INSTANCE-DEPENDENT NOISE ROBUST HASHING FOR MULTI-LABEL IMAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Supervised hashing has achieved remarkable progress in image retrieval, yet its performance heavily relies on high-quality labels and is therefore vulnerable to label noise. Existing robust hashing methods mainly address class-conditional noise (CCN), while instance-dependent noise (IDN), which better reflects real-world noise patterns, remains unexplored in multi-label hashing retrieval. We identify two fundamental corruptions of pairwise supervision caused by multi-label IDN: additive IDN introduces spurious semantic similarities, causing false attraction, whereas subtractive IDN destroys genuine semantic correlations, resulting in false repulsion. To address them, we propose **I**nstance-**D**ependent **N**oise **R**obust **H**ashing (**IDNRH**). It consists of two complementary components: Sample-to-Label Truncation (SLT), which uses a GMM to partition potentially noisy samples, reconstructs label confidence through multi-view neighbor consensus, and truncates unreliable labels to mitigate false attraction; and Repulsion-Filtered Soft Contrastive Learning (RFSCL), which suppresses false repulsion using shared-neighbor similarity while adaptively modulating attraction with soft label similarity. Extensive experiments on widely used datasets and a real-world noisy dataset demonstrate that IDNRH consistently outperforms state-of-the-art hashing methods under multi-label IDN.

1 INTRODUCTION

Driven by the explosive growth of digital media, the influx of massive image data across the internet has significantly complicated efficient information retrieval. Content-based image retrieval tackles this challenge by analyzing underlying visual features. For scalable and efficient large-scale search under data-intensive environments, image hashing transforms high-dimensional visual representations into a shared Hamming space through compact binary codes, thereby enabling rapid retrieval with minimal computational and storage overhead (Yuan et al., 2020; Hoe et al., 2021).

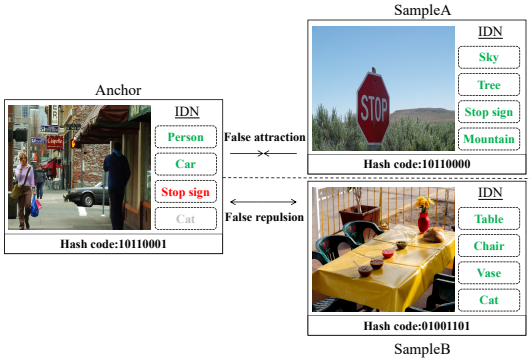


Figure 1: Multi-label instance-dependent noise. Green indicates correct labels, red additive IDN, and gray subtractive IDN.

Supervised image hashing generally outperforms unsupervised approaches because explicit semantic guidance facilitates the learning of discriminative hash codes, whereas unsupervised methods, despite eliminating labeling costs, lack such guidance (Peng et al., 2023; Sun et al., 2022; Shen et al., 2022). However, this advantage relies heavily on clean and fully annotated training data, while real-world labels are predominantly acquired through human labeling and inevitably contain label noise. Such noise is commonly categorized into class-conditional noise (CCN), where corruption depends only on class categories, and instance-dependent noise (IDN), where corruption is induced by instance-specific visual features; recent studies further show that real-world label errors are often better characterized by IDN (Chen et al., 2021; Xia et al., 2020; Cheng et al., 2021). As illustrated by the anchor in Figure 1, multi-label IDN causes two severe problems for retrieval: a red pedestrian signal can be incorrectly annotated as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

a "Stop Sign," creating additive IDN and false attraction between the anchor and Sample A, whereas an inconspicuous "Cat" can be missed, producing subtractive IDN and false repulsion between the anchor and Sample B. Consequently, these noise-corrupted pairwise relations induce misleading hash code similarities, ultimately impairing retrieval ranking performance.

Nevertheless, existing noise-robust hashing methods are still mainly designed for CCN, largely overlooking the more realistic IDN setting. Specifically, DIOR performs sample-level reliability partitioning via perturbed views (Wang et al., 2024a), while SEGA calibrates noisy supervision using dynamic class prototypes and instance uncertainty (Gu et al., 2025). However, multi-label IDN often corrupts only a subset of labels within an instance, making sample-level reliability insufficient to localize corrupted labels. CogniTrust further refines label reliability with structural evidence and semantic prototypes (Gu et al., 2026), yet does not explicitly address pairwise relation corruption caused by missing labels, allowing subtractive IDN to misidentify semantically related samples as negatives and induce false repulsion. Moreover, IDN has been actively studied in image classification and has recently begun to extend toward multi-label classification, but most existing solutions remain designed for single-label classification (Chen et al., 2023; Hou et al., 2025; Heidari et al., 2025). Their transfer to multi-label hashing retrieval is non-trivial: beyond handling multiple labels, classification targets accurate label prediction, whereas hashing retrieval must learn compact binary codes that preserve discriminative and ranking-aware pairwise similarity.

To overcome these limitations, we propose Instance-Dependent Noise Robust Hashing (IDNRH), a method that jointly leverages Sample-to-Label Truncation (SLT) and Repulsion-Filtered Soft Contrastive Learning (RFSCL) to enhance retrieval robustness under multi-label IDN. Specifically, SLT first performs a sample-level partition driven by the small-loss criterion to isolate potentially noisy instances. It then aggregates reliable neighbors via multi-view consistency to dynamically reconstruct trustworthy soft label confidences, and selectively truncates low-confidence labels in noisy samples. RFSCL builds pairwise supervision from the resulting truncated labels. For sample pairs with low label overlap but high shared-neighbor similarity, it filters the repulsion strength using neighbor-set overlap, while the attraction term preserves graded semantic similarity through the truncated-label soft similarity matrix. This synergistic optimization mitigates false attraction caused by additive IDN and false repulsion caused by subtractive IDN.

The main contributions of this paper are: ❶ We study the unexplored problem of image hashing retrieval under multi-label IDN. We characterize multi-label IDN through two fundamental forms: additive IDN and subtractive IDN. These two types of noise affect hashing supervision in different ways, providing a principled basis for addressing their distinct effects. ❷ We propose a novel integrated method, IDNRH. It combines SLT and RFSCL to address the effects of multi-label IDN. SLT mitigates false attraction through sample-level partition, multi-view neighbor aggregation, and selective label truncation. RFSCL mitigates false repulsion through a shared-neighbor filter, while soft attraction preserves graded semantic relationships. ❸ Extensive experiments on three widely used benchmark datasets and one real-world noisy dataset demonstrate consistent improvements over state-of-the-art baselines.

2 Related Work

Learning from noisy labels. Learning from noisy labels (LNL) aims to train reliable models when supervision is corrupted (Song et al., 2023). Label noise is commonly categorized into CCN, where the corruption depends on the class, and IDN, where the corruption is related to the visual content of each instance (Chen et al., 2021; Xia et al., 2020; Berthon et al., 2021). Classical LNL methods mainly address the CCN setting (Ghosh et al., 2017; Zhang & Sabuncu, 2018). Early studies reveal the memorization effect of deep networks, i.e., clean patterns are fitted before noisy labels (Arpit et al., 2017; Liu et al., 2020). Based on this observation, sample-selection methods such as Co-teaching select small-loss samples for robust training (Han et al., 2018; Jiang et al., 2018), while DivideMix models the loss distribution with a Gaussian mixture model and treats noisy-label learning as semi-supervised learning (Li et al., 2020). Recent studies show that real-world label errors are often better described by IDN (Chen et al., 2021; Xia et al., 2020; Cheng et al., 2021). LRA-Diffusion formulates label denoising from a generative perspective and augments diffusion models with retrieved label information (Chen et al., 2023); DLD further introduces directional label diffusion to model sample-specific noise transitions (Hou et al., 2025); and Learning to Clean

casts noisy-label correction as a reinforcement learning problem and learns a policy for iterative correction (Heidari et al., 2025). Beyond single-label classification, MLDN models feature- and label-dependent noise in multi-label data, while RIDE combines label-matrix decomposition with an explicit feature-induced IDN mapping (Wang et al., 2024b; Wu et al., 2025). Nevertheless, these classification-oriented methods recover or predict labels rather than learn binary codes and preserve ranking-aware pairwise similarity; consequently, multi-label IDN remains unexplored in robust hash learning.

Hashing for image retrieval. Hashing for image retrieval can be categorized from two perspectives: supervision signals and noise learning. From the supervision perspective, unsupervised methods learn compact codes without manual labels, therefore IDN is not involved. Recent methods maximize bit entropy to obtain balanced binary codes (Li & van Gemert, 2021), use a contrastive information bottleneck to preserve discriminative information (Qiu et al., 2021), mine semantic concepts as pseudo-supervision (Tu et al., 2023), or calibrate similarity distributions to mitigate similarity collapse in the hash space (Ng et al., 2023). Nevertheless, without explicit semantic supervision, unsupervised hashing generally underperforms supervised hashing. Supervised methods directly exploit labels or pairwise relations to learn discriminative hash codes (Cao et al., 2017): CSQ maps samples to category centers to ease quantization (Yuan et al., 2020), OrthoHash optimizes hashing with a unified cosine-similarity objective (Hoe et al., 2021), and DRMH models dependencies among multiple labels to preserve fine-grained semantic relations (Peng et al., 2023). From the noise-learning perspective, conventional supervised hashing assumes reliable labels, while noise-robust hashing seeks to reduce the influence of corrupted labels (Gu et al., 2025; 2026). DIOR separates reliable and unreliable supervision through dual partition and contrastive learning (Wang et al., 2024a); SEGA shapes the semantic geometry of the hash space with dynamic prototypes and soft uncertainty calibration (Gu et al., 2025). However, existing noise-robust hashing methods are mostly built on the class-conditional noise assumption, which rarely holds in real-world data.

3 Preliminaries

For ease of presentation, we first give some definitions of multi-label IDN hashing. Given a multi-label training set with N images, $\mathcal{D} = \{(x_i, \bar{y}_i)\}_{i=1}^{N}$, where x_i denotes the i-th image and $\bar{y}_i \in \{0,1\}^C$ is its observed multi-hot label vector over C classes. Let $y_i \in \{0,1\}^C$ denote the latent clean label vector, which is unavailable during training. We consider the IDN setting, where the label corruption probability depends on the visual content of each individual instance x_i.

For the c-th class of sample x_i, we define the instance-dependent probabilities of additive and subtractive IDN as $\eta_{ic}^{+} = P(\bar{y}_{ic} = 1 \mid y_{ic} = 0, x_i)$ and $\eta_{ic}^{-} = P(\bar{y}_{ic} = 0 \mid y_{ic} = 1, x_i)$, respectively. η_{ic}^{+} denotes the probability that an absent label is incorrectly added, namely *additive IDN*, while η_{ic}^{-} denotes the probability that a true label is incorrectly removed, namely *subtractive IDN*. Accordingly, the probability of observing a positive label for class c can be written as:

$$P(\bar{y}_{ic} = 1 \mid y_{ic}, x_i) = (1 - y_{ic})\eta_{ic}^{+} + y_{ic}(1 - \eta_{ic}^{-}). \tag{1}$$

Since η_{ic}^{+} and η_{ic}^{-} vary with x_i, the same class may exhibit different corruption probabilities across different images. Therefore, mitigating the erroneous pairwise relationships caused by η_{ic}^{+} and η_{ic}^{-} is crucial for robust multi-label hashing under IDN.

We aim to learn a deep hashing function $\Phi(\cdot)$ that maps images into a compact B-bit Hamming space. For image x_i, weak and strong augmentations produce x_i^w and x_i^s, whose ℓ_2-normalized continuous hash representations are $u_i^w = \Phi(x_i^w)$ and $u_i^s = \Phi(x_i^s)$. The final binary hash code is obtained as $b_i = \text{sign}(u_i^w) \in \{-1, +1\}^B$. Meanwhile, a classification head parameterized by $W \in \mathbb{R}^{B \times C}$ produces the class prediction probabilities $z_i = \sigma(u_i^w W) \in [0,1]^C$, where $\sigma(\cdot)$ denotes the element-wise sigmoid function.

4 Methodology

Figure 2 gives an overview of the IDNRH method. It integrates SLT and RFSCL to improve the robustness of multi-label hash learning under IDN.

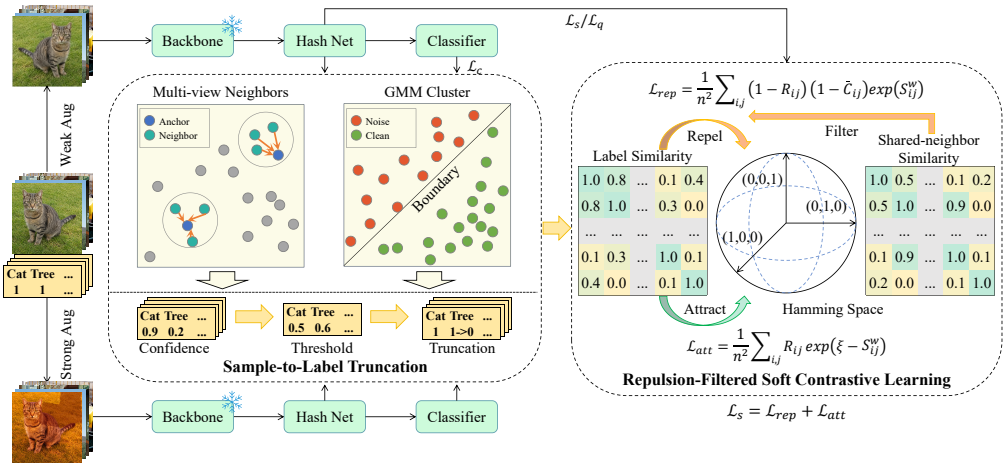


Figure 2: The illustration of the proposed IDNRH. Weak and strong views are encoded into features for pointwise and pairwise supervision. SLT turns sample partition guided by a Gaussian mixture model (GMM) and multi-view neighbor consensus into truncated labels. RFSCL builds pairwise supervision from these labels: shared-neighbor similarity filters false repulsion caused by subtractive IDN, whereas truncated-label soft similarity mitigates false attraction caused by additive IDN. Classification, soft contrastive, and quantization losses jointly optimize the method.

4.1 Sample-to-Label Truncation

Inspired by Neighbor-Aware Confidence Reconstruction (NACR) (Su et al., 2025), we reconstruct label confidence from neighborhood information to truncate additive IDN. Since additive IDN and neighborhood relations are both feature-dependent, single-view neighbors may be unreliable. We therefore introduce multi-view neighbor consistency to identify more reliable semantic neighbors. A GMM is further used for coarse sample partitioning, restricting label truncation to potentially noisy samples and thus preserving clean supervision.

Coarse-Grained Sample Partition. Following the small-loss criterion (Han et al., 2018), we use the BCE loss of the weak view to identify samples likely to contain noisy labels:

$$l_i = -\frac{1}{C}\sum_{c=1}^{C}\Big(\bar{y}_{ic}\log(z_{ic}) + (1-\bar{y}_{ic})\log(1-z_{ic})\Big). \tag{2}$$

We then fit a two-component GMM to the loss distribution $\{l_i\}_{i=1}^{N}$ and regard the component with the smaller mean as the clean component. The posterior probability $P(\text{clean} \mid l_i)$ is used to partition samples into clean and potentially noisy sets, yielding the binary sample-level mask:

$$m_i = \begin{cases} 1, & \text{if } P(\text{clean} \mid l_i) \geq 0.5, \\ 0, & \text{otherwise.} \end{cases} \tag{3}$$

The mask restricts subsequent truncation to potentially noisy samples ($m_i = 0$), while $P(\text{clean} \mid l_i)$ serves as a weight measuring sample cleanliness for subsequent confidence reconstruction.

Fine-Grained Confidence Aggregation. The coarse-grained sample partition cannot further identify specific label-level errors, while IDN may also make ordinary neighborhood relations unreliable. Therefore, we reconstruct label confidence from multi-view-consistent neighbors. We maintain a soft label confidence matrix $P \in \mathbb{R}^{N\times C}$ and initialize p_i as the observed label $\bar{y}_i$.

Next, we extract continuous features u_i^w and u_i^s from the weak and strong views to compute their respective cosine similarity matrices S^w and S^s, and fuse them as $S^f = \frac{1}{2}(S^w + S^s)$. Based on S^f, we retrieve the top-K most relevant semantic neighbors for x_i, forming the neighborhood set $\mathcal{N}_i$. Neighbor ranking is determined by the fused similarity, while the propagation of neighbor information is further modulated by multi-view consistency:

$$\omega_{ij} = \exp\left(-|S^w_{ij} - S^s_{ij}|\right), \tag{4}$$

where ω_{ij} reflects the consistency of the neighborhood relation between samples i and j across different views, suppressing neighbors whose relations are unstable across views.

Furthermore, the normalized neighbor-information propagation weight a_{ij} jointly considers fused feature similarity and cross-view consistency, allowing neighbors that are semantically closer to the current sample and more stable across different views to receive higher propagation weights:

$$a_{ij} = \frac{S_{ij}^{f}\omega_{ij}}{\sum_{j' \in \mathcal{N}_i} S_{ij'}^{f}\omega_{ij'}}. \tag{5}$$

We further reuse the sample cleanliness probability $P(\text{clean} \mid l_i)$ obtained from the GMM to control the confidence correction strength for each sample. We first represent the neighbor information as a residual correction:

$$\Delta p_i = \sum_{j \in \mathcal{N}_i} a_{ij}(p_j - p_i), \tag{6}$$

whose direction points from the confidence of the current sample toward the weighted neighborhood consensus. We then update the label confidence of the sample as:

$$p_i \leftarrow p_i + \Big(1 - P(\text{clean} \mid l_i)\Big)\Delta p_i. \tag{7}$$

Thus, cleaner samples retain more of their own confidence, whereas potentially noisy samples absorb more stable neighbor information through multi-view neighborhood consensus.

Instance-Aware Selective Truncation. To remove additive IDN while preserving reliable labels, we compute a threshold t_k from the non-zero confidence scores of class k within the clean sample set $\mathcal{C} = \{i \mid m_i = 1\}$:

$$t_k = \frac{\sum_{j \in \mathcal{C}} p_{jk} \cdot \mathbb{I}(p_{jk} > 0)}{\sum_{j \in \mathcal{C}} \mathbb{I}(p_{jk} > 0)}, \tag{8}$$

where $\mathbb{I}(\cdot)$ denotes the indicator function. We then perform label truncation only on samples assigned to the noisy set. The truncation is determined by the reconstructed confidence of each instance. Clean samples retain their original observed labels, while a positive label of a noisy sample is preserved only when its reconstructed confidence reaches the corresponding threshold:

$$\hat{y}_{ik} = \begin{cases} 0, & \text{if } m_i = 0 \text{ and } p_{ik} < t_k, \\ \bar{y}_{ik}, & \text{otherwise.} \end{cases} \tag{9}$$

In this way, SLT converts sample-level partition and multi-view neighborhood consensus into explicit label-level truncation. This process reduces unnecessary truncation on clean samples while removing additive IDN from potentially noisy samples. The resulting truncated labels $\hat{y}$ are then used for subsequent hashing contrastive learning.

4.2 Repulsion-Filtered Soft Contrastive Learning

Although SLT mitigates additive IDN, it neither addresses subtractive IDN nor explicitly constrains the pairwise relations required for hashing retrieval, as commonly modeled in supervised contrastive learning (Khosla et al., 2020; Li et al., 2022). Therefore, we propose RFSCL, consisting of a shared-neighbor-filtered repulsion term and a soft-similarity-guided attraction term, which respectively mitigate excessive repulsion caused by subtractive IDN and preserve graded semantic relations.

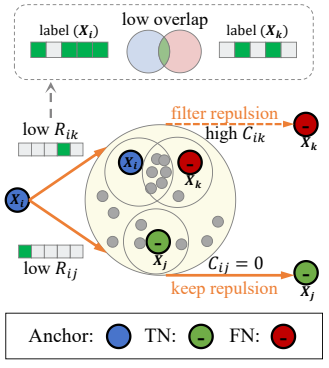


Figure 3: Shared-neighbor filter. It suppresses false repulsion caused by false-negative pairs.

Shared-Neighbor-Filtered Repulsion. We first compute the soft similarity matrix R from the truncated labels $\hat{y}$. The soft label similarity between two samples is defined as:

$$R_{ij} = \frac{|\hat{y}_i \cap \hat{y}_j|}{|\hat{y}_i \cup \hat{y}_j|}. \tag{10}$$

A larger R_{ij} indicates stronger semantic overlap. However, subtractive IDN may remove true positive labels and thus underestimate R_{ij} between semantically related samples, leading to false repulsion. To identify such potentially related samples, we further exploit neighborhood structure. Let $\mathcal{N}_i$ denote the neighborhood set of sample x_i. The shared-neighbor similarity is defined as:

$$C_{ij} = \frac{|\mathcal{N}_i \cap \mathcal{N}_j|}{|\mathcal{N}_i \cup \mathcal{N}_j|}. \tag{11}$$

We then scale the shared-neighbor similarity within the current batch as $\bar{C}_{ij} = \frac{\gamma C_{ij}}{\max_{p,q} C_{pq}}$, where γ controls the filtering strength. Let $S_{ij}^w = (u_i^w)^\top u_j^w$ denote the continuous feature similarity between samples i and j in the weak view. We use the shared-neighbor filtering factor $1 - \bar{C}_{ij}$ to modulate repulsion, and for a mini-batch of n samples, define the repulsion term as:

$$\mathcal{L}_{rep} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} (1 - R_{ij})(1 - \bar{C}_{ij}) \exp(S_{ij}^w). \tag{12}$$

As shown in Figure 3, TN and FN denote true negative and false negative, respectively. Although x_k has low label overlap with the anchor x_i, a high shared-neighbor similarity C_{ik} indicates potential semantic relevance, thus $1 - \bar{C}_{ik}$ suppresses their false repulsion. In contrast, x_j shares no neighbor information with the anchor, yielding $\bar{C}_{ij} = 0$ and preserving the original repulsion strength.

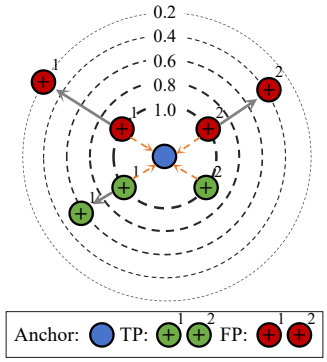


Figure 4: Hard versus soft attraction. Graded attraction mitigates false-positive attraction.

Soft-Similarity-Guided Attraction. The attraction term uses R_{ij} as a continuous weight to provide graded semantic supervision and hierarchical attraction strengths between samples:

$$\mathcal{L}_{att} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} R_{ij} \exp(\xi - S_{ij}^w), \tag{13}$$

where ξ is a constant controlling the base attraction strength and is set to 1 by default. As shown in Figure 4, the dashed circles indicate attraction strength, while the gray arrows connect the strength of each sample pair under hard and soft attraction. Hard attraction applies the same attraction strength to all positive pairs, whereas soft attraction adjusts the strength according to label overlap: true positive (TP) pairs with higher similarity are pulled closer, while false positive (FP) pairs with lower label overlap receive weaker attraction. This preserves finer-grained differences in the Hamming space and mitigates false attraction caused by additive IDN.

Finally, the RFSCL objective combines the two complementary terms as $\mathcal{L}_s = \mathcal{L}_{rep} + \mathcal{L}_{att}$. By suppressing unreliable repulsion with shared-neighbor similarity and modulating attraction strength with soft semantic overlap, RFSCL mitigates false repulsion while preserving discriminative and graded structures in the Hamming space.

4.3 Optimization

We aim to employ reconstructed P to guide the hash network training and improve discrimination. Specifically, we employ the following cross-entropy loss for the weakly augmented view $\mathcal{L}_c = -\frac{1}{n} \sum_{i=1}^{n} \sum_{c=1}^{C} \Big(p_{ic} \log(z_{ic}) + (1 - p_{ic}) \log(1 - z_{ic}) \Big)$. In addition, we employ the ℓ_1 loss to control quantization error between the continuous feature and hash code: $\mathcal{L}_q = \frac{1}{n} \sum_{i=1}^{n} \|u_i^w - b_i\|_1$. By combining the classification loss, soft contrastive loss, and quantization loss, the final loss of the proposed IDNRH is defined as follows:

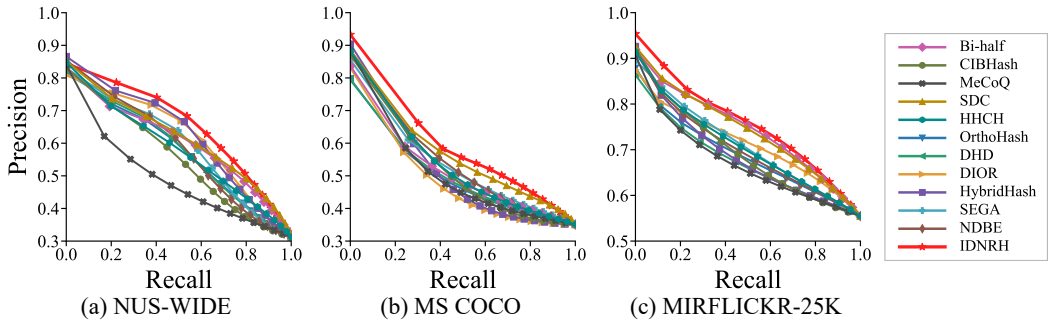
$$\mathcal{L}_{total} = \mathcal{L}_c + \alpha \mathcal{L}_s + \beta \mathcal{L}_q, \tag{14}$$

where the hyperparameters α and β weight the three losses. The training procedure is summarized in Algorithm 1, with details in Appendix A.

5 Experiments

5.1 Experimental Settings

Datasets. We evaluate on four multi-label image retrieval datasets. NUS-WIDE (Chua et al., 2009), MS COCO (Lin et al., 2014), and MIRFLICKR-25K (Huiskes & Lew, 2008) are three widely used benchmarks, with details in Appendix D.1. Following the protocol in (Kim et al., 2024), we inject multi-label IDN at noise rates of 0.2, 0.4, and 0.6; the injection procedure is summarized in Algorithm 2, with details in Appendix A. We also use the real-world noisy dataset NUS-WIDE-Web-21,



(a) NUS-WIDE (b) MS COCO (c) MIRFLICKR-25K

Figure 5: Precision-Recall curves of all the hashing methods under 128-bit hash codes and a 60% noise rate.

whose construction is described in Appendix C.1. Its label noise is further shown to exhibit instance-dependent characteristics from both visual and quantitative evidence, with details in Appendix C.2. **Baselines.** We compare eleven hashing baselines, including five unsupervised methods, i.e., Bi-half (Li & van Gemert, 2021), CIBHash (Qiu et al., 2021), MeCoQ (Wang et al., 2022), SDC (Ng et al., 2023), and HHCH (Wei et al., 2024), and six supervised methods, i.e., OrthoHash (Hoe et al., 2021), DHD (Jang et al., 2022), DIOR (Wang et al., 2024a), HybridHash (He & Wei, 2024), SEGA (Gu et al., 2025), and NDBE (Qin et al., 2026), with details in Appendix D.2. **Implementation.** The frozen CLIP ViT-B/32 (Radford et al., 2021) and a hash projection head are employed in the proposed method, and the details are provided in Appendix D.3. **Evaluation metric.** The mean Average Precision (mAP) and Precision-Recall (PR) curves are employed for performance evaluation.

5.2 Comparison with State-of-the-Art Methods

Synthetic label-noise benchmarks. Table 1 leads to the following key findings. ❶ IDNRH achieves the best mAP in all 36 experimental settings, indicating that its performance advantage is not limited to specific datasets, noise rates, or code lengths. ❷ Even under a severe noise rate of 60%, IDNRH still maintains a clear performance advantage, demonstrating strong robustness against severe instance-dependent label noise. ❸ As the code length increases, the performance of IDNRH generally improves and remains highly competitive with longer hash codes, indicating that it can effectively exploit the larger Hamming space for semantic representation. ❹ At low and moderate noise rates, supervised methods are generally more competitive because the labels still provide relatively reliable semantic supervision; as the noise rate increases, their dependence on labels gradually becomes a limitation, while unsupervised methods become relatively more competitive. Figure 5 shows PR results consistent with the previously reported mAP results. IDNRH maintains higher precision across broad recall ranges on all three datasets. Additionally, to exclude backbone effects, we conduct a backbone-controlled comparison in Appendix B.1.

Real-world label-noise benchmark. Figure 6 shows that IDNRH achieves the highest mAP at all four code lengths on the real-world NUS-WIDE-Web-21 dataset. Overall, unsupervised methods perform better than supervised baselines on this dataset. This trend is similar to that observed under high noise rates on the three synthetic label-noise benchmarks, where unsupervised methods become more competitive. This suggests that NUS-WIDE-Web-21 contains a substantial level of IDN.

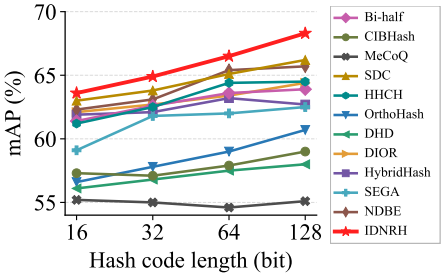


Figure 6: mAP (%) on NUS-WIDE-Web-21 across four code lengths.

5.3 Further Analysis

Ablation Study. To verify the contribution of each component, we construct four variants by removing SLT, removing the GMM-based sample partition, removing the shared-neighbor (SN) filter, and replacing soft attraction with hard attraction. Table 2 shows that the full model achieves the best performance in all settings. More importantly, the contributions of different components vary with

Table 1: mAPs (%) on the three benchmarks. The best and second best are in bold and underlined.

Dataset	Method	Reference	20%				40%				60%			
			16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit	16 bit	32 bit	64 bit	128 bit
NUS-WIDE	Bi-half	AAAI'21	61.2	63.5	64.1	64.1	61.2	63.5	64.1	64.1	61.2	63.5	64.1	64.1
	CIBHash	IJCAI'21	54.4	55.2	56.4	57.9	54.4	55.2	56.4	57.9	54.4	55.2	56.4	57.9
	MeCoQ	AAAI'22	50.3	50.2	50.1	50.0	50.3	50.2	50.1	50.0	50.3	50.2	50.1	50.0
	SDC	BMVC'23	58.4	61.1	63.3	64.8	58.4	61.1	63.3	64.8	58.4	61.1	63.3	64.8
	HHCH	TIP'24	51.4	54.7	58.1	60.7	51.4	54.7	58.1	60.7	51.4	54.7	58.1	60.7
	OrthoHash	NeurIPS'21	65.2	69.2	71.4	72.6	62.7	66.7	68.9	70.1	56.3	59.0	60.2	62.2
	DHD	ECCV'22	66.0	69.0	70.5	71.7	63.0	65.2	67.2	67.6	56.5	58.0	58.8	59.5
	DIOR	TKDE'24	68.1	69.5	72.7	74.6	67.2	68.3	71.6	73.0	61.5	63.7	64.9	66.9
	HybridHash	ICMR'24	73.0	73.7	74.9	75.4	69.3	70.4	72.1	73.4	61.3	63.6	64.6	65.5
	SEGA	NeurIPS'25	70.9	71.3	73.5	74.1	66.4	68.7	70.0	70.5	60.8	61.1	62.6	64.1
	NDBE	TMM'26	70.1	71.6	73.4	74.9	68.5	70.8	71.4	71.9	57.9	59.5	61.0	62.3
	IDNRH	**Ours**	**74.7**	**75.6**	**76.4**	**77.9**	**70.4**	**72.8**	**73.5**	**75.4**	**63.4**	**66.5**	**69.0**	**69.8**
MS COCO	Bi-half	AAAI'21	52.3	54.8	54.9	54.8	52.3	54.8	54.9	54.8	52.3	54.8	54.9	54.8
	CIBHash	IJCAI'21	53.6	55.5	56.4	57.1	53.6	55.5	56.4	57.1	53.6	55.5	56.4	57.1
	MeCoQ	AAAI'22	54.1	54.0	53.2	53.9	54.1	54.0	53.2	53.9	54.1	54.0	53.2	53.9
	SDC	BMVC'23	54.6	57.3	60.0	60.4	54.6	57.3	60.0	60.4	54.6	57.3	60.0	60.4
	HHCH	TIP'24	51.3	53.6	57.2	59.3	51.3	53.6	57.2	59.3	51.3	53.6	57.2	59.3
	OrthoHash	NeurIPS'21	51.6	56.3	61.2	63.1	46.6	51.0	56.1	58.0	44.7	48.7	52.3	53.3
	DHD	ECCV'22	56.5	60.4	63.3	65.4	52.3	56.1	58.4	60.3	48.4	51.4	52.8	53.9
	DIOR	TKDE'24	63.9	64.6	67.2	69.1	57.2	58.9	62.2	64.0	48.9	49.2	50.7	51.3
	HybridHash	ICMR'24	63.6	66.9	69.0	70.1	57.2	61.4	63.0	63.8	47.9	50.5	54.1	56.3
	SEGA	NeurIPS'25	61.2	62.3	64.4	65.8	57.5	58.4	60.1	61.5	49.9	51.5	54.4	56.4
	NDBE	TMM'26	63.8	65.8	68.3	70.2	58.3	60.7	63.1	63.6	52.4	53.5	55.9	57.9
	IDNRH	**Ours**	**64.9**	**68.1**	**70.0**	**71.3**	**60.3**	**62.8**	**64.7**	**67.8**	**55.3**	**59.1**	**60.7**	**62.8**
MIRFLICKR-25K	Bi-half	AAAI'21	68.8	70.7	72.2	72.7	68.8	70.7	72.2	72.7	68.8	70.7	72.2	72.7
	CIBHash	IJCAI'21	67.1	68.2	69.2	70.1	67.1	68.2	69.2	70.1	67.1	68.2	69.2	70.1
	MeCoQ	AAAI'22	67.2	66.9	67.3	67.5	67.2	66.9	67.3	67.5	67.2	66.9	67.3	67.5
	SDC	BMVC'23	69.1	70.5	72.5	72.4	69.1	70.5	72.5	72.4	69.1	70.5	72.5	72.4
	HHCH	TIP'24	67.8	68.4	70.5	71.8	67.8	68.4	70.5	71.8	67.8	68.4	70.5	71.8
	OrthoHash	NeurIPS'21	73.6	75.8	78.8	79.3	70.4	72.5	74.3	74.8	66.8	69.2	68.7	69.8
	DHD	ECCV'22	75.1	77.0	78.7	79.3	70.8	72.2	73.9	74.5	64.9	67.3	68.2	68.3
	DIOR	TKDE'24	71.9	72.8	74.5	76.1	70.6	70.9	72.5	74.2	67.5	68.4	70.1	72.2
	HybridHash	ICMR'24	82.5	83.1	85.5	86.3	75.3	78.1	78.8	79.6	64.7	65.8	67.3	69.7
	SEGA	NeurIPS'25	76.8	77.1	79.6	80.2	73.2	73.8	75.9	77.0	68.4	69.0	70.9	72.1
	NDBE	TMM'26	78.1	79.3	81.5	82.8	74.5	74.9	76.1	77.8	67.4	68.3	70.1	71.1
	IDNRH	**Ours**	**83.5**	**85.4**	**86.3**	**87.1**	**78.8**	**80.4**	**81.3**	**82.6**	**70.2**	**72.1**	**73.9**	**75.1**

the noise rate. Under relatively low noise, the GMM-based sample partition and shared-neighbor filter play more prominent roles, indicating the importance of avoiding unnecessary correction of relatively reliable samples and suppressing false repulsion. As the noise rate increases, label truncation and soft attraction become increasingly important, with hard attraction causing the largest performance degradation under 60% noise. These results demonstrate that the components address different effects of label noise and complement each other across different noise rates, jointly improving the robustness of the model. The corresponding ablation results on MIRFLICKR-25K are provided in Appendix B.2.

Label truncation analysis. Table 3 compares the label-truncation quality of our SLT and the NACR component of NACD (Su et al., 2025). P, R, and FTR respectively measure the purity of removed labels, the fraction of additive IDN removed, and the fraction of observed true labels removed in error; a lower FTR is better because it indicates less erroneous removal of observed true labels. Detailed definitions are provided in Appendix B.3. Across all three datasets, SLT achieves substantially higher P and lower FTR, indicating more accurate identification of additive IDN with less damage to clean labels. This benefit comes from the GMM-based coarse sample partition and multi-view neighbor consensus, which reduce unnecessary correction and provide more reliable label confidence.

Repulsion suppression analysis. Figure 9 compares the proportions of suppressed repulsion among randomly sampled equal numbers of false-negative (FN) and true-negative (TN) pairs, with details in Appendix B.4. A pair is regarded as having suppressed repulsion when its shared-neighbor similarity satisfies $C_{ij} > 0$, such that the shared-neighbor filter attenuates its repulsion strength. The suppression ratio of FN pairs is substantially higher than that of TN pairs, reaching 12.1, 7.9, and 24.7 times the corresponding TN ratios on NUS-WIDE, MS COCO, and MIRFLICKR-25K, respectively,

Table 2: Ablation mAP (%) results with 128-bit codes. Best and second-best are bold and underlined.

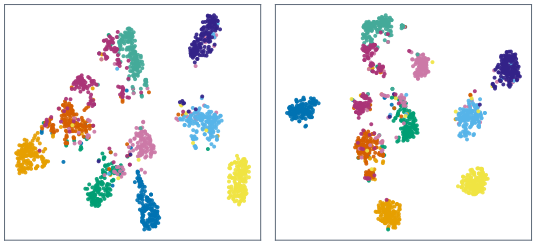
Method	NUS-WIDE			MS COCO		
	20%	40%	60%	20%	40%	60%
w/o SLT	76.7	73.5	66.8	69.9	65.1	59.3
w/o GMM	74.6	72.8	67.9	69.1	<u>66.1</u>	60.9
w/o SN Filter	75.0	<u>74.1</u>	<u>69.0</u>	68.4	64.9	<u>61.3</u>
Hard Attraction	<u>77.7</u>	73.1	65.4	<u>70.9</u>	63.6	57.0
IDNRH	**77.9**	**75.4**	**69.8**	**71.3**	**67.8**	**62.8**

Table 3: Label-truncation quality (%) at 60% noise with 128-bit codes.

Dataset	Module	P ↑	R ↑	FTR ↓
NUS-WIDE	NACR	78.3	35.7	23.4
	SLT	**96.6**	**36.9**	**1.4**
MS COCO	NACR	74.8	38.0	20.3
	SLT	**98.7**	**41.5**	**1.7**
MIRFLICKR-25K	NACR	65.1	45.4	23.9
	SLT	**92.2**	**49.3**	**3.7**

while the TN suppression ratio remains low. This demonstrates that the shared-neighbor filter mainly suppresses false repulsion caused by subtractive IDN, with limited impact on genuine negative pairs.

Soft-attraction analysis. Figure 7 compares hard attraction and soft attraction on MS COCO, showing that soft attraction produces clearer semantic structures. Quantitative results further show that it preserves a monotonic relation between label overlap and Hamming distance, whereas hard attraction does not. This confirms that soft attraction better preserves graded multi-label semantics. Detailed results are provided in Appendix B.5.



(a) Hard attraction (b) Soft attraction

Figure 7: t-SNE (van der Maaten & Hinton, 2008) of 128-dimensional features under 0.6 label noise on MS COCO.

Case study. As shown in Figure 8, all Top-10 results are relevant, making $P@10$ unable to distinguish retrieval quality. IDNRH retrieves more samples with higher R^c_{qj} values, where R^c_{qj} denotes the clean-label Jaccard overlap, and achieves a DCG@10 of 3.70, compared with 1.60 for SDC and 1.08 for NDBE. This indicates that IDNRH better ranks samples with stronger semantic overlap under IDN-corrupted supervision. Detailed experimental settings are provided in Appendix B.9.

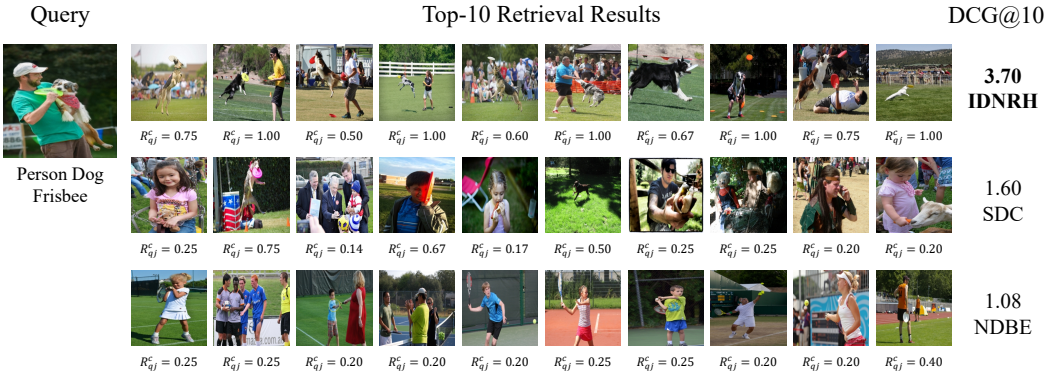


Figure 8: Top-10 retrieval results for an MS COCO query under 60% IDN with 128-bit hash codes. The value below each retrieved image is its clean-label Jaccard overlap R^c_{qj} with the query.

6 Conclusion

This paper investigates multi-label image hashing retrieval under IDN and proposes IDNRH. SLT converts sample partitioning and multi-view neighbor consensus into instance-aware label truncation, while RFSCL employs shared-neighbor filtering and soft attraction to mitigate false repulsion and false attraction, respectively. Extensive experiments demonstrate the robustness of IDNRH on both synthetic and real-world noise datasets. In future work, we further consider extending IDN-robust hashing to cross-modal retrieval and developing an explicit probabilistic model of IDN.

AI Use Statement

Generative AI tools were used for polishing the manuscript, retrieving and discovering relevant literature, generating portions of the experimental code, and generating synthetic datasets. For the other tasks subject to mandatory disclosure under the ICLR 2027 policy, generative AI tools were not used, or those tasks were not applicable to this work. All AI-assisted suggestions were rigorously reviewed and verified by the authors against the original literature, implementation code, and experimental records. The authors take full responsibility for the final text, scientific claims, experimental design, implementations, and reported results.

Reproducibility Statement

We provide detailed experimental settings, implementation details, training procedures, and the IDN injection algorithm in the main text and Appendix to facilitate reproduction of the reported results.

Ethics Statement

Our research adheres to the ICLR Code of Ethics. All datasets used in this study are publicly available research benchmarks. The work neither collects nor releases personal data, and the injected label corruption is used solely to evaluate robustness under controlled conditions. Code and related materials will be released in accordance with the applicable dataset licenses to support transparent and reproducible research.

References

Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S. Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, and Simon Lacoste-Julien. A closer look at memorization in deep networks. In *Proceedings of the International Conference on Machine Learning*, volume 70, pp. 233–242, 2017.

Antonin Berthon, Bo Han, Gang Niu, Tongliang Liu, and Masashi Sugiyama. Confidence scores make instance-dependent label-noise learning possible. In *Proceedings of the International Conference on Machine Learning*, volume 139, pp. 825–836, 2021.

Zhangjie Cao, Mingsheng Long, Jianmin Wang, and Philip S. Yu. Hashnet: Deep learning to hash by continuation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5609–5618, 2017.

Jian Chen, Ruiyi Zhang, Tong Yu, Rohan Sharma, Zhiqiang Xu, Tong Sun, and Changyou Chen. Label-retrieval-augmented diffusion models for learning from noisy labels. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66499–66517, 2023.

Pengfei Chen, Junjie Ye, Guangyong Chen, Jingwei Zhao, and Pheng-Ann Heng. Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11442–11450, 2021.

Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *Proceedings of the International Conference on Learning Representations*, 2021.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 48:1–48:9, 2009.

Aritra Ghosh, Himanshu Kumar, and P. S. Sastry. Robust loss functions under label noise for deep neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, pp. 1919–1925, 2017.

Yiyang Gu, Bohan Wu, Qinghua Ran, Rong-Cheng Tu, Xiao Luo, Zhiping Xiao, Wei Ju, Dacheng Tao, and Ming Zhang. SEGA: Shaping semantic geometry for robust hashing under noisy supervision. In *Advances in Neural Information Processing Systems*, volume 38, pp. 4212–4235, 2025.

Yiyang Gu, Bohan Wu, Yifang Qin, Jiaru Tang, Rong-Cheng Tu, Zhiping Xiao, Taian Guo, Junyu Luo, Wei Ju, Xiao Luo, Dacheng Tao, and Ming Zhang. CogniTrust: Cognitive memory-driven verifiable supervision for robust hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 4339–4347, 2026.

Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor W. Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8536–8546, 2018.

Chao He and Hongxi Wei. Hybridhash: Hybrid convolutional and self-attention deep hashing for image retrieval. In *Proceedings of the International Conference on Multimedia Retrieval*, pp. 824–832, 2024.

Marzi Heidari, Hanping Zhang, and Yuhong Guo. Learning to clean: Reinforcement learning for noisy label correction. In *Advances in Neural Information Processing Systems*, volume 38, pp. 93198–93218, 2025.

Jiun Tian Hoe, Kam Woh Ng, Tianyu Zhang, Chee Seng Chan, Yi-Zhe Song, and Tao Xiang. One loss for all: Deep hashing with a single cosine similarity based learning objective. In *Advances in Neural Information Processing Systems*, volume 34, pp. 24286–24298, 2021.

Senyu Hou, Gaoxia Jiang, Jia Zhang, Shangrong Yang, Husheng Guo, Yaqing Guo, and Wenjian Wang. Directional label diffusion model for learning from noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25738–25748, 2025.

Mark J. Huiskes and Michael S. Lew. The mir flickr retrieval evaluation. In *Proceedings of the ACM International Conference on Multimedia Information Retrieval*, pp. 39–43, 2008.

Young Kyun Jang, Geonmo Gu, Byungsoo Ko, Isaac Kang, and Nam Ik Cho. Deep hash distillation for image retrieval. In *Proceedings of the European Conference on Computer Vision*, volume 13674, pp. 354–371, 2022.

Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *Proceedings of the International Conference on Machine Learning*, volume 80, pp. 2304–2313, 2018.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 18661–18673, 2020.

Youngwook Kim, Sehwan Kim, Youngmin Ro, and Jungwoo Lee. Instance-dependent multilabel noise generation for multilabel remote sensing image classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:17087–17098, 2024.

Junnan Li, Richard Socher, and Steven C. H. Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. In *Proceedings of the International Conference on Learning Representations*, 2020.

Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu. Selective-supervised contrastive learning with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 316–325, 2022.

Yunqiang Li and Jan van Gemert. Deep unsupervised image hashing by maximizing bit entropy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2002–2010, 2021.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of the European Conference on Computer Vision*, pp. 740–755, 2014.

Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20331–20342, 2020.

Kam Woh Ng, Xiatian Zhu, Jiun Tian Hoe, Chee Seng Chan, Tianyu Zhang, Yi-Zhe Song, and Tao Xiang. Unsupervised hashing with similarity distribution calibration. In *Proceedings of the British Machine Vision Conference*, pp. 53–69, 2023.

Liangkang Peng, Jiangbo Qian, Zhengtao Xu, Yu Xin, and Lijun Guo. Multi-label hashing for dependency relations among multiple objectives. *IEEE Transactions on Image Processing*, 32:1759–1773, 2023.

Qibing Qin, Mingkun Dou, Wenfeng Zhang, Lei Huang, and Jie Nie. Deep neighbor discriminant binary embedding for multi-label image retrieval. *IEEE Transactions on Multimedia*, 28:5835–5848, 2026.

Zexuan Qiu, Qinliang Su, Zijing Ou, Jianxing Yu, and Changyou Chen. Unsupervised hashing with contrastive information bottleneck. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 959–965, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Xiaobo Shen, Guohua Dong, Yuhui Zheng, Long Lan, Ivor W. Tsang, and Quan-Sen Sun. Deep co-image-label hashing for multi-label image retrieval. *IEEE Transactions on Multimedia*, 24: 1116–1126, 2022.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 34(11):8135–8153, 2023.

Chao Su, Likang Peng, Yuan Sun, Dezhong Peng, Xi Peng, and Xu Wang. Neighbor-aware contrastive disambiguation for cross-modal hashing with redundant annotations. In *Advances in Neural Information Processing Systems*, volume 38, pp. 108581–108603, 2025.

Jinan Sun, Haixin Wang, Xiao Luo, Shikun Zhang, Wei Xiang, Chong Chen, and Xian-Sheng Hua. HEART: Towards effective hash codes under label noise. In *Proceedings of the ACM International Conference on Multimedia*, pp. 366–375, 2022.

Rong-Cheng Tu, Xian-Ling Mao, Kevin Qinghong Lin, Chengfei Cai, Weize Qin, Wei Wei, Hongfa Wang, and Heyan Huang. Unsupervised hashing with semantic concept mining. *Proceedings of the ACM on Management of Data*, 1(1):3:1–3:19, 2023.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

Haixin Wang, Huiyu Jiang, Jinan Sun, Shikun Zhang, Chong Chen, Xian-Sheng Hua, and Xiao Luo. Dior: Learning to hash with label noise via dual partition and contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*, 36(4):1502–1517, 2024a.

Jinpeng Wang, Ziyun Zeng, Bin Chen, Tao Dai, and Shu-Tao Xia. Contrastive quantization with code memory for unsupervised image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2468–2476, 2022.

Yejiang Wang, Yuhai Zhao, Zhengkui Wang, Wen Shan, and Xingwei Wang. Limited-supervised multi-label learning with dependency noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 15662–15670, 2024b.

Rukai Wei, Yu Liu, Jingkuan Song, Yanzhao Xie, and Ke Zhou. Exploring hierarchical information in hyperbolic space for self-supervised image hashing. *IEEE Transactions on Image Processing*, 33:1768–1781, 2024.

You Wu, Yabo Shi, Yizhang Zou, and Peipei Li. Robust multi-label learning with instance-dependent label noise. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pp. 3458–3467, 2025.

Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7597–7610, 2020.

Li Yuan, Tao Wang, Xiaopeng Zhang, Francis E. H. Tay, Zequn Jie, Wei Liu, and Jiashi Feng. Central similarity quantization for efficient image and video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3083–3092, 2020.

Zhilu Zhang and Mert R. Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems*, volume 31, pp. 8792–8802, 2018.

A ALGORITHM

We summarize the training procedure of the proposed method in Algorithm 1. It alternates mini-batch optimization with epoch-level multi-view neighbor reconstruction and GMM-based sample screening. Algorithm 2 details the multi-label instance-dependent noise injection procedure used to corrupt the training labels.

Algorithm 1 Training procedure of IDNRH

Require: Observed noisy training set $\mathcal{D} = \{(x_i, \bar{y}_i)\}_{i=1}^{N}$; frozen encoder f; hash head h_θ; epochs T; warm-up epochs T_w
Ensure: Trained hash head h_θ

Initialize h_θ, classification head W, and label confidence $P \leftarrow \bar{Y}$
for $t = 1$ to T **do**
  Extract global features and construct fused k-NN sets $\{\mathcal{N}_i\}_{i=1}^{N}$
  if $t > T_w$ **then**
    Fit a two-component GMM to the preceding epoch's losses $\{l_i\}_{i=1}^{N}$; obtain $\{m_i\}_{i=1}^{N}$ and $\{P(\text{clean} \mid l_i)\}_{i=1}^{N}$
    Update P by multi-view neighbor consensus with correction strength $1 - P(\text{clean} \mid l_i)$
    Derive class-specific truncation thresholds $\{t_k\}_{k=1}^{C}$ from nonzero confidences of samples with $m_i = 1$
  end if
  for each mini-batch $\mathcal{B}$ **do**
    Generate weak and strong views; obtain $U^w = h_\theta(f(X^w))$, $U^s = h_\theta(f(X^s))$
    Predict $Z \leftarrow \sigma(U^w W)$
    Compute weak-view per-sample BCE losses $\{l_i\}_{i \in \mathcal{B}}$ from Z and $\bar{Y}_\mathcal{B}$
    if $t \leq T_w$ **then**
      $\hat{Y}_\mathcal{B} \leftarrow \bar{Y}_\mathcal{B}$
    else
      Obtain truncated labels $\hat{Y}_\mathcal{B} \leftarrow \text{SLT}(\bar{Y}_\mathcal{B}, P_\mathcal{B}, \{t_k\}_{k=1}^{C}, m_\mathcal{B})$
    end if
    Compute soft label similarity $R_\mathcal{B}$ and shared-neighbor similarity $C_\mathcal{B}$
    Compute $\mathcal{L}_c$, $\mathcal{L}_s = \mathcal{L}_{rep}(R_\mathcal{B}, C_\mathcal{B}) + \mathcal{L}_{att}(R_\mathcal{B})$, and $\mathcal{L}_q$
    Update θ and W using $\mathcal{L}_{total} = \mathcal{L}_c + \alpha\mathcal{L}_s + \beta\mathcal{L}_q$
  end for
end for

Algorithm 2 Multi-label IDN injection

Require: Clean training set $\mathcal{D}_{\text{clean}} = \{(x_i, y_i)\}_{i=1}^{N}$; class prompts $\{d_c\}_{c=1}^{C}$; percentile rate $r \in [0, 100]$
Ensure: Observed noisy training set $\mathcal{D} = \{(x_i, \bar{y}_i)\}_{i=1}^{N}$

Encode class prompts and images using CLIP; obtain visual-text logits g_{ic} for every (i, c)
$q^- \leftarrow \text{Percentile}(\{g_{ic} \mid y_{ic} = 1\}, r)$; $q^+ \leftarrow \text{Percentile}(\{g_{ic} \mid y_{ic} = 1\}, 100 - r)$
for each sample i and class c **do**
  $\bar{y}_{ic} \leftarrow y_{ic}$
  if $y_{ic} = 0$ and $g_{ic} > q^+$ **then**
    $\bar{y}_{ic} \leftarrow 1$ ▷ additive IDN
  else if $y_{ic} = 1$ and $g_{ic} < q^-$ **then**
    $\bar{y}_{ic} \leftarrow 0$ ▷ subtractive IDN
  end if
end for
for each sample i with $\|\bar{y}_i\|_1 = 0$ **do**
  $c^* \leftarrow \arg\max_{c: y_{ic}=1} g_{ic}$; set $\bar{y}_{ic^*} \leftarrow 1$
end for
Return $\mathcal{D} = \{(x_i, \bar{y}_i)\}_{i=1}^{N}$

B Additional Experiments

B.1 Backbone-Controlled Comparison

The main experiments in Table 1 compare methods using their original implementations, whose visual backbones are not fully aligned, whereas IDNRH uses pretrained CLIP ViT-B/32. We therefore conduct a backbone-controlled comparison to verify that the gains arise from IDN-robust learning rather than from a stronger CLIP representation.

Experimental setting. We select the representative Bi-half, SDC, HHCH, DIOR, SEGA, and NDBE baselines. All selected methods use the same pretrained CLIP ViT-B/32 initialization while retaining their original training strategies. We use the same data splits, IDN labels, and evaluation protocol as in the main experiments at 128 bits and noise rates of 20%, 40%, and 60%. HybridHash is excluded because replacing its convolution–self-attention hybrid backbone would alter a core component of the method. Since Bi-half, SDC, and HHCH are unsupervised and do not use training labels, their results are identical across the three noise rates. Table 4 reports the results.

Table 4: mAP (%) with a unified pretrained CLIP ViT-B/32 backbone at 128 bits. Best and second-best results are in bold and underlined.

Method	Reference	Dataset	20%	40%	60%
Bi-half	AAAI'21	NUS-WIDE	66.6	66.6	<u>66.6</u>
		MS COCO	54.5	54.5	54.5
		MIRFLICKR-25K	74.1	74.1	<u>74.1</u>
SDC	BMVC'23	NUS-WIDE	62.7	62.7	62.7
		MS COCO	58.1	58.1	58.1
		MIRFLICKR-25K	72.3	72.3	72.3
HHCH	TIP'24	NUS-WIDE	62.2	62.2	62.2
		MS COCO	59.0	59.0	59.0
		MIRFLICKR-25K	73.9	73.9	73.9
DIOR	TKDE'24	NUS-WIDE	73.5	71.2	65.8
		MS COCO	68.9	63.9	51.7
		MIRFLICKR-25K	75.8	72.4	71.4
SEGA	NeurIPS'25	NUS-WIDE	<u>76.2</u>	<u>72.3</u>	65.7
		MS COCO	<u>69.9</u>	<u>64.8</u>	<u>60.5</u>
		MIRFLICKR-25K	<u>83.8</u>	<u>80.0</u>	73.0
NDBE	TMM'26	NUS-WIDE	69.5	65.6	56.9
		MS COCO	62.4	56.2	52.3
		MIRFLICKR-25K	77.5	73.4	67.6
IDNRH	**Ours**	NUS-WIDE	**77.9**	**75.4**	**69.8**
		MS COCO	**71.3**	**67.8**	**62.8**
		MIRFLICKR-25K	**87.1**	**82.6**	**75.1**

Analysis. IDNRH achieves the best mAP in all settings under the unified CLIP ViT-B/32 backbone, showing that its gains cannot be attributed simply to CLIP. Compared with their original settings, SEGA, Bi-half, and HHCH improve overall, whereas DIOR, NDBE, and SDC decline overall. Thus, CLIP does not uniformly benefit every hashing method; final performance also depends on its interaction with the hashing learning strategy, while IDNRH retains its advantage under a shared strong CLIP representation.

B.2 Supplementary Ablation Study on MIRFLICKR-25K

Table 5 complements the main ablation study with MIRFLICKR-25K results under the same 128-bit setting. At 40% and 60% noise, the full model achieves the best performance. Notably, at 20% noise, hard attraction instead outperforms the full model by 0.3 mAP points. With limited corruption, most observed positives remain reliable, so the risk of false attraction introduced by hard attraction is relatively low, while its uniform strong attraction can more rapidly reinforce existing semantic associations. As the noise rate increases, however, this advantage disappears; at 60% noise, hard

attraction instead causes the largest performance drop of 4.5 points. This trend is consistent with our design motivation: soft attraction may provide limited benefit under low noise, but becomes increasingly important for suppressing false attraction as erroneous positive relations increase. These results further corroborate the complementary roles of label truncation, shared-neighbor filtering, and soft attraction across different noise levels.

Table 5: Supplementary ablation mAP (%) results with 128-bit codes on MIRFLICKR-25K. Best and second-best are bold and underlined.

Method	20%	40%	60%
w/o SLT	86.9	80.5	71.7
w/o GMM	82.8	79.2	73.4
w/o SN Filter	84.6	<u>81.4</u>	<u>73.5</u>
Hard Attraction	**87.4**	80.2	70.6
IDNRH	<u>87.1</u>	**82.6**	**75.1**

B.3 Label-Truncation Quality Experimental Setting

Experimental setting. For Table 3, we evaluate the truncation outcome at the best-mAP epoch of each run. Clean labels are used only as an offline reference and never participate in training. Let $\mathcal{M} = \{(i,c) : \bar{y}_{ic} = 1, \hat{y}_{ic} = 0\}$ be the set of removed observed positives, $\mathcal{A} = \{(i,c) : \bar{y}_{ic} = 1, y_{ic} = 0\}$ the injected additive-noise positions, and $\mathcal{O} = \{(i,c) : \bar{y}_{ic} = 1, y_{ic} = 1\}$ the observed true-positive positions. We compute $P = \frac{|\mathcal{M} \cap \mathcal{A}|}{|\mathcal{M}|}$, $R = \frac{|\mathcal{M} \cap \mathcal{A}|}{|\mathcal{A}|}$, and $FTR = \frac{|\mathcal{M} \cap \mathcal{O}|}{|\mathcal{O}|}$. Here, P quantifies the purity of removed labels and R the fraction of additive IDN removed; higher values of both indicate more effective truncation. FTR measures the fraction of observed true labels removed in error, for which lower values are preferable.

B.4 Repulsion Suppression Experimental Setting

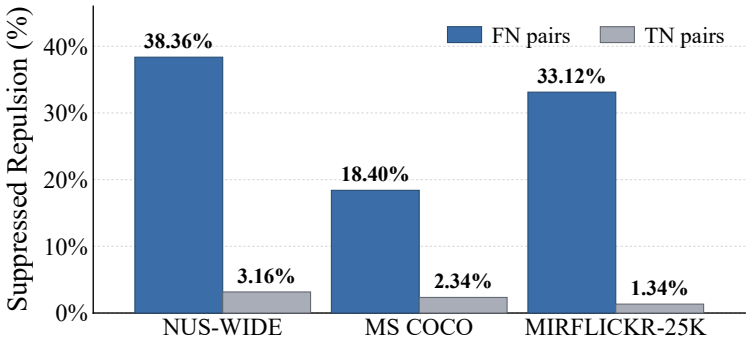


Figure 9: Repulsion suppression ratios of false-negative (FN) and true-negative (TN) pairs.

Experimental setting. For each dataset, we randomly sample equal numbers of false-negative (FN) and true-negative (TN) pairs from training mini-batches. FN pairs have no overlap under the observed labels $\bar{y}$ but positive similarity under the clean labels y, corresponding to false repulsion caused by subtractive IDN, whereas TN pairs have no overlap under either $\bar{y}$ or y. Clean labels are used only for this offline grouping and never participate in training. Following the main text, a pair is considered to have suppressed repulsion when $C_{ij} > 0$. We report the suppression ratio as the proportion of such pairs within the FN and TN groups, respectively.

B.5 Soft-Attraction Quantitative Calibration Analysis

To quantitatively examine whether soft attraction better preserves graded multi-label similarity, we analyze the relation between label overlap and Hamming distance under soft and hard attraction.

Experimental setting. Under 60% noise with 128-bit codes and seed 2025, we evaluate the best-mAP model of each attraction mode. We randomly sample one million training pairs, retain pairs

with positive truncated-label Jaccard similarity $R_{ij} > 0$, divide them into four similarity intervals, and report the mean normalized Hamming distance in each interval.

Analysis. As shown in Table 6, soft attraction exhibits a clearer distance gradient on all three datasets: the mean Hamming distance consistently decreases as truncated-label overlap increases, whereas hard attraction shows a distance increase in the highest-similarity interval on MS COCO and MIRFLICKR-25K. Moreover, from the lowest to the highest similarity interval, the Hamming distance under soft attraction decreases by 25.14, 22.07, and 19.69 percentage points, respectively, substantially larger than the corresponding decreases of 16.12, 13.77, and 7.01 points under hard attraction. These results indicate that soft attraction better maps continuous multi-label semantic similarity into the Hamming space, while hard attraction tends to treat positive relations with different strengths more uniformly. Meanwhile, in the lowest-similarity interval, soft attraction preserves larger Hamming distances than hard attraction, preventing weak-overlap pairs from being pulled excessively close and thereby alleviating false attraction caused by erroneous positive relations introduced by IDN.

Table 6: Mean Hamming distance (%) of positive pairs in truncated-label similarity bins.

Dataset	**Mode**	$(0, .25]$	$(.25, .5]$	$(.5, .75]$	$(.75, 1]$
NUS-WIDE	Hard	32.28	25.28	20.67	16.16
	Soft	39.11	28.45	24.24	13.97
MS COCO	Hard	29.56	19.94	13.81	15.79
	Soft	38.86	26.19	17.01	16.79
MIRFLICKR-25K	Hard	32.63	28.53	25.16	25.62
	Soft	44.48	34.52	26.28	24.79

B.6 Random-Seed Stability Analysis

To assess sensitivity to random initialization, we repeat the 128-bit experiment at the 0.6 noise rate with five seeds (2023–2027), keeping all other settings fixed. Table 7 reports the resulting mAP statistics.

Table 7: Five-seed stability results (mAP, %).

Dataset	**mAP (mean $\pm$ std)**
NUS-WIDE	69.76 ± 0.18
MS COCO	62.18 ± 0.39
MIRFLICKR-25K	74.84 ± 0.21

Analysis. All standard deviations are below 0.4 percentage points, and the best–worst spreads are only 0.5, 1.0, and 0.5 points on NUS-WIDE, MS COCO, and MIRFLICKR-25K. The main results are therefore stable across random initializations.

B.7 Hyperparameter Sensitivity Analysis

To examine the sensitivity of IDNRH to hyperparameters, we investigate the contrastive weight α, quantization weight β, and shared-neighbor filter strength γ under 0.6 noise rate with 128-bit codes.

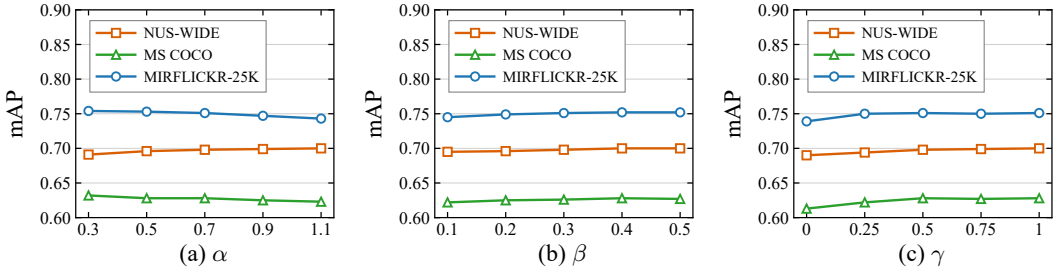


Figure 10: mAP sensitivity to α, β, and γ under the 0.6 noise rate with 128-bit codes.

Analysis. As shown in Figure 10, the mAP fluctuation for every dataset remains within 1.6 percentage points throughout each sweep. Both $\alpha \in [0.5, 0.9]$ and $\beta \in [0.1, 0.5]$ yield stable perfor-

mance. Increasing γ from 0 to 0.5 consistently improves retrieval, validating the shared-neighbor filter, whereas larger values bring only marginal changes. We therefore set $\alpha = 0.7$, $\beta = 0.3$, and $\gamma = 0.5$ as the default values.

B.8 Time Efficiency

To verify the retrieval-efficiency advantage of hash representations, we compare binary hash retrieval with same-dimensional dense continuous retrieval in terms of query time.

Experimental setting. Experiments are conducted on a single core of an Intel Xeon Gold 6530 CPU using gallery sizes matching those of NUS-WIDE and MS COCO, containing 193,734 and 117,218 samples, respectively. We evaluate 16-, 32-, 64-, and 128-bit representations. For each setting, 1,000 independent Top-100 queries are performed after 20 warm-up queries. Binary retrieval uses packed `uint64` codes with XOR and `POPCNT` for Hamming-distance computation, whereas dense retrieval uses same-dimensional normalized `float32` vectors with inner-product similarity. Image decoding, feature extraction, and code generation are excluded from the timing.

Table 8: Average native Top-100 retrieval time per query (ms; 10^3 trials) across code lengths.

	NUS-WIDE (193,734 images)				**MS COCO** (117,218 images)			
	16 Bit	32 Bit	64 Bit	128 Bit	16 Bit	32 Bit	64 Bit	128 Bit
Binary Hash	0.707	0.710	0.538	0.624	0.337	0.341	0.339	0.383
Dense Vector	2.163	4.100	7.658	19.580	1.032	2.046	4.656	11.701
Speed-Up	3.058×	5.778×	14.229×	31.401×	3.064×	6.005×	13.749×	30.544×

Analysis. Binary hash retrieval is consistently faster than same-dimensional dense retrieval across all code lengths, and the speed-up becomes more pronounced as the representation dimension increases. This advantage arises from the compact binary representation and efficient XOR–`POPCNT` operations, whereas the computational cost of dense-vector comparison grows with dimensionality. These results demonstrate the computational efficiency of hash representations for large-scale retrieval.

B.9 Case Study Experimental Setting

Experimental setting. Under the 60% IDN setting on MS COCO with 128-bit hash codes, we compare IDNRH with SDC, a strong unsupervised method, and NDBE, a strong supervised method. We randomly select a query from those for which all three methods achieve $P@10 = 1$, thereby eliminating differences in binary relevance. For a query q and its j-th retrieved image, we define the clean-label Jaccard overlap as $R^c_{qj} = \frac{|y_q \cap y_j|}{|y_q \cup y_j|}$ to measure graded semantic relevance. We further adopt Discounted Cumulative Gain (DCG) (Järvelin & Kekäläinen, 2002):

$$\mathrm{DCG@10} = \sum_{j=1}^{10} \frac{R^c_{qj}}{\log_2(j+1)}, \tag{15}$$

which jointly considers semantic overlap and retrieval rank.

C Real-World Validation on NUS-WIDE-Web-21

C.1 NUS-WIDE-Web-21: Construction and Noise Statistics

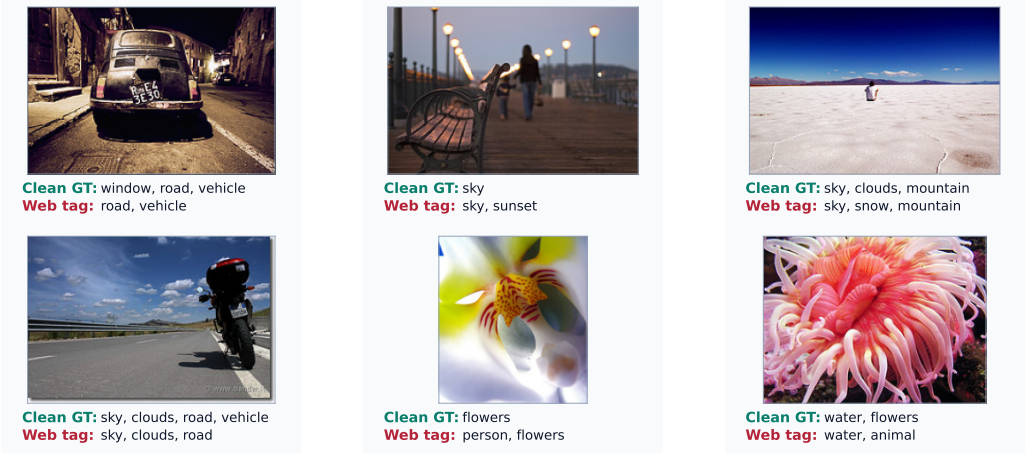
Data source. NUS-WIDE-Web-21 is a real-world multi-label retrieval benchmark constructed from NUS-WIDE (Chua et al., 2009) without introducing synthetic label corruption. For each image, we retain two aligned 21-dimensional label views: the clean labels are obtained from the official NUS-WIDE ground-truth concept labels, while the observed labels are taken from the official Web tags associated with the same image (`AllTags81`). We first dynamically select the 21 most frequent concepts according to their frequencies in the clean GT and extract the corresponding dimensions from both official 81-dimensional label sets. Therefore, the discrepancies between the clean and Web labels arise from real Web annotations rather than artificially injected label flips.

Dataset construction. We first retain readable images that contain at least one positive clean label among the selected 21 concepts, i.e., at least one entry in their 21-dimensional clean GT is positive. With the random seed fixed to 2025, we randomly select 100 previously unselected clean-positive images for each concept as queries, resulting in 2,100 query images, while the remaining candidate images form the retrieval database. The training set is further sampled from the database. We first exclude images whose Web labels are all zero, ensure that each concept contains at least 500 clean-positive training samples, and then randomly fill the training set to 10,500 unique images from the remaining candidates. For the same training images, we separately store their clean GT labels and the unmodified official Web labels. The query and database sets retain the clean GT labels for retrieval evaluation.

Noise statistics. The resulting training set contains both additive and subtractive real-world label errors. An additive error occurs when a concept is negative in the clean GT but is incorrectly marked as positive in the Web labels, whereas a subtractive error occurs when a concept is positive in the clean GT but is omitted from the Web labels. By comparing the clean and Web labels of the 10,500 training images, we identify 13,564 additive errors and 7,584 subtractive errors, accounting for 64.14% and 35.86% of the 21,148 total label discrepancies, respectively. On average, each image contains 1.292 additive errors and 0.722 subtractive errors. In addition, the average numbers of clean and Web labels per image are 2.0452 and 2.6148, respectively, indicating that the observed Web annotations have a higher label cardinality overall.

C.2 Label Noise in NUS-WIDE-Web-21 Exhibits Instance Dependence

Visual evidence. Figure 11 provides visual evidence that the observed Web-label errors depend on image-specific visual cues. In Figure 11(a), *window* is omitted when window-related details are dark and peripheral in a night street scene, while *vehicle* is omitted for a motorcycle shown in an unusual rear-facing view and located at the image boundary. Thus, the missing-label errors vary with relative salience, location, and viewpoint, rather than being caused solely by the corresponding clean category. In Figure 11(b), a row of warm, glowing bulbs in a dusky scene is confused with *sunset*, and the face-mask-like morphology of an orchid is mistaken for *person*. Figure 11(c) exhibits both directions simultaneously: tiny clouds near the horizon are missed whereas the bright arid land is mistaken for *snow*; the tentacle-like petals of the second flower lead to an omitted *flowers* tag and an added *animal* tag. These examples tie both false positives and false negatives to local appearance, scale, position, illumination, viewpoint, and shape. Taken together, this visual evidence confirms that NUS-WIDE-Web-21 indeed contains instance-dependent label noise.



(a) Subtractive IDN (b) Additive IDN (c) Mixed IDN

Figure 11: Qualitative examples of naturally occurring label discrepancies in NUS-WIDE-Web-21. Each subfigure contains two images from one noise type. Green *Clean GT* labels are official ground-truth labels and red *Web tag* labels are the observed human-provided Web labels.

Quantitative evidence. We further examine whether this instance dependence holds across the entire training set rather than only in a few visual examples. Specifically, we use a frozen CLIP ViT-B/32 to encode each image and the class prompt "a photo of a {class}", and compute the cosine similarity

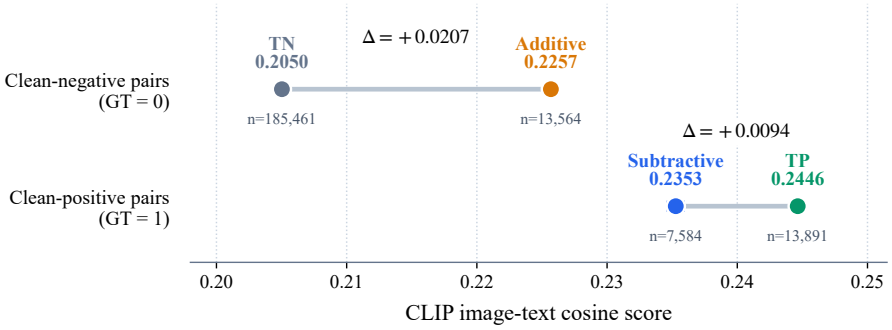


Figure 12: Comparison of CLIP scores across the four clean/Web label combinations in the NUS-WIDE-Web-21 training set. Points indicate mean CLIP scores, horizontal error bars indicate 95% confidence intervals, and n denotes the number of image–class pairs.

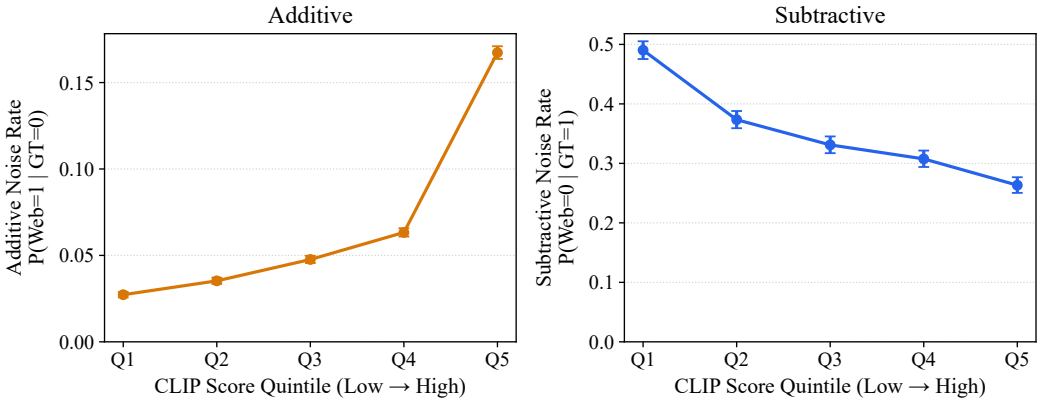


Figure 13: Real Web-label error rates across CLIP-score intervals in NUS-WIDE-Web-21. The left panel shows the additive error rate for pairs with clean $= 0$, while the right panel shows the subtractive error rate for pairs with clean $= 1$. Q1–Q5 denote five intervals from low to high CLIP scores, and error bars indicate 95% Wilson confidence intervals.

between their ℓ_2-normalized embeddings, denoted as the CLIP score. A higher CLIP score indicates that the corresponding class is more visually evident in the image. We divide all 220,500 image–class pairs into four groups according to their clean GT and Web labels: TP (clean $= 1$, Web $= 1$), TN (clean $= 0$, Web $= 0$), Additive (clean $= 0$, Web $= 1$), and Subtractive (clean $= 1$, Web $= 0$). As shown in Figure 12, among image–class pairs with clean $= 0$, the mean CLIP score of Additive pairs is 0.2257, higher than 0.2050 for TN pairs. Among pairs with clean $= 1$, the mean score of Subtractive pairs is 0.2353, lower than 0.2446 for TP pairs. This indicates that an actually absent class is more likely to be incorrectly added when it appears more visually plausible in the image, whereas an actually present class is more likely to be omitted when its visual evidence is weaker.

To further examine whether the label error rate changes systematically with the CLIP score, we separately analyze additive and subtractive errors. For additive errors, we select all image–class pairs with clean $= 0$, sort them in ascending order of CLIP score, and divide them into five equal-sized intervals Q1–Q5. We then compute the proportion of pairs whose Web labels are incorrectly marked as 1 in each interval. For subtractive errors, we select all pairs with clean $= 1$, divide them into Q1–Q5 in the same manner, and compute the proportion whose Web labels are incorrectly marked as 0.

As shown in Figure 13, as the CLIP score increases, the additive error rate rises monotonically from 2.73% in Q1 to 16.73% in Q5. In contrast, the subtractive error rate decreases monotonically from 49.03% to 26.33%. These opposite monotonic trends show that the probability of Web-label errors varies with the visual information of individual image–class pairs, providing dataset-level evidence that the label noise exhibits instance dependence.

D Implementation Details

D.1 Dataset Details

We evaluate IDNRH on three widely used multi-label image-retrieval benchmarks, NUS-WIDE, MS COCO, and MIRFLICKR-25K, and the real-world noisy benchmark NUS-WIDE-Web-21. Their statistics and data partitions are summarized in Table 9.

- **NUS-WIDE** (Chua et al., 2009) contains 269,648 Web images annotated with 81 concepts. Following the standard hashing protocol, we retain 195,834 images associated with the 21 most frequent concepts. We randomly select 2,100 images as queries, use the remaining 193,734 images as the retrieval database, and sample 10,500 database images for training.
- **MS COCO** (Lin et al., 2014) consists of 82,783 training images and 40,504 validation images annotated with 80 object categories. Following the standard protocol, 122,218 images are retained for the experiments. We randomly select 5,000 images as queries, use the remaining 117,218 images as the retrieval database, and sample 10,000 database images for training.
- **MIRFLICKR-25K** (Huiskes & Lew, 2008) consists of 25,000 Flickr images with 24 semantic labels. Following the standard protocol, 20,015 images are retained. We randomly select 2,000 images as queries, use the remaining 18,015 images as the retrieval database, and sample 10,000 database images for training.
- **NUS-WIDE-Web-21** (Chua et al., 2009) is a real-world noisy multi-label benchmark derived from NUS-WIDE. It uses the 21 most frequent concepts and retains 195,834 readable images with at least one clean concept. We randomly select 2,100 images as queries, use the remaining 193,734 images as the retrieval database, and select 10,500 database images with non-empty Web labels for training.

Table 9: Statistics and data partitions of the four benchmarks. The training set is sampled from the retrieval database.

Dataset	Classes	Used images	Training	Query	Database
NUS-WIDE	21	195,834	10,500	2,100	193,734
MS COCO	80	122,218	10,000	5,000	117,218
MIRFLICKR-25K	24	20,015	10,000	2,000	18,015
NUS-WIDE-Web-21	21	195,834	10,500	2,100	193,734

D.2 Baseline Details

We compare eleven advanced hashing baselines, separated into unsupervised and supervised groups.
Unsupervised methods.

- **Bi-half** (Li & van Gemert, 2021) maximizes bit entropy for unsupervised image hashing.
- **CIBHash** (Qiu et al., 2021) combines contrastive learning with an information-bottleneck objective.
- **MeCoQ** (Wang et al., 2022) performs contrastive quantization with a code memory.
- **SDC** (Ng et al., 2023) calibrates the similarity distribution to obtain unsupervised hash codes.
- **HHCH** (Wei et al., 2024) explores hierarchical information in hyperbolic space for self-supervised image hashing.

Supervised methods.

- **OrthoHash** (Hoe et al., 2021) learns hash codes with a single cosine-similarity objective and orthogonal class prototypes.
- **DHD** (Jang et al., 2022) distills retrieval knowledge into compact binary representations.
- **DIOR** (Wang et al., 2024a) partitions samples and network parameters to mitigate label noise through contrastive semi-supervised learning.
- **HybridHash** (He & Wei, 2024) combines convolutional and self-attention representations for image retrieval.

- **SEGA** (Gu et al., 2025) shapes semantic geometry with dynamic prototypes and soft uncertainty calibration under noisy supervision.
- **NDBE** (Qin et al., 2026) preserves neighborhood structure through multi-center discriminant embedding and class-center aggregation.

For each baseline, we use the authors' official implementation when available and otherwise follow the released hyperparameters and evaluation protocols. All methods use the same train/query/database partitions and hash lengths for a fair comparison.

D.3 Training Details

Training details. We use CLIP ViT-B/32 as a frozen image feature extractor and train the hash projection head on top of the extracted features. Our method is optimized using Adam with a learning rate of 1×10^{-5} and a weight decay of 1×10^{-5} for 50 epochs. The batch size is fixed to 128. Unless otherwise specified, the hyperparameters are fixed as $K = 20$, $\xi = 1.0$, $\alpha = 0.7$, $\beta = 0.3$, and $\gamma = 0.5$. The implementation uses PyTorch, and all experiments are conducted on a single NVIDIA RTX 4090 GPU.