CogniTrust: Cognitive Memory-Driven Verifiable Supervision for Robust Hashing

Yiyang Gu[1], Bohan Wu[1], Yifang Qin[1], Jiaru Tang[2], Rong-Cheng Tu[3], Zhiping Xiao[4,*], Taian Guo[1], Junyu Luo[1], Wei Ju[1], Xiao Luo[5,*], Dacheng Tao[3], Ming Zhang[1,*]

[1]State Key Laboratory for Multimedia Information Processing,
School of Computer Science, PKU-Anker LLM Lab, Peking University
[2]School of Psychological and Cognitive Sciences, Peking University
[3]Nanyang Technological University
[4]University of Washington
[5]University of Wisconsin–Madison
{yiyanggu, wxtpku, qinyifang, juwei, mzhang_cs}@pku.edu.cn,
{tang_psy, taianguo, luojunyu}@stu.pku.edu.cn
{rongcheng.tu,dacheng.tao}@ntu.edu.sg, patxiao@uw.edu, xiao.luo@wisc.edu

Abstract

In this paper, we study the problem of robust multi-label hashing, where label noise hinders the learning of a reliable semantic structure from data. Many existing methods rely on heuristic sample selection or consistency-based training, but lack a unified mechanism to validate and refine supervision across structural and semantic levels. Inspired by cognitive theories of human memory, we propose a novel framework called CogniTrust that unifies verifiable supervision with a triadic memory model: a) In episodic memory, feature activations are decomposed into spatial patterns that support the assessment of structural evidence and the estimation of label reliability; b) Semantic memory keeps track of class-level prototypes from structurally attentive regions to estimate the semantic plausibility of labels; c) Reconstructive memory simulates memory recall through interpolation between images using a diffusion-based mixup process, which enriches the training signals for semantically uncertain regions. These components work together, allowing supervision to be refined through the joint consideration of spatial structure and semantic information. Extensive experiments on noisy hashing benchmarks demonstrate that CogniTrust consistently outperforms a range of state-of-the-art baselines. Our results show that cognitive memory mechanisms offer a principled basis for more reliable label denoising and robust hashing.

Introduction

Learning to hash with noisy supervision has become increasingly crucial for large-scale retrieval tasks in real-world multimedia applications (Wang et al. 2017). In multi-label tasks like semantic tagging, scene understanding, and instance-level retrieval, labels are often missing or noisy. This is usually caused by weak annotations, incomplete metadata, or overlapping meanings between categories (Hu et al. 2021; Zhang et al. 2021). Such problems reduce the quality of supervision. As a result, deep hashing models struggle to learn binary codes that reflect true semantic similarities.

Despite the impressive progress in supervised hashing (Li, Wang, and Kang 2016; Cao et al. 2017a; Yuan et al. 2020), most existing approaches implicitly assume access to fully trustworthy labels during training. However, when supervision is noisy, directly fitting the observed labels can lead to memorization of spurious patterns (Arpit et al. 2017; Zhang et al. 2021), representation drift, and fragmented decision boundaries (Somepalli et al. 2022), ultimately degrading generalization. The discrete form of hash codes makes the system even more fragile. Because each bit operates as a hard decision, and small supervision errors can flip code assignments, which creates a vulnerability that is difficult to mitigate in real applications.

Previous work has explored several ways to reduce the impact of noisy labels, including sample reweighting (Chang, Learned-Miller, and McCallum 2017), loss correction (Patrini et al. 2017), consistency training (Iscen et al. 2022), and co-training (Han et al. 2018). These methods are helpful in some cases, but mainly rely on heuristic rules or confidence estimation. More recent studies are in a different direction. DIOR (Wang et al. 2023) separates clean and noisy samples before employing more reliable supervision, and STAR (Long et al. 2025) captures the structure of semantic similarity through contrastive learning. However, they still lack the cognitive mechanisms that humans use when dealing with uncertainty.

To address this challenge, we propose a new framework, called *CogniTrust*, inspired by the way human memory systems handle uncertainty. Unlike earlier models that focus on pairwise similarity and label correction, CogniTrust uses a triadic memory mechanism to interpret, calibrate, and reconstruct supervision signals. This design generalizes concepts in cognitive science (Tulving et al. 1972; Tulving 1983; McClelland, McNaughton, and O'Reilly 1995; Schacter, Norman, and Koutstaal 2000; Hemmer and Steyvers 2009), where different types of memories serve different cognitive functions: a) *Episodic memory* captures visual details associated with particular contexts. It evaluates the class relevance and spatial specificity of different structural factors to

*Corresponding authors.

identify reliable supervision. b) *Semantic memory* stores abstract concepts. It learns class prototypes and class-specific features dynamically to guide the calibration of noisy labels. c) *Reconstructive memory* can synthesize plausible experiences by integrating past episodes. We utilize a diffusion model to structurally blend information from observed instances, thereby filling in semantically uncertain areas. These three memories leverage information from both structural and semantic aspects, and provides a cognitive perspective that connects structural patterns with semantic concepts. They together form a verifiable supervision system. It first verifies the reliability of supervision using factorized structural attention, then calibrates the noisy labels through structural factor-guided prototype reasoning, and finally enriches the supervision signals of semantically uncertain regions through diffusion-based mixup. As a consequence, CogniTrust generates strong hash codes against label noise, which reflect the semantic structure of data. To summarize, the main contributions of this paper are as follows:

- **Cognitive Perspective:** We introduce a cognitive view of learning hash codes from noisy labels. We utilize episodic, semantic, and reconstructive functions to refine supervision signals rather than conventional denoise techniques.
- **Memory-Driven Framework:** We develop a new memory system CogniTrust, where episodic and semantic memory verify the spatial and semantic reliability of supervision respectively, while reconstructive memory supplements the supervision signals for uncertain regions.
- **Comprehensive Experiments:** We conduct extensive experiments to demonstrate that CogniTrust can provide superior performance over competitive baseline methods on widely-used noisy hashing benchmarks.

Related Work

Deep Hashing Learning. Deep hashing has risen as a powerful technique for efficient retrieval. It can embed data into binary codes that reflect the semantic structure behind the data. Most supervised methods rely on label-based similarity to build pairwise, triplet, or category-level objectives (Li, Wang, and Kang 2016; Lai et al. 2015; Cao et al. 2017a). To handle the challenge of discrete codes, GreedyHash (Su et al. 2018) utilizes a greedy principle for optimization. HashNet (Cao et al. 2017b) leverages a continuation method with convergence guarantees to learn hash codes. Unsupervised methods exploit intrinsic semantic structures via contrastive learning (Qiu et al. 2021), entropy maximization (Li and van Gemert 2021), and variational modeling (Dai et al. 2017). Though effective under clean labels, they are vulnerable to noisy supervision. Recent work, such as DIOR (Wang et al. 2023) and STAR (Long et al. 2025), employs partitioning-based or contrast-based regularization to reduce the impact of label noise. However, they are limited to single-view denoising and overlook inconsistencies between structural and semantic cues. We propose a memory-driven framework to refine supervision via structurally semantic alignment and generative enhancement.

Learning with Label Noise. Learning under label noise has inspired a variety of robust training paradigms, ranging from noise-transition modeling (Patrini et al. 2017), to sample-level uncertainty estimation (Zhang et al. 2023; Gu et al. 2025), and implicit or explicit regularization (Zhang et al. 2018; Zhou et al. 2021; Liu et al. 2020). Co-teaching (Han et al. 2018) leverages small-loss samples to teach peer networks, and DivideMix (Li, Socher, and Hoi 2021) further incorporates Gaussian mixture modeling and semi-supervised bootstrapping. MixUp (Zhang et al. 2018) and related interpolation-based techniques improve generalization by improving decision boundaries and mitigating overfitting to noisy labels. Despite the progress in classification tasks, robust learning for structured prediction problems such as hashing remains underexplored, especially under noisy multi-label supervision. Our work fills this gap by designing a closed-loop framework that couples label verification, prototype-driven uncertainty estimation, and diffusion-based mixup, thereby promoting consistency and robustness in the learned binary embedding space.

Methodology

Problem Definition

We consider the problem of learning robust hash codes under label noise. Formally, given a dataset $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{N}$ with N examples, each input $\boldsymbol{x}_i \in \mathbb{R}^{H\times W\times 3}$ is an image, and $\boldsymbol{y}_i \in \{0,1\}^C$ is a multi-hot label vector over C semantic classes. However, the labels in $\mathcal{D}$ may be corrupted, incomplete, or ambiguous due to imperfect annotation. The goal is to learn a hash encoder $\Phi : \boldsymbol{x} \mapsto \boldsymbol{b} \in \{-1,+1\}^L$ that maps each input into an L-bit binary code, such that semantically similar inputs yield similar codes in the Hamming space, despite the presence of unreliable labels.

Framework Overview

CogniTrust introduces a cognitively inspired framework for robust multi-label hashing under noisy supervision, which is grounded in a triadic memory system. It builds on a neural network that maps each input $\boldsymbol{x}$ to a continuous embedding $\boldsymbol{h} \in \mathbb{R}^L$ via a learnable function $g(\boldsymbol{x})$, then binarizes it as $\boldsymbol{b} = \text{sign}(\boldsymbol{h})$ for compact hash codes. Episodic memory identifies structurally grounded label signals by factorizing activations into spatially disentangled patterns, producing interpretable trust scores. Semantic memory keeps class-level prototypes derived from structurally attentive regions. These prototypes help estimate semantic consistency by checking how well features align with them. Reconstructive memory then creates new samples through a diffusion-based mixup process. It enriches the supervision signals for uncertain regions and improves the decision boundaries. These modules together form a closed loop for supervision refinement. Through this loop, CogniTrust learns hash codes that capture the semantic structure of data. An overview of the framework CogniTrust is shown in Figure 1.

Episodic Memory Modeling via Factorized Structural Attention

In order to instantiate the cognitive mechanism of episodic memory, which retrieves perceptually grounded experiences and allows verification through spatial details, we propose

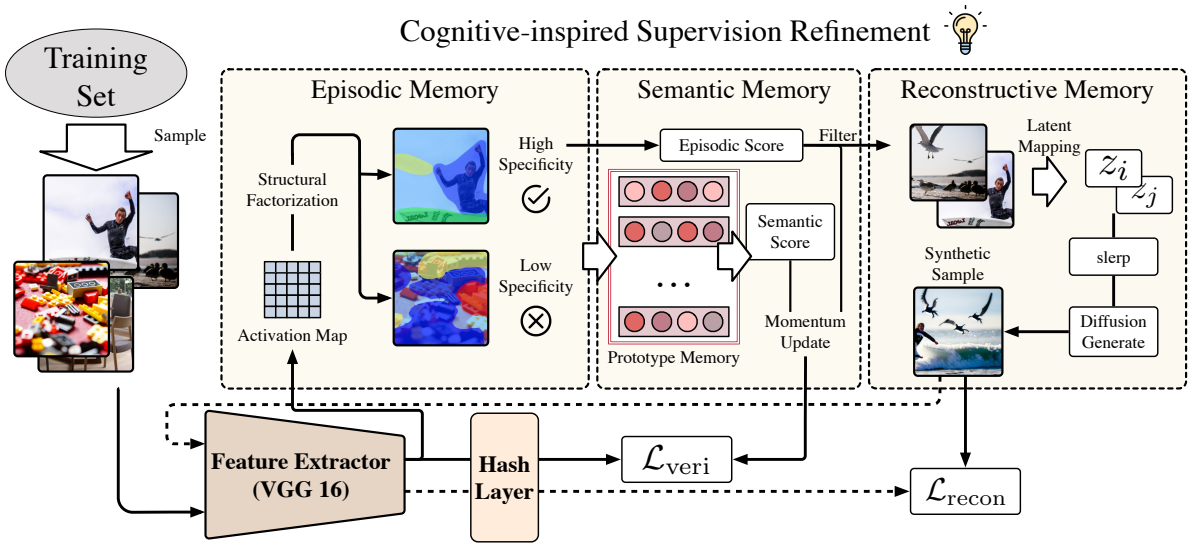


Figure 1: Overview of CogniTrust. The framework brings episodic memory in through factorized structural attention, semantic memory through prototype reasoning, and reconstructive memory via diffusion-based mixup. All of them together form a cognitively inspired closed-loop system to make multi-label hashing robust to noisy supervision.

an episodic verification module that estimates the spatial verifiability of predicted labels relative to the activation patterns of the input. Specifically, we apply a *structural factorization* to the activation tensor of the final convolutional layer in order to expose semantically coherent and spatially disentangled visual components, which serve as evidence to support the relevant labels.

Formally, given the activation tensor $\boldsymbol{G}_i \in \mathbb{R}^{D\times H\times W}$ from the final convolutional layer for image i, we treat $\boldsymbol{G}_i$ as a structured encoding of mid-level perceptual patterns. We reshape it into a 2D matrix $\boldsymbol{G}_i^{\text{flat}} \in \mathbb{R}^{D\times HW}$ and apply Non-negative Matrix Factorization (NMF) (Collins, Achanta, and Susstrunk 2018):

$$\boldsymbol{G}_i^{\text{flat}} \approx \boldsymbol{U}_i \boldsymbol{V}_i, \quad \text{where } \boldsymbol{U}_i \in \mathbb{R}^{D\times K},\ \boldsymbol{V}_i \in \mathbb{R}^{K\times HW}. \tag{1}$$

Each row of $\boldsymbol{V}_i$, reshaped to $H \times W$, yields an attention heatmap $\boldsymbol{A}_i^{(k)}$ representing the k-th structural factor. Collectively, the K heatmaps form an explanation tensor $\boldsymbol{A}_i \in \mathbb{R}^{K\times H\times W}$, which serves as a factorized and interpretable decomposition of the original activation space.

To assess the relevance of each structural factor k to label c, we construct a class-response matrix $\boldsymbol{P}_i \in \mathbb{R}^{K\times C}$ by projecting the representation $\boldsymbol{U}_i^{(k)}$ of each factor into the classification space:

$$P_i[k,c] = \text{softmax}\left(\boldsymbol{W}_{\text{cls}} \cdot \boldsymbol{U}_i^{(k)}\right)[c], \tag{2}$$

where $\boldsymbol{W}_{\text{cls}}$ are weights of the classification layer. We define a structural exclusivity score for each factor to reflect its spatial specificity:

$$\text{Excl}_i^{(k)} = \sum_{u,v} \underbrace{\frac{A_i^{(k)}(u,v)}{\sum_{k'} A_i^{(k')}(u,v) + \epsilon}}_{\text{pixel exclusivity}} \cdot \underbrace{\frac{A_i^{(k)}(u,v)}{\sum_{u,v} A_i^{(k)}(u,v) + \epsilon}}_{\text{region activation}}, \tag{3}$$

where ϵ is a small constant. Combining class relevance and structural specificity, we compute a *episodic verifiability score* for each label c as:

$$\mathcal{T}_i^{\text{Epi}}[c] = \text{Norm}\left(\sum_{k=1}^{K} P_i[k,c] \cdot \text{Excl}_i^{(k)}\right). \tag{4}$$

Here, $\text{Norm}(\cdot)$ is a row-wise min-max normalization over all labels. The outputs $\mathcal{T}_i^{\text{Epi}}[c] \in [0,1]$ reveal the class relevance and structural attention support for the label c. The episodic trust mechanism is beneficial for capturing spatially reliable labels during training. It supports meaningful ongoing corrections over time without using rigid label filtering.

Earlier attempts often leverage class-wise backpropagation in the production of activation maps. The attention heatmaps and class-response matrices generated using the NMF approximation instead project structural factors into the classification space. This allows for a single forward pass that produces relevance scores for all labels. The design enhances both efficiency and stability of computation, which is particularly beneficial in multi-label settings with noisy supervision. Episodic memory provides interpretable spatial signals for the estimation of label reliability. It is of a strongly structural nature, which will serve as a base for later semantic alignment and generative supervision.

Semantic Memory via Structural Factor-Guided Prototype Reasoning

Episodic memory reduces the impact of noisy labels that lack spatial factor-based evidence. In addition, semantic memory uses class-level prototypes to verify semantic consistency between images and labels. From the cognitive perspective, this is indicative of long-term semantic knowledge consolidation, thereby enabling humans to validate current semantic interpretations with prototypical representations developed through prior perceptual experience.

We introduce a memory bank of prototypes $\{\boldsymbol{r}_c\}_{c=1}^{C}$, where $\boldsymbol{r}_c \in \mathbb{R}^D$ is the prototype for the class c. We utilize the structure-aware representations extracted through spatial factors to update the prototypes dynamically during training. For a given image $\boldsymbol{x}_i$, we calculate its class-specific representation by applying the factor-aware class attention map $\boldsymbol{W}_i[c] = \sum_k P_i[k,c] \cdot A_i^{(k)} \in \mathbb{R}^{H\times W}$ to the final convolutional feature map $\boldsymbol{F}_i \in \mathbb{R}^{D\times H\times W}$:

$$\boldsymbol{f}_i[c] = \sum_{u,v} \tilde{W}_{i,uv}[c] \cdot \boldsymbol{F}_{i,:,uv}, \tag{5}$$

where $\tilde{W}_i[c]$ is the normalized attention map over spatial positions. We calculate the mean of the class-specific representations $\boldsymbol{f}_i[c]$ within a mini-batch for each class c, i.e., $\bar{\boldsymbol{f}}_c = \frac{1}{|\mathcal{I}_c|}\sum_{i\in\mathcal{I}_c} \boldsymbol{f}_i[c]$, where $\mathcal{I}_c$ denotes the set of instances with label c. We then utilize $\bar{\boldsymbol{f}}_c$ to update the prototype representation $\boldsymbol{r}_c$ via the exponential moving average (Zhang et al. 2023):

$$\boldsymbol{r}_c \leftarrow \lambda \boldsymbol{r}_c + (1-\lambda)\cdot \bar{\boldsymbol{f}}_c, \tag{6}$$

where $\lambda \in (0,1)$ is the momentum coefficient. This update strategy facilitates stable prototype refinement during training and mitigates the influence of label noise.

We then compare class-specific features $\boldsymbol{f}_i[c]$ with corresponding class prototypes to assess the semantic plausibility of labels. It explores semantic structures in the representation space to enhance the verification of supervision signals in the presence of multi-label noise. Specifically, we define a *semantic verifiability score* as

$$\mathcal{T}_i^{\text{Sem}}[c] = \cos\left(\boldsymbol{f}_i[c], \hat{\boldsymbol{r}}_c\right). \tag{7}$$

This semantic score provides a complementary signal to episodic verification. We combine them to form a joint trust estimation,

$$s_i[c] = \mathcal{T}_i^{\text{Epi}}[c] \cdot \mathcal{T}_i^{\text{Sem}}[c]. \tag{8}$$

We maintain a momentum-based memory $\boldsymbol{w}_i \in [0,1]^C$ for the trust score of each instance to enhance its temporal stability and avoid rapid changes during training,

$$\boldsymbol{w}_i \leftarrow \mu \cdot \boldsymbol{w}_i + (1-\mu)\cdot \boldsymbol{s}_i, \tag{9}$$

where $\mu \in (0,1)$ controls the update rate. This momentum update allows the supervision signals to be adjusted progressively. Let $\boldsymbol{y}_i \in \{0,1\}^C$ denote the original label vector. The calibrated supervision target is then obtained by

$$\tilde{\boldsymbol{y}}_i = \boldsymbol{y}_i \odot \boldsymbol{w}_i. \tag{10}$$

Finally, we compute a weighted cross-entropy loss using the calibrated targets,

$$\mathcal{L}_{\text{veri}} = -\frac{1}{|\mathcal{D}|}\sum_{(\boldsymbol{x}_i,\boldsymbol{y}_i)\in\mathcal{D}}\sum_{c=1}^{C} \tilde{\boldsymbol{y}}_i[c]\cdot\log\left(\sigma(f(\boldsymbol{x}_i))[c]\right), \tag{11}$$

where $\sigma(\cdot)$ denotes the softmax function and $f(\boldsymbol{x}_i) \in \mathbb{R}^C$ is the classification logit obtained from projecting the hash embeddings. This objective integrates both structural and semantic evidence and provides a reliable mechanism for calibrating the noisy supervision.

Reconstructive Memory via Diffusion-Based Mixup

Episodic memory reduces the impact of unreliable labels through structural attention and semantic memory verifies the semantic reliability of labels using structural factor-guided prototypes. Human cognition often supplements these mechanisms with reconstructive memory. It is the ability to infer missing content by utilizing prior knowledge from multiple senses. In particular, the brain can synthesize reasonable experiences by interpolating between semantically related episodes when faced with ambiguous or insufficient information situations. Inspired by this generative principle, we introduce a *diffusion-based mixup* strategy that simulates memory reconstruction through cross-instance semantic blending guided by structural and semantic cues.

Specifically, we leverage a pretrained diffusion model (Ramesh et al. 2022; Wang et al. 2024) to provide rich priors for generating structurally and semantically reasonable interpolations. Given two input images $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ with calibrated labels $\tilde{\boldsymbol{y}}_i$ and $\tilde{\boldsymbol{y}}_j$, we first utilize the model's image encoder to embed both images into a latent semantic space:

$$\boldsymbol{z}_i = \text{enc}(\boldsymbol{x}_i), \quad \boldsymbol{z}_j = \text{enc}(\boldsymbol{x}_j), \tag{12}$$

where $\boldsymbol{z}_i, \boldsymbol{z}_j \in \mathbb{R}^d$ denote latent representations, into which the model incorporates both structural and semantic priors. We then perform spherical linear interpolation (slerp) (Ramesh et al. 2022) between the two latent representations to generate a mixed representation,

$$\hat{\boldsymbol{z}} = \text{slerp}(\boldsymbol{z}_i, \boldsymbol{z}_j, \lambda), \quad \lambda \sim \text{Beta}(\alpha,\alpha), \tag{13}$$

where λ is sampled from a symmetric Beta distribution with parameter of α and determines the relative contribution of each source instance. We then decode the interpolated latent representation $\hat{\boldsymbol{z}}$ into a synthetic image $\hat{\boldsymbol{x}}$ through a multi-step denoising diffusion process:

$$\hat{\boldsymbol{x}} = \text{diff\_dec}(\hat{\boldsymbol{z}}). \tag{14}$$

The synthetic image $\hat{\boldsymbol{x}}$ is assigned a soft label $\hat{\boldsymbol{y}}$, which reflects the semantic mix:

$$\hat{\boldsymbol{y}} = \lambda\tilde{\boldsymbol{y}}_i + (1-\lambda)\tilde{\boldsymbol{y}}_j. \tag{15}$$

The synthesized data pairs are incorporated into the training set and used to further optimize the model with a cross-entropy objective:

$$\mathcal{L}_{\text{recon}} = -\frac{1}{|\hat{\mathcal{D}}|}\sum_{(\hat{\boldsymbol{x}},\hat{\boldsymbol{y}})\in\hat{\mathcal{D}}}\sum_{c=1}^{C} \hat{\boldsymbol{y}}[c]\cdot\log\left(\sigma(f(\hat{\boldsymbol{x}}))[c]\right), \tag{16}$$

where σ means the softmax function and $\hat{\mathcal{D}}$ means the set of synthesized data pairs. This module leverages pretrained knowledge to perform structure-aware reconstructions. These reconstructions enrich the supervision signals for semantically uncertain areas and enhance the robustness of the model against label noise. It enables the model to generalize across sparsely sampled regions and improve the model's decision boundaries. The reconstructive memory complements the other two memory mechanisms to form a closed-loop memory system for robust multi-label learning.

Method	CIFAR-10				FLICKR25K				NUS-WIDE				MS COCO			
Bits	16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
DPSH	51.89	54.04	54.41	56.32	58.29	59.72	59.89	59.65	42.98	44.32	45.74	46.17	32.69	33.47	34.07	34.33
HashNet	44.27	45.03	46.96	47.20	56.83	57.77	58.34	60.63	43.61	44.17	45.21	47.01	31.93	32.09	32.58	36.17
SPQ	41.11	42.35	43.87	45.67	52.75	53.21	54.18	54.56	40.03	41.20	41.75	42.31	27.98	28.35	28.97	30.17
DCH	54.17	54.28	55.27	57.17	58.87	59.54	60.20	60.09	43.31	44.39	45.69	46.29	34.52	35.15	35.81	36.09
FSDH	53.89	54.55	55.26	56.89	57.77	58.59	59.60	60.04	43.28	44.31	44.52	44.87	34.08	34.39	35.12	35.37
GreedyHash	50.97	52.14	53.41	54.18	57.98	58.04	59.05	59.77	43.78	44.26	44.49	44.69	34.17	34.21	34.31	34.66
JMLH	49.93	51.58	54.48	56.01	58.54	58.36	58.61	58.71	43.15	43.89	44.81	45.14	34.18	34.22	34.39	34.81
DPN	49.77	51.02	54.12	56.21	59.23	60.12	59.98	60.54	44.02	44.23	45.34	45.98	34.23	35.01	35.77	36.21
WGLHH	52.57	53.18	55.18	56.43	60.02	59.10	59.72	60.27	43.79	46.08	46.29	46.56	35.33	35.14	36.32	36.87
CSQ	53.13	54.27	55.33	58.25	59.47	60.02	60.38	61.37	43.89	43.67	44.92	46.13	34.27	34.12	34.60	34.78
OrH	54.13	55.15	57.67	58.37	58.39	60.29	59.54	60.19	44.44	44.74	44.45	46.61	34.84	34.90	34.88	34.97
REL	56.19	58.58	60.99	61.74	59.76	61.27	61.46	61.96	44.76	45.02	45.96	46.21	34.85	34.97	35.04	35.12
Jo-SRC	55.75	57.43	59.91	60.12	60.12	60.59	60.42	61.27	44.97	45.36	46.12	47.78	35.12	35.46	35.79	35.67
DIOR	60.96	67.89	69.36	71.17	64.36	65.44	66.78	68.21	50.02	51.40	52.26	52.64	37.14	38.22	38.78	38.92
STAR	69.05	69.82	70.52	71.83	70.54	70.76	71.40	71.96	56.15	56.77	57.69	58.17	41.72	42.04	42.86	43.21
CogniTrust (Ours)	**70.55**	**70.91**	**71.68**	**73.89**	**71.54**	**72.72**	**74.62**	**75.75**	**59.39**	**62.29**	**62.16**	**65.27**	**44.99**	**46.07**	**46.16**	**47.11**

Table 1: MAP scores on four benchmark datasets with pairflip label noise.

Unified Training Objective

CogniTrust's memory mechanisms are integrated into a joint training objective combining two complementary losses that verify the reliability of supervision and enrich the supervision signals for uncertain regions, respectively:

$$\mathcal{L} = \mathcal{L}_{\text{veri}} + \mathcal{L}_{\text{recon}}. \tag{17}$$

Each component plays a crucial role in obtaining reliable supervision under multi-label noise. $\mathcal{L}_{\text{veri}}$ calibrates noisy labels by employing structurally interpretable attention and prototype-based semantic grounding, which guarantees that structurally and semantically plausible signals are exploited in the supervision. $\mathcal{L}_{\text{recon}}$ implements generative regularization through diffusion-based mixup, making it possible to interpolate reasonable instances and improve the class boundaries. The two losses together form a closed-loop learning system that dynamically verifies, calibrates, and reconstructs the supervising signals, guaranteeing that CogniTrust learns compact yet strong hash codes when facing the multifacets of label ambiguity and corruption.

Experiment

Experimental Setup

Datasets. We test our model on four image retrieval benchmarks: CIFAR-10 (Krizhevsky, Hinton et al. 2009), Flickr25k (Huiskes and Lew 2008), NUS-WIDE (Chua et al. 2009), and MS COCO (Lin et al. 2014). To simulate real-world label noise during training, we introduce synthetic corruption into training labels using two noise types: symmetric noise (uniformly reassigns labels) and pairflip noise (swaps labels between semantically similar categories). The intensity of this introduced noise varied from 20% to 80%, increasing in 20% increments, allowing for a comprehensive analysis of our model's performance under different levels of label corruption.

Baseline Methods. We compare CogniTrust with a series of baselines grouped into three categories: standard deep hashing methods, noisy label learning methods, and noise-robust hashing methods. The standard deep hashing methods include WGLHH (Tu et al. 2021), DPSH (Li, Wang, and Kang 2015), HashNet (Cao et al. 2017b), DCH (Cao et al. 2018), FSDH (Gui et al. 2017), SPQ (Jang and Cho 2021), GreedyHash (Su et al. 2018), JMLH (Shen et al. 2019), DPN (Fan et al. 2020), OrH (Hoe et al. 2021), and CSQ (Yuan et al. 2020). Noisy label learning methods include REL (Xia et al. 2021) and Jo-SRC (Yao et al. 2021). While noise-robust hashing methods include DIOR (Wang et al. 2023) and STAR (Long et al. 2025).

Evaluation Metric and Implementation Details. We adopt Mean Average Precision (MAP) to evaluate the performance, which is a standard retrieval metric. Experiments are conducted using PyTorch on a single NVIDIA A40 GPU. We adopt VGG-16 (Simonyan and Zisserman 2014) as the backbone network for fair comparison with baselines. The training process was configured with a batch size of 24 and a learning rate of 0.001 using the stochastic gradient descent (SGD). The number of structural factors is set to 5, and the mixup parameter α is set to 0.4 by default. For the pretrained diffusion model, we utilize karlo-v1-alpha-image-variations[1], a text-conditional diffusion model based on unCLIP (Ramesh et al. 2022).

Experimental Results

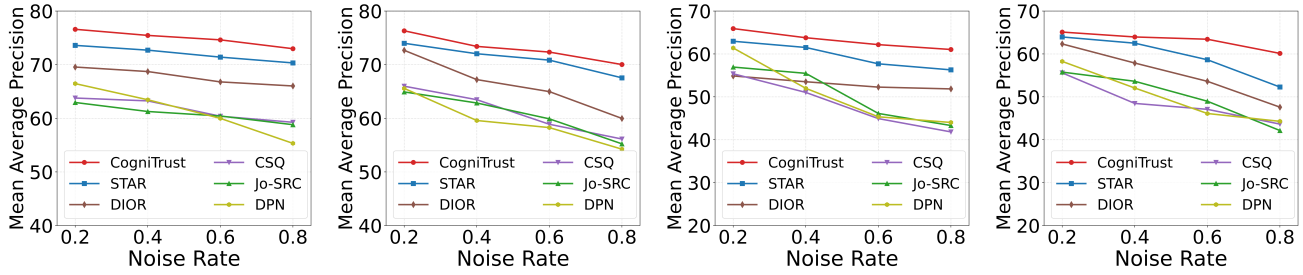
Performance Comparison. We report the retrieval performance of all methods on four benchmark datasets under two types of label noise in Tables 1 and 2. Noise rates of both types are 60%. Overall, our framework CogniTrust achieves the best performance against other hashing methods on all four benchmark datasets. In particular, CogniTrust outperforms the closest competitor on NUS-WIDE by 12.2% (65.27 vs. 58.17) and on MS COCO by 9.6% (46.92 vs. 42.79) at 128 bits under pairflip and symmetric noise respectively, which demonstrates the remarkable capability of our framework for robust hashing under label corruption.

Furthermore, CogniTrust offers several advantages: (1) CogniTrust surpasses existing noise-robust hashing methods on all the datasets, highlighting the necessity of a cognitive trust mechanism for robust hashing with noisy labels; (2) our

[1]https://huggingface.co/kakaobrain/karlo-v1-alpha-image-variations.

Method	CIFAR-10				FLICKR25K				NUS-WIDE				MS COCO			
Bits	16	32	64	128	16	32	64	128	16	32	64	128	16	32	64	128
DPSH	45.36	47.35	48.27	49.06	56.67	57.09	57.86	58.24	44.03	44.82	45.60	45.97	30.29	31.21	31.87	32.46
HashNet	42.54	43.24	44.47	45.29	53.47	54.69	56.03	56.87	43.21	44.37	45.12	46.05	27.17	27.69	28.54	28.89
DCH	45.29	47.75	48.19	49.24	57.02	58.20	58.44	59.12	44.36	44.21	44.98	45.88	31.55	32.39	32.99	33.41
GreedyHash	42.73	45.96	47.51	49.81	56.89	57.21	58.01	58.95	43.17	43.82	44.45	45.00	30.70	31.24	31.97	32.63
JMLH	46.58	48.15	48.74	48.89	57.53	58.55	59.13	60.94	44.47	45.56	45.92	46.02	31.38	32.64	33.19	33.77
DPN	44.57	47.08	48.09	48.56	56.82	57.43	58.26	59.83	44.87	45.72	46.11	46.19	31.11	31.65	31.59	32.13
WGLHH	47.92	49.83	50.19	52.35	57.17	57.99	58.46	59.23	45.11	45.91	46.68	47.32	32.21	33.13	33.74	34.11
CSQ	50.08	51.75	54.21	55.39	57.38	58.15	58.88	59.12	46.01	46.55	47.08	47.65	31.97	32.45	33.71	34.60
OrH	49.87	50.65	51.98	53.26	57.16	58.73	59.12	60.04	46.58	47.13	47.86	49.24	31.76	32.74	33.06	33.89
REL	50.39	51.21	52.64	53.97	58.68	59.02	59.46	60.35	47.05	47.67	48.18	48.86	32.47	33.27	34.29	35.01
Jo-SRC	50.82	51.42	51.96	53.62	58.12	58.91	59.89	60.87	47.79	48.14	48.97	49.34	32.59	33.19	34.61	35.11
DIOR	58.76	59.30	59.82	60.99	63.83	64.05	64.97	65.40	52.26	52.78	53.61	54.06	35.13	36.33	37.01	38.22
STAR	64.85	65.03	65.52	66.76	69.57	70.11	70.84	71.53	57.24	57.83	58.64	59.89	40.80	41.22	41.87	42.79
CogniTrust (Ours)	**65.56**	**66.02**	**67.78**	**68.12**	**70.69**	**71.59**	**72.34**	**73.91**	**60.97**	**61.06**	**63.42**	**64.93**	**43.98**	**45.30**	**45.75**	**46.92**

Table 2: MAP scores on four benchmark datasets with symmetric label noise.



(a) Flickr25K (pair) (b) Flickr25K (sym) (c) NUS-WIDE (pair) (d) NUS-WIDE (sym)

Figure 2: Performance comparison under different noise levels on Flickr25K and NUS-WIDE.

proposed CogniTrust demonstrates robustness across various data distributions and noise patterns on large-scale and complex datasets like MS COCO and NUS-WIDE; (3) CogniTrust consistently shows superiority and robustness across longer code lengths (64 and 128 bits), proving its scalability to various dimensions of hash codes.

Effects of Different Noisy Rates. Figure 2 illustrates CogniTrust's robustness on Flickr25K and NUS-WIDE under increasing noise levels for 64-bit hashing. While most baselines, including CSQ (Yuan et al. 2020), Jo-SRC (Yao et al. 2021), and DPN (Fan et al. 2020), show significant performance deterioration, especially beyond 0.4 noise rate, CogniTrust consistently maintains high MAP scores with less degradation even at 0.8 noise rate. For instance, on NUS-WIDE with pairflip noise (Figure 2(c)), our proposed CogniTrust's MAP remains above 60, significantly outperforming STAR and DIOR as noise intensifies. This superiority becomes evident in high-noise scenarios when other methods fail. It validates the fact that the joint modeling of three cognitive memory mechanisms provides improved robustness in extremely noisy environments.

Ablation Study

We investigate the performance of each major component of the model with respect to four aspects: Episodic Verifiability Score (EVS), Semantic Verifiability Score (SVS), combined episodic-semantic verifiability mechanism (EVS&SVS), and the Diffusion-based Mixup module (DM). The results are summarized in Tables 3.

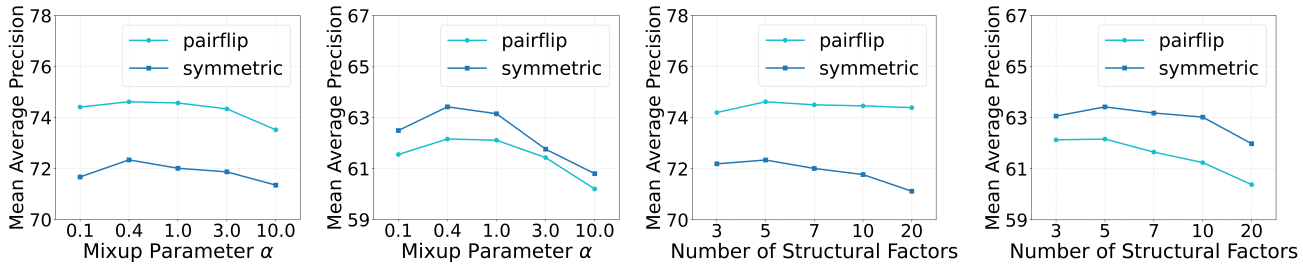
Method	CIFAR-10				Flickr25K			
Noise Type.Bits	pair.32	sym.32	pair.64	sym.64	pair.32	sym.32	pair.64	sym.64
CogniTrust w/o EVS	69.97	65.11	70.74	65.90	71.31	71.39	72.72	71.40
CogniTrust w/o SVS	70.41	65.23	69.79	66.37	70.83	70.68	71.89	70.97
CogniTrust w/o EVS&SVS	69.55	64.77	69.08	65.09	70.09	69.42	71.92	70.65
CogniTrust w/o DM	68.03	64.95	70.22	66.14	70.44	69.77	70.53	71.02
CogniTrust (Ours)	**70.91**	**66.02**	**71.68**	**67.78**	**72.72**	**71.59**	**74.62**	**72.34**

Table 3: Ablation studies on CIFAR-10 and Flickr25K under pairflip and symmetric label noise.

Effect of Episodic Verifiability Score. Removing EVS from our pipeline causes a marked drop off in performance over all datasets and noise settings, showing the essential nature of the episodic verifiability score computation for reliable label identification. A significant drop in performance is observed on the challenging Flickr25K dataset (e.g., 72.72 vs 74.62 for 64-bit codes under pairflip noise), hence showing the value of factorized structural attention in providing robust performance against label noise.

Effect of Semantic Verifiability Score. The removal of SVS reduces the model's performance for both datasets in all noise scenarios. This stable decline shows that SVS plays an essential role in guiding the model to learn from noisy labels through prototype-based reasoning and helps maintain reliable semantic structure during training.

Effect of Diffusion-Based Mixup Module. Omitting the diffusion-based mixup module also leads to consistent performance degradation across all the noise scenarios, which proves that reconstructive memory is beneficial to the



(a) Flickr25K (b) NUS-WIDE (c) Flickr25K (d) NUS-WIDE

Figure 3: Sensitivity analysis of performance with respect to two hyperparameters, mixup parameter α and number of structural factors, with 64-bit hash codes under pairflip and symmetric noise.

model's robustness. Maybe the reason is that diffusion-based mixup provides effective boundary regularization and improves the learning process in uncertain regions.

Sensitivity Analysis

We also investigate the sensitivity of our proposed CogniTrust to the hyperparameters. Specifically, we look into the effect of the mixup parameter α for reconstructive memory and the number of structural factors K for episodic memory.

Analysis of Mixup Parameter α. As illustrated in Figure 3(a) and Figure 3(b), our proposed CogniTrust shows optimal performance when α is 0.4. Performance declines when α is too high (10.0), as excessive interpolation introduces artificial noise. Conversely, too low (0.1) α leads to insufficient data augmentation, limiting generalization. This demonstrates that our diffusion-based mixup strategy can achieve optimal performance when balancing conservative and aggressive augmentation.

Effect of Number of Structural Factor K. Figure 3(c) and Figure 3(d) reveal how K influences the model's performance. CogniTrust achieves peak performance when K is 5. Too few components ($K = 3$) are insufficient to capture diverse semantic patterns, whereas an overly large decomposition ($K = 20$) leads to redundant factors that disrupt meaningful structure. We find that a moderate number of components provides better performance, indicating that appropriate factor decomposition is essential for effective trust modeling in our cognitive memory framework.

Visualization and Case Study

To examine how our proposed CogniTrust strengthens supervision in noisy settings, we visualize the behavior of the diffusion-based mixup module. Figure 4 shows four cases where interpolation in the diffusion latent space produces a semantically coherent intermediate sample from two input images. The intermediate images form coherent transitions between source instances, which preserve structural and semantic consistency rather than collapsing into pixel-level mixtures. These synthetic samples enrich the training distribution with plausible variations and provide additional guidance when labels are noisy. Therefore, it helps the model stabilize decision boundaries and improve robustness under noisy multi-label supervision.

Figure 4: Case studies of the diffusion-based mixup module. This module interpolates between two original images to generate a semantically coherent intermediate sample.

Conclusion

In this paper, we introduce a general framework CogniTrust for robust multi-label hashing against noisy supervision. Our design is guided by cognitive theories of human memory and centers on three interacting components: episodic, semantic, and reconstructive mechanisms. This triadic memory system dynamically evaluates label credibility and produces reliable supervision based on both structural and semantic perspectives. Extensive experiments on a range of widely-used hashing datasets demonstrate the superior performance and robustness of our proposed CogniTrust under symmetric and pairflip noise, especially in challenging high-noise conditions. Beyond offering a feasible approach for robust hashing, our proposed CogniTrust transforms cognitive understanding into algorithmic principles and provides a foundation for developing more trustworthy, interpretable, and broadly applicable machine learning systems.

Acknowledgements

Ming Zhang and Yiyang Gu are supported by grants from the National Key Research and Development Program of China with Grant No. 2023YFC3341203 and the National Natural Science Foundation of China (NSFC Grant Number 62276002). Dr Tao's research is partially supported by NTU RSR and Start Up Grants.

References

Arpit, D.; Jastrzębski, S.; Ballas, N.; Krueger, D.; Bengio, E.; Kanwal, M. S.; Maharaj, T.; Fischer, A.; Courville, A.; Bengio, Y.; et al. 2017. A closer look at memorization in deep networks. In *Proceedings of the International Conference on Machine Learning*.

Cao, Y.; Long, M.; Liu, B.; and Wang, J. 2018. Deep cauchy hashing for hamming space retrieval. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1229–1237.

Cao, Y.; Long, M.; Wang, J.; and Liu, S. 2017a. Deep visual-semantic quantization for efficient image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1328–1337.

Cao, Z.; Long, M.; Wang, J.; and Yu, P. S. 2017b. Hashnet: Deep learning to hash by continuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Chang, H.-S.; Learned-Miller, E.; and McCallum, A. 2017. Active bias: Training more accurate neural networks by emphasizing high variance samples. In *Proceedings of the Conference on Neural Information Processing Systems*.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y. 2009. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, 1–9.

Collins, E.; Achanta, R.; and Susstrunk, S. 2018. Deep feature factorization for concept discovery. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 336–352.

Dai, B.; Guo, R.; Kumar, S.; He, N.; and Song, L. 2017. Stochastic generative hashing. In *International Conference on Machine Learning*, 913–922. PMLR.

Fan, L.; Ng, K. W.; Ju, C.; Zhang, T.; and Chan, C. S. 2020. Deep Polarized Network for Supervised Learning of Accurate Binary Hashing Codes. In *IJCAI*, volume 825.

Gu, Y.; Wu, B.; Ran, Q.; Tu, R.-C.; Luo, X.; Xiao, Z.; Ju, W.; Tao, D.; and Zhang, M. 2025. SEGA: Shaping Semantic Geometry for Robust Hashing under Noisy Supervision. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Gui, J.; Liu, T.; Sun, Z.; Tao, D.; and Tan, T. 2017. Fast supervised discrete hashing. *IEEE transactions on pattern analysis and machine intelligence*, 40(2): 490–496.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Proceedings of the Conference on Neural Information Processing Systems*.

Hemmer, P.; and Steyvers, M. 2009. A Bayesian account of reconstructive memory. *Topics in cognitive science*, 1(1): 189–202.

Hoe, J. T.; Ng, K. W.; Zhang, T.; Chan, C. S.; Song, Y.-Z.; and Xiang, T. 2021. One Loss for All: Deep Hashing with a Single Cosine Similarity based Learning Objective.

Hu, P.; Peng, X.; Zhu, H.; Zhen, L.; and Lin, J. 2021. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5403–5413.

Huiskes, M. J.; and Lew, M. S. 2008. The mir flickr retrieval evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, 39–43.

Iscen, A.; Valmadre, J.; Arnab, A.; and Schmid, C. 2022. Learning with neighbor consistency for noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4672–4681.

Jang, Y. K.; and Cho, N. I. 2021. Self-supervised product quantization for deep unsupervised image retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, 12085–12094.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

Lai, H.; Pan, Y.; Liu, Y.; and Yan, S. 2015. Simultaneous feature learning and hash coding with deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3270–3278.

Li, J.; Socher, R.; and Hoi, S. C. 2021. Dividemix: Learning with noisy labels as semi-supervised learning. In *Proceedings of the International Conference on Learning Representations*.

Li, W.-J.; Wang, S.; and Kang, W.-C. 2015. Feature learning based deep supervised hashing with pairwise labels. *arXiv preprint arXiv:1511.03855*.

Li, W.-J.; Wang, S.; and Kang, W.-C. 2016. Feature learning based deep supervised hashing with pairwise labels. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 1711–1717.

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

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision*, 740–755. Springer.

Liu, S.; Niles-Weed, J.; Razavian, N.; and Fernandez-Granda, C. 2020. Early-learning regularization prevents memorization of noisy labels. *Advances in neural information processing systems*, 33: 20331–20342.

Long, Q.; Wang, H.; Sun, J.; Xiang, W.; Xiao, Y.; Zhao, Y.; and Luo, X. 2025. Learning Resistant Binary Descriptors Against Noise for Efficient Image Retrieval. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

McClelland, J. L.; McNaughton, B. L.; and O'Reilly, R. C. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3): 419.

Patrini, G.; Rozza, A.; Krishna Menon, A.; Nock, R.; and Qu, L. 2017. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Qiu, Z.; Su, Q.; Ou, Z.; Yu, J.; and Chen, C. 2021. Unsupervised Hashing with Contrastive Information Bottleneck. In *Proceedings of the International Joint Conference on Artificial Intelligence*.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical Text-Conditional Image Generation with CLIP Latents. *arXiv e-prints*, arXiv–2204.

Schacter, D. L.; Norman, K. A.; and Koutstaal, W. 2000. The cognitive neuroscience of constructive memory. *False-memory creation in children and adults*, 136–175.

Shen, Y.; Qin, J.; Chen, J.; Liu, L.; Zhu, F.; and Shen, Z. 2019. Embarrassingly simple binary representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 0–0.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Somepalli, G.; Fowl, L.; Bansal, A.; Yeh-Chiang, P.; Dar, Y.; Baraniuk, R.; Goldblum, M.; and Goldstein, T. 2022. Can neural nets learn the same model twice? investigating reproducibility and double descent from the decision boundary perspective. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 13699–13708.

Su, S.; Zhang, C.; Han, K.; and Tian, Y. 2018. Greedy hash: Towards fast optimization for accurate hash coding in cnn. *Advances in neural information processing systems*, 31.

Tu, R.-C.; Mao, X.-L.; Kong, C.; Shao, Z.; Li, Z.-L.; Wei, W.; and Huang, H. 2021. Weighted gaussian loss based hamming hashing. In *Proceedings of the 29th ACM International Conference on Multimedia*, 3409–3417.

Tulving, E. 1983. Elements of episodic memory.

Tulving, E.; et al. 1972. Episodic and semantic memory. *Organization of memory*, 1(381-403): 1.

Wang, H.; Jiang, H.; Sun, J.; Zhang, S.; Chen, C.; Hua, X.-S.; and Luo, X. 2023. DIOR: Learning to hash with label noise via dual partition and contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*, 36(4): 1502–1517.

Wang, J.; Zhang, T.; Sebe, N.; Shen, H. T.; et al. 2017. A survey on learning to hash. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 769–790.

Wang, Z.; Wei, L.; Wang, T.; Chen, H.; Hao, Y.; Wang, X.; He, X.; and Tian, Q. 2024. Enhance image classification via inter-class image mixup with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 17223–17233.

Xia, X.; Liu, T.; Han, B.; Gong, C.; Wang, N.; Ge, Z.; and Chang, Y. 2021. Robust early-learning: Hindering the memorization of noisy labels. In *Proceedings of the International Conference on Learning Representations*.

Yao, Y.; Sun, Z.; Zhang, C.; Shen, F.; Wu, Q.; Zhang, J.; and Tang, Z. 2021. Jo-src: A contrastive approach for combating noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Yuan, L.; Wang, T.; Zhang, X.; Tay, F. E.; Jie, Z.; Liu, W.; and Feng, J. 2020. Central similarity quantization for efficient image and video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3083–3092.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations*.

Zhang, J.; Huang, J.; Jiang, X.; and Lu, S. 2023. Black-box unsupervised domain adaptation with bi-directional atkinson-shiffrin memory. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11771–11782.

Zhou, X.; Liu, X.; Wang, C.; Zhai, D.; Jiang, J.; and Ji, X. 2021. Learning with noisy labels via sparse regularization. In *Proceedings of the IEEE/CVF international conference on computer vision*, 72–81.