DiCA: Disambiguated Contrastive Alignment for Cross-Modal Retrieval with Partial Labels

Chao Su [1], Huiming Zheng [2], Dezhong Peng [1,2], Xu Wang [1]*

[1]The College of Computer Science, Sichuan University, Chengdu, China
[2]Sichuan National Innovation New Vision UHD Video Technology Co., Ltd., Chengdu, China
suchao@stu.scu.edu.cn, michaelzheng@uptcsc.com, pengdz@scu.edu.cn, wangxu.scu@gmail.com

Abstract

Cross-modal retrieval aims to retrieve relevant data across different modalities. Driven by costly massive labeled data, existing cross-modal retrieval methods achieve encouraging results. To reduce annotation costs while maintaining performance, this paper focuses on an untouched but challenging problem, i.e., cross-modal retrieval with partial labels (PLCMR). PLCMR faces the dual challenges of annotation ambiguity and modality gap. To address these challenges, we propose a novel method termed disambiguated contrastive alignment (DiCA) for cross-modal retrieval with partial labels. Specifically, DiCA proposes a novel non-candidate boosted disambiguation learning mechanism (NBDL), which elaborately balances the trade-off between the losses on candidate and non-candidate labels that eliminate label ambiguity and narrow the modality gap. Moreover, DiCA presents an instance-prototype representation learning mechanism (IPRL) to enhance the model by further eliminating the modality gap at both the instance and prototype levels. Thanks to NBDL and IPRL, our DiCA effectively addresses the issues of annotation ambiguity and modality gap for cross-modal retrieval with partial labels. Experiments on four benchmarks validate the effectiveness of our proposed method, which demonstrates enhanced performance over existing state-of-the-art methods.

Code — https://github.com/Rose-bud/DiCA.

Introduction

With the rapid growth of multimedia data on the Internet (Wang et al. 2019, 2020, 2023a,b; Su et al. 2023), the limitations of single-modal retrieval systems becoming increasingly apparent. Consequently, cross-modal retrieval has garnered significant interest due to its ability to facilitate flexible searches across different data modalities (Cao et al. 2022, 2023, 2024). Existing methods can be divided into two broad categories: supervised and unsupervised cross-modal retrieval. The supervised cross-modal retrieval methods (Zhen et al. 2019; Wang and Peng 2021; Sun et al. 2023, 2024a,b) learn modality-invariant features by leveraging the correct label information. However, all of the above methods typically require massive labeled data, which pose formidable obstacles to data collection. Additionally, data annotation in the real world can be naturally subject to inherent label ambiguity and noise. To the end, numerous unsupervised cross-modal retrieval methods (Hardoon, Szedmak, and Shawe-Taylor 2004; Andrew et al. 2013; Wang et al. 2015; Liu et al. 2023) have been proposed, which project data from different modalities into a common space by maximizing the correlation. Although these methods can reduce annotation costs, the modality gap still has significant room for improvement, primarily due to the challenges posed by the lack of label information.

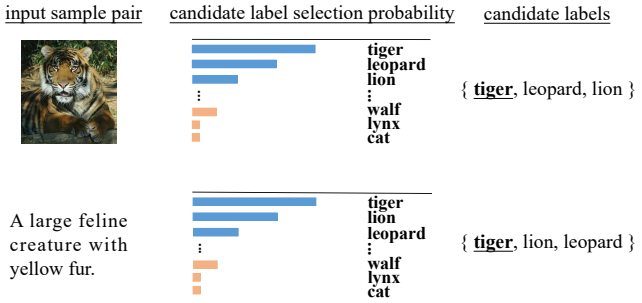


Figure 1: An input sample pair with three candidate labels, where the ground truth is **tiger**.

To reduce the costs of annotation while maintaining the performance of cross-modal retrieval, we propose a novel paradigm named cross-modal retrieval with partial labels (PLCMR). In PLCMR, annotators only need to provide a set of candidate labels and ensure that the ground truth label is included. As a result, the annotation costs can be significantly reduced. This paradigm can effectively handle difficult-to-distinguish sample pairs, such as in Fig. 1, where annotators are uncertain whether to label the image and text as a tiger, lion, or leopard. In such case, annotators can provide a candidate label set including all these three labels for training. Compared to supervised and unsupervised counterparts, the challenges of PLCMR lie in learning discriminative representations from partial labels with ambiguity while eliminating heterogeneous gap across modalities.

To tackle these challenges, we propose a novel method termed disambiguated contrastive alignment (DiCA), which can learn cross-modal invariant features when trained solely

*Corresponding author.

with a set of candidate labels. DiCA effectively integrates two distinct mechanisms, i.e., the non-candidate boosted disambiguation learning mechanism (NBDL) and the instance-prototype representation learning mechanism (IPRL), to enhance the performance of PLCMR task. Specifically, inspired by the leveraged weighted (LW) loss (Wen et al. 2021), NBDL is proposed to leverage the complementary information in the non-candidate label set. However, unlike LW, our NBDL projects data from different modalities into a common space, which helps resolve label ambiguity and reduces the modality gap. To further eliminate the modality gap, we additionally design a novel instance-prototype representation learning mechanism (IPRL). IPRL endows the model the ability to learn modal-invariant features at the instance and prototype levels through two modules, i.e., the instance-wise cross-modal contrast (ICC) module and the prototype-wise cross-modal alignment (PCA) module. In specific, ICC uses pseudo labels to guide the model in learning invariant features across modalities. PCA eliminates modality gap by minimizing the similarity differences between the learned representations and the prototype vectors across distinct modalities. With the support of ICC and PCA, the proposed IPRL reduces the modality gap and enhances the model's ability to learn a more accurate label distribution, thereby improving label disambiguation in NBDL. In return, NBDL provides clearer label information, which guides and further strengthens the IPRL process.

The main contributions are summarized as follows: (1) We propose a novel method called DiCA to tackle an untouched problem, i.e., cross-modal retrieval with partial labels. To the best of our knowledge, this work could be the first study on this problem. (2) A novel non-candidate boosted disambiguation learning mechanism (NBDL) is presented to consider the trade-off between the loss on candidate and non-candidate labels to promote label disambiguation and narrow the modality gap. (3) A novel instance-prototype representation learning mechanism (IPRL) is proposed to learn modal-invariant information and further eliminate the modality gap. (4) Extensive experiments on four widely-used benchmarks demonstrate the effectiveness of the proposed DiCA.

Related Work

Partial Label Learning

In partial label learning, each instance is associated with a set of candidate labels with only one being the ground truth and others being the ambiguity. To tackle this challenge, existing methods can be broadly divided into two categories: averaging-based and identification-based approaches. For the averaging-based approaches (Hüllermeier and Beringer 2006; Cour, Sapp, and Taskar 2011; Zhang and Yu 2015), all candidate labels are treated equally and the prediction is made by averaging the results of their modeling outputs. However, these methods are often limited in performance due to being misled by false positive labels. For the identification-based disambiguation approaches (Jin and Ghahramani 2002; Nguyen and Caruana 2008; Liu and Dietterich 2012; Zhang, Zhou, and Liu 2016; Wang et al. 2021), the ground truth label is treated as a latent variable and can be identified through iterative optimization procedure, such as Expectation-Maximization (EM) algorithm.

Recently, deep-learning-based partial label learning methods (Lv et al. 2020; Feng et al. 2020; Wen et al. 2021; Wang et al. 2022; Xia et al. 2023; Liu et al. 2024) have made some new progress. For instance, Lv et al. (2020) approximately minimizes a risk estimator to identify the true label seamlessly. Motivated by contrastive learning, PiCO (Wang et al. 2022) contrastively learns representations and introduces a novel class prototype-based label disambiguation approach. Additionally, Si et al. (2024) introduces a novel partner classifier and proposes a novel "mutual supervision" paradigm, thereby improving the disambiguation capability of the base classifier. However, these methods are all specifically designed for unimodal scenarios, which may not be satisfactory for multimodal tasks due to the huge modality gap.

Cross-Modal Retrieval

Cross-modal retrieval is a task that aims to find relevant data from different modalities. The critical challenge for the task lies in how to bridge the modality gap. To tackle this challenge, numerous approaches have been proposed, which can be divided into two categories: supervised cross-modal retrieval methods and unsupervised cross-modal retrieval methods. More specifically, 1) The supervised cross-modal retrieval methods (Zhen et al. 2019; Wang and Peng 2021; Sun et al. 2023; Wang et al. 2024b) leverage the annotated labels to learn a discriminative common space for different modalities. For instance, Sun et al. (2023) proposes a coarse-to-fine hierarchical hashing strategy to fully leverage the hierarchical feature information across various modalities. However, these methods typically require massive labeled data, which pose formidable obstacles to data collection. 2) The unsupervised cross-modal retrieval methods (Harold 1936; Andrew et al. 2013; Wang et al. 2015; Liu et al. 2023) learn modality-specific transformations by maximizing correlations between different modalities without relying on label information. For example, SCL (Liu et al. 2023) uses unsupervised contrastive learning to cultivate more discriminative representations, capitalizing on the relationships among intra- and inter-modality instances. However, the aforementioned methods all suffer from performance drops due to the lack of supervision.

In this paper, we focus on an untouched but meaningful problem, i.e., cross-modal retrieval with partial labels (PLCMR), which could reduce annotation costs and maintain performance in cross-modal retrieval task.

Methodology

Problem Formulation

Notations. For a clear presentation, we first give some definitions for notations in this paper. Denote $\mathcal{X}$ as the input space, and $\mathcal{Y} = \{1, 2, ..., K\}$ be the label space where K denotes the number of classes. The training dataset $\mathcal{D}_I = \{(\boldsymbol{x}_i^I, Y_i^I)\}_{i=1}^N$ denotes the image modality and $\mathcal{D}_T = \{(\boldsymbol{x}_i^T, Y_i^T)\}_{i=1}^N$ denotes the text modality where N is the total number of sample pairs. $\boldsymbol{x}_i$ and Y_i represent the i-th

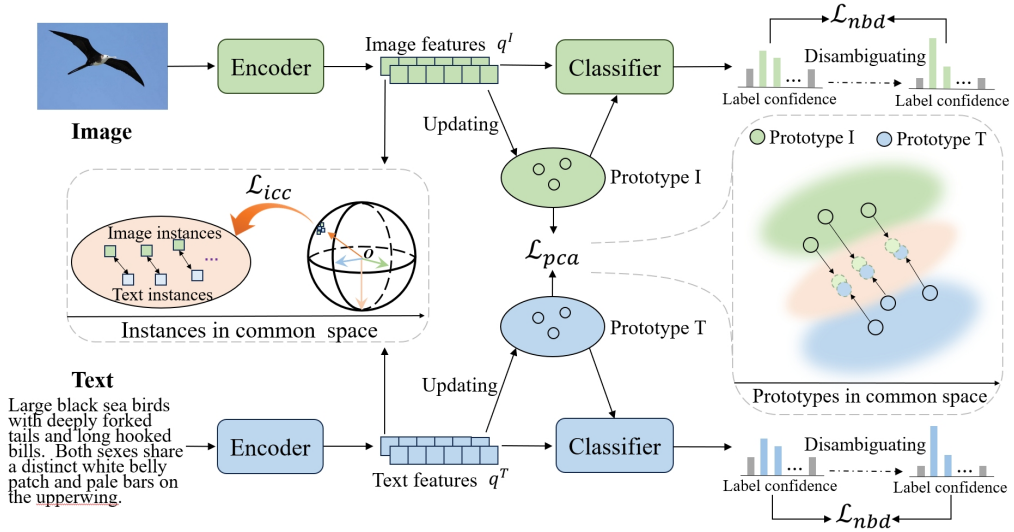


Figure 2: The pipeline of the proposed method DiCA for cross-modal retrieval with partial labels. NBDL ($\mathcal{L}_{nbd}$) balances the trade-off between the losses on candidate and non-candidate labels that eliminate label ambiguity and narrow the modality gap. Meanwhile, IPRL ($\mathcal{L}_{icc}$ and $\mathcal{L}_{pca}$) enhances the model by further eliminating the modality gap at both the instance and prototype levels.

sample pair and candidate label set, respectively. Further, the vector representation of Y is defined as $\boldsymbol{y} \in \mathbb{R}^K$, where the element corresponding to the class in Y is 1 and the others are 0. In partial learning tasks, each sample $\boldsymbol{x} \in \mathcal{X}$ is input with a candidate label set $Y \in \mathcal{Y}$. In the training process, the true label is hidden within a set of candidate labels. The goal of DiCA is to identify the true label from inherent ambiguities. Through an encoder $f(\cdot)$, the feature vector $\boldsymbol{q} \in \mathbb{R}^L$ can be computed by $\boldsymbol{q} = f(\boldsymbol{x})$, where L denotes the dimension of the common space. Moreover, a softmax function $g(\cdot)$ is applied to obtain the probability distribution $\boldsymbol{z} \in \mathbb{R}^K$ as $\boldsymbol{z} = g(\boldsymbol{q})$. Furthermore, we define the representations of the image and text modalities as $\mathcal{Q}^I$ and $\mathcal{Q}^T$, respectively. Meanwhile, the probability distributions of the image and text modalities are defined as $\mathcal{Z}^I$ and $\mathcal{Z}^T$.

Overview. The key challenge of PLCMR is to identify the ground truth label from the candidate set while eliminating the modality gap. As shown in Fig. 2, we present a non-candidate boosted disambiguation learning mechanism (NBDL) to resolve label ambiguity and narrow the modality gap. Furthermore, we propose an instance-prototype representation learning mechanism (IPRL) to further eliminate the modality gap at the instance and prototype levels. The overall objective function is formulated as:

$$\mathcal{L} = \underbrace{\mathcal{L}_{nbd}}_{NBDL} + \underbrace{\alpha \cdot \mathcal{L}_{icc} + \beta \cdot \mathcal{L}_{pca}}_{IPRL}, \tag{1}$$

where $\mathcal{L}_{nbd}$ is the loss function adopted by NBDL, $\mathcal{L}_{icc}$ and $\mathcal{L}_{pca}$ are the loss functions employed by IPRL. Additionally, α and β are the hyperparameters. Our DiCA is trained in a batch-by-batch manner by descending Eq. (1) with stochastic gradient descent. In the following subsections, we will elaborate on each component of the proposed DiCA.

Non-candidate Boosted Disambiguation Learning

To find the disambiguated label $\hat{\boldsymbol{y}}$ from the candidate label $\boldsymbol{y}$, many methods have been proposed by studying the candidate label set. However, they did not notice the simultaneous utilization of the non-candidate label set. Inspired from LW (Wen et al. 2021), we propose a new identification-based mechanism termed non-candidate boosted disambiguation learning (NBDL). NBDL is proposed to leverage the complementary information in the non-candidate label set. Different from LW, our NBDL projects data from different modalities into a common space, which achieves label disambiguation while reducing the modality gap. We formulate the label disambiguation process as follows:

$$\hat{\boldsymbol{y}}^I = \frac{\boldsymbol{z}^I \circ \boldsymbol{y}^I}{\|\boldsymbol{z}^I \circ \boldsymbol{y}^I\|}, \quad \hat{\boldsymbol{y}}^T = \frac{\boldsymbol{z}^T \circ \boldsymbol{y}^T}{\|\boldsymbol{z}^T \circ \boldsymbol{y}^T\|}, \tag{2}$$

where $\boldsymbol{z}^I \in \mathcal{Z}^I$ and $\boldsymbol{z}^T \in \mathcal{Z}^T$ represent the probabilities that the sample $\boldsymbol{x}^I$ and $\boldsymbol{x}^T$ belong to their respective class in the image and text modalities, respectively. By using the Hadamard product $\circ$, we can obtain the new disambiguated label $\hat{\boldsymbol{y}}$ as the guidance for subsequent training.

To reveal the true label from the candidate label set, NBDL encourages the model to reduce the predicted probability of non-candidate labels while disambiguating the candidate label set. By increasing the loss on non-candidate labels and penalizing high predicted probabilities for these labels, the false positive rate can be significantly reduced. The NBDL losses of the image and text modalities are as follows:

$$\begin{aligned}\mathcal{L}^I_{nbd} = &-\sum_{\boldsymbol{x}^I}\sum_{c \in Y^I} p(\boldsymbol{y}^I_c = 1 \mid \boldsymbol{x}^I) \log(p(\boldsymbol{y}^I_c = 1 \mid \boldsymbol{x}^I, f(\boldsymbol{x}^I)) \\ &+\lambda \sum_{\bar{c} \notin Y^I} p(\boldsymbol{y}^I_{\bar{c}} = 0 \mid \boldsymbol{x}^I, f(\boldsymbol{x}^I))^2,\end{aligned} \tag{3}$$

$$\mathcal{L}_{nbd}^{T} = -\sum_{\boldsymbol{x}^T}\sum_{c\in Y^T} p(\boldsymbol{y}_c^T = 1 \mid \boldsymbol{x}^T)\log(p(\boldsymbol{y}_c^T = 1 \mid \boldsymbol{x}^T, f(\boldsymbol{x}^T)) + \lambda \sum_{\overline{c}\notin Y^T} p(\boldsymbol{y}_{\overline{c}}^T = 0 \mid \boldsymbol{x}^T, f(\boldsymbol{x}^T))^2, \tag{4}$$

Where λ is the trade-off parameter between the loss on candidate and non-candidate labels. $f(\boldsymbol{x})$ denotes the output of the model. $p(\boldsymbol{y}_c = 1 \mid \boldsymbol{x})$ is the label confidence that the model assigns to the sample $\boldsymbol{x}$ belonging to class c. $p(\boldsymbol{y}_c = 1 \mid \boldsymbol{x}, f(\boldsymbol{x}))$ and $p(\boldsymbol{y}_{\overline{c}} = 0 \mid \boldsymbol{x}, f(\boldsymbol{x}))$ denote the predicted probabilities on the candidate and non-candidate categories after applying the softmax function.

Finally, the total loss of NBDL ($\mathcal{L}_{nbd}$) can be written as:

$$\mathcal{L}_{nbd} = \mathcal{L}_{nbd}^{I} + \mathcal{L}_{nbd}^{T}. \tag{5}$$

Instance-Prototype Representation Learning

With non-candidate boosted disambiguation learning, the model can solve label ambiguity and reduce the modality gap. However, due to the presence of label ambiguity, the modality gap still exists. To further reduce the modality gap, we design a new instance-prototype representation learning mechanism (IPRL). IPRL endows the model the ability to jointly learn modal-invariant features at the instance and prototype levels through two modules, i.e., the instance-wise cross-modal contrast (ICC) module and the prototype-wise cross-modal alignment (PCA) module.

Instance-wise Cross-Modal Contrast. Contrastive learning techniques are commonly used in learning multimodal representations. However, in partial label learning, the true label is hidden among a set of candidate labels. To address this challenge, we propose an instance-wise cross-modal contrast (ICC) module. Specifically, we use the pseudo label $\hat{k}$ to extract features across different modalities. The pseudo label $\hat{k}$ is formulated as follows:

$$\hat{k}^I = argmax(\hat{\boldsymbol{y}}^I), \quad \hat{k}^T = argmax(\hat{\boldsymbol{y}}^T). \tag{6}$$

To learn shared representations, we maximize the agreement between different modalities in a common space. By maximizing the distance between instances with the same pseudo label $\hat{k}$ across different modalities, the probability of an instance $\boldsymbol{x}_j^i$ belonging to the *j*-th instance across m modalities can be defined as follows:

$$p\left(j \mid \boldsymbol{x}_j^i\right) = \frac{\sum_{l=1}^{m} \mathbb{I}(\hat{k}_l = \hat{k}_j)\exp\left(\frac{1}{\tau} sim(\boldsymbol{z}_j^l, \boldsymbol{z}_j^i)\right)}{\sum_{l=1}^{m}\sum_{t=1}^{N}\exp\left(\frac{1}{\tau} sim(\boldsymbol{z}_t^l, \boldsymbol{z}_j^i)\right)}, \tag{7}$$

where τ is a temperature parameter (Wu et al. 2018; Caron et al. 2020), m denotes the number of modalities. The value of l is set to 1 for the image modality and 2 for the text modality. $\mathbb{I}(\cdot)$ is the indicator function. And $sim(\cdot)$ denotes the similarity calculation function.

To learn the modality-invariant features and bridge the modality gap, we apply cross-modal contrastive learning to instances from different modalities that share the same pseudo labels. Finally, the loss of ICC module is defined as:

$$\mathcal{L}_{icc} = -\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{N}\log\left(p\left(j \mid \boldsymbol{x}_j^i\right)\right). \tag{8}$$

By minimizing Eq. (8), the model improves the ability to learn cross-modal representations, thereby further enhancing the quality of label disambiguation.

Prototype-wise Cross-Modal Alignment. With instance-wise cross-modal contrast, the model learns more discriminative representations at the instance level, narrowing the gap between different modalities. However, this approach does not reduce modality gap at the prototype level. Therefore, we additionally propose a prototype-wise cross-modal alignment approach.

First, we maintain a prototype embedding vector V for each category in the dataset. The prototype can also be treated as a sample feature that best represents the category $k \in \{1, 2, 3, ..., K\}$. Prototype vectors for image and text modalities can be written as follows:

$$\boldsymbol{V}^I = \left[\begin{array}{cccc} \boldsymbol{v}_0^I & \boldsymbol{v}_1^I & \cdots & \boldsymbol{v}_K^I \end{array}\right], \tag{9}$$

$$\boldsymbol{V}^T = \left[\begin{array}{cccc} \boldsymbol{v}_0^T & \boldsymbol{v}_1^T & \cdots & \boldsymbol{v}_K^T \end{array}\right], \tag{10}$$

where $\boldsymbol{V}^I$ and $\boldsymbol{V}^T$ represent the prototypes for image and text modalities respectively. $\boldsymbol{v}_c^I$ and $\boldsymbol{v}_c^T$ represent the prototype vectors for category c in the image and text modalities, respectively, and $c \in \{1, 2, 3, ..., K\}$.

To update the prototypes stably, we use the pseudo label $\hat{k}$ and normalized representations $\boldsymbol{q}$ to update the prototype $\boldsymbol{v}_k$ for class k in a moving-average style:

$$\boldsymbol{v}_k^I = \omega(t)\boldsymbol{v}_k^I + (1-\omega(t))\boldsymbol{q}^I, \quad \text{if } \hat{k}^I = k, \tag{11}$$

$$\boldsymbol{v}_k^T = \omega(t)\boldsymbol{v}_k^T + (1-\omega(t))\boldsymbol{q}^T, \quad \text{if } \hat{k}^T = k, \tag{12}$$

where $\omega(t)$ is the dynamic parameter that controls the balance between the current prototype vectors $\boldsymbol{v}$ and the representation vectors $\boldsymbol{q}$, and its value gradually decreases as the epoch t increases. Initially, the representation vectors $\boldsymbol{q}$ learned by the model tend to show significant deviations because the disambiguation effect is not yet pronounced. Therefore, the influence of these representation vectors $\boldsymbol{q}$ should be considered less. As the training progresses, the improved representation vectors $\boldsymbol{q}$ should increasingly influence the update of the prototype vectors $\boldsymbol{v}$.

To eliminate the cross-modal gap, we use the Mean Absolute Error (MAE) loss to minimize the similarity differences between the normalized representation $\boldsymbol{q}^I$ and their corresponding prototype $\boldsymbol{v}_k^I$ and $\boldsymbol{v}_k^T$. Similarly, we calculate the differences for $\boldsymbol{q}^T$ and their corresponding prototype $\boldsymbol{v}_k^I$ and $\boldsymbol{v}_k^T$. Therefore, the prototype-wise cross-modal alignment loss for the image and text modalities are as follows:

$$\mathcal{L}_{pca}^{I} = \sum_{\boldsymbol{q}^I\in\mathcal{Q}^I}\sum_{k\in\mathcal{Y}}\left|\boldsymbol{q}^I\cdot\boldsymbol{v}_k^I - \boldsymbol{q}^I\cdot\boldsymbol{v}_k^T\right|, \tag{13}$$

$$\mathcal{L}_{pca}^{T} = \sum_{\boldsymbol{q}^T\in\mathcal{Q}^T}\sum_{k\in\mathcal{Y}}\left|\boldsymbol{q}^T\cdot\boldsymbol{v}_k^T - \boldsymbol{q}^T\cdot\boldsymbol{v}_k^I\right|. \tag{14}$$

Finally, the total prototype-wise cross-modal alignment loss ($\mathcal{L}_{pca}$) is formulated as follows:

$$\mathcal{L}_{pca} = \mathcal{L}_{pca}^{I} + \mathcal{L}_{pca}^{T}. \tag{15}$$

Method	Ref.	Wikipedia								NUS-WIDE							
		Image → Text				Text → Image				Image → Text				Text → Image			
		0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
MCCA	SIGKDD2010	0.202	0.202	0.202	0.202	0.189	0.189	0.189	0.189	0.523	0.523	0.523	0.523	0.539	0.539	0.539	0.539
PLS	CVPR2011	0.337	0.337	0.337	0.337	0.320	0.320	0.320	0.320	0.498	0.498	0.498	0.498	0.517	0.517	0.517	0.517
DCCA	ICML2013	0.281	0.281	0.281	0.281	0.260	0.260	0.260	0.260	0.527	0.527	0.527	0.527	0.537	0.537	0.537	0.537
DCCAE	ICML2015	0.308	0.308	0.308	0.308	0.286	0.286	0.286	0.286	0.529	0.529	0.529	0.529	0.538	0.538	0.538	0.538
SCL	TMM2023	0.386	0.386	0.386	0.386	0.357	0.357	0.357	0.357	0.596	0.596	0.596	0.596	0.603	0.603	0.603	0.603
CC	NIPS2020	0.456	0.409	0.404	0.376	0.401	0.358	0.367	0.342	0.691	0.687	0.684	0.676	0.689	0.680	0.678	0.676
RC	NIPS2020	0.432	0.370	0.341	0.294	0.376	0.364	0.314	0.298	0.690	0.680	0.677	0.667	0.683	0.677	0.672	0.655
PRODEN	ICML2020	0.448	0.391	0.371	0.317	0.398	0.348	0.354	0.272	0.687	0.685	0.684	0.671	0.685	0.683	0.679	0.672
LWS	ICML2021	0.433	0.369	0.325	0.252	0.376	0.341	0.299	0.278	0.687	0.678	0.671	0.659	0.685	0.671	0.666	0.655
PaPi	CVPR2023	0.426	0.367	0.346	0.354	0.410	0.362	0.306	0.202	0.479	0.508	0.499	0.576	0.629	0.604	0.593	0.586
SCARCE	ICML2024	0.363	0.336	0.316	0.292	0.346	0.321	0.283	0.272	0.660	0.668	0.635	0.635	0.653	0.638	0.643	0.643
DiCA	Ours	**0.505**	**0.481**	**0.487**	**0.469**	**0.449**	**0.419**	**0.432**	**0.413**	**0.697**	**0.693**	**0.687**	**0.682**	**0.690**	**0.685**	**0.682**	**0.681**

Table 1: Performance comparison in terms of mAP scores under different partial rates of 0.1, 0.2, 0.3, and 0.4 on the Wikipedia and NUS-WIDE datasets. The highest mAP score is shown in bold.

Method	Ref.	INRIA-Websearch								XMediaNet							
		Image → Text				Text → Image				Image → Text				Text → Image			
		0.01	0.02	0.03	0.04	0.01	0.02	0.03	0.04	0.01	0.02	0.03	0.04	0.01	0.02	0.03	0.04
MCCA	SIGKDD2010	0.275	0.275	0.275	0.275	0.277	0.277	0.277	0.277	0.233	0.233	0.233	0.233	0.249	0.249	0.249	0.249
PLS	CVPR2011	0.387	0.387	0.387	0.387	0.398	0.398	0.398	0.398	0.276	0.276	0.276	0.276	0.266	0.266	0.266	0.266
DCCA	ICML2013	0.188	0.188	0.188	0.188	0.182	0.182	0.182	0.182	0.152	0.152	0.152	0.152	0.162	0.162	0.162	0.162
DCCAE	ICML2015	0.167	0.167	0.167	0.167	0.164	0.164	0.164	0.164	0.149	0.149	0.149	0.149	0.159	0.159	0.159	0.159
SCL	TMM2023	0.329	0.329	0.329	0.329	0.337	0.337	0.337	0.337	0.116	0.116	0.116	0.116	0.137	0.137	0.137	0.137
CC	NIPS2020	0.521	0.503	0.482	0.460	0.546	0.528	0.504	0.490	0.554	0.493	0.446	0.391	0.555	0.484	0.434	0.383
RC	NIPS2020	0.502	0.473	0.431	0.403	0.530	0.491	0.454	0.431	0.496	0.401	0.341	0.278	0.492	0.392	0.340	0.276
PRODEN	ICML2020	0.520	0.499	0.479	0.459	0.546	0.525	0.501	0.483	0.544	0.479	0.427	0.369	0.547	0.469	0.414	0.361
LWS	ICML2021	0.515	0.493	0.471	0.438	0.542	0.520	0.491	0.470	0.535	0.455	0.372	0.305	0.528	0.437	0.339	0.252
PaPi	CVPR2023	0.544	0.526	0.516	0.488	0.544	0.532	0.512	0.479	0.490	0.328	0.297	0.278	0.492	0.311	0.292	0.273
SCARCE	ICML2024	0.346	0.292	0.254	0.239	0.420	0.373	0.357	0.297	0.305	0.296	0.289	0.282	0.356	0.331	0.299	0.286
DiCA	Ours	**0.565**	**0.555**	**0.547**	**0.542**	**0.581**	**0.575**	**0.565**	**0.560**	**0.561**	**0.512**	**0.474**	**0.427**	**0.561**	**0.507**	**0.459**	**0.417**

Table 2: Performance comparison in terms of mAP scores under different partial rates of 0.01, 0.02, 0.03, and 0.04 on the INRIA-Websearch and XMediaNet datasets. The highest mAP score is shown in bold.

Experiments

Datasets

To evaluate the effectiveness of our method, we conduct extensive comparison experiments on four cross-modal retrieval benchmark datasets. These datasets are introduced as follows: 1) **Wikipedia** contains 2,866 image-text pairs that belong to 10 classes. We follow the previous work (Feng, Wang, and Li 2014) divide the dataset into 3 subsets: 2,173, 231, and 462 pairs for training, validation, and testing sets, respectively. 2) **INRIA-Websearch** consists of 71,478 images and 71,478 text descriptions. Following (Wei et al. 2017), We use the subset of INRIA-Websearch which selects 14,698 samples of 100 largest classes from the original set. We follow the previous work (Hu et al. 2021) divide the dataset into three subsets: 9,000, 1,332 and 4,366 image-text pairs for training, validation and testing sets, respectively. 3) **NUS-WIDE** consists of about 270,000 images that belong to 81 categories. Following the previous work (Peng et al. 2018), we use a subset of NUS-WIDE which has ten classes. Moreover, we split the dataset into three subsets, i.e., 42,941; 5,000; and 23,661 image-text pairs for training, validation, and testing sets, respectively. 4) **XMediaNet** is a large-scale multimodal dataset comprising 200 categories. We select data of the image and text modalities from this dataset then divide them into 32,000, 4,000, and 4,000 pairs for training, validation, and testing sets, respectively.

Implementation Detail

In this work, we adopt the Adam (Kingma and Ba 2014) optimizer with a learning rate 0.0001 to update the parameters. For all datasets, we set the maximum number of training epochs to 100. The training batch size is set to 32 for Wikipedia dataset, and to 512 for the other datasets. Furthermore, to maintain consistency, the batch size during validation and testing is uniformly set to 256. For the datasets of Wikipedia, NUS-WIDE, and XMediaNet, we utilize the pretrained VGG-19 (Karen 2014) model as the convolutional

neural network (CNN) backbone for processing images. Additionally, we employ the pre-trained Doc2Vec model (Lau and Baldwin 2016) as the textual backbone for handling text data. As for the INRIA-Websearch dataset, we adopt a pre-trained AlexNet (Krizhevsky, Sutskever, and Hinton 2012) of ImageNet and LDA to process image and text data, respectively. Our DiCA is implemented on the PyTorch framework, and all experiments are conducted on four Nvidia GeForce RTX 3090 GPUs.

Experimental Setup

To verify the effectiveness of our proposed method, we compare DiCA with six partial label learning methods and five unsupervised cross-modal retrieval methods. Specifically, we compare DiCA with the following partial label learning methods: **CC** and **RC** (Feng et al. 2020), **PRODEN** (Lv et al. 2020), **LWS** (Wen et al. 2021), **PaPi** (Xia et al. 2023), **SCARCE**[1] (Wang et al. 2024a) . For the unsupervised cross-modal retrieval methods, we chose **MCCA** (Rupnik and Shawe-Taylor 2010), **PLS** (Sharma and Jacobs 2011), **DCCA** (Andrew et al. 2013), **DCCAE** (Wang et al. 2015), **SCL** (Liu et al. 2023). Similar to the previous work (Lv et al. 2020; Wen et al. 2021; Wang et al. 2022), we generate partially labeled datasets by flipping negative labels to false positive labels with a certain probability. Considering the number of categories contained in different datasets, we have set different partial label rates for each dataset, i.e., $\{0.1, 0.2, 0.3, 0.4\}$ for **Wikipedia** and **NUS-WIDE**, and $\{0.01, 0.02, 0.03, 0.04\}$ for **INRIA-Websearch** and **XMediaNet**. Moreover, we employ mean average precision (mAP) on all retrieved results as our evaluation.

Comparison with State-of-the-Art Methods

We apply cross-modal retrieval with partial labels on four datasets to evaluate the performance of our DiCA and the baselines. The experimental results in terms of mAP scores across different partial rates are reported in Table 1 and Table 2 for four datasets, respectively. As shown in these tables, our DiCA outperforms all the baselines on all the datasets and various partial rates. From the experimental results, we can draw the following observations: 1) Some existing unsupervised cross-modal retrieval methods achieve relatively good performance when trained with a set of candidate labels. This is because they employ an unsupervised component that learns modality-specific transformations by maximizing correlations between different modalities. 2) Partial rates remarkably influence the performance of partial label learning methods. With the partial rate increasing in labels, their accuracies will typically decrease fast. In contrast, unsupervised methods do not face this issue. 3) In most cases, partial label learning methods achieve better retrieval results than unsupervised cross-modal retrieval methods because they utilize label information more effectively. This is due to their use of a label disambiguation component, similar to that in DiCA, which enhances the handling of candidate labels and highlights the critical role of label disambiguation. 4) As shown in Table 1 and Table 2, our DiCA outperforms other baselines on all datasets with different partial rates. For example, with a partial rate of 0.4, the proposed DiCA exceeds SCL and PaPi by 8.3% and 11.5% on Wikipedia dataset for the image-to-text retrieval, respectively. This is due to the fact that DiCA not only has a label disambiguation loss that considers both candidate and non-candidate labels, but also has two cross-modal disparity elimination components designed for cross-modal retrieval.

[1]SCARCE is trained with the complement of partial labels.

Ablation Study

In this section, we explore the contribution of each component (i.e., $\mathcal{L}_{nbd}$, $\mathcal{L}_{icc}$, $\mathcal{L}_{pca}$) for cross-modal retrieval with partial labels. To achieve this, we conduct three variants of the proposed DiCA: 1) DiCA with $\mathcal{L}_{nbd}$; 2) DiCA with $\mathcal{L}_{nbd}$ and $\mathcal{L}_{icc}$; 3) DiCA with $\mathcal{L}_{nbd}$ and $\mathcal{L}_{pca}$. All the compared methods are trained with the same settings on the Wikipedia and INRIA-Websearch datasets for a fair comparison. From the experimental results shown in Table 3 and Table 4, one can observe that $\mathcal{L}_{icc}$ can dramatically boost the performance, which indicates its effectiveness in excavating the instance-level discrimination. Meanwhile, the results also demonstrate that $\mathcal{L}_{pca}$ further reduces the modality gap. In conclusion, our DiCA can effectively enhance the performance on distinct datasets under different partial rates.

Method	Image → Text			
	0.1	0.2	0.3	0.4
DiCA (w. $\mathcal{L}_{nbd}$)	0.456	0.416	0.406	0.361
DiCA (w. $\mathcal{L}_{nbd}$ & $\mathcal{L}_{icc}$)	0.460	0.427	0.424	0.389
DiCA (w. $\mathcal{L}_{nbd}$ & $\mathcal{L}_{pca}$)	0.494	0.456	0.457	0.430
Full DiCA	**0.505**	**0.481**	**0.487**	**0.469**
	Text → Image			
DiCA (w. $\mathcal{L}_{nbd}$)	0.410	0.379	0.385	0.365
DiCA (w. $\mathcal{L}_{nbd}$ & $\mathcal{L}_{icc}$)	0.418	0.379	0.385	0.371
DiCA (w. $\mathcal{L}_{nbd}$ & $\mathcal{L}_{pca}$)	0.440	0.395	0.408	0.378
Full DiCA	**0.449**	**0.419**	**0.432**	**0.413**

Table 3: Comparison between our DiCA (full version) and its three variants under the various partial rates of 0.1, 0.2, 0.3, and 0.4 on the Wikipedia dataset. The highest performance is shown in bold.

Effect of Coefficient λ

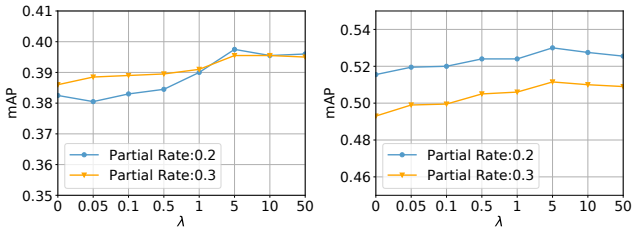
To analyse the impact of the coefficient λ in Eq. (3) and Eq. (4), we conduct parameter analysis experiments on Wikipedia and INRIA-Websearch datasets under different partial rates. As shown in Fig. 3, we plot the retrieval results (mAP) with different parameters of λ. Based on the results, we can observe that the non-candidate boosted disambiguation loss achieves the best performance when the trade-off parameter λ in the range of $[1, 10]$. This is accomplished by effectively balancing the loss on candidate labels or non-candidate labels.

Effect of Coefficient α and β

To evaluate the impact of the coefficient α and β in Eq. (1), we conduct parameter analysis experiments on Wikipedia

Method	Image → Text			
	0.01	0.02	0.03	0.04
DiCA (w. $\mathcal{L}_{nbd}$)	0.531	0.519	0.500	0.489
DiCA (w. $\mathcal{L}_{nbd}$ & $\mathcal{L}_{icc}$)	0.537	0.528	0.514	0.504
DiCA (w. $\mathcal{L}_{nbd}$ & $\mathcal{L}_{pca}$)	0.555	0.548	0.538	0.527
Full DiCA	**0.565**	**0.555**	**0.547**	**0.542**
	Text → Image			
DiCA (w. $\mathcal{L}_{nbd}$)	0.555	0.541	0.523	0.509
DiCA (w. $\mathcal{L}_{nbd}$ & $\mathcal{L}_{icc}$)	0.560	0.548	0.535	0.523
DiCA (w. $\mathcal{L}_{nbd}$ & $\mathcal{L}_{pca}$)	0.572	0.567	0.557	0.547
Full DiCA	**0.581**	**0.575**	**0.565**	**0.560**

Table 4: Comparison between our DiCA (full version) and its three variants under the various partial rates of 0.01, 0.02, 0.03, and 0.04 on the INRIA-Websearch dataset. The highest performance is shown in bold.



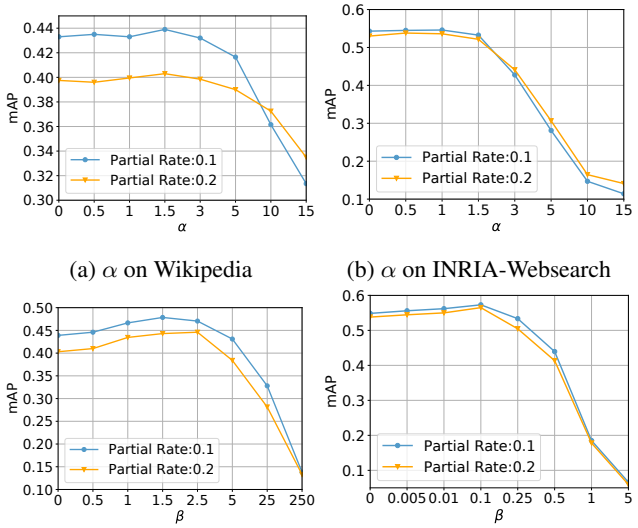
(a) λ on Wikipedia (b) λ on INRIA-Websearch

Figure 3: The performance of DiCA with only $\mathcal{L}_{nbd}$ in terms of mAP scores versus different values of λ on Wikipedia and INRIA-Websearch datasets.

and INRIA-Websearch datasets. As shown in Fig. 4, we can see that both instance-wise cross-modal contrast loss ($\mathcal{L}_{icc}$) and prototype-wise cross-modal alignment loss ($\mathcal{L}_{pca}$) contribute to enhance the model's representational learning capability, which is consist with our ablation study. However, the contributions of each component are distinct for different datasets, which may be caused by the difficulty level of the datasets (e.g., the more classes there are, the more difficult it will be). To be specific, the model can obtain stable performance when the value of α is in range $[1, 3]$ on both Wikipedia and INRIA-Websearch datasets. As for β, the proposed DiCA yields a stable performancee in the range $[1, 2.5]$ on Wikipedia, and in the range $[0.01, 0.25]$ on INRIA-Websearch, respectively.

Effect of Modality Gap Elimination

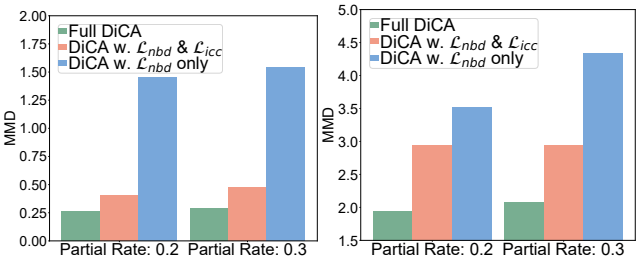
To visually investigate the impact of $\mathcal{L}_{icc}$ and $\mathcal{L}_{pca}$ on eliminating modality gap, we present the modality gap in terms of Maximum Mean Discrepancy (MMD) for our DiCA model, the DiCA model with $\mathcal{L}_{nbd}$ and $\mathcal{L}_{icc}$, and the DiCA model with $\mathcal{L}_{nbd}$ alone, under different partial rates. As shown in Fig. 5, both $\mathcal{L}_{icc}$ and $\mathcal{L}_{pca}$ could narrow the modality gap, demonstrating the effectiveness of the proposed IPRL mechanism.



(a) α on Wikipedia (b) α on INRIA-Websearch

(c) β on Wikipedia (d) β on INRIA-Websearch

Figure 4: The performance of DiCA in terms of mAP scores versus different values of α and β on Wikipedia and INRIA-Websearch datasets.



(a) Wikipedia (b) INRIA-Websearch

Figure 5: The MMD between image and text modalities on Wikipedia and INRIA-Websearch datasets under partial rates of 0.2 and 0.3.

Conclusion

In this paper, we study a new problem, i.e., cross-modal retrieval with partial labels (PLCMR). To this end, we propose a novel method named DiCA to learning discriminative representations from partial labels with ambiguity while eliminating modality gap. To better perform label disambiguation and narrow modality gap, we propose a non-candidate boosted disambiguation learning mechanism (NBDL) that thoughtfully balances the trade-off between the losses on candidate and non-candidate labels. Meanwhile, we introduce a novel instance-prototype representation learning mechanism (IPRL) to enhance the model by further eliminating the modality gap at both the instance and prototype levels. Comprehensive experiments are conducted comparing to several state-of-the-art approaches on four multimodal datasets, demonstrating the effectiveness of our DiCA.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62306197, 62372315), China Postdoctoral Science Foundation (2021TQ0223, 2022M712236), Sichuan Science and Technology Planning Project (2024YFG0007, 2024YFHZ0144, 2024YFHZ0089, 2024NSFTD0049, 2024ZDZX0004), Chengdu Science and Technology Project (2024-YF05-00687-SN, 2023-XT00-00004-GX, 2021-JB00-00025-GX), Postdoctoral Joint Training Program of Sichuan University (SCDXLHPY2307).

References

Andrew, G.; Arora, R.; Bilmes, J.; and Livescu, K. 2013. Deep canonical correlation analysis. In *International conference on machine learning*, 1247–1255. PMLR.

Cao, M.; Bai, Y.; Cao, Z.; Nie, L.; and Zhang, M. 2023. Efficient Image-Text Retrieval via Keyword-Guided Pre-Screening. *IEEE Transactions on Circuits and Systems for Video Technology*.

Cao, M.; Bai, Y.; Zeng, Z.; Ye, M.; and Zhang, M. 2024. An empirical study of clip for text-based person search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 465–473.

Cao, M.; Li, S.; Li, J.; Nie, L.; and Zhang, M. 2022. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*.

Caron, M.; Misra, I.; Mairal, J.; Goyal, P.; Bojanowski, P.; and Joulin, A. 2020. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in neural information processing systems*, 33: 9912–9924.

Cour, T.; Sapp, B.; and Taskar, B. 2011. Learning from partial labels. *The Journal of Machine Learning Research*, 12: 1501–1536.

Feng, F.; Wang, X.; and Li, R. 2014. Cross-modal Retrieval with Correspondence Autoencoder. In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, 7–16. Association for Computing Machinery.

Feng, L.; Lv, J.; Han, B.; Xu, M.; Niu, G.; Geng, X.; An, B.; and Sugiyama, M. 2020. Provably consistent partial-label learning. *Advances in neural information processing systems*, 33: 10948–10960.

Hardoon, D. R.; Szedmak, S.; and Shawe-Taylor, J. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12): 2639–2664.

Harold, H. 1936. Relations between two sets of variables. *Biometrika*, 28(3): 321–377.

Hu, P.; Peng, X.; Zhu, H.; Zhen, L.; and Lin, J. 2021. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5403–5413.

Hüllermeier, E.; and Beringer, J. 2006. Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5): 419–439.

Jin, R.; and Ghahramani, Z. 2002. Learning with multiple labels. *Advances in neural information processing systems*, 15.

Karen, S. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv: 1409.1556*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Lau, J. H.; and Baldwin, T. 2016. An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, 78–86.

Liu, H.; Ma, Y.; Yan, M.; Chen, Y.; Peng, D.; and Wang, X. 2024. DiDA: Disambiguated Domain Alignment for Cross-Domain Retrieval with Partial Labels. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 3612–3620.

Liu, L.; and Dietterich, T. 2012. A conditional multinomial mixture model for superset label learning. *Advances in neural information processing systems*, 25.

Liu, Y.; Wu, J.; Qu, L.; Gan, T.; Yin, J.; and Nie, L. 2023. Self-Supervised Correlation Learning for Cross-Modal Retrieval. *IEEE Transactions on Multimedia*, 25: 2851–2863.

Lv, J.; Xu, M.; Feng, L.; Niu, G.; Geng, X.; and Sugiyama, M. 2020. Progressive identification of true labels for partial-label learning. In *international conference on machine learning*, 6500–6510. PMLR.

Nguyen, N.; and Caruana, R. 2008. Classification with partial labels. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 551–559.

Peng, Y.; Qi, J.; Huang, X.; and Yuan, Y. 2018. CCL: Cross-modal Correlation Learning With Multigrained Fusion by Hierarchical Network. *IEEE Transactions on Multimedia*, 20(2): 405–420.

Rupnik, J.; and Shawe-Taylor, J. 2010. Multi-view canonical correlation analysis. In *Conference on data mining and data warehouses (SiKDD 2010)*, volume 473, 1–4.

Sharma, A.; and Jacobs, D. W. 2011. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *CVPR 2011*, 593–600. IEEE.

Si, C.; Jiang, Z.; Wang, X.; Wang, Y.; Yang, X.; and Shen, W. 2024. Partial Label Learning with a Partner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15029–15037.

Su, C.; Li, Z.; Lei, T.; Peng, D.; and Wang, X. 2023. MetaVG: A Meta-Learning Framework for Visual Grounding. *IEEE Signal Processing Letters*.

Sun, Y.; Dai, J.; Ren, Z.; Chen, Y.; Peng, D.; and Hu, P. 2024a. Dual Self-Paced Cross-Modal Hashing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 15184–15192.

Sun, Y.; Liu, K.; Li, Y.; Ren, Z.; Dai, J.; and Peng, D. 2024b. Distribution Consistency Guided Hashing for Cross-Modal Retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 5623–5632.

Sun, Y.; Ren, Z.; Hu, P.; Peng, D.; and Wang, X. 2023. Hierarchical consensus hashing for cross-modal retrieval. *IEEE Transactions on Multimedia*, 26: 824–836.

Wang, H.; Qiang, Y.; Chen, C.; Liu, W.; Hu, T.; Li, Z.; and Chen, G. 2021. Online partial label learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Part II*, 455–470. Springer.

Wang, H.; Xiao, R.; Li, Y.; Feng, L.; Niu, G.; Chen, G.; and Zhao, J. 2022. Pico: Contrastive label disambiguation for partial label learning. In *International conference on learning representations*.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015. On deep multi-view representation learning. In *International conference on machine learning*, 1083–1092. PMLR.

Wang, W.; Ishida, T.; Zhang, Y.-J.; Niu, G.; and Sugiyama, M. 2024a. Learning with Complementary Labels Revisited: The Selected-Completely-at-Random Setting Is More Practical. In *Forty-first International Conference on Machine Learning*.

Wang, X.; Hu, P.; Liu, P.; and Peng, D. 2020. Deep semisupervised class-and correlation-collapsed cross-view learning. *IEEE transactions on cybernetics*, 52(3): 1588–1601.

Wang, X.; Peng, D.; Hu, P.; Gong, Y.; and Chen, Y. 2023a. Cross-domain alignment for zero-shot sketch-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(11): 7024–7035.

Wang, X.; Peng, D.; Hu, P.; and Sang, Y. 2019. Adversarial correlated autoencoder for unsupervised multi-view representation learning. *Knowledge-Based Systems*, 168: 109–120.

Wang, X.; Peng, D.; Yan, M.; and Hu, P. 2023b. Correspondence-free domain alignment for unsupervised cross-domain image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 10200–10208.

Wang, Y.; and Peng, Y. 2021. MARS: Learning modality-agnostic representation for scalable cross-media retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7): 4765–4777.

Wang, Z.; Xu, X.; Wei, J.; Xie, N.; Yang, Y.; and Shen, H. T. 2024b. Semantics Disentangling for Cross-Modal Retrieval. *IEEE Transactions on Image Processing*, 33: 2226–2237.

Wei, Y.; Zhao, Y.; Lu, C.; Wei, S.; Liu, L.; Zhu, Z.; and Yan, S. 2017. Cross-Modal Retrieval With CNN Visual Features: A New Baseline. *IEEE Transactions on Cybernetics*, 47(2): 449–460.

Wen, H.; Cui, J.; Hang, H.; Liu, J.; Wang, Y.; and Lin, Z. 2021. Leveraged weighted loss for partial label learning. In *International conference on machine learning*, 11091–11100. PMLR.

Wu, Z.; Xiong, Y.; Yu, S. X.; and Lin, D. 2018. Unsupervised feature learning via non-parametric instance discrimination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3733–3742.

Xia, S.; Lv, J.; Xu, N.; Niu, G.; and Geng, X. 2023. Towards effective visual representations for partial-label learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15589–15598.

Zhang, M.-L.; and Yu, F. 2015. Solving the partial label learning problem: An instance-based approach. In *IJCAI*, 4048–4054.

Zhang, M.-L.; Zhou, B.-B.; and Liu, X.-Y. 2016. Partial label learning via feature-aware disambiguation. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1335–1344.

Zhen, L.; Hu, P.; Wang, X.; and Peng, D. 2019. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10394–10403.