Learning Semantic Boundaries: An Adaptive Structural Loss for Multi-Label Contrastive Learning

Ning Chen[1], Shen-Huan Lyu[1,2,3*], Yanyan Wang[1,3*], Bin Tang[1]
[1]Key Laboratory of Water Big Data Technology of Ministry of Water Resources, College of Computer Science and Software Engineering, Hohai University, Nanjing, China
[2]Department of Computer Science, City University of Hong Kong, Hong Kong, China
[3]National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China
[*]Corresponding authors: lvsh@hhu.edu.cn (Shen-Huan Lyu), yanyan.wang@hhu.edu.cn (Yanyan Wang)

***Abstract*—Multi-Label Contrastive Learning (MLCL) seeks to pull samples with shared labels closer in an embedding space. However, existing methods primarily adjust attractive forces without explicitly shaping a geometric structure that captures complex label semantics. This work targets learning an embedding space isomorphic to the semantic label structure, a challenge complicated by *ranking noise* arising from dense positives and *sampling distortion* caused by finite queue sizes. To address these problems, we propose Hierarchical Boundary Learning (HBL), a novel structured regularization loss. HBL partitions positive samples into soft and hard subsets based on Jaccard similarity, then enforces a dual-boundary constraint: a relative boundary between soft and hard positives to mitigate ranking noise, and an absolute boundary to anchor hard positives, preventing *positive sample expulsion*. A reliability gating mechanism further counters sampling distortion. Experiments on diverse multi-label datasets show that MLCL methods using HBL achieve significant improvements over prior methods across multiple evaluation metrics.**

Keywords-component; multi-label contrastive learning; structual loss; hierarchical boundary learning

I. INTRODUCTION

Contrastive learning [1, 2] has become a cornerstone of representation learning, excelling in single-label scenarios by pulling similar samples together and pushing dissimilar ones apart. However, when directly extended to the more general and challenging multi-label setting [3], its simple binary logic encounters a fundamental challenge. This is because, in a multi-label context, the relationship between two samples is not strictly binary; they can share a subset of labels, creating a continuous, non-binary degree of similarity [3]. This partial similarity renders the traditional contrastive objective ambiguous. Therefore, the central task of Multi-Label Contrastive Learning (MLCL) [4] is to learn an embedding space that faithfully captures this rich, structured semantic similarity.

To address this challenge, recent MLCL methods [4-6] primarily focus on modulating the attractive forces between anchors and positive samples—either by decomposing the loss or weighting positives based on label overlap. However, these approaches share a critical limitation: they emphasize how to pull but overlook how to arrange. That is, they fail to address a deeper geometric optimization question—how to safely and robustly organize positive samples in the embedding space under these complex attractive forces. As a result, the learned geometry often remains flat and unstructured. Worse still, naive attempts to impose internal ordering may backfire: in trying to separate closely and distantly related positives, the model may inadvertently repel the weaker positives too far, even beyond negatives. This contradicts the core goal of contrastive learning and degrades representational quality.

We posit that an ideal multi-label representation space should possess an intrinsic manifold structure isomorphic to the semantic structure of the label space. This implies that, beyond merely attracting all positive samples, our objective is to sculpt a geometric structure that reflects their varying semantic distances within a safe and consistent framework. However, translating this ideal into practice requires navigating a series of interconnected challenges rooted in the characteristics of real-world data. We identify three core problems: (1) *ranking noise*, which arises from the need to differentiate subtle semantic differences; (2) *positive sample expulsion*, a more perilous risk caused by improper optimization constraints; and (3) *sampling distortion*, introduced by finite queue sizes, which further exacerbates the former two issues.

Therefore, we propose a novel and adaptive structured regularization loss, termed Hierarchical Boundary Learning (HBL). At its core lies a dual-boundary constraint mechanism: a relative boundary that sculpts the embedding structure at a macroscopic level to suppress ranking noise, and an absolute boundary that anchors positive samples to prevent positive sample expulsion. To ensure this mechanism remains stable under the statistical uncertainty introduced by sampling distortion, HBL further incorporates a third key component—a reliability gate. This gate evaluates the reliability of the sampled batch and activates the structured loss only when the signal is sufficiently stable, thereby enhancing the overall robustness of the learning process.

Our main contributions are manifold: We not only propose the robust HBL method, with its dual-boundary constraint and reliability gate, but more importantly, we systematically identify, define, and address the three core challenges of ranking noise, positive sample expulsion, and sampling distortion when introducing structured information into MLCL. Experimental results on multiple multi-label datasets validate the effectiveness of our approach.

II. Related Work

Multi-Label Learning (MLL) aims to associate a single sample with multiple relevant labels [3]. Traditional methods mainly focus on label transformation and model adaptation, establishing the foundational methodology for multi-label learning [7-9]. However, they are limited in capturing label dependencies, lack end-to-end optimization capabilities, and often result in suboptimal feature representations. In recent years, deep learning-based methods have made notable progress by incorporating structures such as attention mechanisms or graph neural networks to explicitly model label co-occurrence in the label space [10, 11]. However, these approaches mainly focus on structural modeling in the output space or impose explicit graph constraints, making it difficult to capture latent high-order semantic relationships among labels within the representation space.

In contrast, contrastive learning has achieved remarkable breakthroughs in the field of representation learning [1, 2, 12-18]. By constructing effective positive and negative sample pairs and enforcing similarity-based constraints, it significantly enhances the semantic discriminability of the representation space. This line of success has also inspired recent efforts to integrate contrastive learning into multi-label learning. A key challenge in MLCL is how to define positive and negative samples, as label relationships are not simply binary or mutually exclusive. At the same time, effectively representing such complex relationships in the embedding space remains a critical problem. Current approaches [4-6, 19-21] primarily address the challenges of partial similarity and label structure modeling in multi-label contrastive learning through strategies such as loss function decomposition, positive sample reweighting, or predefined label hierarchies. However, these methods predominantly focus on the attractive forces between anchors and positive samples, overlooking a deeper geometric optimization problem: under such complex pulling dynamics, how should positive samples ultimately be arranged in the embedding space in a safe and robust manner? To this end, we propose Hierarchical Boundary Learning (HBL).

III. Method

This chapter systematically presents our method. We begin by reviewing MulSupCon [4] and introducing our central motivation: learning a representation space that is semantically isomorphic to the label space. Building on this foundation, we explore three core challenges that arise in practice. First, we examine two fundamental geometric issues—*ranking noise* and *positive sample expulsion*—followed by the statistical challenge that amplifies them: *sampling distortion*. We then introduce our key contribution: the adaptive HBL method. We show how its dual-boundary constraint mechanism addresses the two geometric challenges, and how the integrated reliability gate effectively mitigates sampling distortion, ensuring robust and reliable learning.

A. Preliminaries: Multi-Label Supervised Contrastiv Learning

For a given anchor sample i with its label set $\boldsymbol{y}^{(i)}$, MulSupCon [4] decomposes its learning objective across each of the labels it possesses, $y_j^{(i)} \in \boldsymbol{y}^{(i)}$. Ultimately, the loss for anchor i is:

$$\mathcal{L}_{\text{Con}}^{(i)} = \sum_{y_j^{(i)} \in \mathbf{y}^{(i)}} \frac{-1}{|\mathcal{P}_j^{(i)}|} \sum_{p \in \mathcal{P}_j^{(i)}} \log \frac{\exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p/\tau)}{\sum_{a \in \mathcal{A}^{(i)}} \exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_a/\tau)}, \quad (1)$$

where $\boldsymbol{z}$ represents the sample's embedding, $\mathcal{P}_j^{(i)}$ is the set of positive samples sharing label $y_j^{(i)}$ with anchor i, $\mathcal{A}^{(i)}$ is the full set of all samples, and τ is the temperature hyperparameter that balances the model's ability to distinguish between positive and negative sample. Although this method is effective, its strategy of indiscriminately attracting all samples that share a common label ignores the rich, non-binary semantic similarities inherent in multi-label scenarios. This results in a learned geometric structure of the embedding space that is flat. More importantly, any naive attempt to impose an internal ordering directly upon this foundation risks the over-expulsion of some positive samples, which in turn undermines the very foundation of contrastive learning.

B. Motivation: From Similarity Aggregation to Geometric Manifold Sculpting

The core motivation for our work stems from the manifold hypothesis [22]. We posit that an ideal multi-label representation space should not merely aggregate similar samples but should arrange them on a manifold whose geometric structure is isomorphic to the semantic structure of the labels. The key to achieving this goal lies in preserving the ordinal relations among samples. Traditional contrastive learning [1, 2] typically employs cosine similarity on L_2-normalized embeddings, confining the geometric space to a unit hypersphere. This metric primarily focuses on vector direction, with the ultimate goal of collapsing all positive samples to a single point on the sphere. This fundamentally conflicts with our objective of sculpting a local manifold with hierarchical and distance-based gradients, as it cannot express notions of proximity and distance. Fortunately, when feature vectors are L_2-normalized onto a unit hypersphere, a strictly monotonic relationship exists between Euclidean distance and cosine similarity. For two unit vectors $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$, their squared Euclidean distance $d^2(\boldsymbol{z}_i, \boldsymbol{z}_j)$ and cosine similarity $\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)$ satisfy:

$$d^2(\boldsymbol{z}_i, \boldsymbol{z}_j) = ||\boldsymbol{z}_i - \boldsymbol{z}_j||^2 = 2 - 2 \cdot \text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j). \quad (2)$$

Equation (2) explicitly demonstrates that minimizing Euclidean distance is equivalent to maximizing cosine similarity. This crucial mathematical equivalence allows us to have the best of both worlds: we can leverage the training stability of modern contrastive learning frameworks operating on the unit hypersphere, while indirectly yet precisely controlling the Euclidean distances between samples by optimizing their cosine similarity. This lays a solid theoretical and practical foundation for our proposed Hierarchical Boundary Learning.

TABLE I. STATISTICS OF POSITIVE SAMPLES PER ANCHOR ACROSS VARIOUS DATASETS (COMPUTED OVER THE ENTIRE TRAINING SET). THE UNDERSAMPLING RATE IS DEFINED AS THE RATIO OF THE AVERAGE NUMBER OF POSITIVES TO OUR FIXED QUEUE SIZE (4096).

Dataset	Min Positives	Max Positives	Mean Positives	Std Dev	Undersampling Rate	Complexity Profile
Scene	165	630	232.4	66.7	0.06	Moderate density, complete structure
Yeast	215	1481	1174.6	282.0	0.29	High density, complete structure
PASCAL	97	3194	1222.2	943.7	0.30	High density, complete structure
MIRFLICKR	792	19202	10977.8	3299.5	2.68	Extreme density, undersampled

C. *From Ideal to Reality: The Threefold Challenge in Practice*

In translating the ideal of learning a semantically isomorphic space into practice, we identify three interconnected and escalating challenges through an in-depth analysis of multi-label datasets (as shown in Table I).

1) Ranking Noise from Positive Crowding: The first challenge arises from the crowding phenomenon within the positive sample space. In these highly dense local regions, differences in label similarity among many positive samples become extremely subtle (e.g., 0.31 vs. 0.32). When the model is forced to distinguish such minute differences—which are likely attributable to statistical noise—and to generate gradients accordingly, a phenomenon we refer to as ranking noise emerges. As the model attempts to satisfy a multitude of low signal-to-noise ratio ranking constraints, its learning dynamics become disrupted, potentially compromising the acquisition of more meaningful macroscopic class-level separability in favor of preserving inconsequential local orderings.

2) The Fundamental Risk of Positive Sample Expulsion: However, circumventing ranking noise via macroscopic partitioning (e.g., categorizing positives as strong or weak) introduces a deeper systemic challenge: *positive sample expulsion*. In separating strong from weak positives in the embedding space, the model may resort to pushing weak positives away from the anchor. If left unconstrained, this expulsive force risks driving true positives beyond their intended attractive region—sometimes even farther than relevant negatives—thereby compromising the representational consistency central to contrastive learning.

3) Sampling Distortion as a Risk Multiplier: Finally, these geometric risks are further exacerbated by *sampling distortion*. As shown by the undersampling rate in Table I, large datasets like MIRFLICKR (rate = 2.68) suffer from severely biased and sparse sampling, where the queue fails to capture most positives. In contrast, smaller datasets (e.g., Scene, Yeast) remain adequately covered (rate < 1). This distortion amplifies earlier risks and can trigger catastrophic positive expulsion from a single poor mini-batch.

D. *The Solution: HBL*

To systematically address the three aforementioned challenges, we propose Hierarchical Boundary Learning (HBL). The core of HBL is the introduction of a novel composite structural loss. This loss function, through its three key internal components, precisely and separately resolves the problems of ranking noise, positive sample expulsion, and sampling distortion.

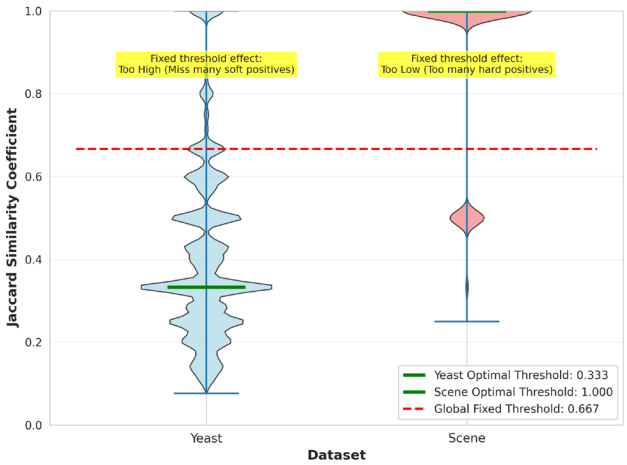

Figure 1. The failure of fixed thresholds across datasets with diverse Jaccard distributions. Violin plots show the Jaccard similarity densities of positive pairs in Yeast and Scene. A global fixed threshold (red dashed line) poorly fits both: it's too high for Yeast—missing most data—and too low for Scene—failing to separate clusters. This underscores the need for an adaptive strategy like our dynamic median threshold.

1) Countering Ranking Noise via Macroscopic Partitioning: To circumvent ranking noise, our first step is to perform a macroscopic, rather than microscopic, partitioning. The key to this partitioning is establishing a robust division threshold. A seemingly straightforward solution is to use a fixed threshold based on global statistics. However, as illustrated in Fig. 1, this "one-size-fits-all" strategy proves fundamentally ineffective when applied across different datasets. Moreover, during training, we are dealing with dynamic and localized data distributions. To address this, we adopt a dynamic median-thresholding strategy, which intelligently determines a tailored boundary for each anchor's local neighborhood. For each anchor i and its set of positive samples $\mathcal{P}^{(i)}$ within the current queue, this strategy dynamically computes the median of the local Jaccard similarity distribution and uses it as the anchor's own tailored division boundary, θ_{dynamic}:

$$\theta_{\text{dynamic}} = \text{Median}\{ S_{\text{Jaccard}}(\mathbf{y}^{(i)}, \mathbf{y}^{(p)}) \mid p \in \mathcal{P}^{(i)} \}, \tag{3}$$

where Jaccard similarity is $S_{\text{Jaccard}}(\mathbf{y}_a, \mathbf{y}_b) = \dfrac{|\mathbf{y}_a \cap \mathbf{y}_b|}{|\mathbf{y}_a \cup \mathbf{y}_b|}$. This adaptive strategy tailors the most reasonable boundary

between strong and weak relationships for each anchor. Based on this, we partition the positive samples into:

a) Soft Positive Set ($\mathcal{P}_s^{(i)}$): All positive samples satisfying $S_{\text{Jaccard}}(\mathbf{y}^{(i)}, \mathbf{y}_p) \geq \theta_{\text{dynamic}}$.

b) Hard Positive Set ($\mathcal{P}_h^{(i)}$): All positive samples satisfying $S_{\text{Jaccard}}(\mathbf{y}^{(i)}, \mathbf{y}_p) < \theta_{\text{dynamic}}$.

2) Countering Ranking Noise and Positive Sample Expulsion via a Dual-Boundary Constraint: After performing the macroscopic partitioning, we design a dual-boundary constraint mechanism to simultaneously address the two core geometric challenges. We operate in the L_2-normalized embedding space and use cosine similarity to indirectly control Euclidean distance.

a) Relative Boundary Loss ($\mathcal{L}_{\text{relative}}$): This constraint acts directly upon the macroscopic partitions, requiring the group of soft positives to be closer to the anchor than the group of hard positives. Specifically, it mandates that the most similar hard positive (the one with the highest similarity) must still have a lower similarity to the anchor than the least similar soft positive (the one with the lowest similarity). This group-wise constraint fundamentally avoids microscopic ranking of individual samples, thereby resolving the ranking noise problem.

$$\mathcal{L}_{\text{relative}}^{(i)} = \max(0, \max_{p_h \in \mathcal{P}_h^{(i)}} \text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_{p_h}) - \min_{p_s \in \mathcal{P}_s^{(i)}} \text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_{p_s}) + m_{\text{rel}}) \quad (4)$$

b) Absolute Boundary Loss ($\mathcal{L}_{\text{absolute}}$): To provide a safety net for our structural adjustments, we introduce an absolute boundary loss. It ensures that even the most distantly related hard positive maintains a higher similarity to the anchor than the most relevant negative sample (the hard negative). This anchoring operation, by establishing an inviolable bottom line, fundamentally resolves the risk of positive sample expulsion.

$$\mathcal{L}_{\text{absolute}}^{(i)} = \max(0, \max_{n \in \mathcal{N}^{(i)}} \text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_n) - \min_{p_h \in \mathcal{P}_h^{(i)}} \text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_{p_h}) + m_{\text{abs}}), \quad (5)$$

where m_{rel} and m_{abs} represent the margins of the two boundaries in (4) and (5), respectively, and $\mathcal{N}^{(i)}$ is the set of negative samples for anchor i.

3) Countering Sampling Distortion via a Reliability Gate: The effectiveness of this sophisticated geometric constraint mechanism depends entirely on the reliability of the statistical signal. However, as revealed in our analysis of the datasets (Table I), the sampling distortion caused by the finite queue can produce biased and unstable statistics (such as the median), severely disrupting the learning process. Therefore, to counter the statistical challenge of sampling distortion, the third key component of HBL is a reliability gate. This gate ensures that the dual-boundary loss is activated only when the statistics are reliable—that is, when the total number of positive samples for an anchor i, $|\mathcal{P}^{(i)}|$, exceeds a preset threshold $k_{\min}$.

$$\mathcal{L}_{\text{HBL}}^{(i)} = \begin{cases} \mathcal{L}_{\text{relative}}^{(i)} + \gamma \cdot \mathcal{L}_{\text{absolute}}^{(i)}, & |\mathcal{P}^{(i)}| \geq k_{\min} \\ & |\mathcal{P}_s^{(i)}| > 0,\ |\mathcal{P}_h^{(i)}| > 0 \\ 0, & \text{otherwise} \end{cases}, \quad (6)$$

where γ balance the relative force of hierarchical shaping and the anchoring force that prevents expulsion. This conditional strategy ensures that the model is not misled by noise introduced from distorted samples, thereby guaranteeing the robustness of the entire method.

4) Final Objective Function: Finally, our HBL loss serves as an auxiliary structural term, integrated with a base contrastive loss $\mathcal{L}_{\text{Con}}$ (such as MulSupCon). For a mini-batch of size N, the total loss is:

$$\mathcal{L}_{\text{total}} = \frac{1}{N} \sum_{i=1}^{N} (\mathcal{L}_{\text{Con}}^{(i)} + \lambda \cdot \mathcal{L}_{\text{HBL}}^{(i)}), \quad (7)$$

where λ balances the indiscriminate aggregating force provided by the basic contrastive loss and the internal structure shaping force imposed by the HBL loss.

IV. Experiments

In this section, we validate the effectiveness of our proposed HBL method through a series of extensive experiments. We aim to answer the following core research questions (RQs):

- **RQ1:** Does HBL achieve significant performance improvements over an existing SOTA method?
- **RQ2:** Is each carefully designed component of HBL—adaptive partitioning, the dual-boundary constraint, and the reliability gate—indispensable?
- **RQ3:** How sensitive is our method to its key hyperparameters?

A. Experimental Setup

1) Datasets: To comprehensively evaluate the performance and generalization capability of our method, we select four widely-used benchmark datasets with diverse characteristics and scales: Scene [7], Yeast [23], MIRFLICKR [24], and PASCAL-VOC [25]. These datasets cover a diverse range of scenarios, from bioinformatics and scene images to large-scale web images.

2) Evaluation Protocol and Principle of Fair Comparison: We follow the standard evaluation protocol in the fields of self-supervised and contrastive learning, employing linear probing to measure the quality of the pretrained representations. After the pretraining stage is complete, we freeze all parameters of the backbone encoder and train only a simple linear classifier on top of it. We choose linear probing over full-network fine-tuning because it provides a purer measure of the representation's intrinsic quality, rather than the network's fine-tuning capability on downstream tasks. This

TABLE II. MAIN RESULTS ON FOUR BENCHMARK DATASETS. THE BEST PERFORMANCE FOR EACH METRIC WITHIN EACH BASELINE COMPARISON GROUP (E.G., ANY-LOSS VS. ANY-LOSS+HBL) IS HIGHLIGHTED IN BOLD. THE UPWARD ARROW (↑) INDICATES THAT HIGHER IS BETTER FOR ALL METRICS.

Dataset	Method	p@1 ↑	mAP ↑	HA ↑	ebF1 ↑	maF1 ↑	miF1 ↑
Scene	MulSupCon	0.7935	0.8401	0.9213	0.7821	0.7810	0.7764
	MulSupCon+HBL	**0.8069**	**0.8459**	**0.9221**	**0.7860**	**0.7876**	**0.7821**
Yeast	MulSupCon	0.7415	0.4916	0.8000	0.6574	0.4780	0.6664
	MulSupCon+HBL	**0.7557**	**0.4944**	**0.8032**	**0.6579**	**0.4803**	**0.6677**
PASCAL-VOC	MulSupCon	0.6662	0.5325	**0.9453**	0.5631	0.5136	0.5850
	MulSupCon+HBL	**0.6749**	**0.5398**	0.9449	**0.5709**	**0.5156**	**0.5868**
MIRFLICKR	MulSupCon	0.8363	0.6035	0.9032	0.6406	0.5645	0.6689
	MulSupCon+HBL	**0.8379**	**0.6106**	**0.9033**	**0.6420**	**0.5713**	**0.6705**
Scene	Any-Loss	0.7993	0.8440	**0.9232**	0.7859	0.7767	0.7732
	Any-Loss+HBL	**0.8094**	**0.8478**	0.9213	**0.7865**	**0.7832**	**0.7790**
Yeast	Any-Loss	0.7535	0.4767	0.7950	0.6365	0.4645	0.6495
	Any-Loss+HBL	**0.7688**	**0.4792**	**0.7995**	**0.6465**	**0.4661**	**0.6536**
PASCAL-VOC	Any-Loss	0.6268	0.4974	0.9414	0.5326	0.4730	0.5530
	Any-Loss+HBL	**0.6581**	**0.5048**	**0.9433**	**0.5511**	**0.4917**	**0.5728**
MIRFLICKR	Any-Loss	**0.8326**	0.5624	0.8981	**0.6244**	0.5240	0.6475
	Any-Loss+HBL	0.8310	**0.5654**	**0.8984**	0.6225	**0.5254**	**0.6494**

aligns with the standard evaluation paradigm in the contemporary contrastive learning literature.

3) Evaluation Metrics: To comprehensively evaluate the performance of our multi-label classification model, we adopt six standard metrics: mean average precision (mAP), precision@1 (p@1), macro-F1 (maF1), micro-F1 (miF1), example-based F1 (ebF1), and Hamming Accuracy (HA). These metrics jointly assess ranking quality (mAP, p@1), label-level performance (mi/ma-F1), instance-level accuracy (ebF1), and overall prediction consistency (HA).

4) Implementation Details: Our method is implemented as an auxiliary structural loss built upon MulSupCon [4] and a SupCon [12] variant, Any-Loss. We use a simple multi-layer perceptron as the backbone for the Yeast and Scene datasets, and ResNet-50 [26] for PASCAL-VOC and MIRFLICKR. The embedding dimension is fixed at 128 for all models. Optimization is performed using the AdamW [27] optimizer with a CosineAnnealingWarmRestarts [28] learning rate scheduler. The feature queue size is set to 4096. To ensure fair comparison and optimal performance, we apply Optuna [29] for automated hyperparameter optimization. The search space includes: temperature $\tau \in \{0.01, 0.1\}$, momentum update rate $m \in \{0.99, 0.999, 0.9999\}$, learning rate and weight decay sampled log-uniformly from $[1\times10^{-6}, 4\times10^{-4}]$ and $[1\times10^{-7}, 1\times10^{-5}]$, respectively; scheduler cycle length $T_0 \in [5, 50]$ (log-scaled integers), and batch size from $\{32, 64, 128, 256\}$. For our proposed HBL method, we additionally tune: structural loss weight $\lambda \in \{0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1.0\}$, internal balancing factor $\gamma \in \{0.5, 0.7, 0.8, 1.0, 1.5, 2.0\}$, relative boundary margin $m_{\text{rel}} \in \{0.05, 0.1\}$, absolute boundary margin $m_{\text{abs}} \in \{0.2, 0.3\}$, and reliability gate threshold $k_{\min} \in \{64, 128, 256\}$.

B. Performance Comparison (RQ1)

We integrate HBL into two representative and evolutionarily related methods: Any-Loss and MulSupCon. The former is a direct extension of SupCon to multi-label learning, embodying a basic strategy of indiscriminate aggregation. The latter adopts a decomposed aggregation mechanism and represents the current SOTA in MLCL. This focused comparative setup is designed to answer two key questions: (1) Can HBL significantly enhance the performance of a foundational method? (2) Can HBL still provide meaningful improvements when applied to an existing SOTA method? By answering both questions affirmatively, we demonstrate that the geometric structure sculpting philosophy introduced by HBL marks a substantial advance over existing aggregation-based approaches and offers a promising direction for achieving better performance. To ensure absolute fairness in our comparison, we strictly adhere to the principle of freezing the baseline's optimal configuration and tuning only the parameters of the newly added module. We first reproduce the baseline performance. Then, building upon its optimal configuration, we keep all shared parameters strictly identical and exclusively tune the hyperparameters introduced by our HBL method.

As shown in Table Ⅱ, our proposed HBL consistently brings significant performance improvements across all

TABLE III. ABLATION STUDY OF HBL COMPONENTS ON PASCAL-VOC.

Method	p@1 ↑	mAP ↑	HA ↑	ebF1 ↑	maF1 ↑	miF1 ↑
Any-Loss	0.6268	0.4974	0.9414	0.5326	0.4730	0.5530
w/ $\mathcal{L}_{\text{relative}}$	0.6351	0.4667	0.9403	0.5226	0.4652	0.5462
w/o Gate	0.6492	0.4987	0.9430	0.5500	0.4895	0.5677
w/all	**0.6581**	**0.5048**	**0.9433**	**0.5511**	**0.4917**	**0.5728**

baseline methods and datasets: (1) Enhancing a simple baseline: Any-Loss+HBL substantially outperforms the original Any-Loss, demonstrating that incorporating structural information is effective even on top of the most basic aggregation strategy. (2) Further boosting a SOTA method: MulSupCon+HBL surpasses the original MulSupCon on multiple metrics across all datasets, with particularly notable gains on more challenging datasets like PASCAL and MIRFLICKR. This highlights that the geometric structure modeling provided by HBL complements existing methods and offers a critical advantage that is otherwise missing from current SOTA approaches.

C. *Ablation Study (RQ2)*

To investigate the contribution of each design component in HBL and better understand its underlying mechanisms, we conduct a series of ablation studies on the PASCAL-VOC dataset using Any-Loss as the base model. We sequentially evaluate the following four configurations: (1) Any-Loss: Serves as our performance baseline. (2) w/ $\mathcal{L}_{\text{relative}}$: Adds only our relative boundary loss on top of the baseline, designed to counteract ranking noise. (3) w/o Gate: Extends configuration (2) by incorporating the absolute boundary loss, while omitting the reliability gate. (4) w/all: Our full HBL method, incorporating all components.

The results, presented in Table III, clearly illustrate the distinct and complementary contributions of each component. Adding only the relative boundary loss $\mathcal{L}_{\text{relative}}$ causes a notable drop in mAP (from 0.4974 to 0.4667), confirming the positive sample expulsion risk—without a stabilizing force, hard positives may be overly pushed away, disrupting the representation space. Introducing the absolute boundary loss $\mathcal{L}_{\text{absolute}}$ immediately reverses this degradation, lifting mAP to 0.4987. This validates $\mathcal{L}_{\text{absolute}}$ as a necessary safety anchor that enables effective structural shaping by $\mathcal{L}_{\text{relative}}$. Finally, incorporating the reliability gate yields further gains (mAP 0.5048), underscoring its role in mitigating sampling distortion and stabilizing training. In summary, the ablation study demonstrates that the three key components of HBL play complementary roles in addressing ranking noise, positive sample expulsion, and sampling distortion. Together, they form the core mechanism of HBL as a structured regularization term, providing effective structural enhancement for MLCL methods.

D. *Parameter Sensitivity Analysis (RQ3)*

To evaluate the robustness of HBL and analyze the roles of its key hyperparameters, we conduct a sensitivity analysis on

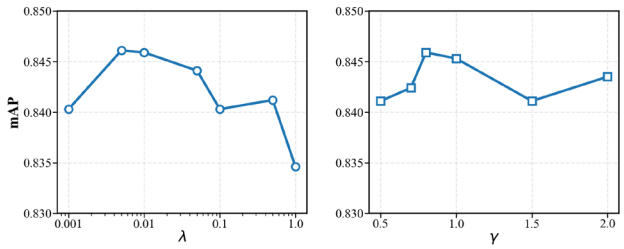

Figure 2. Hyperparameter sensitivity analysis of HBL on the Scene dataset. We investigate the impact of the structural loss weight and the internal balance factor on the mAP performance. The x-axis for λ is plotted on a logarithmic scale. Both parameters show a clear optimal range, demonstrating the robustness of our method.

the structural loss weight λ and the internal balance factor γ on the Scene dataset, using mAP as the primary metric. This analysis reveals the following:

- **Impact of λ (Fig. 2, left):** The performance exhibits a non-monotonic trend with a clear optimal range. Increasing λ from 0.001 to 0.01 improves mAP from 0.8403 to 0.8461, validating the effectiveness of the structured regularization. Further increases in λ lead to performance degradation, indicating that excessively strong local constraints can disrupt the global representations learned by the base contrastive loss.
- **Impact of γ (Fig. 2, right):** γ governs the trade-off between the relative and absolute losses, regulating the balance between structure sculpting and representation stability. As γ increases from 0.5 to 0.8, mAP steadily rises to 0.8459. However, a further increase to 1.5 degrades performance, suggesting that excessive anchoring may limit modeling of intra-class structure. The performance rebound at $\gamma = 2.0$ suggests the presence of a suboptimal equilibrium.

In summary, these experiments validate HBL's design philosophy: through coordinated tuning of λ and γ, HBL effectively balances global and local learning objectives, as well as sculpting and stability of representations.

V. CONCLUSIONS

This work tackles a core challenge in multi-label contrastive learning: moving from simple similarity aggregation to precise geometric manifold sculpting. We identify three key challenges—ranking noise, positive sample expulsion, and sampling distortion—that hinder fine-grained structure learning. To address these, we propose Hierarchical Boundary Learning (HBL), a novel composite structural loss designed to augment existing contrastive losses. HBL uses dual-boundary constraints to distinguish soft and hard positives and anchor hard positives, combined with adaptive semantic partitioning and reliability gating to combat sampling distortion and ensure robustness. Experiments on four datasets show that HBL consistently improves performance when integrated with leading methods. Ablations and sensitivity analyses confirm

the necessity of each component and the importance of balancing global representation and local structure.

In summary, HBL provides an effective method and validated design principles for robust geometric structure learning in multi-label contrastive learning. Future work will focus on scaling, efficiency, and handling complex noise to further enhance generalization and robustness.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (No. 62306104, 62102079), Hong Kong Scholars Program (No. XJ2024010), Research Grants Council of the Hong Kong Special Administrative Region, China (GRF Project No. CityU11212524), China Postdoctoral Science Foundation (No. 2023TQ0104), Natural Science Foundation of Jiangsu Province (No. BK20230949), Jiangsu Association for Science and Technology (No. JSTJ2024285).

REFERENCES

[1] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9726–9735.

[2] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning (ICML)*, 2020, pp. 1597–1607.

[3] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 8, pp. 1819–1837, 2014.

[4] P. Zhang and M. Wu, "Multi-label supervised contrastive learning," in *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, 2024, pp. 16786–16793.

[5] H. Li, M. Fang, X. Li, B. Chen, and G. Wang, "Hierarchical multi-granular multi-label contrastive learning," *Pattern Recognition*, vol. 164, p. 111567, 2025.

[6] S. Zhang, R. Xu, C. Xiong, and C. Ramaiah, "Use all the labels: A hierarchical multi-label contrastive learning framework," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 16639–16648.

[7] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern Recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.

[8] J. Read, B. Pfahringer, G. Holmes, and E. Frank, "Classifier chains for multi-label classification," *Machine Learning*, vol. 85, no. 3, pp. 333–359, 2011.

[9] M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern Recognition*, vol. 40, no. 7, pp. 2038–2048, 2007.

[10] J. Yuan, S. Chen, Y. Zhang, Z. Shi, X. Geng, J. Fan, and Y. Rui, "Graph attention transformer network for multi-label image classification," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 4, pp. 1–16, 2023.

[11] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5172–5181.

[12] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," in *Advances in Neural Information Processing Systems 33*, 2020, pp. 18661–18673.

[13] Y.-C. Wu, S.-H. Lyu, H. Shang, X. Wang, and C. Qian, "Confidence-aware contrastive learning for selective classification," in *Proceedings of the 41st International Conference on Machine Learning*, 2024, pp. 53706–53729.

[14] C. Duan, Y. Jiao, L. Kang, J. Z. Yang, and F. Zhou, "Deep contrastive representation learning for supervised tasks," *Pattern Recognition*, vol. 161, p. 111309, 2025.

[15] T. Gao, X. Yao, D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 6894–6910.

[16] L. Xu, H. Xie, Z. Li, F. L. Wang, W. Wang, and Q. Li, "Contrastive learning models for sentence representations," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 4, p. 67, 2023.

[17] P. Sun, H. Wu, Y. Wang, C. Liu, and Y. Liu, "Neighborhood-enhanced supervised contrastive learning for collaborative filtering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 5, pp. 2069–2081, 2023.

[18] Y. Yan, R. Li, S. Wang, F. Zhang, W. Wu, and W. Xu, "ConSERT: A contrastive framework for self-supervised sentence representation transfer," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 5065–5075.

[19] N. Chen, S.-H. Lyu, T.-S. Wu, Y. Wang, and B. Tang, "Improving multi-label contrastive learning by leveraging label distribution," *arXiv preprint* arXiv:2501.19145, 2025.

[20] R. Gupta et al., "Class prototypes based contrastive learning for classifying multi-label and fine-grained educational videos," in *Proceedings of the 41st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19923–19933.

[21] G. Tian, J. Wang, R. Wang, G. Zhao, and C. He, "A multi-label social short text classification method based on contrastive learning and improved ml-KNN," *Expert Systems*, vol. 41, no. 7, 2024,

[22] P. Hou, X. Geng, and M.-L. Zhang, "Multi-label manifold learning," in *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*, 2016, pp. 1680–1686.

[23] A. Elisseeff and J. Weston, "A kernel method for multi-labelled classification," *Advances in Neural Information Processing Systems*, vol. 14, pp. 681–687, 2001.

[24] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, 2008, pp. 39–43.

[25] M. Everingham, L. Van Gool, C. K. I. Williams, J. M. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2010.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 29th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

[28] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[29] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2019, pp. 2623–2631.