# Disentangling, Embedding and Ranking Label Cues for Multi-Label Image Recognition

Zhao-Min Chen, Quan Cui, Xiu-Shen Wei, *Member, IEEE*, Xin Jin, and Yanwen Guo, *Member, IEEE*

***Abstract*—Multi-label image recognition is a fundamental but challenging computer vision and multimedia task. Great progress has been achieved by exploiting label correlations among these multiple labels associated with a single image, which is the most crucial issue for multi-label image recognition. In this paper, to explicitly model label correlations, we propose a unified deep learning framework to Disentangle, Embed and Rank (DER) the corresponding label cues. Specifically, we first obtain class-aware disentangled maps (CADMs) by reforming deep activations in accordance with the class-specific recognition weights. Then, after transforming CADMs into the corresponding label vectors, we propose an embedding operation from a metric learning perspective to pull the relevant label vectors together and push irrelevant label vectors away. Furthermore, a ranking operation is employed, which aims to accurately and robustly measure the similarity/dissimilarity of these label vectors. Our model can be trained in an end-to-end manner with only image-level supervision, during which the proposed embedding and ranking operations can contribute to the CADMs learning through back-propagation. In addition, the obtained CADMs are aggregated and further used as an essential feature stream for the final multi-label classification. We conduct extensive experiments on three commonly used multi-label benchmark datasets. Quantitative results show that our model can significantly and consistently outperform previous competitive methods. Moreover, qualitative analysis of our DER proposal also reveals the effectiveness of our proposed model.**

***Index Terms*—Multi-label image recognition, deep learning, label correlation, CNNs, disentangling, embedding, ranking.**

## I. Introduction

RECOGNIZING multiple labels of an image is an important and practical problem in computer vision and multimedia fields, as real-world images always contain rich and diverse semantic information. Considerable efforts for multi-label image classification have been devoted to various research directions, including scene recognition [1], human attribute recognition [2], decision tree optimization [3], image annotation [4], retail product recognition [5], etc. In contrast with general image classification, multi-label classification methods should be capable of modeling label correlations, *i.e.*, identifying and recovering the co-occurrence of multiple labels.

Actually, most recent researches on multi-label image recognition were mainly focused on capturing label correlations from different perspectives. Some works [6] tackled this problem by leveraging bounding box annotations; however, these require additional expensive annotations. Some works [7], [8] demonstrated promising results by implicitly establishing the label correlations with attention mechanisms. On the other hand, some researchers proposed to model the label correlations directly with structure learning models, e.g., graph convolutional networks (GCNs) [9] or recurrent neural networks (RNNs) [10]. However, it is a nontrivial task to define a graph structure that is capable of disentangling category-specific information from a classification network and then appropriately modeling the correlations between them.

In this paper, we propose a unified multi-label image classification framework consisting of three key operations, named DER, *i.e.*, *disentangling*, *embedding* and *ranking*. Our DER operations are capable of modeling the label correlations explicitly, without replying on additional annotations or complicated graph structures. Specifically, we define the label correlations locally at the image level, rather than globally at the dataset level. That is, for one image, we consider the labels of objects of interest that appear in the image as correlated, while all other labels are considered to be uncorrelated.

The architecture of our model is shown in Fig. 1, which consists of three key operations, *i.e.*, *disentangling*, *embedding* and *ranking*. 1) The *disentangling* operation is first employed to generate class-aware disentangled maps (CADMs) for all object categories defined by the dataset for each input image. Particularly, we reform the deep activations of the classification network with category-specific recognition weights. Each CADM contains the semantic information about the corresponding label, as well as the spatial contextual information (cf. Fig. 2), which is helpful to improve the multi-label image recognition performance. 2) The *embedding* operation is designed to explicitly model label correlations from a metric learning perspective, which enriches the aforementioned CADMs with information of label correlations. Specifically, we transform CADMs into label embedding vectors, where the discriminative ability of CADMs is preserved. In the label vector space, we first compute a "positive" dummy cluster centroid among all correlated label vectors and then define the

• Z.-M. Chen is with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China. Q. Cui is with the Graduate School of Information, Production and Systems, Waseda University, Japan. X.-S. Wei is with PCA Lab, Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China. J. Xin is with Megvii Research Nanjing, Megvii Technology, Nanjing, China. Y. Guo is with the National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China and Nanjing Lanzhong Intelligent Technology Co., Ltd.

• Corresponding authors: Xiu-Shen Wei (Nanjing University of Science and Technology) and Yanwen Guo (Nanjing University).

• Emails: {chenzhaomin123, weixs.gm}@gmail.com, cui-quan@toki.waseda.jp, jinxin@megvii.com, ywguo@nju.edu.cn

• This research was supported by the Fundamental Research Funds for the Central Universities 020914380080, the National Natural Science Foundation of China under Grants 61772257 and 61672279, the National Key R&D Program of China (No. 2017YFA0700800) and "111" Program B13022.

distance between a label vector and the dummy cluster centroid as a quantitative measure of label co-occurrence. Specifically, we design a label correlation embedding loss that encourages the relevant/positive label vectors to gather closely around the dummy cluster centroid while encouraging irrelevant/negative label vectors to locate far from the cluster centroid. As a result, if two objects are strongly correlated, the emergence of one object may serve as a useful cue to activate another object region in the corresponding CADM. 3) The *ranking* operation is proposed for more accurately and robustly measuring the similarity/dissimilarity of these label vectors and further refine the CADMs. Distances between irrelevant label vectors and the "positive" dummy cluster centroid are designed to be larger than those between relevant label vectors and the centroid. It is worth mentioning that while the *embedding* and *ranking* operations are not involved in the inference phase, they advance the CADMs of both correlated and uncorrelated object categories during training and thus can boost the performance of multi-label classification. The entire network is end-to-end trainable with these three losses, *i.e.*, conventional multi-label classification loss, the proposed label correlation embedding loss and label ranking loss.

It is worth noting that obtained CADMs are also aggregated and further used as an essential feature stream for the final multi-label classification (cf. Fig. 1). One of our ablation studies shows that these two streams are interdependent, and both general deep feature maps and CADMs possess critical information for good multi-label image recognition results. In addition, the recognition accuracy from the first stream is always higher than that of the baseline method, which also proves that the proposed model can enhance the representation learning of the backbone network.

Our contributions can be summarized as follows:

- We deal with the challenge of multi-label image classification by proposing a unified framework, called DER, for explicitly modeling label correlations. The network can be trained in an end-to-end fashion with only image-level supervisions.
- We propose *disentangling*, *embedding* and *ranking* operations. The disentangling operation is capable of producing CADMs which are enriched with class-specific properties, as well as spatial contextual information. The label correlation embedding and label ranking operations collaborate closely for generating compact but representative label vectors. The *disentangling* operation serves as the basis for the *embedding* and *ranking* operations in order to explicit model label correlations.
- We conduct comprehensive experiments on three widely-used multi-label image classification datasets (MS-COCO, VOC 2007, NUS-WIDE) and achieve consistent performance improvement over the state-of-the-art approaches on all of these datasets. Furthermore, ablation studies and qualitative analysis are performed to verify the effectiveness of our model.

This paper is an extension based on our previous work [11] published in the proceedings of the International Conference on Multimedia and Expo (ICME) 2019 as an oral presentation. In this paper, better recognition accuracy is achieved by a novel operation, (*i.e.*, label ranking in Section III-D), and extensive experiments on the VOC2007 dataset are provided to verify the advantages of our method. In addition, more detailed ablation studies are conducted to verify the effectiveness of each module in our model. The rest of the paper is organized as follows. Section II retrospectively reviews the related works. Section III details the proposed model. Experiments and analysis are provided in Section IV, followed by the conclusion in Section V.

## II. Related work

In this section, we will review two aspects of the relevant works: multi-label image recognition and deep metric learning.

### A. Multi-label image recognition

Owing to the establishment of large-scale labeled datasets (e.g., MS-COCO [12] and ImageNet [13]) and the rapid development of deep CNNs [14], [15], rapid advancements in image classification have been achieved in recent years. In parallel with conventional single-label image classification, many researchers have attempted to adapt the deep CNNs to the multi-label image recognition problem and achieved good recognition performance.

A simple and straightforward method for multi-label recognition is to train one binary deep classifier for each label. However, the major challenge of learning from multi-label data lies in the potentially tremendously-sized output space. Here, the number of possible label sets to be predicted grows exponentially as the number of class labels increases. For example, a label space with a moderate number of 20 class labels will lead to more than 1 million (*i.e.*, $2^{20}$) possible label sets. Thus, many label sets will rarely have examples appearing in the training set, leading to poor performance if they are learned separately.

To overcome the challenges of such an enormous output space, some researchers have used proposal generators to degrade multi-label learning into single label learning. For example, Wei *et al.* [16] proposed the Hypotheses-CNN-Pooling network to aggregate the label scores of each of the specific object hypotheses to achieve the final multi-label predictions. Yang *et al.* [6] treated images as a bag of instances/proposals and solved a multi-instance learning problem. However, these aforementioned methods ignored the label correlation when degrading it into the single-label task. In the following, researchers focused on exploiting the label correlation to facilitate the learning process [17]–[19]. In the literature, Gong *et al.* [20] evaluated various loss functions and found that weighted approximate ranking loss worked best with deep CNNs. Additionally, Hu *et al.* [21] proposed employing a structured inference neural network to model the label correlation of multiple labels. Li *et al.* [22] leveraged probabilistic graphical models to capture the label correlation dependency. Furthermore, Wang *et al.* [10] directly utilized recurrent neural networks (RNNs) to exploit higher-order label relationships. Liu *et al.* [23] proposed the easy-to-hard learning

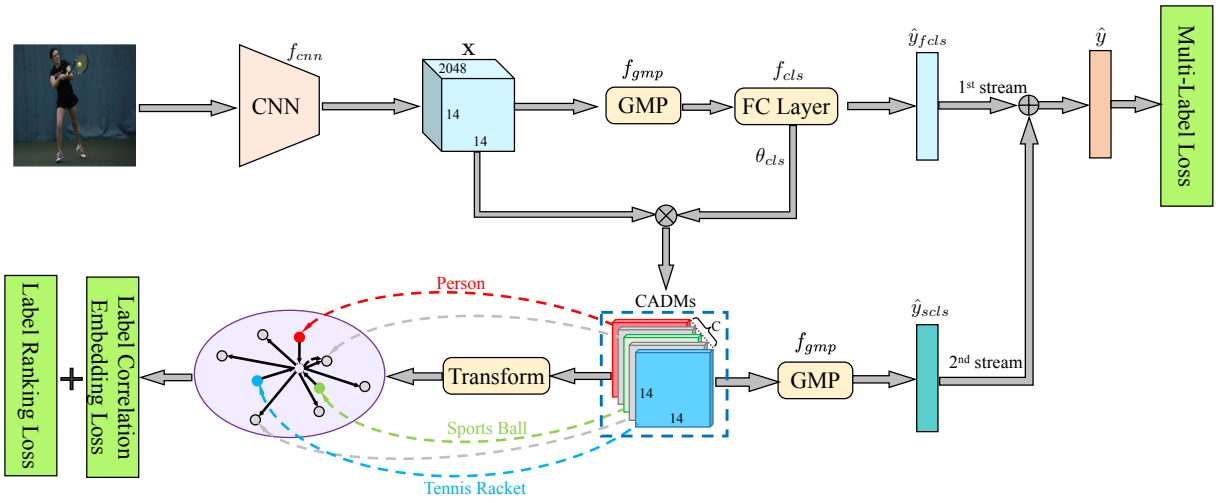



Figure 1. Overall framework of our proposed model for multi-label image recognition during training phase. The input image is first fed to backbone CNNs ($f_{\text{cnn}}$) for the deep activations, (*i.e.*, $\mathbf{X}$) of the last convolutional layer. Next, we utilize global max-pooling (GMP) to obtain the image-level features and then conduct multi-label classification ($f_{\text{fcls}}$) based on these features. In the following, we employ the classification weights ($\theta_{\text{fcls}}$) on $\mathbf{X}$ to generate the CADMs, which could disentangle class-aware specific regions/maps corresponding to multiple image labels. After transforming CADMs into the corresponding label vectors, the embedding operation is performed from a metric learning perspective to pull relevant label vectors ("`person`" "`sports ball`" and "`tennis racket`") together and push irrelevant label vectors away. Furthermore, a ranking operation is employed, which aims at measuring the similarity/dissimilarity of these label vectors accurately and robustly. The entire model is end-to-end trainable and driven by multi-label classification loss, label correlation embedding loss and label ranking loss with only image-level supervisions. (Best viewed in color.)

paradigm for multi-label classification to automatically identify easy and hard labels.

Recently, researchers attempted to model label correlation with region-based multi-label approaches. Some works attempted to apply the attention mechanism to discover the label correlations among different attentional regions, e.g., [7], [8]. In [7], the authors developed the spatial regularization net to focus on the objectiveness regions and further learned label correlations of these regions by self-attention. Meanwhile, Wang *et al.* [8] proposed the spatial transformer to first capture the objectiveness regions and then use LSTMs to handle the label correlation. Furthermore, some works utilized graph structure to model the label correlations, e.g., [9], [24], [25]. Chen *et al.* [9] employed a graph convolution network (GCN) to encode the relationship between categories, where each node is represented by a category-specific word embedding vector and the edge of the graph characterizes the correlations of different categories. Chen *et al.* [25] introduced an RNN as a graph for label correlation modeling, which considers the node as a category-specific feature of each image based on word embedding, and the edge as the relationship between different categories. Lee *et al.* [24] attempted to describe the label relationships by incorporating knowledge graphs. Although the above approaches can achieve good multi-label recognition accuracy, all of them employ word embedding vectors as auxiliary information, assuming that all object categories have corresponding word embedding vectors. However, this assumption might not be satisfied for some unusual categories. Furthermore, the word embedding vectors might impose biased priors for the learning of the graph model, since they have no direct relation with the dataset of interest.

Compared to previous works, our proposed DER model can generate the class-aware disentangled maps (rather than local regions) corresponding to each label of multiple labels. These class-aware disentangled maps contain intact and purely discriminative category-wise information. Moreover, based on these maps, we model the label correlation by formulating it as an effective and efficient label correlation embedding operation, which is a more explicit method for evaluating label co-occurrence. Furthermore, the label ranking operation is also essential, which aims at accurately and robustly measuring the similarity/dissimilarity of these label vectors. The experimental results validate the effectiveness of our proposed model, especially with respect to these three key operations: *Disentangling*, *Embedding* and *Ranking*.

### B. Deep metric learning

In the deep learning era, metric learning aims at learning compact but representative embeddings for samples so that similar/relevant objects are located closely, while dissimilar/irrelevant objects are located far apart. With the advent of CNNs, broad applications have benefited from deep metric learning approaches, including face recognition [26]–[28], image retrieval [29], [30], person re-identification [31]–[33], vehicle re-identification [34], object tracking [35] and many other applications [36], [37].

Contrastive loss [38] was initially proposed for dimensionality reduction and became the footstone in the metric learning field. This loss was optimized by minimizing the distance of positive pairs while keeping negative pair distances larger than a specific margin. Subsequently, better performance was achieved by triplet loss [39] for tackling the problem that positive pairs are always independent from negative pairs

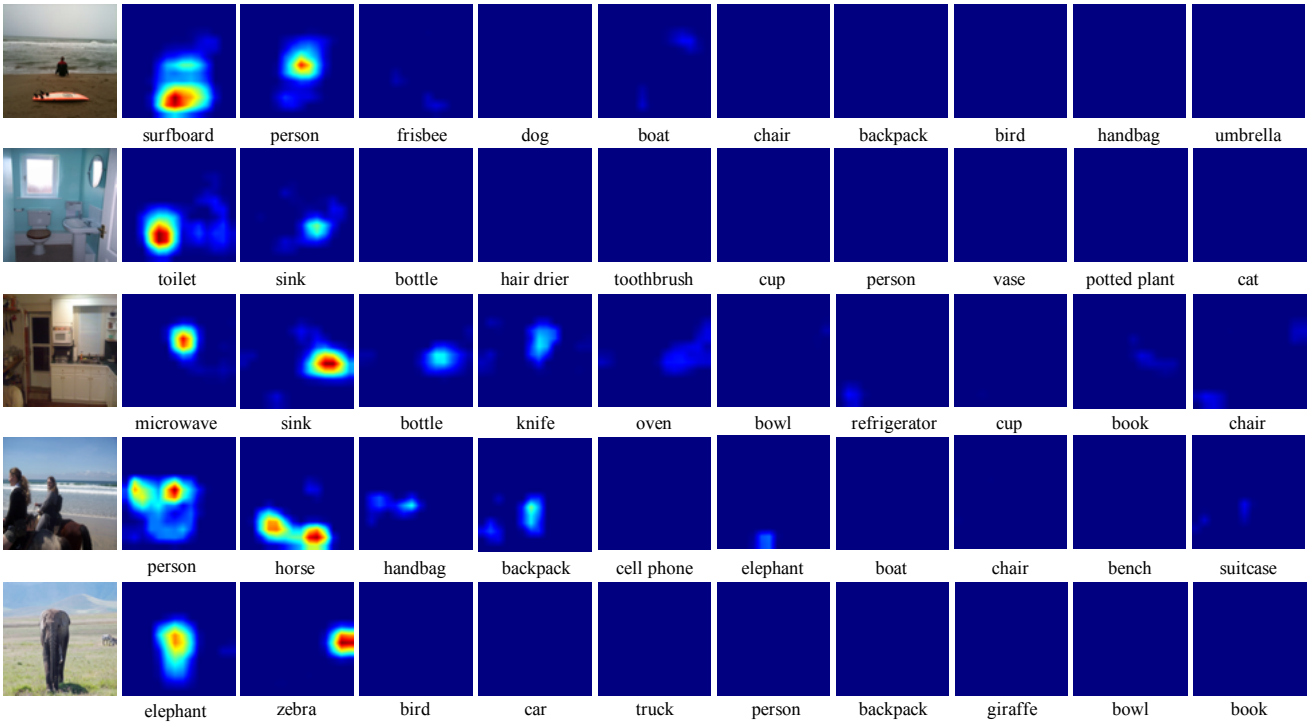



Figure 2. Example images from the *MS-COCO* dataset [12] with the corresponding class-aware disentangled maps (CADMs). For each CADM, we first use linear interpolation to resize the CADM to $1 \times 448 \times 448$. Then, we utilize the visualization tools to convert each CADM to color map. For each image, we first sort the summation activation values of every CADM in descending order and then present the class-aware maps in the same order. It is clear that positive labels correspond to strong activations in their own CADMs, while negative labels activate almost nothing by comparison. (Best viewed in color.)

in the distance calculation process. Specifically, the positive pair distance adds a margin that is forced to be smaller than the negative pair distance. However, triplet loss-based methods often suffer from low convergence and imperfect performance since directly sampling triplets cannot guarantee their effectiveness. To address this issue, semi-hard negative mining [40], [41] was proposed to make the training process more efficient and to improve the performance. In addition, several losses [42]–[44] were proposed in cooperation with the classification loss (cross-entropy loss, binary cross-entropy loss, etc.). Wen *et al.* [42] kept a dummy centroid for every category and proposed center loss to pull data points from the same class close to the corresponding dummy centroid. Liu *et al.* [43] proposed constraining the embeddings learned by the classification loss on a hypersphere. Deng *et al.* [44] improved this kind of method by introducing a margin for generating more highly discriminative features.

In recent years, in contrast with pair-based and triplet-based methods, list-based methods have been widely investigated. To form a more rational list, various ranking losses [45]–[47] were proposed for optimizing the ranking in retrieval result lists. He *et al.* [46] proposed a trainable AP loss calculated by reranking the query results into ideal results. Chen *et al.* [45] proposed a unified deep ranking framework for person Re-ID that directly predicts the similarity of a pair of pedestrian images via joint representation learning. Wang *et al.* [47] incorporated all nontrivial data points and exploited the structure among them by proposing the ranked list loss. Additionally, BIER [48] attempted to increase the robustness of embeddings by dividing the last embedding layer of a deep network into an embedding ensemble. Then, the training of the ensemble was formulated as an online gradient boosting problem. In [49], a large margin metric learning paradigm was proposed. Both the input and output were projected into the same embedding space, and a distance metric was applied on these embeddings to discover output dependency such that instances with similar multiple outputs could gather closely in the embedding space, while those with different outputs could be moved far away.

## III. Proposed Method

We propose a unified framework by disentangling class-aware maps, embedding label correlation information and ranking label vectors to accomplish multi-label image recognition. The class-aware disentangled maps (CADMs) are proposed to assist the multi-label image recognition. With the embedding operation, CADMs are enriched with label correlations. Subsequently, the ranking operation is proposed to model label correlations more accurately and further refine the CADMs. The entire framework consisting of the above three key operations is illustrated in Fig. 1. In this section, we first introduce the notations, then detail these three key operations, and finally provide an overview of the network and the training scheme.

### A. Notations

The following notations are used in the rest of this paper. Let $\mathbf{I}$ denote an input image with ground-truth labels

$\boldsymbol{y} = \left[y^1, y^2, \ldots, y^C\right]^\top$, where $y^c$ is a binary indicator. $y^c = 1$ indicates that image $\mathbf{I}$ is tagged with label $c$, and $y^c = 0$ otherwise. $C$ is the number of all possible labels in the dataset. For multi-label image recognition, the goal is to predict the multi-label vector $\hat{\boldsymbol{y}}$ for a test input $\hat{\mathbf{I}}$. $\mathbf{X}$ represents the activations of the last convolutional layer with shape $d \times h \times w$. For instance, given the widely-used ResNet101 network with $448 \times 448$ input size, the activations' shape of the "`conv5_x`" layer is $2048 \times 14 \times 14$. $f_{\text{cnn}}$ denotes the backbone convolutional neural network with parameters $S = \{j \mid y^j = 1\}$ denotes the correlated label set of the input image $\mathbf{I}$.

### B. Class-aware map disentangling

In this section, we introduce our *disentangling* operation, which is designed to disentangle the class-aware maps from the deep representation. The disentangled maps serve as the basis for the following *embedding* and *ranking* operations.

Based on $\mathbf{X}$, we globally max-pool the image representations into an image-level feature and then conduct one fully connected layer $f_{\text{fcls}}$ with parameters $\theta_{\text{fcls}} \in \mathbb{R}^{d \times C}$ for classification. Inspired by [50], we can utilize $\theta_{\text{fcls}}$ to disentangle $C$ class-aware maps from these distributed representations of $\mathbf{X}$ [51], [52]. However, in contrast with the global average-pooling used in [50], here we employ global max-pooling to maintain the highlighted activations of small-scale objects that frequently emerge in multi-label images.

Concretely, $\theta^c_{\text{fcls}}$ denotes the classification weights w.r.t. the $c$-th label. From another perspective, $\theta^c_{\text{fcls}}$ can be treated as the filter to filter out class-specific discriminative information for the $c$-th label from $\mathbf{X}$. We omit the bias term here because it exerts little impact on the classification performance.

We denote $\boldsymbol{A}_c$ as the corresponding class-aware disentangled map for class $c$, which can be obtained by

$$\boldsymbol{A}_c = {\theta^c_{\text{fcls}}}^\top \cdot \mathbf{X} \in \mathbb{R}^{h \times w} . \tag{1}$$

Each $\boldsymbol{A}_c$ is disentangled for its corresponding $c$-th label. Thus, by collecting all $C$ disentangled maps, we obtain

$$\mathbf{A} = {\theta_{\text{fcls}}}^\top \cdot \mathbf{X} \in \mathbb{R}^{C \times h \times w} . \tag{2}$$

In fact, the class-aware disentangled maps (CADMs) $\mathbf{A}$ are simply weighted linear sums of the presence of these visual patterns at different spatial locations, and $\mathbf{A}$ of all CADMs shares the same parameters $\theta_{\text{fcls}}$. In Fig. 2, several qualitative visualization results of CADMs for multi-label images are provided. As shown in that figure, each CADM corresponds to one specific and independent label meaning. Moreover, it is apparent that the positive label has stronger activations in its class-aware map, and the negative labels have much weaker, or even no activations. These observations verify that the class-aware map disentangling approach can both decouple label semantic information and localize class-specific regions at the same time.

### C. Label correlation embedding

After disentangling, the obtained CADMs absorb category-specific semantic information. Then, we propose to model the label correlations explicitly via an *embedding* operation in a metric learning fashion to enrich CADMs with the information of label correlations. We design a label correlation embedding loss, which is motivated by the fact that co-occurring labels in an image probably share a compact embedding space, while irrelevant label vectors ought to be located far apart.

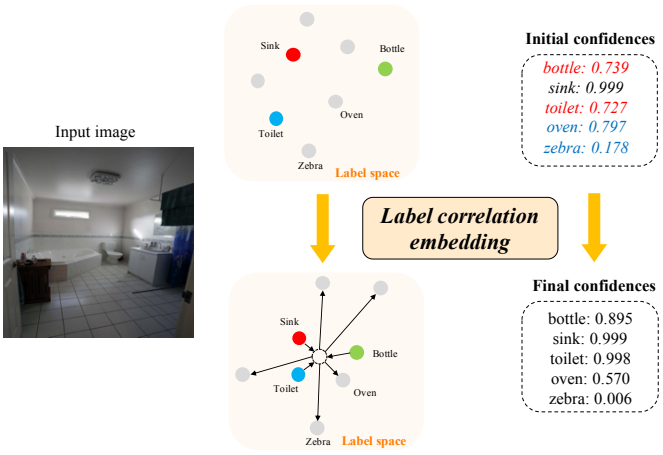



Figure 3. Illustration of our proposed label correlation embedding for improving multi-label image recognition performance. We explicitly model the label correlation in a metric learning paradigm, which can benefit multi-label recognition, although only with image-level supervisions. (Best viewed in color and zoomed in.)

In our model, we embed the class-aware region maps associated with an image $\mathbf{I}$ into a multidimensional *label space*, where each label corresponds to its fixed size label vector $\boldsymbol{a}_c$. Therefore, the co-occurrence of two related labels (*i.e.*, label vectors) can be measured by their distance in this label space. More intuitively, in the multi-label scenario, these correlated labels (*i.e.*, label vectors) could be clustered, while the uncorrelated labels should be apart from the dummy cluster, cf. Fig. 3.

Specifically, to obtain the label vectors, we first flatten the class-aware disentangled map $\boldsymbol{A}_c$ into a single vector $f_{\text{flat}}(\boldsymbol{A}_c) \in \mathbb{R}^{1 \times (h \times w)}$. Then, we introduce a nonlinear transformation $f_{\text{embed}}(\cdot; \theta_{\text{embed}})$ on $f_{\text{flat}}(\boldsymbol{A}_c)$ for embedding it into $\boldsymbol{a}_c$ in the label vectors space:

$$\boldsymbol{a}_c = f_{\text{embed}}(f_{\text{flat}}(\boldsymbol{A}_c); \theta_{\text{embed}}) , \tag{3}$$

where $\theta_{\text{embed}}$ represents the embedding parameters.

Thus, the objective of label correlation embedding becomes minimization of the summation of the pairwise Euclidean distances of correlated label vectors:

$$\min_{\theta_{\text{embed}}} \sum_{j \in S} \sum_{(k<j, k \in S)} \|\boldsymbol{a}_j - \boldsymbol{a}_k\|_2^2 . \tag{4}$$

where the correlated label set $S = \{j \mid y^j = 1\}$. However, for a large-scale number of labels, Eq. (4) exhibits computational redundancy. By some transformations of the term in Eq. (4),

we have

$$
\begin{aligned}
&\sum_{j\in S}\sum_{(k<j,k\in S)}\|\boldsymbol{a}_j-\boldsymbol{a}_k\|_2^2\\
&=\sum_{j\in S}\left(|S|-1\right)\boldsymbol{a}_j^2-2\sum_{j\in S}\sum_{j\neq k}\boldsymbol{a}_j\boldsymbol{a}_k\\
&=|S|\left(\sum_{j\in S}\boldsymbol{a}_j^2-\frac{1}{|S|}\sum_{j\in S}\boldsymbol{a}_j^2-\frac{2}{|S|}\sum_{j\in S}\sum_{k\neq j}\boldsymbol{a}_j\boldsymbol{a}_k\right)\\
&=|S|\left(\sum_{j\in S}\boldsymbol{a}_j^2-\frac{1}{|S|}\left(\sum_{j\in S}\boldsymbol{a}_j\right)^2\right)\\
&=|S|\left(\sum_{j\in S}\boldsymbol{a}_j^2+|S|\left(\frac{\sum_{k\in S}\boldsymbol{a}_k}{|S|}\right)^2-\frac{2}{|S|}\left(\sum_{j\in S}\boldsymbol{a}_j\right)^2\right)\\
&=|S|\sum_{j\in S}\left(\boldsymbol{a}_j^2+\left(\frac{\sum_{k\in S}\boldsymbol{a}_k}{|S|}\right)^2-\frac{2}{|S|}\boldsymbol{a}_j\sum_{k\in S}\boldsymbol{a}_k\right)\\
&=|S|\sum_{j\in S}\left(\boldsymbol{a}_j-\frac{\sum_{k\in S}\boldsymbol{a}_k}{|S|}\right)^2\\
&=|S|\sum_{j\in S}\|\boldsymbol{a}_j-\bar{\boldsymbol{a}}\|_2^2\\
&\propto\sum_{j\in S}\|\boldsymbol{a}_j-\bar{\boldsymbol{a}}\|_2^2\,.
\end{aligned}
\tag{5}
$$

Thus, the optimization problem in Eq. (4) can be written as

$$
\min_{\theta_{\text{embed}}}\sum_{j\in S}\|\boldsymbol{a}_j-\bar{\boldsymbol{a}}\|_2^2\,,
\tag{6}
$$

where $\bar{\boldsymbol{a}}=\frac{1}{|S|}\sum_{j\in S}\boldsymbol{a}_j$ is the mean label vector of all of the correlated labels. Compared with Eq. (4), Eq. (6) is computationally efficient and could contribute to rapid model convergence. Furthermore, considering that the uncorrelated label vectors should be apart from the label mean, the final label correlation embedding loss function becomes

$$
\mathcal{L}_{\text{lce}}=\sum_{j\in S}\|\boldsymbol{a}_j-\bar{\boldsymbol{a}}\|_2^2+\sum_{k\in\overline{S}}\left[1-\|\boldsymbol{a}_k-\bar{\boldsymbol{a}}\|_2^2\right]_+\,,
\tag{7}
$$

where the $[\cdot]_+$ operation indicates the hinge function $\max(0,\cdot)$, and the uncorrelated label set $\overline{S}=\{k\mid y^k=0\}$. By introducing the second term $\left[1-\sum_{k\in\overline{S}}\|\boldsymbol{a}_k-\bar{\boldsymbol{a}}\|_2^2\right]_+$ into Eq. (7), the relationships of correlated labels and uncorrelated labels can be considered at the same time, which can better capture the label co-occurrence from these two different perspectives. Furthermore, this approach can prevent obtaining the trivial solution [53], *i.e.*, $\boldsymbol{a}_c=f_{\text{embed}}(f_{\text{flat}}(\boldsymbol{A}_c);\theta_{\text{embed}})=\mathbf{0}$.

### D. Label ranking

One limitation of label correlation embedding loss, however, is that it cannot guarantee a reasonable ranking of the distances between relevant/irrelevant labels and the dummy cluster. Intuitively, the distances between irrelevant labels and the dummy cluster should be larger than those of relevant labels, cf. Fig. 4. We tackle this problem by proposing a ranking loss that encourages shorter distances for correlated label vectors than for uncorrelated label vectors, such as:

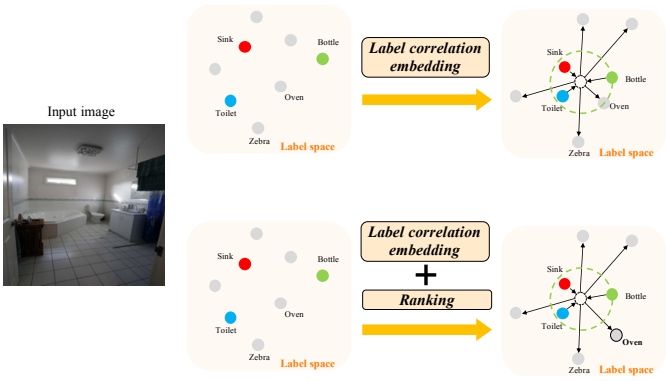



Figure 4. Illustration of label ranking for accurately and robustly measuring the similarity/dissimilarity of these label vectors. The distances between irrelevant labels and the dummy cluster are driven to be larger than the distances among relevant labels and the dummy cluster; *i.e.*, the distance corresponding to category "Oven" becomes greater than that of "Bottle". (Best viewed in color and zoomed in.)

$$
\forall_{(j\in S,k\in\overline{S})}\,\|\boldsymbol{a}_j-\bar{\boldsymbol{a}}\|_2^2<\|\boldsymbol{a}_k-\bar{\boldsymbol{a}}\|_2^2\,,
\tag{8}
$$

where $\|\boldsymbol{a}_j-\bar{\boldsymbol{a}}\|_2^2$ is the distance from correlated label vector $(\boldsymbol{a}_j)$ to the mean label vector $(\bar{\boldsymbol{a}})$, while $\|\boldsymbol{a}_k-\bar{\boldsymbol{a}}\|_2^2$ is the distance from the uncorrelated label vector to the mean label vector.

However, this function is not computationally friendly because it calculates distances of all label vectors and centroid pairs. Hence, we propose the following label ranking function:

$$
\max_{j\in S}\|\boldsymbol{a}_j-\bar{\boldsymbol{a}}\|_2^2<\min_{k\in\overline{S}}\|\boldsymbol{a}_k-\bar{\boldsymbol{a}}\|_2^2\,.
\tag{9}
$$

We enforce ranking only on the maximum distance of correlated label vector and the minimum distance of uncorrelated label vector. Our label ranking loss in the hinge-loss formulation is therefore given as follows:

$$
\mathcal{L}_{\text{lr}}=\left[1+(\max_{j\in S}\|\boldsymbol{a}_j-\bar{\boldsymbol{a}}\|_2^2-\min_{k\in\overline{S}}\|\boldsymbol{a}_k-\bar{\boldsymbol{a}}\|_2^2)\right]_+\,.
\tag{10}
$$

This loss can make the distribution label vectors more compact.

### E. Overall network and training scheme

As shown in Fig. 1, for an input multi-label image $\mathbf{I}$, conventional convolutional neural networks are employed to learn a holistic image representation, which can be formulated as:

$$
\mathbf{X}=f_{\text{cnn}}(\mathbf{I};\theta_{\text{cnn}})\in\mathbb{R}^{d\times h\times w}\,,
\tag{11}
$$

where $\mathbf{X}$ includes a set of 2-D feature maps. These feature maps are embedded with rich spatial information and are also known to obtain mid- and high-level information [54].

Note that we aggregate predicted label confidences from two streams for the final prediction.

**Algorithm 1** Training scheme of our proposed method.

**Require:** Input multi-label image dataset $\mathcal{D} = \{(\mathbf{I}, \boldsymbol{y})\}$.
**Ensure:** Prediction $\hat{\boldsymbol{y}}$.

**for** $i = 1$ to Epoches **do**
  **for** each image $\mathbf{I}$ and its ground truth $\boldsymbol{y}$ in $\mathcal{D}$ **do**
    $\mathbf{X} = f_{\text{cnn}}(\mathbf{I}; \theta_{\text{cnn}})$
    $\hat{\boldsymbol{y}}_{\text{fcls}} = f_{\text{fcls}}(f_{gmp}(\mathbf{X}); \theta_{\text{fcls}})$
    $\mathbf{A} = {\theta_{\text{fcls}}}^{\top} \cdot \mathbf{X}$
    $\hat{\boldsymbol{y}}_{\text{scls}} = f_{\text{gmp}}(\mathbf{A})$
    $\hat{\boldsymbol{y}} = 0.5 \cdot (\hat{\boldsymbol{y}}_{\text{fcls}} + \hat{\boldsymbol{y}}_{\text{scls}})$
    $\boldsymbol{a} = f_{\text{embed}}(f_{\text{flat}}(\boldsymbol{A}); \theta_{\text{embed}})$
    $\bar{\boldsymbol{a}} = \frac{1}{|S|} \sum_{j \in S} \boldsymbol{a}_j$ (where $S = \{j | y^j = 1\}$ is the correlated label set)
    Compute loss $\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha \cdot \mathcal{L}_{\text{lce}} + \beta \cdot \mathcal{L}_{\text{lr}}$
    Backward $\mathcal{L}$ and update parameters.
  **end for**
**end for**

In the first stream, we employ global max-pooling on $\mathbf{X}$ to obtain the image-level features, followed by binary classification for each of the $C$ labels:

$$\hat{\boldsymbol{y}}_{\text{fcls}} = f_{\text{fcls}}(f_{gmp}(\mathbf{X}); \theta_{\text{fcls}}) \in \mathbb{R}^C , \tag{12}$$

where $\hat{\boldsymbol{y}}_{\text{fcls}} = \left[\hat{y}^1_{\text{fcls}}, \hat{y}^2_{\text{fcls}}, \ldots, \hat{y}^C_{\text{fcls}}\right]^{\top}$, and each element of $\hat{\boldsymbol{y}}_{\text{fcls}}$ is a confidence score.

For the second stream, after reforming $\mathbf{X}$ into CADMs $\mathbf{A}$ with classification weights $\theta_{\text{fcls}}$, we obtain additional label confidences $\hat{\boldsymbol{y}}_{\text{scls}}$ by directly applying depthwise global max-pooling on $\mathbf{A}$:

$$\hat{\boldsymbol{y}}_{\text{scls}} = f_{\text{gmp}}(\mathbf{A}) \in \mathbb{R}^C . \tag{13}$$

CADMs contain not only the local-level spatial contextual information (*i.e.*, activations) but also the global-level class-aware semantic meaning. To combine both holistic and class-specific information, we aggregate these two label confidences as the final label prediction confidences by

$$\hat{\boldsymbol{y}} = \frac{1}{2}(\hat{\boldsymbol{y}}_{\text{fcls}} + \hat{\boldsymbol{y}}_{\text{scls}}) \in \mathbb{R}^C . \tag{14}$$

For training, $\hat{\boldsymbol{y}}$ will be used to measure the prediction errors w.r.t. the ground-truth labels $\boldsymbol{y}$ as

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{C} y^c \log(\sigma(\hat{y}^c)) + (1 - y^c) \log(1 - \sigma(\hat{y}^c)), \tag{15}$$

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

Beyond $\mathcal{L}_{\text{cls}}$, our model is also driven by two other loss functions, *i.e.*, $\mathcal{L}_{\text{lce}}$ and $\mathcal{L}_{\text{lr}}$, which are elaborated in previous subsections. The total loss function is presented as follows:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \alpha \cdot \mathcal{L}_{\text{lce}} + \beta \cdot \mathcal{L}_{\text{lr}} . \tag{16}$$

Here, $\alpha$ and $\beta$ are trade-off parameters that are set to $0.5$ and $0.05$, respectively, in all experiments. In Algorithm 1, we present the training scheme our DER method for clarity, where the batch size is set to 1 for simplicity.

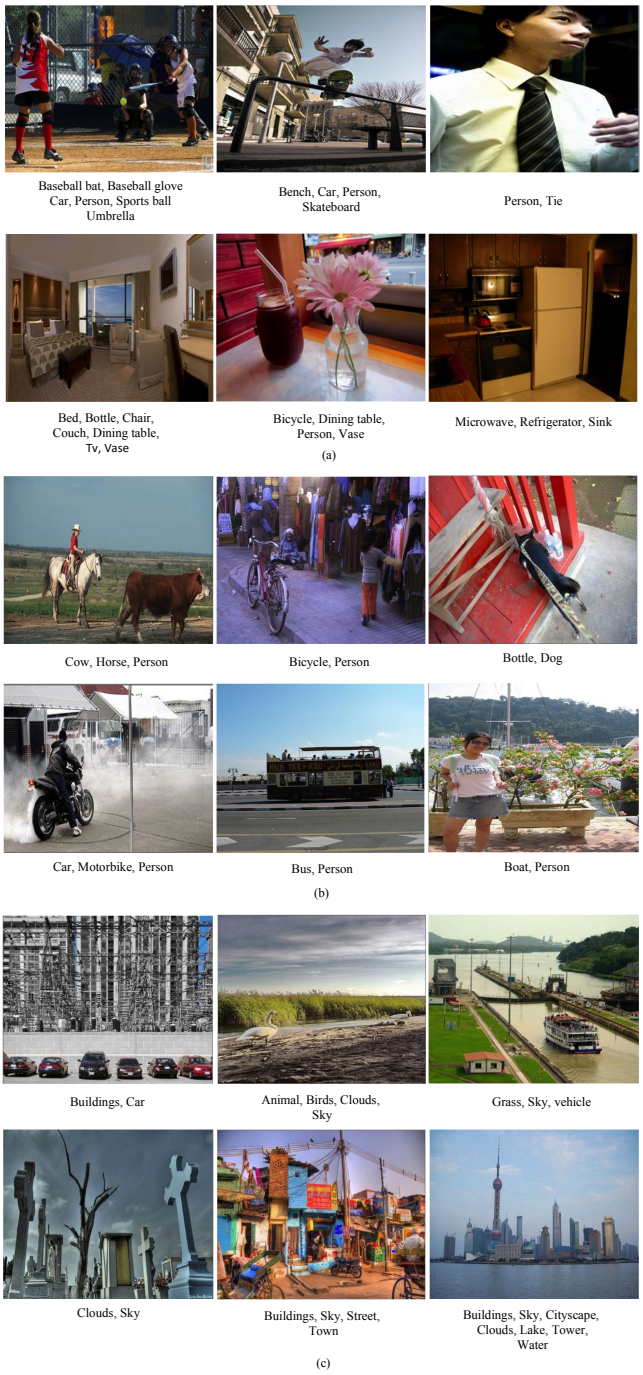



Figure 5. Sampled images from MS-COCO, Pascal VOC 2007 and NUS-WIDE. (a) MS-COCO. (b) Pascal VOC 2007. (c) NUS-WIDE.

## IV. Experiments

In this section, we first describe the evaluation metrics, implementation details and datasets. Then, we report the experimental results on three benchmark multi-label datasets, *i.e.*, *MS-COCO* [12], *NUS-WIDE* [55] and *VOC 2007* [56]. In the next section, we present the ablation studies on the key components of the proposed DER model. Finally, visualization and qualitative analyses are presented.

### A. Evaluation metrics

A comprehensive study of multi-label evaluation metrics is presented in [57]. According to this study, we compute macro/micro precision, macro/micro recall and macro/micro F1-measure for performance evaluation. Specifically, the macro precision (CP), macro recall (CR) and macro F1-measure (CF1) indicate the average pre-class precision, recall and F1-measure, respectively. Meanwhile, the micro precision (OP), micro recall (OR) and micro F1-measure (OF1) represent the average overall precision, recall and F1-measure, respectively. For each image, we assign labels with confidence greater than $0.5$ as positive and compare them with the ground-truth labels. These measures do not require a fixed number of labels per image. In particular, for the *MS-COCO* [12] and *NUS-WIDE* [55] datasets, we also report the results of top-3 labels with highest confidences in order to assure fair comparison with existing state-of-the-art methods, cf. [8], [58].

$$CP = \frac{1}{C}\sum_i \frac{N_i^c}{N_i^p}\,, \qquad OP = \frac{\sum_i N_i^c}{\sum_i N_i^p}\,,$$

$$CR = \frac{1}{C}\sum_i \frac{N_i^c}{N_i^g}\,, \qquad OR = \frac{\sum_i N_i^c}{\sum_i N_i^g}\,,$$

$$CF1 = \frac{2\times CP\times CR}{CP+CR}\,, \qquad OF1 = \frac{2\times OP\times OR}{OP+OR}\,,$$

where $C$ is the number of labels, and $N_i^c$ is the number of correctly predicted images for the $i$-th label. $N_i^g$ is the number of ground-truth images for the $i$-th label, and $N_i^p$ is the number of predicted images for the $i$-th label.

Additionally, in general, the average precision (AP) for each label and the mean average precision (mAP) are important for evaluating multi-label image recognition accuracy, and they are also employed for performance comparison. AP and mAP are computed by using Eq. (17) and Eq. (18), cf. [59].

$$AP(y_c) = \frac{1}{L_{y_c}}\sum_{n=1}^{N} P_{ry_c}(n)\times\left(R_{ry_c}(n)-R_{ry_c}(n-1)\right), \quad (17)$$

where $L_{y_c}$ is the number of images relevant to the label $y_c$, $N$ is the total number of retrieved images for the label $y_c$, $n$ is the rank in the list of retrieved images, and $P_{y_c}(n)$ and $R_{y_c}(n)$ are the precision and recall at rank n.

$$mAP = \frac{1}{C}\sum_{c=1}^{C} AP(y_c)\,. \quad (18)$$

### B. Implementation details

In our experiments, the input images are resized to $512\times512$ and randomly cropped into $448\times448$ with random horizontal flips for data augmentation. The transformation function $f_{\text{embed}}(\cdot;\theta_{\text{embed}})$ is a two-layer fully connected network with ReLU as its activation function. The dimensionality for both of the two fully connected layers in our network is 196, which corresponds to the feature map size of the last convolutional layer ($14\times14$). Following [7], [61], ResNet-101 [14] is selected as the backbone of our proposed model. We utilize the pre-trained model based on ImageNet for model parameter initialization. For optimization, conventional stochastic gradient descent (SGD) with momentum of $0.9$ is selected as the network optimizer. The weight decay is set to $10^{-4}$. The initial learning rate is $0.01$, and it is divided by $10$ every $20$ epochs until $60$, the total number of training epochs. We choose PyTorch[1] for conducting experiments. As described in Section III-E, we set $\alpha$ and $\beta$ to $0.5$ and $0.05$, respectively, in all of the experiments. The entire training process occurs in end-to-end fashion. All of the experiments are run on a computer with an Intel Xeon E5-2660 v4 processor, 120 GB main memory, and eight GTX-1080Ti GPUs.

### C. Datasets

We conduct performance evaluations using three popular benchmark multi-label image datasets: MS-COCO [12], Pascal VOC 2007 [56] and NUS-WIDE [55].

*1) Microsoft COCO dataset:* Microsoft COCO [12] is a widely used dataset for multi-label image recognition. The training set is composed of $82,081$ images, and the validation set consists of $40,504$ images. The dataset covers $80$ common object categories, each image belongs to 1-10 categories and each image contains approximately $3.5$ labels on average. Because the ground-truth labels of the test set are not available, we evaluate the performances of all methods on the validation set instead. For comparison with other methods on the MS-COCO dataset, we report mAP, CP, CR, CF1, OP, OR, and OF1.

*2) Pascal VOC 2007 dataset:* The PASCAL Visual Object Classes Challenge 2007 (*VOC 2007*) dataset [56] contains $9,963$ images from a total of $20$ object categories, each image has 1-5 ground-truth labels, and the average number of ground-truth labels per image is 1.4. VOC 2007 is a popular dataset used as the benchmark for multi-label recognition. *VOC 2007* is divided into train, val, and test sets. Following [10], [58], [62], we train our model on the trainval set and evaluate the recognition performance on the test set. For this dataset, the evaluation metrics are AP and mAP.

*3) NUS-WIDE:* The *NUS-WIDE* dataset [55] is another benchmark dataset for multi-label recognition that contains $269,648$ images with associated tags from `Flickr`[2]. This dataset is manually annotated by $81$ concepts, with $2.4$ concept labels per image on average. Official train/test splits are utilized, *i.e.*, $161,789$ images for training and $107,859$ images for testing. This dataset provides images with four different resolutions, namely, high-resolution images, medium-resolution images, low-resolution images and original resolution images, and we use low-resolution images in our experiment. We utilize the same evaluation metrics as MS-COCO for this dataset.

Some sampled images from the MS-COCO, Pascal VOC 2007 and NUS-WIDE datasets are shown in Fig. 5.

### D. Comparison with state-of-the-art methods

*1) Performance on the* MS-COCO *dataset:* On *MS-COCO*, we compare our proposed model with recent state-of-the-art

[1]PyTorch is available at: http://pytorch.org/

[2]https://www.flickr.com/

Table I
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE *MS-COCO* DATASET. * DENOTES THAT WE FOLLOW THE SETTINGS OF SSGRL [25].

| Methods | All | | | | | | | top-3 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mAP | CP | CR | CF1 | OP | OR | OF1 | CP | CR | CF1 | OP | OR | OF1 |
| WARP [20] | – | – | – | – | – | – | – | 59.3 | 52.5 | 55.7 | 59.8 | 61.4 | 60.7 |
| CNN-RNN [10] | 61.2 | – | – | – | – | – | – | 66.0 | 55.6 | 60.4 | 69.2 | 66.4 | 67.8 |
| RNN-Attention [8] | – | – | – | – | – | – | – | 79.1 | 58.7 | 67.4 | 84.0 | 63.0 | 72.0 |
| Order-Free RNN [60] | – | – | – | – | – | – | – | 71.6 | 54.8 | 62.1 | 74.2 | 62.2 | 67.7 |
| ML-ZSL [24] | – | – | – | – | – | – | – | 74.1 | **64.5** | 69.0 | – | – | – |
| SRN [7] | 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 |
| Multi-Evidence [61] | – | 80.4 | 70.2 | 74.9 | 85.2 | 72.5 | 78.4 | 84.5 | 62.2 | 70.6 | 89.1 | 64.3 | 74.7 |
| ML-GCN [9] | 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** |
| SSGRL [25] | 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 |
| ResNet-101 (Baseline) | 78.3 | 80.2 | 66.7 | 72.8 | 83.9 | 70.8 | 76.8 | 84.1 | 59.4 | 69.7 | 89.1 | 62.8 | 73.6 |
| Our DER | 82.8 | 84.7 | 71.6 | 77.6 | 86.0 | 74.9 | 80.0 | 88.7 | 63.7 | 74.1 | 90.6 | 66.1 | 76.4 |
| Our DER* | **83.9** | 87.2 | 70.7 | **78.1** | 88.4 | 73.5 | **80.3** | 90.4 | 63.9 | **74.8** | 92.0 | 65.8 | **76.7** |

Table II
COMPARISONS OF AP AND MAP WITH STATE-OF-THE-ART METHODS ON THE *VOC 2007* DATASET. * DENOTES THAT WE FOLLOW THE SETTINGS OF SSGRL [25].

| Methods | *aero* | *bike* | *bird* | *boat* | *bottle* | *bus* | *car* | *cat* | *chair* | *cow* | *table* | *dog* | *horse* | *motor* | *person* | *plant* | *sheep* | *sofa* | *train* | *tv* | mAP |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CNN-SVM [63] | 88.5 | 81.0 | 83.5 | 82.0 | 42.0 | 72.5 | 85.3 | 81.6 | 59.9 | 58.5 | 66.5 | 77.8 | 81.8 | 78.8 | 90.2 | 54.8 | 71.1 | 62.6 | 87.4 | 71.8 | 73.9 |
| CNN-RNN [10] | 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 |
| DCNN-VGG [64] | 97.5 | 93.0 | 95.6 | 93.5 | 59.8 | 86.3 | 94.8 | 94.5 | 68.5 | 84.6 | 83.7 | 92.5 | 94.7 | 89.0 | 97.4 | 73.5 | 87.3 | 75.0 | 98.4 | 84.0 | 87.2 |
| RLSD [62] | 96.4 | 92.7 | 93.8 | 94.1 | 71.2 | 92.5 | 94.2 | 95.7 | 74.3 | 90.0 | 74.2 | 95.4 | 96.2 | 92.1 | 97.9 | 66.9 | 93.5 | 73.7 | 97.5 | 87.6 | 88.5 |
| FeV+LV [6] | 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 |
| DCNN-RN [64] | 98.6 | 96.5 | 98.7 | 97.1 | 68.2 | 91.8 | **97.8** | 97.5 | 74.7 | 87.5 | 86.8 | 97.4 | 98.0 | 93.3 | 98.8 | 77.5 | 91.3 | 79.6 | 99.0 | 87.3 | 90.8 |
| HCP [16] | 98.6 | 97.1 | 98.0 | 95.6 | 75.3 | 94.7 | 95.8 | 97.3 | 73.1 | 90.2 | 80.0 | 97.3 | 96.1 | 94.9 | 96.3 | 78.3 | 94.7 | 76.2 | 97.9 | 91.5 | 90.9 |
| RNN-Attention (512) [8] | 98.5 | 96.7 | 95.6 | 95.7 | 73.7 | 92.1 | 95.8 | 96.8 | 76.5 | 92.9 | 87.2 | 96.6 | 97.5 | 92.8 | 98.3 | 76.9 | 91.3 | 83.6 | 98.6 | 88.1 | 91.3 |
| Atten-Reinforce (512) [58] | 98.6 | 96.9 | 96.3 | 94.8 | 74.1 | 91.9 | 96.3 | 97.1 | 76.9 | 91.4 | 86.2 | 96.6 | 96.4 | 93.1 | 98.0 | 79.8 | 91.7 | 83.1 | 98.3 | 88.6 | 91.3 |
| SSGRL [25] | 99.5 | 97.1 | 97.6 | 97.8 | **82.6** | 94.8 | 96.7 | 98.1 | 78.0 | **97.0** | 85.6 | 97.8 | 98.3 | 96.4 | 98.8 | 84.9 | 96.5 | 79.8 | 98.4 | 92.8 | 93.4 |
| ML-GCN [9] | 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 |
| VGG (Baseline) | 98.5 | 96.1 | 96.9 | 93.6 | 76.8 | 88.1 | 96.5 | 96.6 | 74.1 | 86.8 | 79.5 | 95.8 | 94.3 | 93.0 | 98.9 | 77.3 | 87.1 | 75.5 | 97.5 | 88.9 | 89.6 |
| VGG-DER | 98.4 | 97.7 | 98.3 | 94.2 | 79.4 | 92.1 | 97.1 | 98.0 | 79.7 | 89.9 | 87.7 | 97.3 | 96.2 | 95.3 | 99.0 | 81.8 | 86.0 | 81.9 | 98.1 | 91.2 | 92.0 |
| ResNet-101 (Baseline) | 99.5 | 97.7 | 97.8 | 96.4 | 65.7 | 91.8 | 96.1 | 97.6 | 74.2 | 80.9 | 85.0 | 98.4 | 96.5 | 95.9 | 98.4 | 70.1 | 88.3 | 80.2 | 98.9 | 89.2 | 89.9 |
| Our DER | **99.7** | **98.6** | 98.5 | **98.4** | 80.4 | 94.5 | 97.5 | 97.7 | 84.0 | 96.0 | 87.2 | 98.1 | 98.4 | 96.4 | 99.1 | 85.3 | 96.6 | 84.4 | **99.8** | **94.1** | 94.2 |
| Our DER* | **99.7** | 98.3 | **98.8** | 98.2 | 81.5 | **95.6** | 97.7 | **98.4** | **84.1** | 96.5 | **88.9** | **98.8** | **98.6** | **96.7** | **99.3** | **85.4** | **97.0** | **87.1** | 98.7 | 93.8 | **94.6** |

Table III
COMPARISONS WITH STATE-OF-THE-ART METHODS ON THE *NUS-WIDE* DATASET. * DENOTES THAT WE FOLLOW THE SETTINGS OF SSGRL [25].

| Methods | All | | | | | | | top-3 | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | mAP | CP | CR | CF1 | OP | OR | OF1 | CP | CR | CF1 | OP | OR | OF1 |
| KNN [55] | – | – | – | – | – | – | – | 32.6 | 19.3 | 24.3 | 43.9 | 53.4 | 47.6 |
| WARP [20] | – | – | – | – | – | – | – | 31.7 | 35.6 | 33.5 | 48.6 | 60.5 | 53.9 |
| CNN-RNN [10] | – | – | – | – | – | – | – | 40.5 | 30.4 | 34.7 | 49.9 | 61.7 | 55.2 |
| Order-Free RNN [60] | – | – | – | – | – | – | – | 59.4 | 50.7 | 54.7 | 69.0 | **71.4** | 70.2 |
| ML-ZSL [24] | – | – | – | – | – | – | – | 43.4 | 48.2 | 45.7 | – | – | – |
| SRN [7] | 62.0 | 65.2 | 55.8 | 58.5 | 75.5 | 71.5 | 73.4 | 48.2 | **58.8** | 48.9 | 56.2 | 69.6 | 62.2 |
| ResNet-101 (Baseline) | 60.4 | 63.1 | 55.5 | 59.1 | 74.3 | 71.7 | 72.9 | 64.9 | 48.3 | 55.3 | 76.8 | 62.1 | 68.7 |
| Our DER | 63.0 | 64.2 | **57.9** | 60.9 | 75.5 | **73.0** | **74.2** | 66.5 | 49.2 | 56.5 | 78.3 | 63.2 | **70.7** |
| Our DER* | **63.3** | **65.5** | 57.5 | **61.2** | **76.6** | 72.0 | **74.2** | **67.6** | 48.6 | **56.6** | **79.2** | 63.6 | 70.5 |

methods. The comparison results are reported in Table I. It is clearly observed that our DER method outperforms the previous state-of-the-art methods with the same experimental settings. For example, our DER can obtain $+5.7\%$ mAP improvement over the SRN method. For comparisons with ML-GCN and SSGRL, which adopt more complex data augmentation strategies and higher resolutions for training images, we strictly follow the settings of SSGRL to reimplement our method (DER*), and our method can also obtain comparable results.

*2) Performance on the* VOC 2007 *dataset:* We compare our method with the recent state-of-the-art methods on *VOC 2007*. The experimental results are presented in Table II. Since many previous methods used the VGG model as their base model, for fair comparisons, we also report the results using VGG models. It is apparent to observe that in comparison with the previous methods, our proposed method offers significant improvement: we achieve $94.2\%$ mAP on *VOC 2007*, which outperforms the state of the art. Even when using the VGG model as the base model, we can still obtain $92.0\%$ mAP,

Table IV
MAP PERFORMANCE ON THREE DATASETS FOR DIFFERENT OPERATION COMBINATIONS.

| Operations | | | mAP | | |
|---|---|---|---|---|---|
| Disentangle | Embed | Rank | COCO | VOC 2007 | NUS-WIDE |
| | | | 78.3 | 89.9 | 60.4 |
| ✓ | | | 79.9 | 91.7 | 61.6 |
| ✓ | ✓ | | 82.3 | 93.7 | 62.8 |
| ✓ | ✓ | ✓ | **82.8** | **94.2** | **63.0** |

Table V
INFLUENCE OF DIFFERENT CONFIDENCE.

| Methods | mAP | | |
|---|---|---|---|
| | COCO | VOC 2007 | NUS-WIDE |
| ResNet-101 (Baseline) | 78.3 | 89.9 | 60.4 |
| First confidence ($\hat{\boldsymbol{y}}_{\text{fcls}}$) | 80.5 | 92.6 | 61.5 |
| Second confidence ($\hat{\boldsymbol{y}}_{\text{scls}}$) | 79.3 | 92.3 | 54.9 |
| Joint confidence ($\hat{\boldsymbol{y}}$) | **82.8** | **94.2** | **63.0** |

which is $0.7\%$ higher than that of the state of the art. We also reimplement our method following the settings of SSGRL and achieve $94.6\%$, outperforming ML-GCN and SSGRL by $0.6\%$ and $1.2\%$, respectively.

*3) Performance on the* NUS-WIDE *dataset:* Table III shows the comparisons with recent state-of-the-art methods on the *NUS-WIDE* dataset. We can obviously find that our method achieves the best multi-label recognition performance compared with that of the previous state-of-the-art methods, especially for the evaluations of mAP, CF1 (All), OF1 (All), CF1 (top-3) and OF1 (top-3).

### *E. Ablation studies*

In this section, we perform ablation studies from four different aspects, including the impact of different operation combinations on accuracy, effects of different confidences, the effect of $\alpha$ for label correlation embedding and the effect of $\beta$ for label ranking.

*1) The impact of different operation combinations:* We perform an ablation study to analyze the impact of different operation combinations in Table IV. The setting details are as follows:

- ResNet-101 (Baseline): We use ResNet-101 with global max pooling as our baseline.
- Baseline + *Disentangle*: By setting the trade-off parameters in Eq. (16) to 0, the proposed losses will make no contribution to the network learning.
- Baseline + *Disentangle* + *Embed*: Setting $\alpha = 0.5$ and $\beta = 0$ activates the label correlation embedding operation.
- Baseline + *Disentangle* + *Embed* + *Rank*: With $\alpha = 0.5$ and $\beta = 0.05$, both operations will function jointly.

The experiment results show that introducing the class-aware map disentangling operation leads to a slight improvement ($+1.0\%$). This operation can be considered a naïve attention module for the general deep feature. When the label correlation embedding operation and disentangling operation function jointly, there is a significant performance boost ($+2.0\%$) for all benchmarks. The improvement proves the effectiveness

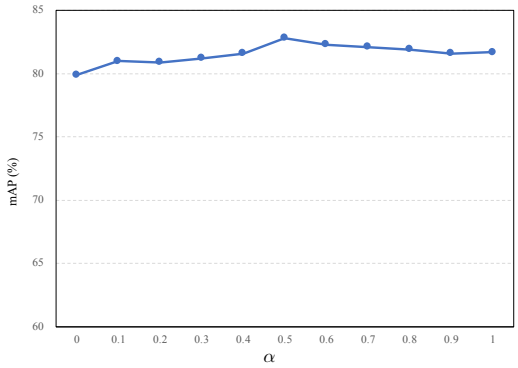



Figure 6. Accuracy comparisons with different values of $\alpha$ on the MS-COCO dataset.

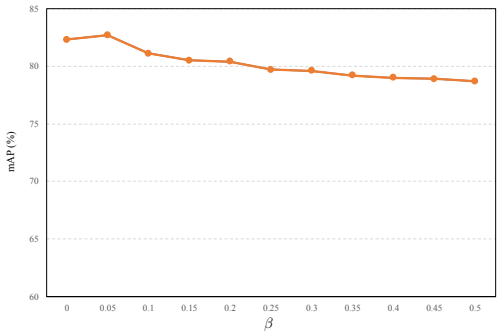



Figure 7. Accuracy comparisons with different values of $\beta$ on the MS-COCO dataset.

of the proposed operations for modeling the label correlation and the operations and further refines the backbone for better representation ability. In addition, bringing in the label ranking loss achieves better mAP because it helps to accurately and robustly measure the similarity/dissimilarity of these label vectors.

*2) Effects of different confidences:* To demonstrate the integrity of our method, we test the multi-label classification performance according to prediction score from only the first or the second stream. It is worth noting that the prediction scores are extracted from a well-trained network under the following settings:

- The first stream confidence means that we only use $\hat{\boldsymbol{y}}_{\text{fcls}}$ in Eq. (12) to calculate mAP.
- The second stream confidence denotes that we use $\hat{\boldsymbol{y}}_{\text{scls}}$ in Eq. (13) to obtain the result.
- Joint confidence, which is our final result, is determined using $\hat{\boldsymbol{y}}$ in Eq. (14) to obtain the accuracy.

The results are shown in Table V. The mAP score of the first stream is always higher than that of the second stream, but both scores are inferior to the jointly produced score, showing that both scores are interdependent. In particular, the first scores on three benchmarks are better than those produced by the baseline method, which indicates that our method greatly benefits the representation learning.

*3) Effects of different $\alpha$ for label correlation embedding:* To explore the effects of different $\alpha$ in Eq. (7) on multi-label

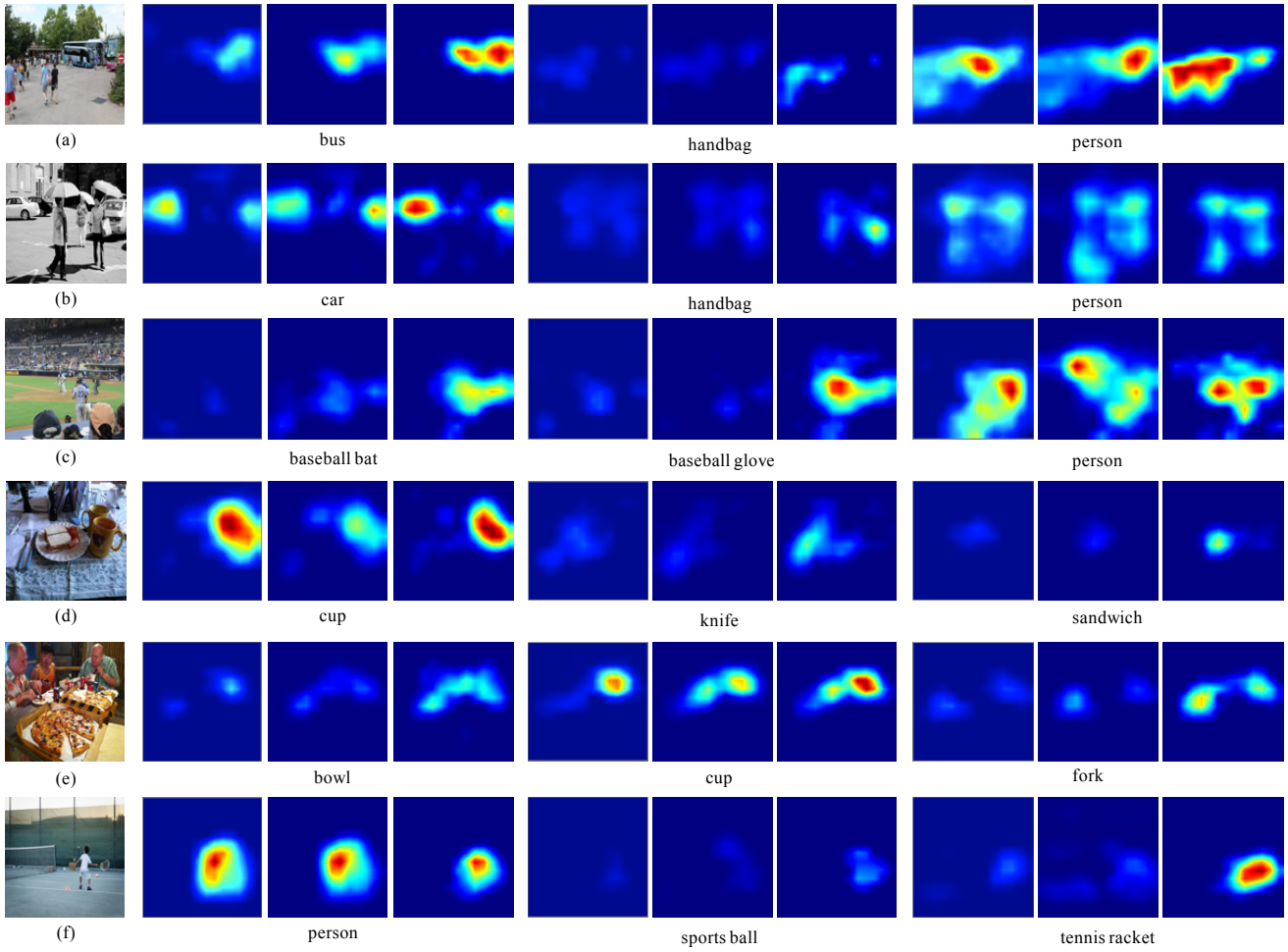



Figure 8. Visualization and comparisons of class-aware disentangled maps (CADMs). For each CADM, we first use linear interpolation to resize the CADM to $1 \times 448 \times 448$. Then, we utilize the visualization tools to convert each CADM to color map. For each category, the left CADM is generated by vanilla ResNet without disentangling during training, the middle and right CADMs are generated with the proposed disentangling step, where the middle map is generated without our metric learning component, and the right map is generated in cooperation with proposed metric learning. (Best viewed in color.)

image recognition performance, we change the values of $\alpha$ in a set of $\{0.0, 0.1, 0.2, 0.3, ..., 1.0\}$, as depicted in Fig. 6. We observe that when $\alpha = 0.5$, it can achieve the best performance on MS-COCO. Either increasing or decreasing the value of $\alpha$ will reduce the mAP. A low $\alpha$ reduces the impact of the label correlation embedding operation, and when $\alpha = 0$, the label correlation embedding operation has failed. Compared with the results of $\alpha = 0.5$, the model without class-aware disentangling and label correlation embedding exhibits a significant performance drop, *i.e.*, a $2.8\%$ lower mAP than that of our proposal. A high $\alpha$ will make representation learning more difficult.

*4) Effects of different $\beta$ values:* To study how parameter $\beta$ affects the performance accuracy, we conduct experiments using the MS-COCO dataset and set $\beta$ to $\{0.0, 0.05, 0.10, 0.15, ..., 0.5\}$ in turn. The mAP scores are shown in Fig. 7. From Fig. 7, we can obtain the optimal result when $\beta = 0.05$, which proves that the label ranking can make the distribution label vectors more compact and actualize an improvement in performance. However, when $\beta > 0.05$, the accuracy decreases as $\beta$ increases; one possible reason for this is that the high value of $\beta$ will impact the label correlation embedding and make representation learning more difficult.

Table VI
COMPARISONS WITH TRIPLET LOSS.

| Methods | mAP | | |
|---|---|---|---|
| | COCO | VOC 2007 | NUS-WIDE |
| ResNet-101 (Baseline) | 78.3 | 89.9 | 60.4 |
| D + Triplet Loss | 81.0 | 92.5 | 61.2 |
| Our DER | **82.8** | **94.2** | **63.0** |

*5) Comparison with triplet loss:* The triplet loss [39] with proper sampling strategy can also pull correlated label vectors together and push uncorrelated label vectors away. However, there are two differences between triplet loss and our method. First, triplet loss has high computational complexity due to its sampling strategy, while for our correlation embedding loss and ranking loss, all label embedding vectors are utilized for computation in one pass, making the optimization process more efficient. Second, if we treat all correlated label embeddings as one group and the uncorrelated label embeddings as the another group, triplet loss tends to pull the samples in each

group together (including the uncorrelated labels), which would, however, not be in accordance with our motivation. Our method is designed to push uncorrelated labels away from the "positive" dummy cluster centroid, rather than to pull them together. To compare the effectiveness between our proposed method and triplet loss, we have reimplemented triplet loss and reported the results in Table VI. While triplet loss achieves a certain degree of accuracy gain over the baseline method, it is still inferior to our method.

### F. Visualization and analyses

In this section, we validate the effectiveness of our proposed key operations (especially for label correlation embedding) according to visualization results from the qualitative perspective. We show the class-aware disentangled maps (CADMs) in Fig. 8 for comparison. Six sampled input images with the corresponding CADMs are presented in each subfigure. We select three of the multiple image labels that are apparent to be observed in the input image. For each category, the left CADM is generated by vanilla ResNet without disentangling during training, the middle and right CADMs are generated with the proposed disentangling step, where the middle map is generated without our metric learning component (*i.e.*, $\alpha = 0$ and $\beta = 0$), and the right map is generated in cooperation with proposed metric learning (*i.e.*, $\alpha = 0.5$ and $\beta = 0.05$). From these figures, it is clearly observed that utilizing our label correlation embedding could significantly strengthen the activations of these relevant labels' CADMs, e.g., "`bus`" of Fig. 8 (a), "`handbag`" of Fig. 8 (b), "`baseball bat`" and "`baseball glove`" of Fig. 8 (c), and "`tennis racket`" of Fig. 8 (f). It is reasonable to benefit from the recognition of the labels whose original CADM is weak. In addition, CADMs without our correlation embedding operation offer no obvious advantages over vanilla ResNet. Therefore, this consideration could provide an intuitive and straightforward explanation about why our model achieves the best multi-label image recognition accuracy on the three aforementioned benchmark datasets.

## V. Conclusion

In this paper, we proposed a unified framework for multi-label image recognition, which was composed of *disentangling*, *embedding* and *ranking* operations. The *disentangling* operation served as the foundation of explicitly modeling label correlations by producing the CADMs. Depending on the label vectors transformed from CADMs, the *embedding* operation pulled together relevant label vectors and pushed away irrelevant vectors in a metric learning fashion. The *ranking* operation accurately and robustly measured the similarity/dissimilarity of these label vectors. With only image-level supervision, our model could be trained in an end-to-end manner. The experimental results on three popular multi-label image recognition datasets and visualization analysis validated the effectiveness of the proposed model from both quantitative and qualitative perspectives. In the future, appropriate handling of the noisy and missing label problem [65] with our DER model merits further investigation.

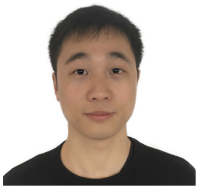
## Acknowledgement

Z.-M. Chen and Q. Cui's contributions were made when they were interns at Megvii Research Nanjing. This work was done when X.-S. Wei served as the Research Director in Megvii Research Nanjing. The authors would like to thank the editor and the anonymous reviewers for their critical and constructive comments and suggestions.

## References

[1] J. Shao, C.-C. Loy, K. Kang, and X. Wang, "Slicing convolutional neural network for crowd video understanding," in *CVPR*, 2016, pp. 5620–5628.
[2] Y. Li, C. huang, C. C. Loy, and X. Tang, "Human attribute recognition by deep hierarchical contexts," in *ECCV*, 2016, pp. 684–700.
[3] W. Liu and I. W. Tsang, "Making decision trees feasible in ultrahigh feature and label dimensions," *JMLR*, vol. 18, no. 1, pp. 2814–2849, 2017.
[4] M. Ivasic-Kos, M. Pobar, and S. Ribaric, "Two-tier image annotation model based on a multi-label classifier and fuzzy-knowledge representation scheme," *PR*, vol. 52, pp. 287–305, 2016.
[5] X.-S. Wei, Q. Cui, L. Yang, P. Wang, and L. Liu, "RPC: A large-scale retail product checkout dataset," *arXiv preprint arXiv:1901.07249*, 2019.
[6] H. Yang, J. Tianyi Zhou, Y. Zhang, B.-B. Gao, J. Wu, and J. Cai, "Exploit bounding box annotations for multi-label object recognition," in *CVPR*, 2016, pp. 280–288.
[7] F. Zhu, H. Li, W. Ouyang, N. Yu, and X. Wang, "Learning spatial regularization with image-level supervisions for multi-label image classification," in *CVPR*, 2017, pp. 5513–5522.
[8] Z. Wang, T. Chen, G. Li, R. Xu, and L. Lin, "Multi-label image recognition by recurrently discovering attentional regions," in *ICCV*, 2017, pp. 464–472.
[9] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *CVPR*, 2019, pp. 5177–5186.
[10] J. Wang, Y. Yang, J. Mao, Z. Huang, C. Huang, and W. Xu, "CNN-RNN: A unified framework for multi-label image classification," in *CVPR*, 2016, pp. 2285–2294.
[11] Z.-M. Chen, X.-S. Wei, X. Jin, and Y. Guo, "Multi-label image recognition with joint class-aware map disentangling and label correlation embedding," in *ICME*, 2019, p. in press.
[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014, pp. 740–755.
[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *CVPR*, 2009, pp. 248–255.
[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, pp. 1–14, 2014.
[16] Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "HCP: A flexible cnn framework for multi-label image classification," *IEEE TPAMI*, vol. 38, no. 9, pp. 1901–1907, 2016.
[17] Y. Zhu, J. Kwok, and Z.-H. Zhou, "Multi-label learning with global and local correlation," *IEEE TKDE*, vol. 30, no. 6, pp. 1081–1094, 2018.
[18] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE TKDE*, vol. 26, no. 8, pp. 1819–1837, 2014.
[19] C. Liu, P. Zhao, S.-J. Huang, Y. Jiang, and Z.-H. Zhou, "Dual set multi-label learning," in *AAAI*, 2018, pp. 3635–3642.
[20] Y. Gong, Y. Jia, T. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multi-label image annotation," *arXiv preprint arXiv:1312.4894*, pp. 1–9, 2013.
[21] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, and G. Mori, "Learning structural inference neural networks with label relations," in *CVPR*, 2016, pp. 2960–2968.
[22] Q. Li, M. Qiao, W. Bian, and D. Tao, "Conditional graphical LASSO for multi-label image classification," in *CVPR*, 2017, pp. 2977–2986.
[23] W. Liu, I. W. Tsang, and K.-R. Müller, "An easy-to-hard learning paradigm for multiple classes and multiple labels," *JMLR*, vol. 18, no. 1, pp. 3300–3337, 2017.
[24] C.-W. Lee, W. Fang, C.-K. Yeh, and Y.-C. F. Wang, "Multi-label zero-shot learning with structured knowledge graphs," in *CVPR*, 2018, pp. 1576–1585.

[25] T. Chen, M. Xu, X. Hui, H. Wu, and L. Lin, "Learning semantic-specific graph representation for multi-label image recognition," in *ICCV*, 2019, pp. 522–531.
[26] B. Cao, N.-N. Wang, X.-B. Gao, and J. Li, "Asymmetric joint learning for heterogeneous face recognition," in *AAAI*, 2018, pp. 6682–6689.
[27] Z. Dong, S. Jia, C. Zhang, M. Pei, and Y. Wu, "Deep manifold learning of symmetric positive definite matrices with application to face recognition," in *AAAI*, 2017, pp. 4009–4015.
[28] C. Ding and D. Tao, "Robust face recognition via multimodal deep face representation," *IEEE TMM*, vol. 17, no. 11, pp. 2049–2058, 2015.
[29] L. Dong, Y. Liang, G. Kong, Q. Zhang, X. Cao, and E. Izquierdo, "Holons visual representation for image retrieval," *IEEE TMM*, vol. 18, no. 4, pp. 714–725, 2016.
[30] I. González-Díaz, M. Birinci, F. Díaz-de-María, and E. J. Delp, "Neighborhood matching for image retrieval," *IEEE TMM*, vol. 19, no. 3, pp. 544–558, 2017.
[31] W. Chen, X. Chen, J. Zhang, and K. Huang, "A multi-task deep network for person re-identification," in *AAAI*, 2017, pp. 3988–3994.
[32] J. Jiao, W.-S. Zheng, A. Wu, X. Zhu, and S. Gong, "Deep low-resolution person re-identification," in *AAAI*, 2018, pp. 6967–6974.
[33] Z. Wang, R. Hu, C. Liang, Y. Yu, J. Jiang, M. Ye, J. Chen, and Q. Leng, "Zero-shot person re-identification via cross-view consistency," *IEEE TMM*, vol. 18, no. 2, pp. 260–272, 2016.
[34] Y. Bai, Y. Lou, F. Gao, S. Wang, Y. Wu, and L.-Y. Duan, "Group-sensitive triplet embedding for vehicle reidentification," *IEEE TMM*, vol. 20, no. 9, pp. 2385–2399, 2018.
[35] Z. Liu, Z. Lin, X. Wei, and S. Chan, "A new model-based method for multi-view human body tracking and its application to view transfer in image-based rendering," *IEEE TMM*, vol. 20, no. 6, pp. 1321–1334, 2018.
[36] Y. Chen, N. Wang, and Z. zhang, "Darkrank: Accelerating deep metric learning via cross sample similarities transfer," in *AAAI*, 2018, pp. 2852–2859.
[37] H. Wang, L. Feng, J. Zhang, and Y. Liu, "Semantic discriminative metric learning for image similarity measurement," *IEEE TMM*, vol. 18, no. 8, pp. 1579–1589, 2016.
[38] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *CVPR*, 2006, pp. 1735–1742.
[39] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *CVPR*, 2015, pp. 815–823.
[40] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," *arXiv preprint arXiv:1703.07737*, pp. 1–8, 2017.
[41] W. Ge, "Deep metric learning with hierarchical triplet loss," in *ECCV*, 2018, pp. 269–285.
[42] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *ECCV*, 2016, pp. 499–515.
[43] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *CVPR*, 2017, pp. 212–220.
[44] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *CVPR*, 2019, pp. 4690–4699.
[45] S.-Z. Chen, C.-C. Guo, and J.-H. Lai, "Deep ranking for person re-identification via joint representation learning," *IEEE TIP*, vol. 25, no. 5, pp. 2353–2367, 2016.
[46] K. He, F. Cakir, S. Adel Bargal, and S. Sclaroff, "Hashing as tie-aware learning to rank," in *CVPR*, 2018, pp. 4023–4032.
[47] E. K. G. H. R. G. N. M. R. Xinshao Wang, Yang Hua, "Ranked list loss for deep metric learning," in *CVPR*, 2019, pp. 5207–5216.
[48] M. Opitz, G. Waltner, H. Possegger, and H. Bischof, "Deep metric learning with bier: Boosting independent embeddings robustly," *TPAMI*, p. in press, 2018.
[49] W. Liu, D. Xu, I. W. Tsang, and W. Zhang, "Metric learning for multi-output tasks," *TPAMI*, vol. 41, no. 2, pp. 408–422, 2018.
[50] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016, pp. 2921–2929.
[51] G. E. Hinton, "Learning distributed representations of concepts," in *CogSci*, 1986, pp. 1–12.
[52] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE TPAMI*, vol. 35, no. 8, pp. 1798–1828, 2013.
[53] K. Sohn, "Improved deep metric learning with multi-class N-pair loss objective," in *NIPS*, 2016, pp. 1857–1865.
[54] L. Liu, C. Shen, and A. van den Hengel, "Cross-convolutional-layer pooling for image recognition," *IEEE TPAMI*, vol. 39, no. 11, pp. 2305–2313, 2016.
[55] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *CIVR*, 2009, pp. 1–9.
[56] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010.
[57] X.-Z. Wu and Z.-H. Zhou, "A unified view of multi-label performance measures," in *ICML*, 2017, pp. 3780–3788.
[58] T.-S. Chen, Z.-X. Wang, G.-B. Li, and L. Lin, "Recurrent attentional reinforcement learning for multi-label image recognition," in *AAAI*, 2018, pp. 6730–6737.
[59] L. Song, J. Liu, B. Qian, M. Sun, K. Yang, M. Sun, and S. Abbas, "A deep multi-modal cnn for multi-instance multi-label image classification," *IEEE TIP*, vol. 27, no. 12, pp. 6025–6038, 2018.
[60] S.-F. Chen, Y.-C. Chen, C.-K. Yeh, and Y.-C. F. Wang, "Order-free RNN with visual attention for multi-label classification," in *AAAI*, 2018, pp. 6714–6721.
[61] W. Ge, S. Yang, and Y. Yu, "Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning," in *CVPR*, 2018, pp. 1277–1286.
[62] J. Zhang, Q. Wu, C. Shen, J. Zhang, and J. Lu, "Multi-label image classification with regional latent semantic dependencies," *IEEE TMM*, vol. in press, 2018.
[63] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *CVPR Workshop*, 2014, pp. 512–519.
[64] W. Shi, Y. Gong, X. Tao, and N. Zheng, "Training dcnn by combining max-margin, max-correlation objectives, and correntropy loss for multilabel image classification," *IEEE TNNLS*, vol. 29, no. 7, pp. 2896–2908, 2018.
[65] H.-F. Yu, P. Jain, P. Kar, and I. Dhillon, "Large-scale multi-label learning with missing labels," in *ICML*, 2014, pp. 593–601.



**Zhao-Min Chen** received the B.S. degree from Hunan University and is now a Ph.D. candidate of computer science and technology from Nanjing University. He has published several academic papers on international conferences, such as ICME, CVPR, etc. His research interets are deep learning, computer vision and multi-label image recognition.

**Quan Cui** received his B.S. degree in Electronic Engineering from Southeast University. He is now a Ph.D. candidate of computer science from Waseda University. He has published several papers on premier conferences, such as CVPR. His research interests are computer vision and pattern recognition.

**Xiu-Shen Wei** (M'18) received his BS degree in computer science, and received his Ph.D. degree in computer science and technology from Nanjing University. He is a Professor at Nanjing University of Science and Technology. Before joining NJUST, he served as the Research Director of Megvii Research Nanjing, Megvii Technology. He has published more than twenty academic papers on the top-tier international journals and conferences, such as IEEE TPAMI, IEEE TIP, IEEE TNNLS, IEEE TKDE, Machine Learning, CVPR, ICCV, ECCV, IJCAI, ICDM, ACCV, etc. He achieved the first place in the iWildCam competition (in association with CVPR 2020), the first place in the iNaturalist competition (in association with CVPR 2019), the first place in the Apparent Personality Analysis competition (in association with ECCV 2016) and the first runner-up in the Cultural Event Recognition competition (in association with ICCV 2015) as the team director. He also received the Presidential Special Scholarship (the highest honor for Ph.D. students) in Nanjing University, and received the Outstanding Reviewer Award in CVPR 2017. His research interests are computer vision and machine learning. He has served as a PC member of CVPR, ICCV, ECCV, NeurIPS, IJCAI, AAAI, etc. He is a member of the IEEE.

**Xin Jin** received the B.Sc., M.Sc., and Ph.D. degrees from the Department of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China, in 2009, 2012, and 2017, respectively. He is currently a Computer Vision Researcher with Megvii Research Nanjing. His research interests include computer vision and deep learning, especially focusing on face landmark detection and general object detection.

**Yanwen Guo** received the Ph.D degree in applied mathematics from the State Key Lab of CAD&CG, Zhejiang University, China, in 2006. He is currently a professor at the National Key Lab for Novel Software Technology, Department of Computer Science and Technology, Nanjing University, Jiangsu, China. He worked as a visiting professor in the Department of Computer Science and Engineering, The Chinese University of Hong Kong, in 2006 and 2009, respectively, and the Department of Computer Science, The University of Hong Kong, in 2008, 2012, and 2013, respectively. He was a visiting scholar in the Department of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, from 2013 to 2015. His research interests include image and video processing, vision, and computer graphics.