Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.059102

ARTICLE

Multi-Scale Feature Fusion and Advanced Representation Learning for Multi Label Image Classification

Naikang Zhong[1], **Xiao Lin**[1,2,3,4,*], **Wen Du**[5] **and Jin Shi**[6]

[1]Institute of Artificial Intelligence on Education Research, College of Information, Mechanical and Electrical Engineering, Shanghai Normal University, Shanghai, 200234, China
[2]Lab for Educational Big Data and Policymaking, Ministry of Education, Shanghai Normal University, Shanghai, 200234, China
[3]Shanghai Intelligent Education Big Data Engineering Technology Research Center, Shanghai Normal University, Shanghai, 200234, China
[4]Shanghai Online Education Research Base for Primary and Secondary Schools, Shanghai, 200234, China
[5]DS Information Technology Co., Ltd., Shanghai, 200032, China
[6]Faculty of Innovation Engineering, Macau university of Science and Technology, Macau, 999078, China

*Corresponding Author: Xiao Lin. Email: lin6008@shnu.edu.cn

Received: 28 September 2024; Accepted: 10 January 2025; Published: 06 March 2025

ABSTRACT: Multi-label image classification is a challenging task due to the diverse sizes and complex backgrounds of objects in images. Obtaining class-specific precise representations at different scales is a key aspect of feature representation. However, existing methods often rely on the single-scale deep feature, neglecting shallow and deeper layer features, which poses challenges when predicting objects of varying scales within the same image. Although some studies have explored multi-scale features, they rarely address the flow of information between scales or efficiently obtain class-specific precise representations for features at different scales. To address these issues, we propose a two-stage, three-branch Transformer-based framework. The first stage incorporates multi-scale image feature extraction and hierarchical scale attention. This design enables the model to consider objects at various scales while enhancing the flow of information across different feature scales, improving the model's generalization to diverse object scales. The second stage includes a global feature enhancement module and a region selection module. The global feature enhancement module strengthens interconnections between different image regions, mitigating the issue of incomplete representations, while the region selection module models the cross-modal relationships between image features and labels. Together, these components enable the efficient acquisition of class-specific precise feature representations. Extensive experiments on public datasets, including COCO2014, VOC2007, and VOC2012, demonstrate the effectiveness of our proposed method. Our approach achieves consistent performance gains of 0.3%, 0.4%, and 0.2% over state-of-the-art methods on the three datasets, respectively. These results validate the reliability and superiority of our approach for multi-label image classification.

KEYWORDS: Image classification; multi-label; multi scale; attention mechanisms; feature fusion

1 Introduction

In a single-label image classification task, each image is assigned only one most relevant label, such as cat, dog, or airplane. However, in the real world, an image often contains multiple objects. This makes multi-label image classification not only more challenging but also more relevant to practical scenarios. In particular, addressing multi-label classification is crucial for complex real-world applications where multiple objects or attributes co-exist within a single image. Such tasks require models to capture richer contextual information

and manage intricate relationships among labels. Currently, multi-label image classification has been widely applied in various fields such as medical diagnosis recognition [1,2], object detection [3], image retrieval [4,5], and image processing [6]. For instance, in medical diagnosis, identifying multiple co-existing diseases from an X-ray image is critical for comprehensive treatment, while in autonomous driving, detecting multiple objects such as vehicles, pedestrians, and traffic signs is crucial for decision-making. These applications highlight the far-reaching implications of effective multi-label image classification. At the same time, multi-label image classification presents significant challenges. Currently, there are three major challenges in this field: (i) efficiently establishing the relationship between image features and labels, particularly in capturing high-level feature representations across different scales; (ii) effectively leveraging label correlations to enhance classification performance through contextual information; (iii) an overemphasis on the single deep feature, with limited attention to multi-scale features, which are vital for recognizing objects of varying sizes within an image.

The early work on multi-label image classification transformed the problem into a single-label image classification task, primarily employing region-based approaches [7,8]. For example, they leveraged the convolutional layers of convolutional neural networks (CNNs) to extract features from input images, generating corresponding feature maps. These feature maps were then used to predict and determine the presence of target categories within the images.

With the introduction of attention mechanisms, researchers attempted to integrate CNNs with attention mechanisms. For example, the following works [9,10] capitalized the robust feature extraction capabilities of CNNs, while they simultaneously utilized the attention mechanism's capacity to select information focus. This way, the model performance is improved significantly. These methods address the challenge of feature representation to a certain extent and are capable of establishing associations between local features and labels. However, by relying solely on traditional CNNs and attention mechanisms, they struggle to capture global image information. As a result, it becomes difficult to capture the correlations between different regions of the same image, and the contextual information of the image remains elusive. Consequently, the obtained feature representations are incomplete. To address this issue, this paper introduces a Global Feature Enhancement Module, designed through the integration of encoders, to compensate for the lack of global features and the incomplete feature representations found in these approaches.

On the other hand, with the development of graph convolutional networks (GCNs), some other researchers focused on leveraging GCNs for multi-label image classification [11–13]. GCNs typically use the graph structures to model the relationships between different labels in the image. Most of these methods explicitly model label correlations using graph structures, which can help leverage label dependencies to some extent. However, there are two key challenges with this approach: (1) Label co-occurrence can lead to biased graph structures, especially in small or imbalanced datasets, affecting the model's generalization. (2) Some label relationships are complex and nonlinear, making them difficult to represent with simple graph structures. To address the challenge of effectively leveraging label correlations, this paper does not use graph structures for label correlation modeling. Instead, we learn label embeddings and use the Transformer's cross-attention mechanism to implicitly capture complex label relationships, avoiding the biases and limitations of graph structures.

In addition, we observe that many mainstream methods highly rely on a single deep feature, particularly those extracted before the max pooling layer in ResNet-101. However, objects within an image vary in scale, necessitating the use of multi-scale feature maps to capture both small and large objects effectively. This is particularly critical in real-world applications, such as traffic monitoring, where vehicles and pedestrians differ greatly in size, or in satellite imagery, where both expansive landscapes and small buildings are present in the same frame. Therefore, relying on a single deep feature for all objects is inadequate. Existing studies,

such as FL-Tran [14] and DATran [15], have addressed critical issues in multi-scale feature handling for multi-label image classification by incorporating multi-scale feature fusion and attention mechanisms. These approaches effectively address challenges like identifying small objects in images and extracting useful features that may be obscured by more dominant ones. However, these approaches still face limitations, such as how to enhance the flow of information across different scales and how to maintain spatial correlations. To address these challenges, our work introduces a multi-branch Transformer model with the following contributions: We incorporate a hierarchical scale attention module to strengthen the information flow between features of different scales. Additionally, by integrating encoders, the model establishes correlations between different regions of the image, significantly improving its ability to capture spatial dependencies. Furthermore, we propose a novel three-branch structure, specifically designed for different scale features, which employs lightweight cross-attention mechanisms to accurately capture feature representations of objects at varying scales, which is an aspect not covered in the aforementioned literature.

To this end, this paper presents a novel multi-scale feature fusion model for multi-label image classification. The model leverages a three-branch structure to capture relationships between objects of varying scales and correlations between features across different image regions. Initially, we incorporate hierarchical scale attention to facilitate information exchange between features of different scales, effectively combining high-level semantic information with low-level details. A global feature enhancement module is then introduced, utilizing Transformer encoders for robust global feature extraction, thereby uncovering intrinsic connections across image regions. Finally, a lightweight cross-attention mechanism is employed to refine feature representations for specific object categories at different scales, leveraging the rich features obtained from the global enhancement module. Simply speaking, in this work we primarily address three key issues to enhance the capabilities of our model: (i) We conducted multi-scale image feature extraction and designed a three-branch structure tailored to different feature scales, addressing the common limitation of relying solely on a single deep feature in most existing methods. Additionally, we introduced a hierarchical scale attention module to enhance the flow of information across scales, enabling more effective cross-scale feature integration. (ii) By learning label embeddings and employing the cross-attention mechanism of Transformers, we implicitly captured complex label relationships. This approach not only mitigates the biases and constraints of graph-based structures but also enables the model to acquire precise class-specific feature representations. (iii) We incorporated multiple encoders and developed a global feature enhancement module, significantly improving the connectivity of information across different image regions. This design enhances the model's global feature extraction capability, addressing the incomplete feature representations inherent in traditional convolutional and attention-based methods.

To summarize, the main contributions of this work are as follows:

- We propose a novel three-branch multi-scale Transformer model for multi-label image classification. Specifically, we introduce a hierarchical scale attention to facilitate the flow of feature information across different scales and integrate it into the Transformer architecture. This design enables the model to capture class-specific feature representations at various scales, significantly improving classification performance.
- We revisit the Transformer architecture and design a global feature enhancement module via integrated encoders, which establishes correlations between features across different regions of the image. Additionally, we simplify the decoder structure by employing the lightweight cross-attention mechanism to efficiently capture cross-modal relationships between image features at different scales and textual label information.

- We conducted extensive experiments on several widely used benchmarks including COCO 2014, VOC 2007, and VOC 2012. The results demonstrate that our proposed model can achieve competitive performance, compared against a set of state-of-the-art models.

2 Related Work

2.1 Multi-Label Classification

In recent years, the challenge of multi-label image classification has garnered significant attention. Proposed methods generally fall into three main categories: (1) Locating regions of interest, (2) Label correlation, and (3) Multi-scale feature fusion.

2.1.1 Locating Regions of Interest

Early works [7,8] leveraged the powerful feature extraction capabilities of convolutional neural networks (CNNs) to obtain regions of interest (ROI) corresponding to specific categories from input image. With the advent of attention mechanisms, some studies [9,16–19] began to integrate CNNs with attention mechanisms to emphasize and focus on important local features. For example, Guo et al. [9] proposed a dual-branch network and attention consistency loss to precisely locate ROI regions through attention consistency. You et al. [16] introduced cross-modal attention and semantic graph embedding for multi-label classification. Liu et al. [17] employed Transformer decoders to adaptively find ROI regions within images. Ridnik et al. [19] redesigned the decoder architecture and introduced a novel grouped decoding method. However, existing methods typically rely on CNNs for feature extraction and use attention mechanisms to focus on local features. These approaches may fail to fully leverage global features and the relationships between different regions, resulting in incomplete feature representations.

2.1.2 Label Correlation

With the rise of GCNs [20], some studies [11–13] have begun to utilize graph structures to establish correlations between labels, providing additional semantic information and demonstrating strong generalization capabilities. For instance, Chen et al. [13] constructed a directed graph over object labels using GCNs and proposed a novel re-weighting scheme to create an effective label correlation matrix. Ye et al. [11] proposed an attention-based dynamic GCN that decomposes input features into category-specific representations and models their interactions. However, label co-occurrence can introduce bias into graph structures, particularly in small or imbalanced datasets, which in turn affects the model's ability to generalize. Furthermore, certain label relationships are complex and nonlinear, making it challenging to accurately represent them with simplistic graph-based approaches.

2.1.3 Multi-Scale Feature Fusion

Multi-label image classification typically involves the classification of multiple objects, which may exist at different scales. Consequently, some studies [15,21] have focused on using multi-scale feature fusion methods to address multi-label image classification tasks. Zhou et al. [15] proposed an innovative Dual Attention Transformer model, which effectively captures both high-level semantics and low-level details in images through multi-scale feature fusion and a dual-stream architecture. Ye et al. [21] proposed a model on multi-scale fusion and adaptive label correlations. This model enhances the feature information of small targets by fusing multi-scale feature maps. Beisdes, it utilizes a graph attention network to adaptively explore category correlations within the image. Current approaches often prioritize feature fusion but fail to address

the flow of information between features of different scales and the accurate class-specific representation for each particular scale.

2.2 Application of Transformers in Visual Tasks

Early Transformer models were predominantly utilized in the domain of natural language processing [22–24]. Recently, the Transformer architecture has gained widespread application in the field of computer vision [25–28]. For example, Alexey [25] introduced the Transformer architecture into the field of computer vision, proposing the Vision Transformer. Liu et al. [26] proposed the Swin Transformer, which employs hierarchical window attention by applying attention mechanisms within local patches and progressively enlarging the window size at each layer. These models have demonstrated substantial improvements in accuracy and efficiency for image classification tasks.

Recently, some studies [17,19] have begun to employ Transformers for multi-label image classification. However, these methods have some issues: (1) They focused on leveraging Transformer decoders to establish correlations between image features and labels. However, the decoder does not effectively distinguish between objects of different scales when querying the ROI regions. (2) The encoder, as a powerful global feature extractor, is often overlooked in its role. Compared to these methods, our model employs multi-scale feature fusion and hierarchical scale attention to differentiate objects of varying sizes. This approach aids the feature-label interaction module in accurately locating category-specific regions of interest for different scales. Additionally, we utilize a global feature enhancement module to enrich the model with valuable global information.

3 Proposed Model

In this section, we introduce a novel approach for multi-label image classification, which addresses key challenges such as multi-scale feature learning, label correlation, and class-specific precise feature representations. Our model consists of two stages, as illustrated in Fig. 1.

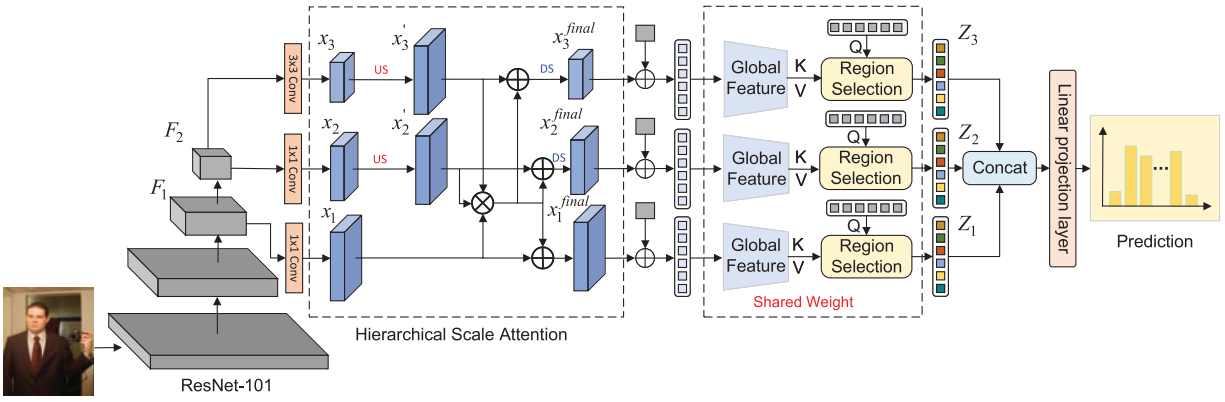

Figure 1: Overview of our framework. ResNet101 and Hierarchical Scale Attention first perform multi-scale feature extraction and cross-scale information flow. Feature reshaping is then applied to incorporate positional encoding. The subsequent global feature enhancement module and region selection module efficiently extract global information and generate class-specific feature representations for each branch. Finally, a linear projection layer and sigmoid function produce the final prediction. "US" represents the upsampling operation, and "DS" represents the downsampling operation

The first phase, as shown on the left side of Fig. 1, comprises two components: (i) Multi-scale Feature Extraction and (ii) Hierarchical Scale Attention. We begin by employing the ResNet101 network [29] for multi-scale feature extraction, generating feature maps at large, medium, and small scales. This enables the model to handle objects of varying sizes more effectively-larger feature maps capture fine-grained details of small objects, while smaller maps are better suited for larger objects. To overcome the common limitation in existing methods that neglect the information flow between features of different scales, our approach integrates multi-scale feature maps using a hierarchical scale attention module, thereby enhancing the flow of information across scales. This design ensures more effective cross-scale feature fusion, improving the model's overall performance in multi-scale object recognition.

The second stage includes three components: (i) Feature Reshaping, (ii) Global Feature Enhancement Module, and (iii) Region Selection Module. Feature reshaping aligns the dimensions and shapes of the extracted features with the requirements of subsequent modules. The Global Feature Enhancement Module improves the model's ability to establish global connections across different image regions, thereby addressing the incomplete feature representations often seen in traditional convolutional and attention-based methods. In the Region Selection Module, we employ multiple cross-attention mechanisms to capture class-specific features based on learnable label embeddings, which implicitly model complex label relationships. This approach not only mitigates biases present in graph-based methods but also allows for more precise class-specific feature representation, significantly enhancing the model's ability to handle complex multi-label classification tasks.

Finally, the outputs from the region selection module are concatenated along the channel dimension and passed through a linear projection layer, producing the final prediction scores. The model's efficiency is further enhanced by using weight-sharing across the three branches, reducing the number of parameters while maintaining high accuracy.

3.1 Multi-Scale Feature Extraction

We first employ ResNet101 for initial feature extraction. Specifically, we obtain the output features from Layer3 and Layer4, which are denoted as $F_1 \in \mathbb{R}^{\frac{d_{\text{model}}}{2} \times 2H \times 2W}$ and $F_2 \in \mathbb{R}^{d_{\text{model}} \times H \times W}$, respectively. Here, d_{model} represents the number of channels, while H and W denote the height and width of the feature maps.

Existing approaches often rely solely on a single deep feature (e.g., features from F_2). To effectively capture multi-scale features, we aim to construct feature maps larger than F_2 for detecting small objects and smaller than F_2 for detecting large objects. To this end, we propose utilizing features from F_1 (which provides a larger feature map than F_2) and applying convolutional operations to F_2 (to generate a smaller feature map) to extract features at different scales. Subsequently, the following operations are applied to F_1 and F_2:

$$\begin{cases} x_1 = \text{Conv}_{1\times1}(F_1), & x_1 \in \mathbb{R}^{\frac{d_{\text{model}}}{4} \times 2H \times 2W}, \\ x_2 = \text{Conv}_{1\times1}(F_2), & x_2 \in \mathbb{R}^{\frac{d_{\text{model}}}{4} \times H \times W}, \\ x_3 = \text{Conv}_{3\times3,\,\text{stride}=2}(F_2), & x_3 \in \mathbb{R}^{\frac{d_{\text{model}}}{4} \times \frac{H}{2} \times \frac{W}{2}}. \end{cases} \tag{1}$$

where 1 × 1 convolution serves to adjust the number of channels, while the 3 × 3 convolution not only modifies the channel dimensions but also further extracts smaller feature maps, enabling the prediction of larger objects.

These convolutional operations produce three feature maps x_1, x_2, and x_3 at different scales. The channel dimensions are unified to $\frac{d_{\text{model}}}{4}$, which aligns with the hidden dimension required by subsequent transformer encoders. These multi-scale feature maps facilitate the prediction of objects of varying sizes, with each scale corresponding to a distinct branch in the model.

3.2 Hierarchical Scale Attention

To address the challenge of effectively integrating features at different scales and enhancing the flow of information between features at various scales, we introduce a hierarchical scale attention module. This module aims to dynamically enhance the representation of multi-scale features by leveraging interpolation-based transformations and element-wise feature interactions. Through this design, features from different scales are enriched with complementary information, enabling the model to better capture both local and global image contexts.

3.2.1 Multi-Scale Feature Alignment

After extracting multi-scale features using ResNet101, we denote the feature maps at three scales as x_1, x_2, and x_3. Here, x_1 corresponds to the largest scale feature map with spatial dimensions $h_{\max} \times h_{\max}$. To align feature maps spatially, bilinear interpolation is applied to upsample x_2 and x_3 to match the resolution of x_1:

$$\begin{aligned} x_2' &= \text{Interpolate}\,(x_2, (h_{\max}, h_{\max})), \\ x_3' &= \text{Interpolate}\,(x_3, (h_{\max}, h_{\max})). \end{aligned} \tag{2}$$

where $\text{Interpolate}\,(x, (h, w))$ indicates that the tensor x is interpolated to a target size of $h \times w$. This operation ensures that features from different scales are spatially consistent, facilitating cross-scale information interaction.

3.2.2 Cross-Scale Feature Fusion

To capture common patterns across scales, element-wise multiplication is performed on the aligned feature maps x_1, x_2', and x_3':

$$m = x_1 \odot x_2' \odot x_3', \tag{3}$$

where $\odot$ represents element-wise multiplication. The resulting tensor m highlights regions with strong responses across all three scales, serving as a shared representation.

Subsequently, m is added back to each scale's feature map to integrate the shared information while preserving scale-specific characteristics:

$$\begin{aligned} x_1^{\text{updated}} &= x_1 + m, \\ x_2^{\text{updated}} &= x_2' + m, \\ x_3^{\text{updated}} &= x_3' + m. \end{aligned} \tag{4}$$

3.2.3 Scale Restoration

Since x_2^{updated} and x_3^{updated} are obtained through upsampling, they are downsampled back to their original resolutions using bilinear interpolation:

$$\begin{aligned} x_2^{\text{final}} &= \text{Interpolate}\left(x_2^{\text{updated}}, (h_2, h_2)\right), \\ x_3^{\text{final}} &= \text{Interpolate}\left(x_3^{\text{updated}}, (h_3, h_3)\right). \end{aligned} \tag{5}$$

here, $x_1^{\text{final}} = x_1^{\text{updated}}$, as no scale transformation is applied to x_1.

This hierarchical fusion process not only preserves the original scale-specific feature but also effectively integrates complementary information from other scales. By dynamically combining multi-scale feature representations, the model enhances its capacity to perceive diverse objects across different scales in complex scenes. The enriched feature map captures both local and global contexts, improving the robustness and generalization of the model while strengthening the relationships between multi-scale features.

3.3 Feature Reshaping

To incorporate positional information into the image features and meet the input requirements of the global feature enhancement module, feature reshaping is essential. The feature reshaping process primarily consists of two steps: (i) adding learnable positional encodings and (ii) flattening the image features into sequential features, as illustrated in Fig. 2.

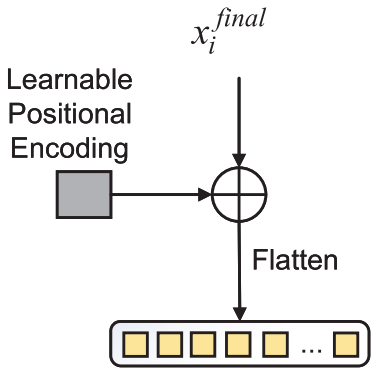

Figure 2: The process of feature reshaping. The input feature x_i^{final} is transformed from a three-dimensional tensor $d_{model} \times H \times W$ to a two-dimensional tensor $d_{model} \times L$, where $L = H \times W$

Learnable positional encoding is a technique for modeling positional information in deep learning models. Unlike traditional absolute positional encodings based on sinusoidal functions, learnable positional encoding does not rely on a predefined mapping from positions to vectors. Instead, it learns vector representations for each position during the training process, providing the model with more flexible and adaptive positional information.

For a given position pos, the corresponding positional encoding PE_{pos} is a trainable parameter. The integration of the positional encoding with the image feature vector x_i^{final} is performed through a simple vector addition:

$$F_{\text{combined}} = x_i^{final} + PE_{\text{pos}} \tag{6}$$

where F_{combined} represents the image feature vector that has been augmented with positional information.

The flattening process is primarily used to reshape the image features from a three-dimensional tensor $d_{model} \times H \times W$ to a two-dimensional tensor $d_{model} \times L$, where $L = H \times W$. The reshaped features Y can be obtained using Eq. (7):

$$Y = Reshape\left(F_{\text{combined}}\right) \tag{7}$$

where $Y \in \mathbb{R}^{d_{model} \times L}$ denotes the reshaped sequential features, $Reshape$ denotes the process of the flattening process.

3.4 Global Feature Enhancement Module

To provide the model with stronger global context, we introduces a Global Feature Enhancement Module to establish relationships between different regions in the image. The Global Feature Enhancement Module consists of two transformer encoders, which leverage self-attention mechanisms to enhance the model's ability to establish the relationships between different regions within the image. The key component of this module is the multi-head self-attention mechanism. Fig. 3 illustrates the workflow of the global feature enhancement module and the multi-head self-attention module.

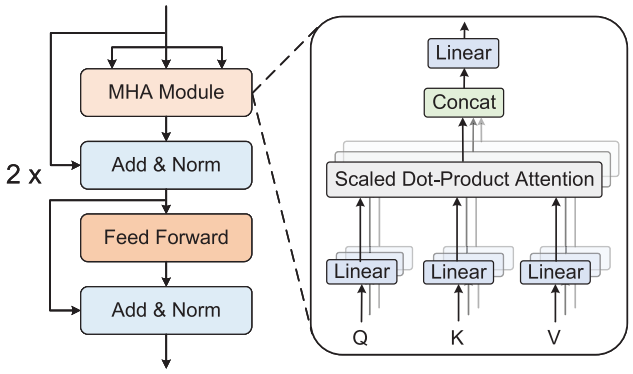

Figure 3: Global feature enhancement module and self attention module

Multi-head self-attention is the result of concatenating the outputs of multiple self-attention mechanisms. Specifically, multi-head self-attention is defined as follows:

$$\text{MultiHead}\left(Q,\ K,\ V\right) = \text{Concat}\left(\text{head}_1, \ldots,\ \text{head}_h\right) W^O \tag{8}$$

where $\text{head}_i = \text{Attention}\left(QW_i^Q,\ KW_i^K,\ VW_i^V\right)$, Q, K, V respectively denote the linear transformation matrices corresponding to Q, K and V. W^O is the linear transformation matrix corresponding to the concatenated long vector, which is used to map the concatenated vector back to the model's dimensional space.

The multi-head self-attention layer, as the first sub-layer of the Transformer encoder, is primarily used for learning global image feature information. The second sub-layer is a feedforward fully connected layer, which enhances the model's expressive power and ability to handle complex features through non-linear transformations. The feedforward fully connected layer is defined as:

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{9}$$

where W_1, W_2, b_1, b_2 are the trainable weights and biases. Each sublayer is followed by a normalization layer and a residual connection, which are used to adjust the distribution of the output, thereby obtaining enhanced image feature information for each branch.

3.5 Region Selection Module

To obtain class-specific feature representation, establishing the relationship between image regions and label information is essential. Cross-attention, as a key component of the Transformer decoder, has a strong capability to establish the relationship between label information and image regions. For this purpose, we introduce a Region Selection Module, which consists of three simplified transformer decoders (as shown

in Fig. 4). Its function is to utilize cross attention to establish the relationship between the ROI regions and the labels, which improved model's representation ability.

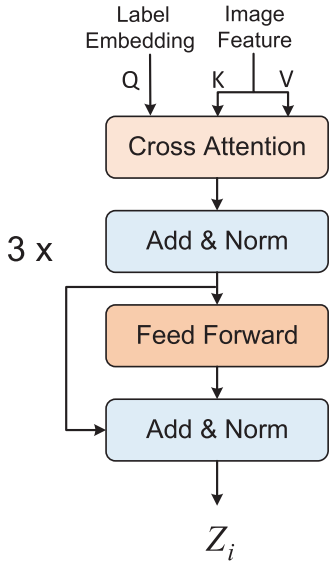

Figure 4: Region selection module. The region selection module consists of three simplified Transformer decoders, where the learnable label embeddings serve as the query, and the enhanced image features are used as the key and value

The input to the region selection module consists of two parts: learnable label embeddings (Query) and image features (Key and Value). The image features are derived from the output of the global feature enhancement module. The computation of cross-attention is essentially the same as that of multi-head self-attention, with the primary difference being the source of the input. Specifically, Eq. (10) represents the update process of the query in the region selection module:

$$Q_j = \text{MultiHead}\left(Q_{j-1},\ F_t,\ F_t\right) \tag{10}$$

where Q_j denotes the query input, when $j = 1$, Q_0 denotes the original label embedding, F_t represents the global image feature obtained through the global feature enhancement module.

Similar to the encoder part, after computing the cross attention, the output distribution is adjusted through a normalization layer and a feed forward neural network, resulting in the final output Z_i for each branch. $Z_i \in \mathbb{R}^{K\times d}$ denotes the final output for the ith branch.

It is important to note that a standard transformer decoder typically includes a self-attention mechanism. However, before computing cross-attention, the input features undergo a linear transformation, which allows the input to adaptively fit the expected output of the model. Therefore, the decoder in this paper does not include a self-attention module.

3.6 Final Classification and Loss Function

Final Classification: To fuse multi-scale feature information and enable the model to better distinguish objects of different sizes in the images, the final feature information from the three branches is concatenated:

$$Z' = \text{Concat}\left(Z_1,\ Z_2,\ Z_3\right) \tag{11}$$

where $Z' \in \mathbb{R}^{K\times 3d}$ represents the result of concatenating the outputs of each branch along the channel dimension. d represents the channel dimension of each branch. K denotes the number of classes. Finally, a linear projection layer followed by a sigmoid function is used to obtain the prediction scores corresponding to each class.

Loss Function: The traditional binary cross-entropy loss function can be applied to this framework. To address the issue of positive-negative sample imbalance, the Asymmetric Loss function is adopted. The Asymmetric Loss function is defined as:

$$Loss = \frac{1}{C}\sum_{c=1}^{C}\begin{cases} (1-p_c)^{\alpha^+}\log(p_c), & y_c = 1, \\ (p_c)^{\alpha^-}\log(1-p_c), & y_c = 0, \end{cases} \tag{12}$$

where y_c represents the true label of the ith sample, p_c represents the predicted probability of the ith sample being in the positive class, α^+ and α^- are the hyperparameters. The default settings are $\alpha^+ = 0$ and $\alpha^- = 4$.

4 Experiment

4.1 Dataset and Competitors

4.1.1 Dataset

To validate the effectiveness of the proposed model, experiments were conducted on the multi-label image classification public datasets: MS-COCO and PASCAL-VOC.

MS-COCO is a large-scale dataset that has been widely used in recent years for evaluating multi-label image classification. This paper uses the COCO 2014 version of the dataset, which contains 82,783 training images and 40,775 test images, with a total of 80 label categories. On average, each image has 2.9 labels.

PASCAL VOC is a large-scale image dataset. This paper conducts experiments primarily on two versions: VOC2007 and VOC2012.

(1) The VOC2007 dataset comprises 5011 training images and 4952 test images, covering 20 common object categories. The average number of labels per image is 1.6.

(2) The VOC2012 dataset includes 11,540 training images and 10,991 test images, covering 20 common object categories. Each image has an average of 1.4 labels.

4.1.2 Competitors

To examine the competitiveness of our proposed method, we compared it with state-of-the-art methods (e.g., MSFA (2024) [30], DRGN (2024) [31], C-TMS (2024) [32], ML-AGCN (2024) [33], FL-Tran (2023) [14], DATran (2023) [15], IDA (2023) [34], DA-GAT (2023) [35], MulCon (2023) [36], etc.) These competitors can be categorized into three classes:

- **Region of Interest based methods:** ResNet101 [29], SRN [37], KSSNet [38], Q2L [17], SRDL [39], IDA [34], HCP [7], RNN-Att [18], RARL [40], VeryDeep [41], Fev+Lv [8], MCAR [42], and C-TMS [32].
- **Label Correlation methods:** CADM [43], ML-GCN [13], MS-CMA [16], CCD [10], MulCon [36], SSGRL [44], ADD-GCN [11], C-Trans [45], TDRG [46], DA-GAT [35], CNN-RNN [47], KGGR [48], P-GCN [49], LDR [50], CPCL [51], FL-Net [52], DSDL [53], and ML-AGCN [33].
- **Multi-scale feature fushion:** FL-Tran [14], DRGN [31], DATran [15], MSFA [30], and MS-SGA [54].

4.2 Evaluation Metrics

In terms of evaluation metrics, this paper utilizes Average Precision (AP) for each individual class and Mean Average Precision (mAP) for the evaluation across all classes:

$$mAP = \frac{\sum_{k=1}^{K} AP_k}{K} \tag{13}$$

where k denotes the current specific class and K represents the total number of classes. In addition, in multi-label image classification, metrics such as Class Precision (CP), Class Recall (CR), and Class F1 Score (CF1), as well as Overall Precision (OP), Overall Recall (OR), and Overall F1 Score (OF1) are commonly used as supplementary evaluation metrics. The definitions of these supplementary metrics are as follows:

$$\mathrm{CP} = \frac{1}{K} \sum_{l} \frac{M_k^l}{M_p^l} \tag{14}$$

$$\mathrm{CR} = \frac{1}{K} \sum_{l} \frac{M_k^l}{M_g^l} \tag{15}$$

$$\mathrm{CF1} = \frac{2 \times \mathrm{CP} \times \mathrm{CR}}{\mathrm{CP} + \mathrm{CR}} \tag{16}$$

$$\mathrm{OP} = \frac{\sum_l M_k^l}{\sum_l M_p^l} \tag{17}$$

$$\mathrm{OR} = \frac{\sum_l M_k^l}{\sum_l M_g^l} \tag{18}$$

$$\mathrm{OF1} = \frac{2 \times \mathrm{OP} \times \mathrm{OR}}{OP + OR} \tag{19}$$

4.3 Implement Details

The experiments in this paper were conducted on a platform running Ubuntu 18.04 with an Intel Xeon Gold 6330 @ 2.0 GHz processor (60 cores), 240 GB of RAM, and six Nvidia GeForce RTX 3090 GPUs.

The network model was built using the Pytorch [55] deep learning framework, and the AdamW [56] optimizer was used for optimization with an initial learning rate of 1×10^{-4} and a weight decay rate of 1×10^{-2}. Data augmentation techniques including Randaugment [57] and Cutout [58] were applied. The model was trained for a total of 80 epochs.

Our method employs a multi-branch Transformer architecture with multi-scale feature fusion, resulting in a model with 69.5 million parameters. This complexity enables rich feature representation but may challenge deployment on resource-constrained devices. Training requires significant resources; on the COCO2014 dataset, the model was trained for 80 epochs on 6 × RTX3090 GPUs, completing in 7.5 h (adjust for your setup). For larger datasets, training time scales with dataset size and model complexity. To mitigate computational demands, we utilized data parallelism and distributed training for acceleration.

To provide a clearer understanding of the model's training process and feature extraction workflow, we present a detailed explanation of the procedure as follows: Using ResNet-101 with a resolution of 448 for COCO2014 (80 labels) as an example, the resulting output features of backbone are denoted as $F_1 \in 1024 \times 28 \times 28$ and $F_2 \in 2048 \times 14 \times 14$. Subsequently, a 1×1 convolution is applied to F_1, while F_2

undergoes both a 1×1 convolution and a 3×3 convolution. This process generates the desired feature maps at three different scales: $x_1 \in 512 \times 28 \times 28$, $x_2 \in 512 \times 14 \times 14$ and $x_3 \in 512 \times 7 \times 7$. After obtaining feature maps at three different scales, they are processed through the hierarchical scale attention mechanism. The input and output feature dimensions remain consistent throughout this process. Then, three class-specific representations, $Z_i \in 80 \times 512$ are obtained through an encoder-decoder structure. To integrate the information from the three feature representations for the final classification stage, the feature maps are concatenated along the channel dimensions, resulting in a composite feature representation $Z' \in 80 \times 1532$. Finally, a linear layer is applied to generate the prediction scores corresponding to 80 categories.

4.4 Comparison with State-of-the-Art Methods

To validate the effectiveness of the proposed model, mainstream models in the multi-label image classification task were selected for comparison. To ensure fairness, the proposed model and the selected comparison methods used the same backbone, and the image crop resolution sizes were clearly noted.

4.4.1 COCO2014

In the COCO2014 dataset, the proposed model was compared with several mainstream models, including DRGN (TMM 2024), C-TMS (TMM 2024), IDA (ICLR 2023), FL-Tran (PR 2023), etc. Table 1 reports the experimental results of the proposed model on the COCO2014 dataset at resolutions of 448 and 576. Since there is a trade-off between precision and recall, the F1 score provides a more comprehensive evaluation. Therefore, this paper focuses on three key metrics: mAP, CF1, and OF1. It is evident that the proposed model consistently outperforms other comparison methods in terms of mAP, CF1, and OF1.

Table 1: Comparison results on the MS-COCO dataset

Method	Resolutions	mAP	CP	CR	CF1	OP	OR	OF1
ResNet101 [29]	224 × 224	78.3	80.2	66.7	72.8	83.9	70.8	76.8
SRN [37]	224 × 224	77.1	81.6	65.4	71.2	82.7	69.9	75.8
KSSNet [38]	448 × 448	83.7	84.6	73.2	77.2	87.8	76.2	81.5
CADM [43]	448 × 448	82.3	82.5	72.2	77.0	84.0	75.6	79.6
ML-GCN [13]	448 × 448	83.0	85.1	72.0	78.0	85.8	75.4	80.3
MS-CMA [16]	448 × 448	83.8	82.9	74.4	78.4	84.4	77.9	81.0
MCAR [42]	448 × 448	83.8	85.0	72.1	78.0	88.0	73.9	80.3
Q2L [17]	448 × 448	84.9	84.8	74.5	79.3	86.6	76.9	81.5
CCD [10]	448 × 448	84.0	87.2	70.9	77.3	88.8	74.6	81.1
SRDL [39]	448 × 448	82.9	85.4	70.8	77.4	87.1	74.8	80.5
FL-Tran [14]	448 × 448	84.0	84.9	73.5	78.8	86.0	76.3	80.9
IDA [34]	448 × 448	84.8	–	–	78.7	–	–	80.9
MulCon [36]	448 × 448	84.9	84.0	74.8	79.2	85.6	78.0	81.6
DRGN [31]	448 × 448	84.9	86.3	73.8	79.6	87.4	76.6	81.6
DATran [15]	448 × 448	84.9	84.8	74.9	79.6	86.0	77.6	81.6
MSFA [30]	448 × 448	83.0	87.0	70.2	77.7	89.7	72.7	80.3
Ours	448 × 448	**85.1**	84.8	75.3	**79.8**	85.8	78.1	**81.8**
SSGRL [44]	576 × 576	83.8	89.9	68.5	76.8	91.3	70.8	79.7
ADD-GCN [11]	576 × 576	85.2	84.7	75.9	80.1	84.9	79.4	82.0
C-Trans [45]	576 × 576	85.1	86.3	74.3	79.9	87.7	76.5	81.7
TDRG [46]	576 × 576	86.0	87.0	74.7	80.4	87.5	77.9	82.4
DA-GAT [35]	576 × 576	84.8	87.0	74.2	80.1	87.3	77.5	82.1
FL-Tran [14]	576 × 576	85.4	84.5	76.3	80.2	85.7	78.9	82.2
MulCon [36]	576 × 576	86.3	84.7	77.3	80.8	85.9	79.9	82.8
Ours	576 × 576	**86.6**	86.6	76.2	**81.0**	87.7	78.7	**83.0**

Note: Bold numbers indicate the optimal values of the mAP, CF1 and OF1.

Compared to Region-based methods, such as C-TMS, IDA, ResNet101, SRN, etc., our model incorporates a Transformer encoder to perform self-attention computations. This significantly enhances the ability to model relationships between different regions in an image, effectively addressing the limitations of traditional convolution- and attention-based models in feature representation and global feature extraction. Compared to graph-based methods for modeling label dependencies, such as DA-GAT, ADD-GCN, TDRG, etc., our approach employs cross-attention mechanisms to directly query class-specific representations, focusing on identifying regions of interest across multiple scales. Without relying on graph networks, our method effectively captures implicit label relationships while mitigating the issue of learning spurious label correlations that can arise from insufficient label statistical data. Compared to other multi-scale feature-based models, such as DATran, DRGN, FL-Tran, MSFA, etc., our approach utilizes a three-branch Transformer structure to perform precise feature selection and region refinement across three different scales. Through a hierarchical scale attention mechanism, it facilitates effective information flow between features of different scales. Our method emphasizes hierarchical multi-scale feature fusion and cross-modal relationship modeling, which distinguishes it from the aforementioned methods. Furthermore, experimental results demonstrate that our method outperforms these approaches, showcasing superior robustness and generalization capabilities.

Notably, by utilizing multi-scale feature fusion, the proposed model demonstrates stronger robustness and generalization ability compared to the Q2L model, which also leverages the cross attention mechanism to establish relationships between image regions and labels. The proposed model achieves improvements of 0.2%, 0.5%, and 0.3% on mAP, CF1, and OF1, respectively.

4.4.2 VOC2007

Table 2 presents the experimental results of the proposed model on the VOC2007 dataset at a resolution of 448, comparing its performance with state-of-the-art methods, including DRGN (TMM 2024) and FLNet (CC 2023). The table reports the average precision (AP) for each of the 20 categories and the mean average precision (mAP).

Table 2: Comparison results on the VOC2007 dataset

Class	HCP [7]	CNN-RNN [47]	RNN-Att [18]	RARL [40]	SSGRL [44]	ML-GCN [13]	KGGR [48]	P-GCN [49]	LDR [50]	MS-SGA [54]	CPCL [51]	FLNet [52]	DRGN [31]	ML-AGCN [33]	Ours
Aero	98.6	96.7	98.6	98.6	99.5	99.5	99.3	99.6	99.6	99.6	99.6	99.6	99.8	99.9	**100.0**
Bike	97.1	83.1	97.4	97.1	97.1	98.5	98.6	98.6	98.3	98.3	98.6	98.7	98.6	98.0	**99.2**
Bird	98.0	94.2	96.3	97.1	97.6	98.6	97.9	98.4	98.0	98.0	98.5	**98.9**	98.3	98.5	**98.9**
Boat	95.6	92.8	96.2	95.5	97.8	98.1	98.4	98.7	98.2	97.5	**98.8**	97.9	98.6	98.0	98.6
Bottle	75.3	61.2	75.2	75.6	82.6	80.8	**86.2**	81.5	78.2	81.0	81.9	84.6	81.8	81.6	85.1
Bus	94.7	82.1	92.4	92.8	94.8	94.6	97.0	94.8	94.2	93.1	95.1	95.3	95.5	96.8	**97.8**
Car	95.8	89.1	96.5	96.8	96.7	97.2	98.0	97.6	97.0	97.5	97.8	96.2	97.6	96.6	**98.3**
Cat	97.3	94.2	97.1	97.3	98.1	98.2	**99.2**	98.2	97.8	98.5	98.2	96.5	98.0	98.2	99.0
Chair	73.1	64.2	76.5	78.3	78.0	82.3	82.6	83.1	80.8	**86.3**	83.0	85.6	83.9	85.6	82.2
Cow	90.2	83.6	92.0	92.2	97.0	95.7	98.3	96.0	94.9	88.3	95.5	96.1	94.9	**99.4**	98.6
Table	80.0	70.0	87.7	87.6	85.6	86.4	87.5	87.1	84.9	**89.2**	85.5	87.2	87.5	88.2	88.2
Dog	97.3	92.4	96.8	96.9	97.8	98.2	99.0	98.3	97.7	95.5	98.4	97.7	98.4	**99.2**	98.7
Horse	96.1	91.7	97.5	96.5	98.3	98.4	98.9	98.5	97.5	98.0	98.5	98.6	97.8	**99.0**	98.9
Motor	94.9	84.2	93.8	93.6	96.4	96.7	97.4	96.3	96.6	96.1	97.0	97.0	97.4	96.5	**97.5**
Person	96.3	93.7	98.5	98.5	98.1	99.0	99.1	99.1	98.7	98.3	99.0	98.1	98.8	98.8	**99.2**
Plant	78.3	59.8	81.6	81.6	84.9	84.7	86.9	87.3	85.0	**89.0**	86.6	86.5	86.6	84.8	88.4
Sheep	94.7	93.2	93.7	93.1	96.5	96.7	98.2	95.5	96.2	96.7	97.0	97.4	96.2	**99.5**	99.1
Sofa	76.2	75.3	82.8	83.2	79.8	84.3	84.1	85.4	83.2	**91.6**	84.9	86.5	85.6	88.1	86.3
Train	97.9	**99.7**	98.6	98.5	98.4	98.9	99.0	98.9	98.5	97.9	99.1	98.8	99.4	98.9	99.6

(Continued)

Table 2 (continued)

Class	HCP [7]	CNN-RNN [47]	RNN-Att [18]	RARL [40]	SSGRL [44]	ML-GCN [13]	KGGR [48]	P-GCN [49]	LDR [50]	MS-SGA [54]	CPCL [51]	FLNet [52]	DRGN [31]	ML-AGCN [33]	Ours
Tv	91.5	78.6	89.3	89.3	92.8	93.7	**95.0**	93.6	92.6	92.3	94.3	90.8	94.9	94.5	94.0
mAP	90.9	84.0	91.9	92.0	93.4	94.0	95.0	94.3	93.4	94.2	94.4	94.4	94.5	95.0	**95.4**

Note: Bold numbers indicate the optimal values in each row, while underlined numbers represent the second-best values.

The proposed model achieves superior average precision in 7 categories and ranks as the second-best in 8 other categories. Notably, for small and challenging objects, such as birds and plants, the proposed model achieves significant improvements. Besides, our method achieves an mAP increase of 0.9% over DRGN and 1.0% over FLNet. This demonstrates the model's ability to effectively handle objects of varying scales, where existing SOTA methods encounter limitations.

Compared to DRGN, which focuses on cross-image semantic learning and intra-image spatial relations, our approach excels in accurately distinguishing features at different scales. DRGN struggles with small objects due to its limited emphasis on fine-grained details within images. In contrast, the hierarchical scale attention mechanism in our model enhances the information flow and sharing between features at different scales, effectively bridging the gap between global context and local details. This synergy enables precise recognition of small and intricate objects.

Similarly, FLNet integrates semantic embeddings with visual features using a CNN-GCN framework, but it heavily relies on the association between labels and image regions without sufficiently addressing feature-level scale variations. Our model surpasses FLNet by leveraging multi-scale feature fusion and cross-attention mechanisms to refine category-specific features at multiple scales, thereby achieving better performance for objects with significant size differences.

These results highlight the strengths of the proposed multi-scale feature fusion framework, particularly in leveraging hierarchical scale attention to effectively balance local detail preservation and global context extraction. This allows the model to robustly locate and classify objects of varying sizes, outperforming SOTA methods in challenging scenarios such as recognizing small or overlapping objects.

4.4.3 VOC2012

In the VOC2012 dataset, the proposed model was compared with several other models, including VeryDeep [41], Fev+Lv [8], HCP [7], MCAR [42], SSGRL [44], ADD-GCN [11], KGGR [48], DSDL [53], C-TMS [32].

Table 3 reports the experimental results of our model on the VOC2012 dataset, showing the average precision (AP) for 20 categories as well as the mean average precision (mAP). Except for SSGRL (576), ADD-GCN (512), and KGGR (576), all other models used an input resolution of 448. We applied the same experimental settings as on the VOC2007 dataset. As seen from the results, our model achieved the highest AP in 11 categories, while 2 of the remaining 9 categories yielded second-best results. Notably, even though SSGRL, ADD-GCN, and KGGR used higher resolutions, our model still attained the highest mAP.

Table 3: Comparison results on the VOC2012 dataset

Class	VeryDeep [41]	Fev+Lv [8]	HCP [7]	MCAR [42]	SSGRL [44]	ADD-GCN [11]	KGGR [48]	DSDL [53]	C-TMS [32]	Ours
Aero	99.1	98.4	99.1	99.6	99.7	99.5	**99.8**	**99.8**	99.4	**99.8**

(Continued)

Table 3 (continued)

Class	VeryDeep [41]	Fev+Lv [8]	HCP [7]	MCAR [42]	SSGRL [44]	ADD-GCN [11]	KGGR [48]	DSDL [53]	C-TMS [32]	Ours
Bike	88.7	92.8	92.8	97.1	96.1	97.1	97.3	95.3	96.2	**97.7**
Bird	95.7	93.4	97.4	98.3	97.7	**98.6**	98.4	97.6	97.6	98.3
Boat	93.9	90.7	94.4	96.6	96.5	96.8	97.1	95.7	96.5	**97.6**
Bottle	73.1	74.9	79.9	87.0	86.9	**89.4**	87.9	83.5	86.4	87.4
Bus	92.1	93.2	93.6	95.5	95.8	97.1	**97.3**	94.8	95.8	96.8
Car	84.8	90.2	89.8	94.4	95.0	96.5	96.5	93.9	95.2	**96.8**
Cat	97.7	96.1	98.2	98.8	98.9	**99.3**	**99.3**	98.5	98.9	99.0
Chair	79.1	78.2	78.2	87.0	88.3	89.0	89.4	85.7	88.7	**90.2**
Cow	90.7	89.8	94.9	96.9	97.6	97.7	97.8	94.5	97.5	**98.1**
Table	83.2	80.6	79.8	85.0	87.4	87.5	88.7	83.8	87.6	**89.0**
Dog	97.3	95.7	97.8	98.7	99.1	99.2	**99.4**	98.4	99.2	99.1
Horse	96.2	96.1	97.0	98.3	99.2	99.1	**99.4**	97.7	99.2	99.0
Motor	94.3	95.3	93.8	97.3	97.3	97.7	**97.9**	95.9	97.3	97.5
Person	96.9	97.5	96.4	99.0	99.0	99.1	**99.2**	98.5	99.0	**99.2**
Plant	63.4	73.1	74.3	83.8	84.8	86.3	86.3	80.6	85.0	**87.5**
Sheep	93.2	91.2	94.7	96.8	98.3	**98.8**	**98.8**	95.7	98.1	98.1
Sofa	74.6	75.4	71.9	83.7	85.8	87.0	86.3	82.3	86.1	**89.1**
Train	97.3	97.0	96.7	98.3	99.2	99.3	**99.7**	98.2	99.1	99.5
Tv	87.9	88.2	88.6	93.5	94.1	95.4	95.2	93.2	94.2	**95.9**
mAP	89.0	89.4	90.5	94.3	94.8	95.5	95.6	93.2	94.8	**95.8**

Note: Bold numbers indicate the optimal values in each row, while underlined numbers represent the second-best values.

Since the official test set annotations were not fully available, all results were evaluated on the official VOC2012 evaluation server. We have shared the link for anonymously viewing the results.

4.5 Ablation Study

In this section, we present a series of ablation experiments conducted on the COCO 2014 dataset to evaluate the effectiveness of key model components. Specifically, we analyze: (1) the contribution of individual modules to performance, (2) the impact of different multi-scale feature fusion strategies, (3) the choice of interpolation methods, (4) the role of self-attention in the decoder, and (5) the effect of varying encoder and decoder layer numbers. These experiments help validate the design choices and optimize model performance.

4.5.1 Comparison of Different Module in the Model

This analysis evaluates the contribution of individual modules within the model, aiming to quantify their respective impact on overall performance and validate their necessity.

Table 4 reports the detailed results of these ablation experiments. It can be observed that each module in the proposed model contributes to the improvement of the model's performance.

Table 4: Ablation analysis of the validity of the three modules in our model

Hierarchical scale attention	Global feature enhancement module	Region selection module	mAP
	✓	✓	84.57
✓		✓	84.89
✓	✓		84.03
✓	✓	✓	85.14

The hierarchical scale attention module integrates features from different scales and establishes connections between them, enabling the model to better handle objects of varying sizes. When the hierarchical scale attention module is removed, there is a 0.57% drop in the mAP score, demonstrating the module's importance in improving the model's ability to detect objects of different sizes.

The global feature enhancement module improves the model's ability to extract global image features. This demonstrates the necessity of further global feature extraction after the image passes through the convolutional neural network. Without the global feature enhancement module, the mAP score decreases by 0.25%, underscoring its significance.

The region selection module in this model utilizes the built-in cross-attention mechanism of the decoder to get the precise feature representation. This enables the model to quickly and adaptively locate the class-specific ROI regions. Without the region selection module, the mAP score drops by 1.11%.

4.5.2 Analysis of Feature Fusion Methods

In the first phase, this paper considers three modes for feature fusion. To select the optimal fusion method, a comparative experimental analysis was conducted on the three modes. The results of these three fusion methods are reported in Table 5.

Table 5: Analysis of the number of encoder and decoder layers

	Mode 1	**Mode 2**	**Mode 3**
Layer 2	/	3 × 3	/
Layer 3	1 × 1 3 × 3	3 × 3	1 × 1
Layer 4	3 × 3	3 × 3	1 × 1 3 × 3
mAP	84.37	84.76	85.14

It can be observed that the feature fusion method in Mode 3 performs the best, with improvements of 0.77% and 0.38% compared to Mode 1 and Mode 2. Additionally, compared to the traditional method (which only extracts results from the 1 × 1 convolution in Layer 4), the feature fusion method in Mode 3 not only leverages 3 × 3 convolutions to obtain smaller feature maps for better handling of large object targets but also utilizes the output from Layer 3 to capture relatively larger feature maps for better handling of small object targets.

4.5.3 Analysis of Interpolation Strategies for Feature Processing

Interpolation techniques play a critical role in feature processing tasks such as image resizing and transformation. In the design of the Hierarchical Scale Attention, as discussed in Section 3.2, interpolation techniques are employed. The choice of interpolation strategy can significantly impact both the quality of feature extraction and model performance. Due to its importance, we have conducted experiments on different interpolation methods suitable for 4D features.

Based on the results presented in Table 6, it can be observed that Bilinear interpolation outperforms the other methods in terms of mAP, achieving a value of 85.15%. This is slightly higher than the Bicubic (85.12%), Area (85.11%), and Nearest (85.08%) interpolation methods. Given this marginal but consistent improvement, it is evident that Bilinear interpolation is the most effective interpolation strategy for this task.

It is worth noting that during the experiments, we found that the Bicubic interpolation method, while providing competitive performance, incurs a significant computational overhead. Compared to the other

interpolation methods, Bicubic requires approximately 4–5 times more processing time. This additional computational cost may be a limiting factor in scenarios where efficiency is critical, making Bilinear interpolation a more balanced choice, as it offers the best performance with relatively lower computational demand.

Table 6: Ablation study on interpolation strategies

Interpolation method	mAP
Nearest	85.08
Bilinear	85.15
Bicubic	85.12
Area	85.11

4.5.4 Analysis of the Necessity of Self-Attention in the Decoder

The self-attention layer in the Transformer decoder is often believed to enhance the embedding representations of labels, and previous works that use the standard decoder typically incorporate this layer. However, in Section 3.5, we argue that the self-attention layer in the decoder may not be as beneficial as commonly assumed. To validate this, we conducted experiments on the MS-COCO dataset, and the results are shown in Table 7. In our experiments, we further analyzed four settings: using self-attention in the first decoder layer, using self-attention in other layers, and using self-attention in all or none of the decoder layers. By comparing these settings, we aimed to assess the role of self-attention at different layers and its impact on model performance. From the table, we observe that removing self-attention layers from the decoder reduces the parameter count with little to no impact on model performance.

Table 7: Ablation study on the usage of self-attention layers in different decoders. "✓" denotes the use of self-attention in the corresponding decoder layer, while the absence of "✓" means self-attention is not used. The "Parameters" column shows the model parameter count for each setting

Decoder Layer 1	Decoder Layer 2	Decoder Layer 3	mAP	Parameters
✓	✓	✓	85.08	72.7 M
	✓	✓	85.01	71.7 M
✓			85.14	70.6 M
			85.15	69.5 M

One of possible reasons is that, applying self-attention to these learnable label embeddings forces the labels to learn inter-relationships, which may result in spurious label correlations. Another possible reason is that, before computing cross-attention (recall Fig. 1), the input label embeddings undergo a linear transformation, which enables the input to adaptively fit the cross-attention module's expected input, as mentioned in Section 3.5; in other words, removing self-attention layer in decoders makes less or even no negative impact.

4.5.5 Analysis of the Number of Encoder and Decoder Layers

To explore the sensitivity of the encoder-decoder layers in the global feature enhancement and region selection modules, and to identify the optimal model parameters, we conducted experiments with varying layer numbers in these modules. Due to the substantial computational cost associated with a large number

of encoder-decoder layers, only configurations with up to 3 layers were considered. Fig. 5 presents the experimental results for the analysis of encoder and decoder layer counts.

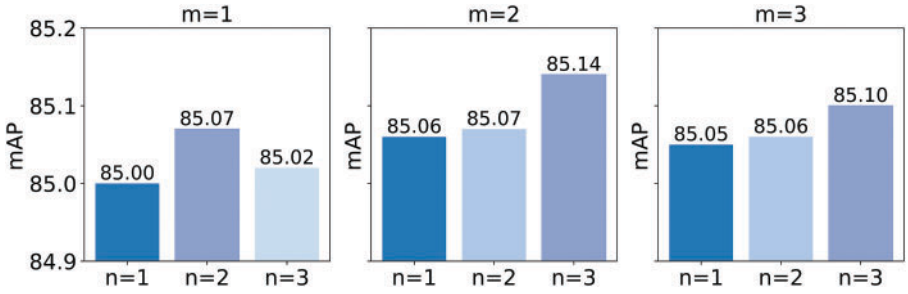

Figure 5: Analysis of the Number of Encoder and Decoder layers. *m* denotes the number of encoder layers in the global feature enhancement module and *n* represents the number of simplified decoder layers in the region selection module

It is clear that the mAP value does not increase proportionally with the number of encoder-decoder layers. Among the nine different combinations, the highest mAP value, indicating the most optimal classification performance, was achieved when the number of encoder layers was set to 2 and the number of decoder layers was set to 3.

5 Conclusion

This paper presents a novel multi-label image classification framework that integrates multi-scale feature fusion with cross-modal representation learning. By incorporating a three-branch Transformer model with hierarchical scale attention and a lightweight cross-attention mechanism, the proposed model efficiently captures class-specific features at different scales and establishes strong correlations between image regions and textual label information. Extensive experimental results on benchmark datasets such as MS-COCO and PASCAL VOC demonstrate that the proposed approach achieves competitive performance, surpassing existing state-of-the-art methods.

Despite its effectiveness, the model introduces certain computational complexities due to the multi-branch structure and attention mechanisms. While shared weights across branches help reduce parameters, this design slightly compromises the ability to capture scale-specific features. Future work will address these limitations by exploring lightweight architectures, such as pruning or knowledge distillation, and enhancing branch-specific designs to achieve better trade-offs between efficiency and performance.

Acknowledgement: We would like to extend our heartfelt thanks to the editors and each reviewer for their diligent efforts and expert guidance.

Funding Statement: This work was supported by the National Natural Science Foundation of China (62302167, 62477013), Natural Science Foundation of Shanghai (No. 24ZR1456100), Science and Technology Commission of Shanghai Municipality (No. 24DZ2305900) and the Shanghai Municipal Special Fund for Promoting High-Quality Development of Industries (2211106).

Author Contributions: The authors confirm contribution to the paper as follows: Naikang Zhong: Designing methodologies, developing network modules, and composing the thesis. Xiao Lin: Analyzing the results, reviewing, supervising, and securing funding. Wen Du: Guiding the work and analyzing the theoretical aspects of the module. Jin Shi: Developing network modules and composing the thesis. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in MS-COCO and PASCAL-VOC at https://cocodataset.org and http://host.robots.ox.ac.uk/pascal/VOC (accessed on 24 October 2024).

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Jenifer PIR, Kannan S. Deep learning with optimal hierarchical spiking neural network for medical image classification. Comput Syst Sci Eng. 2023;44(2):1081–97. doi:10.32604/csse.2023.026128.
2. Zhou YT, Yang XM, Yin JP, Liu SQ. Research on multi-scale feature fusion network algorithm based on brain tumor medical image classification. Comput Mater Contin. 2024;79(3):5313–33. doi:10.32604/cmc.2024.052060.
3. Wang Z-J, Ma L, Lin X, Wu X. MSGC: a new bottom-up model for salient object detection. In: 2018 IEEE International Conference on Multimedia and Expo (ICME); 2018; San Diego, CA, USA: IEEE. p. 1–6.
4. Zeng M, Yao B, Wang ZJ, Shen Y, Li F, Zhang J, et al. CATIRI: an efficient method for content-and-text based image retrieval. J Comput Sci Tech. 2019;34(2):287–304. doi:10.1007/s11390-019-1911-2.
5. Alshahrani AA, Jaha ES, Alowidi N. Fusion of hash-based hard and soft biometrics for enhancing face image database search and retrieval. Comput Mater Contin. 2023;77(3):3489–509. doi:10.32604/cmc.2023.044490.
6. Lin X, Xu DJ, Tan PW, Ma LZ, Wang ZJ. Image deraining based on dual-channel component decomposition. Comput Graph. 2023;116:93–101. doi:10.1016/j.cag.2023.08.010.
7. Wei Y, Xia W, Lin M, Huang J, Ni B, Dong J, et al. HCP: a flexible CNN framework for multi-label image classification. IEEE Trans Pattern Anal Mach Intell. 2015;38(9):1901–7. doi:10.1109/TPAMI.2015.2491929.
8. Yang H, Zhou JT, Zhang Y, Gao B-B, Wu J, Cai J. Exploit bounding box annotations for multi-label object recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 280–8.
9. Guo H, Zheng K, Fan X, Yu H, Wang S. Visual attention consistency under image transforms for multi-label image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA, USA. p. 729–39.
10. Liu R, Liu H, Li G, Hou H, Yu T, Yang T. Contextual debiasing for visual recognition with causal mechanisms. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 12755–65.
11. Ye J, He J, Peng X, Wu W, Qiao Y. Attention-driven dynamic graph convolutional network for multi-label image recognition. In: Computer Vision—ECCV 2020, 16th European Conference on Computer Vision; 2020 Aug 23–28; Glasgow, UK. Berlin/Heidelberg, Germany: Springer. p. 649–65.
12. Nguyen HD, Vu X-S, Le D-T. Modular graph transformer networks for multi-label image classification. Proc AAAI Conf Artif Intell. 2021;35(10):9092–100. doi:10.1609/aaai.v35i10.17098.
13. Chen Z-M, Wei X-S, Wang P, Guo Y. Multi-label image recognition with graph convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA, USA. p. 5177–86.
14. Zhou W, Dou P, Su T, Hu H, Zheng Z. Feature learning network with transformer for multi-label image classification. Pattern Recognit. 2023;136:109203. doi:10.1016/j.patcog.2022.109203.
15. Zhou W, Zheng ZJ, Su T, Hu HF. DATran: dual attention transformer for multi-label image classification. IEEE Trans Circuits Syst Video Technol. 2024;34(1):342–56. doi:10.1109/TCSVT.2023.3284812.
16. You R, Guo Z, Cui L, Long X, Bao Y, Wen S. Cross-modality attention with semantic graph embedding for multi-label classification. Proc AAAI Conf Artif Intell. 2020;34(7):12709–16. doi:10.1609/aaai.v34i07.6964.
17. Liu S, Zhang L, Yang X, Su H, Zhu J. Query2label: a simple transformer way to multi-label classification. arXiv:2107.10834. 2021.

18. Wang Z, Chen T, Li G, Xu R, Lin L. Multi-label image recognition by recurrently discovering attentional regions. In: Proceedings of the IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice, Italy. p. 464–72.
19. Ridnik T, Sharir G, Ben-Cohen A, Ben-Baruch E, Noy A. Ml-decoder: scalable and versatile classification head. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision; 2023 Jan 2–7; Waikola, HI, USA. p. 32–41.
20. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv:1609.02907. 2016.
21. Ye J, Jiang L, Xiao S, Zong Y, Jiang A. Multi-label image classification model based on multiscale fusion and adaptive label correlation. J Shanghai Jiaotong Univ. 2024;34:1–10. doi:10.1007/s12204-023-2688-6.
22. Kenton JDM-WC, Toutanova LK. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT; 2019 Jun 2–7; Minneapolis, MN, USA. Vol. 1, p. 2.
23. Yang Z. XLNet: generalized autoregressive pretraining for language understanding. arXiv:1906.08237. 2019.
24. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1–67.
25. Alexey D. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
26. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, BC, Canada. p. 10012–22.
27. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H. Training data-efficient image transformers & distillation through attention. In: Proceedings of the 38th International Conference on Machine Learning; 2021 Jul 18–24. p. 10347–57.
28. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: simple and efficient design for semantic segmentation with transformers. Adv Neural Inf Process Syst. 2021;34:12077–90.
29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8.
30. Chen J, Xu F, Zeng T, Li X, Chen S, Yu J. MSFA: multi-stage feature aggregation network for multi-label image recognition. IET Image Process. 2024;18(7):1862–77. doi:10.1049/ipr2.13068.
31. Zhou W, Jiang W, Chen D, Hu H, Su T. Mining semantic information with dual relation graph network for multi-label image classification. IEEE Trans Multimed. 2023;26:1143–57. doi:10.1109/TMM.2023.3277279.
32. Wu Y, Feng S, Zhao G, Jin Y. Transformer driven matching selection mechanism for multi-label image classification. IEEE Trans Circ Syst Video Technol. 2024;34(2):924–37. doi:10.1109/TCSVT.2023.3288205.
33. Singh IP, Ghorbel E, Oyedotun O, Aouada D. Multi-label image classification using adaptive graph convolutional networks: from a single domain to multiple domains. Comput Vis Image Underst. 2024;247:104062. doi:10.1016/j.cviu.2024.104062.
34. Liu R, Huang J, Li TH, Li G. Causality compensated attention for contextual biased visual recognition. In: 11th International Conference on Learning Representations; 2023 May 1–5; Kigali, Rwanda.
35. Zhou W, Xia Z, Dou P, Su T, Hu H. Double attention based on graph attention network for image multi-label classification. ACM Trans Multimed Comput Commun Appl. 2023;19(1):1–23. doi:10.1145/3519030.
36. Dao SD, Zhao H, Phung D, Cai J. Contrastively enforcing distinctiveness for multi-label image classification. Neurocomputing. 2023;555:126605. doi:10.1016/j.neucom.2023.126605.
37. Zhu F, Li H, Ouyang W, Yu N, Wang X. 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 Jul 21–26; Honolulu, HI, USA. p. 5513–22.
38. Liu Y, Sheng L, Shao J, Yan J, Xiang S, Pan C. Multi-label image classification via knowledge distillation from weakly-supervised detection. In: Proceedings of the 26th ACM International Conference on Multimedia; 2018 Oct; Seoul, Republic of Korea. p. 700–8.
39. Pu T, Sun M, Wu H, Chen T, Tian L, Lin L. Semantic representation and dependency learning for multi-label image recognition. Neurocomputing. 2023;526:121–30. doi:10.1016/j.neucom.2023.01.018.
40. Chen T, Wang Z, Li G, Lin L. Recurrent attentional reinforcement learning for multi-label image recognition. Proc AAAI Conf Artif Intell. 2018;32(1):6730–7. doi:10.1609/aaai.v32i1.12281.

41. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. 2014.
42. Gao B-B, Zhou H-Y. Learning to discover multi-class attentional regions for multi-label image recognition. IEEE Trans Image Process. 2021;30:5920–32. doi:10.1109/TIP.2021.3088605.
43. Chen Z-M, Wei X-S, Jin X, Guo Y. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In: 2019 IEEE International Conference on Multimedia and Expo (ICME); 2019 Jul 8–12; Shanghai, China. p. 622–7.
44. Chen T, Xu M, Hui X, Wu H, Lin L. Learning semantic-specific graph representation for multi-label image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 522–31.
45. Lanchantin J, Wang T, Ordonez V, Qi Y. General multi-label image classification with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 16478–88.
46. Zhao J, Yan K, Zhao Y, Guo X, Huang F, Li J. Transformer-based dual relation graph for multi-label image recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision; 2021 Oct 11–17; Montreal, BC, Canada. p. 163–72.
47. Wang J, Yang Y, Mao J, Huang Z, Huang C, Xu W. Cnn-rnn: a unified framework for multi-label image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 2285–94.
48. Chen T, Lin L, Chen R, Hui X, Wu H. Knowledge-guided multi-label few-shot learning for general image recognition. IEEE Trans Pattern Anal Mach Intell. 2020;44(3):1371–84. doi:10.1109/TPAMI.2020.3025814.
49. Chen Z-M, Wei X-S, Wang P, Guo Y. Learning graph convolutional networks for multi-label recognition and applications. IEEE Trans Pattern Anal Mach Intell. 2021;45(6):6969–83. doi:10.1109/TPAMI.2021.3063496.
50. Hassanin M, Radwan I, Khan S, Tahtali M. Learning discriminative representations for multi-label image recognition. J Vis Commun Image Rep. 2022;83:103448. doi:10.1016/j.jvcir.2022.103448.
51. Xu J, Huang S, Zhou F, Huangfu L, Zeng D, Liu B. Boosting multi-label image classification with complementary parallel self-distillation. arXiv:2205.10986. 2022.
52. Sun D, Ma L, Ding Z, Luo B. An attention-driven multi-label image classification with semantic embedding and graph convolutional networks. Cognit Comput. 2023;15:1308–19. doi:10.1007/s12559-021-09977-9.
53. Zhou F, Huang S, Xing Y. Deep semantic dictionary learning for multi-label image classification. Proc AAAI Conf Artif Intell. 2021;35(4):3572–80. doi:10.1609/aaai.v35i4.16472.
54. Liang J, Xu F, Yu S. A multi-scale semantic attention representation for multi-label image recognition with graph networks. Neurocomputing. 2022;491:14–23. doi:10.1016/j.neucom.2022.03.057.
55. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: an imperative style, high-performance deep learning library. In: Advances in neural information processing systems. Vancouver, BC, Canada: Curran Associates Inc.; 2019. Vol. 32.
56. Loshchilov I. Decoupled weight decay regularization. arXiv:1711.05101. 2017.
57. Cubuk ED, Zoph B, Shlens J, Le QV. Randaugment: practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops; 2020 Jun 14–19; Seattle, WA, USA. p. 702–3.
58. DeVries T. Improved regularization of convolutional neural networks with cutout. arXiv:1708.04552. 2017.