Deep Evidential Learning with Noisy Correspondence for Cross-modal Retrieval

Yang Qin
qinyang.gm@gmail.com
Sichuan University
Chengdu, China

Dezhong Peng*
pengdz@scu.edu.cn
Sichuan University
Chengdu, China

Xi Peng
pengx.gm@gmail.com
Sichuan University
Chengdu, China

Xu Wang
wangxu.scu@gmail.com
Sichuan University
Chengdu, China

Peng Hu†
penghu.ml@gmail.com
Sichuan University
Chengdu, China

ABSTRACT

Cross-modal retrieval has been a compelling topic in the multimodal community. Recently, to mitigate the high cost of data collection, the co-occurred pairs (e.g., image and text) could be collected from the Internet as a large-scaled cross-modal dataset, e.g., Conceptual Captions. However, it will unavoidably introduce noise (i.e., mismatched pairs) into training data, dubbed noisy correspondence. Unquestionably, such noise will make supervision information unreliable/uncertain and remarkably degrade the performance. Besides, most existing methods focus training on hard negatives, which will amplify the unreliability of noise. To address the issues, we propose a generalized Deep Evidential Cross-modal Learning framework (DECL), which integrates a novel Cross-modal Evidential Learning paradigm (CEL) and a Robust Dynamic Hinge loss (RDH) with positive and negative learning. CEL could capture and learn the uncertainty brought by noise to improve the robustness and reliability of cross-modal retrieval. Specifically, the bidirectional evidence based on cross-modal similarity is first modeled and parameterized into the Dirichlet distribution, which not only provides accurate uncertainty estimation but also imparts resilience to perturbations against noisy correspondence. To address the amplification problem, RDH smoothly increases the hardness of negatives focused on, thus embracing higher robustness against high noise. Extensive experiments are conducted on three image-text benchmark datasets, i.e., Flickr30K, MS-COCO, and Conceptual Captions, to verify the effectiveness and efficiency of the proposed method. The code is available at https://github.com/QinYang79/DECL.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; *Retrieval models and ranking*; Similarity measures.

*Dezhong Peng is also with Chengdu Ruibei Yingte Information Technology Co., Ltd.
†Corresponding author.

MM '22, October 10–14, 2022, Lisboa, Portugal.

ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3547922

KEYWORDS

Cross-modal retrieval, Image-Text matching, Evidential learning, Noisy correspondence

ACM Reference Format:
Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. 2022. Deep Evidential Learning with Noisy Correspondence for Cross-modal Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547922

1 INTRODUCTION

Cross-modal retrieval aims to retrieve the relevant instances across different modalities, which is receiving more and more attention from academics and industrial sectors [6, 13, 14, 20, 45]. The core of cross-modal retrieval is to measure the similarity between different modalities and to retrieve the most relevant samples from other modalities according to the similarities.

To address the issue, most existing methods exploit different techniques to maximize the similarity of positive cross-modal pairs while minimizing that of the negatives, such as common representation learning [20, 37] and similarity learning[6, 22]. Although these approaches have achieved promising performance, most of them heavily rely on large-scaled well-annotated data [14]. It is expensive and even impossible to collect extremely clean large-scaled data. To alleviate the collection cost, one could collect the co-occurred image-text pairs from the Internet as a cheap large-scaled cross-modal dataset, e.g., Conceptual Captions [32]. However, without carefully manually annotating, it is inevitable to introduce noise (i.e., mismatched pairs) into collected data, a.k.a noisy correspondence, which will undoubtedly make cross-modal correspondence unreliable/uncertain and remarkably degrade the retrieval performance. Especially for the widely-used hinge-based triplet ranking loss with hard negatives [6, 7, 20, 22], the hard learning paradigm would be more sensitive to the noise, and the uncertainty could remarkably degrade the retrieval performance as shown in our experiments (e.g., Tables 1 and 2).

The most similar paradigm to noisy correspondence might be learning with noisy labels [3, 10, 26]. In the decades, numerous methods are proposed to alleviate the negative influence of noisy labels for classification tasks, e.g., co-teaching [10], robust loss functions [26], etc. Obviously, the noisy correspondence refers to the wrong alignments of cross-modal pairs, which is different

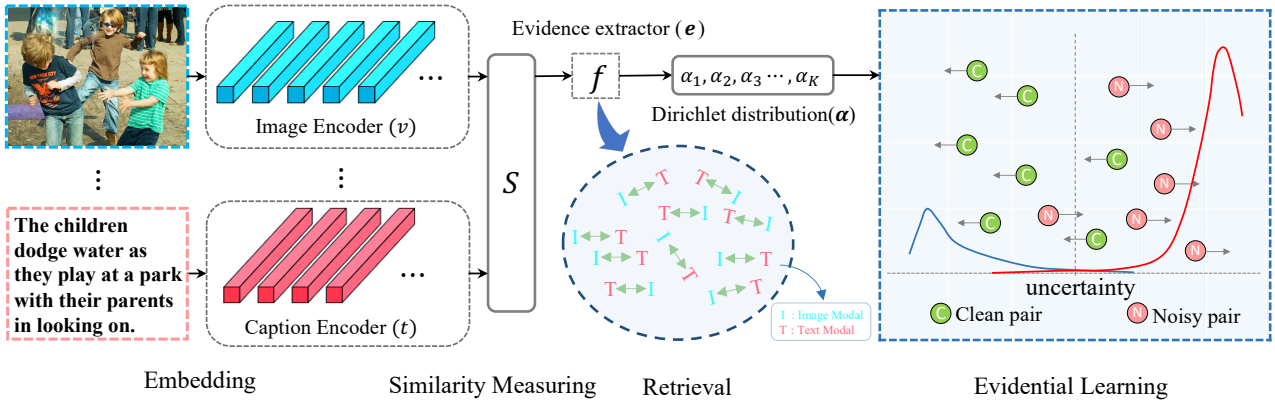

Figure 1: The overview of our Deep Evidential Cross-modal Learning framework (DECL). The images and texts are first encoded to feature representations, and then the similarities across different modalities are computed by similarity measure functions (e.g., cosine [20] and SGR [6]). Second, the evidence extractor f is exploited to collect the non-negative evidence parameterized into Dirichlet distribution α, which models the second-order query probabilities and the overall uncertainty. Finally, DECL integrates CEL and RDH to train the model on clean and noisy sets separated by evidence with positive and negative learning.

from noisy labels of classification tasks [10, 26], thus making these robust learning methods incapable of tacking noisy correspondence cases. To be specific, noisy correspondence focuses on the uncertain instance-level matching instead of the category-level unreliability. Thus, noisy correspondence will be more challenging than noisy labels because the number of instances is much larger than that of categories.

To tackle the challenge, we propose a Deep Evidential Cross-modal Learning framework (DECL) to robustly learn cross-modal similarity with noisy correspondence for cross-modal retrieval as shown in Figure 1. To be specific, our DECL exploits a novel Cross-modal Evidential Learning paradigm (CEL) to capture and leverage the uncertainty brought by noise to isolate the noisy pairs, thus addressing the unreliability problem. Furthermore, to address the hard negative problem, RDH smoothly increases the hardness of negatives during training to mitigate the adverse effects of unreliable hard negatives caused by noisy correspondence. With the integration of CEL and RDH, our DECL could accurately separate and employ the clean and noisy data for learning with noisy correspondence. Specifically, to capture the uncertainty caused by the noisy correspondence, CEL first models and parameterizes the bidirectional evidence into Dirichlet distribution according to the cross-modal similarity. Second, the learned evidence is exploited to dynamically separate the training data into a matched set with clean pairs and a mismatched set with noisy pairs. Finally, by integrating CEL and RDH losses, DECL trains the cross-modal model on the matched and mismatched set with positive and negative learning, respectively. Our main contributions can be summarized as follows:

- We propose a generalized Deep Evidential Cross-modal Learning framework (DECL) to provide trusted retrieval in an effective and efficient way. Our DECL could be directly applied to existing methods to robustly learn with noisy correspondence for cross-modal retrieval.
- A Cross-modal Evidential Learning paradigm (CEL) is presented to capture the uncertainty brought by noise to improve the robustness and reliability. In addition, the learned uncertainty could be exploited to self-estimate the retrieval reliability for some insightful evaluation. To the best of our knowledge, our CEL could be the first method that endows evidential learning with cross-modal retrieval.
- A Robust Dynamic Hinge loss (RDH) is proposed to mitigate the adverse effects of unreliability caused by noisy correspondence. Specifically, RDH smoothly increases the hardness of negatives during training to improve the robustness against noisy correspondence.
- Extensive experiments verify that the proposed method improves the robustness against noisy correspondence, especially the high noise rate. Moreover, we provide insightful analysis that the learned uncertainty could reduce the negative impacts of noisy correspondences, improving the robustness.

2 RELATED WORKS

In this section, we briefly review some most related works from three aspects, i.e., deep cross-modal retrieval, learning with noisy labels, and uncertainty-based learning.

2.1 Deep Cross-Modal Retrieval

Most existing cross-modal methods attempt to exploit the mutual information of cross-modal pairs (e.g., image and text) to learn a common space, wherein different modalities could be comparable by manual or learnable similarity metrics for retrieval. These methods could be roughly categorized into two groups, i.e., coarse-grained matching [7, 18, 37] and fine-grained matching [4, 20, 25]. Specifically, coarse-grained matching methods aim at excavating the global cross-modal correspondence by projecting the whole images and the full texts into a common discriminant space [7, 18, 37].

However, these approaches ignore the intrinsic local correspondence between the image objects and the words in cross-modal pairs. To address the issue, fine-grained matching methods attempt to establish the local connections between the objects of images and words of sentences to infer the latent fine-grained visual-semantic relationships. However, these methods implicitly assume that all positive cross-modal pairs are correctly aligned in the training data, which is impractical in practice due to the high cost of collection and annotation. To tackle the partially aligned data caused by noisy correspondence, Huang et al. [14] present a Noisy Correspondence Rectifier method (NCR) to learn with noisy correspondence. In brief, NCR separates the training data into clean and noisy pairs and then rectifies the cross-modal correspondence in a co-teaching manner. However, it is hard to accurately divide the noisy data only relying on traditional loss, especially for extremely high noise. Moreover, its "co-learning" training manner will exponentially increase the training overhead.

2.2 Learning with Noisy Labels

To tackle the ubiquitous unreliable annotations, numerous robust training methods are proposed to learn from noisy labels for classification tasks and have achieved promising results [33, 34, 36]. One of the typical techniques is the correction strategy, which could be classified into two groups: label correction [23, 35] and loss correction [10, 29]. Label correction methods aim to improve the training robustness by correcting the wrong labels [23, 35], but they require some additional clean data or expensive detection processes for estimating noise, which remarkably limits their practicability. Moreover, loss correction methods estimate the noise transition matrix to correct the loss for robust training [10, 29]. However, accurately estimating the transition matrix is challenging in practice [21]. Different from correction strategy, some works attempt to design adaptive training strategies to improve the training robustness against noisy labels. Specifically, these methods learn a weighting scheme to re-weight samples so that degrade the contributions of noisy samples to the training [15]. Recently, the memorization nature of deep neural networks (DNNs) [3] has been exploited to select clean samples from noisy training data for robust learning [2, 21, 42]. As aforementioned, the noisy cross-modal correspondence is more challenging than noise labels, and the above robust training methods cannot tackle it.

2.3 Uncertainty-based Learning

Although DNNs have achieved great success in various applications [43, 44], most deep models are deterministic predictions and cannot assess the uncertainty of their decisions like humans. To achieve that, some early works endow DNNs with uncertainty by using distributions instead of the deterministic weight parameters [8, 17, 28], but leading to huge extra computation overhead. Subsequently, some studies explore different methods to estimate the uncertainty of the model predictions [9, 19, 27]. Specifically, Sensoy et al. [31] employed the Subjective Logic (SL) theory [16] to explicitly model the uncertainty of DNNs by placing the Dirichlet distribution on the class probability, which greatly improves the endurance against adversarial perturbations. Unlike some multi-view methods[39, 40], Han et al. [11] endowed multi-view classification with uncertainty to dynamically integrate different views at an evidence level, which greatly improves the reliability and robustness of multi-view classification. For object detection tasks, Wang et al. [38] introduce uncertainty quantification into training to avoid overconfident predictions and achieve robust semi-supervised learning against noise. Different from these methods of estimating class uncertainty, we aim to model the uncertainty of cross-modal correspondence to achieve robust and reliable performance in cross-modal retrieval.

3 THE PROPOSED METHOD

3.1 Problem Formulation

Let $\mathcal{P} = \{P_i\}_{i=1}^{N}$ be the annotated cross-modal pairs for training, where $P_i = (I_i, T_i)$ is the ith visual I_i and textual T_i pair, and N is the data size. The cross-modal similarity of P_i could be measured through $S(v(I_i), t(T_i))$, where v and t are the modality-specific encoders that embed the visual and textual modalities into feature representations, respectively. Notably, the similarity function $S(\cdot,\cdot)$ could be a non-parametric [7, 20] or parametric [6]. To simplify the representation, we denote $S(v(I_i), t(T_i))$ as $S(I_i, T_i)$ in the following.

In practice, it is unavoidable to introduce noise into the training data, leading to mismatched pairs, a.k.a noisy correspondence, which makes supervision information unreliable/uncertain in cross-modal learning. To tackle the noisy correspondence, we propose a generalized Deep Evidential Cross-modal Learning framework (DECL). The overall objective function for a mini-batch with K pairs is shown as follows:

$$\mathcal{L}_{overall} = \frac{1}{K}\sum_{i=1}^{K}\mathcal{L}_e(I_i, T_i, l_i) + \lambda_1 \sum_{i=1}^{K} l_i \mathcal{L}_h(I_i, T_i), \tag{1}$$

where $\mathcal{L}_e$ is the loss function for bidirectional evidential learning, $\mathcal{L}_h$ is the proposed RDH loss function, λ_1 is a positive balance factor, and l_i is the predicted correspondence label according to its support evidence. l_i could be obtained as follows:

$$l_i = \begin{cases} 1, & \text{if } i = \arg\max(\boldsymbol{e}_i^{i2t} + \boldsymbol{e}_i^{t2i}), \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $\boldsymbol{e}_i^{i2t}$ and $\boldsymbol{e}_i^{t2i}$ are two K-dimensional evidence vectors inferred through the cross-modal similarities of a given visual I_i and textual T_i samples, respectively. The evidence could measure the amount of support collected from data in favor of a retrieved cross-modal sample to be correlated to the given query. Therefore, l_i could be exploited to determine whether DECL performs positive learning or negative learning on the pair (I_i, T_i) in Equation (1).

3.2 Uncertainty Modeling

In this section, we utilize the Dempster-Shafer Theory of Evidence to model the cross-modal uncertainty, which follows the principles of Subjective Logic [16]. For ease of representation, we only consider the computation in a given mini-batch with K pairs in the following. Given a pair (I_i, T_j), the evidence could be extracted by the evidence extractor f (See Figure 1), which is defined as

$$e_{ij} = f(S(I_i, T_j)) = \exp^{(\tanh(S(I_i, T_j)/\tau)}, \tag{3}$$

where $0 < \tau < 1$ is a scaling parameter. Thus, the evidence vector $\boldsymbol{e}_i^{i2t}$ of a given visual query I_i could be extracted from the corresponding cross-modal similarities through Equation (3), i.e., $\boldsymbol{e}_i^{i2t} = [e_{i1}, e_{i2}, \cdots, e_{iK}]$. Similarly, the evidence vector $\boldsymbol{e}_i^{t2i}$ of a given textual query T_i could be obtained by $\boldsymbol{e}_i^{t2i} = [e_{1i}, e_{2i}, \cdots, e_{Ki}]$.

Subjective Logic tries to assign a belief mass to each query and an overall uncertainty mass based on the collected cross-modal evidence (e.g., $\boldsymbol{e}_i^{i2t}$ and $\boldsymbol{e}_i^{t2i}$), which could be defined as

$$b_{ij} = \frac{e_{ij}}{L_i} = \frac{\alpha_{ij} - 1}{L_i} \quad \text{and} \quad u_i = \frac{K}{L_i}, \tag{4}$$

where $L_i = \sum_{j=1}^{K}(e_{ij} + 1)$ and $u + \sum_{j=1}^{K} b_{ij} = 1$. The L_i could be regarded as the Dirichlet distribution strength, and the belief mass assignment $\boldsymbol{b}_i = [b_{i1}, b_{i2}, \cdots, b_{iK}]$ could be viewed as subjective opinions corresponding to the Dirichlet distribution with parameters $\boldsymbol{\alpha}_i = [\alpha_{i1}, \alpha_{i2}, \cdots, \alpha_{iK}]$, where $\alpha_{ij} = e_{ij} + 1$.

Intuitively, cross-modal retrieval is similar to classifying instances (i.e., pairs), the query similarity corresponds to the probability alignment. The Dirichlet distribution parametrized over evidence represents the density of each such probability assignment. Thus, $\boldsymbol{\alpha}_i$ models second-order probabilities and uncertainty [16]. The density function is defined as

$$D(\boldsymbol{p}_i|\boldsymbol{\alpha}_i) = \begin{cases} \frac{1}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} & \text{for } \boldsymbol{p}_i \in \mathcal{S}_K \\ 0 & \text{otherwise} \end{cases}, \tag{5}$$

where $\boldsymbol{p}_i \in \mathbb{R}^K$ are the query probabilities, $B(\boldsymbol{\alpha}_i)$ is the K-dimensional multinomial beta function, and $\mathcal{S}_K$ is the K-dimensional unit simplex [11].

3.3 Cross-modal Evidential Learning

Inspired by *Contrast Learning* [5, 12], we can intuitively regard cross-modal retrieval as K-way classification that a query is classified to its positive cross-modal counterpart. Given a query I_i or T_i, its retrieval ground-truth $\boldsymbol{y}_i$ could be defined as a K-dimensional vector whose ith element is l_i and the rest are 0. The least-squares loss (LS) is exploited to make the query probabilities $\boldsymbol{p}_i$ approach the ground-truth $\boldsymbol{y}_i$. Notably, the density of $\boldsymbol{p}_i$ follows the parameterized Dirichlet distribution $\boldsymbol{\alpha}_i$. Thus, the loss could be formulated as follows:

$$\begin{aligned} \mathcal{L}_m(\boldsymbol{\alpha}_i, \boldsymbol{y}_i) &= \int \left\|\boldsymbol{y}_i - \boldsymbol{p}_i\right\|_2^2 \frac{1}{B\left(\boldsymbol{\alpha}_i\right)} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} d\boldsymbol{p}_i \\ &= \sum_{j=1}^{K} \left[\left(y_{ij} - \mathbb{E}(p_{ij})\right)^2 + \operatorname{Var}\left(p_{ij}\right) \right] \\ &= \sum_{j=1}^{K} \left(y_{ij} - \frac{\alpha_{ij}}{L_i} \right)^2 + \frac{\alpha_{ij}\left(L_i - \alpha_{ij}\right)}{L_i^2\left(L_i + 1\right)} \end{aligned} \tag{6}$$

where $\mathbb{E}(p_{ij})$ and $\operatorname{Var}\left(p_{ij}\right)$ are the expected value and the variance of p_{ij}, respectively. The expected probability $\mathbb{E}(p_{ij})$ could be estimated by $\frac{\alpha_{ij}}{L_i}$ [31]. More derivation details could be found in the supplementary material[1]. From Equation (6), one could find that minimizing $\mathcal{L}_m$ will make the expected probabilities $\mathbb{E}(\boldsymbol{p}_i)$ approach the ground-truth $\boldsymbol{y}_i$, while reducing the variance $\operatorname{Var}(\boldsymbol{p}_i)$.

[1] https://github.com/QinYang79/DECL/blob/main/supplementary_material.pdf

Although Equation (6) could ensure that evidence for positive pairs is higher than evidence for negative pairs, it is unable to guarantee that negative pairs generate zero evidence. Thus, we enforce the total evidence of negative samples to shrink to zero. To this end, a Kullback-Leibler (KL) divergence term is introduced to penalize the divergences from the uncertain retrieval, which is formulated as

$$\begin{aligned} &\mathcal{L}_{kl}(\boldsymbol{\alpha}_i, \boldsymbol{y}_i) \\ &= KL\left[D\left(\boldsymbol{p}_i \mid \tilde{\boldsymbol{\alpha}}_i\right) \| D\left(\boldsymbol{p}_i \mid \mathbf{1}\right)\right] \\ &= \log\left(\frac{\Gamma\left(\sum_{j=1}^{K} \tilde{\alpha}_{ij}\right)}{\Gamma(K) \prod_{j=1}^{K} \Gamma\left(\tilde{\alpha}_{ij}\right)}\right) + \sum_{j=1}^{K}\left(\tilde{\alpha}_{ij} - 1\right)\left[\psi\left(\tilde{\alpha}_{ij}\right) - \psi\left(\sum_{j=1}^{K} \tilde{\alpha}_{ij}\right)\right], \end{aligned} \tag{7}$$

where $\tilde{\boldsymbol{\alpha}}_i = \boldsymbol{y}_i + (1 - \boldsymbol{y}_i) \odot \boldsymbol{\alpha}_i$ is the Dirichlet parameters after removing the unreliable evidence from the predicted Dirichlet distribution, and $\Gamma(\cdot)$ and $\psi(\cdot)$ are the gamma and digamma functions, respectively.

For image-to-text retrieval (i.e., retrieving texts with an image query), one could obtain a query evidence vector $\boldsymbol{e}_i^{i2t} \in \mathbb{R}^K$ for a given query I_i by Equation (3). Then, the corresponding Dirichlet distribution $\boldsymbol{\alpha}_i^{i2t} = \boldsymbol{e}_i^{i2t} + 1$. The image query evidential loss is

$$\mathcal{L}_e^{i2t}(I_i, l_i) = \mathcal{L}_m(\boldsymbol{\alpha}_i^{i2t}, l_i) + \lambda_2 \mathcal{L}_{kl}(\boldsymbol{\alpha}_i^{i2t}, l_i), \tag{8}$$

where $0 < \lambda_2 < 1$ is a balance factor. Similarly, the evidential loss $\mathcal{L}_e^{t2i}(T_i, l_i)$ for text-to-image retrieval could also be obtained like the above equation. Therefore, the bidirectional evidential loss could be formulated as

$$\mathcal{L}_e(I_i, T_i, l_i) = \mathcal{L}_e^{i2t}(I_i, l_i) + \mathcal{L}_e^{t2i}(T_i, l_i). \tag{9}$$

3.4 Robust Dynamic Hinge Loss

The early cross-modal methods mainly employ a hinge-based triplet ranking loss (HTR) with a margin for Image-Text matching, which makes the model pay attention to all negatives in a mini-batch. To improve the performance of HTR, Faghri et al. [7] incorporated hard negative pairs in HTR, dubbed Max of Hinge loss (MH), which only focuses on the hardest negatives instead of all ones. However, we observe that MH performs very poorly in the presence of noisy correspondence (see Tables 1, 2 and 4), probably because the hardest negatives will over-amplify the wrong gradients caused by the noisy correspondence. To mitigate the overamplification influence of mismatched pairs, we present a Robust Dynamic Hinge loss (RDH) to dynamically increase the hardness of negatives, which could be defined as

$$\mathcal{L}_h(I, T) = \frac{1}{n}\left(\sum_{j=1}^{n}[\gamma - S(I, T) + S(I, \hat{T}_j)]_+ + \sum_{j=1}^{n}[\gamma - S(I, T) + S(\hat{I}_j, T)]_+\right), \tag{10}$$

where $[x]_+ \equiv \max(x, 0)$ and n is the number of selected hardest negatives, i.e., $\{\hat{I}_j\}_{j=1}^{n}$ and $\{\hat{T}_j\}_{j=1}^{n}$. Note that n will dynamically decrease during training by

$$n = \max(\lceil K - \eta * Step \rceil, \mu) \tag{11}$$

where K is the size of mini-batch, η is the annealing coefficient, $Step$ is the current training step, $\lceil x \rceil$ is the rounding down operation, and μ is the lower bound of n.

4 EXPERIMENTS

In this section, we conduct extensive experiments to verify the effectiveness and efficiency of our DECL. In the experiments, we use three widely-used datasets, i.e., Flickr30K [41], MS-COCO [24], and Conceptual Captions [32]. We artificially simulate the noisy correspondence to comprehensively evaluate the robustness of compared methods against different degrees of noise on two well-annotated datasets, i.e., Flickr30K and MS-COCO. Besides, to evaluate the performance under real noisy correspondence, we conduct comparison experiments on Conceptual Captions collected from the wild.

4.1 Datasets and Performance Measurements

To comprehensively evaluate the performance of our method, we use three widely-used image-text datasets in the experiments. A brief introduction of these datasets is given as follows:

- **Flickr30K [41]** has 31,000 images collected from the Flickr website and each image corresponds to five captions. We follow the data partition of [20]: 1,000 images for validation, 1,000 images for testing, and the rest for training.
- **MS-COCO [24]** consists of 123,287 images with five captions each. Following [20], 5,000 images are selected for validation, 5,000 images for testing, and the rest for training.
- **Conceptual Captions [32]** is a large-scale image-text dataset with 3.3M images collected from the Internet and each image has a single caption. This dataset consists of 3% $\sim$ 20% pairs with noisy correspondence [32]. Following [14], we use a subset of Conceptual Captions, i.e., CC152K, to conduct experiments. Specifically, 150,000 images for training, 1,000 images for validation, and 1,000 images for testing.

We need to perform preprocessing for all original images to obtain feature vectors of regions [1]. More specifically, we employ the Faster-RCNN [30] detector to extract 36 region proposals from each original image and encode each proposal into a 2048-dimensional feature vector. For each caption, following [6], the word embedding size is 300, and the dimensionality of joint embedding is 1024. We follow [14, 20] to compute Recall at K (R@K) on all retrieval results as the measurement of performance. In our experiments, we report R@1, R@5, R@10, and their sum to evaluate the performance of our DECL.

4.2 Implementation Details

Our DECL could be directly applied to almost all cross-modal retrieval methods to improve their robustness against noisy correspondence. Without losing generality, we apply our DECL to SGR, SAF, and SGRAF [6] to comprehensively verify the effectiveness of our framework. Specifically, we directly perform our DECL on the similarity output of these models without any changes to their models. We train our framework using the same settings as [6], except for our specific parameters, whose settings are given in the supplementary material[1].

4.3 Comparison with State-of-the-Arts

In this section, we conduct comparison experiments in terms of cross-modal retrieval on three datasets to evaluate the performance of our DECL. The baselines are SCAN (i-t AVG) [20], VSRN [22], IMRAM (Text) [4], SGRAF [6], and NCR [14] respectively, wherein NCR is the robust learning method against noisy correspondence. To verify the robustness of our framework, we inject noisy correspondence in Flickr30K and MS-COCO datasets by randomly shuffling images for a specific percentage (i.e., 20%, 40% 60%, and 80%), which is more strict than [14]. Due to the space limitation, we move the comparison results and the discussion of 0% noise to the supplementary material[1]. Unlike them, CC152K naturally contains 3% $\sim$ 20% pairs with noisy correspondence. For this, no artificial noise needs to be injected.

4.3.1 Comparisons on Flickr30. Table 1 reports the experimental results on the 1K test images of Flickr30K. The experimental results show that DECL-SGRAF not only achieves the best overall performance under low noise, but also maintains superior robustness under extreme high noise (80%). Thanks to our DECL, one could greatly improve the ability of baselines (e.g., SGR and SAF) to resist noise. Specifically, DECL improves R@1 of SAF by **10.6%**, **62.7%**, **56.5%**, and **46.9%** for retrieving texts under different noise rates, respectively. For SGR, DECL improves R@1 by **18.6%**, **64.9%**, **63.0%**, and **44.2%** for retrieving images under different noise rates, respectively. Thus, our DECL could greatly improve the robustness of the baselines against noisy correspondence. Moreover, our DECL-SGRAF could outperform the best baseline NCR in terms of sum for retrieving texts under different noise rates by exceeding **3.6**, **12.5**, **266.0**, and **373.4**, respectively.

4.3.2 Comparisons on MS-COCO. Tables 1 and 2 show the bidirectional retrieval results on the MS-COCO dataset. Following [20], we report the average results over 5 folds of 1K test images for a comprehensive comparison as shown in Table 1. From the experimental results, one could see that our DECL could improve the robustness of baselines against noisy correspondence, e.g., SAF and SGR. Similarly, NCR could also improve the performance of SGR under noisy correspondence. However, we found that NCR will fail to handle the noisy correspondence with high noise rates, such as higher than 50%. One reason might be that it is hard for NCR to divide the noisy and clean pairs only relying on the different distribution of losses under high noise rates because the dominance of noise makes the distribution indistinguishable. On the contrary, our DECL shows strong robustness against noise and outperforms all baselines. This might benefit from the uncertainty learning, which could alleviate the impact of noise, thus leading to robustness. Besides, we also evaluate the retrieval performance of compared methods for 5K test images without additional noise. The experimental results are reported in Table 2, wherein some results are brought from the corresponding papers. From the results, one could see that our DECL-SGRAF achieves competitive results, with the best results for retrieving image.

4.3.3 Comparisons on CC152K. In addition to the manual noise, we also evaluate the proposed method under the real noisy correspondence of CC152K. The experimental results are reported in Table 2. From the results, one could observe that our DECL could achieve competitive performance under real noise. Specifically, almost all results of DECL-SGRAF are superior to that of all baselines for cross-modal retrieval, except for R@1 of retrieving text. DECL-SGRAF is **1.6%** and **1.7%** higher than the best baseline NCR in terms

		Flickr30K							MS-COCO 1K						
		Image → Text			Text → Image				Image → Text			Text → Image			
Noise	Methods	R@1	R@5	R@10	R@1	R@5	R@10	Sum	R@1	R@5	R@10	R@1	R@5	R@10	Sum
20%	SCAN [20]	58.5	81.0	90.8	35.5	65.0	75.2	406.0	62.2	90.0	96.1	46.2	80.8	89.2	464.5
	VSRN [22]	33.4	59.5	71.3	25.0	47.6	58.6	295.4	61.8	87.3	92.9	50.0	80.3	88.3	460.6
	IMRAM [4]	22.7	54.0	67.8	16.6	41.8	54.1	257.0	69.9	93.6	97.4	55.9	84.4	89.6	490.8
	SAF [6]	62.8	88.7	93.9	49.7	73.6	78.0	446.7	71.5	94.0	97.5	57.8	86.4	91.9	499.1
	SGR [6]	55.9	81.5	88.9	40.2	66.8	75.3	408.6	25.7	58.8	75.1	23.5	58.9	75.1	317.1
	NCR [14]	73.5	93.2	96.6	**56.9**	**82.4**	**88.5**	491.1	76.6	95.6	98.2	60.8	88.8	95.0	515.0
	DECL-SAF	73.4	92.0	96.4	53.6	79.7	86.4	481.5	74.4	95.3	98.2	59.8	88.3	94.8	510.8
	DECL-SGR	74.5	92.9	**97.1**	53.6	79.5	86.8	484.4	75.6	95.1	98.3	59.9	88.3	94.7	511.9
	DECL-SGRAF	**77.5**	**93.8**	97.0	56.1	81.8	**88.5**	**494.7**	**77.5**	**95.9**	**98.4**	**61.7**	**89.3**	**95.4**	**518.2**
40%	SCAN [20]	26.0	57.4	71.8	17.8	40.5	51.4	264.9	42.9	74.6	85.1	24.2	52.6	63.8	343.2
	VSRN [22]	2.6	10.3	14.8	3.0	9.3	15.0	55.0	29.8	62.1	76.6	17.1	46.1	60.3	292.0
	IMRAM [4]	5.3	25.4	37.6	5.0	13.5	19.6	106.4	51.8	82.4	90.9	38.4	70.3	78.9	412.7
	SAF [6]	7.4	19.6	26.7	4.4	12.2	17.0	87.3	13.5	43.8	48.2	16.0	39.0	50.8	211.3
	SGR [6]	4.1	16.6	24.1	4.1	13.2	19.7	81.8	1.3	3.7	6.3	0.5	2.5	4.1	18.4
	NCR [14]	68.1	89.6	94.8	51.4	78.4	84.8	467.1	74.7	94.6	98.0	**59.6**	88.1	94.7	509.7
	DECL-SAF	70.1	90.6	94.4	49.7	76.6	84.1	465.5	73.3	94.6	98.1	57.9	87.2	94.1	505.2
	DECL-SGR	69.0	90.2	94.8	50.7	76.3	84.1	465.1	73.6	94.6	97.9	57.8	86.9	93.9	504.7
	DECL-SGRAF	**72.7**	**92.3**	**95.4**	**53.4**	**79.4**	**86.4**	**479.6**	**75.6**	**95.5**	**98.3**	59.5	**88.3**	**94.8**	**512.0**
60%	SCAN [20]	13.6	36.5	50.3	4.8	13.6	19.8	138.6	29.9	60.9	74.8	0.9	2.4	4.1	173.0
	VSRN [22]	0.8	2.5	5.3	1.2	4.2	6.9	20.9	11.6	34.0	47.5	4.6	16.4	25.9	140.0
	IMRAM [4]	1.5	8.9	17.4	1.9	5.0	7.8	42.5	18.2	51.6	68.0	17.9	43.6	54.6	253.9
	SAF [6]	0.1	1.5	2.8	0.4	1.2	2.3	8.3	0.1	0.5	0.7	0.8	3.5	6.3	11.9
	SGR [6]	1.5	6.6	9.6	0.3	2.3	4.2	24.5	0.1	0.6	1.0	0.1	0.5	1.1	3.4
	NCR [14]	13.9	37.7	50.5	11.0	30.1	41.4	184.6	0.1	0.3	0.4	0.1	0.5	1.0	2.4
	DECL-SAF	56.6	82.5	89.7	40.4	66.6	76.6	412.4	68.6	92.9	97.4	54.1	84.9	92.7	490.6
	DECL-SGR	64.5	85.8	92.6	44.0	71.6	80.6	439.1	69.7	93.4	97.5	54.5	85.2	92.6	492.9
	DECL-SGRAF	**65.2**	**88.4**	**94.0**	**46.8**	**74.0**	**82.2**	**450.6**	**73.0**	**94.2**	**97.9**	**57.0**	**86.6**	**93.8**	**502.5**
80%	SCAN [20]	1.1	5.0	8.7	0.4	1.3	2.3	18.8	10.2	29.9	42.0	0.1	0.7	1.1	84.0
	VSRN [22]	0.3	1.4	2.1	0.6	2.0	3.3	9.7	1.4	5.3	8.8	0.7	2.8	5.4	24.4
	IMRAM [4]	0.1	1.2	2.7	0.3	1.0	1.5	6.8	1.3	5.0	8.3	0.2	0.6	1.3	16.7
	SAF [6]	0.0	0.8	1.2	0.1	0.5	1.1	3.7	0.2	0.8	1.4	0.1	0.5	1.0	4.0
	SGR [6]	0.2	0.3	0.5	0.1	0.6	1.0	2.7	0.2	0.6	1.0	0.1	0.5	1.0	3.4
	NCR [14]	1.5	6.2	9.9	0.3	1.0	2.1	21.0	0.1	0.3	0.4	0.1	0.5	1.0	2.4
	DECL-SAF	46.9	73.7	83.0	32.1	59.0	69.4	364.1	59.3	87.9	94.8	46.3	79.1	88.9	456.3
	DECL-SGR	44.4	72.6	82.0	33.9	59.5	69.0	361.4	60.0	88.7	94.5	45.9	78.8	88.3	456.2
	DECL-SGRAF	**53.4**	**78.8**	**86.9**	**37.6**	**63.8**	**73.9**	**394.4**	**64.8**	**90.5**	**96.0**	**49.7**	**81.7**	**90.3**	**473.0**

Table 1: Performance comparison with different mismatching rates (MRate) on Flickr30K and MS-COCO 1K.

	CC152K							MS-COCO 5K						
	Image → Text			Text → Image				Image → Text			Text → Image			
Methods	R@1	R@5	R@10	R@1	R@5	R@10	Sum	R@1	R@5	R@10	R@1	R@5	R@10	Sum
SCAN [20]	30.5	55.3	65.3	26.9	53.0	64.7	295.7	44.7	75.9	86.6	33.3	63.5	75.4	379.4
VSRN [22]	32.6	61.3	70.5	32.5	59.4	70.4	326.7	53.0	81.1	89.4	40.5	70.6	81.1	415.7
IMRAM [4]	33.1	57.6	68.1	29.0	56.8	67.4	312.0	53.7	83.2	91.0	39.7	69.1	79.8	416.5
SAF [6]	31.7	59.3	68.2	31.9	59.0	67.9	318.0	53.3	-	90.1	39.8	-	80.2	-
SGR [6]	11.3	29.7	39.6	13.1	30.1	41.6	165.4	56.9	-	90.5	40.2	-	79.8	-
SGRAF [6]	-	-	-	-	-	-	-	57.8	-	**91.6**	**41.9**	-	81.3	-
NCR [14]	**39.5**	64.5	73.5	40.3	64.6	73.2	355.6	58.2	84.2	91.5	41.7	**71.0**	**81.3**	427.9
DECL-SAF	36.6	63.0	73.3	38.5	63.2	73.5	348.1	57.1	83.1	90.8	39.8	68.9	79.4	419.1
DECL-SGR	36.2	63.6	73.2	37.1	63.6	73.7	347.4	56.7	83.4	90.5	40.2	68.9	79.9	419.6
DECL-SGRAF	39.0	**66.1**	**75.5**	**40.7**	**66.3**	**76.7**	**364.3**	**59.2**	**84.5**	91.5	41.7	70.6	81.1	**428.6**

Table 2: Performance comparison on CC152K and MS-COCO 5K.

of R@5 in text and image retrieval, respectively. Moreover, the sum score of DECL-SGRAF outperforms all the baselines, indicating that our DECL could improve the overall robustness against noisy correspondence. Besides, for the exiting cross-modal methods (e.g., SAF and SGR), our DECL could greatly improve the robustness of their original models against real noise.

4.4 Ablation Study

In this section, we conduct ablation studies under 40% noise on Flickr30K to investigate the contributions of different proposed components in our DECL, i.e., $\mathcal{L}_h$, $\mathcal{L}_m$, and $\mathcal{L}_{kl}$. For a fair comparison, each component is evaluated on the same training and testing settings. The experimental results are reported in Table 3. From the results, one could observe that the full version of our DECL could achieve the best overall performance of cross-modal retrieval. Thus, each component contributes to the robustness of cross-modal retrieval against noisy correspondence. Compared with the results of the standard SGR in Table 1, one could find that both the proposed CEL and RDH could remarkably improve the robustness against noisy correspondence, which demonstrates the effectiveness of our CEL and RDH.

	Evidential loss		Image → Text			Text → Image		
$\mathcal{L}_h$	$\mathcal{L}_m$	$\mathcal{L}_{kl}$	R@1	R@5	R@10	R@1	R@5	R@10
✓	✓	✓	**69.0**	**90.2**	**94.8**	**50.7**	**76.3**	**84.1**
	✓	✓	68.9	89.4	94.1	48.8	75.1	83.3
✓	✓		66.7	89.2	93.6	49.0	75.8	83.6
✓			67.5	89.1	94.1	48.9	**76.3**	83.7
	✓		68.0	89.5	94.1	48.7	74.5	82.1

Table 3: Ablation studies for DECL-SGR on Flickr30K with 40% noise.

4.5 Robustness Analysis

In this section, we conduct some comparison experiments with different hinge-based losses (i.e., HTR, MH, and RDH) to quantitatively investigate the robustness of our RDH against noisy correspondence as shown in Table 4. From the experimental results, one could find that the original SGR with MH performs poorly under noisy correspondence, probably because that MH pays too much attention to the hardest negative samples to capture the correct correspondence in clean data. That is to say, the hardest negative samples of noisy queries usually produce large wrong gradients because it's harder to optimize irregular noisy correspondence. Thanks to focusing on all negative samples, SGR′ with HTR could improve the robustness of SGR against noisy correspondence, which indicates that the bottom easy negatives contain more correct information contributing to robust training than the top hard negatives in the retrieved ranking list. However, this indiscriminative usage of all negatives will prevent the model from focusing on the key points. By dynamically increasing the hardness of negatives, our RDH could remarkably improve the robustness of HTR against different levels of noise.

4.6 Visualizing Uncertainty

In this section, we carry out experiments under 20% noise on Flickr30K to visually investigate the evolution of uncertainty in the training process. As shown in Figure 3, one could observe the polarizing evolution: the uncertainty of clean pairs gradually becomes smaller and smaller, while that of noisy ones increases higher and higher, which verifies the effectiveness of uncertainty estimation for noisy correspondence. Therefore, the uncertainty could naturally be exploited to classify the noisy and clean pairs, boosting the robustness against noisy correspondence.

Noise	Methods	Image → Text			Text → Image			Sum
		R@1	R@5	R@10	R@1	R@5	R@10	
20%	SGR	30.9	59.8	67.4	26.7	50.7	61.0	296.5
	SGR′	62.9	87.1	93.5	45.8	73.0	81.5	443.8
	SGR*	**68.2**	**90.2**	**95.5**	**50.8**	**77.0**	**84.4**	**466.1**
40%	SGR	13.1	33.7	45.4	15.1	34.5	45.3	187.1
	SGR′	51.6	78.5	87.0	34.3	62.5	72.5	386.4
	SGR*	**63.7**	**85.6**	**91.9**	**46.5**	**72.1**	**77.2**	**437.0**
60%	SGR	0.3	1.4	3.1	0.2	1.0	1.8	7.8
	SGR′	34.7	62.4	72.8	23.0	46.4	57.1	296.4
	SGR*	**42.6**	**70.7**	**81.6**	**29.7**	**57.3**	**67.9**	**349.8**
80%	SGR	0.0	0.2	0.6	0.1	0.6	1.0	2.5
	SGR′	8.1	21.7	30.9	5.4	16.3	24.0	106.4
	SGR*	**25.1**	**52.3**	**64.4**	**16.9**	**40.2**	**52.9**	**251.8**

Table 4: Performance comparison of SGR with different loss functions on Flickr30K under different noise rates. SGR [6] is the original version with Max of Hinge loss (MH) focused on the hardest negatives. SGR′ trains SGR with the Hinge-based Triplet Ranking (HTR) loss focused on all negatives. SGR* trains SGR with our Robust Dynamic Hinge (RDH) loss focused on the top-n (smoothly attenuating) hard negatives.

4.7 Qualitative Result

To illustrate the advantages of our DECL, some retrieved examples for cross-modal retrieval are presented in Figure 2. Different from the existing methods, our DECL could not only retrieve samples across different modalities but also estimate the uncertainty of the retrieved results as shown in Figure 2. From the qualitative examples, one could see that the overall uncertainty negatively reflects the global retrieval performance. Specifically, the amount and ranking of correctly retrieved results will impact the score of uncertainty. The more and higher up the correct retrieved results are, the lower the uncertainty is, and vice versa, e.g., Figure 2(a,d) and Figure 2(c,f). However, the deterministic similarity is not enough to reflect the confidence of cross-modal retrieval. That is to say, high similarity might infer incorrect retrieval, but the uncertainty could provide correct evaluations, e.g., incorrect retrieval with high similarity and high uncertainty in Figure 2(c). Hence, by quantifying the retrieval uncertainty, our DECL could greatly increase the interpretability of the cross-modal model.

4.8 Efficiency Analysis

To verify the efficiency of our DECL, we conduct an efficiency analysis by comparing the best baseline (NCR) and our DECL. Specifically, we record the graphics memory (NVIDIA Tesla V100S-32 GB), the memory, and the average training time per epoch on the Flickr30K dataset with 20% noise as shown in Table 5. From the experimental results, one could see that our DECL achieves a remarkable improvement in training time by **40.25%**.

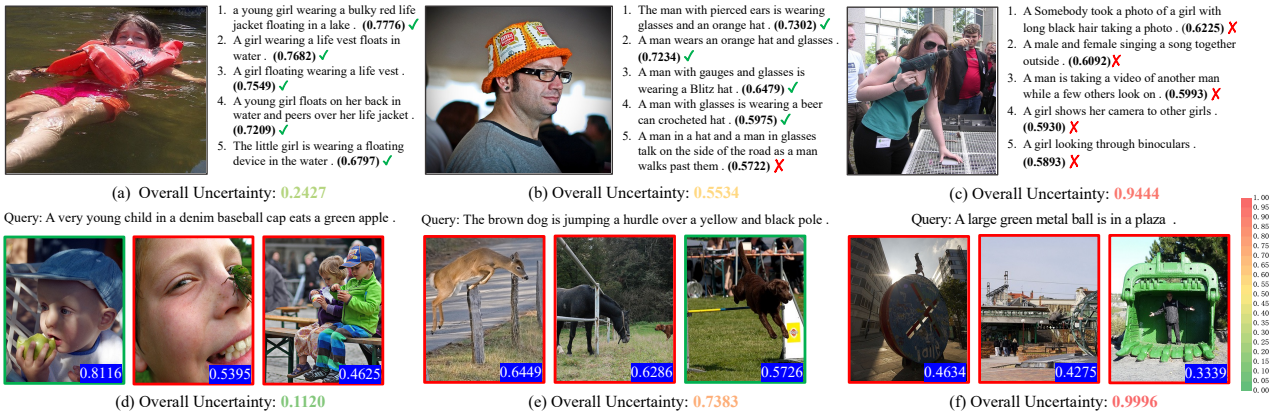

Figure 2: Some retrieved examples of cross-modal retrieval on Flickr30K under 40% noise. For each image query, we show the top-5 ranked sentences, i.e., (a)-(c). The correctly matched texts are marked with a green tick, otherwise the red cross. For each sentence query, we show the top-3 ranked images from left to right, i.e., (d)-(f). We outline the correctly matched images in green boxes and incorrectly matched ones in red boxes. Estimated uncertainty and similarity (i.e., bold font with bracket in sentences, and white font with blue background in images) are given in sub-captions and exemplars, respectively.

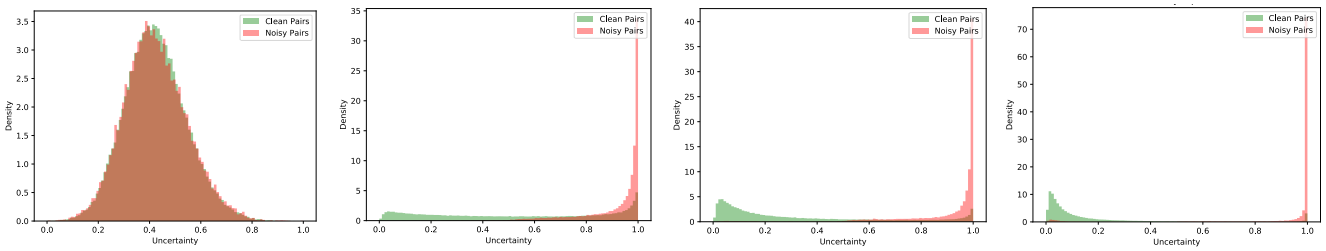

Figure 3: We visualize the uncertainty distribution of clean and noisy pairs at different training stages of our DECL, which is conducted on Flickr30K under 20% noise. Thanks to our DECL, the uncertainty of clean pairs gradually approaches the left (low) and the uncertainty of noisy pairs tightly gathers to the right (high).

Methods	G-memory (*MB*)	Memory (*GB*)	Time (*S*)
NCR	12,619 *MB*	13.60 *GB*	5155.95 *S*
DECL-SAF	11,661 *MB*	13.61 *GB*	1343.56 *S*
DECL-SGR	12,607 *MB*	13.62 *GB*	1736.97 *S*
DECL-SGRAF	12,607 *MB*	13.62 *GB*	3080.53 *S*

Table 5: Training efficiency comparison in terms of graphics memory, memory, and time cost. The reported per-epoch time is the average time for 40 epochs.

5 CONCLUSION

This paper studies a challenging paradigm of noisy labels, i.e., noisy correspondence, which will introduce mismatched pairs into the training data leading to performance degradation. To address this problem, we present a generalized Deep Evidential Cross-modal Learning framework (DECL) to capture the uncertainty of noise with the CEL and be immune to the noisy perturbation using the proposed RDH, thus embracing the robustness against noisy correspondence. Furthermore, extensive experiments are carried out to verify the effectiveness of our DECL in mitigating noisy correspondence.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grants No. U19A2078, U21B2040, 62176171, and 62102274), Sichuan Science and Technology Planning Project (Grants No. 2022YFQ0014, 2022YFH0021, 2022YFQ0014, and 2021YFS0389), Chengdu Science and Technology Project under grant 2021-JB00-00025-GX, Scu&Zigong Cooperation Project under Grant 2021CDZG-5, China Postdoctoral Science Foundation (No. 2021M692270), and Open Research Projects of Zhejiang Lab under Grant 2021KH0AB02.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 6077–6086.

[2] Eric Arazo, Diego Ortego, Paul Albert, Noel O'Connor, and Kevin McGuinness. 2019. Unsupervised label noise modeling and loss correction. In *International conference on machine learning*. PMLR, 312–321.

[3] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. 2017. A closer look at memorization in deep networks. In *International conference on machine learning*. PMLR, 233–242.

[4] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12655–12663.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[6] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. *arXiv preprint arXiv:2101.01368* (2021).

[7] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).

[8] Yarin Gal and Zoubin Ghahramani. 2015. Bayesian convolutional neural networks with Bernoulli approximate variational inference. *arXiv preprint arXiv:1506.02158* (2015).

[9] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*. PMLR, 1050–1059.

[10] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems* 31 (2018).

[11] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2021. Trusted multi-view classification. *arXiv preprint arXiv:2102.02051* (2021).

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9729–9738.

[13] Peng Hu, Hongyuan Zhu, Jie Lin, Dezhong Peng, Yin-Ping Zhao, and Xi Peng. 2022. Unsupervised Contrastive Cross-modal Hashing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).

[14] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. 2021. Learning with Noisy Correspondence for Cross-modal Matching. *Advances in Neural Information Processing Systems* 34 (2021).

[15] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. 2018. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International Conference on Machine Learning*. PMLR, 2304–2313.

[16] Audun Jsang. 2016. Subjective Logic: A formalism for reasoning under uncertainty. *Springer Verlag* (2016).

[17] Durk P Kingma, Tim Salimans, and Max Welling. 2015. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems* 28 (2015).

[18] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* (2014).

[19] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems* 30 (2017).

[20] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–216.

[21] Junnan Li, Richard Socher, and Steven CH Hoi. 2020. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394* (2020).

[22] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International conference on computer vision*. 4654–4662.

[23] Yuncheng Li, Jianchao Yang, Yale Song, Liangliang Cao, Jiebo Luo, and Li-Jia Li. 2017. Learning from noisy labels with distillation. In *Proceedings of the IEEE International Conference on Computer Vision*. 1910–1918.

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

[25] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. 2020. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10921–10930.

[26] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey. 2020. Normalized loss functions for deep learning with noisy labels. In *International Conference on Machine Learning*. PMLR, 6543–6553.

[27] Andrey Malinin and Mark Gales. 2018. Predictive uncertainty estimation via prior networks. *Advances in neural information processing systems* 31 (2018).

[28] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. 2017. Variational dropout sparsifies deep neural networks. In *International Conference on Machine Learning*. PMLR, 2498–2507.

[29] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. 2013. Learning with noisy labels. *Advances in neural information processing systems* 26 (2013).

[30] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39 (2015), 1137–1149.

[31] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems* 31 (2018).

[32] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

[33] Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[34] Daiki Tanaka, Daiki Ikami, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Joint optimization framework for learning with noisy labels. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5552–5560.

[35] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. 2017. Learning from noisy large-scale datasets with minimal supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 839–847.

[36] Hao Wang, Bing Liu, Chaozhuo Li, Yan Yang, and Tianrui Li. 2019. Learning with noisy labels for sentence-level sentiment classification. *arXiv preprint arXiv:1909.00124* (2019).

[37] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5005–5013.

[38] Zhenyu Wang, Ya-Li Li, Ye Guo, and Shengjin Wang. 2021. Combating Noise: Semi-supervised Learning by Region Uncertainty Quantification. *Advances in Neural Information Processing Systems* 34 (2021).

[39] Jie Wen, Yong Xu, and Hong Liu. 2018. Incomplete multiview spectral clustering with adaptive graph learning. *IEEE transactions on cybernetics* 50, 4 (2018), 1418–1429.

[40] Jie Wen, Zheng Zhang, Zhao Zhang, Lunke Fei, and Meng Wang. 2020. Generalized incomplete multiview clustering with flexible locality structure diffusion. *IEEE transactions on cybernetics* 51, 1 (2020), 101–114.

[41] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[42] Xingrui Yu, Bo Han, Jiangchao Yao, Gang Niu, Ivor Tsang, and Masashi Sugiyama. 2019. How does disagreement help generalization against label corruption?. In *International Conference on Machine Learning*. PMLR, 7164–7173.

[43] Ji Zhang, Jingkuan Song, Lianli Gao, Ye Liu, and Heng Tao Shen. 2022. Progressive Meta-learning with Curriculum. *IEEE Transactions on Circuits and Systems for Video Technology* (2022).

[44] Ji Zhang, Jingkuan Song, Yazhou Yao, and Lianli Gao. 2021. Curriculum-based meta-learning. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1838–1846.

[45] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng. 2019. Deep supervised cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10394–10403.