Global Meets Local: Effective Multi-Label Image Classification via Category-Aware Weak Supervision

Jiawei Zhan*
Jun Liu*
Tencent Youtu Lab, China

Wei Tang*
Institute of Automation, Chinese Academy of Sciences, China

Guannan Jiang
Xi Wang
Contemporary Amperex Technology Co., Limited, China

Bin-Bin Gao
Tianliang Zhang
Wenlong Wu
Tencent Youtu Lab, China

Wei Zhang
Contemporary Amperex Technology Co., Limited, China

Chengjie Wang†
Tencent Youtu Lab, China

Yuan Xie†
East China Normal University, China

ABSTRACT

Multi-label image classification, which can be categorized into label-dependency and region-based methods, is a challenging problem due to the complex underlying object layouts. Although region-based methods are less likely to encounter issues with model generalizability than label-dependency methods, they often generate hundreds of meaningless or noisy proposals with non-discriminative information, and the contextual dependency among the localized regions is often ignored or over-simplified. This paper builds a unified framework to perform effective noisy-proposal suppression and to interact between global and local features for robust feature learning. Specifically, we propose category-aware weak supervision to concentrate on non-existent categories so as to provide deterministic information for local feature learning, restricting the local branch to focus on more high-quality regions of interest. Moreover, we develop a cross-granularity attention module to explore the complementary information between global and local features, which can build the high-order feature correlation containing not only global-to-local, but also local-to-local relations. Both advantages guarantee a boost in the performance of the whole network. Extensive experiments on two large-scale datasets (MS-COCO and VOC 2007) demonstrate that our framework achieves superior performance over state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies** → **Supervised learning by classification**; **Object recognition**.

KEYWORDS

multi-label classification, image recognition, region proposal, weak supervision, self-attention

1 INTRODUCTION

Multi-label image classification [9, 20, 25], which aims at predicting multiple labels for an image, is a fundamental task in the field of computer vision and multimedia. It has a wide range of applications in fields such as image retrieval [5], attribute recognition [24], and automatic image annotation [29]. Compared with single-label image classification, multi-label classification is more complex and challenging, as it is deals with complex underlying object layouts such as variations in location and scale and the difference between intra-class and inter-class.

As one of the widely adopted solutions, region-based multi-label learning first generates a large number of proposals using methods such as selective search [26], edge box [47] and BING [7]. Binary cross-entropy loss is then used for each proposal instead of common softmax loss. More recent methods have begun to adopt long short-term memory [3, 33] or utilize strong supervision [39] such as bounding boxes with RPN [22] to generate more accurate proposals.

Although acceptable results have been achieved, region-based methods still face two limitations: they generate a large number of noisy region proposals [21, 34, 35], and the relationship among region proposals has not been thoroughly explored [1, 3]. Specifically, to achieve a high recall rate, region-based methods mainly produce proposals using object-detection techniques. These methods usually generate many noisy region proposals, which are not only computationally inefficient for multi-label learning, but also detrimental to the performance due to the background interference and inaccurate border of the proposals. Moreover, as some methods are multi-stage [3, 45] and do not explore the correlation among labels, region-based methods usually lack a thorough understanding of the global information of an image, and the multi-label information is not effectively utilized to learn the semantic relationships between regions. If the relationship is well established, the regional proposal can be further constrained. Efficiently addressing the limitations described above in a unified framework is therefore a crucial step in boosting performance.

To generate region proposals more effectively, we propose category-aware weak supervision, accompanied with energy-based region-of-interest (ROI) selection to suppress noisy region proposals jointly. Since the ground-truth label only provides an indication of an existing category, rather than the accurate location of instances, current region-based methods are more likely to generate undesirable region proposals. Thus, we designed a weakly supervised loss function by only considering the feature map activation of non-existent categories, which can suppress most of the background information from an image through a weakly supervised learning process, reducing the number of proposals from thousands to dozens. Moreover, unlike the traditional area-based selection

*indicates equal contributions.
†indicates co-corresponding authors.

schema, the energy-based criterion can efficiently generate meaningful region proposals, thus enhancing their quality.

To establish the correlation between region proposals, we introduce a cross-granularity attention module, which can integrate feature information of instances at different granularities. The determination of how many categories are presented in an image, as distinct from instance localization, is a comprehensive global process, whereas region proposal is a local procedure. With the help of the cross-granularity attention module, our approach enables the construction of a high-order global-to-local interaction, which not only builds the connection between global features and local features (global-to-local), but also explores the association between region features (local-to-local) to improve representation capability and further suppress noisy region proposals. Both advantages guarantee a boost in the performance of the framework.

Overall, the major contributions of this work can be summarized as follows:

- A novel and unified framework for multi-label image classification tasks is proposed that includes category-aware weak supervision and a cross-granularity attention module, thus significanlty reducing the number of proposals and enabling the modeling of correlations among regions.
- Unlike traditional methods of supervision, which focus on the instances of presented categories, our proposed category-aware weak supervision concentrates on non-existent categories to provide deterministic information in the process of learning local features, thus effectively suppressing noisy region proposals.
- The cross-granularity attention module can not only capture global-to-local relationships, but also mine local-to-local correlations. To our knowledge, this is the first region-based method that can explore high-order global-to-local correlations among region proposals.
- Extensive experiments were conducted on two challenging large-scale public datasets (MS-COCO [18] and VOC 2007 [8]), and the results demonstrate that our framework achieves superior performance over state-of-the-art methods.

2 RELATED WORK

Previous works on this task mainly follow two directions: label-dependency methods and region-based methods.

2.1 Label-dependency Methods

To solve the problem that the correlation among labels is usually ignored in multi-label classifications, label-dependency methods start with labels and establish correlation across categories through the co-occurrence between different labels. [30] jointly learns image features and label correlations in a unified framework composed of a CNN module and an LSTM layer. Some works [3, 33, 42] further took advantage of proposal generation/visual attention mechanism and LSTM to explicitly model label dependencies. But it requires an explicit module for removing duplicate prediction labels and needs a threshold for stopping the sequence outputs.

Recently, due to the effectiveness of the Graph Convolutional Network in establishing correlations between labels, [4, 6, 32, 40] try to model the label dependency with a graph to boost the performance of multi-label image classification. While reasonably effective, these methods suffer from high computational costs or manually defined adjacency matrices. Besides, it is also arguable that it may learn spurious correlations when the label statistics are insufficient.

As a result, since label-dependency methods rely on the prior of the training data to a great extent, these methods may degrade the model's generalizability when faced with domain shift.

2.2 Region-based Methods

Region-based methods are another hot branch for multi-label image classification. Most existing region-based methods focus on locating informative regions (e.g., proposal candidates [35, 39, 42], attentive regions [1, 33, 46], random regions [31]) to cover all possible existing objects and aggregate local discriminative features to facilitate recognizing multiple labels of the given image.

Proposal candidates normally rely on object detection techniques [11], typically including BING [7, 34], EdgeBox [21, 35, 47] or Selective Search [26, 39] to generate an arbitrary number of object segment hypotheses, which often obtain hundreds of meaningless or noisy proposals with non-discriminative information, and the modeling of spatial contextual dependency among localized regions is often ignored or over-simplified [1, 3]. However, these methods model the spatial contextual dependency of different regions by using the category-agnostic mechanism, leading to generating noisy ROI proposals inevitably, as label knowledge is underutilization.

Both [39] and [42] utilize strong supervision (ground-truth bounding box annotations) to enhance the feature discriminative power. Despite the performance is improved, it will increase the cost of data acquisition. There have been some other attempts on multi-label researches, such as attention-based methods [12, 45, 46], dictionary learning [44] and few-shot multi-label classification [2].

In this paper, we aim to improve the performance of multi-label recognition with only image semantic information and propose category-aware weak supervision combined with an energy-based ROI selection schema to effectively suppress a large number of noisy region proposals. In addition, inspired by the methodology of label-dependency, by introducing cross-granularity attention, we establish global-to-local interaction to build the relationship not only between global and local features but also within different categories in an implicit way.

3 METHODOLOGY

3.1 Framework Overview

The framework consists of two branches: a global branch and a local branch. The global branch is used to generate the basic prediction for multi-label classification, and the local branch aims to provide category-specific (not only one category, but several saliency categories) prediction under category-aware weak supervision and energy-based ROI selection. The two branches interact with each other with the help of the cross-granularity attention module. Fig. 1 provides an overview of our proposed framework.

In the global branch, the samples are augmented and first passed through the backbone to produce feature maps with size $\frac{1}{16}$ and $\frac{1}{32}$, denoted by $\boldsymbol{F_{16}}$ and $\boldsymbol{F_{32}}$, respectively. The $\boldsymbol{F_{32}}$ features are processed through the global self-attention module, the fully connected layer,

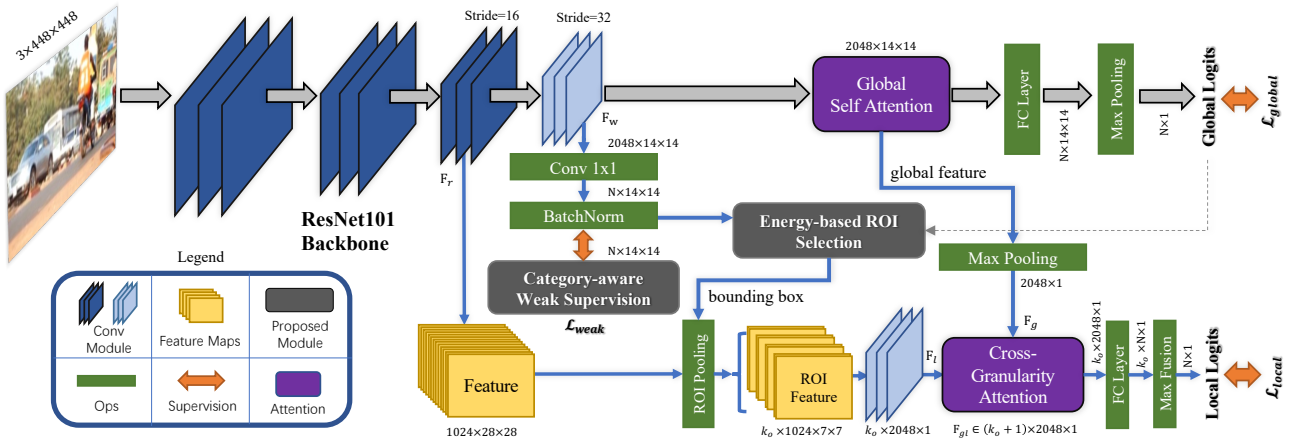

Figure 1: The overview of our proposed framework.

and a max-pooling layer to produce the prediction. Binary cross-entropy is then used as the objective function to train this branch:

$$\mathcal{L}_{global} = -\sum_{j}^{L}(y_j \log(\overline{y}_j) + (1-y_j)\log(1-\overline{y}_j)), \qquad (1)$$

where $y_j \in \{0,1\}^L$ is the ground-truth of the j-th category of the image, which is a multi-hot binary vector with a set of L labels in total. $\overline{y}_j$ is the prediction of the global branch.

In the local branch, the main idea is to extract high-quality region proposals while eliminating large-scale noisy ones by using the proposed category-aware weak supervision and energy-based ROI selection mechanism. In detail, let $\boldsymbol{F_w} \in \mathbb{R}^{C_w \times H_w \times W_w}$ denote the feature maps that will be downsampled and used for weak supervision, where H_w, W_w and C_w are the height, width, and channels of the feature map, respectively. In our default experimental setup, we make $\boldsymbol{F_w} = \boldsymbol{F_{32}}$. The $\boldsymbol{F_w}$ feature map is then connected to a 1×1 convolution, reducing the number of channels to the number of categories N, and then connected to the batch normalization layer.

After batch normalization, the pipeline splits into two modules: category-aware weak supervision and an energy-based ROI selection. The weak supervision is used to improve feature representations of local regions and then suppress the noisy ROI proposals generated from non-existent categories. The energy-based ROI selection allows for the automatic selection of ROIs based on the criterion of energy value rather than area size, which further filters noisy ROIs. Next, the box coordinates generated by the energy-based ROI selection are fed into the ROI pooling layer to crop $\boldsymbol{F_r}$ and produce the output as local features. Here we have $\boldsymbol{F_r} = \boldsymbol{F_{16}}$.

Finally, by taking as input $\boldsymbol{F_g}$ (global feature) and $\boldsymbol{F_l}$ (local feature associated with several categories), the proposed cross-granularity attention module can effectively explore the complementary information between global and local regions, making possible high-order interaction, including global-to-local and local-to-local (see the bottom right of Fig. 1).

3.2 Category-aware Weak Supervision

This module aims to improve feature representations of local regions by introducing auxiliary weak supervision. Since the dimensions of the feature maps have been reduced to the number of categories N, each feature map can be considered as an indicator of whether the corresponding category is active. In other words, the activation value of feature maps corresponding to non-existent categories should be suppressed, and, conversely, feature maps of existing categories should achieve a relatively high activation in the region near an instance.

Since we are unable to accurately determine the location of instances at this stage, forcing a restriction on feature map activation tends to mislead the model into a sub-optimal situation. In any case, however, feature maps of non-existent categories should always be inactive. Thus, we designed a weakly supervised loss function that considers only feature maps of non-existent categories: when some categories do not exist, we expect the activation value of each pixel in their corresponding feature maps to be 0, as illustrated in Fig. 2. Otherwise, we do not compute the loss for existing categories.

$$\mathcal{L}_{weak} = -\frac{1}{H \cdot W}\sum_{i=1}^{H}\sum_{j=1}^{W}\frac{1}{N}\sum_{k\in\neg gt} log\left(1 - S(\hat{p}_{ij}^{k}) + \delta\right). \qquad (2)$$

This is a typical binary classification loss, where H and W are the height and width of the image, respectively, $\hat{p}_{ij}^{k}$ is the activation value of the corresponding pixel on the k-th category feature map, $S(\cdot)$ represents the sigmoid function, gt stands for the set containing existing categories, N indicates the number of non-existent categories, and δ is a small value to prevent math domain errors.

The weakly supervised loss function focuses on filtering noisy activation regions that correspond to non-existent categories, which creates an inaccurate estimation of the foreground. Thus, fewer proposals are generated, and the selection of regions is more precise than those without weak supervision.

3.3 Energy-based ROI Selection

The primary purpose of energy-based ROI selection is to extract the bounding box of the local region so that the classification accuracy can be improved when the features corresponding to the ROI

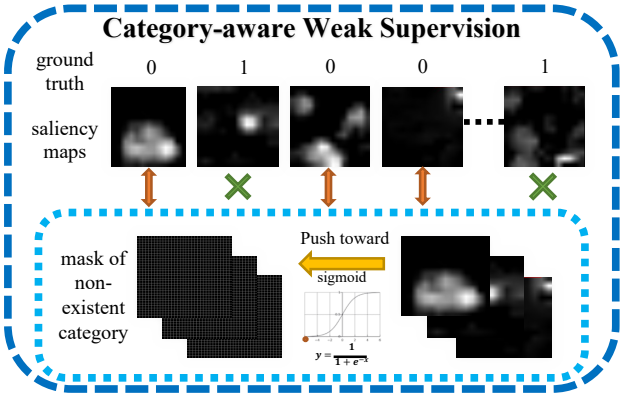

Figure 2: The Category-aware Weak Supervision.

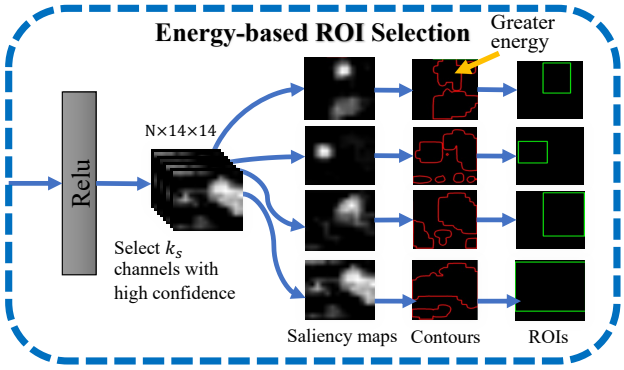

Figure 3: The Energy-based ROI Selection.

are jointly classified again. For this purpose, we need regions with positive activation values as candidates, which indicate a higher probability of the region containing the target. Specifically, the batch normalization can be regarded as an operation that automatically learns local region thresholds, while the ReLU can be seen as a filter that eliminates negative feature values so that feature maps with only positive values can be obtained. We then sort the classification scores (the logit from the global branch prediction) in descending order and select ROI candidates for each selected feature map in the top k_s feature maps.

A common strategy for selecting ROIs is to choose regions with large ambiguous areas corresponding to connected regions [42, 43] but small overall activation values. This strategy, however, leads to the incorrect area and potential performance degradation. The region with the higher overall activation is more likely to be the target we need. To this end, we chose to select the key ROIs by comparing the energy values $\mathbb{E}$ within the regions rather than by the region size: $\mathbb{E} = \sum_{i \in \mathcal{R}} \hat{p}_i$, where $\hat{p}_i$ is the i-th pixel activation within a region $\mathcal{R}$, which represents the minimum enclosing rectangle of the activated region boundary.

Once we obtain the energy value of each connected region in the feature map, we can generate high-quality ROIs by selecting the maximum energy and using the corresponding bounding boxes. As illustrated in Fig. 3, by treating k_s feature maps as hints of objectness, we find all the contours in each feature map and take the rectangular region of the contours with greater energy as the candidate. Note that we use the top k_e region of the energy value to improve the coverage for the relevant target instance, and obtain k_r bounding boxes of different sizes in the candidate energy region (further described in the supplementary material A since they are implementation details or tricks). In the end, we can obtain k_o candidate regions $\mathcal{R}$, where $k_o = k_s \times k_r \times k_e$.

3.4 Cross-granularity Attention Module

The information spread between the global and local branches has usually been ignored in previous work. Since the global and local feature maps are misaligned in the spatial dimension, simple fusion does not provide a performance boost. We therefore resort to the self-attention mechanism to achieve global-to-local interaction.

First, we perform self-attention on the global branch (see the purple rectangle at the top of Fig. 1) to capture non-local dependencies to generate a more high-level semantic feature. Let $Q_g = \phi_g^q(\boldsymbol{F_{32}}), K_g = \phi_g^k(\boldsymbol{F_{32}}), V_g = \phi_g^v(\boldsymbol{F_{32}})$, where $\phi(\cdot)$ indicates a linear projection. The global self-attention maps can be calculated as

$$\boldsymbol{A_g} = Softmax(\frac{\boldsymbol{Q_g^\top K_g}}{\sqrt{C}}), \boldsymbol{A_g} \in \mathbb{R}^{HW \times HW}. \tag{3}$$

where the dot product is scaled by a factor $\frac{1}{\sqrt{C}}$ (where C is the dimension of a query vector Q_g and key vector K_g) to prevent its result from being too large [28].

A max-pooling operation is conducted on the optimized features $\boldsymbol{A_g V_g}$ along the spatial dimension to obtain global features (i.e., $\boldsymbol{F_g} \in \mathbb{R}^{C \times 1}$), which are one of the inputs of the cross-granularity attention module.

Another input of the cross-granularity attention module comes from the feature maps of selected ROIs (i.e., $\boldsymbol{F_l} \in \mathbb{R}^{k_o \times C \times 1}$), which is obtained by passing extracted features through a series of transform layers. By concatenating the features $\boldsymbol{F_g}$ and $\boldsymbol{F_l}$ at the batch dimension (i.e., $\boldsymbol{F_{gl}} \in \mathbb{R}^{(k_o+1) \times (C \times 1)}$), we can calculate the self-attention to force information spreading between branches. Similar to Eq. (3), let $\boldsymbol{Q_{gl}} = \phi_{gl}^Q(\boldsymbol{F_{gl}}), \boldsymbol{K_{gl}} = \phi_{gl}^K(\boldsymbol{F_{gl}}), \boldsymbol{V_{gl}} = \phi_{gl}^V(\boldsymbol{F_{gl}})$. The cross-granularity attention maps can be calculated as,

$$\boldsymbol{A_{gl}} = Softmax(\boldsymbol{Q_{gl} K_{gl}^\top}), \boldsymbol{A_{gl}} \in \mathbb{R}^{(k_o+1) \times (k_o+1)}. \tag{4}$$

The output of the cross-granularity attention module, $\boldsymbol{A_{gl} V_{gl}}$, is then further sliced and summed as needed for the local branch and fed into the fully connected layer for classification to obtain local results. With the help of category-aware weak supervision and energy-based ROI selection, which restricts the local branch to concentrating on high-quality ROIs, noisy ROIs are significantly suppressed. These regions contain detailed information that the global branch cannot provide. By taking as inputs $\boldsymbol{F_g}$ and $\boldsymbol{F_l}$, the cross-granularity attention module can effectively explore the complementary information between global and local features. Moreover, self-attention allows the correlations to be captured within features of different categories, which build the label-dependency in an implicit manner.

Table 1: Comparisons of AP and mAP in % with state-of-the-art methods on the Pascal VOC 2007 dataset. Input denotes the input image size during the inference. $*$ *denotes that the model is pretrained on the MS-COCO dataset.* The best results and sub-optimal results are highlighted in red and blue, respectively. Best viewed in color.

Method	Bone	Input	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	moto	person	plant	sheep	sofa	train	tv	mAP
CNN-RNN [30]	vgg-16	224	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	64.2	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	99.7	78.6	84.0
RMIC [14]	vgg-16	224	97.1	91.3	94.2	57.1	86.7	90.7	93.1	63.3	83.3	83.3	92.8	94.4	91.6	95.1	92.3	59.7	86.0	69.5	96.4	79.0	84.5
RLSD [42]	vgg-16	224	96.4	92.7	93.8	94.1	71.2	92.5	94.2	95.7	74.3	74.3	74.2	95.4	96.2	92.1	97.9	66.9	93.5	73.7	97.5	87.6	88.5
VeryDeep [23]	vgg-16	224	98.9	95.0	96.8	95.4	69.7	90.4	93.5	96.0	74.2	74.2	87.8	96.0	96.3	93.1	97.2	70.0	92.1	80.3	98.1	87.0	89.7
ResNet101 [13]	res-101	448	99.1	97.3	96.2	94.7	68.3	92.9	95.9	94.6	77.9	77.9	85.1	94.7	96.8	94.3	98.1	80.8	93.1	79.1	98.2	91.1	90.8
HCP [35]	vgg-16	-	98.6	97.1	98.0	95.6	75.3	94.7	95.8	97.3	73.1	73.1	80.0	97.3	96.1	94.9	96.3	78.3	94.7	76.2	97.9	91.5	90.9
RDAR [33]	res-101	448	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	76.5	87.7	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
FeV+LV [39]	vgg-16	224	98.2	96.9	97.1	95.8	74.3	94.2	96.7	96.7	76.7	76.7	88.0	96.9	97.7	95.9	98.6	78.5	93.6	82.4	98.4	90.4	92.0
RARL [3]	res-101	448	98.6	97.1	97.1	95.5	75.6	92.8	96.8	97.3	78.3	78.3	87.6	96.9	96.5	93.6	98.5	81.6	93.1	83.2	98.5	89.3	92.0
RCP [31]	vgg-16	-	99.3	97.6	98.0	96.4	79.3	93.8	96.6	97.1	78.0	78.0	87.1	97.1	96.3	95.4	99.1	82.1	93.6	82.2	98.4	92.8	92.5
ML-GCN [6]	res-101	448	99.5	98.5	98.6	98.1	80.8	94.6	97.2	98.2	82.3	82.3	86.4	98.2	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
SSGRL [4]	res-101	448	99.5	97.1	97.6	97.8	82.6	94.8	96.7	98.1	78.0	97.0	85.6	97.8	98.3	96.4	98.8	84.9	96.5	79.8	98.4	92.8	93.4
A-GCN [40]	res-101	448	99.3	98.6	97.8	96.0	78.8	92.9	97.0	97.3	80.8	95.3	82.8	97.6	98.5	95.5	98.6	83.4	97.5	83.8	98.8	91.8	93.1
CoP [36]	res-101	448	99.9	98.4	97.8	98.8	81.2	93.7	97.1	98.4	82.7	94.6	87.1	98.1	97.6	96.2	98.8	83.2	96.2	84.7	99.1	93.5	93.8
DSDL [44]	res-101	448	99.8	98.7	98.4	97.9	81.9	95.4	97.6	98.3	83.3	95.0	88.6	98.0	97.9	95.8	99.0	86.6	95.9	86.4	98.6	94.4	94.4
MCAR [1]	res-101	448	99.7	99.0	98.5	98.2	85.4	96.9	97.4	98.9	83.7	95.5	88.8	99.1	98.2	95.1	99.1	84.8	97.1	87.8	98.3	94.8	94.8
Ours	res-101	448	99.9	98.7	96.2	99.7	85.7	97.3	98.3	99.1	76.0	95.2	95.1	99.4	99.9	97.6	99.0	80.5	96.8	93.7	99.2	96.6	95.2
Ours*	res-101	448	99.9	99.1	97.0	99.5	89.1	99.1	99.5	98.4	81.0	99.9	93.8	99.2	99.9	98.1	99.1	87.4	99.2	92.5	99.5	99.1	96.5

3.5 Network Training

There are three kinds of losses combined to train the whole network. ① $\mathcal{L}_{global}$ defined in Eq. (1). ② $\mathcal{L}_{weak}$ to suppress noisy proposals of ROI, which are described in Eq. (2). ③ $\mathcal{L}_{local}$ is a cross entropy loss used to calculate the loss of local branches as follows,

$$\mathcal{L}_{local} = -\sum_{j}^{L}(y_j \log(\widetilde{y}_j) + (1-y_j)\log(1-\widetilde{y}_j)), \tag{5}$$

where $y_j \in \{0,1\}^L$ is the ground-truth of the j-th category of the image, which is actually a multi-hot binary vector with a set of L labels in total. $\overline{y}_j$ is the prediction after layers of cross-granularity attention, linear projection and max fusion in the local branch.

Finally, the overall loss function is,

$$\mathcal{L}_{total} = \mathcal{L}_{global} + \mathcal{L}_{weak} + \mathcal{L}_{local}. \tag{6}$$

The proposed framework is trained by reducing $\mathcal{L}$ in an end-to-end fashion. The final result is the average of the prediction results for both global and local branches.

4 EXPERIMENT

In this section, we compare our proposed model with the state-of-the-art multi-label classification methods on two popular benchmark datasets: Pascal VOC 2007 [8] and Microsoft COCO [18]. In addition, comprehensive ablation and qualitative studies of the proposed method are also provided.

Implementation Details. We adopt ResNet101 [13] as the backbone, which is pre-trained on ImageNet [16]. In the training phase, the image is first scaled to (N+64)×(N+64), and then cropped at five scales [1.0, 0.875, 0.75, 0.66, 0.5] as suggested in [6] to avoid over-fitting. Finally, the cropped patches are further resized to N×N. During inference, the image is directly scaled to the size of N×N, normalized for evaluation. In addition, we set k_s=4, k_r=3, k_e=2 in our experiments. Our network was trained on 8 GPUs using the SGD algorithm with a total of 50 epochs, a batch size of 16, and initial learning rates of 0.05 and 0.0125 for VOC 2007 and MS-COCO, respectively, with a multi-step learning rate decay (i.e., [30, 40]).

Evaluation Metrics. We employ the average precision (AP) for each category, and the mean average precision (mAP) overall categories to evaluate all the methods. In the experiments, we computed the overall precision, recall, F1 (OP, OR, OF1) and per-class precision, recall, F1 (CP, CR, CF1) for comparison.

4.1 Comparisons with State-of-the-Arts

Result on VOC 2007. The Pascal VOC 2007 dataset contains 9,963 images of 20 object categories, with about 1.5 labels per image. We compare our method against the following SoTA methods: CNN-RNN [30], RMIC [14], RLSD [42], VeryDeep [23], ResNet101 [13], HCP [35], RDAR [33], FeV+LV [39], RARL [3], RCP [31], ML-GCN [6], SSGRL [4], A-GCN [40], CoP [36], DSDL [44] and MCAR [1].

As shown in Tab. 1, with an input size of 448×448, our method achieves 95.2% mAP, *a new state-of-the-art performance on VOC 2007 dataset.* It outperforms the previous methods such as SSGRL [4] by 1.8%, CoP [36] by 1.4%, ML-GCN [6] by 1.2%. Finally, our method achieves 96.5% mAP with the COCO pretrained model.

Result on MS-COCO. MS-COCO contains 122,218 images of 80 object labels, with about 2.9 labels per image. We followed the official split of 82,081 images for training and 40,137 images for testing. We compare our method against the following SoTA methods: CNN-RNN [30], RLSD [42], RDAR [33], RARL [3], DELTA [41], ResNet101 [13], SRN [45], Acfs [12], MultiEvid [10], DecoupleNet [19], ML-GCN [6], CoP [36], DSDL [44], CSRA [46], MCAR [1], SSGRL [4], KGGR [2], A-GCN [40] and C-Tran [17].

As shown in Tab. 2, our method achieves 85.3% mAP, which outperforms CoP [36] by 4.2%, DSDL [44] by 3.6%, ML-GCN [6] by 2.3%, CSRA [46] by 1.8%, respectively. Finally, our method achieves 88.8% mAP with a semi-weakly supervised [38] ResNeXt-101 32x8d [37] as the visual feature extractor, surpassing other methods including the transformer-based. The remarkable improvement owes to our proposed framework, which guarantees the generation of

Table 2: Comparisons with state-of-the-art methods on the MS-COCO dataset. Input denotes the input image size during the inference. $*$ *denotes that the backbone is replaced with ResNeXt-101 32x8d pre-trained on ImageNet using semi-supervised methods.* The best results and sub-optimal results are highlighted in red and blue, respectively. Best viewed in color.

Methods	Bone	Input	mAP	ALL						TOP3					
				CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
CNN-RNN [30]	vgg-16	224	-	-	-	-	-	-	-	66.0	55.6	60.4	69.2	66.4	67.8
RLSD [42]	vgg-16	224	-	-	-	-	-	-	-	67.6	57.2	62.0	70.1	63.4	66.5
RDAR [33]	vgg-16	448	-	-	-	-	-	-	-	79.1	58.7	67.4	84.0	63.0	72.0
RARL [3]	vgg-16	448	-	-	-	-	-	-	-	78.8	57.2	66.2	84.0	61.6	71.1
ResNet101 [13]	res-101	224	77.3	80.2	66.7	72.8	83.9	70.8	76.8	84.1	59.4	69.7	89.1	62.8	73.6
SRN [45]	res-101	224	77.1	81.6	65.4	71.2	82.7	69.9	75.8	85.2	58.8	67.4	87.4	62.5	72.9
Acfs [12]	res-101	224	77.5	77.4	68.3	72.2	79.8	73.1	76.3	85.2	59.4	68.0	86.6	63.3	73.1
MultiEvid [10]	res-101	448	-	80.4	70.2	74.9	85.2	72.5	78.4	84.5	62.2	70.6	89.1	64.3	74.7
DecoupleNet [19]	res-101	448	82.2	83.1	71.6	76.3	84.7	74.8	79.5	-	-	-	-	-	-
ML-GCN [6]	res-101	448	83.0	85.1	72.0	78.0	85.8	75.4	80.3	89.2	64.1	74.6	90.5	66.5	76.7
CoP [36]	res-101	448	81.1	81.2	70.8	75.8	83.6	73.3	78.1	86.4	62.9	72.7	88.7	65.1	75.1
DSDL [44]	res-101	448	81.7	84.1	70.4	76.7	85.1	73.9	79.1	88.1	62.9	73.4	89.6	65.3	75.6
CSRA [46]	res-101	448	83.5	84.1	72.5	77.9	85.6	75.7	80.3	88.5	64.2	74.4	90.4	66.4	76.5
MCAR [1]	res-101	448	83.8	85.0	72.1	78.0	88.0	73.9	80.3	88.1	65.5	75.1	91.0	66.3	76.7
Ours	res-101	448	85.3	86.3	74.4	79.9	87.5	77.6	82.3	89.6	65.5	75.7	91.7	67.9	78.1
SSGRL [4]	res-101	576	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 [2]	res-101	576	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
MCAR [1]	res-101	576	84.5	84.3	73.9	78.7	86.9	76.1	81.1	87.8	65.9	75.3	90.4	67.1	77.0
A-GCN [40]	res-101	576	85.2	84.7	75.9	80.1	84.9	79.4	82.0	88.8	66.2	75.8	90.3	68.5	77.9
C-Tran [17]	res-101	576	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
Ours	res-101	576	86.7	86.1	75.9	80.6	88.5	78.9	83.4	89.0	66.1	75.8	92.3	68.6	78.7
Ours*	next-101	576	88.8	88.8	75.9	81.9	88.9	79.2	83.8	91.3	66.1	76.7	92.3	69.1	79.0

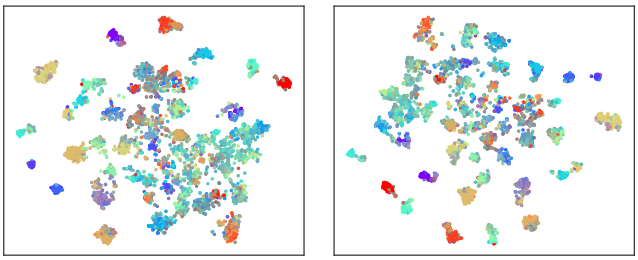

Figure 4: t-SNE visualization for models trained without weak supervision (left) and with the combination of weak supervision (right). Each dot represents one label-level embedding of a unique tag and each color represents one class.

high-quality region proposals efficiently as well as the exploration of the complementary information between global and local.

4.2 Ablation Studies

To explain how our method works, we conducted exhaustive experiments to study the influence of different modules on the performance of multi-label image classification tasks.

Weak Supervision. To demonstrate the effectiveness of our weak supervision, we show the t-SNE [27] visualization of the label-level embeddings of samples on test sets of the MS-COCO dataset [18] in Fig. 4. These embeddings are trained with weak supervision (i.e., Eq. (2)) and without weak supervision, respectively. From Fig 4, additional weak supervision allows the embeddings of the same label to fall into more compact clusters, which are better separated from the clusters of other labels. This clearly shows the enforced distinctiveness of the embeddings with weak supervision.

Table 3: Comparison of mAP (%) of our model (Original Model), our model without cross-granularity attention (Our w/o CA), our model without cross-granularity attention and global self-attention (Our w/o CA, GA).

Methods	VOC 2007			MS-COCO		
	mAP	CF1	OF1	mAP	CF1	OF1
Original Model	95.2	89.0	90.7	85.3	79.1	82.2
Our w/o CA	94.5	88.1	90.2	83.3	77.0	80.3
Our w/o CA&GA	94.1	87.3	89.2	81.7	75.1	78.7

Cross-granularity Attention. To illustrate the effectiveness of our proposed attention, we tested three different variants separately on VOC 2007 and MS-COCO as shown in Tab. 3. In Our w/o CA, F_l is directly used to obtain the classification results, and no F_g is involved in the process. In Our w/o CA&GA, F_{32} is directly used for classification instead of feeding into the self-attention module in the global branch.

Cross-granularity attention can effectively explore the complementary information between branches, thus enabling global to local interaction. In addition, inter-pixel correlations can also be captured due to global self-attention. Consequently, both advantages guarantee a boost in the performance, which verified the effectiveness of both modules.

Number of Feature Maps k_s Chosen for ROI Selection. As previously described, our proposed method needs fewer region proposals to achieve better performance. We choose the feature maps ranking top-k_s scores from the global branch to find region proposals. Tab. 4 shows the performance of our model under the different k_s selected. The base method represents the result of no region selected. Since the VOC dataset contains an average of 1.5

categories per image and a maximum of 6 categories and the COCO dataset contains an average of 3.5 categories per image, increasing k_s will result in more redundant non-existence categories, which will distract the model's attention and degrade its performance. From Tab. 4, our model achieves the best result with k_s=4. This proves that our method for region generation is very efficient.

Table 4: Ablation study on the impact of the number of high-scoring heat maps on the performance. Experiments are conducted on the COCO dataset.

Num.	mAP	ALL					
		CP	CR	CF1	OP	OR	OF1
Base	83.2	84.0	71.2	77.1	86.6	75.2	80.5
Top2	84.1	83.8	73.5	78.3	86.1	77.1	81.3
Top4	85.3	86.3	74.4	79.9	87.6	77.6	82.3
Top8	84.7	86.4	73.6	79.5	86.4	76.7	81.3

Training Loss from Different branches. Our whole network has three kinds of training loss: $\mathcal{L}_{global}$, $\mathcal{L}_{weak}$, and $\mathcal{L}_{local}$. Tab. 5 shows the results of different combinations of losses. The mAP on both datasets improves 1.9% and 2.1%, respectively, when utilizing the category-aware weak supervision and the energy-based ROI selection.

Table 5: Ablation study on the impact of the loss on the performance. Experiments are conducted on both VOC and COCO dataset.

Training Loss	VOC	COCO
$\mathcal{L}_{global}$	93.21	83.25
$\mathcal{L}_{global} + \mathcal{L}_{local}$	94.27	84.60
$\mathcal{L}_{global} + \mathcal{L}_{local} + \mathcal{L}_{weak}$	95.17	85.31

Feature Maps Chosen for F_r, F_w. In the global branch, the augmented samples are fed into ResNet101 [13] for visual feature extraction, extracted feature maps with resolutions 1/8, 1/16 and 1/32, denoted by F_8, F_{16} and F_{32}, respectively. As for the choice of F_w, since F_{32} contains more high-level semantic information, which is more conducive to the layer-by-layer mapping and gradient back-propagation of our category-aware weak supervision, we conjecture that F_w should be chosen as F_{32} to obtain the best performance. Regarding the choice of F_r, we refer to the common practice in object detection methods [22], as a result, F_{16} can be the only choice to obtain the best performance. In order to analyze what size of the feature map should be selected for F_w, F_r, we assign F_8, F_{16}, F_{32} to F_r, and also assign F_8, F_{16}, F_{32} to F_w, respectively. Then, the parameters of the input sizes of the convolutional and fully connected layers in the network structure are changed simultaneously to train the network and obtain the mAP values of the final experimental results. According to Tab. 6, choosing F_{16} as F_r and choosing F_{32} as F_w will produce the best performance, which not only proves our conjecture but also further points out the validity and necessity of the proposed pipeline structure.

Visualization of Results. Figure 5 illustrates how the model can produce performance improvements. On the left is the original image, and the four images on the right are cropped areas used for local re-classification. These four crops were sent to the proposed local branch, where they were re-classified and made a correction for our final results. The strategy allows the model to concentrate on smaller objects and complex samples than the baseline and obtain higher classification performance, which further illustrates the effectiveness of the proposed method in this paper.

Table 6: Ablation study on the impact of the chosen for F_r, F_w on the performance. Experiments are conducted on the VOC dataset.

F_r, F_w	$F_w \leftarrow F_8$	$F_w \leftarrow F_{16}$	$F_w \leftarrow F_{32}$
$F_r \leftarrow F_8$	92.89	92.01	92.81
$F_r \leftarrow F_{16}$	91.00	93.27	**95.17**
$F_r \leftarrow F_{32}$	93.02	93.46	94.00

Despite the impressive performance improvements of our method, the computational cost (MACs) increases by only 14% compared to the baseline [13]. In addition, to further analyze the effectiveness of each module, we conducted extensive experiments, including experiments on comparison of energy-based strategy versus area-size-based strategy, the specific descriptions and discussions of which are presented in supplementary material B.

5 CONCLUSION

In this work, we propose a novel multi-label image classification method to alleviate the limitations of previous region-based approaches. Specifically, this paper proposes category-aware weak supervision as well as energy-based ROI selection to generate region proposals. Moreover, by using cross-granularity attention to explore the correlation between global and local, label dependencies can be established implicitly. Experimental results on two benchmark datasets demonstrate that our framework achieves superior performance over state-of-the-art methods. In addition, the effectiveness of each component was also carefully studied.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (2021ZD0111000), National Natural Science Foundation of China (62176092), Shanghai Science and Technology Commission (21511100700), Natural Science Foundation of Shanghai (20ZR1417700).

ground truth: motorbike, person
baseline: person (0.9901),
proposed method: motorbike (0.5166), person (0.9928),

(a)

ground truth: chair, person, tvmonitor
baseline: person (0.9989),
proposed method: person (0.9937), tvmonitor (0.5259),

(b)

ground truth: bicycle, motorbike, person
baseline: bicycle (1.0), person (0.9981),
proposed method: bicycle (1.0), motorbike (0.9944), person (0.8625),

(c)

ground truth: cat, chair, person, pottedplant, sofa
baseline: person (1.0),
proposed method: cat (0.5681), person (0.9993),

(d)

Figure 5: Visualization of results. The first column shows the original input image, and the second to fifth columns show the visualisation of the five activation maps of the categories corresponding to the highest scores in the global branch, arranged in descending order. Take (b) as an illustration. In the global branch, the model can only detect the category of person, while the confidence scores for TV monitor is only 0.33, which is less than 0.5, leading to the multi-label classification result being significantly different compared to the ground-truth. However, after local cropping and feeding into the local branch for re-classification, the final confidence of the TV monitor improves from 0.33 to 0.53. We can observe that the re-classification of the model for the local region can effectively improve the multi-label classification performance.

REFERENCES

[1] Hong-Yu Zhou Bin-Bin Gao. 2021. Learning to Discover Multi-Class Attentional Regions for Multi-Label Image Recognition. *TIP* 30 (2021), 5920–5932.

[2] Tianshui Chen, Liang Lin, Riquan Chen, Xiaolu Hui, and Hefeng Wu. 2022. Knowledge-Guided Multi-Label Few-Shot Learning for General Image Recognition. *TPAMI* 44, 3 (2022), 1371–1384. https://doi.org/10.1109/TPAMI.2020.3025814

[3] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin. 2018. Recurrent attentional reinforcement learning for multi-label image recognition. *AAAI* 32, 1 (2018).

[4] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. 2019. Learning semantic-specific graph representation for multi-label image recognition. In *CVPR*. 522–531.

[5] Wei Chen, Yu Liu, Weiping Wang, Erwin Bakker, Theodoros Georgiou, Paul Fieguth, Li Liu, and Michael S Lew. 2021. Deep Image Retrieval: A Survey. *arXiv: 2101.11282* (2021).

[6] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. 2019. Multi-label image recognition with graph convolutional networks. In *CVPR*. 5177–5186.

[7] Ming-Ming Cheng, Ziming Zhang, Wen-Yan Lin, and Philip Torr. 2014. BING: Binarized normed gradients for objectness estimation at 300fps. In *CVPR*. 3286–3293.

[8] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2015. The pascal visual object classes challenge: A retrospective. *IJCV* 111, 1 (2015), 98–136.

[9] Dhatri Ganda and Rachana Buch. 2018. A survey on multi label classification. *Recent Trends in Programming Languages* 5, 1 (2018), 19–23.

[10] Weifeng Ge, Sibei Yang, and Yizhou Yu. 2018. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *CVPR*. 1277–1286.

[11] Ross Girshick. 2015. Fast r-cnn. In *ICCV*. 1440–1448.

[12] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. 2019. Visual attention consistency under image transforms for multi-label image classification. In *CVPR*. 729–739.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*. 770–778.

[14] Shiyi He, Chang Xu, Tianyu Guo, Chao Xu, and Dacheng Tao. 2018. Reinforced multi-label image classification by exploring curriculum. *AAAI* 32, 1 (2018).

[15] Mate Kisantal, Zbigniew Wojna, Jakub Murawski, Jacek Naruniec, and Kyunghyun Cho. 2019. Augmentation for small object detection. *arXiv preprint arXiv:1902.07296* (2019).

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. *NeurIPS* 25 (2012), 1097–1105.

[17] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. 2021. General Multi-Label Image Classification With Transformers. In *CVPR*. 16478–16488.

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*. Springer, 740–755.

[19] Luchen Liu, Sheng Guo, Weilin Huang, and Matthew R Scott. 2019. Decoupling category-wise independence and relevance with self-attention for multi-label image classification. In *ICASSP*. IEEE, 1682–1686.

[20] Weiwei Liu, Xiaobo Shen, Haobo Wang, and Ivor W Tsang. 2020. The Emerging Trends of Multi-Label Learning. *arXiv: 2011.11197* (2020).

[21] Yongcheng Liu, Lu Sheng, Jing Shao, Junjie Yan, Shiming Xiang, and Chunhong Pan. 2018. Multi-label image classification via knowledge distillation from weakly-supervised detection. In *ACMMM*. 700–708.

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *TPAMI* 39, 06 (2017), 1137–1149.

[23] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv: 1409.1556* (2014).

[24] Nathan Thom and Emily M Hand. 2020. Facial Attribute Recognition: A Survey. *Computer Vision: A Reference Guide* (2020), 1–13.

[25] Vaishali S Tidake and Shirish S Sane. 2018. Multi-label classification: a survey. *International Journal of Engineering and Technology* 7, 4.19 (2018), 1045–1054.

[26] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. 2013. Selective search for object recognition. *IJCV* 104, 2 (2013), 154–171.

[27] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*. 6000–6010.

[29] Changhu Wang, Shuicheng Yan, Lei Zhang, and Hong-Jiang Zhang. 2009. Multi-label sparse coding for automatic image annotation. In *CVPR*. IEEE, 1643–1650.

[30] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. 2016. Cnn-rnn: A unified framework for multi-label image classification. In *CVPR*. 2285–2294.

[31] Meng Wang, Changzhi Luo, Richang Hong, Jinhui Tang, and Jiashi Feng. 2016. Beyond object proposals: Random crop pooling for multi-label image recognition. *TIP* 25, 12 (2016), 5678–5688.

[32] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. 2020. Multi-label classification with label graph superimposing. *AAAI* 34, 07 (2020), 12265–12272.

[33] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. 2017. Multi-label image recognition by recurrently discovering attentional regions. In *ICCV*. 464–472.

[34] Yunchao Wei, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. 2014. CNN: Single-label to multi-label. *arXiv: 1406.5726* (2014).

[35] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. 2015. HCP: A flexible CNN framework for multi-label image classification. *TPAMI* 38, 9 (2015), 1901–1907.

[36] Shiping Wen, Weiwei Liu, Yin Yang, Pan Zhou, Zhenyuan Guo, Zheng Yan, Yiran Chen, and Tingwen Huang. 2020. Multilabel image classification via feature/label co-projection. *TSMC* 51, 11 (2020), 7250–7259.

[37] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *CVPR*. 1492–1500.

[38] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546* (2019).

[39] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. 2016. Exploit bounding box annotations for multi-label object recognition. In *CVPR*. 280–288.

[40] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. 2020. Attention-Driven Dynamic Graph Convolutional Network for Multi-Label Image Recognition. In *ECCV*. Springer, 649–665.

[41] Wan-Jin Yu, Zhen-Duo Chen, Xin Luo, Wu Liu, and Xin-Shun Xu. 2019. DELTA: A deep dual-stream network for multi-label image classification. *Pattern Recognition* 91 (2019), 322–331.

[42] Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, and Jianfeng Lu. 2018. Multi-label image classification with regional latent semantic dependencies. *TMM* 20, 10 (2018), 2801–2813.

[43] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *CVPR*. 2921–2929.

[44] Fengtao Zhou, Sheng Huang, and Yun Xing. 2020. Deep Semantic Dictionary Learning for Multi-label Image Classification. *arXiv: 2012.12509* (2020).

[45] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. 2017. Learning spatial regularization with image-level supervisions for multi-label image classification. In *CVPR*. 5513–5522.

[46] Ke Zhu and Jianxin Wu. 2021. Residual Attention: A Simple but Effective Method for Multi-Label Recognition. *CoRR* abs/2108.02456 (2021). arXiv:2108.02456 https://arxiv.org/abs/2108.02456

[47] C Lawrence Zitnick and Piotr Dollár. 2014. Edge boxes: Locating object proposals from edges. In *ECCV*. Springer, 391–405.

SUPPLEMENTARY MATERIAL

A ADDITIONAL DESCRIPTIONS

The strategy of Energy-based ROI generation. Examples of our energy-based ROI selection are shown in Figure 6. It can be demonstrated that more optimal regions can be selected by energy-based selection than by area-size-based selection.

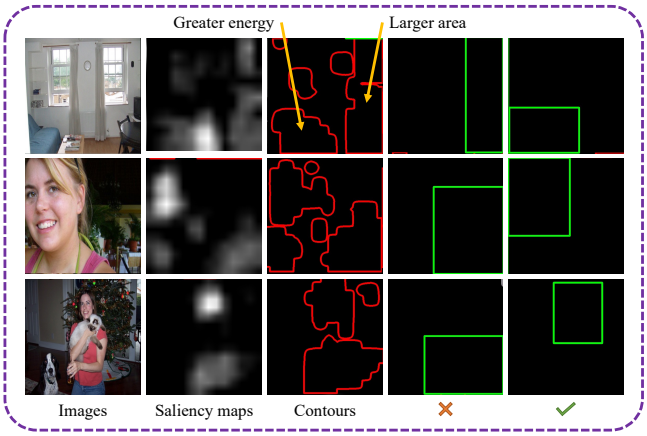

Figure 6: The strategy of Energy-based ROI generation. The first column on the left is the original input image, the second column is the saliency maps, the third column is the contours, the fourth is the sub-optimal selection strategy (based on area size) and the fifth column is the optimal selection strategy (based on energy).

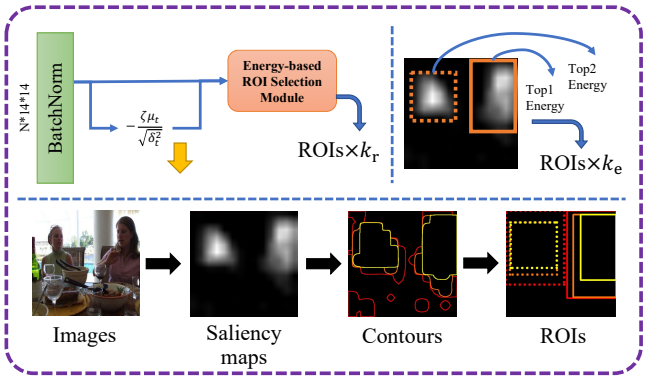

Figure 7: The ROI selection strategy when dealing with small objects. The red, orange and yellow rectangles represent the different proposals obtained by scaling strategy in the same region, while the solid line represents the candidate with the highest energy value and the dashed line represents the candidate with the second highest energy value.

To ensure that more semantic information about small objects is available for local branches since MS-COCO contains a large number of small objects [15, 18], more regions are needed to improve the coverage of small areas. Scaling to obtain more bounding boxes of smaller size can solve our needs. Since we applied the *ReLU* activation function (in the front of the energy-based ROI selection) to filter the negative activation values and use the regions with positive activation values as candidate regions, we can change the size of the bounding box to obtain more bounding boxes by adjusting each pixel value in the cropped feature map, instead of scaling the size of bounding boxes directly. In addition, to scale the size of bounding boxes directly, manual specifying thresholds for each separate dataset is inevitable, which may reduce the model's generalizability and also be very labor-consuming and computationally demanding.

Specifically, our proposed scaling method is adaptive to the selected feature maps. First, for each mini-batch, Batch Normalization performs a scaling operation on the input features.

$$\widehat{x_i} = \frac{x_i - \mu_B}{\sqrt{\delta_B^2 + \epsilon}}, \tag{7}$$

where $\mu_B = \frac{1}{m}\sum_i^m x_i$, $\delta_B^2 = \frac{1}{m}\sum_i^m (x_i - \mu_B)^2$. To shift the features after scaling, we reduce the running mean μ_t to obtain a feature map with a smaller connected region. To ensure that the final output is on the same scale, we also divide the value of the offset by the running variance $\sqrt{\delta_t^2}$. Finally, the extra output features became,

$$\widehat{x_i} = \frac{x_i - \mu_B}{\sqrt{\delta_B^2 + \epsilon}} - \frac{\zeta\mu_t}{\sqrt{\delta_t^2}}, \tag{8}$$

where ζ is the empirical value we use to adjust the magnitude of the offset.

Since we consider the connected areas of positive values in the feature map as indicators for obtaining the contours, when the number of positive pixels within the feature map decreases, the size of the corresponding bounding box decreases accordingly. Then, we find the minimum enclosing rectangles for each contour and save them as bounding boxes. In this way, k_r pooled feature maps are obtained, where $(kr-1)$ is generated by our method and the other 1 is the feature map of the original features after pooling. In our experiments, ζ will be given two different values to obtain k_r=3 bounding boxes in each high energy region.

In summary, for each image, we pick the top k_s feature maps of categories with the highest confidence score in the global branch; for each feature map, we choose the top k_e connected region with the highest energy value; for each connected region, we perform scaling to obtain the k_r bounding box, as shown in the Figure 7. In the end, we can acquire $k_o = k_s \times k_r \times k_e$ number of candidate regions.

B ADDITIONAL EXPERIMENTS

To further analyze the effectiveness of each module, we conducted extensive experiments.

Energy-based ROI selection vs. Area-size-based ROI selection. To further illustrate the necessity of our proposed strategy of energy-based ROI selection, which has been explained in the previous analysis, we conducted ablation experiments on the VOC dataset. The experiments compared energy-based ROI selection (the region selection strategy based on energy value) with area-size-based ROI selection (the region selection strategy based on

Figure 8: Comparison between our proposed ROI selection method and object detection technique, selective search The green bounding boxes on the left are the proposals generated by the selective search, while the bounding boxes on the right are the proposals produced by our method, where red, orange and yellow bounding boxes represent the different proposals obtained by scaling strategy in the same region. Besides, the solid line represents the option with the highest energy value and the dashed line represents the option with the second highest energy value.

Table 8: Experiments on computational cost and efficiency analysis of the proposed model. The experiments are based on 448-size image input to count the number of parameters and computational cost of each comparison model during inference. mAP values are obtained by inference on the MS-COCO 2014 dataset.

Methods	MACs(G)	Params(M)	mAP
ResNet101	31.39	44.55	79.31
Our w/o CA&GA	32.43	57.63	81.67
Full model	35.98	74.42	85.27

Table 7: Ablation study on the impact of the chosen for Energy-based ROI selection vs. Area-size-based ROI selection on the performance. Experiments are conducted on both VOC and COCO datasets.

Methods	Energy-based	Area-size-based
VOC	95.17	95.00
COCO	85.31	84.23

area size, i.e., a preference for large sized bounding boxes), and the results are shown in the Table 7.

As can be seen in Table 7, the classification performance of the energy-based ROI selection strategy is still improved over that of the area-based strategy, although the difference is not significant. Since we want to obtain the best performance in our experiments, we decided to use the energy-based ROI selection strategy.

Computational Costs. To illustrate the efficiency and computational costs of our model, we counted the number of multiply-accumulate (MAC) of resnet101, the model without attention (Our w/o CA&GA), and the full model.

As can be seen from Table 8, there is only a slight increase in the computational cost ($\approx$ 14%), despite the acceptable increase in the number of parameters of the model (mainly on the cross-granularity attention), which further illustrates the great efficiency of our model. This small increase in computation greatly improves the performance, making the model more practical. Additionally, we analyze the growth of the corresponding computational cost of varying the input sizes. As can be seen from Figure 10, the additional increase in computational cost of our model is small in proportion compared to the increase in the backbone computation due to the resolution change in input image.

Visualization of Region Proposals. To demonstrate the advantage of our method in generating region proposals, Figure 8 gives a visual comparison between the proposals generated by our method and that generated by selective search [26]. It is manifest from the Figure 8 that the proposals generated by our method are more accurate and efficient (less number is needed) compared to selective search when locating possible objects in a given image,

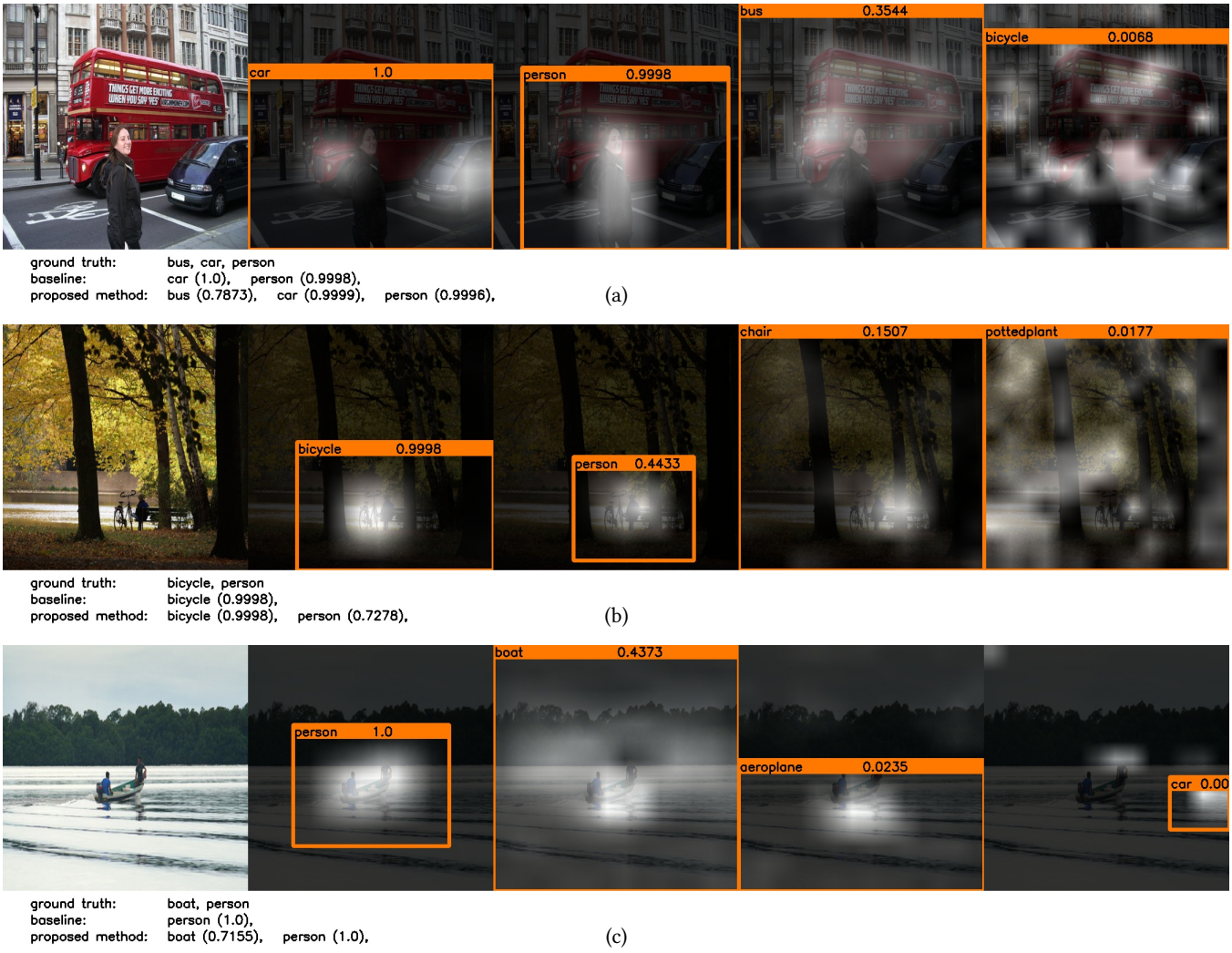

Figure 9: Visualization of results. The first column shows the original input image, and the second to fifth columns show the visualization of the five activation maps of the categories corresponding to the highest scores in the global branch, arranged in descending order. Although the recognition is relatively accurate, some bounding boxes are too large, which affects the efficiency of the methods.

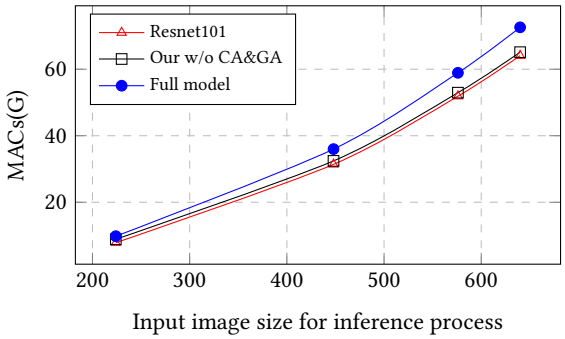

Figure 10: Experiment on the computational cost of the model. The experiments are based on different sizes of image input and are used to compare the effect of image size on the computational cost.

which illustrates the effectiveness of category-aware weak supervision and energy-based ROI selection. Within the visualization results (Figure 9), we discovered several examples that could be limitations of our methods. Since the weakly supervised loss only constrains the non-existence/absence category and the threshold selection for the foreground is automatic, the region proposal's bounding box may be slightly larger. Although it is not a major issue, large local regions can occasionally interfere with the reclassification of small objects and influence the final fusion results, which can be an area that future work may address.