






Modeling Label Correlations with Latent Context for Multi-Label Recognition

Zhaomin Chen¹, Quan Cui², Ruoxi Deng¹, Jie Hu¹, and Guodao Zhang^{3,*}

¹ Key Laboratory of Intelligent Informatics for Safety & Emergency of Zhejiang Province, Wenzhou University

² Waseda University

³ Institute of Intelligent Media Computing, Hangzhou Dianzi University
{chenzhaomin123, ruoxii.deng}@gmail.com, cui-quan@toki.waseda.jp,
israel1987@126.com, guodaozhang@hdu.edu.cn

Abstract. Label dependencies have been widely studied in multi-label image recognition for improving performances. Previous methods mainly considered label co-occurrences as label correlations. In this paper, we show that label co-occurrences may be insufficient to represent label correlations, and modeling label correlations relies on latent context information. To this end, we propose a latent context embedding information network for multi-label image recognition. Our proposal is straightforward and contains three key modules to correspondingly tackle three questions, *i.e.*, where to locate the latent context information, how to utilize the latent context information, and how to model label correlations with context-aware features. First, the multi-level context feature fusion module fuses the multi-level feature pyramids to obtain sufficient latent context information. Second, the latent context information embedding module aggregates the latent context information into categorical features, and thus the label correlation can be directly established. Moreover, we use the label correlation capturing module to model label correlations with full and partial manners, respectively. Comprehensive experiments validate the correctness of our arguments and the effectiveness of our method. In both generic multi-label classification and partial-label multi-label classification, our proposed method consistently achieves promising results.

Keywords: Multi-label · Label correlation · Latent context information

1 Introduction

Multi-label image recognition is a fundamental and practical problem in computer vision, as real-world images generally contain rich and diverse semantic objects. In the literature, conventional multi-label image classification methods mainly take the label co-occurrence as the label correlation, and then utilize the

* Corresponding authors

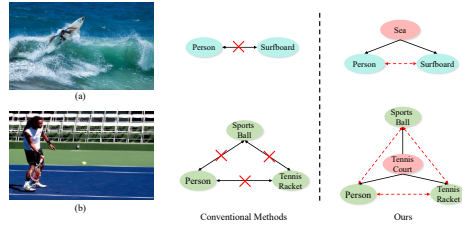


Fig. 1: Illustration of our motivation. A solid arrow indicates a direct correlation between the two labels, while the dashed arrow indicates an indirect correlation. The conventional methods usually consider the label co-occurrence as label correlation. However, we argue that the label correlations depend on latent context information.

graph structures [8, 9, 37] or attention mechanism [27, 48] to capture the label correlation. Recently, some researchers [25, 26] found that the contextual bias may impact the performance, since the modeling of label correlations depends on the co-occurrence context. And then utilizing the causal theory to alleviate the influence of contextual bias.

Although those previous methods achieve the considerable improvements, however, we argue that the label co-occurrence may be insufficient to represent label correlations. And the label correlations can be constructed through reasonable use of the latent context information, even the latent context information may have contextual bias. For instance, we count the co-occurrence frequencies of labels in the MS-COCO [24] dataset, and then select images with the highest label co-occurrence frequencies, as shown in Fig. 1. The labels of “*Person*” and “*Surfboard*” frequently appear together. “*Person*” and “*Sports Ball*” also frequently appear together. Conventional methods consider such labels have a direct correlation, and directly model the correlation between labels. However, multi-label images could be human-centered, and the label “*Person*” could frequently co-occur with other labels. Inferring other labels with “*Person*” cannot lead to accurate results due to the high co-occurrence frequency. Therefore, we conjecture that direct correlations may not always exist between the labels without context, and directly learning label co-occurrence is an inferior solution in such cases. Based on the above analysis, we propose that the correlation between labels depends on the latent context information. Here, the latent context information exists in various formats, *e.g.*, the background, scene, pose and so on. For example, as shown in Fig. 1, “*Person*” and “*Surfboard*” are related because they are both on the “*Sea*” (background). “*Person*”, “*Tennis Racket*”, and “*Sports Ball*” are correlated since they appear near the “*Tennis Court*” (scene).

In this paper, differing from the traditional label correlation assumption, we propose a novel latent context information embedding framework for multi-label image recognition. Our framework contains three modules, *i.e.*, the multi-level context feature fusion module, the latent context information embedding module, and the label correlation capturing module. Firstly, it is difficult to define the specific context information, since we assume that latent context informa-

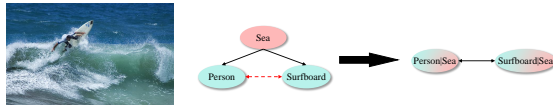


Fig. 2: Illustration of embedding the latent context information to obtain direct label correlations. A solid arrow indicates a direct correlation between the two labels, while the dashed arrow indicates an indirect correlation.

tion appears in various forms (*e.g.*, scene, background, pose, texture, and color). These high-level and low-level latent context information are contained in different levels of features [46], and all kinds of context information may influence the label correlation modeling. Therefore, to obtain sufficient latent context information, we design a multi-level context feature fusion module, which fuses the features from different network levels to extract the context information from low-level to high-level.

Then, we design a latent context information embedding module based on the multi-head cross-attention mechanism [39]. It can embed the latent context information into label features to bring the potential for directly modeling label correlations from an input image. For instance, as shown in Fig. 2, assume that there was no direct correlation between “*Person*” and “*Surfboard*”, but after embedding the context information “*Sea*”, a direct correlation rises between the two labels. Finally, the label correlation capturing module is designed to capture the label directly correlation in full label and partial label manners. Qualitative and quantitative results prove our arguments and evaluate the effectiveness of our method.

2 Related Work

2.1 Multi-Label Image Recognition

A straightforward way for multi-label image recognition is to utilize the independent binary classifiers for each label. Thanks to the great success of CNNs [19, 33, 36] in recent years, the performance of the binarization solutions has been significantly advanced [17, 23, 42]. Although these methods have achieved improvement, they still suffer from optimization difficulties because of the tremendous optimization space. As the number of categories increases, the combinations of labels show exponential growth.

To overcome the above shortcomings, researchers focused on exploiting the label correlation to facilitate the learning process [3, 4, 8, 9, 18, 20, 31, 32, 37, 38, 41, 44, 48]. For instance, Chen *et al.* [8, 9] employed Graph Convolutional Network (GCN) to propagate the information from each category to capture label dependencies. Chen *et al.* [37] proposed a Semantic-Specific Graph Representation Learning (SSGRL) framework utilized the RNN to correlate semantic-specific representations by building the graph model. Zhao *et al.* [47] built a structural relation graph and a semantic relation graph. Both graphs try to obtain context

information from other co-occurrence labels. The above methods rely on graph structures, while some other works directly use class-aware features to model correlations. For example, Chen *et al.* [6] utilized metric learning for pulling correlated label vectors together and pushing uncorrelated label vectors away. Jack *et al.* [21] employed a Transformer encoder to embed the correlations into label embeddings. CCD [26] and IDA [25] focused on utilizing the causal theory to alleviate the influence of contextual bias. For example, CCD [26] only considered object representations to predictions. Thus, the context was defined as the prior context knowledge, and the authors used the causal theory to avoid the influence of this context, and made the image-specific context and object representations to directly affect the predictions. So they built context from the entire dataset, and then removed the prior context knowledge information by backdoor adjustment.

The previous methods assume a direct correlation between labels. However, we argue that there is no guarantee of such direct correlation between labels without latent context information. In this paper, we embed the latent context information into label embeddings containing direct correlations, and then use Transformer to build label correlations. Extensive experiments show that our method achieves competitive results with previous state-of-the-art methods. While previous methods may indirectly leverage the latent context information, we explicitly present this viewpoint and designs corresponding network structures to address it. And we further study the effects of context by proposing a novel multi-label image classification method.

2.2 Transformer in Computer Vision

In recent years, the Transformer [39] model has contributed to great success in Natural Language Processing (NLP) field due to the ability to capture long-range dependencies [1, 12, 28]. Inspired by the great success of the Transformer in the natural language processing field, several works attempted to migrate the Transformer to Computer Vision field. And recent works have verified the effectiveness of the Transformer in various computer vision tasks, *e.g.*, the image recognition [13, 29], object detection [2], and video processing [22, 45]. In this paper, we explore how to utilize the Transformer’s outstanding ability to capture long-distance correlations to embed latent context information and build label correlations.

3 Approach

3.1 Motivation

In the literature [6–8, 37, 48], it has been verified that the critical challenge of the multi-label image recognition task is to capture label correlations. Previous methods were all based on the same assumption that the label co-occurrence can be assumed as the label correlation, and directly utilized various neural networks (*e.g.*, CNN and GCN) to capture the label correlations, which is intuitive

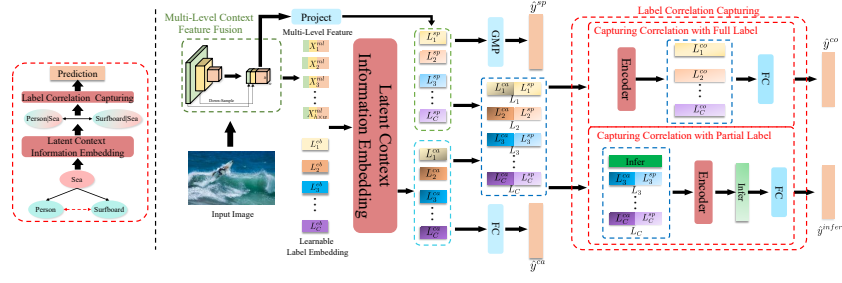


Fig. 3: Overview of the proposed method. **Left:** Our motivation. **Right:** Detailed framework of our method. Specifically, the latent context information embedding module will embed the latent context information into label features to bring the potential for directly modeling label correlations from an input image. And then, label correlation capturing module will model the correlations.

but suboptimal. We argue that there may be an indirect correlation between labels, and that establishing label correlation depends on latent context information. For instance, there is no direct correlation between the labels of “*Person*” and “*Surfboards*” without considering the context information “*Sea*” or “*Beach*”. The labels “*Person*” and “*Surfboards*” often jointly appear. It is attributed to that both labels are correlated to the context information about scenes (*i.e.*, “*Sea*”). Motivated by this, we attempt to embed the latent context information into deep features to obtain the context-aware categorical features. Categorical features (embedded with context information) may bring the potential for directly modeling label correlations, as shown in the Fig. 2.

3.2 Where to Locate the Latent Context Information?

It is difficult to explicitly define the specific context information, since we assume that latent context information appears in various forms (*e.g.*, scene, background, pose, texture, and color). Features of different layers contain different levels of context information [46], *i.e.*, the high-resolution features have low-level appearance information (*e.g.*, texture, and color), and the low-resolution maps have the opposite (*e.g.*, scene, background and pose), which can provide complementary information and help improve the performance. Thus, we fuse these multi-level features to obtain sufficient contextual information.

As shown in Fig. 3, for ResNet [19] we denote the output of three last residual blocks as $\{X_3, X_4, X_5\}$ for “*conv3*”, “*conv4*” and “*conv5*” outputs, and note that they have strides of $\{8, 16, 32\}$ pixels with respect to the input image. We do not include “*conv1*” and “*conv2*” due to the computing resource limitation. Then, we utilize bilinear interpolation to down-sample X_3 and X_4 to the same size as X_5 to reduce computation. After aligning the resolutions of these features, we utilize the concatenation operation to fuse these pyramid features to obtain the multi-level features X^{ml} :

$$X^{ml} = [f_{ds}(X_3) : f_{ds}(X_4) : X_5], \quad (1)$$

where $f_{\text{ds}}(\cdot)$ is the down-sample operation.

3.3 How to Embed the Context Information?

To embed the latent context information into categorical features, we develop our model based on the design principle of Transformer decoder blocks. The standard Transformer decoder contains three modules, *i.e.*, a self-attention module, a cross-attention module, and a feed-forward network (FFN). However, our proposed Latent Context Information Embedding module does not involve a self-attention module, since this module would capture label correlations from labels without direct correlations. Therefore, we use learnable label embeddings as queries and perform cross-attention module and FFN to obtain the context-aware categorical features. Thanks to the cross-attention mechanism of Transformer decoder, it can adaptively embed the latent context information into the context-aware categorical features from the global multi-level features, among which direct label correlations exist.

Formally, let $\mathbf{L}^{\text{eb}} \in \mathbb{R}^{C \times D}$ denote the learnable label embeddings, where C is the number of categories, D means the dimensionality of the label embeddings, which is equal to the channels of the multi-level feature \mathbf{X}^{ml} . Then, the context-aware categorical features can be obtained as follows:

$$\mathbf{L}^{\text{ca}} = f_{\text{lcie}}(\mathbf{X}^{\text{ml}}, \mathbf{L}^{\text{eb}}) \in \mathbb{R}^{C \times D}, \quad (2)$$

where $f_{\text{lcie}}(\cdot)$ is the latent context information embedding module containing a cross-attention module and FFN. Since \mathbf{L}^{ca} is derived from label embedding \mathbf{L}^{eb} through query multi-level feature \mathbf{X}^{ml} , so this feature contains both category information and latent contextual information. Therefore, \mathbf{L}^{ca} can represent the context-aware categorical features. Lastly, we can obtain the prediction confidence of the context-aware categorical features $\hat{\mathbf{y}}^{\text{ca}}$:

$$\hat{\mathbf{y}}^{\text{ca}} = f_{\text{fc}}(\mathbf{L}^{\text{ca}}) \in \mathbb{R}^C, \quad (3)$$

where $f_{\text{fc}}(\cdot)$ denotes a fully-connected layer.

Although the context-aware categorical features directly contain the label correlations due to the latent context information, it has no spatial information. Therefore, we disentangle the class-aware features \mathbf{L}^{sp} from the multi-level features \mathbf{X}_{ml} to enrich the spatial information (see Fig. 3). Inspired by WildCat [15], we first simply apply a 1×1 convolution layer on the multi-level features \mathbf{X}_{ml} to obtain the class-aware features:

$$\mathbf{L}^{\text{sp}} = f_{\text{conv}}(\mathbf{X}_{\text{ml}}) \in \mathbb{R}^{C \times H \times W}, \quad (4)$$

where H and W represent the spatial dimensions (height and width), $f_{\text{conv}}(\cdot)$ denotes the 1×1 convolution layer. Similarly, we can also obtain the prediction confidence of the class-aware features $\hat{\mathbf{y}}^{\text{sp}}$:

$$\hat{\mathbf{y}}^{\text{sp}} = f_{\text{gmp}}(\mathbf{L}^{\text{sp}}) \in \mathbb{R}^C, \quad (5)$$

where $f_{\text{gmp}}(\cdot)$ indicates global max pooling.

Lastly, we leverage the concatenation operation to fuse the context-aware categorical features \mathbf{L}^{ca} and the class-aware features \mathbf{L}^{sp} to obtain the fusion features \mathbf{L} :

$$\mathbf{L} = [\mathbf{L}^{\text{ca}} : f_{\text{flatten}}(\mathbf{L}^{\text{sp}})] \in \mathbb{R}^{C \times (D+HW)}, \quad (6)$$

where $f_{\text{flatten}}(\cdot)$ is the flatten operation, which can flatten the class-aware features \mathbf{L}^{sp} into a single vector $f_{\text{flatten}}(\mathbf{L}^{\text{sp}}) \in \mathbb{R}^{C \times HW}$.

3.4 How to Model Correlations with Context-Aware Features?

Modelling Label Correlations with Full Labels: In the literature, capturing label correlations has effectively improved the performance of multi-label image recognition [8, 20, 37]. Nonetheless, the previous methods directly exploited the class-aware features to model the label correlations, among which direct correlations may not exist. To solve the above problem, we utilize the fusion features, which directly contain label correlations by embedding latent context information, instead of the class-aware features, to build the label correlations.

Without bells and whistles, we employ a standard Transformer encoder to capture the label correlations. Specifically, as illustrated in Fig. 3, the Transformer encoder is first applied to fusion features to establish the label correlations, the operation can be written as

$$\mathbf{L}^{\text{co}} = f_{\text{encoder}}(\mathbf{L}) \in \mathbb{R}^{C \times (D+HW)}, \quad (7)$$

where \mathbf{L}^{co} denotes the label correlation features. After capturing the label correlation, we employ a fully-connected layer to project the label correlation features \mathbf{L}^{co} to the prediction confidence $\hat{\mathbf{y}}_{\text{co}}$:

$$\hat{\mathbf{y}}^{\text{co}} = f_{\text{fc}}(\mathbf{L}^{\text{co}}) \in \mathbb{R}^C. \quad (8)$$

Modelling Label Correlations with Partial Labels: To further capture the label correlations, we utilize partial labels to infer the full label correlations. As shown in Fig. 3, similar to capturing label correlations with full labels, we employ a Transformer encoder to establish the label correlations with partial labels in a popular ‘‘Masked Modeling’’ manner [12]. It is worth noting that we share the parameters of the Transformer encoder with Sec. 3.4.

There are two advantages of sharing the Transformer encoder. Firstly, sharing the Transformer encoder can reduce the parameters of our method. Secondly, we find the shared Transformer encoder benefits from both full and partial labels. In other words, modeling both types of correlations complements each other and further improves performance (demonstrated later by experiments, cf. Table 3).

Specifically, let m denotes mask ratio and $\mathbf{L}^{\text{infer}} \in \mathbb{R}^{1 \times (D+HW)}$ denotes the inference token. First, we randomly mask the fusion features \mathbf{L} based on mask ratio m and ignore these masked features during the training process to obtain

the masked fusion features $\mathbf{L}^{\text{mask}} \in \mathbb{R}^{N \times (D+HW)}$, where N represents the number of remaining unmasked features. Then, we concatenate the inference token $\mathbf{L}^{\text{infer}}$ and masked fusion feature \mathbf{L}^{mask} , and employ the Transformer encoder to model the masked label correlation. The process can be formulated as:

$$\mathbf{L}^{\text{inf}} = f_{\text{select}}(f_{\text{encoder}}([\mathbf{L}^{\text{infer}} : \mathbf{L}^{\text{mask}}])), \quad (9)$$

where \mathbf{L}^{inf} denote the inference token processed by the Transformer encoder. $f_{\text{select}}(\cdot)$ means that selecting \mathbf{L}^{inf} from all inputs.

Lastly, we apply the full-connected layer to project \mathbf{L}^{inf} to the predicted inference confidence $\hat{\mathbf{y}}_{\text{infer}}$:

$$\hat{\mathbf{y}}^{\text{infer}} = f_{\text{fc}}(\mathbf{L}^{\text{inf}}) \in \mathbb{R}^C. \quad (10)$$

Note that the parameters of this fully-connected layer do not share with the co-occurrence correlations method, since \mathbf{L}^{inf} and \mathbf{L}^{co} are in different feature spaces.

3.5 Loss Function

We aggregate the $\hat{\mathbf{y}}^{\text{ca}}$, $\hat{\mathbf{y}}^{\text{sp}}$ and $\hat{\mathbf{y}}^{\text{co}}$ to obtain the final label confidence:

$$\hat{\mathbf{y}} = \hat{\mathbf{y}}^{\text{ca}} + \hat{\mathbf{y}}^{\text{sp}} + \hat{\mathbf{y}}^{\text{co}}, \quad (11)$$

where $\hat{\mathbf{y}}$ denotes the final label confidence. One advantage of linearly aggregating predictions is that, the multi-label image recognition loss (binary cross entropy loss) could be directly applied to the context-aware categorical features and the class-aware feature maps. It has great benefits for embedding latent context information and disentangling categorical features.

The training process of the entire network is end-to-end, we assume the ground truth label of an image is \mathbf{y} , and the multi-label image recognition loss can be written as:

$$\mathcal{L}_{\text{cls}} = - \sum_{c=1}^C y^c \log(\sigma(\hat{y}^c)) + (1 - y^c) \log(1 - \sigma(\hat{y}^c)), \quad (12)$$

where $\sigma(\cdot)$ is the sigmoid function.

We only use $\hat{\mathbf{y}}$ as the final prediction and apply the multi-label loss to $\hat{\mathbf{y}}_{\text{infer}}$ as an auxiliary learning target. One reason is that using $\hat{\mathbf{y}}_{\text{infer}}$ is equivalent to using partial labels for training, which will introduce noise to the predictions and drop the performance. Finally, capturing label correlation with partial labels can advance the fusion features and the Transformer encoder during training, complementing to full labels and further boosting the performance. The overall loss is computed as follows:

$$\mathcal{L} = \alpha \times \mathcal{L}_{\text{cls}} + \beta \times \mathcal{L}_{\text{infer}}, \quad (13)$$

where $\mathcal{L}_{\text{infer}}$ is the loss calculated by $\hat{\mathbf{y}}_{\text{infer}}$. α and β are the hyperparameters set to 1.0 and 0.1, respectively.

4 Experiments

4.1 Generic Multi-Label Image Recognition

Evaluation Metrics. We strictly follow evaluation metrics in previous works [34, 40, 48] to compare with other existing methods. We compute the mean average precision (mAP) overall categories for accuracy evaluation. For more comprehensive comparisons, we compute the average per-class precision (CP), recall (CR), F1 (CF1) and the average overall precision (OP), recall (OR), F1 (OF1). Besides, we also report the results of top-3 labels with the highest scores for further comparing with existing state-of-the-art methods.

Table 1: Comparisons with state-of-the-art methods on the MS-COCO dataset. * denotes that we re-implement the results using the input image with 448×448 resolutions for fair comparisons.

Methods	Resolution	All			Top-3	
		mAP	CF1	OF1	CF1	OF1
ResNet-101* [19]	448×448	79.1	74.1	77.9	70.9	74.7
SSGRL* [37]	448×448	81.9	76.6	78.6	73.5	75.2
DER [6]	448×448	82.8	77.6	80.0	74.1	76.4
ADD-GCN* [20]	448×448	82.8	77.7	80.0	74.3	76.4
ML-GCN [8]	448×448	83.0	78.0	80.3	74.6	76.7
C-Tran* [21]	448×448	83.1	77.7	79.6	74.5	76.1
P-GCN [9]	448×448	83.2	78.3	80.5	74.8	76.7
MCAR [16]	448×448	83.8	78.0	80.3	75.1	76.7
MS-CMA [32]	448×448	83.8	78.4	81.0	74.9	77.1
CCD [26]	448×448	84.0	77.3	81.1	72.9	77.2
SST [7]	448×448	84.2	78.5	80.8	74.8	76.9
TDRG [47]	448×448	84.6	79.0	81.2	75.0	77.2
IDA [25]	448×448	84.8	78.7	80.9	73.6	77.4
Q2L [27]	448×448	84.9	79.3	81.5	73.3	75.4
Ours	448×448	86.8	80.2	82.3	76.1	78.0
SSGRL [37]	576×576	83.8	76.8	79.7	72.7	76.2
C-Tran [21]	576×576	85.1	79.9	81.7	76.0	77.6
CCD [26]	576×576	85.3	80.2	82.1	76.0	77.9
TDRG [47]	576×576	86.0	80.4	82.4	76.2	78.1
SST [7]	576×576	85.8	80.2	82.2	76.0	77.9
IDA [25]	576×576	86.3	80.4	82.5	76.4	78.2
Q2L [27]	576×576	86.5	81.0	82.8	76.5	78.3
Ours	576×576	87.9	81.3	83.0	76.9	78.6

Implementation Details. We use random horizontal flips for data augmentation. Following [8, 37, 48], ResNet-101 [19] is selected as the backbone of our proposed model, which is pre-trained on ImageNet [11] for model parameter

Table 2: Comparisons with state-of-the-art methods on the NUS-WIDE dataset. * means that we re-implement the results in our experiment environment.

Methods	mAP	All		Top-3	
		CF1	OF1	CF1	OF1
CNN-RNN [40]	–	–	–	34.7	55.2
ResNet-101* [19]	59.7	57.5	72.7	53.9	68.4
CMA [32]	60.8	60.4	73.7	55.5	70.0
MS-CMA [32]	61.4	60.5	73.8	55.7	69.5
SRN [48]	62.0	58.5	73.4	48.9	62.2
P-GCN [9]	62.8	60.4	73.4	57.0	69.1
Q2L [27]	65.0	63.1	75.0	–	–
CCD [26]	65.1	61.3	75.0	–	–
Ours	67.1	63.4	75.6	58.6	71.2

Table 3: Impacts of different modules on the MS-COCO dataset. “MLFF” and “LCIE” mean the Multi-Level Feature Fusion module and the latent Context Information Embedding module, respectively. “Spatial” denotes that we utilize the spatial feature. “Full” and “Partial” denote capturing the full label and partial label correlations, respectively. “o” means that “Full” and “Partial” do not share parameters.

MLFF	LCIE	Spatial	Full	Partial	All			Top-3	
					mAP	CF1	OF1	CF1	OF1
					79.1	74.1	77.9	70.9	74.7
✓					80.9	75.3	78.8	71.8	75.5
✓	✓				83.2	77.6	80.4	74.0	76.9
✓	✓		✓		84.0	78.6	81.0	74.7	77.2
✓		✓	✓		82.7	76.5	80.4	73.0	76.9
✓	✓	✓	✓		84.7	79.2	81.4	75.3	77.6
✓	✓	✓	✓	o	85.2	79.9	81.4	75.8	77.5
✓	✓	✓	✓	✓	86.8	80.2	82.3	76.1	78.0

initialization. Traditional stochastic gradient descent (SGD) with a momentum of 0.9 is selected as the model optimizer, and the weight decay is set to 10^{-4} . The initial learning rate is 0.01 decaying by 0.1 every 30 epochs, and we train our model for 100 epochs in total. The batch size of each GPU is 16. We use 4 attention heads and 3 layers Transformer decoder and encoder to learn context-aware categorical features and capture the label correlations, respectively. All probabilities of dropout [35] are set to 0.1. The mask ratio is set to 10%, and we utilize the bilinear interpolation to resize the features.

Datasets. We evaluate our proposed method in two popular benchmark multi-label datasets: MS-COCO 2014 [24] and NUS-WIDE [10].

MS-COCO 2014 Dataset: The MS-COCO 2014 Dataset [24] was originally constructed for object detection and segmentation, but it is also widely used in multi-label image recognition tasks because of the high-quality annotation. This

dataset contains 122,218 images divided into 82,081 images as the training set and 40,504 images as the validation set. It covers 80 common categories with about 2.9 object labels per image.

NUS-WIDE Dataset: The NUS-WIDE Dataset [10], which is a Real-World Web Image Dataset collected from the web *Flickr*, is another popular benchmark dataset for multi-label image recognition. It has 269,648 images and 81 concepts, with an average of 2.4 concept labels per image. This dataset is divided into two parts: a training set of 161,789 images and a test set of 107,859 images. Since image labels are collected based on the associated tags of images, these labels contain a lot of noise information, causing NUS-WIDE more challenging.

Quantitative Results. In this part, we report the quantitative results in MS-COCO 2014 dataset and NUS-WIDE dataset.

Performance on the MS-COCO 2014 Dataset: Quantitative results with different input resolutions on MS-COCO are reported in Table 1. Note that SSGRL, ADD-GCN and C-Tran utilize the input image with resolutions of 576×576 , for a fair comparison, we re-implement these methods using an input image with 448×448 resolutions, and denote them as SSGRL*, ADD-GCN* and C-Tran*. It is obvious to see that our method outperforms previous methods on most metrics. Specifically, our proposed method obtains a +2.2% mAP improvement over the TDRG based on 448×448 resolutions. It is worth mentioning that Q2L also utilizes the Transformer to capture the label correlation, however, our method outperforms it by 1.9% and 1.4% mAP using the same input size respectively. This result further supports our motivation, *i.e.*, we should firstly embed context information into features and secondly model correlations based on context-aware deep features.

Performance on the NUS-WIDE Dataset: The results for NUS-WIDE are presented in Table 2. Similar to the MS-COCO dataset, our proposed method performs better than previous ones on most metrics. Specifically, we obtain 67.1% mAP, 63.4% CF1 and 75.6% OF1, which outperforms another state-of-the-art CCD [26] by 2.0%, 2.1% and 0.6%, respectively. As mentioned in Section 4.1, the annotations of the NUS-WIDE dataset are collected from the tags of images, since these labels contain a lot of noise information. These results show that our method is robust under noisy labels.

Ablation Studies. Unless otherwise specified, the following experiments use ResNet-101 as the backbone and evaluations are based on the MS-COCO dataset.

The Effect of Different Modules: We investigate the impacts of key modules in our framework. Specifically, there are two essential modules: (1) The multi-level feature fusion module (denoted as “MLFF”) utilizes the feature pyramids to enrich the latent context information, in Section 3.2. (2) The latent context information embedding module (denoted as “LCIE”) is employed to obtain the context-aware categorical features, which directly contain label correlations in Section 3.3. Besides, we also integrate (3) the spatial features and the context-

Table 4: Impacts of different backbones. Noting that the backbones noted with 21K are pretrained on the ImageNet-21K dataset.

Backbone	Ours	All			Top-3	
		mAP	CF1	OF1	CF1	OF1
ResNet-101(21K)		84.8	78.8	81.2	74.7	77.4
	✓	87.8	81.5	83.1	77.3	78.8
Swin-L (21K)		89.0	81.6	82.0	78.9	80.1
	✓	90.9	83.7	84.5	79.8	80.8
CvT-w24 (21K)		89.6	81.9	82.4	79.0	80.2
	✓	91.4	84.2	85.0	80.2	80.9

aware categorical features to model (4) full label correlations (denoted as “Full”) and (5) partial label correlations (denoted as “Partial”).

Table 3 shows the performance by progressively integrating the above five modules. Solely applying MLFF on the backbone gives a 1.8% mAP improvement. This result verifies that the low-level information is complementary to the high-level information since it can enrich the latent context information. Then, the context-aware categorical features, which can be directly utilized to predict the final result, bring another 2.3% mAP. This result suggests that the latent context information can make labels have a direct correlation and effectively improve accuracy. Subsequently, leveraging the Transformer encoder to capture the correlations with full labels improves by 0.8%. Due to a lack of spatial information, we attempt to fuse the spatial features into categorical features and obtain 84.7% mAP. It is worth noting that we also directly utilize the spatial features to model the co-occurrence (fifth row of Table 3), and it only achieves 82.7% mAP. This result verifies our motivation, *i.e.*, there may not be had a direct correlation between these labels. Finally, capturing the label correlation with the partial label can achieve 86.8% mAP. Besides, we also find that sharing the parameters of “Full” and “Partial” can further improve the performance, suggesting that both types of correlations are complementary.

The Effect of Large-Scale Backbones and Pretrained Model: To explore the impact of large-scale backbones and pretrained model, we conduct experiments using ResNet-101 (21K), Swin-L (21K) [29] and CvT-w24 (21K) [43] as backbone, respectively. Noting that the backbones noted with 21K are pretrained on the ImageNet-21K dataset. The results are shown in Table 4, which reveal that even with stronger backbone and pretrain models, our proposed method can still improve performance. For example, our method with Swin-L (21K) can achieve 90.9 mAP, which outperforms the baseline by 1.9% mAP. This result shows that our method can be generalized to stronger backbone and pre-trained models.

Visualization and Analyses. In this section, we visualize the learned context-aware categorical features to show if these features contain direct correlations. In Fig. 4, we randomly sample 500 images from the MS-COCO val dataset, and

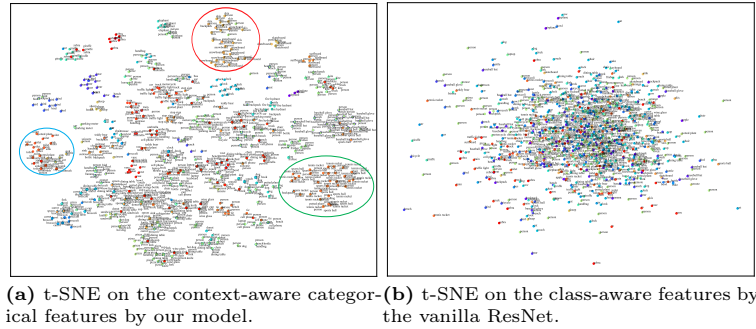


Fig. 4: Visualization of the context-aware categorical features by our model and the class-aware features of the vanilla ResNet on MS-COCO val.

Table 5: The mAP results for the partial-label problem on MS-COCO. $n\%$ denotes the percentage of discarded labels.

Method	10%	20%	30%	40%	50%	60%	70%	80%	90%
BCE [14]	78.5	78.4	78.2	77.7	77.2	76.3	74.1	70.5	61.6
SST [5]	79.9	79.6	79.2	78.9	78.1	77.3	75.9	73.5	68.1
P-GCN [9]	81.6	80.7	79.7	78.6	77.9	76.3	74.4	72.3	68.3
Baseline	76.8	75.5	74.0	72.7	69.3	68.1	64.8	63.5	56.6
Ours	83.5	82.8	81.9	81.2	80.4	78.7	77.3	75.7	72.0
	(+6.7)	(+7.3)	(+7.9)	(+8.5)	(+11.1)	(+10.6)	(+12.5)	(+12.2)	(+15.4)

apply our method to obtain context-aware categorical features. Then, we adopt the t -SNE [30] to visualize the context-aware categorical features associated with the ground-truth label. Besides, we also visualize the class-aware features of vanilla ResNet for comparison. After embedding the latent context information with our method, the context-aware categorical features maintain meaningful semantic topology, *i.e.*, the context-aware categorical features that contain a direct correlation exhibit cluster patterns. As shown in Fig. 4 (a), (1) in the red circle, “Person”, “Skis” and “Snowboard” tend to be close because they are both related to the “Snow” (context). (2) In the green circle, “Person”, “Tennis Racket” and “Sports Ball” tend to be close since they appear near the “Tennis Court” (context). On the contrary, no meaningful topology could be observed from class-aware features learned by vanilla ResNet. The visualizations further verify that labels do not contain a direct correlation, and establishing label correlation depends on latent context information. However, we also observe some negative cases. For instance, “Person”, “Toilet”, “Sink” in the blue circle clustered together, and we conjecture that these labels appear in “Shower Room”. However, “TVs” are rarely observed in shower rooms, and a possible reason is that the contextual bias might be contained in the multi-level features [26].

4.2 Partial-label Multi-label Classification

Experimental Settings. The purpose of partial labeling is to train a model on partially labeled data and then test it on all labeled data. However, current popular multi-label image recognition datasets are fully labeled. Therefore, by following [5, 9, 14], based on the MS-COCO dataset, we construct a partially labeled dataset by randomly discarding some labels on the whole training set. The proportion of discarded labels is between 10% and 90%. If all the labels of an image are completely discarded, we ignore this image. As with the generic multi-label image recognition on the MS-COCO dataset, we evaluate the performance on the validation set with full annotations. We leverage vanilla ResNet-101 [19] as our baseline, and utilize mAP as the evaluation metric. The other settings are the same as those for generic multi-label image recognition in Sec. 4.1.

Quantitative Results. Quantitative results of partial-label multi-label classification are reported in Table 5. As reflected, the mAPs of state-of-the-arts and our method decrease as the proportion of the discarded labels increases. However, under each proportion, our method obviously outperforms the previous methods, which demonstrates the effectiveness of our method in the partial label scenario. Besides, compared with the baseline method, we observe that as the proportion becomes larger, the improvement of our method is more significant. Concretely, our method outperforms the baseline method by +6.7% when the proportion is equal to 10%, while when the proportion is equal to 90%, our approach still can achieve 72.0% mAP, which outperforms the baseline by +15.4%. This observation can further verify the practicality and stability of the proposed method for multi-label image recognition in real-world applications.

5 Conclusions

In this work, we demonstrate that label co-occurrence may be insufficient to represent label correlations, and propose a novel latent context information embedding framework for multi-label image recognition. Our framework mainly contains three modules, *i.e.*, the multi-level context feature fusion module, latent context information embedding module and label correlation capturing module. Firstly, the multi-level context feature fusion module is for obtaining sufficient latent context information. Then, the latent context information embedding module embeds context information into features to obtain the context-aware categorical features. Finally, the label correlation capturing module utilizes context-aware categorical features to establish label correlations. Both quantitative and qualitative results validate the effectiveness of our method. However, we also observe that our method may suffer from contextual bias. Therefore, we will focus on alleviating contextual bias in the future.

Acknowledgements

This research was supported by the National Natural Science Foundation of China under Grant No. 62202337, 62201401, 62201400, the Fundamental Research Funds for the Provincial Universities of Zhejiang GK239909299001-019.

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Nee-lakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020)
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: *ECCV*. pp. 213–229. Springer (2020)
3. Chen, S.F., Chen, Y.C., Yeh, C.K., Wang, Y.C.F.: Order-Free RNN with visual attention for multi-label classification. In: *AAAI*. pp. 6714–6721 (2018)
4. Chen, T., Lin, L., Hui, X., Chen, R., Wu, H.: Knowledge-guided multi-label few-shot learning for general image recognition. *IEEE TPAMI* p. in press (2020)
5. Chen, T., Pu, T., Wu, H., Xie, Y., Lin, L.: Structured semantic transfer for multi-label recognition with partial labels. In: *AAAI*. pp. 339–346 (2022)
6. Chen, Z.M., Cui, Q., Wei, X.S., Jin, X., Guo, Y.: Disentangling, embedding and ranking label cues for multi-label image recognition. *IEEE TMM* **23**, 1827–1840 (2020)
7. Chen, Z.M., Cui, Q., Zhao, B., Song, R., Zhang, X., Yoshie, O.: Sst: Spatial and semantic transformers for multi-label image recognition. *IEEE TIP* **31**, 2570–2583 (2022)
8. Chen, Z.M., Wei, X.S., Wang, P., Guo, Y.: Multi-label image recognition with graph convolutional networks. In: *CVPR*. pp. 5177–5186 (2019)
9. Chen, Z., Wei, X.S., Wang, P., Guo, Y.: Learning graph convolutional networks for multi-label recognition and applications. *IEEE TPAMI* p. in press (2021)
10. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.: NUS-WIDE: a real-world web image database from national university of singapore. In: *Proceedings of Conseil Interprofessionnel des Vins du Roussillon*. pp. 1–9 (2009)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: *CVPR*. pp. 248–255 (2009)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020)
14. Durand, T., Mehrasa, N., Mori, G.: Learning a deep convnet for multi-label classification with partial labels. In: *CVPR*. pp. 647–657 (2019)
15. Durand, T., Mordan, T., Thome, N., Cord, M.: Wildcat: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation. In: *CVPR*. pp. 642–651 (2017)
16. Gao, B.B., Zhou, H.Y.: Learning to discover multi-class attentional regions for multi-label image recognition. *IEEE TIP* **30**, 5920–5932 (2021)
17. Gong, Y., Jia, Y., Leung, T., Toshev, A., Ioffe, S.: Deep convolutional ranking for multi-label image annotation. *arXiv preprint arXiv:1312.4894* pp. 1–9 (2013)
18. Guo, H., Zheng, K., Fan, X., Yu, H., Wang, S.: Visual attention consistency under image transforms for multi-label image classification. In: *CVPR*. pp. 729–739 (2019)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *CVPR*. pp. 770–778 (2016)

20. Jin, Y., Junjun, H., Xiaojiang, P., Wenhao, W., Yu, Q.: Attention-driven dynamic graph convolutional network for multi-label image recognition. In: ECCV. pp. 649–665 (2020)
21. Lanchantin, J., Wang, T., Ordonez, V., Qi, Y.: General multi-label image classification with transformers. In: CVPR. pp. 16478–16488 (2021)
22. Li, L., Gao, X., Deng, J., Tu, Y., Zha, Z.J., Huang, Q.: Long short-term relation transformer with global gating for video captioning. *IEEE TIP* **31**, 2726–2738 (2022)
23. Li, Y., Song, Y., Luo, J.: Improving pairwise ranking for multi-label image classification. In: CVPR. pp. 3617–3625 (2017)
24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV. pp. 740–755 (2014)
25. Liu, R., Huang, J., Thomas, H.L., Li, G.: Causality compensated attention for contextual biased visual recognition. In: ICLR. pp. 1–17 (2023)
26. Liu, R., Liu, H., Li, G., Hou, H., Yu, T., Yang, T.: Contextual debiasing for visual recognition with causal mechanisms. In: CVPR. pp. 12755–12765 (2022)
27. Liu, S., Zhang, L., Yang, X., Su, H., Zhu, J.: Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834* (2021)
28. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* pp. 1–13 (2019)
29. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV. pp. 10012–10022 (2021)
30. Maaten, L., Hinton, G.: Visualizing data using t-SNE. *JMLR* **9**(11), 2579–2605 (2008)
31. Nguyen, H.D., Vu, X.S., Le, D.T.: Modular graph transformer networks for multi-label image classification. In: AAAI. pp. 9092–9100 (2021)
32. Renchun, Y., Zhiyao, G., Lei, C., Xiang, L., Yingze, B., Shilei, W.: Cross-modality attention with semantic graph embedding for multi-label classification. In: AAAI. pp. 1–9 (2020)
33. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. pp. 1–8 (2015)
34. Song, L., Liu, J., Qian, B., Sun, M., Yang, K., Sun, M., Abbas, S.: A deep multi-modal cnn for multi-instance multi-label image classification. *IEEE TIP* **27**(12), 6025–6038 (2018)
35. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. *JMLR* **15**(56), 1929–1958 (2014)
36. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR. pp. 2818–2826 (2016)
37. Tianshui, C., Muxin, X., Xiaolu, H., Hefeng, W., Liang, L.: Learning semantic-specific graph representation for multi-label image recognition. In: ICCV. pp. 522–531 (2019)
38. Vacit, Oguz, Y., Abel, G.G., Arnau, R.: Orderless recurrent models for multi-label classification. In: CVPR. pp. 13440–13449 (2020)
39. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: *NeurIPS*. p. 6000–6010 (2017)
40. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., Xu, W.: CNN-RNN: A unified framework for multi-label image classification. In: CVPR. pp. 2285–2294 (2016)

41. Wang, Z., Chen, T., Li, G., Xu, R., Lin, L.: Multi-label image recognition by recurrently discovering attentional regions. In: ICCV. pp. 464–472 (2017)
42. Wei, Y., Xia, W., Huang, J., Ni, B., Dong, J., Zhao, Y., Yan, S.: CNN: single-label to multi-label. arXiv preprint arXiv:1406.5726 pp. 1–14 (2014)
43. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In: ICCV. pp. 22–31 (2021)
44. Xu, J., Tian, H., Wang, Z., Wang, Y., Kang, W., Chen, F.: Joint input and output space learning for multi-label image classification. IEEE TMM **23**, 1696–1707 (2020)
45. Yang, X., Wang, H., Xie, D., Deng, C., Tao, D.: Object-agnostic transformers for video referring segmentation. IEEE TIP **31**, 2839–2849 (2022)
46. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV. pp. 818–833. Springer (2014)
47. Zhao, J., Yan, K., Zhao, Y., Guo, X., Huang, F., Li, J.: Transformer-based dual relation graph for multi-label image recognition. In: ICCV. pp. 163–172 (2021)
48. Zhu, F., Li, H., Ouyang, W., Yu, N., Wang, X.: Learning spatial regularization with image-level supervisions for multi-label image classification. In: CVPR. pp. 5513–5522 (2017)