Towards Space and Semantics: Object-Purified Representation Learning for Multi-Label Image Classification

Haifeng Zhao
School of Computer Science and Technology, Anhui University
Hefei, Anhui, China
senith@163.com

Shuo Xu
School of Computer Science and Technology, Anhui University
Hefei, Anhui, China
xush1020@163.com

Leilei Ma*
School of Computer Science and Technology, Anhui University
Hefei, Anhui, China
xiaoleilei1990@gmail.com

Yufei Zhang
School of Computer Science and Technology, Anhui University
Hefei, Anhui, China
summoning@stu.ahu.edu.cn

Lei Wang
School of Computer Science and Engineering, Nanjing University of Science and Technology
Nanjing, Jiangsu, China
lei_wang@njust.edu.cn

Dengdi Sun*
School of Artificial Intelligence, Anhui University
Hefei, Anhui, China
sundengdi@163.com

Abstract

Multi-label image classification requires simultaneously recognizing multiple objects with complex interdependencies. While existing attention-based methods are prominent, their performance is hampered by two forms of representation entanglement: 1) **Spatial entanglement**, where contextual interference from backgrounds and co-occurring objects confuses specific object representations; 2) **Semantic entanglement**, where models overfit label co-occurrence priors, thereby impairing a genuine semantic understanding of the image. To address these challenges, we propose an Object-Purified Representation Learning framework. Concretely, for spatial entanglement, we propose the Spatial-wise Representation Purification Module that employs Spatial-Purified Attention to eliminate object-irrelevant feature activations for contextual interference reduction, combined with Spatial-Aware Supervision to enhance object perception capability. For semantic entanglement, we develop the Semantic-wise Association Purification Module that synergistically integrates our proposed average message with the original co-occurrence-based message. This design effectively models co-occurrence relationships while preventing their overemphasis. Furthermore, we design the Bidirectional Representation Refinement Module to efficiently enhance representations, further boosting classification performance. Extensive experiments on multiple benchmark datasets with different configurations demonstrate that our proposed method achieves state-of-the-art performance.

*Corresponding Authors: Leilei Ma and Dengdi Sun (Primary contact). All authors except Lei Wang are members of the Anhui Provincial Key Laboratory of Multimodal Cognitive Computation. Dengdi Sun is also with Jianghuai Advance Technology Center.

MM '25, October 27–31, 2025, Dublin, Ireland.

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754878

CCS Concepts

• **Computing methodologies** → **Object recognition**.

Keywords

Multi-Label Image Classification, Object-Purified Representation Learning, Representation Refinement

ACM Reference Format:

Haifeng Zhao, Shuo Xu, Leilei Ma, Yufei Zhang, Lei Wang, and Dengdi Sun. 2025. Towards Space and Semantics: Object-Purified Representation Learning for Multi-Label Image Classification. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754878

1 Introduction

Multi-label image classification (MLC) [27, 52], a fundamental task in computer vision, has garnered increasing attention due to its wide applicability in real-world scenarios, *e.g.*, image retrieval [60], image captioning [18], and medical diagnosis [45]. However, compared to the single-label setting, MLC faces greater challenges owing to complex label relationships and diverse variations among objects.

In MLC, it is crucial to model the correlations between labels. Early researchers employed sequence-based methods [1, 37, 40, 49] to capture the label dependencies. Recently, graph-based methods [4, 19, 39] have become mainstream, typically incorporating label relationships into the training process through graph structures. Additionally, the difficulty in recognizing diverse objects with varying shapes and sizes is also a significant obstacle in MLC. Leveraging effective visual attention, prevailing methods [2, 26, 29, 50, 62] focus on decoupling the original image features into specific label representations for subsequent classification. To further improve prediction accuracy, some methods [10, 57] design extra loss functions or modules to enhance the representations.

Despite the acknowledged success of the aforementioned methods, they still have two major shortcomings: ❶ For existing attention models, accurately perceiving objects within image space is a critical challenge. Particularly, the pervasive presence of co-occurring objects and background information in images inevitably

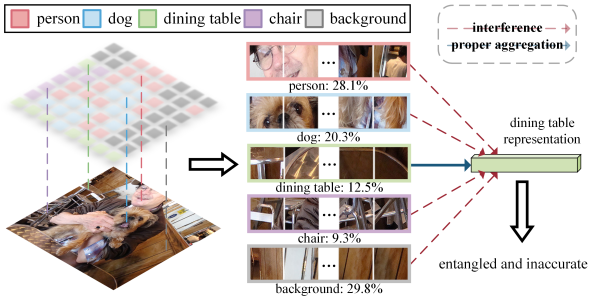


Figure 1: This case illustrates a limitation from the spatial perspective of image content, where acquiring dining table representation is prone to interference from co-occurring objects ("person", "dog", "chair") and background information.

introduces contextual interference, which can mislead the representation of specific labels. ❷ Co-occurrence information of labels establishes a valuable relational prior for MLC models, while recent studies [47, 48] have indicated that models tend to overfit the co-occurrence patterns in the training phase, which in turn impairs their semantic understanding of images.

Regarding the first drawback, Fig. 1 presents an empirical example. For smaller objects such as the dining table (occupying only about 12.5% of image layout), the majority of the image is dominated by co-occurring objects ("person", "dog", "chair", about 57.7%) and backgrounds (about 29.8%). Evidently, the aggregated attention scores assigned to these dining table-irrelevant regions are non-negligible, substantially encroaching on the model's focus on the objects of genuine interest (the "dining table"), which leads to entangled and inaccurate dining table-aware representations. To mitigate this issue, we propose the Spatial-wise Representation Purification Module (SRPM) from the spatial perspective of image content, aiming to systematically eliminate object-irrelevant feature activations during the representation learning phase. Specifically, we propose Spatial-Purified Attention (SPA) that employs saliency-driven spatial masking. The SPA suppresses contextual interference by filtering out non-salient regions in attention maps, thereby reducing the impact of co-occurring objects and background. Building on this, we develop Spatial-Aware Supervision (SAS) to guide attention maps to focus on salient regions while allowing the SPA to retain more beneficial features, thereby accurately perceiving objects in complex multi-label scenarios.

For the second challenge, graphs are widely used to model semantic co-occurrence relationships among labels [11, 50, 56], where the features of each node (label) interact via graph embedding techniques such as Graph Convolutional Networks (GCNs) [19] on the graph structure. However, excessive emphasis on co-occurrence relationships can significantly amplify the influence of each node by its co-occurring nodes, blurring the distinctiveness of label semantic features, which in turn leads to a decline in image understanding performance. Given this, we design a simple Semantic-wise Association Purification Module (SAPM) to appropriately utilize co-occurrence relationships while preventing the model from overemphasizing them. Specifically, we substitute partial layers of the original GCNs with our Average Message Passing (AMP) layers, where all nodes are interconnected and each node updates its state by considering the averaged influence from all other nodes. In contrast to the original layers that focus on learning co-occurrence relationships, AMP prioritizes node-specific feature learning while incorporating weak yet comprehensive influences from all other nodes. By alternately propagating average and co-occurrence-based messages, SAPM prevents overemphasis on co-occurrence patterns, purifying representations from a semantic association perspective. Notably, this approach adds neither complexity nor computational cost to the original GCNs.

The above SRPM and SAPM purify complex image objects from spatial and semantic perspectives, respectively, and synergistically constitute the proposed Object-Purified Representation Learning (OPRL) framework. In addition, to make representations more conducive to classification, we design a Bidirectional Representation Refinement Module (BRRM) based on multilayer perceptrons (MLPs). Conventional MLPs perform unidirectional modeling along the representation dimension with category-specific weights and lack cross-dimensional interaction [35], failing to capture shared patterns across categorical boundaries, such as feature commonality (*e.g.*, attributes like texture or shape shared across labels) and implicit hierarchy (*e.g.*, cars and trucks as "vehicles"). Considering the characteristics of the MLC tasks, our BRRM simultaneously facilitates learning in both the representation and category dimensions, thereby building a better multi-label feature space.

In summary, extensive experiments on diverse benchmarks verify the effectiveness of the proposed OPRL framework, and the main contributions of this work are summarized as follows:

- We propose a Spatial-wise Representation Purification Module, where Spatial-Purified Attention filters out object-irrelevant feature activations to reduce contextual interference, and Spatial-Aware Supervision is designed to enhance object perception capability.
- We introduce the Semantic-wise Association Purification Module, where our average message collaborates with the co-occurrence message to avoid overemphasizing co-occurrence relationships, purifying representations from a semantic association perspective.
- We propose a Bidirectional Representation Refinement Module to efficiently refine representations in both categorical and representational dimensions, further enhancing classification performance.

2 Related Work

Modeling Label Correlations. Real-world images often involve multi-object scenarios, driving researchers to model label dependencies. Early methods like Gong *et al.* [17] and Li *et al.* [22] employed pairwise ranking and tree-structured graphs to augment the original labels, but suffered from inefficiency. To address this, RNN-based methods [37, 40, 49] mapped labels and images into shared embeddings with predefined prediction strategies. However, these methods only captured local label interactions, neglecting global dependencies. This limitation spurred GCN-based approaches: ML-GCN [5] developed label-aware classifiers; ADDGCN [50] introduced dynamic frameworks for content-aware representations; AdaHGNN [41] utilized adaptive hypergraphs to capture high-order semantics, eliminating manual graph construction. Although these GCN-based methods have achieved good results, they are prone to overfitting correlations, which is detrimental to MLC [47, 48].

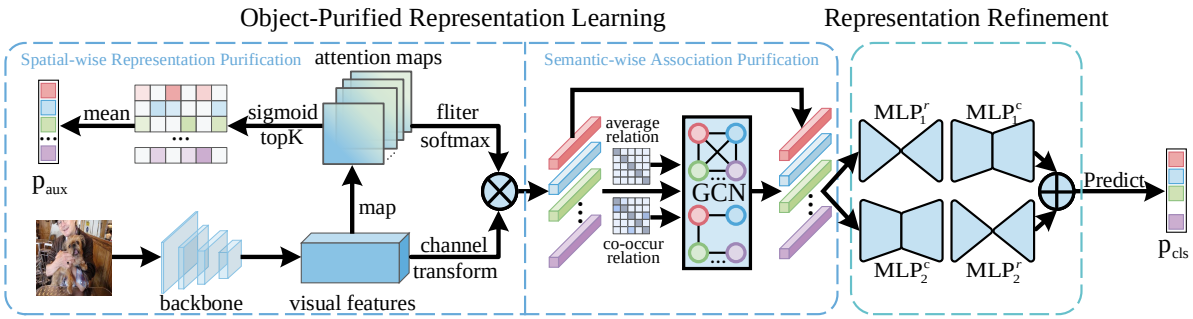


Figure 2: The overview of our framework. The Object-Purified Representation Learning comprises two key components: the Spatial-wise Representation Purification Module and the Semantic-wise Association Purification Module, which purify representations from spatial and semantic perspectives, respectively. Subsequently, the Bidirectional Representation Refinement Module is employed to further enhance these representations. $\boldsymbol{p}_{\text{cls}}$ and $\boldsymbol{p}_{\text{aux}}$ are jointly utilized to supervise the model.

Modeling Spatial Dependencies. In multi-label images with complex scenes and semantically diverse objects, modeling spatial relationships between regions is critical. Wang *et al.* [40] proposed a region-proposal-free framework using spatial transformers and LSTMs to localize objects and predict labels. GM-MLIC [44] decomposed images via pretrained detectors, matching instances to labels via spatial-semantic graphs, later extended by ML-SGM [42] with semantic-guided activation for efficient instance extraction. Zhang *et al.* [51] proposed a novel spatial context-aware mechanism that leverages an adaptive patch expansion strategy to effectively capture objects and their surrounding spatial relationships. Although effective, these methods heavily rely on bounding-box generation and intricate architectures, limiting deployment flexibility.

Attention Models. Attention mechanism and transformer [36] have gained significant traction in computer vision [34], prompting their adaptation to MLC tasks. Early approaches like ResNet-SRN [58] proposed a spatial regularization network to generate attention maps for all labels. Subsequent works improved parameter efficiency: SSGRL [2] and SALGL [63] adopted the linear attention method with fewer parameters to learn specific label representations, and TRM-ML [30] introduced text-region matching for attention generation. Transformer architectures have proven particularly effective for MLC due to their capacity to model long-range dependencies. For example, Query2Label [26], LAGC [46], SADCL [29], TSFormer [61], and ML-Decoder [32] employ label embeddings as queries to identify and pool category-specific features from visual feature maps, subsequently feeding these features into binary classifiers for accurate classification. Although these methods have achieved success, they still face limitations in accurately localizing objects and typically overlook the interference caused by co-occurring objects in representation learning.

3 Preliminaries

Motivations. Recent studies have leveraged attention models to extract label-aware representations for training binary classifiers. Contextual content and co-occurring objects inevitably introduce interference to the specific label representations from both spatial and semantic association perspectives, resulting in suboptimal model performance. This work aims to mitigate these limitations to obtain higher-quality label-aware representations, thereby improving performance for MLC tasks.

Notations. In our work, we adopt the standard setting of MLC. Let $\boldsymbol{i} \in \mathcal{I}$ denote an instance, and $\boldsymbol{y} \in \mathcal{Y}$ is its label, $\mathcal{Y} = \{0, 1\}^C$. C represents the total number of label categories. Considering a specific instance $\boldsymbol{i}$, $y^c = 1$ indicates that the c-th label is associated with the instance $\boldsymbol{i}$ and vice versa. Our goal is to train a model $f(\cdot) : \mathcal{I} \rightarrow \mathcal{Y}$ to predict the presence of each label and we denote $f_c(\boldsymbol{i})$ as the predicted probability of the c-th label for the instance $\boldsymbol{i}$.

Graph Convolutional Network. Graph Convolutional Network [19] is tailored to handle graph-structured data, which can be briefly denoted as: $\widehat{\mathbf{V}} = \delta(\boldsymbol{AVW})$, where δ, $\mathbf{A}$ and $\mathbf{W}$ represent activation function, correlation matrix and linear layer respectively, $\mathbf{V}$ and $\widehat{\mathbf{V}}$ are the input nodes and updated nodes.

Co-occurrence Relation Modeling. Label co-occurrence is a statistical concept that builds the connection between labels in the training set based on conditional probability. Through this data-driven approach, prior knowledge is provided for MLC tasks. It is generally recognized that higher co-occurrence represents a more pronounced correlation among labels. Specifically, the co-occurrence of label pairs is computed to yield matrix $\boldsymbol{M} \in \mathbb{R}^{C \times C}$, and the co-occurrence matrix can be denoted as: $\boldsymbol{A}_i = \boldsymbol{M}_i / \mathrm{N}_i$, where N_i is the occurrence times of label i and $\boldsymbol{A}_i$ denotes the i-th row of the co-occurrence matrix. $\boldsymbol{A}_{i,j}$ indicates the probability that the j-th label appears when the i-th label is present.

4 The Proposed Method

4.1 Overview

In this section, we first provide an overview of the proposed OPRL framework, as illustrated in Fig. 2. First, SRPM extracts visual features of the given image and eliminates feature activations in object-irrelevant regions, purifying label-aware representations from a spatial perspective. Then, SAPM is proposed to model co-occurrence relationships while mitigating over-reliance on them, purifying representations from a semantic association perspective. After that, the representations are fed into BRRM to achieve effective enhancement, rendering them more conducive to classification. Finally, the classification prediction $\boldsymbol{p}_{\text{cls}}$ and the auxiliary scores $\boldsymbol{p}_{\text{aux}}$ are combined to jointly supervise model training.

4.2 Visual Feature Extraction

Given an image $\boldsymbol{i} \in \mathbb{R}^{h\times w\times 3}$, we employ a backbone network (*e.g.*, ResNet [16], ViT [13]) to extract its visual features, as follows:

$$\boldsymbol{F} = \text{Encoder}_v(\boldsymbol{i}) , \tag{1}$$

where $\boldsymbol{i}$ indicates the input image and $\boldsymbol{F} \in \mathbb{R}^{D\times H\times W}$ is the extracted visual features, also called the feature maps.

Due to the constraints of the receptive field size inherent in convolutional operations, convolutional architectures struggle to model global and long-range dependencies, leading to entangled features [8, 13]. Given this, we employ a transformer encoder to capture more comprehensive visual information, denoted as:

$$\overline{\boldsymbol{F}} = \text{Encoder}_t(\boldsymbol{F}) , \tag{2}$$

where Encoder_t denotes a simple transformer block and $\overline{\boldsymbol{F}}$ represents improved visual features.

4.3 Object-Purified Representation Learning

4.3.1 ***Spatial-wise Representation Purification.*** In order to suppress contextual interference from co-occurring objects and backgrounds, we propose the Spatial-wise Representation Purification Module, which primarily consists of Spatial-Purified Attention and Spatial-Aware Supervision.

In comparison to conventional multi-label representation learning, our SPA filters out feature activations in non-salient spatial regions, aiming to mitigate contextual interference from object-irrelevant regions to label-aware representations. Furthermore, our SAS is designed to enhance object perception capacity in complex multi-label scenarios, ensuring more focused attention on relevant objects and mitigating the risk of filtering out beneficial features. The above process can be written as follows:

$$\boldsymbol{S} = \text{Map}(\overline{\boldsymbol{F}}) , \tag{3}$$

where Map denotes the attention generation and is implemented through a convolution operation. $\boldsymbol{S} \in \mathbb{R}^{C\times H\times W}$ represents the generated attention saliency map and $\boldsymbol{C}$ refers to the number of categories. More concretely, $\boldsymbol{S}_c = \{\boldsymbol{s}^c_{i,j} | i \in [1, H], j \in [1, W]\}$ is the map corresponding to category $\boldsymbol{c}$ and $\boldsymbol{s}^c_{i,j}$ denotes the saliency value at spatial location (i, j). Then we employ the SAS as follows:

$$\boldsymbol{\alpha} = \text{Sigmoid}(\boldsymbol{S}) , \quad \boldsymbol{p}_{\text{aux}} = \text{Mean}(\text{TopK}(\boldsymbol{\alpha})) , \tag{4}$$

where $\boldsymbol{\alpha}$ denotes the activated attention map, TopK indicates the selection of the K largest values across spatial locations, and Mean signifies the process of averaging these values. $\boldsymbol{p}_{\text{aux}}$ is an obtained probability value that will subsequently serve as a supervision for object perception. Specifically, if $\boldsymbol{\alpha}^c$ at (i, j) exhibits a larger logistic value, it indicates a higher probability that position (i, j) contains features relevant to category c.

The visual information contained in $\overline{\boldsymbol{F}}$ has already been partially supervised by $\boldsymbol{p}_{\text{aux}}$, in order to provide richer optimization information for subsequent specific label representations, we employ a channel transformation on $\overline{\boldsymbol{F}}$, denoted as:

$$\hat{\boldsymbol{F}} = \text{Ctransform}(\overline{\boldsymbol{F}}) , \tag{5}$$

where Ctransform is the channel transformation through a convolution operation, and $\hat{\boldsymbol{F}}$ is the enriched visual features. The non-salient

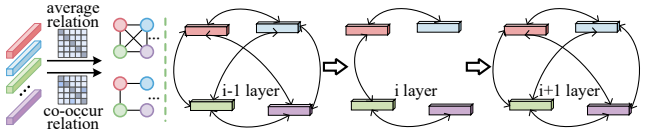


Figure 3: The schematic illustration of our Semantic-wise Association Purification Module (SAPM).

regions of the attention maps are filtered out as follows:

$$\boldsymbol{s}^c_{i,j} = \begin{cases} \boldsymbol{s}^c_{i,j} , & \text{if } \boldsymbol{s}^c_{i,j} \geq \boldsymbol{z}_{\boldsymbol{\tau}} \\ -\infty, & \text{if } \boldsymbol{s}^c_{i,j} < \boldsymbol{z}_{\boldsymbol{\tau}} \end{cases} , \tag{6}$$

where $\boldsymbol{z}_{\boldsymbol{\tau}}$ is a quantile that functions as a threshold, determining the filter ratio assigned to the attention maps. For example, if $\boldsymbol{z}_{\boldsymbol{\tau}}$ is set as the median $\boldsymbol{z}_{50\%}$, then the 50% non-salient regions in the maps will be excluded. More concretely, when $\boldsymbol{\tau} \to 1$, an increasing proportion of non-salient regions will be discarded, while $\boldsymbol{\tau} \to 0$, the situation is reversed.

The attention values of non-salient regions are converted to $-\infty$, which will be ignored in the subsequent softmax processing. Then the label-aware representations can be computed as follows:

$$\hat{\boldsymbol{s}}^c_{i,j} = \frac{\exp(\boldsymbol{s}^c_{i,j})}{\sum_{w,h} \exp(\boldsymbol{s}^c_{w,h})}, \quad \boldsymbol{x}_c = \sum\nolimits_{w,h} \hat{\boldsymbol{s}}^c_{w,h} \cdot \hat{\boldsymbol{F}}_{w,h} , \tag{7}$$

where the upper equation denotes a softmax function employed to normalize the attention coefficients, and the lower equation describes a weighted average pooling operation on the visual features $\hat{\boldsymbol{F}}$ across all positions, guided by the coefficients. Notably, positions excluded in (6) will not be considered in this calculation. We apply the process to all labels, and then obtain all label-aware representation vectors $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_C\}, \boldsymbol{x}_i \in \mathbb{R}^D, i \in [1, C]$.

4.3.2 ***Semantic-wise Association Purification.*** In order to mitigate semantic representation entanglement caused by overfitting co-occurrence relationships, we propose the Semantic Association Purification Module (SAPM), which models co-occurrence relationships while preventing excessive emphasis on them.

Similar to existing methods [11, 50, 56], we leverage Graph Convolutional Networks (GCNs) to model the co-occurrence relationships among labels, with the basic model described as follows:

$$\hat{\boldsymbol{X}} = \boldsymbol{X} + \mathcal{G}(\boldsymbol{X}) \ \leftarrow \boldsymbol{A}^{C\times C} , \tag{8}$$

where $\mathcal{G}$ denotes GCN layers and $\boldsymbol{A}^{C\times C}$ represents the co-occurrence matrix, which is used as a residual connection. In particular, our residual connection essentially implements feature fusion: $\boldsymbol{X}$ corresponds to the vanilla label-aware representations, $\mathcal{G}(\boldsymbol{X})$ encodes its co-occurrence dependencies, and $\hat{\boldsymbol{X}}$ combines both parts.

Innovatively, our SAPM introduces a straightforward approach to prevent GCNs from excessively emphasizing co-occurrence relationships, which purifies representations from a semantic association perspective. Specifically, we propose Average Message Passing (AMP) strategy that collaborates with co-occurrence information to achieve this. Our AMP is to partially replace the co-occurrence

matrix $\boldsymbol{A}^{C\times C}$ with an average matrix $\boldsymbol{K}^{C\times C}$, denoted as:

$$\begin{pmatrix} \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{pmatrix}_{C\times C} \leftarrow \begin{pmatrix} 1 & k & \cdot & k \\ k & 1 & & \cdot \\ \cdot & & \cdot & \cdot \\ k & \cdot & \cdot & 1 \end{pmatrix}_{C\times C}, \tag{9}$$

where the left and right matrices represent $\boldsymbol{A}^{C\times C}$ and $\boldsymbol{K}^{C\times C}$ respectively, and $k = \frac{1}{C-1}$. In this way, each node is equally connected to all other nodes and each receives uniform and equivalent influence from the remaining nodes. Additionally, the values along the main diagonal of the average matrix $\boldsymbol{K}^{C\times C}$ are all set to 1, ensuring the stability of each node's self-feature. Combining the AMP and co-occurrence information, our GCNs propagation is detailed as:

$$\hat{\boldsymbol{X}}_{i+1} = \begin{cases} \delta(\boldsymbol{A}\hat{\boldsymbol{X}}_i\boldsymbol{W}_i), & \text{if } i = 0 \pmod 2 \\ \delta(\boldsymbol{K}\hat{\boldsymbol{X}}_i\boldsymbol{W}_i), & \text{if } i = 1 \pmod 2 \end{cases}, \tag{10}$$

where $\boldsymbol{A}$, $\boldsymbol{K}$, and $\boldsymbol{W}$ denote the predefined co-occurrence matrix, the average matrix of AMP, and learnable linear transformations, respectively, and δ represents an activation function. $\hat{\boldsymbol{X}}_i$ and $\hat{\boldsymbol{X}}_{i+1}$ are the input and output of the i-th GCNs layer.

More concretely, in the co-occurrence-based message layer, GCNs focus on modeling co-occurrence relationships. In the average message layer, GCNs emphasize learning object-self features while incorporating subtle yet comprehensive influences from other objects. Notably, this approach does not complicate the original GCNs or increase computational overhead, effectively preventing GCNs from overly favoring co-occurrence patterns.

4.4 Bidirectional Representation Refinement

To enhance label-aware representations and make them more conducive to classification, we propose a Bidirectional Representation Refinement Module based on MLPs. In contrast to traditional MLPs that utilize category-specific weights for single-directional feature transformation, while neglecting shared patterns across categories, our BRRM simultaneously facilitates learning in both the representation dimension (r-dim) and category dimension (c-dim).

To prevent excessive bias towards a single direction in feature transformation and mitigate the sensitivity to feature processing order, the BRRM is designed as a dual-path architecture to ensure more comprehensive information integration. Specifically, we propagate the label-aware representations through two distinct pathways. In the first pathway, we initially enhance the representation abstraction capability across the r-dim, followed by capturing shared patterns across the c-dim. Conversely, the second pathway reverses the order, performing learning across the c-dim first, and then across the r-dim. The above process is detailed as follows:

$$\hat{\boldsymbol{X}}_1 = \mathrm{LN}(\hat{\boldsymbol{X}}), \quad \hat{\boldsymbol{X}}_2 = \mathrm{LN}(\hat{\boldsymbol{X}})^{\mathrm{T}}, \tag{11}$$

where $\hat{\boldsymbol{X}}$ is the label-aware representations derived from the preceding steps. LN and T denote the layer normalization and tensor transpose operation respectively. Subsequently, $\hat{\boldsymbol{X}}_1$ and $\hat{\boldsymbol{X}}_2$ are processed through the below two pathways as follows:

$$\begin{aligned} &\hat{\boldsymbol{X}}_1' = (\mathrm{LN}(\mathrm{MLP}_1^{\boldsymbol{r}}(\hat{\boldsymbol{X}}_1) + \hat{\boldsymbol{X}}_1))^{\mathrm{T}}, \hat{\boldsymbol{X}}_1'' = (\mathrm{MLP}_1^{\boldsymbol{c}}(\hat{\boldsymbol{X}}_1'))^{\mathrm{T}} + \hat{\boldsymbol{X}}_1', \\ &\hat{\boldsymbol{X}}_2' = \mathrm{LN}((\mathrm{MLP}_2^{\boldsymbol{c}}(\hat{\boldsymbol{X}}_2))^{\mathrm{T}} + \hat{\boldsymbol{X}}_2^{\mathrm{T}}), \hat{\boldsymbol{X}}_2'' = \mathrm{MLP}_2^{\mathrm{r}}(\hat{\boldsymbol{X}}_2') + \hat{\boldsymbol{X}}_2', \end{aligned} \tag{12}$$

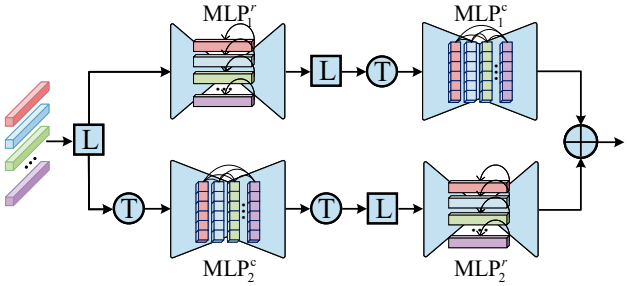


Figure 4: The schematic illustration of our Bidirectional Representation Refinement Module (BRRM).

where $\mathrm{MLP}_1^{\boldsymbol{r}}$, $\mathrm{MLP}_2^{\boldsymbol{r}}$ represent the MLPs executed in the r-dim, while $\mathrm{MLP}_1^{\boldsymbol{c}}$, $\mathrm{MLP}_2^{\boldsymbol{c}}$ are implemented in the c-dim. It is worth noting that the subscripts 1 and 2 denote the first and second pathways, respectively. In addition, LN, +, and T denote layer normalization, residual connection, and tensor transposition, respectively. Finally, we combine the outputs from the two pathways as follows:

$$\bar{\boldsymbol{X}} = \tfrac{1}{2} \cdot (\hat{\boldsymbol{X}}_1'' + \hat{\boldsymbol{X}}_2''), \tag{13}$$

where $\hat{\boldsymbol{X}}_1''$ and $\hat{\boldsymbol{X}}_2''$ denote the outputs of the two paths respectively.

4.5 Joint Learning and Optimization

In this work, we adopt ASL loss function [31], which offers advantages in addressing imbalance issues. Specifically, the main loss (i.e., classification loss) is calculated as follows:

$$\mathcal{L}_{\mathrm{cls}} = \mathrm{ASL}(\boldsymbol{p}_{\mathrm{cls}}, y), \quad \boldsymbol{p}_{\mathrm{cls}} = \mathrm{Sigmoid}(\mathbf{W}_c^{\mathrm{T}}\bar{\boldsymbol{X}}_c + b_c), \tag{14}$$

where $\mathbf{W}_c^{\mathrm{T}}$ and b_c are learnable parameters. $\boldsymbol{p}_{\mathrm{cls}}$ and y denote the predicted probability and true label, respectively. Furthermore, we use the $\boldsymbol{p}_{\mathrm{aux}}$ from Eq. (4) to obtain an auxiliary loss, denoted as:

$$\mathcal{L}_{\mathrm{aux}} = \mathrm{ASL}(\boldsymbol{p}_{\mathrm{aux}}, y). \tag{15}$$

Finally, the overall optimization objective is defined as a joint loss, obtained by taking a weighted sum of $\mathcal{L}_{\mathrm{cls}}$ and $\mathcal{L}_{\mathrm{aux}}$ as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{cls}} + \lambda\mathcal{L}_{\mathrm{aux}}. \tag{16}$$

where λ is a hyper-parameter that determines the weight of the auxiliary loss $\mathcal{L}_{\mathrm{aux}}$. The joint loss $\mathcal{L}$ is used to train our OPRL framework. In general, this joint learning and optimization mechanism benefits from two aspects: 1) Supervised by the auxiliary loss, the attention map $\boldsymbol{S}$ is better equipped to perceive object regions. 2) As an additional classifier, the auxiliary loss $\mathcal{L}_{\mathrm{aux}}$ alleviates certain limitations of the primary classifier [47].

5 Experiments

5.1 Experimental Settings

Dataset To validate the effectiveness of the proposed OPRL framework, we evaluate it on four prevalent MLC benchmarks: (1) PASCAL VOC 2007 (VOC07) [14]; (2) Microsoft COCO 2014 (COCO) [24]; (3) NUS-WIDE (NUS) [7]; (4) Visual Genome (VG) [20]. For VG dataset, where most categories contain only a limited number of samples, we follow prior works [2, 26] and select the 500 most

Table 1: The performance comparison of our method with state-of-the-art models on VOC07. All metrics are in %. Note that the symbol † indicates that the corresponding method utilizes a higher resolution (576×576).

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
SSGRL† [2]	99.5	97.1	97.6	97.8	82.6	94.8	96.7	98.1	78.0	97.0	85.6	97.8	98.3	96.4	98.8	84.9	96.5	79.8	98.4	92.8	93.4
ML-GCN [5]	99.5	98.5	98.6	98.1	80.8	94.6	97.2	98.2	82.3	95.7	86.4	98.2	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
ASL [31]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	94.4
ISLC [55]	99.8	97.6	98.4	98.3	79.2	95.2	97.5	98.5	81.6	95.5	**88.1**	98.7	98.6	96.7	98.7	84.9	96.9	85.1	99.1	93.5	94.1
SST [3]	99.8	98.6	**98.9**	**98.4**	**85.5**	94.7	**97.9**	**98.6**	83.0	96.8	85.7	98.8	98.9	95.7	99.1	85.4	96.2	84.3	99.1	95.0	94.5
DATran [57]	**99.9**	**98.7**	98.6	**98.4**	82.6	96.0	97.7	**98.6**	**85.0**	96.2	84.7	98.5	98.2	96.9	98.8	85.0	97.9	**86.7**	99.2	**95.2**	94.6
OPRL-R101	**99.9**	98.5	98.4	**98.4**	83.9	**97.3**	97.7	**98.6**	80.7	**98.1**	86.9	**99.1**	**99.1**	**98.2**	**99.3**	**87.0**	**99.6**	**86.7**	**99.7**	93.8	**95.1**
SSGRL(pre)† [2]	99.5	97.1	97.6	97.8	82.6	94.8	96.7	98.1	78.0	97.0	85.6	97.8	98.3	96.4	98.8	84.9	96.5	79.8	98.4	92.8	93.4
ASL(pre) [31]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	95.3
ISLC(pre) [55]	**100.0**	98.8	98.8	**98.6**	86.2	97.0	**98.8**	98.9	**86.5**	96.9	89.1	99.0	98.5	97.6	**99.3**	88.1	98.4	**90.0**	99.2	**97.0**	95.8
ADD-GCN(pre)†	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	96.0
OPRL-R101(pre)	**100.0**	**99.5**	**99.3**	**98.6**	**88.4**	**98.6**	98.6	**99.2**	85.2	**98.6**	**91.2**	**99.3**	**99.3**	**99.4**	**99.3**	**90.5**	**99.6**	88.7	**99.7**	95.1	**96.4**

Table 2: The performance comparison of our method with state-of-the-art models on COCO. The backbones noted with 22k are pretrained on the ImageNet 22k dataset.

Methods	Backbone	Res.	mAP	CP	CR	CF1	OP	OR	OF1
ML-GCN [5]	ResNet101	448^2	83.0	85.1	72.0	78.0	85.8	75.4	80.3
MCAR [15]	ResNet101	448^2	83.8	85.0	72.1	78.0	88.0	73.9	80.3
CSRA [59]	ResNet101	448^2	83.5	84.1	72.5	77.9	85.6	75.7	80.3
TDRG [53]	ResNet101	448^2	84.6	86.0	73.1	79.0	86.6	76.4	81.2
Q2L-R101 [26]	ResNet101	448^2	84.9	84.8	74.5	79.3	86.6	76.9	81.5
DATran [57]	ResNet101	448^2	84.9	84.8	74.9	79.6	86.0	77.6	81.6
IDA [25]	ResNet101	448^2	84.8	-	-	78.7	-	-	80.9
C-TMS [43]	ResNet101	448^2	85.1	87.2	74.2	79.1	88.7	76.5	81.4
SpliceMix-CL [38]	ResNet101	448^2	84.9	87.4	73.2	79.7	88.2	76.3	81.8
OPRL-R101	ResNet101	448^2	**85.7**	85.9	75.3	**79.8**	87.1	77.9	**82.3**
SSGRL [2]	ResNet101	576^2	83.8	91.9	62.5	72.7	93.8	64.1	76.2
TDRG [53]	ResNet101	576^2	86.0	87.0	74.7	80.4	87.5	87.9	82.4
Q2L-R101 [26]	ResNet101	576^2	86.5	85.8	76.7	81.0	87.0	78.9	82.8
DATran [57]	ResNet101	576^2	86.3	86.6	76.1	81.0	87.2	78.9	82.8
C-TMS [43]	ResNet101	576^2	86.2	88.2	75.3	80.4	89.6	77.2	82.5
IDA [25]	ResNet101	576^2	86.3	-	-	80.4	-	-	82.5
OPRL-R101	ResNet101	576^2	**87.0**	86.8	77.0	**81.3**	88.1	79.2	**83.4**
ViT-L16 [13]	ViT-L16	448^2	80.4	83.8	67.0	74.5	86.6	72.0	78.6
CSRA [59]	ViT-L16	448^2	86.9	89.1	74.2	81.0	89.6	77.1	82.9
OPRL-ViT-L16	ViT-L16	448^2	**88.1**	89.6	76.6	**81.7**	89.3	80.0	**84.4**
ViT-L16 [13]	ViT-L16(22k)	448^2	89.6	76.7	87.9	81.9	75.9	89.8	82.3
OPRL-ViT-L16	ViT-L16(22k)	448^2	**90.3**	92.5	78.5	**84.3**	92.7	80.2	**86.0**

frequent categories to construct a new dataset, referred to as VG500. Comprehensive statistics for all datasets can be found in Appendix.

Implementation Details. Following previous MLC works [26, 29, 31], we adopt a similar experimental setup. We apply RandAugment [9] and Cutout [12] for data augmentation, while using an exponential moving average (EMA) with a decay factor of 0.9997 to smooth the model parameters. Our model is trained for 40 epochs, and the batch size is set to 64. The training process employs AdamW [28] optimizer and one-cycle learning rate schedule [33], with a maximum learning rate of 5e-5. Some hyper-parameter settings: the number of supervised locations K in Eq. (4) is set to 3, the filter ratio τ in Eq. (6) is set to 0.5, the GCN layers in Eq. (8) is set to 3 , the weight of auxiliary loss λ in Eq.(16) is set to 1.

Evaluation Metric In this work, we evaluate the performance of our proposed OPRL framework using the mean average precision (mAP), which is a widely adopted metric in MLC tasks. Additionally, following prior works [2, 31, 59], we adopt overall precision, recall, F1-measure (OP, OR, OF1) as well as per-class precision, recall, F1-measure (CP, CR, CF1) to conduct a comprehensive comparative analysis. It is worth noting that mAP, OF1, and CF1 are often considered key performance indicators.

5.2 Compared to State-of-the-Art Results

Comparisons on VOC07 Dataset. The comparison results on VOC07 are reported in Table 1, including the average precision (AP) for each category and mAP over all categories. It is evident that our OPRL framework attains the highest AP on a substantial proportion of label categories. The upper section presents the results of these models based on ResNet101, which is pre-trained on the ImageNet-1k dataset. It is worth noting that OPRL achieves 95.1% in mAP, surpassing the suboptimal method, DATran [57], by 0.5%. The lower section demonstrates the performance of these models when pre-trained on the COCO dataset and subsequently fine-tuned on the VOC07 dataset. Notably, the proposed OPRL framework exhibits a significant improvement, with the mAP increasing from 95.1% to 96.4%. Although SSGRL [2] and ADDGCN [50] employ a higher image resolution of 576×576, our OPRL framework still achieves a superior mAP. In summary, regardless of whether the backbone network is pre-trained on ImageNet-1k or COCO datasets, our OPRL framework consistently achieves outstanding performance compared to existing methods.

Comparisons on COCO Dataset. The comparison results on COCO are presented in Table 2. It can be observed that our OPRL framework outperforms current state-of-the-art methods on most metrics. Specifically, leveraging the ResNet-101 pre-trained on the ImageNet 1K dataset as the backbone, our OPRL framework achieves 85.7% in mAP, outperforming the suboptimal method, C-TMS [42], by a margin of 0.6%. With a higher image resolution of 576×576, our OPRL framework consistently shows superior performance, highlighting its scalability. In addition to employing the convolution-based ResNet-101 as the backbone, we also evaluate the performance of our OPRL framework on transformer-based architectures. Concretely, with ViT-L16 [13] as the backbone network,

Table 3: The performance comparison of our method with current state-of-the-art models on NUS.

Methods	mAP	CP	CR	CF1	OP	OR	OF1
SRN [58]	62.0	65.2	55.8	58.5	75.5	71.5	73.4
ASL [31]	63.9	-	-	62.7	-	-	74.6
P-GCN [6]	62.8	64.4	56.8	60.4	75.7	71.2	73.4
Q2L-R101 [26]	65.0	-	-	63.1	-	-	75.0
SST [3]	63.5	67.2	53.5	59.6	77.4	69.4	73.2
VSGCN [11]	63.4	64.2	58.9	61.5	73.6	74.7	74.1
Mulcon [10]	63.9	-	-	61.8	-	-	74.8
C-TMS [43]	62.8	-	-	61.4	-	-	74.6
OPRL-R101	**66.0**	62.1	66.1	**63.3**	72.2	78.1	**75.1**

Table 4: The performance comparison of our method with current state-of-the-art models on VG500.

Methods	mAP	CP	CR	CF1	OP	OR	OF1
ResNet101 [16]	30.9	39.1	25.6	31.0	61.4	35.9	45.4
ML-GCN [5]	32.6	42.8	20.2	27.5	66.9	31.5	42.8
SSGRL [2]	36.6	-	-	-	-	-	-
Q2L-R101 [26]	39.5	-	-	-	-	-	-
ISLC [55]	38.8	49.6	25.6	33.8	69.8	36.6	48.0
KGGR [23]	37.4	47.4	24.7	32.5	66.9	36.5	47.2
C-Tran [21]	38.4	49.8	27.2	35.2	66.9	39.2	49.5
DRGN [54]	39.8	50.6	29.4	37.2	67.1	40.4	50.4
OPRL-R101	**40.6**	51.1	39.6	**44.6**	56.9	53.6	**55.2**

the proposed OPRL framework consistently achieves state-of-the-art performance in multi-label image classification, irrespective of whether it is pre-trained on ImageNet-1K or ImageNet-22K. Overall, across nearly all key metrics (mAP, CF1, OF1), the proposed OPRL framework achieves the best performance when implemented on both convolution-based and transformer-based networks, demonstrating its superiority in multi-label image classification.

Comparisons on NUS Dataset. The comparison results on NUS are summarized in Table 3. As shown, the proposed OPRL framework achieves superior performance compared to existing methods, attaining the highest mAP of 66.0%. This marks a significant improvement over other state-of-the-art methods, such as Q2L-R101(65.0%) [26] and ASL(63.9%) [31]. Moreover, our method achieves the best performance in CF1(63.3%) and OF1(75.1%), further showcasing its effectiveness. Overall, our proposed OPRL framework outperforms existing approaches with notable advancements on the NUS dataset.

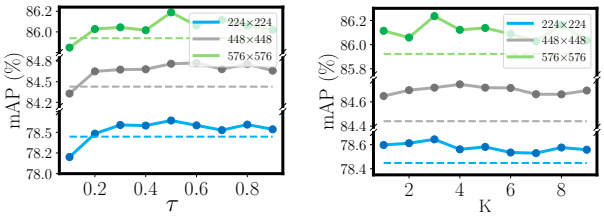
Comparisons on VG500 Dataset. The comparison results on VG500 are presented in Table 4. As demonstrated, the proposed OPRL framework achieves the highest mAP of 40.6 %, outperforming all other state-of-the-art methods, such as Q2L-R101 [26]. Notably, the proposed OPRL framework achieves the highest CF1 (44.6%) and OF1 (55.2%), further demonstrating its effectiveness. In summary, the proposed OPRL framework demonstrates significant improvements over existing methods on the VG500 dataset.

5.3 Diagnostic Experiments

Effect of key components in the OPRL framework. To evaluate the impact of each component in our model, we conducted an ablation study using the COCO and NUS datasets, as shown in Table 5. Starting with the baseline ResNet101 model, which achieves an mAP of 82.9% on COCO and 63.7% on NUS, we observe performance

Table 5: Ablation study of component contributions to our model on COCO and NUS datasets.

Method					COCO			NUS		
ResNet101	T	SRPM	SAPM	BRRM	mAP	CF1	OF1	mAP	CF1	OF1
✓					82.9	77.1	80.9	63.7	60.1	74.2
✓		✓			84.7	78.7	81.6	64.5	61.5	74.8
✓		✓	✓		85.2	79.3	82.0	65.2	62.7	74.9
✓		✓		✓	85.2	79.5	82.0	65.3	62.7	75.0
✓		✓	✓	✓	85.5	79.5	82.1	65.5	62.7	**75.1**
✓	✓	✓	✓	✓	**85.7**	**79.8**	**82.3**	**66.0**	**63.3**	**75.1**



(a) filter ratio (b) supervised locations

Figure 5: The analysis of filter ratio τ and the number of supervised locations K in SRPM. Note that the horizontal dashed lines represent the results directly obtained from the original attention map, *i.e.*, without applying our SRPM.

improvements with the addition of various modules. Incorporating SRPM increases the mAP to 84.7% on COCO and 64.5% on NUS, while adding SAPM further enhances it to 85.2% and 65.2%. The inclusion of BRRM, both independently and alongside SAPM, consistently boosts performance. The full model culminates in a mAP of 85.7% on COCO and 66.0% on NUS. Notably, T in the Table 5 denotes a standard transformer encoder, designed to mitigate the limitations of convolutional architectures. Although not a primary contribution, it still provides additional performance gains. These results confirm the effectiveness of SRPM, SAPM, and BRRM, demonstrating their complementary roles in improving model performance.

Effect of filter ratio τ in SRPM. To explore the impact of the filter ratio τ in Eq. (6) on our SRPM, we conducted extensive experiments, as shown in Fig. 5a. It can be observed that when the filter ratio τ ranges from 0.2 to 0.9, our SRPM outperforms the original attention map results (indicated by the dashed lines). However, setting the filter ratio to 0.1 results in performance degradation. Overall, 0.5 appears to be an optimal choice, especially for higher-resolution settings, where its advantages are more pronounced.

Effect of supervised locations K in SRPM. To investigate the impact of the number of supervised locations K in Eq. (4) on the SRPM, we conducted extensive experiments, with the results presented in Fig. 5b. It is evident that when K ranges from 1 to 9, our SRPM outperforms the original attention map results (indicated by the dashed lines). Notably, the SRPM achieves optimal performance with K = 3 for both 224 × 224 and 576 × 576 resolutions, while K = 4 is optimal for 448 × 448 resolution. In general, the number of supervised locations K does not necessarily improve with larger values, optimal performance is typically achieved around K = 3.

Table 6: The effects of joint learning and optimization on the performance of the proposed model.

Method	mAP			
	COCO	NUS	VOC07	VG500
Ours w/o $\mathcal{L}_{aux}$	85.5	65.7	94.7	40.3
Ours with $\mathcal{L}_{aux}$	85.7	66.0	95.1	40.6

Table 7: The analysis of the hyper-parameter λ in Eq.(16).

Hyper-parameter λ	mAP			
	COCO	NUS	VOC07	VG500
0.1	85.4	65.7	94.8	40.3
0.4	85.5	65.8	94.9	40.4
0.7	85.5	65.8	95.0	40.5
1.0	**85.7**	**66.0**	**95.1**	**40.6**
1.3	85.5	65.9	95.0	**40.6**

Table 8: The model efficiency comparison of our OPRL, Q2L, and the Backbone Network.

Framework	Param	FLOPs	FPS	mAP	
				COCO	NUS
ResNet101	43M	31.5G	99	82.9	63.7
Q2L-R101	143M	36.6G	62	84.9	65.0
OPRL-R101	86M	34.7G	81	85.7	66.0

Effect of the joint learning and optimization. To explore the effect of the joint learning and optimization mechanism, we train the entire model using only $\mathcal{L}_{\text{cls}}$. As shown in Table 6, omitting $\mathcal{L}_{\text{aux}}$ consistently degrades mAP across all four datasets, which demonstrates the beneficial impact of $\mathcal{L}_{\text{aux}}$, validating the effectiveness of the proposed joint learning and optimization mechanism. We also examine the impact of the hyper-parameter λ in Eq. (16). As shown in Table 7, on the VG500 dataset, model performance improves as λ increases and then stabilizes. On the COCO, NUS, and VOC07 datasets, the model reaches peak performance at $\lambda = 1.0$ before declining. Consequently, setting $\lambda = 1.0$ appears to be optimal, indicating that both $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{aux}}$ are significant for the model.

Comparisons on model efficiency. We investigate model efficiency by comparing the proposed OPRL framework with the previous state-of-the-art model, Q2L [26]. The results, including Parameters (Param), Floating Point Operations (FLOPs), and Frames Per Second (FPS), are shown in Table 8. Using ResNet101 as the backbone, our proposed OPRL framework achieves better performance on the COCO and NUS datasets with fewer parameters compared to Q2L (86M vs 143M). Furthermore, our model exhibits lower FLOPs (34.7G vs 36.6G) and higher FPS (81 vs 62), highlighting its low complexity and fast inference capability.

Visualization. Figure 6 presents several examples from the COCO test set. Notably, compared to the baseline model, our OPRL framework better perceives image regions relevant to label semantics. As shown in Fig. 6 (a), OPRL accurately localizes objects such as bicycle, skateboard, person, car, and backpack. In addition, as illustrated in Fig. 6 (d), although the baseline model can locate some objects like spoon, laptop, and backpack, it also pays attention to irrelevant areas. In contrast, OPRL focuses more precisely on each object. We attribute this to our OPRL framework, which suppresses the interference from co-occurring objects and background information, enabling the model to focus more on the objects themselves. As a result, we achieve higher-quality label-aware representations, which enhance subsequent representation interaction and refinement, leading to superior performance in MLC tasks.

Figure 6: The visualization comparison of the OPRL framework and baseline model on several samples from COCO.

6 Conclusion and Discussion

In this work, addressing the challenge of contextual interference in MLC, we propose the Object-Purified Representation Learning framework with two key components. First, our Spatial-wise Representation Purification Module eliminates object-irrelevant feature activations, purifying representations from the spatial perspective of image content. Second, our Semantic-wise Association Purification Module models co-occurrence relationships while preventing their overemphasis, purifying representations from a semantic association perspective. Additionally, our Bidirectional Representation Refinement Module further enhances classification performance. The synergistic integration of these components achieves state-of-the-art results on multiple multi-label benchmarks.

Although this work yields promising results, a potential limitation is the use of a fixed threshold in SRPM to filter object-irrelevant regions. This approach may be overly stringent for small objects and too permissive for large ones in extreme cases. In the future, we will explore more advanced strategies to further improve our framework's ability to handle complex scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62076005, 61860206004, 62472004), the Dreams Foundation of Jianghuai Advance Technology Center (2023-ZM01Z015), Provincial Quality Project of Education in the New Era in 2023 (Postgraduate Education 2023lhpysfjd009), and the University Synergy Innovation Program of Anhui Province, China (GXXT-2021-002, GXXT-2022-029). We sincerely thank Bin Luo for his generous support of this project. We also thank the reviewers for their comments and suggestions. Finally, we acknowledge the High-performance Computing Platform of Anhui University for providing computational resources for this project.

References

[1] Shang-Fu Chen, Yi-Chen Chen, Chih-Kuan Yeh, and Yu-Chiang Wang. 2018. Order-free rnn with visual attention for multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 32. 6714–6721.

[2] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. 2019. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 522–531.

[3] Zhao-Min Chen, Quan Cui, Borui Zhao, Renjie Song, Xiaoqin Zhang, and Osamu Yoshie. 2022. Sst: Spatial and semantic transformers for multi-label image recognition. *IEEE Transactions on Image Processing* 31 (2022), 2570–2583.

[4] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. 2019. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 622–627.

[5] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5177–5186.

[6] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2023. Learning Graph Convolutional Networks for Multi-Label Recognition and Applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 6 (2023), 6969–6983.

[7] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. 2009. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. Santorini, Greece.

[8] Jean-Baptiste Cordonnier, Andreas Loukas, and Martin Jaggi. 2020. On the Relationship between Self-Attention and Convolutional Layers. In *International Conference on Learning Representations (ICLR)*.

[9] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. 2020. Randaugment: Practical automated data augmentation with a reduced search space. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

[10] Son D. Dao, He Zhao, and Jianfei Phung, Dinh Cai. 2023. Contrastively enforcing distinctiveness for multi-label image classification. *Neurocomputing* 555 (2023), 1.1–1.12.

[11] Xiang Deng, Songhe Feng, Gengyu Lyu, Tao Wang, and Congyan Lang. 2023. Beyond Word Embeddings: Heterogeneous Prior Knowledge Driven Multi-Label Image Classification. *IEEE Transactions on Multimedia* 25 (2023), 4013–4025.

[12] Terrance DeVries and Graham W Taylor. 2017. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552* (2017).

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations (ICLR)*.

[14] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2015. The Pascal Visual Object Classes Challenge: A Retrospective. *International Journal of Computer Vision* 111 (2015), 98–136.

[15] Bin-Bin Gao and Hong-Yu Zhou. 2021. Learning to Discover Multi-Class Attentional Regions for Multi-Label Image Recognition. *IEEE Transactions on Image Processing* 30 (2021), 5920–5932.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[17] Sergey Ioffe, Alexander Toshev, Yangqing Jia, Thomas Leung, and Yunchao Gong. 2013. Deep Convolutional Ranking for Multilabel Image Annotation. In *International Conference on Learning Representations (ICLR)*.

[18] Weitao Jiang, Wei Zhou, and Haifeng Hu. 2022. Double-Stream Position Learning Transformer Network for Image Captioning. *IEEE Transactions on Circuits and Systems for Video Technology* 32, 11 (2022), 7706–7718.

[19] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 123 (2017), 32–73.

[21] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. 2021. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16478–16488.

[22] Xin Li, Feipeng Zhao, and Yuhong Guo. 2014. Multi-label Image Classification with A Probabilistic Label Enhancement Model. In *UAI*, Vol. 1. 1–10.

[23] Zechao Li, Hao Tang, Zhimao Peng, Guo-Jun Qi, and Jinhui Tang. 2023. Knowledge-Guided Semantic Transfer Network for Few-Shot Image Recognition. *IEEE Transactions on Neural Networks and Learning Systems* (2023), 1–15.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollar, and Larry Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European Conference on Computer Vision (ECCV)*. Springer, 740–755.

[25] Ruyang Liu, Jingjia Huang, Thomas H. Li, and Ge Li. 2023. Causality Compensated Attention for Contextual Biased Visual Recognition. In *International Conference on Learning Representations (ICLR)*.

[26] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. 2021. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834* (2021).

[27] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W Tsang. 2021. The emerging trends of multi-label learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 11 (2021), 7955–7974.

[28] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*.

[29] Leilei Ma, Dengdi Sun, Lei Wang, Haifeng Zhao, and Bin Luo. 2023. Semantic-Aware Dual Contrastive Learning for Multi-label Image Classification. In *26th European Conference on Artificial Intelligence*. 1656–1663.

[30] Leilei Ma, Hongxing Xie, Lei Wang, Yanping Fu, Dengdi Sun, and Haifeng Zhao. 2024. Text-Region Matching for Multi-Label Image Recognition with Missing Labels. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 6133–6142.

[31] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. 2021. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 82–91.

[32] Tal Ridnik, Gilad Sharir, Avi Ben-Cohen, Emanuel Ben-Baruch, and Asaf Noy. 2023. Ml-decoder: Scalable and versatile classification head. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WCACV)*. 32–41.

[33] Leslie N Smith and Nicholay Topin. 2019. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, Vol. 11006. SPIE, 369–386.

[34] Lichun Tang, Zhaoxia Yin, Hang Su, Wanli Lyu, and Bin Luo. 2024. Wfss: weighted fusion of spectral transformer and spatial self-attention for robust hyperspectral image classification against adversarial attacks. *Visual Intelligence* 2, 1 (2024), 5.

[35] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. 2021. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems (NeurIPS)* 34 (2021), 24261–24272.

[36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)* 30 (2017), 6000–6010.

[37] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2285–2294.

[38] Lei Wang, Yibing Zhan, Leilei Ma, Dapeng Tao, Liang Ding, and Chen Gong. 2025. SpliceMix: A Cross-scale and Semantic Blending Augmentation Strategy for Multi-label Image Classification. *IEEE Transactions on Multimedia* (2025), 1–15.

[39] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. 2020. Multi-label classification with label graph superimposing. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 34. 12265–12272.

[40] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. 2017. Multi-Label Image Recognition by Recurrently Discovering Attentional Regions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 464–472.

[41] Xiangping Wu, Qingcai Chen, Wei Li, Yulun Xiao, and Baotian Hu. 2020. AdaHGNN: Adaptive Hypergraph Neural Networks for Multi-Label Image Classification. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 284–293.

[42] Yanan Wu, Songhe Feng, and Yang Wang. 2023. Semantic-Aware Graph Matching Mechanism for Multi-Label Image Recognition. *IEEE Transactions on Circuits and Systems for Video Technology* (2023), 6788–6803.

[43] Yanan Wu, Songhe Feng, Gongpei Zhao, and Yi Jin. 2024. Transformer Driven Matching Selection Mechanism for Multi-Label Image Classification. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 2 (2024), 924–937.

[44] Yanan Wu, He Liu, Songhe Feng, Yi Jin, Gengyu Lyu, and Zizhang Wu. 2021. GM-MLIC: Graph Matching based Multi-Label Image Classification. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 1179–1185.

[45] Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. 2023. Delving into Masked Autoencoders for Multi-Label Thorax Disease Classification. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 3577–3589.

[46] Ming-Kun Xie, Jiahao Xiao, and Sheng-Jun Huang. 2022. Label-Aware Global Consistency for Multi-Label Learning with Single Positive Labels. *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), 18430–18441.

[47] Ming-Kun Xie, Jia-Hao Xiao, Pei Peng, Gang Niu, Masashi Sugiyama, and Sheng-Jun Huang. 2024. Counterfactual Reasoning for Multi-Label Image Classification via Patching-Based Training. In *Proceedings of the International Conference on Machine Learning (ICML)*. 54576–54589.

[48] Jiazhi Xu, Sheng Huang, Fengtao Zhou, Luwen Huangfu, Daniel Zeng, and Bo Liu. 2022. Boosting Multi-Label Image Classification with Complementary Parallel Self-Distillation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*. 1495–1501.

[49] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. 2020. Orderless recurrent models for multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13440–13449.

[50] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. 2020. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *European Conference on Computer Vision (ECCV)*. Springer, 649–665.

[51] Jialu Zhang, Jianfeng Ren, Qian Zhang, Jiang Liu, and Xudong Jiang. 2023. Spatial Context-Aware Object-Attentional Network for Multi-Label Image Classification. *IEEE Transactions on Image Processing* 32 (2023), 3000–3012.

[52] Min-Ling Zhang and Zhi-Hua Zhou. 2013. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 26, 8 (2013), 1819–1837.

[53] Jiawei Zhao, Ke Yan, Yifan Zhao, Xiaowei Guo, Feiyue Huang, and Jia Li. 2021. Transformer-based dual relation graph for multi-label image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 163–172.

[54] Wei Zhou, Weitao Jiang, Dihu Chen, Haifeng Hu, and Tao Su. 2024. Mining Semantic Information With Dual Relation Graph Network for Multi-Label Image Classification. *IEEE Transactions on Multimedia* 26 (2024), 1143–1157.

[55] Wei Zhou, Zhiwu Xia, Peng Dou, Tao Su, and Haifeng Hu. 2023. Aligning Image Semantics and Label Concepts for Image Multi-Label Classification. *ACM Transactions on Multimedia Computing Communications and Applications* 19, 2, Article 75 (2023).

[56] Wei Zhou, Zhiwu Xia, Peng Dou, Tao Su, and Haifeng Hu. 2023. Double Attention Based on Graph Attention Network for Image Multi-Label Classification. *ACM Transactions on Multimedia Computing Communications and Applications* 19, 1, Article 18 (2023).

[57] Wei Zhou, Zhijie Zheng, Tao Su, and Haifeng Hu. 2024. DATran: Dual Attention Transformer for Multi-Label Image Classification. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 1 (2024), 342–356.

[58] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. 2017. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5513–5522.

[59] Ke Zhu and Jianxin Wu. 2021. Residual attention: A simple but effective method for multi-label recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 184–193.

[60] Lei Zhu, Hui Cui, Zhiyong Cheng, Jingjing Li, and Zheng Zhang. 2021. Dual-Level Semantic Transfer Deep Hashing for Efficient Social Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 4 (2021), 1478–1489.

[61] Xuelin Zhu, Jiuxin Cao, Jiawei Ge, Weijia Liu, and Bo Liu. 2022. Two-stream transformer for multi-label image classification. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*. 3598–3607.

[62] Xuelin Zhu, Jianshu Li, Jiuxin Cao, Dongqi Tang, Jian Liu, and Bo Liu. 2024. Semantic-Guided Representation Enhancement for Multi-Label Image Classification. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 10 (2024), 10036–10049.

[63] Xuelin Zhu, Jian Liu, Weijia Liu, Jiawei Ge, Bo Liu, and Jiuxin Cao. 2023. Scene-aware label graph learning for multi-label image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 1473–1482.