UGNCL: Uncertainty-Guided Noisy Correspondence Learning for Efficient Cross-Modal Matching

Quanxing Zha
Dep. of CS, Huaqiao University &
Xiamen CVPR Key Lab. & Fujian Key
Lab. of Big Data Intell. and Security
Xiamen, China
qxzha@stu.hqu.edu.cn

Xin Liu*
Dep. of CS, Huaqiao University &
Xiamen CVPR Key Lab., Department
of CS, Hong Kong Baptist University
Xiamen, China
xliu@hqu.edu.cn

Yiu-ming Cheung*
Department of Computer Science,
Inst. for Research and Continuing
Edu., Hong Kong Baptist University
Hong Kong SAR, China
ymc@comp.hkbu.edu.hk

Xing Xu
CSE, University of Electronic Science
and Technology of China
Chengdu, China
xing.xu@uestc.edu.cn

Nannan Wang
State Key Lab. of Integrated Services
Networks, Xidian University
Xi'an, China
nnwang@xidian.edu.cn

Jianjia Cao
Artificial Intelligence Research
Institute, Xiamen Wangsu Co., Ltd.
Xiamen, China
caojj@wangsu.com

ABSTRACT

Cross-modal matching has recently gained significant popularity to facilitate retrieval across multi-modal data, and existing works are highly relied on an implicit assumption that the training data pairs are perfectly aligned. However, such an ideal assumption is extremely impossible due to the inevitably mismatched data pairs, a.k.a. noisy correspondence, which can wrongly enforce the mismatched data to be similar and thus induces the performance degradation. Although some recent methods have attempted to address this problem, they still face two challenging issues: 1) unreliable data division for training inefficiency and 2) unstable prediction for matching failure. To address these problems, we propose an efficient *Uncertainty-Guided Noisy Correspondence Learning (UGNCL)* framework to achieve noise-robust cross-modal matching. Specifically, a novel *Uncertainty Guided Division (UGD)* algorithm is reliably designed leverage the potential benefits of derived uncertainty to divide the data into clean, noisy and hard partitions, which can effortlessly mitigate the impact of easily-determined noisy pairs. Meanwhile, an efficient *Trusted Robust Loss (TRL)* is explicitly designed to recast the soft margins, calibrated by confident yet error soft correspondence labels, for the data pairs in the hard partition through the uncertainty, leading to increase/decrease the importance of matched/mismatched pairs and further alleviate the impact of noisy pairs for robustness improvement. Extensive experiments conducted on three public datasets highlight the superiorities of the proposed framework, and show its competitive performance compared with the state-of-the-arts. The code is available at https://github.com/qxzha/UGNCL.

CCS CONCEPTS

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

KEYWORDS

Cross-Modal Matching, Noisy Correspondence Learning, Uncertainty Guided Division, Trusted Robust Loss

ACM Reference Format:
Quanxing Zha, Xin Liu, Yiu-ming Cheung, Xing Xu, Nannan Wang, and Jianjia Cao. 2024. UGNCL: Uncertainty-Guided Noisy Correspondence Learning for Efficient Cross-Modal Matching . In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '24), July 14–18, 2024, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3626772.3657806

1 INTRODUCTION

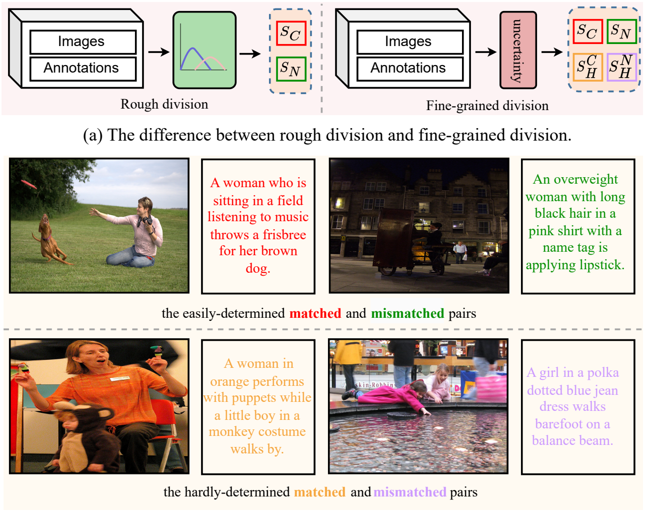
Cross-modal matching, favored for its effectiveness and efficiency, has attracted significant research interest in the field of multi-modal learning. In the literature, existing cross-modal matching methods [2, 8, 14] generally try to embed specific-modality features into a comparable common space, wherein the similarity of matched pairs is maximized and the mismatched ones is minimized. Although these approaches have shown promising progress, their performances are largely relied on massive high-quality aligned data [17]. However, unlike the easily-determined categorical labels of images, the textual descriptions of images are inherently subjective, which makes the annotation or collection of such data pairs very expensive and time-consuming.

In fact, the co-occurring data pairs collected from the public Internet are usually harvested as a cost-effective large-scaled cross-modal dataset. Undoubtedly, it is inevitable to introduce noise (i.e., mismatched data pairs) into collected data, a.k.a. noisy correspondence, which will remarkably degrade the performance of the existing methods [7, 22]. Unlike traditional cross-modal matching [42, 43], cross-modal matching learning with noisy correspondence aims to identify noisy training data pairs and mitigate their adverse impact, thereby enhancing model's generalization capabilities. The

*Xin Liu and Yiu-ming Cheung are the corresponding authors.

SIGIR '24, July 14–18, 2024, Washington, DC, USA

ACM ISBN 979-8-4007-0431-4/24/07
https://doi.org/10.1145/3626772.3657806


Images
Annotations
Rough division
S_C
S_N
Images
Annotations
uncertainty
Fine-grained division
S_C
S_N
S_H^C
S_H^N
A woman who is sitting in a field listening to music throws a frisbree for her brown dog.
An overweight woman with long black hair in a pink shirt with a name tag is applying lipstick.
the easily-determined matched and mismatched pairs
A woman in orange performs with puppets while a little boy in a monkey costume walks by.
A girl in a polka dotted blue jean dress walks barefoot on a balance beam.
the hardly-determined matched and mismatched pairs

(a) The difference between rough division and fine-grained division.

(b) Representative examples divided from the fine-grained division.

Figure 1: Illustration of uncertainty guided division (UGD). (a) clarifies the difference between rough division and the proposed fine-grained division. (b) showcases some representative examples of pairs from these partitions.

key challenge in cross-modal matching learning with noisy correspondence is *how to accurately partition the training data into clean and noisy parts and then estimate precisely soft correspondence labels for those noisy data pairs.*

To date, only a few efforts have been made to address the noisy correspondence problem. NCR [17] is the pioneer work to investigate this problem, which aims to train the model from noisy image-text pairs elastically. Specifically, NCR employs a two-component Guassian Mixture Model (GMM), fitted by per-sample loss, to divide the training data into clean and noisy partitions and then recasts the soft correspondence labels of the noisy partition using similarities predicted by an adaptive model. Latter, MSCN [12] and BiCro [39] replace the GMM using a two-component Beta Mixture Model (BMM) [28] to better accommodate the skewed distribution characteristic of clean data. MSCN leverages a meta network, trained on purely clean data, to produce reliable similarity score, while BiCro employs a bi-directional cross-modal similarity consistency assumption to estimate soft correspondence labels, working solely with noisy pairs. In contrast to these distribution-based methods, some studies [15, 33] start with adopting a robust loss function to prevent over-amplification for incorrect supervision information. However, there still are two limitations among these approaches: 1. ***Distribution Bottleneck:*** the predicted per-sample loss value or similarity score usually deviates from standard GMM or BMM, which inevitably introduce division errors, i.e., truly mismatched data pairs may be mistakenly distinguished as matched ones and vice versa. 2. ***Unstable Prediction:*** the soft correspondence labels, rectified by confident yet wrong predictions, are utilized to recast the soft margins for subsequent training, typically produces overfit on noise and further leads to match failure.

In this paper, we propose a novel ***Uncertainty-Guided Noisy Correspondence Learning framework for Efficient Cross-Modal Matching***, named ***UGNCL*** (see Fig. 2), to address the above challenges. Specifically, we first aggregate the extracted image features into distinct views. Subsequently, an evidence extractor is employed to learn bidirectional cross-modal evidence between the different views. Intuitively, cross-modal matching could be viewed as a K-way classification task, where a query is classified based on its corresponding match in another modality. Therefore, we naturally view the class probability of a query classified to its cross-modal counterpart as their similarity. Then a Dirichlet distribution, parameterized by the bidirectional cross-modal evidence from different views, is utilized to model class probabilities and overall uncertainty. Moreover, a *Uncertainty Guided Division (UGD)* is proposed, leveraging the class probabilities and uncertainty, to provide a fine-grained division (see Fig. 1). Meanwhile, we further propose a *Trusted Robust Loss (TRL)* to provide a reliable recasting for the soft margins, calibrated by the confidence yet error soft correspondence labels, of the pairs in the hard set through the uncertainty to enhance the robustness against noisy correspondence. Specifically, TRL smoothly decreases the values of soft margins for matched data pairs with high uncertainty, and for mismatched pairs with low uncertainty, while simultaneously increasing the values for matched pairs with low uncertainty and mismatched pairs with high uncertainty. The main contributions of this work are summarized as follows:

- We investigate a widely-existing yet challenging problem, i.e., noisy correspondence. To address this problem, we propose an effective and novel framework, termed *Uncertainty-Guided Noisy Correspondence Learning (UGNCL)*, to robustly learn with noise and ensure accurate cross-modal matching.
- We design a novel *Uncertainty Guided Division (UGD)* algorithm to divide the training data into clean, noisy and hard partitions. Thanks to this explicit division, our UGNCL could effortlessly mitigate the impact of easily-determined noise, as well as reliably and accurately recast the soft correspondence labels of ambiguous pairs in the hard partition.
- A *Trusted Robust Loss (TRL)* is presented to reliably recast the soft margins calibrated by the confident yet error soft correspondence labels, which could further improve the robustness against noise. Specifically, TRL smoothly increases the contribution of matched data pairs while decreasing the one of mismatched data pairs in the hard partition.
- Extensive experiments on three widely-used benchmarks, i.e., Flickr 30K [41], MS-COCO [24] and Conceptual Captions [38], verify the effectiveness and robustness of the proposed UGNCL model to against noisy correspondence.

2 RELATED WORK

2.1 Cross-Modal Matching

Cross-modal matching aims to bridge the inter-modality gap and establish semantic correspondence between different modalities [16]. Existing works could be broadly divided into two categories: *embedding-base* and *score-based* matching methods. **Embedding-based**, it typically employs modal-specific feature extractors to learn the global representation of a specific modality, and subsequently aligns different modalities in a common comparable embedding space [40]. Visual Semantic Embedding (VSE) [9] represents a widely-used solution for cross-modal matching, which separately maps each modality into a shared space using modality-specific network and then optimizing via a ranking loss. To further enhancing the matching performance, VSE++ [8] exploited a hard negatives mining strategy, integrating it within the ranking loss to improve the discriminate embedding of each specific modality. VSRN [22] and similar studies [4, 25] incorporate Graph Convolutional Networks (GCNs) to learn global representations that account for intra-modality relationships; VSE∞ [2] generates global representations by aggregating local features using a learnable pooling operation, achieving excellent performance. **Score-based**, it focuses on measuring fine-grained similarities at fragment level and then aggregates these local similarities to determine final overall instance-level similarity [10, 31, 45]. SCAN [20] aligns image regions with sentence words by associating visual semantics locally, and then integrates the semantic similarities between relevant region-word pairs to measure the overall image-text relevance. SGRAF [7] utilizes Graph Convolutional Network to infer local similarity nodes, and then aggregates them into a global node to determine the final similarity. NAAF [44] delves into identifying irrelevant fragments, thereby effectively discerning subtle mismatches across modalities to enhance the accuracy of image-text matching. To mitigate the adverse impact of redundant alignments, CHAN [30] leverages hard assignment code to mine informative region-word pairs and filters out mismatched alignments. While these alignment methods have achieved remarkable results, however, they typically rely on an implicit assumption: all cross-modal pairs are perfectly aligned in training data. In fact, it is an unrealistic expectation due to the high costs of collection and annotation. Therefore, exploring how cross-modal matching learns robustly with noisy correspondence is significant.

2.2 Learning with noisy Correspondence

Noisy Correspondence refers to instances where the data pairs are semantically mismatched but incorrectly considered as matched ones, which is essentially different from the noise label which occurs when a class label is annotated with the wrong category. To the best of our knowledge, Noisy Correspondence Rectify (NCR) [17] is the first attempt to investigate this problem, which is designed for robust training with noisy image-text pairs. Inspired by the prior success in tackling noisy labels [21], NCR divides the training data into clean and noisy partitions based on a two-component Guassian Mixture Model (GMM) [32] fitted by per-sample loss and then rectifies the soft correspondence labels for the noisy partition with similarity predicted by an adaptive model. MSCN [12] trains a meta network using exclusively clean data, which facilitates the provision of reliable soft correspondence labels estimations that are then utilized to rectify the predictions produced by main net. BiCro [39] rectifies the binary noisy correspondence labels into more accurate soft correspondence labels based on a assumption, i.e., similar images should have similar textual descriptions and vice versa. Afterwards, the predicted soft correspondence labels are recast as soft margins of triplet ranking loss to reduce the impact of noisy pairs. Moreover, DECL [33] exploits a robust loss function to enhance the robustness against noisy correspondence. In contrast, our UGNCL employ uncertainty derived from a Dirichlet distribution to divide the training data into three partitions, which helps explicitly mitigate the adverse impact of easily-determined noise. Furthermore, the derived uncertainty could provide trusted soft margins rectification for the triplet loss, thereby facilitating robust noisy correspondence learning to performance improvements.

2.3 Uncertainty-based Learning

Despite the remarkable success of DNNs across a broad domains [23], most are intrinsically deterministic predictions and the provision of uncertainty measure related to their decisions is absent. Some early studies [6, 29] utilized BNNs to introduce uncertainty into deep models by substituting distribution for the deterministic weight parameters. Owing to the limitations of BNNs, which typically suffer with inefficient performance inference and are burdened by considerable computational expenses, a more scalable and practical alternative, MC-dropout [11], was proposed. Subsequently, the Evidential Deep Learning (EDL) approach has been increasingly applied in computer vision and multi-modal task [3, 13] as an alternative to indirectly modeling uncertainty through network weights. Specifically, Sensoy et al. [37] introduces subjective logic theory to directly estimate uncertainty and thus significantly enhances resilience against adversarial perturbations. Han et al. [13] innovatively endowed multi-view classification with uncertainty to dynamically integrate diverse viewpoints at evidence level, which considerably bolsters the reliability and robustness of multi-view classification. Chen et al. [3] fist employ the EDL to quantify the uncertainty at both video- and snippet-level caused by background noise and facilitates the achievement of robust weakly-supervised temporal action localization. Different from these methods, our work leverages the uncertainty derived from modeled Dirichlet distribution to reliably and accurately divide training data into clean, noisy, and hard partitions, which effortlessly mitigates the adverse impact of easily-determined noise for performance improvement and further achieves noise-robust cross-modal matching.

3 METHODOLOGY

3.1 Preliminaries and Problem Statement

Without losing generality, we first use image-text matching task as an example to investigate noisy correspondence problem in cross-modal matching. Given a cross-modal matching training set $U = \{(\mathcal{D}_i, y_i)\}_{i=1}^{N}$, where $\mathcal{D}_i = (I_i, T_i)$ is the $i\text{-}th$ visual I_i and textual T_i pair, N indicates the size of entire training set and $y_i \in \{0, 1\}$ represents the label of $i\text{-}th$ image-text pair. The binary label y_i, represents a hard-labeled correspondence score, indicating whether the pair belongs to same instance ($y_i = 1$) or not ($y_i = 0$). The core of cross-modal matching lies in projecting the different modalities (visual and textual, in this case) into a shared, comparable space wherein matched pairs are identified by either closer feature distances or higher feature similarities and vice verse. Typically, the similarity of given image-text pairs could be measured through $S(\Phi(I), \Psi(T))$, where Φ and Ψ are modal-specific encoders that embed the visual and textual modalities into feature representation, respectively. Notably, the similarity measure function $S(\cdot, \cdot)$ could be a non-parametric [8, 9, 20] or parametric [7]. For simplicity, we denote $S(\Phi(I), \Psi(T))$ as $S(I, T)$ in the following.

Most existing methods have achieved promising performance by optimizing the hinge-based triplet ranking loss with online hard

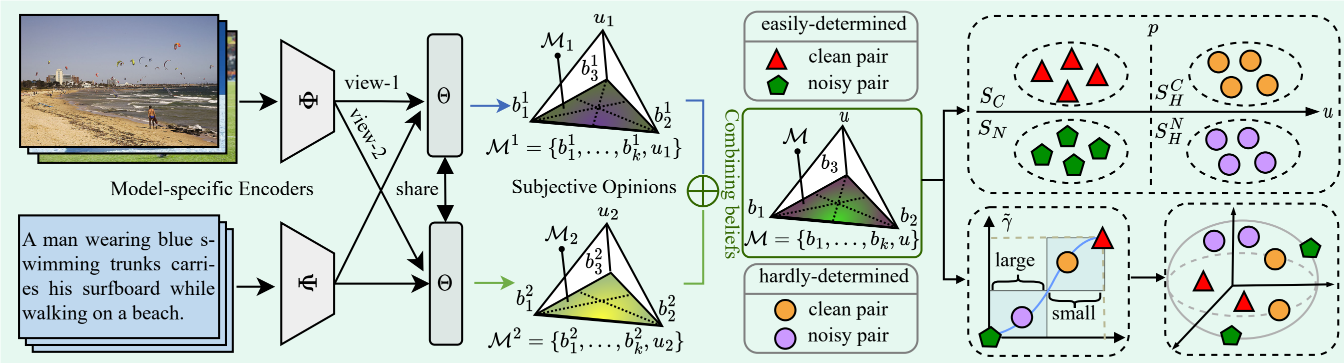


Figure 2: Overview of our Uncertainty-Guided Noisy Correspondence Learning framework (UGNCL).

negative mining proposed by VSE++ [8], the specific matching objective is defined as follows:

$$\mathcal{L}_H(I_i, T_i) = [\gamma - S(I_i, T_i) + S(\hat{I}_h, T_i)]_+ + [\gamma - S(I_i, T_i) + S(I_i, \hat{T}_h)]_+ \quad (1)$$

where $\gamma > 0$ is a scalar hyper-parameter that denotes a specific margin. (I_i, T_i) represents a matched data pair in the training dataset $\mathcal{D}$ and $[x]_+ = max(0, x)$. Additionally, we define $\hat{I}_h = argmax_{I_j \neq I_i} S(I_j, T_i)$ and $\hat{T}_h = argmax_{T_j \neq T_i} S(I_i, T_j)$ as the hardest negative samples within a mini-batch.

However, the solution of Eq. (1) relies on an implicit assumption, i.e., all cross-modal pairs in the training dataset are perfectly aligned. In practice, the cross-modal datasets are often sourced from Internet due to the high collection and annotation costs. Consequently, it is inevitable to introduce noise (i.e., mismatched data pairs) into collected data, a.k.a. noisy correspondence. Obviously, optimizing with Eq. (1) on such datasets will severely hurt the robustness and generalization of cross-modal matching model since its potential overfitting to the noisy samples. To tackle this issue, a common approach [12, 17, 39] is to roughly divide the training data into clean and noisy sets. Moreover, this approach typically incorporates the design of soft margins in order to enhance the robustness of triplet loss against noise, which could be formulated as follows:

$$\mathcal{L}_S(I_i, T_i) = [\hat{\gamma} - S(I_i, T_i) + S(\hat{I}_h, T_i)]_+ + [\hat{\gamma} - S(I_i, T_i) + S(I_i, \hat{T}_h)]_+ \quad (2)$$

where $\hat{\gamma}$ represents the soft margins adaptively computed by $\hat{\gamma} = \frac{m^{\hat{y}_{ii}} - 1}{m - 1}$, m is a curve parameter and $\hat{y}_{ii}$ denotes the rectified soft correspondence labels. However, this approach still has two limitations: 1) the division typically includes twofold errors due to the properties of GMM/BMM distribution, i.e., truly matched data pairs are distinguished as mismatched and vice versa. 2) the rectified labels utilized to recast soft margins are unstable, especially those wrong soft correspondence labels might produce the risk of matching failure in high noise ratio. To overcome these limitations, we propose an *Uncertainty Guided Division (UGD)* strategy to perform explicit division (Sec 3.2) and a *Trusted Robust Loss (TRL)* to provide reliable rectification (Sec 3.3). The details are delineated next.

3.2 Uncertainty Guided Division

3.2.1 Multi-View Bidirectional Evidence Extraction. For each input image, we first use Faster R-CNN [35] model pretrained on Visual Genomes [19] to extract R region-level visual representation $V = \{v_1, v_2, ..., v_i\}_{i=1}^R$, where R represents the number of salient regions in an image. Afterwards, a visual modal feature extractor, denoted as Φ, is employed to obtain multi-view embedding $\{v_{w,i}\}_{w=1}^W = f(v_i)$, where W is the number of visual views and $v_{w,i} \in \mathbb{R}^d$ represents multi-view region-level embedding in a shared space. Likewise, for each input text, we first tokenize it into L words and sequentially feed the word embeddings into a bi-directional GRU [36]. After that, a textual feature extractor Ψ is exploited to project the word embeddings into the shared space, denoted as $T = \{t_j | j \in [1, ..., L], t_j \in \mathbb{R}^d\}$, where L indicates the length of input sentence and t_j represents the j-th word-level embedding. For ease of representation, we only consider the computation in a given mini-batch with K pairs in the following. Considering image-text pair (V_i^w, T_j), the evidence of w-th view could be extracted by the cross-modal evidence extractor Θ, which is defined as:

$$e_{ij}^w = \Theta(S(V_i^w, T_j)) = \exp(softplus(S(V_i^w, T_j))/\tau) \quad (3)$$

where $0 < \tau < 1$ serves as a scaling parameter and $softplus(x) = log(1+exp(x))$ represents a smooth approximation of the ReLU activation function. Therefore, the evidence vector $e_i^{w,i2t} = [e_{i1}^w, e_{i2}^w, ..., e_{iK}^w]$ of a given w-th view visual query I_i^w could be extracted from the corresponding cross-modal similarities through Eq. (3). Likewise, the evidence vector $e_j^{w,t2i} = [e_{1j}^w, e_{2j}^w, ..., e_{Kj}^w]$ of a given textual query T_j could be obtained. Then cross-modal bidirectional evidence vector of the given w-th view image-text pair could be obtained:

$$e^w = e_i^{w,i2t} + e_j^{w,t2i} = [e_1^w, e_2^w, ..., e_K^w] \quad (4)$$

where e^w refers the aggregated bidirectional evidence to support classification, i.e., we regard cross-modal matching task as a K-way classification problem, where the w-th view visual embedding or textual embedding serves as the query to be classified.

3.2.2 Cross-Modal Uncertainty Modeling. Having introduced multi-view bidirectional evidence, we now focus on model the cross-modal uncertainty, which follows the principles of Subjective Logic [18]. For simplicity, we assume that each view has equal contribution. Subjective Logic posits a theoretical framework predicated on evidence sourced from data, aimed to derive the probabilities (belief mass) of different classes and overall uncertainty (uncertainty mass) of the multi-classification problem. Note that the evidence refers to the metrics sourced from input data to support the classification and is closely related to the concentration parameters of Dirichlet distribution, a statistical model utilized to express the likelihood of outcomes based on prior observations. We could intuitively regard cross-modal matching task as K-way classification that a query is classified to its corresponding cross-modal counterpart. Subjective Logic tries to assign a belief mass and overall uncertainty mass to each query based on the multi-view bidirectional evidence. Specifically, for the *w-th* view of visual embedding, the $K+1$ mass values are all non-negative and their sum is one, which could defined as:

$$u^w + \sum_{k=1}^{K} b_k^w = 1 \tag{5}$$

where $u^w \geq 0$ and $b_k^w \geq 0$ indicate the overall uncertainty and the classification probability for the k-th class of w-th view image-text pair, respectively. For the w-th view image-text pair, subjective logic connects the bidirectional evidence e^w to the parameters of the Dirichlet distribution $\alpha^w = [\alpha_1^w, \alpha_2^w, ..., \alpha_K^w]$. Specifically, the parameter α_k^w of the Dirichlet distribution is derived from e_k^w, i.e., $\alpha_k^w = e_k^w + 1$. Then the belief mass b_k^w and the uncertainty u^w are calculated as follows:

$$b_k^w = \frac{e_k^w}{L^w} = \frac{\alpha_k^w - 1}{L^w} \quad \text{and} \quad u^w = \frac{K}{L^w} \tag{6}$$

where $L^w = \sum_{i=1}^{K}(e_i^w + 1) = \sum_{i=1}^{K} \alpha_i^w$ indicates the Dirichlet distribution strength and the belief mass b_k^w could be viewed as subjective opinions corresponding to the Dirichlet distribution. One could easily observe that the more bidirectional evidence supported for the classification of k-th cross-modal counterpart, the greater the probability assigned to the k-th query.

The Dempster-Shafer theory (DST) [5] of evidence is designed to combine evidence from different sources at a degree of belief (also known as the belief function) that comprehensively considers all available evidence. Like TMC[13], leveraging the combination rule of DST, we combine W independent sets of probability mass assignment $\{\mathcal{M}^w\}_{w=1}^{W}$ for each image-text pair, where $\mathcal{M}^w = \{\{b_k^w\}_{k=1}^{K}, u^w\}$, to obtain a joint opinion $\mathcal{M} = \{\{b_k\}_{k=1}^{K}, u\}$. According to Eq. (6), the corresponding joint bidirectional evidence from multiple views and the parameters of Dirichlet distribution are induced as follows:

$$L = \frac{K}{u}, \quad e_k = b_k \times L \quad \text{and} \quad \alpha_k = e_k + 1 \tag{7}$$

based on the above combination rule, we can obtain the estimated multi-view joint bidirectional evidence and the corresponding parameters of joint Dirichlet distribution to produce the final probability (represented by b_k) of each query is classified to its cross-modal counterpart and the overall uncertainty (represented by u).

Intuitively, by setting a threshold p to u, we can divide the training data into two distinct sets: a determined set S_D with definitive hard-labels wherein clean data labeled as $\tilde{y} = 1$ and noisy data labeled as $\tilde{y} = 0$, and an hard set S_H wherein the labels of all pairs are wiped and replaced by the estimated soft corresponding labels, denoted as $\tilde{y} \in [0, 1]$, to accurately depict the corresponding degree of cross-modal pairs. The division could be formulated as:

$$\begin{cases} S_D \supseteq (I_i, T_i), & u_i < p \\ S_H \supseteq (I_i, T_i), & u_i \geq p \end{cases} \tag{8}$$

more specific, for the determined set S_D, the definitive hard-labels could be obtained using the joint belief mass as follows:

$$\tilde{y}_i = \begin{cases} 1, & \text{if } i = \arg\max_k (b_{i,k}) \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where $\tilde{y}_i = 1$ indicates the pair (I_i, T_i) belongs to the determined clean set S_C and another belongs to the determined noisy set S_N. We further formulated it as follows:

$$\begin{cases} S_C \supseteq (I_i, T_i), & \tilde{y}_i = 1 \\ S_N \supseteq (I_i, T_i), & \tilde{y}_i = 0 \end{cases} \tag{10}$$

based on the above division, one could find that the three sets are explicitly divided, and thus we could effortless mitigate the adverse impact of noisy pairs in S_N. For the data pairs of hard set S_H, the most challenge is to accurately estimate its soft corresponding labels. In contrast to most existing methods usually regard the predicted similarity scores as the soft correspondence labels, our UGNCL obtains the labels from the estimated belief mass of Dirichlet distribution, the specific calculation is defined as $\tilde{y}_i = b_{i,i}$. Moreover, the hard set S_H could be divided into two subsets: a hardly-determined clean subset S_H^C and noisy subset S_H^N. The specific division could be formulated as follows:

$$\begin{cases} S_H^C \supseteq (I_i, T_i), & \tilde{y}_i > \epsilon \\ S_H^N \supseteq (I_i, T_i), & \tilde{y}_i \leq \epsilon \end{cases} \tag{11}$$

where the ϵ is a fixed threshold, and $\tilde{y}_i$ could be employed to determine whether UGNCL performs positive or negative learning on the pairs of the hard set S_H. The details are presented in Sec 3.3.

3.2.3 Bidirectional Opinion Learning. In this section, we will discuss how to learn bidirectional cross-modal mass $\mathcal{M}^w$ of each view. We intuitively view cross-modal matching task as K-way classification problem that a query is classified to its corresponding cross-modal counterpart. Given a query I_i or T_i and its retrieval ground truth $\boldsymbol{y_i} \in \mathbb{R}^K$ (wherein the i-th element is $\tilde{y}_i$ and the rest are 0), we can get the cross-modal bidirectional evidence through Eq. (4) and then derive the parameter $\boldsymbol{\alpha_i}$ of Dirichlet distribution. Furthermore, the cross-modal opinions $D(\boldsymbol{p_i}|\boldsymbol{\alpha_i})$ can be formed, where $\boldsymbol{p_i}$ represents the class probabilities and its density follows $\boldsymbol{\alpha_i}$. Leveraging the least-squares loss, we make the query probabilities $\boldsymbol{p_i}$ approach the ground truth $\boldsymbol{y_i}$, which indicates the class probabilities of a query classified its cross-modal counterpart in our case. The specific formulation could be defined as:

$$\begin{aligned} \mathcal{L}_{ls}(\boldsymbol{\alpha_i}, \boldsymbol{y_i}) &= \int ||\boldsymbol{y_i} - \boldsymbol{p_i}||_2^2 \frac{1}{B(\boldsymbol{\alpha_i})} \prod_{j=1}^{K} p_{ij}^{\alpha_{ij}-1} d\boldsymbol{p_i} \\ &= \sum_{j=1}^{K} [(y_{ij} - \mathbb{E}(p_{ij}))^2 + \mathrm{Var}(p_{ij})] \\ &= \sum_{j=1}^{K} (y_{ij} - \frac{\alpha_{ij}}{L_i})^2 + \frac{\alpha_{ij}(L_i - \alpha_{ij})}{L_i^2(L_i + 1)} \end{aligned} \tag{12}$$

where $\mathbb{E}(p_{ij})$ and $\mathrm{Var}(p_{ij})$ are the expected value and the variance of p_{ij}, respectively. Noted that we estimate the expected probability $\mathbb{E}(p_{ij})$ by $\frac{\alpha_{ij}}{L_i}$ [37]. The above loss function ensures that the evidence supported matched pairs could be higher than the one supported mismatched pairs. However, it is challenging to generate scant evidence for mismatched pairs, and ideally, the generated evidence is expected to trend towards zero in this case. Thus the following Kullback-Leiber (KL) divergence term is introduced:

$$\begin{aligned}\mathcal{L}_{kl}(\boldsymbol{\alpha_i}, \boldsymbol{y_i}) &= KL[D(\boldsymbol{p_i}|\tilde{\boldsymbol{\alpha}}_{\boldsymbol{i}})||D(\boldsymbol{p_i}|\mathbf{1})] \\ &= \log(\frac{\Gamma(\sum_{j=1}^{K} \tilde{\alpha}_{ij})}{\Gamma(K)\prod_{j=1}^{K}\Gamma(\tilde{\alpha}_{ij})}) + \sum_{j=1}^{K}(\tilde{\alpha}_{ij} - 1)[\psi(\tilde{\alpha}_{ij}) - \psi(\sum_{j=1}^{K}\tilde{\alpha}_{ij})],\end{aligned} \tag{13}$$

where $\tilde{\boldsymbol{\alpha}}_{\boldsymbol{i}} = \boldsymbol{y_i} + (1 - \boldsymbol{y_i}) \odot \boldsymbol{\alpha_i}$ represents the adjusted Dirichlet parameters with unreliable evidence removed from the predicted Dirichlet distribution and helps avoid the evidence of the ground truth class erroneously reduced to zero. The $\Gamma(\cdot)$ and $\psi(\cdot)$ are gamma and digamma functions, respectively.

Once we obtain the evidence vector $\boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{w},\boldsymbol{i2t}} \in \mathbb{R}^K$, then the corresponding parameters of Dirichlet distribution could be calculated as $\boldsymbol{\alpha}_{\boldsymbol{i}}^{\boldsymbol{w},\boldsymbol{i2t}} = \boldsymbol{e}_{\boldsymbol{i}}^{\boldsymbol{w},\boldsymbol{i2t}} + 1$. Afterwards, the joint belief mass could be derived through the combination rule of DST, then the combined evidence vector $\boldsymbol{e}_{\boldsymbol{i}}^{i2t}$ and the corresponding parameters of Dirichlet distribution $\boldsymbol{\alpha}_{\boldsymbol{i}}^{i2t}$ cloud be induced by Eq. (7). Furthermore, the evidential loss of image-to-text could be defined as:

$$\mathcal{L}_e^{i2t}(I_i, \tilde{y}_i) = \mathcal{L}_{ls}(\boldsymbol{\alpha}_{\boldsymbol{i}}^{\boldsymbol{i2t}}, \tilde{y}_i) + \xi_1\mathcal{L}_{kl}(\boldsymbol{\alpha}_{\boldsymbol{i}}^{\boldsymbol{i2t}}, \tilde{y}_i) \tag{14}$$

where $1 > \xi > 0$ is a balance factor. Likewise, the evidential loss $\mathcal{L}_e^{t2i}(T_i, \tilde{y}_i)$ of text-to-image could also be obtained by the above equation. Finally, the bidirectional evidential loss could be formulated as:

$$\mathcal{L}_e(I_i, T_i, \tilde{y}_i) = \mathcal{L}_e^{i2t}(I_i, \tilde{y}_i) + \mathcal{L}_e^{t2i}(T_i, \tilde{y}_i) \tag{15}$$

3.3 Trusted Robust Loss

As mentioned 3.1, most existing methods tackle the noisy correspondence using the Eq. (2), which easily suffers the inaccurate prediction $\hat{y}_{ii}$ and thus fails to produce well performance. In a word, the prediction is not robust against noise. Considering our explicit division, the easily-determined partitions S_C and S_N could be well-processed using Eq. (1). However, one could find that there are two embarrassing phenomenons in the hard partition S_H: the truly matched pairs are wrongly considered as mismatched pairs and vise verse. To this end, we present a *Trusted Robust Loss (TRL)* to provide a reliable rectification for the calibrated soft margins $\hat{\gamma}$ and further enhance the robustness of models against noise. The specific formulation could be defined as follows:

$$\mathcal{L}_t(I_i, T_i) = \frac{1}{\lambda}\sum_{j=1}^{\lambda}([\tilde{\gamma}_i - S(I_i, T_i) + S(\hat{I}_j, T_i)]_+ + [\tilde{\gamma}_i - S(I_i, T_i) + S(I_i, \hat{T}_j)]_+), \tag{16}$$

where $\hat{I}_j$ and $\hat{T}_j$ represents the selected *top-λ* hardest negatives, and λ could dynamically increase following the training epoch. Specifically, it could be formulated as $\lambda = \max(\lceil K - \delta * Epoch \rceil, \theta)$, where $\lceil x \rceil$ indicates the rounding down operation, δ is a decay factor and θ represents the lower bound of λ. More specific, the trusted soft margin $\tilde{\gamma}_i$ could be adaptively determined by:

$$\tilde{\gamma}_i = \begin{cases} \frac{1}{1+(\frac{1}{u-1})^{-\Delta}}\hat{\gamma}_i, & (I_i, T_i) \in S_H^C \\ \frac{1}{1+(\frac{u}{1-u})^{-\Delta}}\hat{\gamma}_i, & (I_i, T_i) \in S_H^N \end{cases} \tag{17}$$

where $\Delta > 0$ is an empirical parameter and $\hat{\gamma}_i$ represents the soft margins calibrated by estimated soft correspondence labels. S_H^C and S_H^N indicate the hardly-determined clean and noisy partitions, respectively. Theoretically, the above trusted soft margins $\tilde{\gamma}_i$ could be assigned a higher value for the matched data pairs with lower uncertainty and mismatched data pairs with high uncertainty, while ensuring the values lower the initial margin γ, which means that all pairs in the hard partition contribute lower than those in easily-determined clean partition. Finally, the overall objective function of our UGNCL could be formulated as follows:

$$\mathcal{L}_{overall} = \mathcal{L}_e + \xi_2\mathcal{L}_t \tag{18}$$

4 EXPERIMENT

To validate the effectiveness of our proposed UGNCL, this section demonstrates extensive experiments conducted on three widely-used benchmark datasets: Flickr30K[41], MS-COCO[24], and Conceptual Captions[38]. Due to the Flickr30K and MS-COCO are two well-annotated datasets, thus we simulate the noisy correspondence on them. In contrast, the more challenging Conceptual Captions dataset contains real-world noisy correspondence. Furthermore, we also provide detailed descriptions of UGNCL's implementation.

4.1 Datasets and Performance Measurements

The following three widely-used cross-modal matching datasets are utilized in our experiments:

- **Flickr30K** contains 31,000 images collected from the Flickr website with five captions for each image. Following previous work [20], we use 1,000 images for validation, 1,000 images for testing and the rest for training.
- **MS-COCO** contains 123,287 images and each image is annotated with five captions. Following previous work [20], we use 5,000 images for validation, 5,000 images for testing and the rest 113,287 images for training.
- **Conceptual Captions** is a large-scale image-text dataset, with all data pairs sourced from the Internet, which results in approximately 3% ∼ 20% image-text pairs are mismatched, i.e. noisy correspondence. It contains 3.3M images with a single caption per image. For fair comparison, we employ the same subset, dubbed CC152K, following NCR[17]. Specifically, we use 150,000 pairs for training, 1,000 pairs for validation and 1,000 pairs for testing.

Evaluation Protocol. For cross-modal matching task, the Recall at K (R@K) is a widely-used metric for performance evaluation, defined as the percentage of queries correctly matched among the top K retrieval instances [26, 27]. In our experiments, the R@1, R@5, R@10, and their sum are reported to provide a comprehensive evaluation.

4.2 Implementation Details

Our proposed UGNCL could extent the capabilities of almost all cross-modal visual-language methods, significantly improving their robustness against noisy correspondence. For fair comparison, we use the same settings as NCR [17] except the specific parameters of our method, and report the results of SGR, SAF and SGRAF [7] as backbone to comprehensively validate the effectiveness our proposed UGNCL. Noted that we conduct our comparison experiments without any extra preprocess to datasets or using any other data source. Specifically, the joint embedding dimensionality d is 1024 and the mini-batch K is set to 128. Additionally, $m, \tau, \xi_1, \xi_2, \delta$ and θ were empirically set to $\{10, 0.1, 0.1, 0.8, 0.25, 5\}$. After that, we initialize the margin parameter, denoted as γ, to 0.2, and set the empirical parameter $\triangle$ to 10 for reliable rectification of soft margins. Moreover, we set the number of views, represented as W, to 2 and the division threshold p and ϵ to 0.5 for all experiments. For MS-COCO datasets, the training processes contain 20 epochs after a 2 epochs warmup, while other datasets contain 40 epochs following a 5 epochs warmup. Finally, at the inference stage, we average the similarities calculated by diverse views for the retrieval evaluation.

4.3 Comparison with State-of-the-Arts

In this section, we conduct extensive comparisons on three widely-used benchmarks with 8 state-of-the-art methods, i.e., SCAN (ECCV' 2018) [20], VSRN (ICCV' 2019) [22], IMRAM (CVPR' 2020) [1], SGRAF (SGR and SAF, AAAI' 2021) [7], NCR (NeurIPS' 2021) [17], DECL (ACM MM' 2022) [33], MSCN (CVPR' 2023 [12] and BiCro (CVPR' 2023) [39]. As the Flickr30K and MS-COCO are well-annotated datasets, we artificially generate synthetic noisy correspondence by randomly shuffling the captions like [17] for a specific percentage (i.e., 20%, 50% and 70%). For fairness, we directly refer to the results already reported in the corresponding papers and retrain the baselines with the recommended settings to obtain the results that these works did not report. Moreover, we choose the best checkpoint on the validation set and report its performance on the test set for all methods. Following NCR, we also report two strong baselines based SGR, i.e., SGR-C and SGR*. In short, SGR-C selects only clean data for training, and SGR* employs a pre-training process while training without hard negatives to improve robustness. Like [17, 33, 39], the ensemble results of UGNCL-SGRAF, denoted as UGNCL*, are reported in this paper.

4.3.1 Comparisons on Synthetic Noise. Tab. 1 reports the quantitative results on Flickr30K dataset and over 5 folds of 1K test images MS-COCO dataset. From the experimental results, one could observe that UGNCL achieves best overall performance. Specifically, UGNCL not only performs well under low noise ratio, but also its performance significantly exceeds that of other state-of-the-art methods under high noise ratio, especially 70% noise ratio. The reason of all baselines's performance significantly degrades in the high noise ratio may be that these methods adversely suffer the determined noise, while our UGNCL effectively mitigate the impact of the noise due to the explicit division.

4.3.2 Comparisons on Real-world Noise. Tab. 2 shows the quantitative results on more challenging dataset CC152K with real-world noisy correspondence. The results clearly indicate that our UGNCL surpasses all the evaluated baselines, achieving the best overall performance of 373.1%, which strongly demonstrates the stability and robustness against noise of our method. Specifically, UGNCL outperforms the best baseline BiCro with performance improvement of 2.8% and 0.6% in terms of R@1 in text and image retrieval, respectively. Moreover, compared with BiCro filters the noisy data according to their soft correspondence labels, the performance gap shows that our UGD is more effective and robust to mitigate the impact of mismatched data on performance.

4.3.3 Comparisons on Well-annotated Correspondences. For a comprehensive comparison, we also evaluate the performance of UGNCL on MS-COCO dataset without additional noise. From the quantitative results of MS-COCO 5K shown in Tab. 2, one could see that UGNCL achieves state-of-the-art performance in terms of all evaluated metrics, which shows that our UGNCL not only demonstrates superiorities in handling noisy cases but also performs effectively in well-aligned scenarios. Specifically, in comparison to the best baseline BiCro, UGNCL improves the score for retrieving by 1.8%, 1.0%, 0.6%, 1.0%, 0.9% and 0.6% across all metrics.

4.4 Comparison with Pre-trained Model

In this section, we compare our method with the well-known large pre-trained model CLIP [34]. CLIP is trained from scratch on a massive dataset collected from the Internet and thus presumably has a large number of image-text pairs with real-world noisy correspondence. Following the settings of NCR [17], we report the zero-shot and fine-tuning results on MS-COCO 5K dataset in Tab. 3. Noted that we exploit two baselines, i.e., CLIP (ViT-L/14) denoted as CLIP-14 and CLIP (ViT-B/32) denotes as CLIP-32. From the results, one could find that our method achieves the best performance across all metrics and even the retrieval scores of each metric under high noise are higher than the zero-shot result of CLIP-14 by a large margin, which is strongly proved that our method has the potential and robustness to deal with the noisy correspondence.

4.5 Ablation Study

In this section, we conduct ablation studies on Flickr30K with 20% noise to investigate the contributions of different proposed components in our UGNCL, i.e., UGD and TRL. The results are presented in Tab. 4, one could observe that the full version of our UGNCL* achieves the best performance, which indicates that each component contributes to our method for performance improvement. For a fair comparison and effectively explore the effect of UGD, like previous works [17, 39], we divide the training data into two subsets (i.e., clean subset and noisy subset.) according to their predicted class probabilities and further estimate the soft correspondence labels for the pairs in the noisy subset. In comparison to the one without UGD, one could see that our UGNCL* improves 10.5% in the overall performance, which strongly indicates that the provided division of UGD is accurate and reliable, remarkably mitigate the impact of the easily-determined noise. Furthermore, to explore the effect of TRL, we replace it with the widely-used soft triplet loss $\mathcal{L}_S$. From the result of the second line, one could see that our UGNCL* achieves 6.5% improvement in the terms of overall performance, which indicates that our TRL could provide more trusted and suitable soft margins rectification for the pairs in the hard subset and further enhance the robustness of triplet loss against noise.

4.6 Visualization on Uncertainty

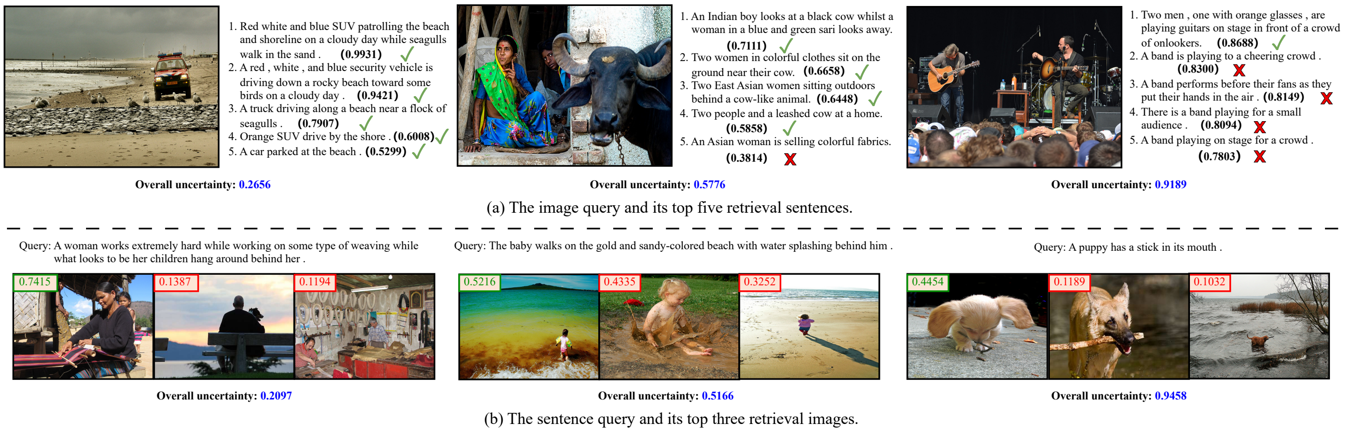
In this section, we showcase some examples form Flickr30K in Fig. 3 to demonstrate the effectiveness and reliability of our method. Noted that the similarity score for each image-text pair and overall

Table 1: Cross-modal retrieval performance comparison under synthetic noise rates of 20%, 50% and 70% on Flickr30K and MS-COCO 1K, and the best results are highlighted on bold.

		Flickr30K							MS-COCO 1K						
		Image to Text			Text to Image				Image to Text			Text to Image			
Noise	Methods	R@1	R@5	R@10	R@1	R@5	R@10	rSum	R@1	R@5	R@10	R@1	R@5	R@10	rSum
20%	SCAN	59.1	83.4	90.4	36.6	67.0	77.5	414.0	66.2	91.0	96.4	45.0	80.2	89.3	468.1
	VSRN	58.1	82.6	89.3	40.7	68.7	78.2	417.6	25.1	59.0	74.8	17.6	49.0	64.1	289.6
	IMRAM	63.0	86.0	91.3	41.4	71.2	80.5	433.1	68.6	92.8	97.6	55.7	85.0	91.0	490.7
	SGR*	62.8	86.2	92.2	44.4	72.3	80.4	438.3	67.8	91.7	96.2	52.9	83.5	90.1	482.2
	SGR-C	72.8	90.8	95.4	56.4	82.1	88.6	486.1	75.4	95.2	97.9	60.1	88.5	94.8	511.9
	NCR	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	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
	MSCN	77.4	94.9	97.6	59.6	83.2	89.2	501.9	78.1	**97.2**	98.8	**64.3**	90.4	95.8	524.6
	BiCro	78.1	94.4	97.5	**60.4**	**84.4**	**89.9**	504.7	78.8	96.1	98.6	63.7	90.3	95.7	523.2
	UGNCL*	**78.4**	**95.8**	**97.8**	59.8	84.3	89.5	**505.6**	**79.5**	**97.2**	**99.0**	63.7	**90.9**	**96.0**	**526.3**
50%	SCAN	27.7	57.6	68.8	16.2	39.3	49.8	259.4	40.8	73.5	84.9	5.4	15.1	21.0	240.7
	VSRN	14.3	37.6	50.0	12.1	30.0	39.4	183.4	23.5	54.7	69.3	16.0	47.8	65.9	277.2
	IMRAM	9.1	26.6	38.2	2.7	8.4	12.7	97.7	21.3	60.2	75.9	22.3	52.8	64.3	296.8
	SGR*	36.9	68.1	80.2	29.3	56.2	67.0	337.7	60.6	87.4	93.6	46.0	74.2	79.0	440.8
	SGR-C	69.8	90.3	94.8	50.1	77.5	85.2	467.7	71.7	94.1	97.7	57.0	86.6	93.7	500.8
	NCR	72.9	93.0	96.3	54.3	79.8	86.5	482.8	74.6	94.7	97.8	59.1	87.8	94.5	508.5
	DECL	72.7	92.0	95.8	54.8	80.4	87.5	483.2	76.1	95.0	98.3	60.5	88.7	94.9	513.5
	MSCN	**74.4**	**93.2**	96.0	55.3	80.4	86.8	486.1	**77.5**	96.2	**98.7**	60.7	89.1	94.9	517.1
	BiCro	73.1	91.4	96.1	53.7	80.2	**87.4**	481.9	76.2	96.0	98.4	**61.6**	89.2	95.1	516.5
	UGNCL*	74.1	93.0	**96.9**	**56.7**	**80.9**	87.3	**488.9**	77.2	**96.4**	**98.7**	**61.6**	**89.7**	**95.3**	**518.9**
70%	SCAN	5.6	19.3	27.4	2.2	8.0	12.8	75.3	18.1	43.1	57.4	0.3	1.3	2.3	122.5
	VSRN	0.8	2.5	4.1	0.5	1.5	2.7	12.1	5.1	15.7	24.6	2.5	8.8	13.3	70.0
	IMRAM	1.3	3.1	3.9	0.3	1.2	2.8	12.6	7.1	20.0	33.4	5.3	15.2	22.0	103.0
	SGR*	17.9	42.1	51.9	14.6	31.0	40.8	198.3	35.7	71.2	85.4	31.6	65.8	79.0	368.7
	SGR-C	65.0	89.3	94.7	48.1	74.5	81.1	452.7	69.8	93.6	97.5	56.5	86.0	93.4	496.8
	NCR	16.1	38.5	52.8	11.0	29.5	41.4	189.3	35.4	69.5	83.4	31.5	66.4	81.1	367.3
	DECL	60.4	85.4	90.8	42.8	71.5	80.7	431.7	70.5	94.7	98.8	56.1	85.3	93.2	498.6
	MSCN	69.0	89.3	93.0	49.2	73.1	79.0	452.6	74.4	94.9	97.7	58.8	87.2	**93.7**	506.7
	BiCro	65.2	88.0	93.2	48.0	74.1	81.6	450.1	73.9	95.5	97.8	59.2	**87.8**	93.6	507.9
	UGNCL*	**70.3**	**91.1**	**96.0**	**52.4**	**77.6**	**83.4**	**470.8**	**75.7**	**95.9**	**98.4**	**59.6**	87.5	92.6	**509.7**

Table 2: Cross-modal retrieval performance comparison on CC152K with real noise and MS-COCO 5K with the evaluated model trained on clean data, and the best results are highlighted in bold.

	CC152K							MS-COCO 5K						
	Image to Text			Text to Image				Image to Text			Text to Image			
Methods	R@1	R@5	R@10	R@1	R@5	R@10	rSum	R@1	R@5	R@10	R@1	R@5	R@10	rSum
SCAN	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	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	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	31.7	59.3	68.2	31.9	59.0	67.9	318.0	55.5	83.8	91.8	40.1	69.7	80.4	421.3
SGR	11.3	29.7	39.6	13.1	30.1	41.6	165.4	57.3	83.2	90.6	40.5	69.6	80.3	421.5
NCR	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	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
MSCN	40.1	65.7	**76.6**	40.6	67.4	76.3	366.7	-	-	-	-	-	-	-
BiCro	40.8	**67.2**	76.1	42.1	67.6	**76.4**	370.2	59.0	84.4	91.7	42.4	71.2	81.7	430.4
UGNCL*	**43.6**	67.1	74.9	**42.7**	**68.4**	**76.4**	**373.1**	**60.8**	**85.4**	**92.3**	**43.4**	**72.1**	**82.3**	**436.2**


1. Red white and blue SUV patrolling the beach and shoreline on a cloudy day while seagulls walk in the sand . **(0.9931)** ✓
2. A red , white , and blue security vehicle is driving down a rocky beach toward some birds on a cloudy day . **(0.9421)** ✓
3. A truck driving along a beach near a flock of seagulls . **(0.7907)** ✓
4. Orange SUV drive by the shore . **(0.6008)** ✓
5. A car parked at the beach . **(0.5299)** ✓

1. An Indian boy looks at a black cow whilst a woman in a blue and green sari looks away. **(0.7111)** ✓
2. Two women in colorful clothes sit on the ground near their cow. **(0.6658)** ✓
3. Two East Asian women sitting outdoors behind a cow-like animal. **(0.6448)** ✓
4. Two people and a leashed cow at a home. **(0.5858)** ✓
5. An Asian woman is selling colorful fabrics. **(0.3814)** ✗

1. Two men , one with orange glasses , are playing guitars on stage in front of a crowd of onlookers. **(0.8688)** ✓
2. A band is playing to a cheering crowd . **(0.8300)** ✗
3. A band performs before their fans as they put their hands in the air . **(0.8149)** ✗
4. There is a band playing for a small audience . **(0.8094)** ✗
5. A band playing on stage for a crowd . **(0.7803)** ✗

Overall uncertainty: 0.2656
Overall uncertainty: 0.5776
Overall uncertainty: 0.9189

(a) The image query and its top five retrieval sentences.

Query: A woman works extremely hard while working on some type of weaving while what looks to be her children hang around behind her .

Query: The baby walks on the gold and sandy-colored beach with water splashing behind him .

Query: A puppy has a stick in its mouth .

0.7415
0.1387
0.1194
0.5216
0.4335
0.3252
0.4454
0.1189
0.1032
Overall uncertainty: 0.2097
Overall uncertainty: 0.5166
Overall uncertainty: 0.9458

(b) The sentence query and its top three retrieval images.

Figure 3: Some retrieved examples of cross-modal retrieval on Flickr30K under 50% noise. The top-5 ranked sentences for each image query are demonstrated in (a). Sentences that match correctly are marked with a green tick, while mismatches are indicated with a red cross. For each sentence query, we show the top-3 ranked images from left to right in (b). We highlight the similarities of correctly matched images with green boxes and texts, the mismatched ones with red boxes and texts. All the estimated overall uncertainties are marked with blue texts.

Table 3: Comparison with CLIP on MS-COCO 5K.

		Image to Text			Text to Image			
Noise	Methods	R@1	R@5	R@10	R@1	R@5	R@10	rSum
0%	CLIP-14	58.4	81.5	88.1	37.8	62.4	72.2	400.4
	CLIP-32	50.2	74.6	83.6	30.4	56.0	66.8	361.6
20%	CLIP-14	36.1	61.3	72.5	22.6	43.2	53.7	289.4
	CLIP-32	21.4	49.6	63.3	14.8	37.6	49.6	236.3
	UGNCL*	**57.5**	**85.3**	**92.2**	**42.3**	**71.5**	**81.6**	**430.4**
50%	CLIP-32	10.9	27.8	38.3	7.8	19.5	26.8	131.1
	UGNCL*	**57.2**	**84.1**	**91.6**	**41.6**	**70.8**	**81.0**	**426.4**

Table 4: Ablation studies of UGD and TRL on Flick30K with 20% noise.

Method		Image to Text			Text to Image			
UGD	TRL	R@1	R@5	R@10	R@1	R@5	R@10	rSum
✓	✓	**78.4**	**95.8**	**97.8**	**59.8**	**84.3**	**89.5**	**505.6**
✓		77.0	94.1	97.0	58.8	83.3	88.7	499.1
	✓	75.1	93.9	97.3	58.6	82.1	88.0	495.1
		75.6	93.7	96.9	57.0	82.3	88.3	493.8

uncertainty are both depicted. One cloud observe that the lower uncertainty leads to more determined retrieval results. Specifically, in cases of results on the left with low uncertainty, the corresponding matches in retrieval typically exhibit high degree of similarity, whereas the mismatched results display markedly lower similarity. From the right cases with high uncertainty, high similarity dose not always guarantee a match and vice verse, which means high similarity but factually mismatched and low similarity but factually matched. That is to say, the deterministic similarity is not enough to reflect the confidence of cross-modal retrieval. Therefore, the proposed division based uncertainty is interpretable and effective, which could enhance the robustness against noisy correspondence.

5 CONCLUSION

This paper investigates the challenging problem of robust cross-modal matching with noisy correspondence, and presents a novel *Uncertainty-Guided Noisy Correspondence Learning (UGNCL)* framework to achieve noise-robust cross-modal matching. Specifically, the proposed *Uncertainty Guided Division (UGD)* is capable of providing the explicit and accurate division, mitigating the impact of easily-determined noise. Furthermore, a *Trusted Robust Loss (TRL)* is exploited to reliably recast the soft margins to enhance the robustness against noise. Extensive experiments conducted on three widely-used datasets demonstrate that the proposed method achieves state-of-the-art effectiveness and robustness in handling both synthetic and real-world noisy correspondence.

ACKNOWLEDGMENTS

This work was supported in part by the NSFC / Research Grants Council (RGC) Joint Research Scheme under grant: N_HKBU214/21, in part by the General Research Fund of RGC under the grants: 12201321, 12202622, 12201323, in part by the RGC Senior Research Fellow Scheme under the grant: SRFS2324-2S02, in part by the National Natural Science Foundation of China under grant: 62222203, in part by the Natural Science Foundation of Fujian Province under the grant: 2021J01317, and in part by the Key Project of Xiamen Science and Technology Plan (Future Industry Field) under the grant: 3502Z20234035.

REFERENCES

[1] 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.* 12652–12660.

[2] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. 2021. Learning the Best Pooling Strategy for Visual Semantic Embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 15789–15798.

[3] Mengyuan Chen, Junyu Gao, Shicai Yang, and Changsheng Xu. 2022. Dual-evidential learning for weakly-supervised temporal action localization. In *Proceedings of the European Conference on Computer Vision.* 192–208.

[4] Yuhao Cheng, Xiaoguang Zhu, Jiuchao Qian, Fei Wen, and Peilin Liu. 2022. Cross-Modal Graph Matching Network for Image-Text Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 4 (2022), 23 pages.

[5] Arthur P Dempster. 2008. Upper and lower probabilities induced by a multivalued mapping. In *Classic works of the Dempster-Shafer theory of belief functions.* 57–72.

[6] John S. Denker and Yann LeCun. 1990. Transforming Neural-Net Output Levels to Probability Distributions. In *Proceedings of the 3rd International Conference on Neural Information Processing Systems.* 853–859.

[7] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI Conference on Artificial Intelligence.* 1218–1226.

[8] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *Proceedings of the British Machine Vision Conference.*

[9] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc'Aurelio Ranzato, and Tomas Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *Proceedings of Advances in Neural Information Processing Systems.*

[10] Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. 2023. Learning Semantic Relationship Among Instances for Image-Text Matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 15159–15168.

[11] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the 33rd International Conference on Machine Learning.* 1050–1059.

[12] Haochen Han, Kaiyao Miao, Qinghua Zheng, and Minnan Luo. 2023. Noisy Correspondence Learning with Meta Similarity Correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 7517–7526.

[13] Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. 2023. Trusted Multi-View Classification With Dynamic Evidential Fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 2 (2023), 2551–2566.

[14] Yi He, Xin Liu, Yiu-Ming Cheung, Shu-Juan Peng, Jinhan Yi, and Wentao Fan. 2021. Cross-graph attention enhanced multi-modal correlation learning for fine-grained image-text retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1865–1869.

[15] Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng. 2023. Cross-Modal Retrieval With Partially Mismatched Pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 8 (2023), 9595–9610.

[16] Zhikai Hu, Xin Liu, Xingzhi Wang, Yiu-ming Cheung, Nannan Wang, and Yewang Chen. 2019. Triplet Fusion Network Hashing for Unpaired Cross-Modal Retrieval. In *Proceedings of ACM International Conference on Multimedia Retrieval.* 141–149.

[17] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. 2021. Learning with Noisy Correspondence for Cross-modal Matching. In *Proceedings of Advances in Neural Information Processing Systems.* 29406–29419.

[18] Audun Jsang. 2018. *Subjective Logic: A formalism for reasoning under uncertainty.* Springer Publishing Company, Incorporated.

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision* 123, 1 (2016), 32–73.

[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.* 201–216.

[21] Junnan Li, Richard Socher, and Steven C. H. Hoi. 2020. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In *Proceedings of 8th International Conference on Learning Representations.*

[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] Mengke Li, Yiu-Ming Cheung, and Yang Lu. 2022. Long-tailed Visual Recognition via Gaussian Clouded Logit Adjustment. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 6919–6928.

[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 *Proceedings of the European Conference on Computer Vision.* 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] Xin Liu, Yi He, Yiu-Ming Cheung, Xing Xu, and Nannan Wang. 2024. Learning Relationship-Enhanced Semantic Graph for Fine-Grained Image–Text Matching. *IEEE Transactions on Cybernetics* 54, 2 (2024), 948–961.

[27] Xin Liu, Zhikai Hu, Haibin Ling, and Yiu-Ming Cheung. 2021. MTFH: A Matrix Tri-Factorization Hashing Framework for Efficient Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 3 (2021), 964–981.

[28] Zhanyu Ma and Arne Leijon. 2011. Bayesian estimation of beta mixture models with variational inference. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 11 (2011), 2160–2173.

[29] David JC MacKay. 1992. A practical Bayesian framework for backpropagation networks. *Neural Computation* 4, 3 (1992), 448–472.

[30] Zhengxin Pan, Fangyu Wu, and Bailing Zhang. 2023. Fine-Grained Image-Text Matching by Cross-Modal Hard Aligning Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 19275–19284.

[31] Shu-Juan Peng, Yi He, Xin Liu, Yiu-ming Cheung, Xing Xu, and Zhen Cui. 2024. Relation-Aggregated Cross-Graph Correlation Learning for Fine-Grained Image–Text Retrieval. *IEEE Transactions on Neural Networks and Learning Systems* 35, 2 (2024), 2194–2207.

[32] Haim Permuter, Joseph Francos, and Ian Jermyn. 2006. A study of Gaussian mixture models of color and texture features for image classification and segmentation. *Pattern Recognition* 39, 4 (2006), 695–706.

[33] 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.* 4948–4956.

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning.* 8748–8763.

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2017. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 6 (2017), 1137–1149.

[36] Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing* 45, 11 (1997), 2673–2681.

[37] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential Deep Learning to Quantify Classification Uncertainty. In *Advances in Neural Information Processing Systems.*

[38] 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.

[39] Kai Wang Yang You Hongxun Yao Tongliang Liu Min Xu Shuo Yang, xu Zhao Pan. 2023. BiCro: Noisy Correspondence Rectification for Multi-modality Data via Bi-directional Cross-modal Similarity Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 19883–19892.

[40] Jinyu Yang, Jiali Duan, Son Tran, Yi Xu, Sampath Chanda, Liqun Chen, Belinda Zeng, Trishul Chilimbi, and Junzhou Huang. 2022. Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 15671–15680.

[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] Tan Yu, Yi Yang, Yi Li, Lin Liu, Hongliang Fei, and Ping Li. 2021. Heterogeneous attention network for effective and efficient cross-modal retrieval. In *Proceedings of the 44th international ACM SIGIR Conference on Research and Development in Information Retrieval.* 1146–1156.

[43] Zhixiong Zeng, Shuai Wang, Nan Xu, and Wenji Mao. 2021. Pan: Prototype-based adaptive network for robust cross-modal retrieval. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1125–1134.

[44] Kun Zhang, Zhendong Mao, Quan Wang, and Yongdong Zhang. 2022. Negative-Aware Attention Framework for Image-Text Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 15661–15670.

[45] Lei Zhang, Min Yang, Chengming Li, and Ruifeng Xu. 2022. Image-Text Retrieval via Contrastive Learning with Auxiliary Generative Features and Support-set Regularization. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.* 1938–1943.