Transformer-based Dual Relation Graph for Multi-label Image Recognition

Jiawei Zhao[1][*]   Ke Yan[2]   Yifan Zhao[1]   Xiaowei Guo[2]   Feiyue Huang[2]   Jia Li[1,3][†]

[1]State Key Laboratory of Virtual Reality Technology and Systems, SCSE, Beihang University

[2]Tencent Youtu Lab, Shanghai, China   [3]Peng Cheng Laboratory, Shenzhen, China

{zhaojiaweii, zhaoyf, jiali}@buaa.edu.cn,   {kerwinyan, scorpioguo, garyhuang}@tencent.com

Abstract

The simultaneous recognition of multiple objects in one image remains a challenging task, spanning multiple events in the recognition field such as various object scales, inconsistent appearances, and confused inter-class relationships. Recent research efforts mainly resort to the statistic label co-occurrences and linguistic word embedding to enhance the unclear semantics. Different from these researches, in this paper, we propose a novel Transformer-based Dual Relation learning framework, constructing complementary relationships by exploring two aspects of correlation, i.e., structural relation graph and semantic relation graph. The structural relation graph aims to capture long-range correlations from object context, by developing a cross-scale transformer-based architecture. The semantic graph dynamically models the semantic meanings of image objects with explicit semantic-aware constraints. In addition, we also incorporate the learnt structural relationship into the semantic graph, constructing a joint relation graph for robust representations. With the collaborative learning of these two effective relation graphs, our approach achieves new state-of-the-art on two popular multi-label recognition benchmarks, i.e. MS-COCO and VOC 2007 dataset.

1. Introduction

Multi-label image recognition aims at assigning multiple labels for multiple objects presented in one natural image. As a fundamental task in computer vision, multi-label image recognition can serve as prerequisites for many applications, such as weakly supervised localization and segmentation [12, 15, 44], attribute recognition [25, 22] , scene understanding [29] and recommendation systems [39, 19]. Benefiting from the development of deep learning techniques [17, 30], recent CNN-based architectures have made significant process in distinguishing multiple objects. But

---

[*]Works done while interning at Tencent Youtu Lab.

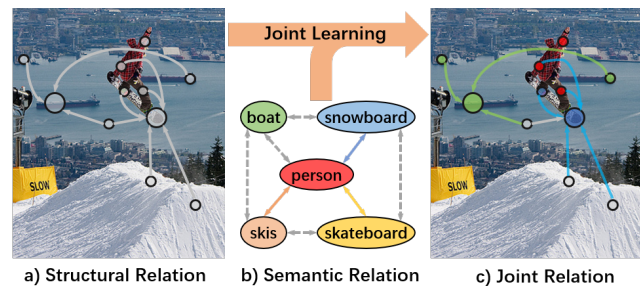[†]Jia Li is the Corresponding author. URL: http://cvteam.net

Figure 1. Motivation of the proposed dual relations. a) Structural relation provides long-term contextual relations for recognizing *snowboard*, while b) semantic relation builds dynamic correlations of co-occurred classes. These two relations jointly form a structural and semantic-aware image understanding.

the accurate parsing of multi-label images still faces great challenges, including various object scales, inconsistent visual appearances, and confused inter-class relationships.

One intuitive solution for discovering visual consistency is to enhance the feature representation with self-attention mechanisms [46, 34, 16]. For example, Wang *et al.* [34] propose to automatically discover the attentional regions with a recurrent neural network, introducing discriminative features for representation learning. Beyond these improvements, Guo *et al.* [16] propose the assumption for visual perception consistency of attention regions and then amplify these regions by a visual consistency loss. Although the spatial representations of CNNs are strengthened by these techniques, the multi-label dependencies are not explicitly modeled, which is crucial for the understanding of multi-label relationships.

To tackle this problem, recent ideas propose to learn the inter-class relationships based on the co-occurrences of multiple classes (*e.g.*, the *snowboard* should be attached with higher confidence if *person* occurs). Pioneer works tend to model this relationship with Recurrent Neural Networks (RNN) [32, 2, 40], while the co-occurred labels can be gradually refined in the sequential predictions. Inspired by the advanced improvements of Graph Convolutional

Network (GCN) [20], tens of works [8, 5, 33, 41, 6, 3] propose to construct label-wise relationships based on the semantic meanings or statistical co-occurrences. For example, Chen *et al.* [8] propose to construct the graph model with the semantic word embeddings, forming a static label-wise relationship. However, this static relationship neglects the characteristics of each image, leading to negative optimization for objects with less-frequent co-occurrences. To solve this problem, some works [5, 41, 45] propose to construct dynamic graph based on the image-specific descriptors of high-level semantic features. Nevertheless, this modeling of multi-label relationship still shows its limitations: 1) the spatial interactions of contextual objects are not implicitly modeled in the label-wise relation, 2) the features of high-level semantics are somewhat unstable and do not reflect specific semantic classes, 3) the representation of long-range context and various object scales are not considered.

To efficiently solve these deficiencies as well as the major challenges in multi-label image recognition, we propose to model a joint structural and semantic relationship of multi-label objects in one image. As in Fig. 1, considering the co-occurrence of semantic labels in vanilla class-wise relation models [8], absent classes would also be hallucinated (*i.e.*, *skis* and *skateboard*). Beyond these demerits, the semantic meaning of one object should be not only decided by its intrinsic attributes but also the contextual information. In Fig. 1 a), the appearance of *snowboard* is visually similar with *skis* and *skateboard*, and also shows high co-occurrence frequencies in Fig. 1 b). But human being can easily recognize it as a *snowboard* based on the long-range contextual information (*snow*) and even *person* appearances. Based on these investigations, we propose a collaborative framework with joint structural and semantic relational graphs in Fig. 1 c), which depict the position-wise and class-wise relationship respectively.

To construct the structural graph, we make the first attempt to introduce the Transformer architecture [31] into multi-label recognition. This new attempt greatly broadens the receptive capability of conventional CNNs and draws position-wise long-term dependencies for object contextual correlations (Fig. 1 a)). Starting from this novel design, we further propose a cross-scale attention module to enhance the perceptive ability of various object scales. For the construction of semantic graph, we aim at constructing dynamic relation which is aware of object emergence and structural embedding. Different from previous works [5, 41] with implicit high-level embedding, the graph nodes in our paper are explicitly constructed with semantic-aware constraints, reflecting features of specific classes. Beyond this explicit class-wise embedding, we incorporate the learnt structural graph embedding into the semantic relation construction from two aspects: adjacent correlation construction and feature-wise complementary. These

two mechanisms efficiently endow the semantic graph with the perception of structural information, generating robust graph relations. With the collaborative learning of proposed structural and semantic relation, our proposed approach achieves state-of-the-art results on two most popular benchmarks, *i.e.*, MS-COCO [23] and PASCAL VOC [13].

In summary, our contribution is three-fold: 1) We propose a novel Transformer-based Dual Relation learning framework for multi-label image recognition tasks, which jointly models the structural and semantic information with Transformer architectures. 2) A transformer-based structural relation graph is constructed to incorporate long-term contextual information, building position-wise spatial relationships across different scales. 3) A semantic relation graph is constructed with explicit class-specific constraints and structure-aware embedding, modeling the dynamic class-wise dependencies.

2. Related Work

Multi-label Recognition. Recently multi-label recognition approaches mainly focus on two aspects, *i.e.*, spatial information and co-occurrence dependency.

Some works [35, 38, 34, 43, 46, 16] devoted to exploit spatial information for improving recognition performance. Previous pioneer works tend to coarsely locate multiple objects for recognition [35, 43, 14]. For example, Wei *et al.* [35] generate multiple object proposals [48] and aggregate their label scores to obtain the final prediction. However, the performance of localization is unstable without additional annotations of proposals. To solve this issue, recent works introduce attention mechanism to implicitly locate attentional regions and enhance the spatial representation [46, 34, 16]. For example, Zhu *et al.* [46] propose to capture spatial relationships between labels with self-attention mechanism. Wang *et al.* [34] utilize a proposal-free pipeline to iteratively locate attentional regions and capture their contextual dependencies.

Some other works [32, 24, 4, 2, 40] devoted to build co-occurrence dependency with Recurrent Neural Network (RNN) [18]. For example, Wang *et al.* [32] combine RNNs with CNNs to capture semantic label dependency and predict labels with a predefined order. Chen *et al.* [2] design an order-free RNN to avoid propagating prediction errors during inference process. However, these RNN-based methods explore limited relationships between labels in a sequential manner, hence recent works introduce Graph Convolutional Network (GCN) [20] to fully exploit pair-wise relationships [8, 5, 33, 41, 42, 36, 7, 6, 3]. For example, Chen *et al.* [8] propose a directed graph over word embedding of labels to model the label correlations. Chen *et al.* [5] build a semantic-specific graph to incorporate high-level features into word embeddings for better semantic-specific features and explore their interactions.
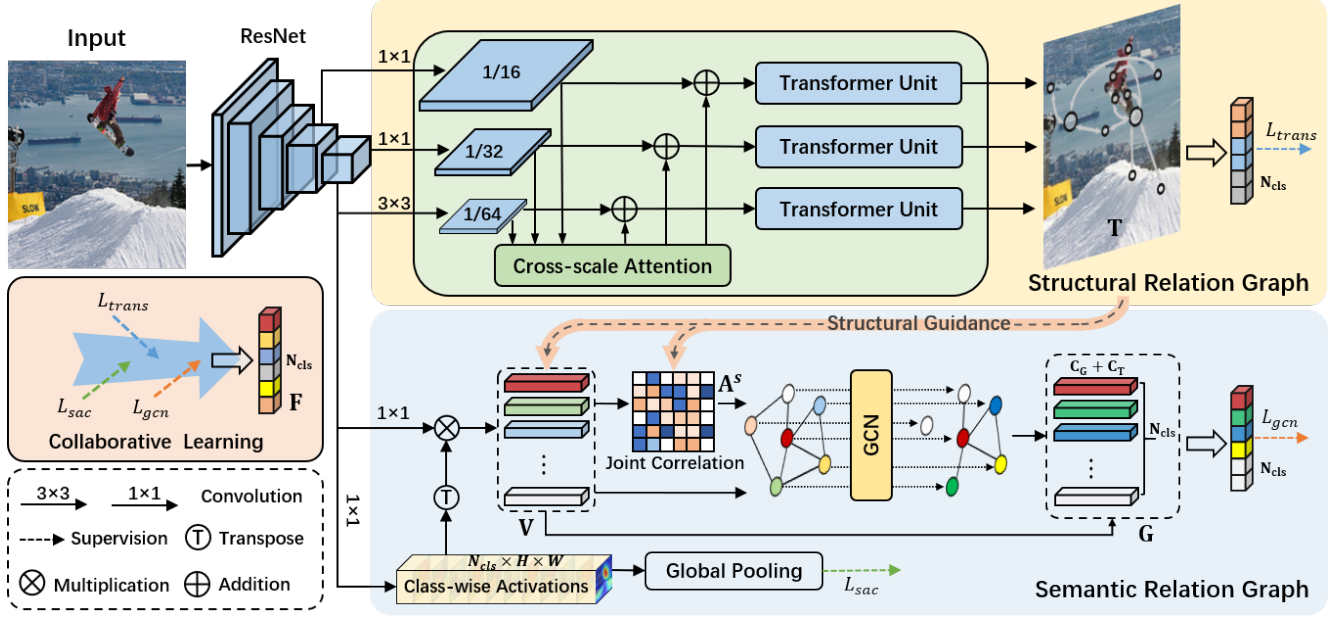
Figure 2. The overall architecture of our proposed Transformer-based Dual Relation Graph (TDRG) network, which consists of two essential modules: the structural relation graph module to incorporate long-term contextual information, the semantic relation graph module to model the dynamic class-wise dependencies.

Relationship Modeling. Different from CNNs and RNNs, transformer is recently proposed to extract intrinsic features with self-attention mechanism [27]. Transformer has demonstrated its success in natural language processing tasks [31, 10]. As the pioneer work, Vaswani *et al.* [31] first propose the vanilla Transformer architecture, which is based on self-attention mechanisms for machine translation. Transformer has not only obtained great breakthrough in NLP tasks, but also shows huge potential in Computer Vision (CV) tasks [11, 1, 47, 37, 21]. For example, recently, Dosovitskiy *et al.* [11] propose a pure transformer architecture on sequential image patches for image recognition task. Carion *et al.* [1] design a fully end-to-end object DEtection TRansformer (DETR), which shows impressive performance on object detection. Zhu *et al.* [47] introduce a deformable attention module to solve the defects of DETR, *e.g.*, poor performance on small objects. However, as an effective architecture for relationship modeling, transformer is less explored in multi-label recognition tasks.

3. Approach

In this section, we introduce a novel collaborative learning framework with joint structural and semantic relational graphs for multi-label recognition, depicting the position-wise and class-wise dependencies respectively in Fig. 2. The first structural relation graph aims to capture long-term contextual information and build spatial relationships across different scales in Section 3.1. In Section 3.2, a semantic relation graph is proposed to exploit dynamic co-

occurrence dependencies with structure-aware embedding. In the end, we joint structural and semantic relations in a collaborative learning manner in Section 3.3.

Given an input image $\mathcal{I}$, let $\{\mathbf{X}_1, \cdots, \mathbf{X}_s\} = \Phi_S(\mathcal{I})$ be the multi-scale features encoded by the backbone network Φ_S with a channel-reduction transformation, *e.g.*, 1×1 and 3×3 convolution. To construct the structural relation graph nodes $\mathbf{T}$, we introduce s transformer units $\mathcal{G}^{trans}$ to capture long-term contextual information and build position-wise relationships with a cross-scale attention module $\Psi_i(\cdot)$:

$$\mathbf{T} = \mathop{\mathrm{concat}}_{i=1}^{s}(\mathcal{G}_i^{trans}(\Psi_i(\mathbf{X}_i; \{\mathbf{X}\}_{k=1}^s))) \in \mathbb{R}^{N_T \times C_T},$$
(1)

where N_T and C_T denote the number and dimension of structural relation nodes $\mathbf{T}$ respectively. To construct the nodes of semantic relation graph $\mathbf{G}$, we model dynamic class-wise dependencies with explicit semantic-aware constraints and structural guidance:

$$\mathbf{G} = \mathcal{G}^{sem}((\mathcal{C}(\mathbf{X}), \mathbf{T}); \mathcal{A}(\mathbf{T}, \mathcal{C}(\mathbf{X}))) \in \mathbb{R}^{N_{cls} \times (C_G + C_T)},$$
(2)

where $\mathcal{G}^{sem}$ denotes the semantic graph neural network, $\mathcal{C}(\cdot)$ denotes the semantic-aware constraints, $\mathcal{A}(\cdot)$ denotes the joint relational correlation matrices of $\mathcal{G}^{sem}$, N_{cls} and C_G denote the number of categories and the dimension of semantic-specific vectors. With these two complementary relation graphs, we further conduct a collaborative learning manner to get the final prediction $\mathbf{F}$:

$$\mathbf{F} = \psi_t(\mathrm{GMP}(\mathbf{T})) \boxplus \psi_g(\mathbf{G}) \in \mathbb{R}^{N_{cls}},$$
(3)

where $\psi_{\{t,g\}}$ denote the category classifier for structural and semantic relation graph respectively, $\texttt{GMP}(\cdot)$ denotes the global max-pooling operation, $\uplus$ denotes the weighted fusion operation of two relation graphs.

3.1. Structural Relation Graph

As aforementioned, one crucial problem in multi-label recognition is to capture long-term contextual information and build structural interactions between different objects. Due to the intrinsic flaws of CNN-based architectures, the position correlations are locally conducted without perceiving global contextual information. To mitigate this issue, we make the first attempt to introduce Transformer into multi-label recognition tasks to capture long-term contextual information and build position-wise spatial relationships.

Revisiting Transformer for Structural Relation. In the field of natural language processing, the conventional transformers [31] take language sentences as input and build relationships between different semantic words from global perspectives. Different from language processing, images cannot be directly converted into sequence form. Hence, there are two popular ways to apply transformer on images, embedding transformer into CNN backbones [1] and applying transformer on sequential embedding features of image patches [11]. The latter leads to a high computation burden for network optimization with limited data. Hence in our framework, we adopt the former scheme to capture global contextual information.

We apply standard transformer encoder structure as transformer unit $\mathcal{G}_i^{trans}$ to build long-term relationships between pair-wise positions. As shown in Fig. 3, each transformer unit consists of n groups multi-head self-attention modules and feed forward networks, which is composed of two linear transformation layers. The detailed structure of multi-head self-attention module is illustrated in Fig. 3. For each head, we first adopt relative positional encoding $\mathcal{E}(\cdot)$ on the channel-wise transformed feature $\phi(\mathbf{X})$ to keep the position information:

$$\mathbf{X}_e = \mathcal{R}(\phi(\mathbf{X})) + \mathcal{E}(\mathcal{R}(\phi(\mathbf{X}))) \in \mathbb{R}^{HW \times C_T}, \quad (4)$$

where $\mathcal{R}(\cdot)$ denotes the reshape operation, which squeezes the spatial dimensions into one dimension. Then we respectively obtain query, key and value projections of encoded features by linear transformation layers. To build and enhance global position relationships, we calculate the positional correlation matrix $\mathbf{A}^p$ by query and key, and reweight value with $\mathbf{A}^p$ by multiplication:

$$\mathbf{A}^p = \texttt{softmax}(\frac{\mathbf{X}_e \mathbf{W}_Q (\mathbf{X}_e \mathbf{W}_K)^\top}{\sqrt{C_T}}),$$
$$\mathbf{H} = \mathbf{A}^p \mathbf{X}_e \mathbf{W}_V, \quad (5)$$

where $\mathbf{W}_{\{Q,K,V\}}$ are the learnable weights of query, key and value projections, $\mathbf{H}$ is the enhanced feature of one
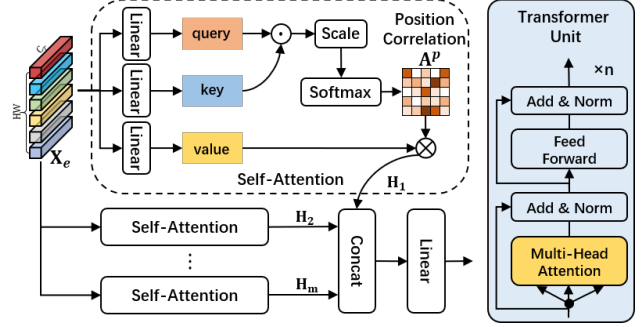

Figure 3. Illustration of transformer unit and multi-head self-attention module. Each head provides structural self-attention mechanism by multiplying query, key and value features.

head. Different heads could mine different structural relationships due to different projections. Hence employing multiple heads could capture more comprehensive structural relationships to enrich the representations. For multi heads, we concatenate results from multi heads and fuse them with a linear transformation layer.

Cross-scale Attention Transformers. Small objects tend to have lower performance in multi-label recognition due to the position information of small objects may lost in low-resolution features especially for challenging datasets, *e.g.*, MS-COCO. To address this issue, a natural idea is to consider more high-resolution features to retain the position information of small objects. In fact, high-resolution features indeed improve the performance of small objects, but also introduce more computation burden and noise, which lower the performance of other objects. Towards this concern, we propose a simple yet effective cross-scale attention module as a trade-off between performance and computation cost, which effectively improves the capacity of our structural relation graph.

To suppress noise between different scales and enhance the structure information of small objects, we propose cross-attention feature fusion strategy $\Psi_i(\cdot)$. We extract the common positions while alleviating the ambiguous ones by position-wise multiplication operation after up-sampling features of different scales. To enhance the positional information, the extracted feature is respectively down-sampled and enhanced with position-wise addition operation. Thus the structural feature $\mathbf{T}_i$ of ith scale can be formed by this serialized operation:

$$\mathbf{T}_i = \mathcal{G}_i^{trans}(\mathcal{D}(\prod_i^s \mathcal{U}(\mathbf{X}_i)) + \mathbf{X}_i), \quad (6)$$

where $\mathcal{U}(\cdot)$ and $\mathcal{D}(\cdot)$ denote up- and down-sampling operation. Then we input enhanced features into weight-sharing transformer unit to respectively capture the structural relationships with different scales, and the final feature $\mathbf{T}$ in Eq. (1) can be obtained by concatenating each $\mathbf{T}_i$.

3.2. Semantic Relation Graph

Motivation and Discussions. Motivated by the co-occurrence dependencies in multi-label learning, existing works usually resort to graph networks to model this relationship into deep CNNs. Pioneer works [8] in Fig. 4 a) tend to build static correlations of different linguistic word embedding from statistic priors. However, in this label graph, the characteristic of each image are less taken into consideration, which would lead to the hallucination of nonexistent objects and suppression of less common co-occurrences. To this end, in Fig. 4 b), semantic graphs in [5] are built to incorporate high-level features into word embedding. Despite the additional dependencies on word embedding and dataset statistics, high-level features only present *implicit* semantics for graph construction and detailed relationship between multiple objects are still neglected.

To revisit the graph construction of multi-label learning, here we propose a joint relation graph in Fig. 4 c), involving two meaningful cues for object relation discovery: 1) introducing *explicit* semantic-aware high-level features via an auxiliary semantic-aware constraint; 2) incorporating structural relationship for graph nodes as well as correlations. The former cue helps to construct an *explicit* relationship of semantic classes while the latter cue provides the graph a spatial awareness of contextual objects.

Semantic-aware Constraints. Different from previous researches using high-level features $\mathbf{X}$ for implicit semantic, here we introduce the explicit embedding with class-specific vectors $\mathbf{M} = \phi_m(\mathbf{X}) \in \mathbb{R}^{N_{cls} \times H \times W}$, which is regularized by explicit classification constraints. ϕ_m denotes the learnable 1×1 convolutional layer. Hence we conduct a high-order fusion to form the semantic-aware features $\mathbf{V}_G$:

$$\mathbf{V}_G = \mathcal{R}(\mathbf{M})\phi_g(\mathcal{R}(\mathbf{X})^\top) \in \mathbb{R}^{N_{cls} \times C_G}, \quad (7)$$

where $\phi_g(\cdot)$ denotes a dimension-reduction operation from C to C_G. However, how to ensure the representation quality of $\mathbf{V}_G$ for each class is less explored, which affects the subsequent modeling process as an important precondition. To solve this issue, we adopt a global pooling operation, *i.e.* top-k max-pooling on $\mathbf{M}$ to squeeze spatial dimensions and then apply an auxiliary loss on $\mathtt{KMP}(\mathbf{M}) \in \mathbb{R}^{N_{cls}}$ to constraint $\mathbf{M}$ for learning more accurate initial activation maps and less noise for each class.

Besides the semantic-aware vectors, we also introduce structural information from structural relation graphs to incorporate long-term contextual information and position-wise relationships:

$$\mathbf{V}_T = \mathcal{R}(\mathtt{GAP}(\mathbf{T})) \in \mathbb{R}^{N_{cls} \times C_T}, \quad (8)$$

where $\mathbf{V}_T$ denotes the structure-aware vectors and $\mathtt{GAP}$ denotes the global average pooling operation.
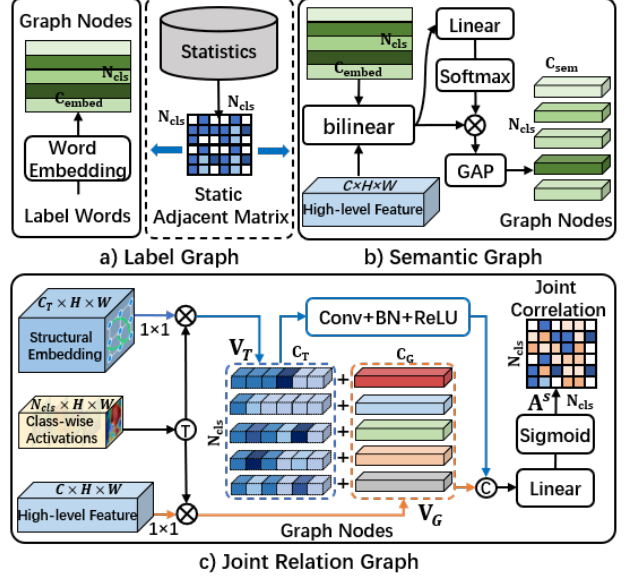

Figure 4. Illustrations of three typical graph constructions. a) Label graph [8]: building graph based on the statistical priors of label co-occurrences. b) Semantic graph [5]: incorporating high-level features beyond word embeddings. c) Our joint relation graph: building graph nodes by joint structural embedding and semantic-aware constraints and dynamically constructing the correlation matrix in a learnable manner.

Joint Relation Graph. Graph neural networks propagate messages between adjacent nodes based on correlation matrix. As in Fig. 4 c), we build the joint correlation matrix $\mathbf{A}^s$ from two aspects, *i.e.* semantic correlations $\mathbf{V}_G$ and structural correlations $\mathbf{V}_T$ in a learnable manner:

$$\mathbf{A}^s = \mathtt{sigmoid}(\varphi_c(\mathtt{concat}(\varphi_t(\mathbf{V}_T), \mathbf{V}_G))) \in \mathbb{R}^{N_{cls} \times N_{cls}}, \quad (9)$$

where $\varphi_{\{c,t\}}$ denote the learnable dimension transformation operation, *e.g.*, 1×1 convolutional layer.

Obtaining graph nodes $\mathbf{V} = \mathtt{concat}(\mathbf{V}_G, \mathbf{V}_T)$ and correlation matrix $\mathbf{A}^s$, we further model the joint co-occurrence dependencies between joint structural and semantic-aware vectors based on correlation matrix using Kipf *et al.*'s [20] Graph Convolutional Networks, which can be formulated as:

$$\mathbf{G} = \delta(\mathbf{A}^s\mathbf{V}\mathbf{W}_G) + \mathbf{V} \in \mathbb{R}^{N_{cls} \times (C_G + C_T)}, \quad (10)$$

where $\mathbf{G}$ denotes the updated semantic relation graph, $\mathbf{W}_G \in \mathbb{R}^{(C_G + C_T) \times (C_G + C_T)}$ is the learnable graph weights. $\delta(\cdot)$ denotes the LeakyReLU [26] activation function.

3.3. Learning Objective

With structural and semantic relation graphs obtained, we further joint their predictions for training in a collaborative learning manner (see Fig. 2). We adopt $\mathcal{L}_{sac}$ to

Table 1. Comparisons with state-of-the-art methods on the MS-COCO dataset. The performance of our approach based on three resolution settings are reported. R_{train} and R_{test} denote resolution used in training and testing stage. * denotes the performance of our implementation.

Methods	(R_{train}, R_{test})	mAP	All						Top 3					
			CP	CR	CF1	OP	OR	OF1	CP	CR	CF1	OP	OR	OF1
CNN-RNN [32]	$(-,-)$	61.2	-	-	-	-	-	-	66.0	55.6	60.4	69.2	66.4	67.8
RNN-Attention [34]	$(-,-)$	-	-	-	-	-	-	-	79.1	58.7	67.4	84.0	63.0	72.0
Order-Free RNN [2]	$(-,-)$	-	-	-	-	-	-	-	71.6	54.8	62.1	74.2	62.2	67.7
SRN [46]	$(224, 224)$	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
PLA [40]	$(288, 288)$	-	80.4	68.9	74.2	81.5	73.3	77.2	-	-	-	-	-	-
ResNet-101 *[17]	$(448, 448)$	78.6	82.4	65.5	73.0	86.0	70.4	77.4	85.9	58.6	69.7	90.5	62.8	74.1
ML-GCN [8]	$(448, 448)$	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
KSSNet [33]	$(448, 448)$	83.7	84.6	**73.2**	77.2	**87.8**	76.2	**81.5**	-	-	-	-	-	-
Ours	$(448, 448)$	**84.6**	86.0	73.1	79.0	86.6	**76.4**	81.2	**89.9**	64.4	75.0	91.2	67.0	77.2
ADD-GCN [41]	$(448, 576)$	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
Ours	$(448, 576)$	**85.8**	87.9	73.6	80.1	87.9	77.3	82.3	91.3	64.8	75.8	92.0	67.6	77.9
SSGRL [5]	$(576, 576)$	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
C-Tran [21]	$(576, 576)$	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	$(576, 576)$	**86.0**	87.0	**74.7**	**80.4**	87.5	77.9	**82.4**	90.7	65.6	**76.2**	91.9	68.0	**78.1**

constraint semantic-aware features in Section 3.2. To accelerate the convergence process, we adopt $\mathcal{L}_{trans}$ and $\mathcal{L}_{gcn}$ on the prediction result of structural and semantic relation graphs respectively. Besides, we adopt $\mathcal{L}_{joint}$ for the final prediction result. All these loss functions are supervised with typical multi-label classification entropy $\mathcal{L}_{joint}$ $= -\sum_{i=1}^{N_{cls}} \mathbf{y}_i \log(\mathbf{p}_i), \mathbf{p}, \mathbf{y} \in \mathbb{R}^{N_{cls}}, \mathbf{y} \in \{0,1\}$. Hence the final learning objective $\mathcal{L}_{sum}$ can be formulated as:

$$\mathcal{L}_{sum} = \mathcal{L}_{joint} + \mathcal{L}_{sac} + \mathcal{L}_{trans} + \mathcal{L}_{gcn}. \quad (11)$$

With this collaborative regularization, the final classification embedding in Eq. (3) can jointly be aware of structural and semantic information for multi-label understanding.

4. Experiments

4.1. Datasets and Evaluation Metrics

MS-COCO Benchmark. Microsoft COCO [23] is a widely-used benchmark for many vision tasks, *e.g.* object detection, segmentation and multi-label recognition. It contains 82,081 images in train set and 40,137 images in validation set from 80 common object categories. On average, each image has 2.9 labels. Especially, it contains a large number of small objects, which is more challenging for multi-label recognition. Following [8, 41, 5], we evaluate the performance of all methods on the validation set.

VOC 2007 Benchmark. PASCAL VOC 2007 [13] is another popular benchmark for multi-label recognition. It contains 5,011 images in train and validation set and 4,952 images in test set from 20 common object categories. On average, each image has 1.4 labels. Following [8], we train our approach on the trainval set and evaluate on the test set.

Evaluation Metrics. To quantitatively evaluate the performance of our approach and state-of-the-art methods, we adopt the average per-class precision (CP), recall (CR), F1 (CF1), the average overall precision (OP), recall (OR), F1 (OF1) and mean average precision (mAP) as evaluation metrics. For fair comparisons, we also report top-3 results. Notably, precision/recall/F1-score may be affected by the threshold, which is set as 0.5 in our setting. Among all metrics, AP and mAP are the most important metrics which can provide a more comprehensive comparison.

4.2. Implementation Details

We adopt ResNet-101 [17] pre-trained on ImageNet [9] as our backbone. In the training stage, input images are first resized into 512×512, then random cropped and resized into 448 × 448 with random horizontal flips for augmentation. In the testing stage, input images are resized into 448 × 448. We use the SGD optimizer with momentum of 0.9 and weight decay of 1e-4. The initial learning rate is 0.01 for VOC 2007 and 0.03 for MS-COCO, which decays by a factor of 10 for every 30 epochs. The batch size is set as 16 for VOC 2007 and 32 for MS-COCO on each GPU. The network converge quickly and only needs 50 epochs in total for training. Detailed experiments on hyper-parameters can be found in supplementary materials. We set hidden dimension $C_T = 512$ and $C_G = 512$. The transformer unit consists of 3 layers and each layer has 4 attention heads. The semantic graph neural network has one layer. As mentioned in Eq. (3), We apply a weight coefficient α on structural relation graph and $(1-\alpha)$ on semantic relation graph. We set $\alpha = 0.7$ to achieve the best performance. All experiments are conducted on two NVIDIA Tesla V100 GPUs.

Table 2. Comparisons with state-of-the-art methods on the VOC 2007 dataset. The performance of our approach based on resolution 448×448 is reported. * denotes the performance of our implementation.

Methods	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
CNN-RNN [32]	96.7	83.1	94.2	92.8	61.2	82.1	89.1	94.2	64.2	83.6	70.0	92.4	91.7	84.2	93.7	59.8	93.2	75.3	**99.7**	78.6	84.0
RNN-Attention [34]	98.6	97.4	96.3	96.2	75.2	92.4	96.5	97.1	76.5	92.0	87.7	96.8	97.5	93.8	98.5	81.6	93.7	82.8	98.6	89.3	91.9
Fev+Lv [38]	97.9	97.0	96.6	94.6	73.6	93.9	96.5	95.5	73.7	90.3	82.8	95.4	97.7	95.9	98.6	77.6	88.7	78.0	98.3	89.0	90.6
Atten-Reinforce [4]	98.6	97.1	97.1	95.5	75.6	92.8	96.8	97.3	78.3	92.2	87.6	96.9	96.5	93.6	98.5	81.6	93.1	83.2	98.5	89.3	92.0
ResNet-101 * [17]	99.8	98.3	98.0	98.0	79.5	93.2	96.8	97.7	79.9	91.0	86.6	98.2	97.8	96.4	98.8	79.4	94.6	82.9	99.1	92.1	92.9
SSGRL [5]	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	98.1	84.9	96.5	79.8	98.4	92.8	93.4
ML-GCN [8]	99.5	98.5	**98.6**	98.1	80.8	94.6	97.2	**98.2**	82.3	95.7	86.4	98.2	98.4	96.7	99.0	84.7	96.7	84.3	98.9	93.7	94.0
ADD-GCN *[41]	99.7	98.5	97.6	98.4	80.6	94.1	96.6	98.1	80.4	94.9	85.7	97.9	97.9	96.4	99.0	80.2	97.3	85.3	98.9	94.1	93.6
Ours	**99.9**	**98.9**	98.4	**98.7**	81.9	**95.8**	**97.8**	98.0	**85.2**	95.6	**89.5**	**98.8**	**98.6**	**97.1**	**99.1**	**86.2**	**97.7**	**87.2**	99.1	**95.3**	**95.0**

Table 3. Ablation study for different components. $\mathcal{R}_{Structural}$ and $\mathcal{R}_{Semantic}$ denote structural and semantic relation graphs. $\mathcal{M}_{Trans}$ denotes transformer units. $\mathcal{M}_{CSA}$ denotes cross-scale attention module. $\mathcal{M}_{GCN}$ denotes graph convolutional network. $\mathcal{R}_{SAC}$ denotes semantic-aware constraints.

$\mathcal{R}_{Structural}$		$\mathcal{R}_{Semantic}$		mAP	
$\mathcal{M}_{Trans}$	$\mathcal{M}_{CSA}$	$\mathcal{M}_{GCN}$	$\mathcal{M}_{SAC}$	COCO	VOC
				78.6	92.9
✓				82.9	94.3
✓	✓			83.9	94.6
		✓		82.5	93.4
		✓	✓	83.5	93.8
✓		✓		84.0	94.6
✓	✓	✓	✓	84.6	95.0

Table 4. Ablation study for cross-scale attention module on MS-COCO. $\mathcal{S}_{\{\frac{1}{32},\frac{1}{64},\frac{1}{16}\}}$ denote features of different scales. $\mathcal{M}_{CA}$ denotes cross attention module. $\mathcal{M}_{Trans}$ denotes transformer units.

$\mathcal{S}_{\frac{1}{32}}$	$\mathcal{S}_{\frac{1}{64}}$	$\mathcal{S}_{\frac{1}{16}}$	$\mathcal{M}_{CA}$	$\mathcal{M}_{Trans}$	mAP $\mathcal{R}_{Structural}$
✓				TR	82.9
✓	✓			TR	83.1 (↑ 0.2)
✓	✓	✓		TR	83.3 (↑ 0.4)
✓	✓	✓	SUM	TR	83.2 (↑ 0.3)
✓	✓	✓	MUL	MLP	83.3 (↑ 0.4)
✓	✓	✓	MUL	TR	83.9 (↑ 1.0)

4.3. Comparison with State-of-the-art

Comparisons on MS-COCO. As shown in Tab. 1, we compare our approach on MS-COCO benchmark with 11 state-of-the-art methods. The most commonly used resolution is 448×448 during both training and testing stage. However, it is worthy to notice that some methods evaluate their performance on different resolutions during training and inference stages, e.g., ADD-GCN [41] and SSGRL [5]. For fair comparisons, we follow their resolution settings [8, 41, 5] and report three results, which achieve a new performance leader-board with a clear margin.

Comparisons on VOC 2007. In Tab. 2, we compare our approach with 8 state-of-the-art methods. For fair comparisons, we report mAP and AP of each class on commonly used 448×448 resolution with only ImageNet pretrained. In terms of mAP, our approach achieves the best performance and outperforms state-of-the-art ML-GCN [8] by 1.0%.

4.4. Performance Analysis

Ablation Studies. To evaluate the effectiveness of our proposed structural relation graph module and semantic relation graph module, we reconstruct our model with different ablation factors in Tab. 3. We first employ ResNet-101 with the identical training protocol as our baseline model in the first row, which reaches a high baseline performance, e.g., 92.9% on VOC 2007. Note that this baseline model

outperforms several state-of-the-art models and our proposed model can steadily improve the performance based on this high baseline. As shown in Tab. 3, our proposed modules make a steady improvement to the final performance, which demonstrates the necessity of the proposed modules to obtain the best classification results.

Effects of Structural Relation. It can be found in Tab. 3 that only adopting Transformer for structural relation can notably improve the performance, e.g., 4.3% on MS-COCO, which demonstrates the effectiveness of long-term contextual information for multi-label recognition task. Besides, the position information of small objects may vanish after down-sampling especially for challenging datasets, e.g., MS-COCO. Our proposed cross-scale attention module could enhance cross-scale features and suppress noises, which further boosts performance by 1.0% on MS-COCO.

To verify the effectiveness of our proposed cross-scale attention module, we explore different scales on MS-COCO as shown in Tab. 4. The default scale is $\frac{1}{32}$ generated by the last stage of our baseline ResNet-101. With more different scales taken into consideration, the performance of structural relation graph module is slightly improved. With our proposed cross attention module with multiplication operation, the performance further boosts to a new level, the performance drops 0.7% when replacing the multiplication with summation, which demonstrates our proposed cross attention module could further effectively strengthen the position information between different scales.

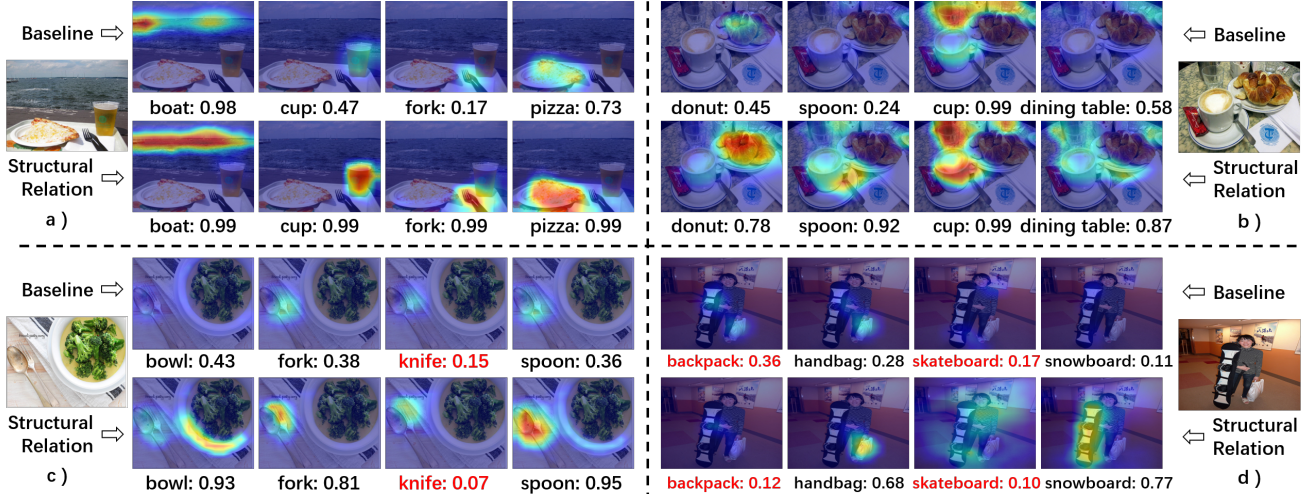To further verify the effectiveness of transformer units on

Figure 5. Visualization analyses of baseline and our proposed structural relation graph module. We present several labels for demonstration and the labels not presented in the image are highlighted in red. Compared to the baseline our structural relation has the ability to handle objects in small scales in a) and b) or with confused appearances in c) and d).

Table 5. Ablation study for proposed joint relation graph on MS-COCO. $\mathcal{G}_{JCM}$ denotes the learnable joint correlation matrix. $\mathcal{G}_{SG}$ denotes structural guidance. $\mathcal{C}_{SAC}$ denotes the semantic-aware constraints.

$\mathcal{G}_{JCM}$	$\mathcal{G}_{SG}$	$\mathcal{C}_{SAC}$	mAP $\mathcal{R}_{Semantic}$
Static			81.9
✓			82.5 (↑ 0.6)
✓	✓		83.0 (↑ 1.1)
✓	✓	GMP	83.5 (↑ 1.6)
✓	✓	GAP	83.6 (↑ 1.7)
✓	✓	$KMP_{5\%}$	83.7 (↑ 1.8)

cross-scale information, we replace transformer units with a simple MLP layer in the 5th row in Tab. 4, the performance of $R_{structural}$ shows a clear drop (0.6%) in mAP, which demonstrates the transformer units could effectively capture the long-term spatial context.

Effects of Semantic Relation. As shown in Tab. 3, building semantic relationships with GCN could notably improve the performance, *e.g.* 3.9% on MS-COCO. Moreover, the performance further boosts for 1% with our proposed semantic-aware constraints, which demonstrates that GCN could achieve better modeling results with more representative semantic-specific vectors. To evaluate the effectiveness of our proposed semantic relation graph module, we conduct detailed ablations on MS-COCO in Tab. 5. We adopt static adjacent matrix used in ML-GCN [8] to form a baseline in the first row. Applying our proposed learnable correlation matrix improves the performance by 0.6%. Besides, joint semantic and structural information could effectively improve the performance by 0.5% with structural guidance. Another main exploration is to find global pool-

ing operations for semantic-aware constraints, our final semantic relation achieves the best performance with top-k max-pooling operation with threshold 5%.

Interpretable Visualizations of Structural Relation. We utilize Grad-CAM [28] to exhibit the visualization results of baseline and the proposed structural relation in Fig. 5. i) Benefiting from the cross-scale attention module, our approach could capture more accurate localization and effectively perceive **small objects**, *e.g. fork* in Fig. 5 a) and *spoon* in Fig. 5 b). ii) As baseline model could not distinguish objects with **similar appearances**, *e.g.* the triplet labels {*fork, knife, spoon*} in Fig. 5 c) and the paired labels {*backpack, handbag*}, {*skateboard, snowboard*} in Fig. 5 d), these issues are well handled by our proposed structural relation module benefiting from the long-term contextual information captured by Transformer-based relation graph.

5. Conclusion

In this paper, we propose a novel Transformer-based Dual Relation Graph (TDRG) framework for multi-label recognition tasks. We make the first attempt to introduce Transformer architecture into multi-label recognition tasks to incorporate long-term contextual information and build position-wise relationships across different scales. Besides, we model dynamic co-occurrences with semantic-aware constraints. With these two complementary relations jointed, our proposed approach achieves new state-of-the-art on two multi-label recognition benchmarks.

Acknowledgments

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 3, 4

[2] Shang-Fu Chen, Yi-Chen Chen, Chih-Kuan Yeh, and Yu-Chiang Wang. Order-free rnn with visual attention for multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 1, 2, 6

[3] Tianshui Chen, Liang Lin, Xiaolu Hui, Riquan Chen, and Hefeng Wu. Knowledge-guided multi-label few-shot learning for general image recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

[4] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin. Recurrent attentional reinforcement learning for multi-label image recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2, 7

[5] Tianshui Chen, Muxin Xu, Xiaolu Hui, Hefeng Wu, and Liang Lin. Learning semantic-specific graph representation for multi-label image recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 522–531, 2019. 2, 5, 6, 7

[6] Zhaomin Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Learning graph convolutional networks for multi-label recognition and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[7] Zhao-Min Chen, Xiu-Shen Wei, Xin Jin, and Yanwen Guo. Multi-label image recognition with joint class-aware map disentangling and label correlation embedding. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 622–627. IEEE, 2019. 2

[8] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5177–5186, 2019. 2, 5, 6, 7, 8

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3, 4

[12] Thibaut Durand, Taylor Mordan, Nicolas Thome, and Matthieu Cord. Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 642–651, 2017. 1

[13] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 2, 6

[14] Bin-Bin Gao and Hong-Yu Zhou. Learning to discover multi-class attentional regions for multi-label image recognition. *IEEE Transactions on Image Processing*, 30:5920–5932, 2021. 2

[15] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1277–1286, 2018. 1

[16] Hao Guo, Kang Zheng, Xiaochuan Fan, Hongkai Yu, and Song Wang. Visual attention consistency under image transforms for multi-label image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 729–739, 2019. 1, 2

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 6, 7

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[19] Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 935–944, 2016. 1

[20] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016. 2, 5

[21] Jack Lanchantin, Tianlu Wang, Vicente Ordonez, and Yanjun Qi. General multi-label image classification with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16478–16488, 2021. 3, 6

[22] Yining Li, Chen Huang, Chen Change Loy, and Xiaoou Tang. Human attribute recognition by deep hierarchical contexts. In *European Conference on Computer Vision*, pages 684–700. Springer, 2016. 1

[23] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 6

[24] Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun. Semantic regularisation for recurrent image annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2872–2880, 2017. 2

[25] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 1

[26] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer, 2013. 5

[27] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016. 3

[28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[29] Jing Shao, Kai Kang, Chen Change Loy, and Xiaogang Wang. Deeply learned attributes for crowded scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4657–4666, 2015. 1

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, May 2015. 1

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017. 2, 3, 4

[32] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. Cnn-rnn: A unified framework for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294, 2016. 1, 2, 6, 7

[33] Ya Wang, Dongliang He, Fu Li, Xiang Long, Zhichao Zhou, Jinwen Ma, and Shilei Wen. Multi-label classification with label graph superimposing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12265–12272, 2020. 2, 6

[34] Zhouxia Wang, Tianshui Chen, Guanbin Li, Ruijia Xu, and Liang Lin. Multi-label image recognition by recurrently discovering attentional regions. In *Proceedings of the IEEE international conference on computer vision*, pages 464–472, 2017. 1, 2, 6, 7

[35] Yunchao Wei, Wei Xia, Min Lin, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. Hcp: A flexible cnn framework for multi-label image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(9):1901–1907, 2015. 2

[36] Jiahao Xu, Hongda Tian, Zhiyong Wang, Yang Wang, Fang Chen, and Wenxiong Kang. Joint input and output space learning for multi-label image classification. *IEEE Transactions on Multimedia*, 2020. 2

[37] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5791–5800, 2020. 3

[38] Hao Yang, Joey Tianyi Zhou, Yu Zhang, Bin-Bin Gao, Jianxin Wu, and Jianfei Cai. Exploit bounding box annotations for multi-label object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–288, 2016. 2, 7

[39] Xitong Yang, Yuncheng Li, and Jiebo Luo. Pinterest board recommendation for twitter users. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 963–966, 2015. 1

[40] Vacit Oguz Yazici, Abel Gonzalez-Garcia, Arnau Ramisa, Bartlomiej Twardowski, and Joost van de Weijer. Orderless recurrent models for multi-label classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13440–13449, 2020. 1, 2, 6

[41] Jin Ye, Junjun He, Xiaojiang Peng, Wenhao Wu, and Yu Qiao. Attention-driven dynamic graph convolutional network for multi-label image recognition. In *European Conference on Computer Vision*, pages 649–665. Springer, 2020. 2, 6, 7

[42] Renchun You, Zhiyao Guo, Lei Cui, Xiang Long, Yingze Bao, and Shilei Wen. Cross-modality attention with semantic graph embedding for multi-label classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12709–12716, 2020. 2

[43] Junjie Zhang, Qi Wu, Chunhua Shen, Jian Zhang, and Jianfeng Lu. Multilabel image classification with regional latent semantic dependencies. *IEEE Transactions on Multimedia*, 20(10):2801–2813, 2018. 2

[44] Yifan Zhao, Jia Li, Yu Zhang, and Yonghong Tian. Multiclass part parsing with joint boundary-semantic awareness. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9177–9186, 2019. 1

[45] Yifan Zhao, Ke Yan, Feiyue Huang, and Jia Li. Graph-based high-order relation discovery for fine-grained recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15079–15088, 2021. 2

[46] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5513–5522, 2017. 1, 2, 6

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3

[48] C Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In *European conference on computer vision*, pages 391–405. Springer, 2014. 2