Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# Dynamic multi-modal hypergraph learning for semi-supervised multi-label image recognition

Chen Zhang [a], Cheng Xu [a], Yu Xie [b],*, Wenjie Mao [a], Bin Yu [a]

[a] *School of Computer Science and Technology, Xidian University, China*
[b] *School of Computer and Information Technology, Shanxi University, China*

ARTICLE INFO

*Keywords:*
Multi-label image recognition
Hypergraph neural networks
Label correlation
Semi-supervised learning

ABSTRACT

In recent years, multi-label image recognition has emerged as a crucial task in computer vision, requiring simultaneous detection of multiple objects or attributes within images. Unlike single-label classification, this task demands explicit modeling of complex label correlations. Existing methods primarily focus on low-order pairwise relationships, failing to capture higher-order dependencies critical for real-world scene understanding. Additionally, the long-tail distribution of labels often causes models to prioritize frequent head labels while neglecting rare tail labels with limited training samples. To address these challenges, we propose a Dynamic Multi-modal Hypergraph Learning (DMHL) framework for semi-supervised multi-label recognition. DMHL constructs adaptive hypergraphs by fusing visual features, co-occurrence statistics, and textual embeddings. The framework dynamically refines these hypergraphs through three novel modules: HyperPrune, which prunes redundant hyperedges; HyperTransform, which generates dynamic hyperedges from node features; and HyperTune, which optimizes hypergraph weights via feature similarity alignment. These dynamic optimization modules enable DMHL to capture intricate high-order label correlations. Furthermore, DMHL employs hypergraph residual concatenation to enhance deep feature representations, which are leveraged for dynamic pseudo-label generation to alleviate label imbalance. Extensive experiments demonstrate DMHL achieves state-of-the-art results across four benchmarks: 86.0% mAP on MS-COCO (0.7% gain over prior SOTA), 96.3%/96.5% on Pascal VOC 2007/2012, and 64.0% on NUS-WIDE. Notably, in semi-supervised settings with 5% labeled data, DMHL surpasses previous methods by over 20% mAP on MS-COCO (70.5%), highlighting its effectiveness in capturing intricate label relationships and improving tail label recognition.

## 1. Introduction

Multi-label image recognition, a pivotal task in computer vision, involves identifying multiple objects or attributes within a single image. Unlike single-label classification, which assigns an image to a predefined category, multi-label recognition annotates an image with multiple labels, reflecting the complexity of real-world scenes. This complexity arises from the diversity of coexisting objects and their intricate relationships. Multi-label image recognition is crucial for applications such as automatic image annotation [1], image retrieval [2], surveillance, and environmental monitoring [3].

Despite its importance, multi-label recognition faces significant challenges. Existing methods [4,5] primarily rely on graph neural networks (GNNs) to model pairwise label correlations, but these approaches overlook higher-order dependencies (e.g., a “bird" often co-occurs with both “sky” and “tree”). Moreover, most methods use manually curated co-occurrence statistics, leading to inflexible models that reinforce frequency bias [6] and struggle with novel label combinations.

Another critical issue is the long-tail distribution of labels, where models prioritize head labels with abundant training data while neglecting tail labels that are underrepresented but essential for comprehensive scene understanding. Although theoretically solvable by large-scale labeled datasets, such resources are scarce in real-world multi-label scenarios, necessitating semi-supervised approaches that leverage unlabeled data to mitigate data scarcity. Existing methods often fail to address this challenge, as they either ignore tail labels entirely or rely on heuristic sampling strategies that do not fundamentally resolve the imbalance.

To address these limitations, we propose Dynamic Multi-modal Hypergraph Learning (DMHL), a semi-supervised framework that explicitly captures high-order label correlations and mitigates label imbalance, centered around three core innovations. Firstly, it constructs

* Corresponding author.
*E-mail address:* sxlljcxy@gmail.com (Y. Xie).

https://doi.org/10.1016/j.patcog.2025.111959
Received 16 December 2024; Received in revised form 12 April 2025; Accepted 2 June 2025
Available online 17 June 2025

adaptive hypergraphs by fusing visual features, co-occurrence statistics, and textual embeddings, enabling comprehensive modeling of label relationships beyond static co-occurrence. Secondly, DMHL dynamically refines hypergraphs through three modules: HyperPrune (pruning redundant hyperedges), HyperTransform (generating dynamic hyperedges from node features), and HyperTune (aligning hypergraph weights with feature similarities). This dynamic optimization contrasts with the static hypergraph structure in AdaHGNN [7], allowing DMHL to adapt to evolving label dependencies. Thirdly, DMHL employs hypergraph residual concatenation to generate dynamic pseudo-labels, guiding unsupervised learning and improving performance on tail labels with limited training data.

The contributions of this work are fourfold:

(1) Multi-Modal Hypergraph Construction: We design a novel framework that adaptively integrates visual, statistical, and textual modalities to capture complex label dependencies.
(2) Dynamic Hypergraph Optimization: The proposed HyperPrune, HyperTransform and HyperTune modules iteratively refine hypergraph structures, enabling real-time adaptation to dynamic label relationships.
(3) Hypergraph Residual-Guided Pseudo-Labeling: Via hypergraph residual concatenation, DMHL enhances deep feature representations, while dynamic pseudo-label generation leverages unlabeled data to elevate tail label recognition under sparse annotations.
(4) Extensive Validation: Comprehensive experiments on four benchmarks (MS-COCO, Pascal VOC 2007/2012, NUS-WIDE) validate DMHL's superiority over existing methods, including hypergraph-based AdaHGNN and graph-based ML-GCN.

The remainder of this paper is structured as follows. Section 2 reviews related work on hypergraph-based and semi-supervised multi-label recognition. Section 3 details the DMHL framework. Section 4 presents experimental results, and Section 5 concludes with future research directions.

## 2. Related work

Multi-label image recognition is a key area in computer vision, distinguishing itself from single-label classification by requiring multiple annotations per image. Early methods trained independent binary classifiers for each label but struggled with scalability due to the exponential growth of possible label combinations. This limitation has spurred the development of dependency modeling techniques to capture label correlations.

### *2.1. Graph-based multi-label image recognition*

To address the scalability issue, early attempts such as CNN-RNN [5] employ recurrent networks to propagate label information, yet these RNN-based approaches have inherent limitations in capturing complex non-sequential relationships among labels. As a more effective alternative, graph-based methods emerge, which model label dependencies through pairwise edges, offering a natural and intuitive way to encode label co-occurrences. Specifically, ML-GCN [4] initiates GCN-based multi-label recognition by iteratively updating label representations, and SSGRL [8] enriches the process by integrating dynamic semantic-region interactions to enhance the modeling of label correlations.

Notwithstanding these advancements, recent methods like MFS-AGD [9] and cfMGNML [10] still rely on pairwise graph structures and static adjacency matrices. For instance, MFS-AGD introduces adaptive graph diffusion to preserve higher-order structural information between instances, while cfMGNML proposes dual-granularity labeling to handle noisy label scenarios. However, their foundational reliance on static pairwise relations inherently limits the ability to capture dynamic label combinations.

This persistent limitation highlights the critical need for models that can transcend pairwise dependencies. Indeed, graph-based methods remain confined to pairwise relationships, failing to capture higher-order dependencies such as the co-occurrence of "bird" with both "sky" and "tree". Moreover, their reliance on static adjacency matrices derived from co-occurrence statistics restricts adaptability to dynamic label combinations, underscoring the necessity for more powerful models capable of handling complex label interactions.

### *2.2. Hypergraph-based multi-label image recognition*

Hypergraph structures address the limitations of graphs by modeling high-order relationships through hyperedges connecting multiple vertices. Early hypergraph approaches use static hypergraphs for multi-label propagation and noise mitigation, but their fixed structures struggle to adapt to evolving data distributions. The introduction of hypergraph neural networks (HGNNs) marks a significant leap. AdaHGNN [7] initializes hypergraphs with label embeddings to capture high-order semantic relations, yet its fixed-dimension incidence matrix and reliance on pre-trained embeddings limit scalability.

Recently, IC-HGT [11] introduces a transformer-based hypergraph architecture with attention mechanisms to propagate label features, while a multi-modal hypergraph method [12] enhances cross-modal correlations through hyperedge convolutions. However, both approaches still rely on pre-defined hypergraph structures, lacking dynamic adaptability to evolving label relationships.

In contrast, our DMHL framework introduces dynamic hypergraph optimization (HyperPrune, HyperTransform, HyperTune) to iteratively refine hypergraph structures. By fusing visual, statistical, and textual modalities, DMHL addresses the scalability issues of static hypergraphs while capturing complex high-order dependencies.

### *2.3. Semi-supervised multi-label image recognition*

Semi-supervised learning (SSL) techniques emerge to leverage unlabeled data in multi-label recognition. Early works focus on global structure learning and manifold embedding, but they primarily adapt single-label SSL frameworks. Guo et al. [13] explore label propagation, yet these methods overlook two critical challenges: higher-order label correlations and long-tail label distributions.

Building upon these foundations, recent studies have addressed specific limitations through domain-specific innovations. For instance, predictive clustering trees [14] enhance label propagation in remote sensing by integrating deep feature extraction with semi-supervised PCT ensembles. Similarly, stacked co-training frameworks [15] combine multiple SSL assumptions (e.g., co-training, clustering, manifold) to improve label correlation modeling. Notwithstanding these improvements, both approaches remain constrained by implicit label dependency modeling and fixed topological assumptions, failing to explicitly capture dynamic high-order relationships.

DMHL advances SSL for multi-label recognition by combining hypergraph residual concatenation with dynamic pseudo-label generation, which effectively aligns supervised deep features with unsupervised shallow features, mitigating label imbalance and enabling better utilization of unlabeled data.

## 3. Methodology

This section introduces the Dynamic Multi-modal Hypergraph Learning (DMHL) framework for semi-supervised multi-label image recognition. The framework addresses three key challenges: (i) capturing high-order label dependencies, (ii) mitigating long-tail label distributions, and (iii) leveraging unlabeled data. It begins with multi-scale feature extraction using ResNet-101, followed by three core modules: multi-modal hypergraph construction, dynamic hypergraph structure optimization, and dynamic pseudo-label generation, detailed in subsequent subsections.

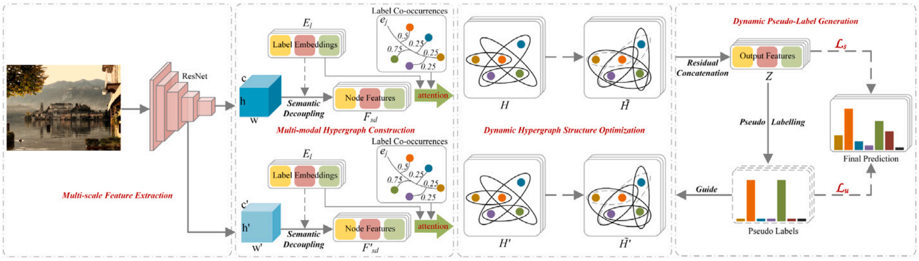


**Fig. 1.** DMHL framework overview. The pipeline includes: (1) multi-scale feature extraction via ResNet-101; (2) multi-modal hypergraph construction integrating visual, statistical, and textual modalities; (3) dynamic hypergraph structure optimization through HyperPrune, HyperTransform, and HyperTune; (4) dynamic pseudo-label generation aligning supervised and unsupervised features. Hypergraph nodes represent labels (e.g., "bird"), and hyperedges encode high-order dependencies where a single hyperedge connects multiple nodes (e.g., "bird" linked to both "sky" and "tree" via a hyperedge, indicating their co-occurrence).

### 3.1. Overview of DMHL

To address the inherent challenges of multi-label image recognition, the proposed DMHL framework employs an innovative fusion of multi-modal information and hypergraph theory. The framework unfolds through several key phases: (1) multi-modal hypergraph construction integrating visual, statistical, and textual modalities; (2) dynamic hypergraph structure optimization via HyperPrune, HyperTransform, and HyperTune; and (3) dynamic pseudo-label generation for semi-supervised learning. These phases guide the process from initial feature extraction to final label prediction in an end-to-end manner.

The overall pipeline is illustrated in Fig. 1. Our DMHL framework begins with ResNet-101 [16] extracting multi-scale visual features. Simultaneously, label embeddings $\mathbf{E}_l$ are obtained from pre-trained word representation models. These features and embeddings undergo semantic decoupling, producing category-specific hierarchical features $\mathbf{F}_{sd}$ and $\mathbf{F}'_{sd}$. Co-occurrence statistics, precomputed from the training dataset, capture label co-occurrence frequencies. A multi-modal hypergraph $\mathbf{H}$ is then constructed, with nodes representing labels and hyperedges encoding interdependencies from visual, textual, and statistical modalities. The hypergraph structure is dynamically optimized through three modules: HyperPrune prunes redundant hyperedges, HyperTransform generates new hyperedges from updated features, and HyperTune adjusts weights to reflect label correlations. Node features are enriched via residual concatenation, resulting in enhanced features $\mathbf{Z}$, which generate dynamic pseudo-labels using confidence-based thresholds to reduce noise propagation for tail labels. The learning process is guided by a total loss function $\mathcal{L}_{\text{total}}$, combining supervised and unsupervised components.

### 3.2. Multi-scale feature extraction

This subsection describes how deep and shallow visual features are extracted, semantically decoupled, and utilized for subsequent hypergraph construction.

The process begins with ResNet-101 extracting multi-scale visual features from the input image. Concurrently, label embeddings $\mathbf{E}_l$ are derived from the pre-trained GloVe [17] word representation model. These multi-scale visual features and label embeddings are then semantically decoupled via a two-branch network, following the method employed in SSGRL. This process generates category-specific hierarchical features $\mathbf{F}_{sd}$ (for hypergraph construction) and $\mathbf{F}'_{sd}$ (for pseudo-label generation).

$\mathbf{F}_{sd}$ and $\mathbf{F}'_{sd}$ are independently employed for hypergraph construction and pseudo-label generation, respectively. During the Dynamic Pseudo-Label Generation phase, these features are aligned to reduce distribution disparity. For simplicity, $\mathbf{F}_{sd}$ is used as the representative feature to illustrate the subsequent hypergraph learning process.

### 3.3. Multi-modal hypergraph construction

The DMHL framework constructs a hypergraph by fusing three complementary modalities: visual features capture image-level semantics, statistical co-occurrences encode dataset-level dependencies, and textual embeddings provide linguistic correlations. This subsection details how these modalities are integrated to form an adaptive hypergraph structure.

**Visual Incidence Matrix $\mathbf{H}_v$** Visual features $\mathbf{F}_{sd} \in \mathbb{R}^{N\times D}$ undergo global average pooling (GAP) to reduce spatial dimensions, followed by a convolutional layer (Conv), batch normalization (BN), and Leaky ReLU activation (LReLU):

$$\mathbf{F}_g = \mathrm{LReLU}\left(\mathrm{BN}\left(\mathrm{Conv}\left(\mathrm{GAP}\left(\mathbf{F}_{sd}\right)\right)\right)\right) \tag{1}$$

where each operation enhances feature robustness. The visual incidence matrix is then computed via:

$$\mathbf{H}_v = \sigma\left(\mathrm{Conv}\left(\mathbf{F}_g\right)\right) \tag{2}$$

with $\sigma$ normalizing weights to [0, 1].

**Statistical Incidence Matrix $\mathbf{H}'_s$** Using precomputed adjacency matrix $\mathbf{A}$ and label frequencies $N$, the statistical matrix $\mathbf{H}_s$ captures pairwise co-occurrences:

$$\mathbf{H}_{s_{ij}} = \begin{cases} \frac{\mathbf{A}_{ij}}{N_j}, & \text{if } j \in \mathcal{V}_i \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Self-connections are added via $\mathbf{H}'_s = \mathbf{H}_s + \mathbf{I}_n$ to encode label self-dependency. This statistical structure is dynamically enhanced during optimization (Section 3.4), enabling the hypergraph to capture both static co-occurrences and dynamic label combinations.

**Textual Incidence Matrix $\mathbf{H}_t$** Pre-trained GloVe embeddings $\mathbf{E}_l \in \mathbb{R}^{N\times d}$ are directly used as $\mathbf{H}_t$ to capture initial semantic relationships among labels. This initialization leverages pre-trained language knowledge to provide a robust starting point for hypergraph construction. Subsequent dynamic optimization (Section 3.4) refines these relationships, ensuring adaptability to dataset-specific label semantics.

**Adaptive Fusion** Modal-specific matrices $\mathbf{H}_v$ (visual), $\mathbf{H}'_s$ (statistical), and $\mathbf{H}_t$ (textual) are dynamically weighted and combined using learnable attention weights. First, each modality $i$ is assigned a scalar weight $w_i$, normalized via sigmoid to obtain attention weights. This weight is replicated to match the number of columns $M_i$ in $\mathbf{H}_i$:

$$\mathbf{w}_i = \mathrm{copy}(\sigma(w_i), M_i) \quad \text{with } \mathbf{w}_i \in \mathbb{R}^{M_i} \tag{4}$$

Next, these weights are concatenated into a diagonal matrix $\mathbf{W}$:

$$\mathbf{W} = \mathrm{diag}(\mathbf{w}) \tag{5}$$

Finally, the modality-specific matrices are concatenated and multiplied by $\mathbf{W}$ to emphasize relevant modalities:

$$\mathbf{H} = (\mathbf{H}_v \oplus \mathbf{H}'_s \oplus \mathbf{H}_t)\mathbf{W} \tag{6}$$

This adaptive fusion mechanism dynamically balances contributions from visual, statistical, and textual modalities, providing a robust foundation for capturing multi-scale label relationships. However, the initial hypergraph structure is further optimized in Section 3.4 to adapt to evolving label dependencies and reduce redundancy, ensuring the hypergraph remains aligned with the dataset's underlying structure.

### 3.4. Dynamic hypergraph structure optimization

Upon establishing the multi-modal hypergraph incidence matrix $\mathbf{H}$, our DMHL framework progresses to the Dynamic Hypergraph Structure Optimization phase. This phase systematically refines the initial hypergraph using three complementary strategies: HyperPrune removes redundant hyperedges, HyperTransform generates new hyperedges from updated features, and HyperTune adjusts hypergraph weights to reflect evolving label correlations. This dynamic optimization process enables the hypergraph to adapt to unseen label combinations and mitigate the impact of noisy annotations, addressing the limitations of static hypergraph models.

**HyperPrune** HyperPrune refines the hypergraph by dynamically pruning redundant hyperedges using attention-based masking. This process reduces noise and computational complexity while preserving high-order label correlations, addressing the sparsity challenge in multi-label recognition.

For each hyperedge $i$, attention weights $m_i$ are computed using a Gumbel-Softmax distribution:

$$m_i = \sigma\left(\frac{\exp\left(\frac{1}{\tau}\left(\mathbf{W}_e\mathbf{e}_i + \mathbf{W}_n\overline{\mathbf{n}}_i + \epsilon\right)\right)}{\sum_{k=1}^{E}\exp\left(\frac{1}{\tau}\left(\mathbf{W}_e\mathbf{e}_k + \mathbf{W}_n\overline{\mathbf{n}}_k + \epsilon\right)\right)}\right) \tag{7}$$

where $\mathbf{e}_i$ and $\overline{\mathbf{n}}_i$ encode hyperedge and node features, respectively. $\mathbf{W}_e$ and $\mathbf{W}_n$ learn feature projections, while $\epsilon$ and $\tau$ ensure differentiable sampling. The sigmoid $\sigma$ normalizes $m_i$ to [0,1], balancing sparsity and stability.

A binary mask $\mathbf{m}$ is generated by applying an adaptive threshold $\gamma_t = \gamma_0 - \alpha \cdot t$, where $\gamma_0$ is the initial threshold, $\alpha$ controls decay rate, and $t$ is the training iteration. This mask is broadcasted across nodes using $\Gamma(\mathbf{m})$ and element-wise multiplied with $\mathbf{H}$:

$$\hat{\mathbf{H}} = \mathbf{H} \odot \Gamma(\mathbf{m}) \tag{8}$$

The resulting pruned matrix $\hat{\mathbf{H}}$ retains only hyperedges with $m_i > \gamma_t$, ensuring sparsity and relevance.

This attention-based pruning enhances model interpretability and efficiency by focusing on informative hyperedges. The adaptive thresholding strategy dynamically balances aggressive pruning early in training with gradual relaxation as the model converges.

**HyperTransform** HyperTransform dynamically expands the hypergraph by generating new hyperedges from updated node features, addressing the challenge of capturing evolving label correlations. This process combines current and emerging data characteristics to enhance the model's adaptability.

First, node features are updated via hypergraph convolution:

$$\mathbf{F}_h = \mathrm{HConv}\left(\mathbf{F}_{sd}, \hat{\mathbf{H}}\right) \tag{9}$$

where $\mathbf{F}_h$ denotes hypergraph-convolved features, and $\hat{\mathbf{H}}$ is the pruned incidence matrix from HyperPrune.

Next, global features $\mathbf{F}_{gh}$ are extracted from $\mathbf{F}_h$ using the same pipeline as in Eq. (1). These features are concatenated with $\mathbf{F}_h$ ($\oplus$), averaged to reduce noise, and convolved to generate dynamic hyperedges:

$$\mathbf{H}_d = \sigma\left(\mathrm{Conv}\left(\mathrm{mean}\left(\mathbf{F}_{gh} \oplus \mathbf{F}_h\right)\right)\right) \tag{10}$$

The sigmoid function normalizes $\mathbf{H}_d$ to [0, 1], representing new hyperedge strengths.

Finally, $\mathbf{H}_d$ is fused with $\hat{\mathbf{H}}$ using the adaptive attention mechanism (Section 3.3), producing the transformed hypergraph $\mathbf{H}^*$. This dynamic generation of new hyperedges captures emerging label correlations not present in the initial statistical co-occurrence matrix, enabling the hypergraph to adapt to unseen label combinations.

**HyperTune** HyperTune dynamically adjusts hyperedge weights based on node-hyperedge similarity, refining label correlations to adapt to evolving data characteristics. This process ensures the hypergraph remains aligned with the most prevalent label interdependencies.

First, cosine similarity between node features $\mathbf{F}_{h_i}$ (from HyperTransform) and hyperedge features $\mathbf{e}_j$ (derived from $\mathbf{H}^*$) is computed:

$$\mathbf{S}_{ij} = \frac{1}{\|\mathbf{F}_{h_i}\|\|\mathbf{e}_j\|}\langle \mathbf{F}_{h_i}, \mathbf{e}_j \rangle \tag{11}$$

Normalized to [−1,1], $\mathbf{S}$ measures semantic alignment between nodes and hyperedges.

A dynamic threshold $\theta = \mu(\mathbf{S}) + \lambda \cdot \sigma(\mathbf{S}) + \mathbf{b}$ is computed, where $\mu/\sigma$ are $\mathbf{S}$'s statistics, $\lambda$ controls sensitivity, and $\mathbf{b}$ is a learned bias. The incidence matrix $\mathbf{H}^*$ is updated as follows:

$$\tilde{\mathbf{H}}_{ij} = \begin{cases} \mathbf{S}_{ij}, & \text{if } \mathbf{S}_{ij} > \theta \text{ and } \mathbf{H}^*_{ij} = 0 \text{ (add new hyperedge)} \\ 0, & \text{if } \mathbf{S}_{ij} < \theta \text{ and } \mathbf{H}^*_{ij} \neq 0 \text{ (remove irrelevant hyperedge)} \\ \mathbf{H}^*_{ij}, & \text{otherwise} \end{cases} \tag{12}$$

This adaptive thresholding ensures $\tilde{\mathbf{H}}$ adapts to emerging label correlations not captured by initial statistical co-occurrences, enhancing generalization to unseen label combinations.

The refined incidence matrix $\tilde{\mathbf{H}}$ reflects current node-hyperedge similarities, improving the model's predictive capability by focusing on task-relevant label dependencies.

The Dynamic Hypergraph Structure Optimization, encompassing HyperPrune, HyperTransform, and HyperTune, collectively refines the hypergraph to more accurately represent label interdependencies. This tripartite refinement approach not only ensures that the DMHL framework's hypergraph is sparsity-optimized but also adaptively reflects the underlying data structure, especially in the face of new label combinations and evolving data patterns. By capturing accurate label relationships and adapting to new data, this refined hypergraph provides a crucial and well-structured input for the subsequent phase: Dynamic Pseudo-Label Generation, which can then leverage these optimized relationships to generate more reliable pseudo-labels.

### 3.5. Dynamic Pseudo-Label Generation

The Dynamic Pseudo-Label Generation phase in our DMHL framework leverages both labeled and unlabeled data for semi-supervised multi-label image recognition. It refines feature representations and narrows the gap between shallow and deep features through hypergraph learning and dynamic pseudo-label generation.

Shallow features from unlabeled data, which capture low-level patterns without strong supervision, can be refined by deep, supervised features. The category-specific hierarchical features $\mathbf{F}_{sd}$ and $\mathbf{F}'_{sd}$, extracted and semantically decoupled during the Multi-scale Feature Extraction phase, undergo independent hypergraph learning through two layers of hypergraph convolutions, resulting in updated features $\mathbf{F}_z$ and $\mathbf{F}'_z$, respectively.

Specifically, we concatenate progressively refined deep features at various levels of the hypergraph convolution to synthesize a composite feature vector, $\mathbf{Z}$, that encapsulates both the initial and evolved deep node representations. The residual concatenation is formally expressed as:

$$\mathbf{Z} = \mathbf{F}_{sd} \oplus \mathbf{F}_h \oplus \mathbf{F}_z \tag{13}$$

Here, $\mathbf{Z}$ is the deep residual feature vector obtained by concatenating the initial category-specific hierarchical features $\mathbf{F}_{sd}$, the output of the first hypergraph convolution layer $\mathbf{F}_h$, and the updated deep features $\mathbf{F}_z$. This residual concatenation helps mitigate over-smoothing.

Subsequently, the refined feature set $\mathbf{Z}$, enriched through deep layer convolutions, is subjected to classification to yield preliminary scores $\mathbf{S}_z$:

$$\mathbf{S}_z = \sigma(\mathbf{W}_c\mathbf{Z} + \mathbf{b}_c) \tag{14}$$

The preliminary scores $\mathbf{S}_z$ are calculated by applying the sigmoid function $\sigma$ to the linear combination of the deep residual feature vector $\mathbf{Z}$ with the weights $\mathbf{W}_c$ and biases $\mathbf{b}_c$ of the classifier layer.

Similarly, the updated shallow features $\mathbf{F}'_z$ are also classified to produce prediction scores $\mathbf{S}'_z$.

Pseudo-labels, serving as soft targets for training on unlabeled data, are dynamically generated based on thresholds that distinguish confident from uncertain predictions. The generation process is governed by:

$$y_u = \begin{cases} 1, & \mathbf{S}_z > T_{\text{pos}} \\ 0, & \mathbf{S}_z < T_{\text{neg}} \\ \text{ignore}, & \text{otherwise} \end{cases} \tag{15}$$

Binary pseudo-labels $y_u$ are assigned based on the comparison between the sigmoid-activated score $\mathbf{S}_z$ and the positive threshold $T_{\text{pos}}$ and negative threshold $T_{\text{neg}}$. Scores between the thresholds are ignored to avoid noise. The positive threshold $T_{\text{pos}}$ is initially set based on the class imbalance ratio, $\eta$, calculated by normalizing class occurrence counts and applying a sigmoid function, and is further scaled by a hyperparameter $\beta$ to prevent extreme values. During training, $T_{\text{pos}}$ can be adjusted based on the training progress, such as decreasing it gradually as the model converges. The negative threshold $T_{\text{neg}}$ is initially set to a fixed value but is adjusted during training according to the distribution of $\mathbf{S}_z$ scores, for example, by increasing it if the number of confidently negative samples is too large.

To enhance the precision of pseudo-labels, a warm-up phase using a subset of labeled data is employed. During this phase, the deep residual feature $\mathbf{Z}$ is pre-trained for a predefined number of epochs $n$ to stabilize the feature space before the introduction of semi-supervised learning.

To compute the weighted unsupervised loss function, which guides the alignment of unsupervised shallow features with supervised deep features, weights are assigned to both positively and negatively predicted instances based on the confidence level of their predictions. For positive predictions, specific weights are calculated as follows:

$$w_{\text{pos}} = \begin{cases} \frac{1}{1+e^{-(\mathbf{S}_z - T_{\text{pos}})}}, & \text{if } \mathbf{S}_z > T_{\text{pos}} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

The weights for positive predictions $w_{\text{pos}}$ are calculated using a sigmoid-like function, which assigns non-zero weights only when $\mathbf{S}_z$ exceeds $T_{\text{pos}}$. And for negative predictions:

$$w_{\text{neg}} = \mathbb{I}(\mathbf{S}_z < T_{\text{neg}}) \tag{17}$$

The weights for negative predictions $w_{\text{neg}}$ are determined by the indicator function $\mathbb{I}$, which assigns a weight of 1 when $\mathbf{S}_z$ is below $T_{\text{neg}}$.

To align unsupervised shallow features $\mathbf{F}'_z$ with high-confidence pseudo-labels, we compute $\mathcal{L}_{\text{u}}$ using weighted binary cross-entropy. The weights $w_{\text{pos}}$ and $w_{\text{neg}}$ dynamically adjust based on prediction confidence, prioritizing reliable pseudo-labels and reducing noise from uncertain predictions. This strategy specifically addresses the long-tail problem by assigning higher weights to rare categories when their predictions exceed $T_{\text{pos}}$:

$$\mathcal{L}_{\text{u}} = -\left(w_{\text{pos}} y_u \log \mathbf{S}'_z + w_{\text{neg}}\left(1 - y_u\right)\log\left(1 - \mathbf{S}'_z\right)\right) \tag{18}$$

This weighted loss ensures the model learns from both common and rare labels, leveraging unlabeled data to mitigate class imbalance.

To prevent overfitting and ensure robust learning, regularization is applied to the unsupervised loss $\mathcal{L}_{\text{u}}$. This helps the model learn the comprehensive information of deep features while retaining the unique spatial positional information of shallow features, thereby enhancing overall model generalization and performance in semi-supervised multi-label image recognition.

By integrating hypergraph residual concatenation with dynamic pseudo-labeling, the Dynamic Pseudo-Label Generation module effectively incorporates substantial unlabeled data during training. This adaptability allows the module to handle different data distributions and label correlations, compensating for the scarcity of well-labeled multi-label datasets that leads to the long-tail distribution problem. Consequently, our DMHL's ability to capture diverse label correlations is enhanced, ensuring the accurate recognition of both common and rare categories.

### 3.6. Learning objective

The learning objective combines supervised and unsupervised components to guide the framework's training. The supervised loss $\mathcal{L}_{\text{s}}$ is defined using Binary Cross-Entropy (BCE) to measure discrepancies between true labels and predictions:

$$\mathcal{L}_{\text{s}} = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij}\log p_{ij} + (1 - y_{ij})\log(1 - p_{ij}) \tag{19}$$

where $y_{ij}$ is the binary true label (1 for presence, 0 for absence), and $p_{ij}$ is the predicted probability generated by the classifier. The classifier takes as input the concatenated feature vector $\mathbf{Z} \oplus \mathbf{F}'_z$, combining deep residual features $\mathbf{Z}$ (encoding high-level semantics) and shallow features $\mathbf{F}'_z$ (preserving low-level details) to capture hierarchical label correlations.

The total loss function balances supervision and pseudo-label guidance:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{s}} + \lambda \cdot \mathcal{L}_{\text{u}} \tag{20}$$

Here, $\lambda$ is a fixed hyperparameter manually tuned to balance the contributions of labeled and unlabeled data. Parameter Sensitivity Analysis validates the optimal value of $\lambda$ in different scenarios, demonstrating its impact on model performance.

By optimizing $\mathcal{L}_{\text{total}}$, the framework addresses the long-tail problem by leveraging pseudo-labels to augment rare categories, ensuring balanced training across all labels. This dual-objective approach harnesses hierarchical feature representations and dynamic hypergraph structures to capture nuanced label correlations.

## 4. Experiments

In this section, we outline the benchmark datasets (MS-COCO, VOC2007, VOC2012, and NUS-WIDE), evaluation metrics, and implementation details. We then compare our DMHL framework with existing methods, followed by ablation studies and hyperparameter sensitivity analysis.

### 4.1. Evaluation metrics

To ensure fair comparison with state-of-the-art models, we employ standard metrics for multi-label image recognition: average per-class precision (CP), recall (CR), F1-score (CF1), and average overall precision (OP), recall (OR), F1-score (OF1). These metrics are calculated by comparing ground truth labels with predicted probabilities, assigning a label as positive if its estimated probability exceeds 0.5:

$$\begin{aligned} \text{OP} &= \frac{\sum_i N_i^c}{\sum_i N_i^p}, & \text{CP} &= \frac{1}{C}\sum_i \frac{N_i^c}{N_i^p}, \\ \text{OR} &= \frac{\sum_i N_i^c}{\sum_i N_i^g}, & \text{CR} &= \frac{1}{C}\sum_i \frac{N_i^c}{N_i^g}, \\ \text{OF1} &= \frac{2 \times \text{OP} \times \text{OR}}{\text{OP} + \text{OR}}, & \text{CF1} &= \frac{2 \times \text{CP} \times \text{CR}}{\text{CP} + \text{CR}}, \end{aligned} \tag{21}$$

**Table 1**
Comparisons with state-of-the-art methods on the MS-COCO dataset under fully supervised settings. The best results are shown in bold. “-” denotes that the metric was not reported.

| Methods | | All | | | | | | Top-3 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mAP | CP | CR | CF1 | OP | OR | OF1 | CP | CR | CF1 | OP | OR | OF1 |
| ResNet-101 (baseline) [16] | 77.3 | 78.9 | 67.7 | 72.9 | 80.8 | 71.7 | 76.0 | 82.4 | 60.2 | 69.6 | 86.6 | 63.5 | 73.3 |
| ML-GCN[a] [4] | 81.2 | 81.2 | 71.5 | 76.0 | 82.9 | 75.4 | 79.0 | 86.1 | 63.3 | 72.9 | 88.5 | 66.1 | 75.7 |
| KSSNet [18] | 83.7 | 84.6 | 73.2 | 77.2 | 87.8 | 76.2 | 81.5 | – | – | – | – | – | – |
| MS-CMA [19] | 83.8 | 82.9 | 74.4 | 78.4 | 84.4 | 77.9 | 81.0 | 88.2 | 65.0 | 74.9 | 90.2 | 67.4 | 77.1 |
| SSGRL [8] | 83.8 | **89.9** | 68.5 | 76.8 | **91.3** | 70.8 | 79.7 | **91.9** | 62.5 | 72.7 | **93.8** | 64.1 | 76.2 |
| KGGR [20] | 84.3 | 85.6 | 72.7 | 78.6 | 87.1 | 75.6 | 80.9 | 89.4 | 64.6 | 75.0 | 91.3 | 66.6 | 77.0 |
| DA-GAT [21] | 84.8 | 87.0 | 74.2 | 80.1 | 87.3 | 77.5 | 82.1 | 89.2 | 65.6 | 75.6 | 91.6 | 67.7 | 77.9 |
| AdaHGNN[a] [7] | 84.5 | 86.4 | 73.0 | 79.1 | 87.3 | 76.0 | 81.3 | 89.7 | 64.5 | 75.0 | 91.3 | 66.7 | 77.1 |
| C-Tran [22] | 85.1 | 86.3 | 74.3 | 79.9 | 87.7 | 76.5 | 81.7 | 90.1 | 65.7 | 76.0 | 92.1 | **71.4** | 77.6 |
| CCD-R101 [23] | 85.3 | 88.3 | 73.1 | 80.2 | 88.8 | 76.3 | 82.1 | 91.0 | 65.2 | 76.0 | 92.3 | 67.3 | 77.9 |
| DMHL (Ours) | **86.0** | 86.9 | **75.4** | **80.8** | 87.6 | **78.3** | **82.7** | 90.2 | **66.2** | **76.3** | 92.0 | 68.1 | **78.2** |

[a] Indicates that the results are reproduced by using the open source code.

where $C$ is the number of labels, $N_i^c$ is the number of images where the $i$th label is correctly predicted, $N_i^p$ is the number of images where the $i$th label is predicted, and $N_i^g$ is the number of ground truth images for the $i$th label.

We also report average precision (AP) for individual labels and mean average precision (mAP) across all categories, which consider the ranked order of label predictions. Additionally, following recent literature [4,7], we examine top-3 label predictions to account for the variable number of relevant labels per image. Among these metrics, we prioritize mAP, OF1, and CF1 as they provide a more comprehensive evaluation.

### 4.2. Datasets

We utilize four benchmark datasets to validate our DMHL framework: MS-COCO, Pascal VOC 2007, Pascal VOC 2012, and NUS-WIDE.

**MS-COCO** [24] contains 82,081 training images and 40,504 validation images across 80 object categories, averaging approximately 2.9 labels per image. Due to unavailable test set labels, evaluation is conducted on the validation set.

**Pascal VOC 2007** [25] includes 9,963 images over 20 categories. We train on the combined train and validation sets (5,011 images) and test on the test set (4,952 images), reporting average precision (AP) and mean average precision (mAP).

**Pascal VOC 2012** [25] has 11,540 trainval images and 10,991 test images across 20 categories. We follow the same training and evaluation protocol as for VOC 2007.

**NUS-WIDE** [26] consists of 269,648 images with 5,018 tags, annotated for 81 concepts, averaging 2.4 labels per image. Following the standard split [27], we use 161,789 images for training and 107,859 for testing, using small-sized images as per the dataset settings.

### 4.3. Implementation details

**Data Preprocessing**

All images are resized to 576 × 576 pixels. During training, we apply cutout augmentation (randomly masking image regions), random scaling, and photometric distortions to enhance robustness. Images are then converted to tensors and normalized using standard mean and standard deviation values. For evaluation, images are resized to 576 × 576 pixels, converted to tensors, and normalized without augmentation.

**Experimental Settings**

We use the ADAM optimizer with momentum parameters 0.999 and 0.9. Our backbone is a pretrained ResNet-101 on ImageNet [28], with the first two stages frozen. The learning rate starts at $10^{-5}$ and is reduced by a factor of 10 when the training loss plateaus. Labels are represented using 300-dimensional GloVe embeddings; multi-word labels are averaged. Features pass through fully connected layers from 3072 to 2048 to 1 dimension for classification. The DMHL model is trained end-to-end with a batch size of 4.

### 4.4. Comparisons with state-of-the-arts

To evaluate the effectiveness of our DMHL framework, we conduct comprehensive comparisons with contemporary state-of-the-art methods that also use ResNet-101 as a backbone. The comparative analysis is performed across four widely recognized datasets. This section details our framework’s performance relative to leading methods, highlighting its advancements in handling high-order label interdependencies and alleviating the long-tail distribution of labels.

**Results on MS-COCO** The performance of our DMHL framework on the MS-COCO dataset is presented in Table 1, where it outperforms state-of-the-art methods under fully supervised settings. DMHL achieves 86.0% mAP, surpassing the previous SOTA (CCD-R101) by 0.7%, driven by its ability to model high-order label dependencies (e.g., “bird-sky-tree” co-occurrence) and mitigate long-tail bias.

Notably, DMHL improves CF1/OF1 by 0.6% over the closest competitor, and achieves 76.3% CF1 in top-3 predictions, indicating robust label prioritization. Compared to traditional graph-based methods (e.g., ML-GCN: 81.2% mAP), DMHL’s hypergraph structure captures complex relationships, yielding a 4.8% mAP gain. Against HGNN-based AdaHGNN (84.5% mAP), DMHL’s dynamic optimization modules (HyperPrune/HyperTransform/HyperTune) further improve performance by 1.5% mAP. These results validate that dynamic multimodal hypergraph modeling outperforms static graph structures and other advanced architectures like Transformers (C-Tran: 85.1% mAP) and causal models (CCD-R101: 85.3% mAP).

**Results on Pascal VOC 2007** The performance of our DMHL framework on Pascal VOC 2007 is summarized in Table 2. DMHL achieves the highest mAP of 96.3%, surpassing leading methods like ADD-GCN and DA-GAT (both 96.0% mAP). This improvement is driven by DMHL’s ability to capture high-order label interdependencies, such as the co-occurrence of “bird” with “sky” and “tree”.

Notably, DMHL excels in challenging categories like “aero” (99.9% AP) and “bike” (99.0% AP), where fine-grained visual features and complex spatial relationships are required. It also outperforms competitors in categories like “table” (92.9% AP) and “sofa” (90.9% AP), which often co-occur with other furniture items. Compared to traditional graph-based methods (e.g., ML-GCN: 94.0% mAP), DMHL’s hypergraph structure yields a 2.3% mAP gain. Against HGNN-based AdaHGNN (95.2% mAP), DMHL’s dynamic optimization modules improve performance by 1.1% mAP, validating the value of adaptive hypergraph modeling.

**Results on Pascal VOC 2012** On the Pascal VOC 2012 dataset, DMHL achieves the highest mAP of 96.5%, surpassing all competing methods ( Table 3). This improvement is attributed to DMHL’s ability to capture high-order label correlations and adapt to complex visual contexts.

DMHL demonstrates outstanding performance in challenging categories like “aero” (99.9% AP), where hierarchical relationships are critical, and “boat” (98.5% AP), requiring modeling of multi-object

**Table 2**
Comparisons of AP and mAP with state-of-the-art methods on the Pascal VOC 2007 dataset under fully supervised settings. The best results are shown in bold.

| Methods | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | motor | person | plant | sheep | sofa | train | tv | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ResNet-101 (baseline) [16] | 99.1 | 97.3 | 96.2 | 94.7 | 68.3 | 92.9 | 95.9 | 94.6 | 77.9 | 89.9 | 85.1 | 94.7 | 96.8 | 94.3 | 98.1 | 80.8 | 93.1 | 79.1 | 98.2 | 91.1 | 90.8 |
| CNN-RNN [5] | 96.7 | 83.1 | 94.2 | 92.8 | 61.2 | 82.1 | 89.1 | 94.2 | 64.2 | 83.6 | 70.0 | 92.4 | 91.7 | 84.2 | 93.7 | 59.8 | 93.2 | 75.3 | 99.7 | 78.6 | 84.0 |
| RMIC [29] | 97.1 | 91.3 | 94.2 | 57.1 | 86.7 | 90.7 | 93.1 | 63.3 | 83.3 | 76.4 | 92.8 | 94.4 | 91.6 | 95.1 | 92.3 | 59.7 | 86.0 | 69.5 | 96.4 | 79.0 | 84.5 |
| RLSD [30] | 96.4 | 92.7 | 93.8 | 94.1 | 71.2 | 92.5 | 94.2 | 95.7 | 74.3 | 90.0 | 74.2 | 95.4 | 96.2 | 92.1 | 97.9 | 66.9 | 93.5 | 73.7 | 97.5 | 87.6 | 88.5 |
| ML-GCN [4] | 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 |
| SSGRL [8] | 99.7 | 98.4 | 98.0 | 97.6 | 85.7 | 96.2 | 98.2 | 98.8 | 82.0 | 98.1 | 89.7 | 98.8 | 98.7 | 97.0 | 99.0 | 86.9 | 98.1 | 85.8 | 99.0 | 93.7 | 95.0 |
| ADD-GCN [6] | 99.8 | 99.0 | 98.4 | 99.0 | 86.7 | 98.1 | 98.5 | 98.3 | 85.8 | 98.3 | 88.9 | 98.8 | 99.0 | 97.4 | 99.2 | 88.3 | 98.7 | 90.7 | 99.5 | **97.0** | 96.0 |
| DA-GAT [21] | 99.9 | 98.7 | 98.3 | 98.9 | 87.5 | 97.5 | 98.3 | **99.3** | 85.8 | 98.4 | 89.5 | 98.9 | 99.0 | 97.9 | 99.0 | **90.2** | 99.0 | 88.8 | 99.2 | 96.1 | 96.0 |
| TDRG [31] | 99.9 | 98.9 | 98.4 | 98.7 | 81.9 | 95.8 | 97.8 | 98.0 | 85.2 | 95.6 | 89.5 | 98.8 | 98.6 | 97.1 | 99.1 | 86.2 | 97.7 | 87.2 | 99.1 | 95.3 | 95.0 |
| AdaHGNN [7] | 99.8 | 98.9 | **98.9** | 98.1 | **88.1** | 97.3 | 98.3 | 98.8 | 81.7 | 98.2 | 87.2 | 99.1 | 99.3 | 97.7 | 99.1 | 87.8 | 98.4 | 82.5 | 99.7 | 94.5 | 95.2 |
| DMHL (Ours) | **99.9** | **99.0** | 98.4 | **99.0** | 87.7 | **98.2** | **98.5** | 98.4 | **85.9** | **98.6** | **92.9** | **99.2** | **99.3** | **98.1** | **99.2** | 88.9 | **99.1** | **90.9** | **99.7** | 95.5 | **96.3** |

**Table 3**
Comparisons of AP and mAP with state-of-the-art methods on the Pascal VOC 2012 dataset under fully supervised settings. The best results are shown in bold.

| Methods | aero | bike | bird | boat | bottle | bus | car | cat | chair | cow | table | dog | horse | motor | person | plant | sheep | sofa | train | tv | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RMIC [29] | 98.0 | 85.5 | 92.6 | 88.7 | 64.0 | 86.8 | 82.0 | 94.9 | 72.7 | 83.1 | 73.4 | 95.2 | 91.7 | 90.8 | 95.5 | 58.3 | 87.6 | 70.6 | 93.8 | 83.0 | 84.4 |
| VeryDeep [32] | 99.1 | 88.7 | 95.7 | 93.9 | 73.1 | 92.1 | 84.8 | 97.7 | 79.1 | 90.7 | 83.2 | 97.3 | 96.2 | 94.3 | 96.9 | 63.4 | 93.2 | 74.6 | 97.3 | 87.9 | 89.0 |
| FeV+LV [33] | 98.4 | 92.8 | 93.4 | 90.7 | 74.9 | 93.2 | 90.2 | 96.1 | 78.2 | 89.8 | 80.6 | 95.7 | 96.1 | 95.3 | 97.5 | 73.1 | 91.2 | 75.4 | 97.0 | 88.2 | 89.4 |
| HCP [34] | 99.1 | 92.8 | 97.4 | 94.4 | 79.9 | 93.6 | 89.8 | 98.2 | 78.2 | 94.9 | 79.8 | 97.8 | 97.0 | 93.8 | 96.4 | 74.3 | 94.7 | 71.9 | 96.7 | 88.6 | 90.5 |
| RCP [35] | 99.3 | 92.2 | 97.5 | 94.9 | 82.3 | 94.1 | 92.4 | 98.5 | 83.8 | 93.5 | 83.1 | 98.1 | 97.3 | 96.0 | 98.8 | 77.7 | 95.1 | 79.4 | 97.7 | 92.4 | 92.2 |
| SSGRL [8] | 99.7 | 96.1 | 97.7 | 96.5 | 86.9 | 95.8 | 95.0 | 98.9 | 88.3 | 97.6 | 87.4 | 99.1 | 99.2 | 97.3 | 99.0 | 84.8 | 98.3 | 85.8 | 99.2 | 94.1 | 94.8 |
| KGGR [20] | 99.6 | 96.8 | 97.9 | 96.7 | 87.3 | 96.5 | 96.2 | 99.1 | 87.9 | 97.7 | 86.8 | 99.3 | 99.3 | 97.5 | 99.1 | 85.4 | 98.8 | 84.9 | 99.6 | 94.4 | 95.0 |
| ADD-GCN [6] | 99.8 | 97.1 | 98.6 | 96.8 | 89.4 | 97.1 | 96.5 | **99.3** | 89.0 | 97.7 | 87.5 | 99.2 | 99.1 | 97.7 | 99.1 | 86.3 | 98.8 | 87.0 | 99.3 | 95.4 | 95.5 |
| DMHL (Ours) | **99.9** | **98.1** | **99.2** | **98.5** | **90.3** | **97.7** | **97.3** | 99.2 | **92.5** | **98.7** | **89.9** | **99.4** | **99.4** | **98.0** | **99.2** | **87.5** | **98.9** | **89.1** | **99.6** | **96.8** | **96.5** |

interactions. It also outperforms competitors in categories like “bottle” (90.3% AP) and “sofa” (89.1% AP), which often co-occur with other objects. Compared to GNN-based methods like SSGRL (94.8% mAP) and ADD-GCN (95.5% mAP), DMHL’s hypergraph structure yields a 1.7% and 1.0% mAP gain, respectively. These results validate the superiority of dynamic hypergraph optimization in handling diverse and overlapping labels.

**Results on NUS-WIDE** On the NUS-WIDE dataset, DMHL achieves a mAP of 64.0%, surpassing the previous best method, AdaHGNN (62.3%), as shown in Table 4. This improvement is driven by DMHL’s ability to model complex label correlations and mitigate noise in the dataset’s high-dimensional tag system.

DMHL also leads in Composite F1 (CF1: 62.8%) and Overall F1 (OF1: 74.7%), outperforming PLA (CF1: 56.2%, OF1: 72.4%). Compared to graph-based DA-GAT (61.9% mAP), DMHL’s hypergraph structure captures intricate relationships in noisy tags, yielding a 2.1% mAP gain. Against HGNN-based AdaHGNN (62.3% mAP), DMHL’s dynamic optimization modules and pseudo-labeling further improve performance by 1.7% mAP. These results validate that dynamic multi-modal hypergraph modeling and confidence-based pseudo-labeling are effective for handling ambiguous and noisy label systems.

### 4.5. Ablation studies

To evaluate the contributions of each module in DMHL, we perform ablation studies on the MS-COCO dataset, focusing on Multi-modal Hypergraph Construction, Dynamic Hypergraph Structure Optimization, and Hypergraph Residual Concatenation. Additionally, we examine the impact of Dynamic Pseudo-Label Generation under varying label coverage in semi-supervised settings. By disabling specific components, we quantify their influence on performance, providing insights into each module’s role in enhancing multi-label image recognition.

**Effect of Multi-modal Hypergraph Construction** To evaluate the impact of multi-modal hypergraph construction, we conduct ablation studies using different initial hypergraph configurations (Table 5). Replacing the multi-modal hypergraph with a randomly initialized matrix **H** reduces mAP by 0.6% (86.0% → 85.4%), highlighting the importance of structured initialization. Individual modalities (e.g., $\mathbf{H}_v$, $\mathbf{H}'_s$, $\mathbf{H}_t$) yield marginal improvements (≤ 0.3% mAP), indicating that multi-modal integration is critical for capturing complementary information.

**Table 4**
Comparisons with state-of-the-art methods on the NUS-WIDE dataset under fully supervised settings. The best results are shown in bold. “-” denotes that the metric was not reported.

| Methods | mAP | CF1 | OF1 |
|---|---|---|---|
| CNN-RNN [5] | – | 34.7 | 55.2 |
| Att-imagine [36] | 49.9 | 43.9 | 59.3 |
| RLSD [30] | 54.1 | 46.9 | 60.3 |
| CNN+RMLC [37] | 58.8 | 48.8 | 61.8 |
| ResNet101-sem [38] | 60.1 | 47.0 | 61.8 |
| DAN [39] | 61.4 | 59.7 | 61.7 |
| PLA [40] | – | 56.2 | 72.4 |
| CMA [19] | 60.8 | 55.5 | 70.0 |
| MS-CMA [19] | 61.4 | 55.7 | 69.5 |
| DA-GAT [21] | 61.9 | 56.2 | 68.9 |
| AdaHGNN [7] | 62.3 | – | – |
| DMHL (Ours) | **64.0** | **62.8** | **74.7** |

**Table 5**
Ablation study of Multi-modal Hypergraph Construction on the MS-COCO dataset. The best results are shown in bold.

| Methods | | All | | Top-3 | |
|---|---|---|---|---|---|
| | mAP | CF1 | OF1 | CF1 | OF1 |
| DMHL w/ random **H** | 85.4 | 79.9 | 82.2 | 75.8 | 77.9 |
| DMHL w/ $\mathbf{H}_v$ | 85.6 | 80.3 | 82.3 | 76.0 | 78.1 |
| DMHL w/ $\mathbf{H}'_s$ | 85.5 | 80.0 | 82.1 | 75.8 | 77.9 |
| DMHL w/ $\mathbf{H}_t$ | 85.6 | 80.3 | 82.4 | 76.0 | 78.0 |
| Direct Concatenation | 85.8 | 80.6 | 82.6 | 76.2 | 78.2 |
| Adaptive Fusion(Ours) | **86.0** | **80.8** | **82.7** | **76.3** | **78.2** |

Notably, our adaptive fusion method achieves the best results (mAP 86.0%), outperforming direct concatenation (85.8% mAP). This improvement is attributed to dynamic modality weighting, which aligns visual, statistical, and textual cues. These findings demonstrate that adaptive multi-modal fusion provides a robust foundation for subsequent dynamic hypergraph optimization.

**Table 6**
Ablation study of Dynamic Hypergraph Structure Optimization on the MS - COCO dataset. The best results are shown in bold.

| Methods | | All | | Top-3 | |
|---|---|---|---|---|---|
| | mAP | CF1 | OF1 | CF1 | OF1 |
| Random **H** w/o DHSO | 83.9 | 77.6 | 80.8 | 73.6 | 76.7 |
| Random **H** w/ DHSO | 85.4 | 79.9 | 82.2 | 75.8 | 77.9 |
| DMHL w/o hypergraph | 84.3 | 78.1 | 81.1 | 74.0 | 77.1 |
| DMHL w/o DHSO | 85.3 | 79.1 | 81.3 | 75.0 | 77.2 |
| DMHL w/ HyperPrune | 85.6 | 80.3 | 82.4 | 76.0 | 78.1 |
| DMHL w/ HyperTransform | 85.4 | 80.3 | 82.2 | 75.8 | 78.0 |
| DMHL w/ HyperTune | 85.7 | 80.5 | 82.5 | 76.3 | 78.1 |
| DMHL(Ours) | **86.0** | **80.8** | **82.7** | **76.3** | **78.2** |

**Table 7**
Ablation study of hypergraph residual concatenation strategies on the MS-COCO dataset. The best results are shown in bold.

| Methods | | All | | Top-3 | |
|---|---|---|---|---|---|
| | mAP | CF1 | OF1 | CF1 | OF1 |
| $\mathbf{F}_z$ | 82.6 | 76.9 | 80.7 | 73.1 | 76.8 |
| $\mathbf{F}_z$ add $\mathbf{F}_h$ | 82.7 | 77.0 | 80.6 | 73.2 | 76.8 |
| $\mathbf{F}_z$ add $\mathbf{F}_{sd}$ | 83.9 | 78.2 | 81.6 | 74.1 | 77.4 |
| $\mathbf{F}_z$ cat $\mathbf{F}_h$ | 83.8 | 78.3 | 81.6 | 74.6 | 77.6 |
| $\mathbf{F}_z$ cat $\mathbf{F}_{sd}$ | 85.7 | 80.5 | 82.5 | 75.9 | 78.0 |
| $\mathbf{F}_z$ cat $\mathbf{F}_h$ cat $\mathbf{F}_{sd}$ | **86.0** | **80.8** | **82.7** | **76.3** | **78.2** |

**Effect of Dynamic Hypergraph Structure Optimization** We conduct ablation studies on the MS-COCO dataset to evaluate the Dynamic Hypergraph Structure Optimization (DHSO) module and its components (Table 6). A random hypergraph **H** achieves an mAP of 83.9% without DHSO, improving to 85.4% with it, demonstrating DHSO's ability to refine non-optimal structures.

To validate the necessity of hypergraph modeling, we introduce an ablation variant "DMHL w/o hypergraph" (mAP 84.3%), which replaces hyperedges with traditional pairwise edges. This configuration underperforms both the full model (86.0% mAP) and the random hypergraph with DHSO (85.4% mAP), confirming that hypergraph structure is critical for capturing high-order dependencies.

Individually, HyperPrune (mAP 85.6%) prunes redundant hyperedges, HyperTransform (mAP 85.4%) adapts to dynamic label relationships, and HyperTune (mAP 85.7%) fine-tunes label interdependencies. Integrating all components yields the best mAP of 86.0%, highlighting their synergistic effect. These results confirm DHSO's necessity for capturing complex multi-label dynamics and enhancing recognition accuracy.

**Effect of Hypergraph Residual Concatenation** We evaluate hypergraph residual concatenation strategies on the MS-COCO dataset to understand their impact on feature integration (Table 7). Concatenating initial ($\mathbf{F}_{sd}$), intermediate ($\mathbf{F}_h$), and final ($\mathbf{F}_z$) features achieves the best results (mAP 86.0%, CF1 80.8%, OF1 82.7%), outperforming additive strategies. This configuration also excels in Top-3 metrics (CF1 76.3%, OF1 78.2%).

Residual concatenation mitigates over-smoothing by retaining hierarchical information. For example, integrating $\mathbf{F}_{sd}$ (semantic features) and $\mathbf{F}_h$ (dynamic hypergraph features) improves CF1 by 3.9% compared to $\mathbf{F}_z$ alone (76.9% → 80.8%). These results validate that residual concatenation enhances representational power by fusing diverse feature sources.

**Effect of Dynamic Pseudo-Label Generation** The Dynamic Pseudo-Label Generation (DPG) module significantly enhances DMHL's performance on the MS-COCO dataset, particularly in semi-supervised scenarios (Table 8). With 5% labeled data, DMHL w/ DPG achieves 70.5% mAP, improving over DMHL w/o DPG (68.9%) and surpassing traditional GNN-based ML-GCN (46.3%) by 24.2% mAP. This advantage diminishes as labeled data increases, with a 0.3% mAP gain under full supervision.

**Table 8**
Ablation study of dynamic pseudo-label generation under semi-supervised settings on the MS-COCO dataset. The best results are shown in bold.

| Methods | Ratio of labeled data | | | | |
|---|---|---|---|---|---|
| | 5% | 15% | 25% | 50% | 100% |
| ResNet-101(baseline) [16] | 55.5 | 66.1 | 70.4 | 74.1 | 77.3 |
| ML-GCN [4] | 46.3 | 71.0 | 75.2 | 78.9 | 81.2 |
| DMHL w/o DPG | 68.9 | 76.9 | 79.5 | 82.5 | 85.7 |
| DMHL(Ours) | **70.5** | **78.6** | **81.1** | **83.3** | **86.0** |

**Table 9**
Comparisons of computational costs between DMHL and state-of-the-art methods on the MS-COCO dataset.

| Methods | Training time (s) | Testing time (s) | Memory usage (GB) |
|---|---|---|---|
| ResNet-101 (baseline) [16] | 1175.53 ± 21.82 | 239.36 ± 54.28 | 5.5 |
| ML-GCN [4] | 2032.45 ± 24.68 | 395.82 ± 42.15 | 6.7 |
| TDRG [31] | 3294.02 ± 258.55 | 1274.74 ± 34.58 | 9.3 |
| AdaHGNN [7] | 2895.71 ± 28.73 | 571.99 ± 33.14 | 7.1 |
| DMHL (Ours) | 2912.48 ± 24.15 | 582.74 ± 30.56 | 7.8 |

DPG dynamically generates pseudo-labels using confidence-based thresholds, reducing noise from ambiguous annotations. The 1.6% mAP gain with 5% labeled data highlights its effectiveness in leveraging unlabeled data to compensate for sparse annotations. These results validate DPG's critical role in improving semi-supervised performance while maintaining robustness under full supervision.

### 4.6. *Parameter sensitivity analysis*

We analyze hyperparameters in the Dynamic Pseudo-Label Generation module: class imbalance ratio scaling factor $\beta$, warm-up epochs $n$, and unsupervised loss balance coefficient $\lambda$, under 5% and 100% labeled data on MS-COCO.

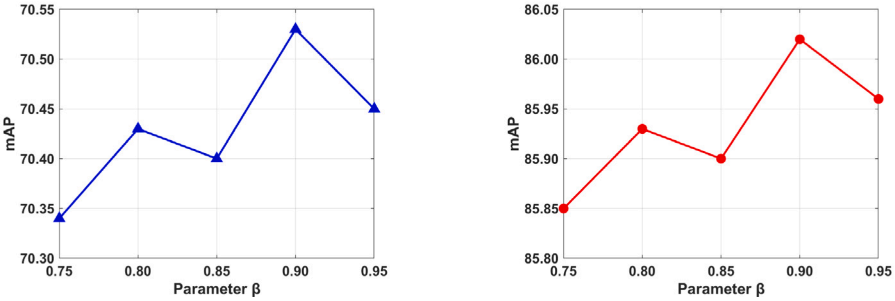
As shown in Fig. 2, $\beta = 0.90$ yields optimal mAP in both sparse (5%) and dense (100%) annotation scenarios, balancing class influence effectively. Fig. 3 reveals $n = 15$ as the optimal warm-up epochs for stabilizing feature representations, with performance declining post-15 epochs in low-label settings. For $\lambda$ (Fig. 4), $\lambda = 0.3$ maximizes mAP under both 5% and 100% labeled data, highlighting its role in tuning unsupervised loss impact.

Overall, $\beta$, $n$ and $\lambda$ significantly affect DMHL's performance, especially in low-label scenarios. This sensitivity analysis validates DMHL's robust adaptability for semi-supervised multi-label recognition, demonstrating fine-grained control over learning dynamics in sparse data environments.

### 4.7. *Computational complexity analysis*

Table 9 compares the computational costs of DMHL with state-of-the-art methods on the MS-COCO dataset. In terms of time complexity, DMHL's total time (training: 2912.48 s, testing: 582.74 s) is longer than ResNet-101 (training: 1175.53 s, testing: 239.36 s) but shorter than the transformer-based TDRG (training: 3294.02 s, testing: 1274.74 s). Although its time is slightly higher than ML-GCN (training: 2032.45 s, testing: 395.82 s) and AdaHGNN (training: 2895.71 s, testing: 571.99 s), the smaller standard deviations (training: ±24.15, testing: ±30.56) indicate greater stability compared to AdaHGNN (training: ±28.73, testing: ±33.14) and ML-GCN (training: ±24.68, testing: ±42.15).

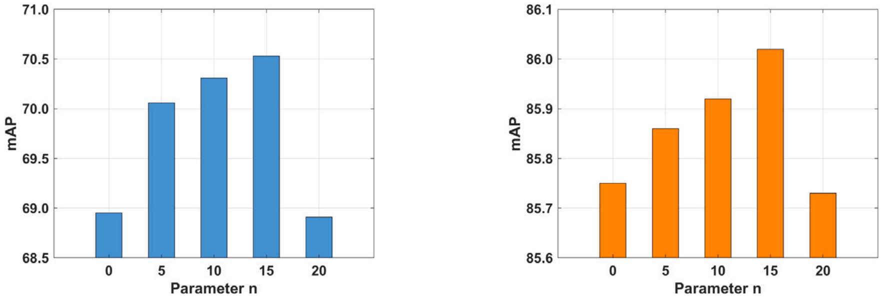
Regarding memory usage, DMHL consumes 7.8 GB, more than ML-GCN (6.7 GB) due to hyperedge processing. However, it is comparable to the HGNN-based AdaHGNN (7.1 GB). This shows that DMHL can maintain stable performance while achieving a balance between accuracy and computational cost.



(a) 5% labeled data

(b) 100% labeled data

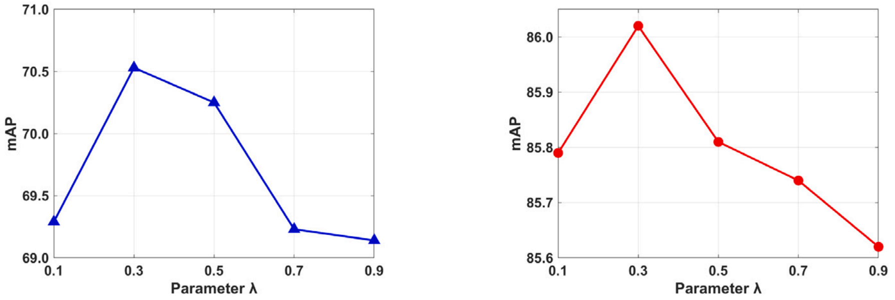
**Fig. 2.** Sensitivity analysis of scaling factor $\beta$ for class imbalance ratio $\eta$ under different labeled data proportions on MS-COCO.



(a) 5% labeled data

(b) 100% labeled data

**Fig. 3.** Impact of warm-up epochs $n$ on DMHL's mAP performance under different labeled data scenarios on MS-COCO.



(a) 5% labeled data

(b) 100% labeled data

**Fig. 4.** mAP variation with balance coefficient $\lambda$ of unsupervised loss on MS-COCO.

### 4.8. Visual annotation results

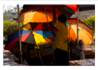
Table 10 compares DMHL with a baseline (ResNet-101), GNN-based (ML-GCN), and HGNN-based (AdaHGNN) methods on three MS-COCO images. DMHL achieves perfect alignment with ground truth across all test cases, eliminating false positives (red) and false negatives (blue). For example, in Image 3, DMHL correctly identifies all 7 ground-truth labels, while the baseline misclassifies "bowl" (red) and misses "person/cup" (blue). The GNN-based ML-GCN reduces errors but still misses "person" (blue), highlighting traditional GNNs' inability to capture high-order label dependencies.

In contrast, the HGNN-based AdaHGNN similarly fails to detect "person" despite eliminating other false positives, demonstrating static hypergraph structures' limitations in handling difficult labels. DMHL addresses these issues through multi-modal hypergraph construction and dynamic structure optimization, which adaptively update hyperedges during training to model complex label interactions. This dynamic design enables DMHL to correctly identify labels like "knife" in Image 2 (missed by AdaHGNN) and "person" in Image 3, validating its effectiveness for accurate multi-label recognition.

## 5. Conclusion

In this paper, we propose Dynamic Multi-modal Hypergraph Learning (DMHL), a semi-supervised framework for multi-label image recognition that explicitly models high-order label correlations and addresses long-tail label imbalance. By fusing visual, statistical, and textual modalities into multi-modal hypergraphs, DMHL captures complex label interactions beyond pairwise relationships. The framework dynamically refines hypergraph structures through three dynamic optimization modules (HyperPrune, HyperTransform, HyperTune) and leverages hypergraph residual concatenation with dynamic pseudo-label generation to enhance feature learning from sparse annotations. Extensive experiments on four benchmark datasets (MS-COCO, Pascal VOC 2007/2012, and NUS-WIDE) validate DMHL's state-of-the-art performance, achieving competitive results in both fully supervised and semi-supervised settings.

**Table 10**
Comparisons of visual annotation results between DMHL and state-of-the-art methods on the MS-COCO dataset. Green: correct; Red: false positive; Blue: false negative.

| Image |  | | |
|---|---|---|---|
| Ground Truth | person, umbrella, bottle, chair | cup, fork, knife, bowl, dining table | person, bottle, cup, fork, knife, pizza, dining table |
| ResNet-101(baseline) [16] | person, umbrella, bottle, chair | cup, fork, knife, bowl, dining table, pizza | person, bottle, cup, fork, knife, pizza, dining table, bowl |
| ML-GCN [4] | person, umbrella, bottle, chair, pizza | cup, fork, knife, bowl, dining table, pizza | person, bottle, cup, fork, knife, pizza, dining table |
| AdaHGNN [7] | person, umbrella, bottle, chair, pizza | cup, fork, knife, bowl, dining table | person, bottle, cup, fork, knife, pizza, dining table |
| DMHL(Ours) | person, umbrella, bottle, chair | cup, fork, knife, bowl, dining table | person, bottle, cup, fork, knife, pizza, dining table |

DMHL's dynamic multi-modal hypergraph modeling holds significant practical potential. In medical imaging, it can simultaneously detect multiple pathologies (e.g., tumors and fractures) and their spatial relationships, improving diagnostic accuracy. For autonomous driving, DMHL could enhance scene understanding by recognizing objects (e.g., pedestrians and traffic signs) alongside contextual interactions (e.g., a pedestrian crossing near a bicycle). These capabilities highlight its value for domains requiring comprehensive visual interpretation.

Despite its advancements, DMHL faces computational challenges due to dynamic hypergraph operations. Future work could focus on developing lightweight hypergraph architectures, such as sparse hyperedge approximations or parameter-sharing strategies, to reduce computational overhead. Integrating self-supervised pre-training could also enhance feature learning while maintaining adaptive hypergraph optimization. Expanding DMHL to video analysis, where temporal label correlations are critical, and adapting it to domains with limited annotations (e.g., remote sensing) are promising directions.

## CRediT authorship contribution statement

**Chen Zhang:** Writing – review & editing, Supervision, Resources, Project administration, Data curation, Conceptualization. **Cheng Xu:** Writing – original draft, Validation, Software, Methodology, Investigation. **Yu Xie:** Writing – review & editing, Supervision, Resources, Project administration, Data curation, Conceptualization. **Wenjie Mao:** Writing – review & editing, Software, Project administration, Conceptualization. **Bin Yu:** Writing – review & editing, Supervision, Resources, Investigation, Funding acquisition.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

This work was supported by the Key Research and Development Program of Shaanxi (Program No. 2022ZDLGY01-13), the National Natural Science Foundation of China (Grant No. 62472270, 62106131), the National Key Research and Development Program of China (Grant No. 2020YFC1523304), the Fund Program for the Scientific Activities of Selected Returned Overseas Professionals in Shanxi Province (Grant No. 20220002), Research Project Supported by Shanxi Scholarship Council of China (Grant No. 2022-007).

## Data availability

Data will be made available on request.

## References

[1] B.-B. Jia, M.-L. Zhang, Multi-dimensional multi-label classification: Towards encompassing heterogeneous label spaces and multi-label annotations, Pattern Recognit. 138 (2023) 109357.

[2] Y. Duan, N. Chen, P. Zhang, N. Kumar, L. Chang, W. Wen, MS2GAH: Multi-label semantic supervised graph attention hashing for robust cross-modal retrieval, Pattern Recognit. 128 (2022) 108676.

[3] M. Minervini, A. Fischbach, H. Scharr, S.A. Tsaftaris, Finely-grained annotated datasets for image-based plant phenotyping, Pattern Recognit. 81 (2016) 80–89.

[4] Z.-M. Chen, X.-S. Wei, P. Wang, Y. Guo, Multi-label image recognition with graph convolutional networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 5177–5186.

[5] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, W. Xu, Cnn-rnn: A unified framework for multi-label image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2285–2294.

[6] J. Ye, J. He, X. Peng, W. Wu, Y. Qiao, Attention-driven dynamic graph convolutional network for multi-label image recognition, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16, Springer, 2020, pp. 649–665.

[7] X. Wu, Q. Chen, W. Li, Y. Xiao, B. Hu, AdaHGNN: Adaptive hypergraph neural networks for multi-label image classification, in: Proceedings of the 28th ACM International Conference on Multimedia, 2020, pp. 284–293.

[8] T. Chen, M. Xu, X. Hui, H. Wu, L. Lin, Learning semantic-specific graph representation for multi-label image recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 522–531.

[9] J. Ma, F. Xu, X. Rong, Discriminative multi-label feature selection with adaptive graph diffusion, Pattern Recognit. 148 (2024) 110154.

[10] Y. Wang, Y. Zhao, Z. Wang, C. Zhang, X. Wang, Robust multi-graph multi-label learning with dual-granularity labeling, IEEE Trans. Pattern Anal. Mach. Intell. (2024).

[11] Z. Gu, S. Fu, D. Wang, S. Xu, Hypergraph transformer for multi-label image classification, in: 2024 International Symposium on Digital Home, ISDH, IEEE, 2024, pp. 79–84.

[12] B. Lu, Q. Fan, X.-l. Zhou, H. Yan, F.-x. Wang, A multimodal multi-label classification method based on hypergraph, Comput. Eng. Sci. 46 (09) (2024) 1667.

[13] B. Guo, H. Tao, C. Hou, D. Yi, Semi-supervised multi-label feature learning via label enlarged discriminant analysis, Knowl. Inf. Syst. 62 (6) (2020) 2383–2417.

[14] M. Stoimchev, J. Levatić, D. Kocev, S. Džeroski, Semi-supervised multi-label classification of land use/land cover in remote sensing images with predictive clustering trees and ensembles, IEEE Trans. Geosci. Remote Sens. (2024).

[15] J. Li, X. Zhu, H. Wang, Y. Zhang, J. Wang, Stacked co-training for semi-supervised multi-label learning, Inf. Sci. 677 (2024) 120906.

[16] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[17] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014, pp. 1532–1543.

[18] Y. Wang, D. He, F. Li, X. Long, Z. Zhou, J. Ma, S. Wen, Multi-label classification with label graph superimposing, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 12265–12272.

[19] R. You, Z. Guo, L. Cui, X. Long, Y. Bao, S. Wen, Cross-modality attention with semantic graph embedding for multi-label classification, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 07, 2020, pp. 12709–12716.

[20] T. Chen, L. Lin, R. Chen, X. Hui, H. Wu, Knowledge-guided multi-label few-shot learning for general image recognition, IEEE Trans. Pattern Anal. Mach. Intell. 44 (3) (2020) 1371–1384.

[21] W. Zhou, Z. Xia, P. Dou, T. Su, H. Hu, Double attention based on graph attention network for image multi-label classification, ACM Trans. Multimed. Comput. Commun. Appl. 19 (1) (2023) 1–23.

[22] J. Lanchantin, T. Wang, V. Ordonez, Y. Qi, General multi-label image classification with transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16478–16488.

[23] R. Liu, H. Liu, G. Li, H. Hou, T. Yu, T. Yang, Contextual debiasing for visual recognition with causal mechanisms, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12755–12765.

[24] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13, Springer, 2014, pp. 740–755.

[25] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2010) 303–338.

[26] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y. Zheng, Nus-wide: a real-world web image database from national university of singapore, in: Proceedings of the ACM International Conference on Image and Video Retrieval, 2009, pp. 1–9.

[27] Z.-M. Chen, X.-S. Wei, P. Wang, Y. Guo, Learning graph convolutional networks for multi-label recognition and applications, IEEE Trans. Pattern Anal. Mach. Intell. 45 (6) (2021) 6969–6983.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[29] S. He, C. Xu, T. Guo, C. Xu, D. Tao, Reinforced multi-label image classification by exploring curriculum, in: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32, no. 1, 2018.

[30] J. Zhang, Q. Wu, C. Shen, J. Zhang, J. Lu, Multilabel image classification with regional latent semantic dependencies, IEEE Trans. Multimed. 20 (10) (2018) 2801–2813.

[31] J. Zhao, K. Yan, Y. Zhao, X. Guo, F. Huang, J. Li, Transformer-based dual relation graph for multi-label image recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 163–172.

[32] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[33] H. Yang, J. Tianyi Zhou, Y. Zhang, B.-B. Gao, J. Wu, J. Cai, Exploit bounding box annotations for multi-label object recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 280–288.

[34] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, S. Yan, HCP: A flexible CNN framework for multi-label image classification, IEEE Trans. Pattern Anal. Mach. Intell. 38 (9) (2015) 1901–1907.

[35] M. Wang, C. Luo, R. Hong, J. Tang, J. Feng, Beyond object proposals: Random crop pooling for multi-label image recognition, IEEE Trans. Image Process. 25 (12) (2016) 5678–5688.

[36] F. Lyu, Q. Wu, F. Hu, Q. Wu, M. Tan, Attend and imagine: Multi-label image classification with visual attention and recurrent neural networks, IEEE Trans. Multimed. 21 (8) (2019) 1971–1981.

[37] H. Cevikalp, B. Benligiray, Ö.N. Gerek, H. Saribas, Semi-supervised robust deep neural networks for multi-label classification, in: CVPR Workshops, 2019, pp. 9–17.

[38] F. Zhu, H. Li, W. Ouyang, N. Yu, X. Wang, Learning spatial regularization with image-level supervisions for multi-label image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5513–5522.

[39] H. Zhao, W. Zhou, X. Hou, H. Zhu, Double attention for multi-label image classification, IEEE Access 8 (2020) 225539–225550.

[40] Y. Ding, Y. Zhou, Y. Zhu, Q. Ye, J. Jiao, Selective sparse sampling for fine-grained image recognition, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 6599–6608.

**Chen Zhang** received the Ph.D. degree in computer application technology from Xidian University, Xi'an, China, in 2012. She is currently an Associate Professor with School of Computer Science and Technology. Her research interests include artificial intelligence and software theory.

**Cheng Xu** was born in 2000. He is currently pursuing his master degree in computer science and technology at Xidian University, Xi'an, China. His research interests include deep learning, graph neural networks, and multi-label image recognition.

**Yu Xie** received the Ph.D. degree in pattern recognition and intelligent systems from Xidian University, Xi'an, China, in 2020. From 2019 to 2020, he was a research assistant with School of Computer Science and Engineering, Nangyang Technological University. He is currently an Associate Professor with School of Computer and Information Technology, Shanxi University. His research interests include graph neural networks and secure artificial intelligence.

**Wenjie Mao** was born in 1997. He is currently pursuing Ph.D. degree in computer science and technology, Xidian University, Xi'an, China. His research interests include deep learning, federated learning and representation learning.

**Bin Yu** received the Ph.D. degree in computer software and theory from Northwest University, Xi'an, China, in 2003. He is currently a Full Professor with School of Computer Science and Technology. His research interests include artificial intelligence and information security.