Modeling Label Correlations with Latent Context for Multi-Label Recognition (Supplementary Material)

Zhaomin Chen[1], Quan Cui[2], Ruoxi Deng[1], Jie Hu[1], 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

1 Generic Multi-Label Image Recognition

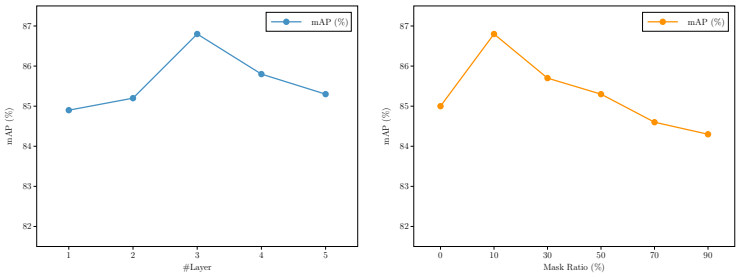
1.1 Ablation Studies

The Effect of Different Depths of Transformer: We show the performance results with different depths of Transformer for our model in Fig. 1 (a). It could be observed that even using a one-layer Transformer, we still achieve competitive results. This result demonstrates the effectiveness and robustness of our proposed method. Furthermore, for transformers with three layers, we can obtain the best result. And when equipped with more layers, the performance drops slightly. We conjecture that more layers will bring more parameters, which can result in over-fitting.

The Effect of Different Mask Ratios: We also analyze the importance of the mask ratio. We change the values of r in a set of $\{0, 0.1, 0.2, ..., 0.9\}$, as depicted in Fig. 1 (b). In experiments, we choose the optimal value of r by cross-validations. We can see that when $r = 10\%$, it can obtain the best performance. If r is too small, such as $r = 0\%$, modelling label correlations with partial label failed. While if r is too large, it will make the inference process more difficult and cause performance degradation.

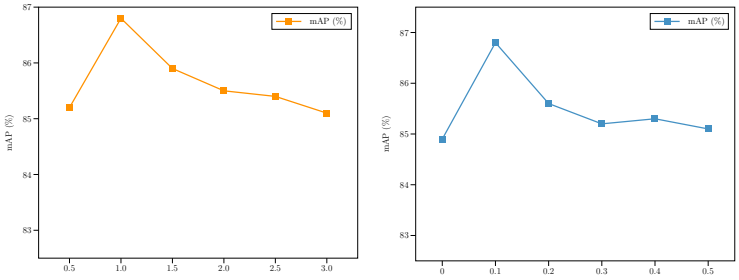
The Effect of Different α and β: To study how the hyperparameters α and β affect the performance accuracy, we conduct the ablation study on the MS-COCO dataset. We change the values of α in a set of $\{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$, as shown in Fig. 2 (a). We observe that when $\alpha = 1.0$, our method can achieve the best performance. Slightly increasing the α value would not influence the performance significantly, which proves the robustness of our method. However,

* Corresponding authors



(a) Different depths of Transformer. **(b)** Different mask ratios.

Fig. 1: Accuracy comparisons with different depths of Transformer and mask ratios on the MS-COCO dataset.



(a) Comparisons of α. **(b)** Comparisons of β.

Fig. 2: Accuracy comparisons with different values of α and β on the MS-COCO dataset.

extremely large or small α will degrade performance. The reason could be that a minimal α value makes the gradient small and training is insufficient, while a large α value leads to unstable training statistics.

Besides, we set the β to $\{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$ in turn, and the mAP scores are shown in Fig. 2 (b). From Fig. 2 (b), we can obtain the optimal result when $\beta = 0.1$, which indicates that the label correlation with full labels and partial labels are complementary, and explains the performance improvement. However, when $\beta > 0.1$, the mAP decreases as β increases.

The Number of Parameters and Computations It's worth noting that our method does not use the branch to build partial label relationships during inference, since the number of parameters and computational complexity do not increase significantly compared to state-of-the-art methods. The results are shown in Table 1. Specifically, the number of parameters and computations of our method is 69.7M and 35.8GFLOPS respectively, which is comparable to state-of-the-art methods such as Q2L (89.5M and 36.6GFLOPS) and TDRG (69.6M and 73.9GFLOPS), yet our method outperforms them in accuracy significantly.

Table 1: The Number of Parameters and Computations.

Method	Param	GFLOPS	mAP
ResNet-101 [1]	44.6M	31.5	79.1
Q2L [2]	89.5M	36.6	84.9
TDRG [3]	69.6M	73.9	84.6
Ours	69.7M	35.8	**86.8**

1.2 Visualization and Analyses

To verify whether the Latent Context Information Embedding module can model latent information, we provide attention visualizations of the vanilla ResNet (*i.e.*, class-aware features) and our method in Fig 3. It is clearly observed that our method activates not only the corresponding label area, but also the contextual area. And the activation of the corresponding label area is stronger than that of the contextual area. For example, in Fig. 3 (a), the vanilla ResNet only activates the area of "Motorcycle" and "Person", while our method not only activates the area of these two labels, but also the contextual area. Additionally, the activation of these two labels is stronger than the context, *e.g.*, the activations of "Motorcycle" and "Background" are weaker than the "Head" and "Body" of "Person" in "Person" label of Fig. 3 (a). This visualization further verifies that our Latent Context Information Embedding module can embed the latent context information into label features.

Besides, we also visualize the class-aware feature, context-aware feature and fusion feature from our method, the results are shown in Fig. 4. We can observe that class-aware features are similar to the vanilla ResNet results (please refer to Fig.4 in the main text), as they do not demonstrate significant topological structure due to a lack of embedded latent contextual information. In contrast, context-aware and fusion features incorporate latent contextual information and therefore exhibit meaningful topological structures. Furthermore, since we construct label relationships through fusion features, the distribution of the visualized results for this feature is more compact compared to context-aware features. This visualization further validates our motivation.

2 The impact of β value for Partial-label Multi-label Classification

In addition, to explore the impact of modeling label correlations with partial label, we conducted ablation study experiments of β value with ratios of 10%, 50% and 90%, the results are shown in Table 2. We found that a beta of 0.1, 0.3 and 0.4 is optimal at 10%, 50% and 90% of ratio respectively. These results reveal that a larger ratio contributes to a better effect of capturing correlation with partial label.

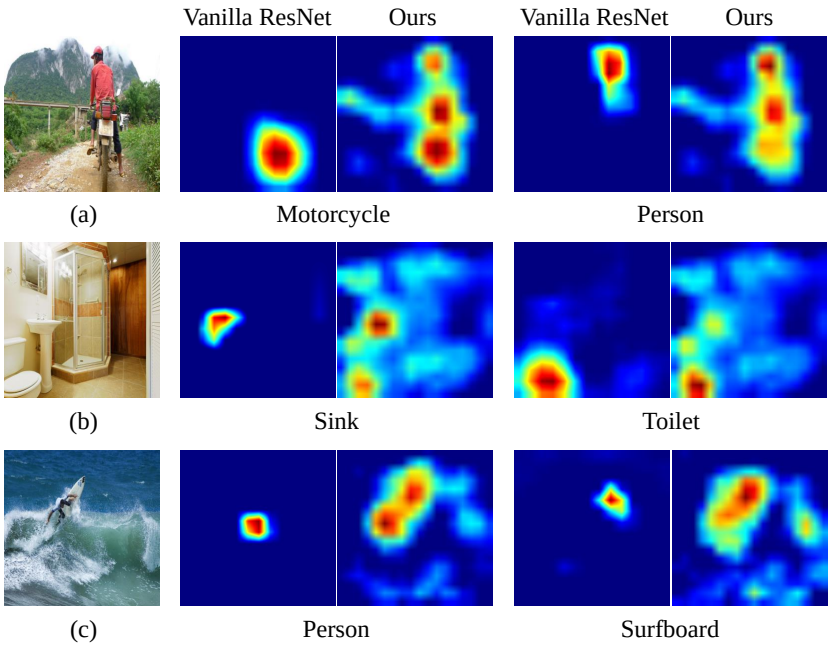


Fig. 3: Visualizations of class-aware features from the vanilla ResNet and the attention features from our method.

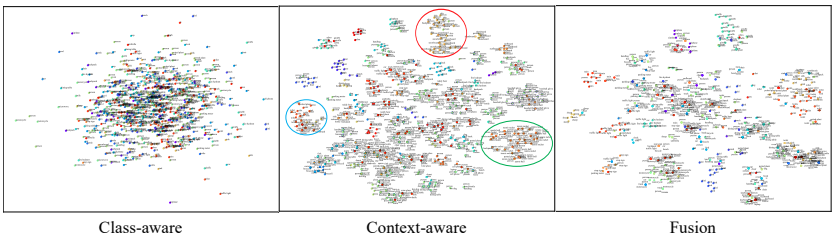


Fig. 4: Visualizations of class-aware features, context-aware feature and fusion feature of our method.

Table 2: Impacts of different β in partial label classification.

Ratio	β	mAP
10%	0.1	83.5
50%	0.1	80.4
50%	0.3	80.9
90%	0.1	72.0
90%	0.4	73.1

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
2. 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)
3. 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)