC patterns, and therefore provides more informative guidance for weight assignment during fusion than Q, which only serves as a proxy for individual view quality.

5.4. Parameter Analysis

Figure 5 and Figure 6 present the sensitivity analysis of two key hyperparameters in BML: the loss coefficient λ and the noise-augmentation ratio ρ.

Balance of $\mathcal{L}_{cls}$ and $\mathcal{L}_w$ (λ). For the loss coefficient λ, the performance curves of nearly all datasets follow a rise-then-fall trend. When λ is too small, the contribution of the TNC-oriented augmentation is underweighted, and BML cannot fully exploit it to identify TNC samples. When λ is too large, the classification objective $\mathcal{L}_{cls}$ fails to converge and the model underfits, which in turn harms performance on downstream tasks.

Noise-augmented proportion (ρ). For the noise-augmentation ratio ρ, we observe that, for almost all datasets, introducing augmentation consistently improves performance under downstream TNC corruption,

since it effectively narrows the task gap. However, on some datasets (*e.g.*, SUN R-D-T), an excessively large augmentation ratio can inject too many spurious noisy correspondences into training. This is because SUN R-D-T is a raw dataset whose backbone model is more complex and trains more slowly than the backbones used on feature-vector datasets, making it less robust to aggressive augmentation. Taking this into account, we set the default noise-augmentation ratio in BML to $\rho = 0.5$, which works reliably across different types of datasets.

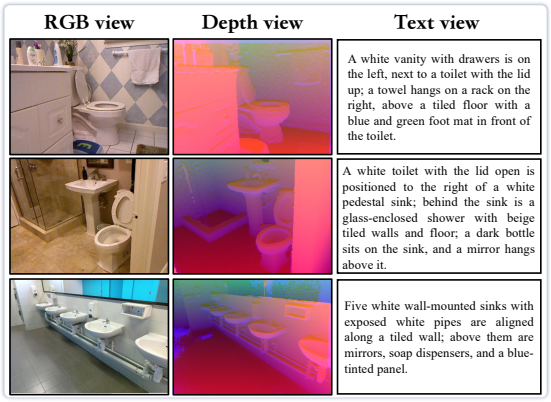


Figure 7. Bathroom

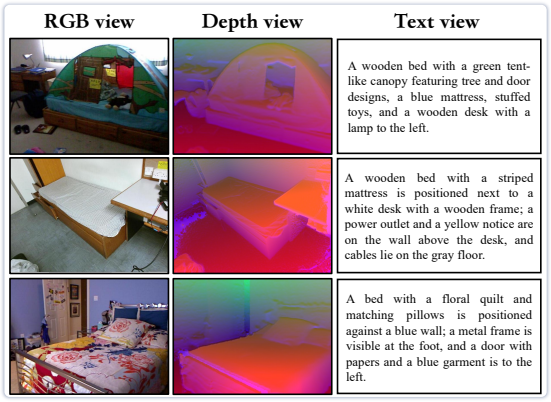


Figure 8. Bedroom

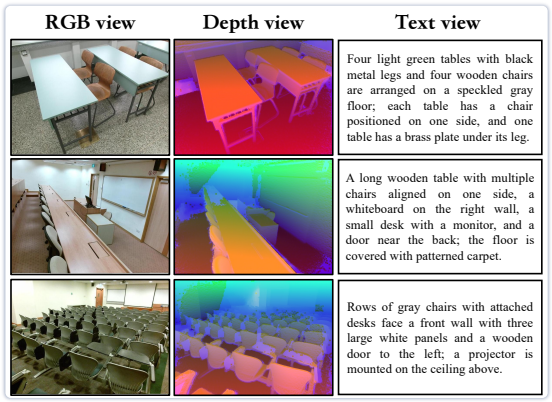


Figure 9. Classroom

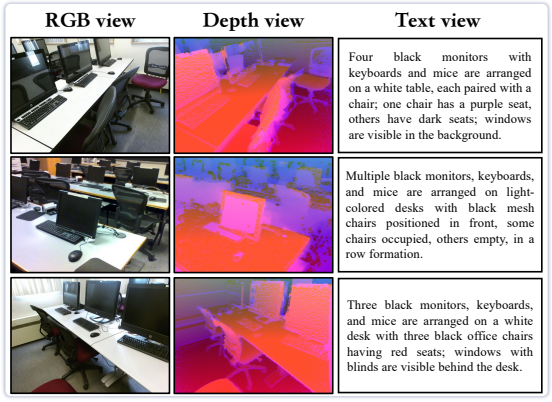


Figure 10. Computer room

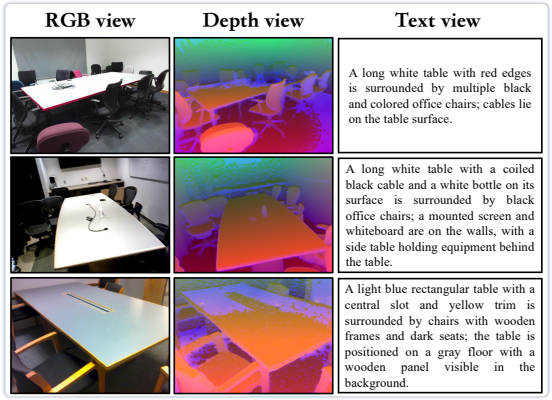


Figure 11. Conference room

References

[1] Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. In *International Conference on Machine Learning*, pages 5608–5628. PMLR, 2024. 2

[2] 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. 4, 5

[3] Robert Duin. Multiple Features. UCI Machine Learning Repository, 1998. 4, 5

[4] Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *2005*

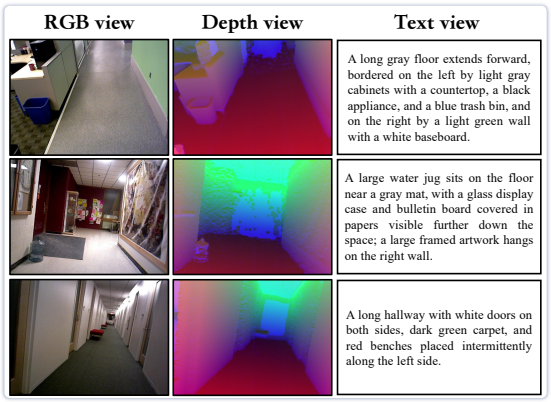


Figure 12. Corridor

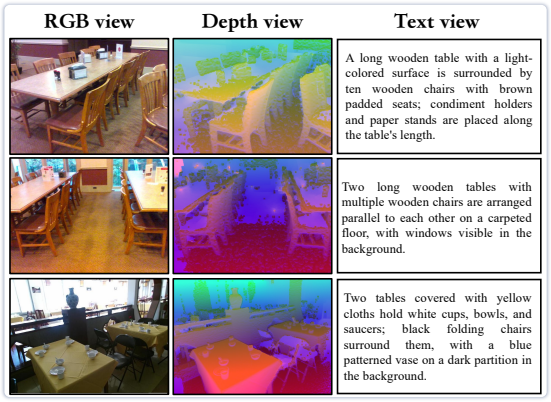


Figure 13. Dining area

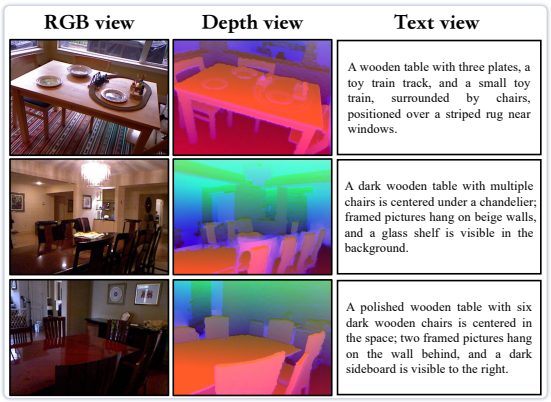


Figure 14. Dining room

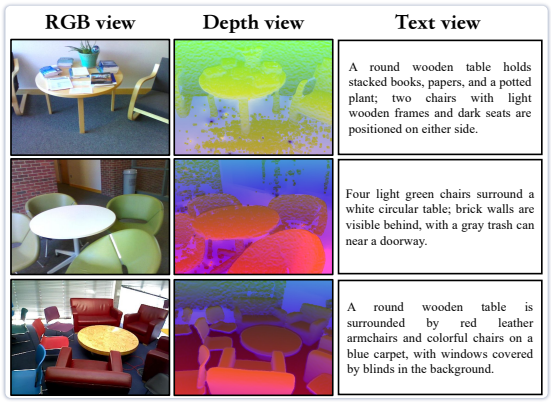


Figure 15. Discussion area

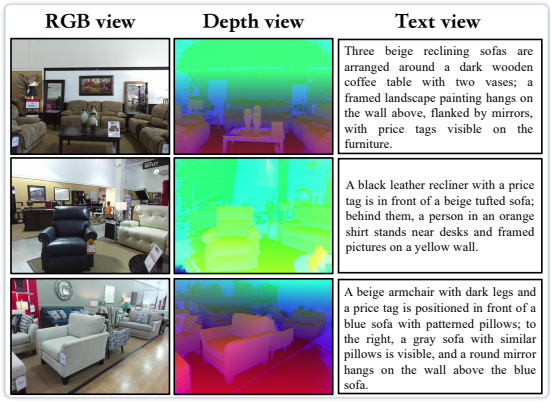


Figure 16. Furniture store

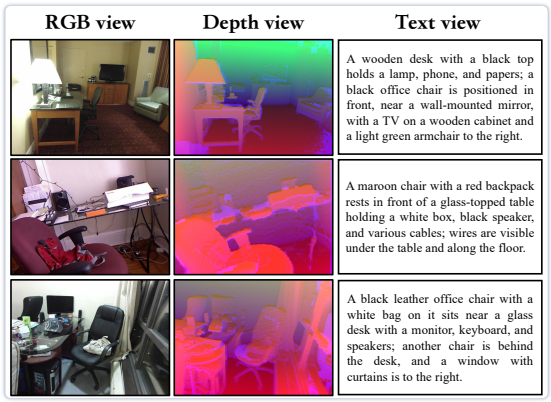


Figure 17. Home office

IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), pages 524–531. IEEE, 2005. 4, 5

[5] 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. 4, 5

[6] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification with dynamic evidential fusion. *IEEE transactions on*

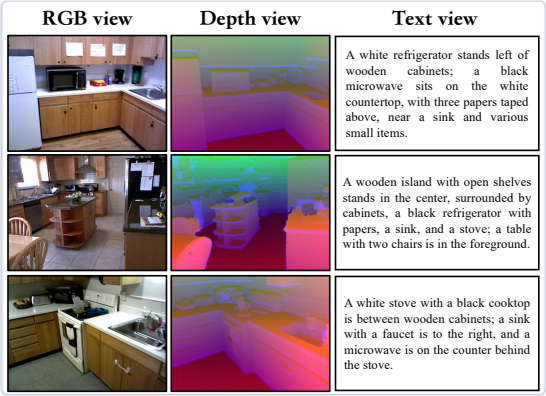


Figure 18. Kitchen

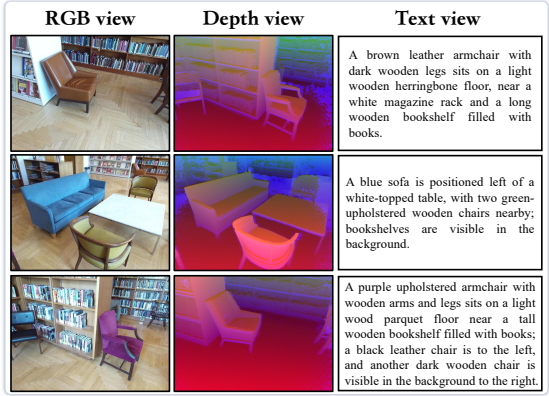


Figure 21. Library

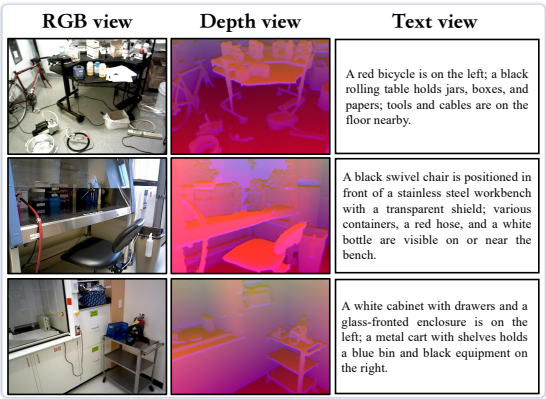


Figure 19. Lab

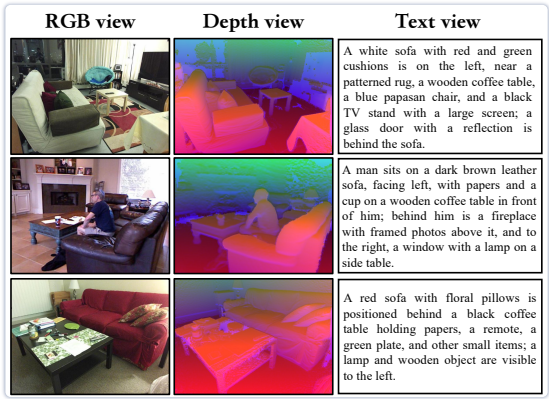


Figure 22. Living room

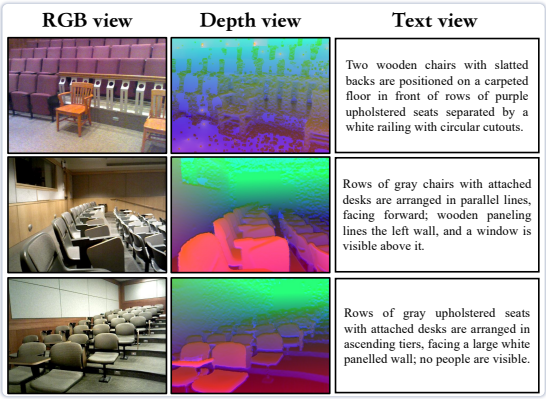


Figure 20. Lecture theatre

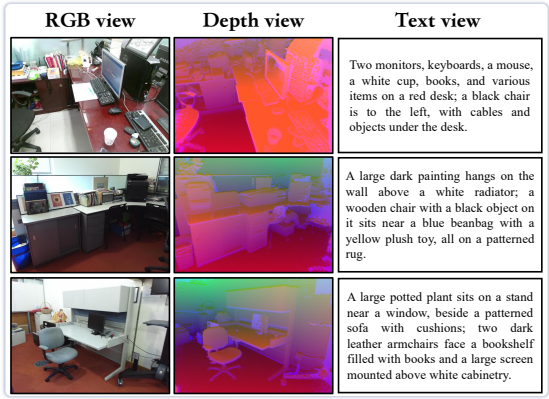


Figure 23. Office

pattern analysis and machine intelligence, 45(2):2551–2566, 2022. 2

[7] 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. 2

[8] 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*

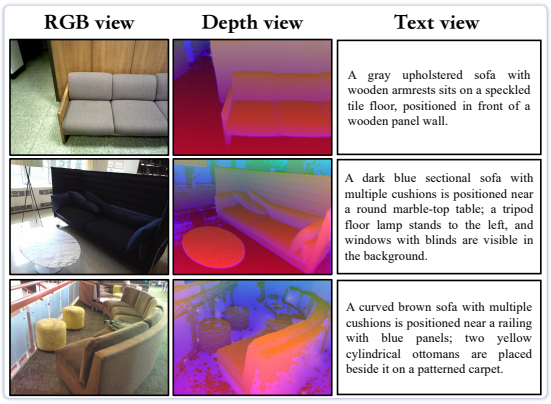


Figure 24. Rest space

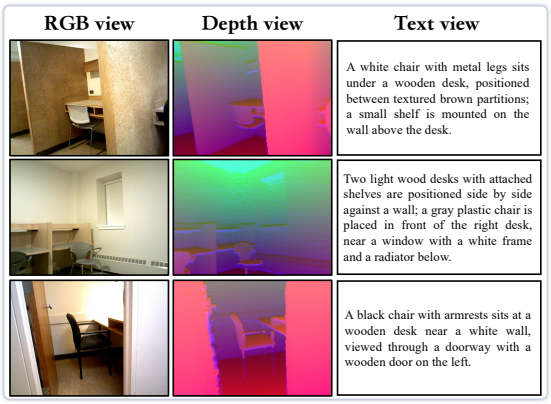


Figure 25. Study space

ACM international conference on multimedia retrieval, pages 1–8, 2011. 4, 5

[9] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023. 2

[10] 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. 4, 5

[11] 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. 2

[12] 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. 2

[13] 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. 4, 5

[14] Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pages 529–534. IEEE, 2011. 4, 5

[15] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 4, 5

[16] 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. 2

[17] Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16051–16060, 2022. 4

[18] 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. 2

[19] 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. 4, 5

[20] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR, 2023. 2