This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3247939

Cross-Modal Retrieval with Partially Mismatched Pairs

Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, Xi Peng

Abstract—In this paper, we study a challenging but less-touched problem in cross-modal retrieval, *i.e.*, partially mismatched pairs (PMPs). Specifically, in real-world scenarios, a huge number of multimedia data (*e.g.*, the Conceptual Captions dataset) are collected from the Internet, and thus it is inevitable to wrongly treat some irrelevant cross-modal pairs as matched. Undoubtedly, such a PMP problem will remarkably degrade the cross-modal retrieval performance. To tackle this problem, we derive a unified theoretical Robust Cross-modal Learning framework (RCL) with an unbiased estimator of the cross-modal retrieval risk, which aims to endow the cross-modal retrieval methods with robustness against PMPs. In detail, our RCL adopts a novel complementary contrastive learning paradigm to address the following two challenges, *i.e.*, the overfitting and underfitting issues. On the one hand, our method only utilizes the negative information which is much less likely false compared with the positive information, thus avoiding the overfitting issue to PMPs. However, these robust strategies could induce underfitting issues, thus making training models more difficult. On the other hand, to address the underfitting issue brought by weak supervision, we present to leverage of all available negative pairs to enhance the supervision contained in the negative information. Moreover, to further improve the performance, we propose to minimize the upper bounds of the risk to pay more attention to hard samples. To verify the effectiveness and robustness of the proposed method, we carry out comprehensive experiments on five widely-used benchmark datasets compared with nine state-of-the-art approaches w.r.t. the image-text and video-text retrieval tasks.

Index Terms—Cross-modal retrieval, mismatched pairs, complementary contrastive learning.

1 Introduction

For a given query of one modality, cross-modal retrieval aims at retrieving the relevant instances from another modality, which has attracted considerable attention from academic and industrial communities [1], [2], [3], [4], [5]. In recent, a large number of approaches have been proposed in the decades, which could be roughly classified into the category of representation learning [2], [3], [6], [7], and similarity learning [4], [5]. Although these methods have achieved promising performance, their success heavily relies on the well-matched cross-modal pairs. In real-world applications, it is extremely expensive and even impossible to collect such clean data [8]. Hence, is it possible to explore an economic way to solve this problem? In this paper, we attempt to answer and address this practical question.

To alleviate the labor-intensive costs in labeling, one possible way is to collect co-occurrent cross-modal pairs from the Internet [8], [9]. For example, an image and its surrounding textual description on the web page could be regarded as an image-text pair in nature. Although such a data collection approach is economic, it will inevitably introduce a lot of mismatched pairs even with rigorous filtering and post-processing steps [10]. To be specific, some irrelevant cross-modal samples will be wrongly treated as the relevant pairs, which will undoubtedly degrade the performance of cross-modal retrieval. Such a PMP problem is less touched so far, to the best of our knowledge.

The most similar paradigm to PMPs might be learning with noisy labels. To eliminate the influence of noisy labels,

Peng Hu, Zhenyu Huang, Dezhong Peng, Xu Wang, and Xi Peng are with the College of Computer Science, Sichuan University, Chengdu 610065, China. Corresponding author: Xi Peng (email: pengx.gm@gmail.com).

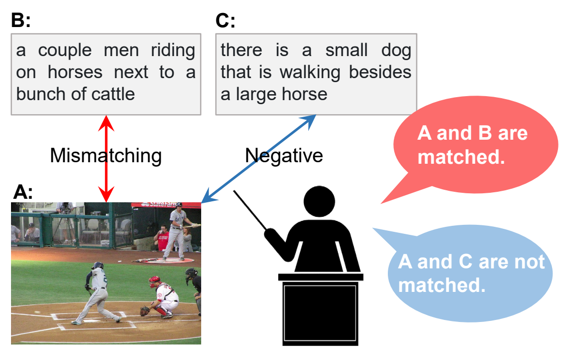


Fig. 1: A toy example to illustrate our idea. Different from Positive Learning (PL) paradigm, our Complementary Contrastive Learning (CCL) solution utilizes negative (see blue balloon) instead of positive (see red balloon) information, thus embracing the robustness against PMPs.

a large number of approaches have been proposed in past years, such as correction methods [11], [12], adaptive training strategies [13], [14], [15], [16], semi-supervised learning paradigms [17], [18], [19], robust loss functions [20], [21], etc. Although these methods have achieved great success in numerous applications, they are always specifically designed for the scenarios of unimodal classification, which cannot handle the multimodal data focused on in this paper. In addition, more distinctively, these studies consider the errors in the category annotation of a given sample, whereas the PMPs focus on the mismatching errors of two associated samples across different modalities. To transform cross-modal retrieval to cross-modal classification, each sample should be compared with all training samples across different modalities. This will remarkably increase computational and storage complexity, and may even be infeasible for complex models and large datasets. Therefore, to solve the PMP problem, one has to simultaneously consider noisy supervision, large "category" size, and cross-modal discrepancy, thus remarkably making the difficulty in cross-modal model optimization.

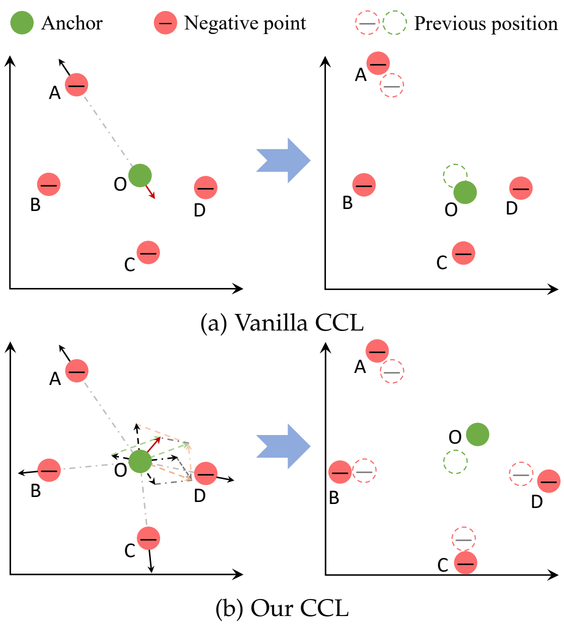


Fig. 2: A toy example to show the challenge of negative learning (NL, a.k.a. complementary learning) for cross-modal retrieval. (a) shows that traditional complementary learning cannot obtain the correct optimization direction, which makes the anchor "O" apart from "A" but close to "C" and "D", because the complementary label is less informative than the ordinary one. In addition, the anchor will suffer from the instability issue as it will only affect by a single negative point at any instant, acting like Brownian motion. More specifically, when the particle is very fine in flowing fluid, there are only a few molecules around to interact with it, thus the random interaction will produce an imbalance force to perturb the particle movement. (b) illustrates that the resultant of all negative information could provide a strong and correct optimization direction, thus helping our method to converge. More intuitively, for larger particles, there are much more molecules all around to interact with them, and thus the interaction forces from all directions will cancel out the inter randomness and produce the correct resultant force along the flowing direction.

To tackle the PMP problem, we propose a general Robust Cross-modal Learning framework (RCL) to learn similarities for cross-modal retrieval as shown in Fig. 3. In brief, RCL achieves cross-modal instance-level retrieval by using a Cross-Modal Contrastive Learning module (CMCL). Due to the existence of PMPs, vanilla contrastive learning (CL) aims to learn common representations by maximizing the mutual information between positive pairs, which would overfit the wrong supervision and thus lead to inaccurate predictions. To tackle this problem, we derive a Complementary Contrastive Learning paradigm (CCL) with an unbiased estimator of the retrieval risk using negative information to enhance the reliability of the supervision. More specifically, different from traditional CL paradigm [2], [3], [22], [23], our CCL paradigm exploits negative (complementary) instead of positive information to train neural networks,*e.g.*, "A and C are not matched" as shown in Fig. 1. Clearly, the complementary information is much more unlikely to provide the false ground truth compared with the positive information, thus avoiding overfitting to false supervision. For example, assuming both the noise and pair selection follow the uniform distribution, then the selected N pairs will consist of one positive pair and $N-1$ negative pairs for a given sample. Let the visual sample $\mathbf{V}$ be wrongly labeled as matching to a textual sample $\mathbf{T}$ in p probability. Hence, one could obtain that $\mathbf{V}$ and $\mathbf{T}$ are correctly labeled as unmatched in $1-\frac{p}{(N-1)^2}\approx 1$ probability. In other words, the correction probability of complementary information is remarkably larger than that of a positive one, *i.e.*, $1-\frac{p}{(N-1)^2} > 1-p$.

In practice, however, it is non-trivial and non-straightforward to employ complementary learning (a.k.a. negative learning) [17], [24], [25] for cross-modal retrieval, especially, in the presence of PMPs. To be specific, almost all existing works mainly study complementary learning in the scenario of classification, and it is still unclear how to exploit its potential in retrieval. Based on the discussion mentioned above, it is hard even impossible to convert cross-modal retrieval into cross-modal classification due to the high computation costs. In addition, once complementary learning is applied to retrieval, the model would underfit the latent data distribution and thus suffer from the convergence issue. In detail, the standard complementary learning will only push away a few selected negative pairs. As a result, the existence of the other massive negative samples will make it difficult in converging. It should be pointed out that, although some complementary learning studies [17], [26] have been conducted to solve the underfitting problem in classification, the proposed strategy is infeasible for the retrieval scenario due to two facts. On the one hand, the convergence of the retrieval models deteriorates more seriously than the classification models with complementary learning. On the other hand, it will take an over-expensive computational cost which is proportional to the number of instances.

Interestingly, the above instability issue is much similar to the motion of particles in slowly flowing fluid [27]. Namely, the large particles will more stably move along the flowing direction compared with the fine particles. To be specific, in the flowing fluid, the liquid molecules have two moving directions: the flowing direction and the random direction of thermal motion. From the view of microscopic particles, for a given very fine particle, it will be only affected by the random interaction of a few molecules at any instant. As a result, a large enough net resultant force will be easily produced to bring the particle towards a random direction, *i.e.*, leading to Brownian motion. In contrast, for a larger particle, there are more molecules around it to produce random interaction forces from all directions to cancel out the randomness, thus leading to the net resultant force in the correct direction, *i.e.*, the direction of fluid flowing [27] as shown Fig. 2. To summarize, more participants will alleviate the randomness brought by different ones, thus enabling large particles to have greater stability. Motivated by the aforementioned relative stability of large particles, we propose directly increasing the number of participants to improve the stability of complementary contrastive learning, *i.e.*, leveraging all available negative pairs to alleviate the randomness caused by few participants in the vanilla complementary learning as shown in Fig. 2. Moreover, to tackle the underfitting issue faced by the estimated risk, we propose to minimize the upper bounds of the risk to pay more attention to hard samples.

The main contributions and novelties of this work could be summarized as follows:

- We derive a general Robust Cross-modal Learning framework (RCL) which is specifically designed to solve the less-touched PMP problem in cross-modal retrieval. The proposed method employs a contrastive learning module (*i.e.*, CMCL) to formulate cross-modal retrieval as an N-way retrieval and a novel complementary learning approach (*i.e.*, CCL) to alleviate the overfitting issue faced by CMCL.
- To address the underfitting issue faced by the vanilla complementary learning methods, CCL employs all available instead of single negative information to achieve convergence, inspired by Brownian motion. Moreover, we propose to minimize the upper bounds of the estimated risk to further alleviate the underfitting problem. Therefore, that makes it possible to apply complementary learning to retrieval.
- To demonstrate the effectiveness of the proposed method, we conducted extensive experiments on three benchmark datasets (MS-COCO, Flickr30K, and CC152K) for image-text matching, and two benchmark datasets (MSVD and MSR-VTT) for video-text retrieval. The experimental results empirically verify that our RCL can boost the existing cross-modal methods by remarkable margins, especially under large mismatching rates.

2 Related Works

In this section, we will briefly review some related works on cross-modal retrieval and noisy label learning.

2.1 Cross-modal Retrieval

Cross-modal retrieval attempts to retrieve the relevant instances from different modalities for a given query, wherein the key is to measure the cross-modal similarity. During decades, a variety of cross-modal retrieval methods have been proposed by resorting to different approaches, *e.g.*, representation learning [2], [3], [7], [28] and similarity learning [4], [5]. More specifically, cross-modal representation learning methods [29], [30] aim at projecting different modalities into a latent common space wherein the representations of distinct modalities can be directly compared to calculate the similarities w.r.t. a distance metric, such as cosine similarity, Euclidean distance, and so on. To exploit existing knowledge in pre-trained embeddings, [6] proposed a Collaborative Experts model (CE) which aggregates the "general" and "specific" information from different pre-trained experts for video-text retrieval. To encode videos and texts into dense representations, [7] proposed a concept-free Dual deep Encoding network (DE). To achieve video-corpus moment retrieval, [31] presents a Retrieval and Localization Network with Contrastive Learning (ReLoCLNet) by maximizing the mutual information between query and candidates at both video- and frame-level. To exploit fine-grained information to improve the discrimination, [2] proposed a Stacked Cross Attention Network method (SCAN) to excavate the full latent object-word alignments between image regions and words. Like [2], [3] proposed Visual Semantic Reasoning Network (VSRN) to enhance visual representations for capturing the key objects and semantic concepts of a scene via region relationship reasoning and global semantic reasoning. To conduct fine-grained video-text retrieval, [32] proposed a Hierarchical Graph Reasoning (HGR) model by performing video-text matching into three hierarchical semantic levels to simultaneously capture global events, local actions, and entities respectively. Although these cross-modal representation learning methods could achieve good performance, the handcrafted similarity may further hinder performance improvements. To overcome such a limitation, some works attempt to learn parametric similarity functions in a data-driven way [4], [5], [33]. In brief, [4] presented a Graph Structured Matching Network (GSMN) to learn the fine-grained correspondence via both node-level matching and structure-level matching. In [5], a novel Similarity Graph Reasoning and Attention Filtration (SGRAF) network is proposed to capture the global- and local-region alignments between images and texts, which consists of a Graph Convolution Neural Network (GCNN) and a Similarity Attention Filtration (SAF) module.

Different from these prior arts that assume the data is with well-established pairs, this study aims to find a solution for PMPs that are less touched before. As the false positive pairs will be inevitably introduced when the data is collected from the Internet, it is reasonable to believe that this study could provide some novel insights to the community of cross-modal retrieval.

2.2 Learning with Noisy Labels

To alleviate or even eliminate the influence of the errors in annotations, a number of works have been carried out during past years [12], [21], [34], [35]. In the scenario of classification, existing methods on noisy labels could be divided into the following groups. The first group is the correction paradigm which alleviates the noisy labels by rectifying the incorrect annotations or the corresponding loss [11], [12], [36]. The major limitation of these methods is that the extra inputs are required to support the correction process, such as the noise transition matrix [37], [38] or some extra clean data [21], [34], [36], [39]. The second group of methods usually elaborately designs some training strategies to automatically adapt the incorrect labels for robust learning, such as MentorNet [14], [40] and Co-teaching [13]. The third group of methods resorts to a variety of approaches to distinguish the correct labels from the noisy ones so that the latter could be discarded or rectified [17], [18], [19], [41]. Different

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3247939

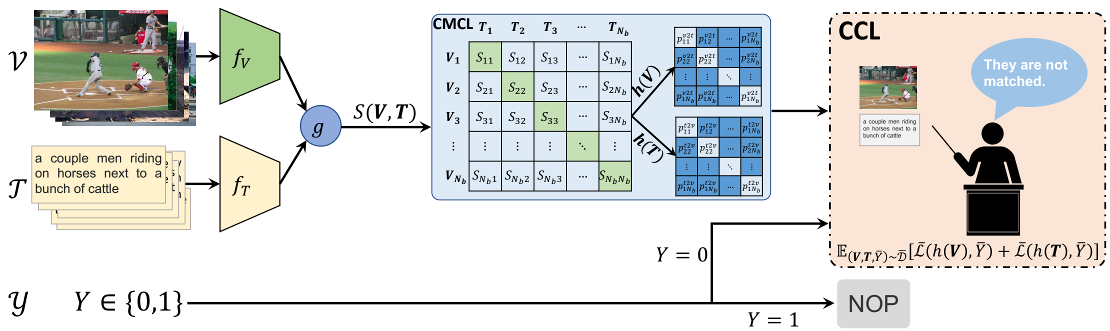


Fig. 3: The framework of the proposed method. First, the visual and textual samples are fed into the corresponding modality-specific networks f_V and f_T to extract the features $f_V(\mathbf{V})$ and $f_T(\mathbf{T})$, respectively. Second, a nonparametric or parametric function g is conducted on the features to measure the cross-modal similarity between $\mathbf{V}$ and $\mathbf{T}$. Then, our Cross-Modal Contrastive Learning module (CMCL) is adopted to compute the cross-modal matching probability, where $\sigma(\cdot)$ is the softmax function. As the mismatched pairs will lead to inaccurate probability prediction, we propose a novel Complementary Contrastive Learning (CCL) loss to solve this problem by only using the negative information ($Y = 0$) to optimize our model. For positive information ($Y = 1$), our CCL will do nothing operation (NOP). Thanks to our complementary contrastive learning paradigm, the proposed method could be robust against PMPs because the negative information is less possible to be false than the positive one.

from the above three kinds of methods, the fourth group of methods usually designs different loss functions which are robust against the noisy labels, such as Mean Absolute Error (MAE) [20], Generalized Cross-Entropy (GCE) [42], Normalization [21], etc.

Although the aforementioned methods have achieved huge success, almost all of them mainly focus on the errors in category-level annotations, while ignoring the instance-level mismatched pairs. In other words, they are specifically designed for classification and cannot be applied for cross-modal retrieval. In addition, it will cost too large memory and computational costs to convert cross-modal retrieval to cross-modal classification, and even be impossible for complex models [2], [4], [5] and strategies [18], [19]. To tackle instance-level errors, recently, Huang *et al.* proposed a Noisy Correspondence Rectifier method (NCR) to adaptively predict the confidence coefficient of cross-modal correspondence to divide the data into clean and noisy partitions in a co-teaching manner [43]. However, NCR needs to simultaneously train two individual networks in the manner of co-teaching, which will introduce extra training overhead. Moreover, it is difficult to correctly predict the confidence coefficient of cross-modal correspondence from PMPs, especially with a high mismatching rate.

3 The Proposed Method

In this section, we will elaborate on the influence of the PMP problem in cross-modal retrieval, and then detail the proposed RCL which consists of CMCL and CCL. More specifically, Section 3.1 will first present the problem formulation of image-text matching in presence of PMPs. After that, Section 3.2 introduces the proposed cross-modal contrastive learning module and Section 3.3 presents the details of our complementary contrastive learning loss.

3.1 Problem Formulation

Cross-modal retrieval aims at retrieving the relevant instances across different modalities for a given query. Formally, take the visual-text retrieval as an example, given a visual-text dataset $\mathcal{D} = \{\mathcal{V}, \mathcal{T}, \mathcal{Y}\}$ with partially mismatching pairs, we use $\mathcal{V} = \{\mathbf{V}_j\}_{j=1}^{N_v}$ to denote the visual training set with N_v visual samples, $\mathcal{T} = \{\mathbf{T}_j\}_{j=1}^{N_t}$ to denote the textual training set with N_t text samples, $\mathcal{V}_j$ and $\mathcal{T}_j$ to represent the j-th visual and textual samples, respectively; In addition, we use the binary set $\mathcal{Y} = \{Y_{jk} | j = 1, 2, \cdots, N_v; k = 1, 2, \cdots, N_t\}$ to indicate whether the corresponding image-text pairs are matched or not, *i.e.*, $Y_{jk} = 1$ if $\mathbf{V}_j$ and $\mathbf{T}_k$ are matched, and 0 otherwise for the visual sample $\mathbf{V}_j$ and the textual sample $\mathbf{T}_k$. As data collection would mistakenly treat some negative pairs as positive, we aim to search the most relevant samples from the textual/visual modality for a given visual/textual query while being immune to the influence of these false positive pairs or so-called partially mismatched pairs.

3.2 Cross-modal Contrastive Learning

The key to cross-modal retrieval is measuring the similarity between different modalities. To this end, most existing methods attempt to learn two modality-specific networks $f_V(\cdot, \Theta_I)$ and $f_T(\cdot, \Theta_T)$ to project the corresponding visual and textual modalities into a latent shared space, where Θ_I and Θ_T are the parameterized models for visual and textual modalities, respectively. In the latent space, there exists a mapping function $S_{jk} = g(f_V(\mathbf{V}_j), f_T(\mathbf{T}_k), \Theta_g)$ to measure the similarity between the visual feature $f_V(\mathbf{V}_j)$ and textual feature $f_T(\mathbf{T}_k)$, where Θ_g is the parameters of the similarity function g. Note that, g could be a nonparametric [2], [3] or parametric function [4], [5]. With the output of these networks, one could obtain retrieval results by simply ranking the computed cross-modal similarities.

Inspired by contrastive learning, we formulate the cross-modal retrieval objective as an N-way retrieval using the softmax criterion. The decision function is N_t-way searcher $h : \mathcal{V} \xrightarrow{\mathcal{T}} \mathbb{R}^{N_t}$ for visual modality, similarly $h : \mathcal{T} \xrightarrow{\mathcal{V}} \mathbb{R}^{N_v}$ for textual modality. Therefore, the cross-modal matching

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3247939

probability of the textual sample $\mathbf{T}_j$ w.r.t. the visual query $\mathbf{V}_i$ could be calculated by:

$$p_{ij}^{v2t} = p(Y_{ij} = 1|\mathbf{V}_i, \mathbf{T}_j) = h(\mathbf{V}_i, \mathbf{T}_j) = \frac{e^{\frac{S_{ij}}{\tau}}}{\sum_{l=1}^{N_t} e^{\frac{S_{il}}{\tau}}}, \quad (1)$$

where τ is a temperature parameter [22], [44], and $h(\mathbf{V}_i, \mathbf{T}_j)$ is the j-th element of $h(\mathbf{V}_i)$. Similarly, the matching probability of the visual query $\mathbf{V}_i$ w.r.t. the textual sample $\mathbf{T}_j$ is obtained by:

$$p_{ij}^{t2v} = p(Y_{ji} = 1|\mathbf{T}_i, \mathbf{V}_j) = h(\mathbf{T}_i, \mathbf{V}_j) = \frac{e^{\frac{S_{ji}}{\tau}}}{\sum_{l=1}^{N_v} e^{\frac{S_{li}}{\tau}}}, \quad (2)$$

where $h(\mathbf{T}_i, \mathbf{V}_j)$ is the j-th element of $h(\mathbf{T}_i)$. However, it is expensive to compute the decision function h on the whole training set. Following [22], [23], we explore Monte Carlo approximation to estimate the softmax criterion by:

$$Z_i \backsimeq N_t \mathbb{E}_{\mathbf{T}_j \sim \mathcal{D}} \left[e^{\frac{S_{ij}}{\tau}} \right] = \frac{N_t}{N} \sum_{k=1}^{N} e^{\frac{S_{ij_k}}{\tau}}, \quad (3)$$

where $Z_i = \sum_{l=1}^{N_t} e^{\frac{S_{il}}{\tau}}$, $\{j_k\}_{k=1}^{N}$ are random indices sampling a subset from the training set, and N could be the size of a mini-batch. Thus, the cross-modal decision function h could be estimated by:

$$h(\mathbf{V}_i, \mathbf{T}_j) \backsimeq \frac{e^{\frac{S_{ij}}{\tau}}}{\frac{N_t}{N} \sum_{k=1}^{N} e^{\frac{S_{ij_k}}{\tau}}}. \quad (4)$$

Similarly, h could be estimated by:

$$h(\mathbf{T}_i, \mathbf{V}_j) \backsimeq \frac{e^{\frac{S_{ji}}{\tau}}}{\frac{N_t}{N} \sum_{k=1}^{N} e^{\frac{S_{j_k i}}{\tau}}}. \quad (5)$$

From the above, one could see that the goal of cross-modal retrieval is learning the projection functions f_V, f_T, and g to separate the positive and negative pairs well. The cross-modal retrieval aims to learn a model that minimizes the risk of decision function h:

$$\begin{aligned} R(h) := & \, \mathbb{E}_{(\mathbf{V}_i, \mathbf{Y}_{i\cdot}) \sim \mathcal{D}} \left[\mathcal{L}(h(\mathbf{V}_i), \mathbf{Y}_{i\cdot}) \right] \\ & + \mathbb{E}_{(\mathbf{T}_i, \mathbf{Y}_{\cdot i}) \sim \mathcal{D}} \left[\mathcal{L}(h(\mathbf{T}_i), \mathbf{Y}_{\cdot i}) \right], \end{aligned} \quad (6)$$

where $\mathbb{E}(\cdot)$ is the expectation operator, and $\mathcal{L}(\cdot, \cdot)$ is a loss function. Given cross-modal pairs $\mathcal{D} = \{\mathbf{V}_i, \mathbf{T}_i, Y_i\}_{i=1}^{N}$, like Equations (4) and (5) the risk could be approximated by:

$$\hat{R}(h, \mathcal{L}) \backsimeq \frac{1}{N} \sum_{i=1}^{N} \left[\mathcal{L}(h(\mathbf{V}_i), \mathbf{Y}_{i\cdot}) + \mathcal{L}(h(\mathbf{T}_i), \mathbf{Y}_{\cdot i}) \right]. \quad (7)$$

As in the usual classification case, some well-known loss functions could be utilized to optimize the cross-modal models. Especially, for the cross-entropy loss function, the risk could be rewritten as:

$$\hat{R}(h) \backsimeq -\frac{1}{N} \left(\sum_{p \in \mathcal{P}_{+}^{v2t}} \log p + \sum_{p \in \mathcal{P}_{+}^{t2v}} \log p \right), \quad (8)$$

where $\mathcal{P}_{+}^{v2t} = \{p_{ij}^{v2t} | Y_{ij} = 1; i, j = 1, \cdots, N\}$ and $\mathcal{P}_{+}^{t2v} = \{p_{ij}^{t2v} | Y_{ji} = 1; i, j = 1, \cdots, N\}$ are the probability sets of positive image-query-text and text-query-image pairs, respectively. Obviously, Equation (8) is the contrastive loss function [22], [23], which could maximize the agreement between positive pairs while minimizing the mutual information between negative pairs.

It should be pointed out that our CMCL is remarkably different from the popular triplet losses [2], [3], [45] in the given aspects. To be specific, the triplet losses aim to enforce the similarity gaps between the positive pair and negative pair to be larger than a given margin, whereas CMCL aims at maximizing the similarity gap between the positive pair and negative pairs as large as possible. Such a difference will bring two benefits which are helpful in alleviating the overfitting of our model to the false positive. On the one hand, our method does not involve specifying the margin, thus avoiding the labor-intensive efforts for the parameter selection and the corresponding overfitting issue. On the other hand, unlike existing methods, we compute each term of the loss by using all instead of one specific negative sample for one given anchor (see Section 4.8 for more detailed discussions).

Such a difference could improve the robustness against mismatched pairs and thus alleviate the overfitting to the false positive pairs since the influence of the mismatched pairs will be weakened.

3.3 Complementary Contrastive Learning

Despite the benefits brought by CMCL, it will overfit the false positive pairs as shown in our ablation study (Section 4.8). Specifically, like cross-entropy loss functions [21], [23], [44], Equation (8) will focus on the optimization of the hard samples that will lead to a relatively large loss. As the false positive pairs will mislead Equation (8) to the wrong optimization direction, thus degrading the performance.

Inspired by complementary learning [24], we employ complementary instead of positive information to provide more accurate supervision. However, the complementary supervision is too weak to train the models, thus it will induce an underfitting problem as the aforementioned. Motivated by the Brownian motion, we employ multiple negatives to enhance the supervision information of complementary learning to address the problem. Our method is derived from the following theorem which allows the unbiased estimation of the retrieval risk from complementarily labeled patterns.

Theorem 1. *For any ordinary distribution $\mathcal{D}$ and complementary distribution $\overline{\mathcal{D}}$ related by Equation* (6) *with decision function h, and loss $\mathcal{L}$, we have*

$$\begin{aligned} R(h; \mathcal{L}) = \overline{R}(h; \overline{\mathcal{L}}) = \mathbb{E}_{(\mathbf{V}, \mathbf{T}, \overline{\mathbf{Y}}) \sim \overline{\mathcal{D}}} \Big[& \overline{\mathcal{L}}(h(\mathbf{V}), \overline{\mathbf{Y}}) \\ & + \overline{\mathcal{L}}(h(\mathbf{T}), \overline{\mathbf{Y}}) \Big], \end{aligned} \quad (9)$$

for the complementary loss

$$\begin{aligned} \overline{\mathcal{L}}(h(\mathbf{X}), \overline{\mathbf{Y}}) = & -\frac{N - |\overline{\mathbf{Y}}| - 1}{|\overline{\mathbf{Y}}|} \sum_{y \in \overline{\mathbf{Y}}} \mathcal{L}(h(\mathbf{X}), y) \\ & + \sum_{y \notin \mathcal{Y}} \mathcal{L}(h(\mathbf{X}), y) \end{aligned} \quad (10)$$

where $\mathbf{X} \in \{\mathbf{V}, \mathbf{T}\}$, $\overline{R}$ is the risk for complementary labels, $\overline{\mathcal{L}}$ is complementary loss, $\overline{\mathbf{Y}}$ is a set of complementary labels indicating

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3247939

multiple negatives, $\overline{Y}_{ij}=1$ indicates that the i-th visual and j-th textual samples are unmatched, and $|\overline{\mathbf{Y}}|$ is the size of the set.

The proof is provided in Appendix A. By using Theorem 1 and Equation (3), we could rewrite the retrieval risk as:

$$R(h;\mathcal{L})=\sum_{k=1}^{N}\overline{q}_k\mathbb{E}_{\overline{\mathbb{P}}_k}\left[\overline{\mathcal{L}}(h(\mathbf{V}),\mathbf{Y}_{\cdot k})+\overline{\mathcal{L}}(h(\mathbf{T}),\mathbf{Y}_{k\cdot})\right] \quad (11)$$

where $\overline{q}_k=P(\overline{Y}=k)$. Given the dataset with $\overline{\mathcal{D}}=\{(\mathbf{V}_i,\mathbf{T}_j,\overline{Y}_{ij})\}_{i,j=1}^{N}$, we could empirically estimate $\overline{q}_k$ by $\frac{|\overline{\mathcal{V}}_k|}{N}$, where $|\mathcal{X}|$ denotes the size of the set $\mathcal{X}$. With Equation (11), we can further obtain the following empirical approximation of the unbiased risk estimator introduced in Theorem 1:

$$\begin{aligned}\hat{R}(h;\mathcal{L})\backsimeq\frac{1}{N}\sum_{k=1}^{N}\Bigg(&\sum_{\mathbf{T}_i\in\overline{\mathcal{T}}_k}\overline{\mathcal{L}}(h(\mathbf{V}_k,\mathbf{T}_i),\overline{Y}_{ki})\\&+\sum_{\mathbf{V}_i\in\overline{\mathcal{V}}_k}\overline{\mathcal{L}}(h(\mathbf{T}_k,\mathbf{V}_i),\overline{Y}_{ik})\Bigg),\end{aligned} \quad (12)$$

where $\overline{\mathcal{V}}_k=\{\mathbf{V}_i|\overline{Y}_{ik}=1;i=1,\cdots,N\}$ and $\overline{\mathcal{T}}_k=\{\mathbf{T}_j|\overline{Y}_{kj}=1;j=1,\cdots,N\}$ denote the visual and textual samples labeled as unmatching with the k-th textual and visual ones, respectively. Inspired by [20], we employ the noise-tolerate Mean Absolute Error (MAE) to approximate the risk. Specifically, by utilizing MAE in Equation (12), we could obtain

$$\begin{aligned}\hat{R}(h;\mathcal{L})\backsimeq\alpha\sum_{k=1}^{N}\Bigg(&\sum_{\mathbf{T}_i\in\overline{\mathcal{T}}_k}h(\mathbf{V}_k,\mathbf{T}_i)\\&+\sum_{\mathbf{V}_i\in\overline{\mathcal{V}}_k}h(\mathbf{T}_k,\mathbf{V}_i)\Bigg)+Z\end{aligned} \quad (13)$$

where $\alpha=\frac{2(N-1)}{C}$, Z is a constant, and $C=|\overline{\mathcal{V}}_k|=|\overline{\mathcal{T}}_k|$ is the number of selected negatives. Minimizing Equation (13) is equivalent to minimizing the following loss function:

$$\mathcal{L}_{\text{mae}}=\frac{1}{N}\sum_{k=1}^{N}\sum_{p\in\overline{\mathcal{P}}_k}p, \quad (14)$$

where $\overline{\mathcal{P}}_k=\overline{\mathcal{P}}_k^{v2t}\cup\overline{\mathcal{P}}_k^{t2v}$, $\overline{\mathcal{P}}_k^{v2t}=\{p_{ki}^{v2t}|\overline{Y}_{ki}=1;i=1,\cdots,N\}$ and $\overline{\mathcal{P}}_k^{t2v}=\{p_{ik}^{t2v}|\overline{Y}_{ik}=1;i=1,\cdots,N\}$ are the probability sets of complementary image-query-text and text-query-image pairs, respectively. Equation (14) is theoretically robust against PMPs, whose proof is provided in Appendix B. However, one could see that $\mathcal{L}_{\text{mae}}$ equally treats each point to make it more robust against noisy labels. However, without focusing on more challenging samples, its noise-tolerate property would make the DNN models difficult to train on complicated datasets [42]. To address this problem, we formulate the inequations of $x\leqslant-\log(1-x)$, $x\leqslant e^{-(1-x)}$, $x\leqslant\frac{1}{q}(1-(1-x)^q)$, and $x\leqslant\tan(x)$ to transform Equation (14) as the following upper bounds of MAE. As a result, the model will focus more on the hard samples while preserving the robustness.

$$\mathcal{L}_{\log}=-\frac{1}{N}\sum_{k=1}^{N}\sum_{p\in\overline{\mathcal{P}}_k}\log(1-p), \quad (15)$$

$$\mathcal{L}_{\exp}=\frac{1}{N}\sum_{k=1}^{N}\sum_{p\in\overline{\mathcal{P}}_k}e^{-(1-p)}, \quad (16)$$

$$\mathcal{L}_{\text{gce}}=\frac{1}{N}\sum_{k=1}^{N}\sum_{p\in\overline{\mathcal{P}}_k}\frac{1}{q}\left(1-(1-p)^q\right), \quad (17)$$

$$\mathcal{L}_{\tan}=\frac{1}{N}\sum_{k=1}^{N}\sum_{p\in\overline{\mathcal{P}}_k}\tan(p), \quad (18)$$

where $q\in(0,1]$. By minimizing these complementary loss functions, we could achieve robust cross-modal retrieval. Specifically, Equation (15) is an instance-level variant of negative learning loss [17] with multiple negatives. Equation (17) is an instance-level complementary variant of Generalized Cross Entropy (GCE) [42]. The basic idea of the above objective functions is employing complementary information to alleviate the influence of mismatched pairs. In brief, complementary contrastive learning will specify an instance to which the given input does not belong.

One major advantage of complementary learning is that collecting the complementary labels would be less laborious than the ordinary labels because it is unnecessary to carefully seek the correct class from a long list of candidate classes. Although complementary learning could avoid the exhaustive accurate data annotation, it will suffer from the following limitations which hinder its application in cross-modal retrieval. First, the standard complementary learning is proposed for multi-class classification, and it is intractable or even infeasible to apply the idea to the retrieval task due to the significant difference between the two tasks. Second, although complementary learning shows potential in solving the PMP problem, simply using the idea will underfit the model to the latent correct distribution of data, thus making it difficult to converge. More specifically, on the one hand, the complementary labels are less informative than the positive ones, thus the convergence of the model is hard to guarantee under such weak supervision. On the other hand, almost all existing complementary learning methods usually construct a complementary label for a given sample, directly adopting the methods for retrieval will result in non-convergence of the model as elaborated in Sections 1 and 4.8.

To address the above instability issue, we propose to use all negative pairs available within the given batch as formulated in Equations (15)–(18), *i.e.*, $|\overline{\mathbf{Y}}|=N_b-1$, where N_b is the size of a mini-batch. The idea comes from the study on Brownian motion. In brief, if only one negative relationship is considered like the standard complementary learning, the anchors will be affected by the negatives in a random way, thus making it difficult in achieving convergence. By simultaneously considering all available negatives, in contrast, one could achieve a steady-state approximation. Notably, although our experimental results will empirically show the stability of such a dynamic system, it is daunting to prove its convergence in theory since the random motion of massive particles is involved in essence.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3247939

4 Experiments

In this section, to evaluate the effectiveness of the proposed method, we conduct extensive experiments with the comparisons of state-of-the-art methods w.r.t. two cross-modal retrieval tasks, *i.e.*, image-text matching, and video-text retrieval. For a comprehensive comparison, our experiments are conducted on three image-text and two video-text databases.

TABLE 1: General statistics of all datasets in the experiments. N_{tr}, N_{va}, and N_{te} are the number of training, validation, and testing sets in the corresponding dataset, respectively.

Dataset	Modality	N_{tr}	N_{va}	N_{te}
MS-COCO	Image	113,287	5,000	5,000
	Text	566,435	25,000	25,000
Flickr30K	Image	29,000	1,000	1,000
	Text	145,000	5,000	5,000
CC152K	Image	150,000	1,000	1,000
	Text	150,000	1,000	1,000
MSVD	Video	1,200	100	670
	Text	48,774	8,100	54,270
MSR-VTT	Video	6,513	497	2,990
	Text	130,260	9,940	59,800

4.1 Datasets

In this section, we will briefly introduce the used five datasets, *i.e.*, MS-COCO [48], Flickr30K [49], CC152K [10], MSVD [50], and MSR-VTT [51]. For clarity, we summarize some statistics of these datasets in Table 1. In brief,

- **MS-COCO [48]** is a large-scale cross-modal dataset, which consists of 123,287 images each of which is described by five sentences. Following [2], in our experiments, the training set consists of $113,287$ images and $566,435$ sentences, the validation set contains $5,000$ images and $25,000$ sentences, and the testing set consists of $5,000$ images and $25,000$ sentences.
- **Flickr30K [49]** consists of 31,000 images with five text annotations for each image. Like MS-COCO, we also use the default splits of [2], *i.e.*, the training set includes 29,000 images and 145,000 texts, the validation set contains 1,000 images and 5,000 texts, and the testing set consists of 1,000 images and 5,000 texts.
- **CC152K [10]** is a subset of Conceptual Captions [10] that comprises 3.3M image-text pairs wherein each image is crawled from the Internet with a text description. In our experiments, we randomly select 150,000, 1,000, and 1,000 pairs from the training, validation, and testing sets.
- **MSVD [50]** comprises 1,970 videos sourced from YouTube, and each video is captioned by around 40 sentences/tags (80,000 English text descriptions in total). In our evaluations, the standard partitions used in [6] are adopted, *i.e.*, 1,200 videos for training, 100 videos for validation, and 670 videos for testing.
- **MSR-VTT [51]** is a large-scale video-caption dataset, which contains about 200,000 unique video-caption pairs including 10,000 web video clips and 200,000 texts. In the dataset, each video is captioned with 20 different description sentences. We use the official data partitions for experiments, *i.e.*, 6,513 videos for training, 497 videos for validation, and the remaining 2,990 videos for testing.

4.2 Experiment Settings

Robust Cross-modal Learning (RCL) is a general framework that could extend most of the existing cross-modal matching approaches to enjoy robustness against PMPs by simply replacing the triplet loss with our loss. To demonstrate the effectiveness and generalization of RCL, we apply it to seven different cross-modal retrieval methods (*i.e.*, VSRN [3], GSMN [4], IMRAM (text) [47], SAF [5], SGR [5], DE [7], and CE [6]). Specifically, the visual regions/frames and sentences are fed into the visual network $f_V(\cdot, \Theta_V)$ and the textual network $f_T(\cdot, \Theta_T)$, respectively. To calculate the cross-modal similarities, the similarity function $g(f_V(\mathbf{V}), f_T(\mathbf{T}), \Theta_g)$ is adopted to measure the similarity score between visual feature $f_V(\mathbf{V})$ and textual feature $f_T(\mathbf{T})$, where g could be nonparametric or parametric. For fair comparisons, our variants adopt the same network structure and setting as the original methods. The temperature τ is set as 0.07 [22]. For convenience, our method uses $\mathcal{L}_{\log}$ unless otherwise specified.

Besides the comparisons with the above seven methods, we also investigate the performance of SCAN (i-t AVG) [2] and PolyLoss [46] as baselines. For a comprehensive performance evaluation, we adopt Recall@K (R@K, higher is better) for different values of K and Median rank (Med r, lower is better) to measure the performance for cross-modal retrieval. In brief, R@K is the percentage of tested queries for which at least one correct item is among the top K ranking results [46], [52]. Med r is the median rank of the first correct item in the retrieved results [7]. Following [4], [5], we report the corresponding results on the testing set when the model achieves the best performance on the validation set in terms of the sum of the evaluation scores.

4.3 Comparisons with State of the Arts

In this section, we conduct comparisons with nine cross-modal retrieval approaches on five benchmark datasets to verify the effectiveness of the proposed method. To comprehensively investigate the robustness of our RCL against PMPs, we carry out experiment under four different settings with the synthesized false positive pairs on MS-COCO [48], Flickr30K [49], MSVD [50], and MSR-VTT [51], *i.e.*, the mismatching rates increases from 0.2 to 0.8 with an interval of 0.2. To be specific, we randomly select a given proportion of visual samples and then randomly permute their all textual counterparts, which is more challenging than the noise injection approach used in [43]. In brief, in [43], although one image may have mismatched texts, it is still likely to have some correctly matched texts, which will lead to semantic leaking, *i.e.*, the vast majority of images still have one or more correctly matched texts with similar semantics, especially for MS-COCO and Flickr30. However, in real-world applications, if the images are inserted into texts incorrectly, all the surrounding texts will be mismatched with the images, *e.g.*, the Conceptual Captions dataset. Therefore,

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3247939

TABLE 2: Image-text matching with different mismatching rates (MRate) on MS-COCO 1K and Flickr30K.

MRate	Method	MS-COCO							Flickr30K						
		Image-to-Text			Text-to-Image			rSum	Image-to-Text			Text-to-Image			rSum
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
0.2	SCAN [2]	62.2	90.0	96.1	46.2	80.8	89.2	464.5	58.5	81.0	90.8	35.5	65.0	75.2	406.0
	PolyLoss [46]	68.4	92.3	96.9	44.4	79.2	88.2	469.4	58.1	83.8	90.6	39.6	68.3	78.3	418.7
	VSRN [3]	61.8	87.3	92.9	50.0	80.3	88.3	460.6	33.4	59.5	71.3	25.0	47.6	58.6	295.4
	GSMN [4]	65.8	91.7	96.6	51.6	83.0	88.8	477.5	54.6	81.2	87.8	32.2	61.0	71.4	388.2
	IMRAM [47]	69.9	93.6	97.4	55.9	84.4	89.6	490.8	59.1	85.4	91.9	44.5	71.4	79.4	431.7
	SAF [5]	71.5	94.0	97.5	57.8	86.4	91.9	499.1	62.8	88.7	93.9	49.7	73.6	78.0	446.7
	SGR [5]	25.7	58.8	75.1	23.5	58.9	75.1	317.1	55.9	81.5	88.9	40.2	66.8	75.3	408.6
	RCL-VSRN	70.8	93.4	**97.6**	57.2	**86.9**	**93.7**	**499.6**	59.6	83.7	89.7	44.2	72.9	**81.6**	431.7
	RCL-GSMN	**76.8**	**95.2**	**98.2**	60.4	**87.1**	**92.4**	**510.1**	**66.6**	87.5	92.4	45.9	73.6	**81.4**	**447.4**
	RCL-IMRAM	**74.1**	**94.9**	**97.9**	**58.9**	**86.4**	**92.6**	**504.8**	**64.0**	**89.5**	**94.7**	**45.9**	**73.8**	**82.5**	**450.4**
	RCL-SAF	**77.1**	**95.5**	**98.2**	**61.0**	**88.8**	**94.6**	**515.2**	**72.0**	**91.7**	**95.8**	**53.6**	**79.9**	**86.7**	**479.7**
	RCL-SGR	**77.0**	**95.5**	**98.1**	**61.3**	**88.8**	**94.8**	**515.5**	**74.2**	**91.8**	**96.9**	**55.6**	**81.2**	**87.5**	**487.2**
0.4	SCAN [2]	42.9	74.6	85.1	24.2	52.6	63.8	343.2	26.0	57.4	71.8	17.8	40.5	51.4	264.9
	PolyLoss [46]	40.4	75.3	85.9	31.1	64.7	77.9	323.0	30.4	61.7	73.3	19.7	44.0	55.6	284.7
	VSRN [3]	29.8	62.1	76.6	17.1	46.1	60.3	292.0	2.6	10.3	14.8	3.0	9.3	15.0	55.0
	GSMN [4]	18.3	43.3	55.0	13.0	39.4	54.9	223.9	31.0	62.0	74.1	19.7	44.3	56.3	287.4
	IMRAM [47]	51.8	82.4	90.9	38.4	70.3	78.9	412.7	44.9	73.2	82.6	31.6	56.3	65.6	354.2
	SAF [5]	13.5	43.8	48.2	16.0	39.0	50.8	211.3	7.4	19.6	26.7	4.4	12.0	17.0	87.1
	SGR [5]	1.3	3.7	6.3	0.5	2.5	4.1	18.4	4.1	16.6	24.1	4.1	13.2	19.7	81.8
	RCL-VSRN	**67.7**	**91.9**	**96.4**	**53.3**	**84.3**	**92.0**	**485.6**	**52.4**	**79.8**	**87.3**	**38.1**	**67.0**	**76.7**	**401.3**
	RCL-GSMN	**74.5**	**94.4**	**97.5**	**58.2**	**85.1**	**91.0**	**500.7**	**59.0**	**84.4**	**90.9**	**41.7**	**65.6**	**72.9**	**414.5**
	RCL-IMRAM	**73.7**	**94.5**	**97.9**	**56.8**	**83.8**	**89.8**	**496.5**	**59.2**	**84.8**	**91.9**	**42.2**	**70.6**	**80.0**	**428.7**
	RCL-SAF	**74.8**	**94.8**	**97.8**	**59.0**	**87.1**	**93.9**	**507.4**	**68.8**	**89.8**	**95.0**	**51.0**	**76.7**	**84.8**	**466.1**
	RCL-SGR	**73.9**	**94.9**	**97.9**	**59.0**	**87.4**	**93.9**	**507.0**	**71.3**	**91.1**	**95.3**	**51.4**	**78.0**	**85.2**	**472.3**
0.6	SCAN [2]	29.9	60.9	74.8	0.9	2.4	4.1	173.0	13.6	36.5	50.3	4.8	13.6	19.8	138.6
	PolyLoss [46]	31.3	66.5	78.7	22.1	49.3	59.7	307.6	18.0	42.0	55.5	3.4	9.9	15.1	143.9
	VSRN [3]	11.6	34.0	47.5	4.6	16.4	25.9	140.0	0.8	2.5	5.3	1.2	4.2	6.9	20.9
	GSMN [4]	4.7	14.7	20.4	2.9	9.9	14.3	66.9	0.0	0.4	0.9	0.1	0.5	1.0	2.9
	IMRAM [47]	18.2	51.6	68.0	17.9	43.6	54.6	253.9	16.4	38.2	50.9	7.5	19.2	25.3	157.5
	SAF [5]	0.1	0.5	0.7	0.8	3.5	6.3	11.9	0.1	1.5	2.8	0.4	1.2	2.3	8.3
	SGR [5]	0.1	0.6	1.0	0.1	0.5	1.1	3.4	1.5	6.6	9.6	0.3	2.3	4.2	24.5
	RCL-VSRN	**61.9**	**88.3**	**94.9**	**46.0**	**79.1**	**88.6**	**458.8**	**42.8**	**70.9**	**81.3**	**29.7**	**56.9**	**68.0**	**349.6**
	RCL-GSMN	**69.9**	**92.7**	**97.1**	**54.8**	**83.7**	**90.9**	**489.1**	**54.3**	**78.5**	**85.8**	**38.2**	**63.0**	**72.3**	**392.1**
	RCL-IMRAM	**68.3**	**92.0**	**96.5**	**53.8**	**82.3**	**89.6**	**482.5**	**53.9**	**80.4**	**87.6**	**37.5**	**64.8**	**74.0**	**398.2**
	RCL-SAF	**70.1**	**93.1**	**96.8**	**54.5**	**84.4**	**91.9**	**490.8**	**63.9**	**84.8**	**91.7**	**43.0**	**71.2**	**79.4**	**434.0**
	RCL-SGR	**71.4**	**93.2**	**97.1**	**55.4**	**84.7**	**92.3**	**494.1**	**62.3**	**86.3**	**92.9**	**45.1**	**71.3**	**80.2**	**438.1**
0.8	SCAN [2]	10.2	29.9	42.0	0.1	0.7	1.1	84.0	1.1	5.0	8.7	0.4	1.3	2.3	18.8
	PolyLoss [46]	11.2	33.5	48.3	0.1	0.6	1.9	95.6	2.2	8.8	13.0	0.1	0.7	1.8	26.6
	VSRN [3]	1.4	5.3	8.8	0.7	2.8	5.4	24.4	0.3	1.4	2.1	0.6	2.0	3.3	9.7
	GSMN [4]	1.5	5.9	10.7	1.5	5.9	10.0	35.5	0.1	0.5	0.8	0.1	0.5	1.0	3.0
	IMRAM [47]	1.3	5.0	8.3	0.2	0.6	1.3	16.7	3.1	9.7	5.2	0.3	0.9	1.9	31.1
	SAF [5]	0.2	0.8	1.4	0.1	0.5	1.0	4.0	0.0	0.8	1.2	0.1	0.5	1.1	3.7
	SGR [5]	0.2	0.6	1.0	0.1	0.5	1.0	3.4	0.2	0.3	0.5	0.1	0.6	1.0	2.7
	RCL-VSRN	**49.8**	**79.7**	**88.9**	**33.8**	**68.1**	**80.9**	**401.2**	**12.3**	**32.0**	**41.5**	**8.3**	**23.7**	**33.8**	**151.6**
	RCL-GSMN	**60.3**	**87.2**	**93.9**	**45.3**	**76.1**	**85.4**	**448.2**	**34.6**	**61.5**	**71.9**	**23.8**	**47.0**	**57.0**	**295.8**
	RCL-IMRAM	**60.1**	**86.6**	**93.3**	**44.6**	**73.9**	**82.9**	**441.4**	**39.5**	**66.3**	**76.0**	**26.7**	**52.1**	**62.2**	**322.8**
	RCL-SAF	**62.9**	**89.3**	**94.9**	**47.1**	**77.9**	**87.4**	**459.5**	**45.0**	**72.8**	**80.8**	**30.7**	**56.5**	**67.3**	**353.1**
	RCL-SGR	**63.2**	**89.3**	**95.2**	**47.6**	**78.7**	**88.0**	**462.0**	**47.1**	**70.5**	**79.4**	**30.3**	**56.1**	**66.3**	**349.7**

the PMPs studied in the paper are more challenging than noisy correspondence [43], which is demonstrated by the following experiments.

4.3.1 *Image-Text Matching with Synthesized Noises*

To verify the robustness of RCL against synthesized mismatched pairs, we carry out experiments on two image-text datasets, *i.e.*, Flickr30K, and MS-COCO. As shown in Table 2, one could see that RCL could remarkably improve the robustness of existing methods, and all extensions with RCL achieve promising performance on the two benchmark datasets. More specifically,

- The PMPs will corrupt the performance of the cross-modal matching modal. With more false positive pairs, the performance of all tested methods will be degraded.
- On the larger-size dataset (*i.e.*, MS-COCO) and the mismatching rate is small (*e.g.*, 20%), some of the baselines (*e.g.*, SAF) could achieve competitive performance, which could attribute to that massive training data would enhance the robustness of the model. However, with more false positives, simply increasing the amount of data cannot benefit stronger

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3247939

TABLE 3: Video-text retrieval with different mismatching rates (MRate) on MSVD and MSR-VTT.

MRate	Method	MSVD								MSR-VTT							
		Video-to-Text				Text-to-Video				Video-to-Text				Text-to-Video			
		R@1	R@5	R@10	Med r↓	R@1	R@5	R@10	Med r↓	R@1	R@5	R@10	Med r↓	R@1	R@5	R@10	Med r↓
0.2	DE [7]	7.4	23.5	34.0	28.0	10.3	21.9	28.4	50.0	4.9	15.4	23.0	69.0	0.3	1.4	2.7	1314.0
	RCL-DE	**10.4**	**30.0**	**42.2**	**16.0**	**11.2**	**27.0**	**35.4**	**33.0**	**6.6**	**19.3**	**28.5**	**40.0**	**0.4**	**2.3**	**3.9**	**756.0**
	CE [6]	14.3	38.7	53.8	9.0	16.7	36.7	47.2	13.0	6.9	22.0	32.3	26.0	9.6	31.0	44.5	14.0
	RCL-CE	**18.8**	**46.7**	**61.4**	**6.0**	**25.8**	**52.8**	**63.7**	**4.5**	**11.2**	**30.8**	**42.5**	**16.0**	**17.7**	**44.2**	**57.2**	**7.0**
0.4	DE [7]	4.4	15.4	24.0	57.0	6.7	15.4	20.3	92.0	2.8	9.9	15.5	160.0	0.2	0.9	2.2	3318.0
	RCL-DE	**7.4**	**23.6**	**35.2**	**24.0**	**10.3**	**21.5**	**29.4**	**59.0**	**4.9**	**15.8**	**24.2**	**49.0**	**0.4**	**1.7**	**3.2**	**992.0**
	CE [6]	6.7	22.0	34.0	21.0	8.1	22.8	31.0	35.5	4.7	15.4	23.6	47.0	7.6	21.8	32.1	26.75
	RCL-CE	**12.7**	**35.4**	**49.8**	**11.0**	**18.7**	**43.4**	**52.8**	**8.0**	**8.9**	**25.5**	**36.2**	**23.0**	**13.7**	**37.4**	**50.6**	**10.0**
0.6	DE [7]	2.1	8.6	14.0	98.0	2.8	8.2	10.9	222.0	0.6	3.0	5.6	248.0	0.1	0.3	0.6	4719.0
	RCL-DE	**4.3**	**15.6**	**23.7**	**49.0**	**7.2**	**13.1**	**16.9**	**159.0**	**3.5**	**12.0**	**18.8**	**81.0**	**0.4**	**1.4**	**2.2**	**1486.0**
	CE [6]	4.4	13.9	21.6	54.0	2.8	11.8	15.1	174.5	2.3	8.6	13.8	123.0	2.7	9.9	15.3	110.0
	RCL-CE	**7.8**	**23.3**	**34.2**	**23.0**	**12.4**	**26.7**	**35.2**	**26.0**	**6.2**	**19.0**	**27.9**	**40.0**	**8.5**	**25.2**	**36.4**	**20.0**
0.8	DE [7]	0.4	2.6	5.6	202.0	1.2	2.5	4.0	960.0	0.0	0.2	0.3	1465.0	0.0	0.0	0.0	14189.0
	RCL-DE	**1.1**	**5.8**	**10.3**	**142.0**	**1.5**	**3.9**	**7.8**	**159.0**	**1.7**	**6.2**	**10.3**	**211.0**	**0.1**	**0.6**	**1.0**	**4939.0**
	CE [6]	1.0	5.4	9.5	120.0	1.2	5.1	7.9	460.0	0.8	3.2	5.3	472.0	0.7	2.7	4.8	1019.5
	RCL-CE	**2.4**	**8.5**	**14.1**	**97.0**	**2.8**	**7.6**	**11.5**	260.25	**2.3**	**8.3**	**13.3**	**146.0**	**2.5**	**9.1**	**14.5**	**114.75**

robustness against PMPs.

- When the mismatching rate increases from 0.2 to 0.6, the extensions with RCL slightly decrease their performance. For example, R@1 of RCL-SAF on MS-COCO decreases from 77.1% to 70.1%, whereas the baseline SGR decreases from 71.5% to 0.1%.
- Under lower MRate, compared with non-parameterized similarity metrics (*e.g.*, SCAN, VSRN, IMRAM, and PolyLoss), the parameterized similarity metrics (*e.g.*, GSMN, SAF, and SGR) could lead to better performance despite the changes in dataset and mismatching rate. The possible reason is that the adaptive similarity metric will enhance the fitting ability of the methods even for noisy data. However, the superior fitting ability will induce models partial to PMPs under higher MRate, leading to performance degradation.
- Our RCL is remarkably superior to its counterparts for image-text matching, especially, when the mismatching rate is high. For example, on the MS-COCO dataset with 80% false positives, our RCL can improve VSRN [3] from 1.4% to 49.8% (R@1) for image-to-text matching and from 0.7% to 33.8% (R@1) for text-to-image matching. It also improves SGR [5] from 0.2% to 63.2% (R@1) for image-to-text matching and from 0.1% to 47.6% (R@1) for text-to-image.

4.3.2 Video-Text Retrieval with Synthesized Noises

In addition to the evaluation for image-text matching, we also conduct experiments for video-text matching on two benchmark datasets. Similarly, we synthesize the false positive pairs for the MSVD [50] and MSR-VTT [51] datasets. As shown in Table 3, one could conclude that RCL remarkably boosts the robustness of the baselines. More specifically,

- Like the observations on image-text matching, when the mismatched pairs become dominant in training data, the video-text matching performance of all baselines will deteriorate dramatically.
- The extensions with RCL remarkably outperform all the baselines under all settings. For example, on the MSVD dataset with 20% noises, RCL improves DE by 40.5% (R@1) for video-to-text retrieval and 8.7% (R@1) for text-to-video matching, and improves CE [6] by 31.5% (R@1) for video-to-text matching and 54.5% (R@1) for text-to-video matching, respectively. Furthermore, on the MSR-VTT dataset with 20% noises, RCL improves DE [7] by 34.7% (R@1) for video-to-text retrieval and 33.3% (R@1) for text-to-video matching, and improves CE [6] by 62.3% (R@1) for video-to-text matching and 84.4% (R@1) for text-to-video retrieval, respectively.

TABLE 4: Image-text matching on CC152K.

Method	Image-to-Text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
SCAN [2]	30.5	55.3	65.3	26.9	53.0	64.7
PolyLoss [46]	31.0	57.8	69.0	30.0	56.5	67.9
VSRN [3]	32.4	60.5	71.6	30.8	61.7	70.9
IMRAM [47]	27.8	52.4	60.9	29.2	51.5	61.2
SAF [5]	32.5	59.5	70.0	32.5	60.7	68.7
SGR [5]	14.5	35.5	48.9	13.7	36.1	47.9
SGRAF [5]	32.5	59.5	70.0	32.5	60.7	68.7
NCR* [43]	36.9	62.4	70.7	34.6	61.4	71.0
NCR [43]	39.5	64.5	73.5	40.3	64.6	73.2
RCL-VSRN	34.4	63.1	73.8	34.4	61.9	**73.6**
RCL-IMRAM	32.9	60.5	69.5	34.9	59.8	68.7
RCL-SAF	37.5	63.0	71.4	37.8	62.4	72.4
RCL-SGR	38.3	63.0	70.4	39.2	63.2	72.3
RCL-SGRAF	**41.7**	**66.0**	**73.6**	**41.6**	**66.4**	**75.1**

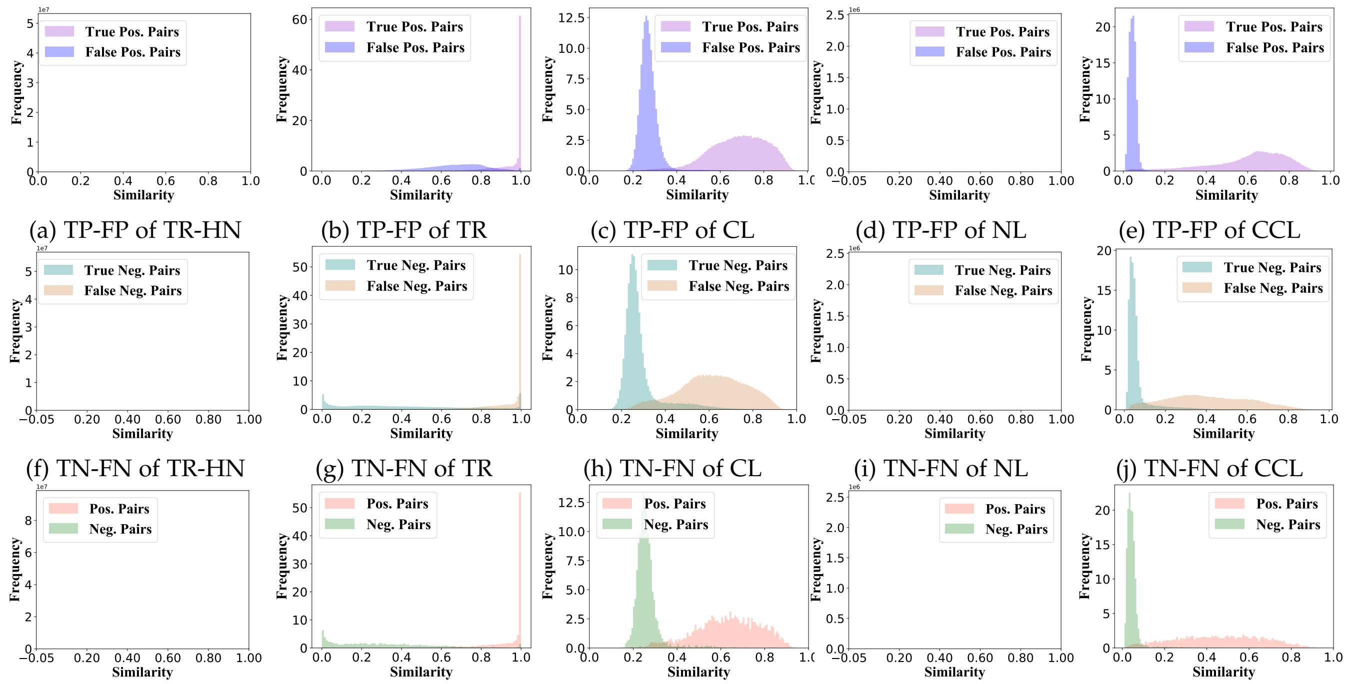
* denotes the results of one single model for NCR.

4.3.3 Image-Text Matching with Real Noises

Besides the above experiments on the synthesized noises, we also conduct comparisons on the dataset which is with real mismatched pairs. To this end, we adopt the CC152K dataset which is collected from the Internet and contains some unknown mismatched pairs. As shown in Table 4, one could see that the extensions with RCL are remarkably superior to the baselines w.r.t. the real noises. The promising

TABLE 5: Image-text matching with the mismatching rate of 0.6 on MS-COCO 1K and Flickr30K.

Method	Loss	MS-COCO							Flickr30K						
		Image-to-Text			Text-to-Image			rSum	Image-to-Text			Text-to-Image			rSum
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
SAF	TR	28.7	61.7	77.4	26.0	59.1	74.8	327.7	28.4	51.6	64.2	16.1	37.9	48.9	247.1
	TR-HN	0.5	1.3	2.4	1.4	5.4	8.8	19.8	0.1	1.2	2.1	0.5	1.1	2.1	7.1
	NL	0.1	0.9	1.3	0.1	0.5	1.0	3.9	0.0	0.5	1.3	0.1	0.5	1.0	3.4
	CL	68.0	92.1	96.7	52.0	83.5	91.5	483.8	53.8	81.3	88.4	39.6	66.6	75.7	405.4
	CCL	**70.1**	**93.1**	**96.8**	**54.5**	**84.4**	**91.9**	**490.8**	**64.5**	**86.6**	**91.6**	**43.9**	**70.0**	**79.2**	**435.8**
SGR	TR	33.7	66.9	80.3	26.2	59.1	73.2	339.4	37.9	64.7	75.0	24.1	47.7	58.3	307.7
	TR-HN	0.1	0.6	1.1	0.1	0.5	1.0	3.4	0.3	1.4	3.1	0.2	1.0	1.8	7.8
	NL	0.0	0.5	0.9	0.1	0.5	1.0	3.0	0.2	0.4	0.6	0.1	0.5	1.0	2.8
	CL	68.7	91.6	96.6	52.3	83.3	91.1	483.6	56.8	81.0	88.4	39.4	66.6	75.8	408.0
	CCL	**71.4**	**93.2**	**97.1**	**55.4**	**84.7**	**92.3**	**494.1**	**65.1**	**86.1**	**92.0**	**44.3**	**71.2**	**79.7**	**438.4**



(a) TP-FP of TR-HN (b) TP-FP of TR (c) TP-FP of CL (d) TP-FP of NL (e) TP-FP of CCL

(f) TN-FN of TR-HN (g) TN-FN of TR (h) TN-FN of CL (i) TN-FN of NL (j) TN-FN of CCL

(k) PP-NP of TR-HN (l) PP-NP of TR (m) PP-NP of CL (n) PP-NP of NL (o) PP-NP of CCL

Fig. 4: Comparison of robustness against PMPs with the mismatching rate of 0.6. This figure shows the pairwise similarity distributions of TP-FP (true positive pairs vs. false positive pairs on the training set of MS-COCO), TN-FN (true negative pairs vs. false negative pairs on the training set of MS-COCO), and PP-NP (positive pairs vs. negative pairs on validation set of MS-COCO) calculated by TR-HN, TR, CL, NL, and CCL, respectively.

performance of our method could be attributed to the that our CCL loss adopts only the negative pairs to avoid using false information, thus embracing better performance. Specifically, RCL improves VSRN [3] by 6.2% (R@1) for image-to-text matching and 11.7% (R@1) for text-to-image matching, IMRAM [47] by 18.3% (R@1) for image-to-text matching and 19.5% (R@1) for text-to-image matching, and SGR [5] by 164.1% (R@1) for image-to-text matching and 186.1% (R@1) for text-to-image matching. The experiments verify that our RCL could provide an effective solution to utilize massive and economical data collected from the Internet while being immune to possible mismatched pairs.

4.4 Comparison with Rectifying Method

In this section, we compare our RCL with the most related method NCR [43] to investigate the effectiveness and efficiency of the proposed learning paradigm. First, NCR requires simultaneously training two individual cross-modal models in a co-teaching manner, which will take a relatively high computational cost. In contrast, our method does not introduce extra training costs into the original cross-modal method, thus embracing higher efficiency. Second, we conduct some comparisons with NCR in Table 6. From the experiments, one could see that both NCR and our RCL achieve comparable retrieval performance in low mismatching rates (*e.g.*, 0.2 and 0.4). However, the performance of NCR will fast degrade with high mismatching rates (*e.g.*, 0.6 and 0.8) because NCR cannot correctly distinguish true positives from false positives when the PMPs dominate in the training data. Furthermore, one could find that NCR achieves worse performance under PMPs comparing the results reported in [43], which demonstrates that our PMP injection approach is more challenging than that used in NCR.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3247939

TABLE 6: Comparison with NCR [43] under different mismatching rates (MRate) on MS-COCO and Flickr30K.

MRate	Method	MS-COCO Image-to-Text R@1	R@5	R@10	Text-to-Image R@1	R@5	R@10	rSum	Flickr30K Image-to-Text R@1	R@5	R@10	Text-to-Image R@1	R@5	R@10	rSum
0.2	NCR* [43]	73.7	94.5	97.7	58.3	88.7	94.0	506.9	69.9	**92.0**	95.4	52.6	79.4	86.8	476.1
	RCL-SAF	**77.1**	**95.5**	**98.2**	**61.0**	**88.8**	**94.6**	**515.2**	**72.0**	91.7	**95.8**	**53.6**	**79.9**	86.7	**479.7**
	RCL-SGR	**77.0**	**95.5**	**98.1**	**61.3**	**88.8**	**94.8**	**515.5**	**74.2**	91.8	**96.9**	**55.6**	**81.2**	**87.5**	**487.2**
	NCR [43]	76.6	95.6	98.2	60.8	88.8	95.0	515.0	73.5	93.2	96.6	56.9	82.4	88.5	491.1
	RCL-SGRAF	**78.9**	**96.0**	**98.4**	**62.8**	**89.9**	**95.4**	**521.4**	**75.9**	**94.5**	**97.3**	**57.9**	**82.6**	**88.6**	**496.8**
0.4	NCR* [43]	71.7	93.9	97.5	56.7	86.8	**94.0**	500.6	61.6	88.3	92.8	46.9	74.5	82.3	446.4
	RCL-SAF	**74.8**	**94.8**	**97.8**	**59.0**	**87.1**	93.9	**507.4**	**68.8**	**89.8**	**95.0**	**51.0**	**76.7**	**84.8**	**466.1**
	RCL-SGR	**73.9**	**94.9**	**97.9**	**59.0**	**87.4**	93.9	**507.0**	**71.3**	**91.1**	**95.3**	**51.4**	**78.0**	**85.2**	**472.3**
	NCR [43]	74.7	94.6	98.0	59.6	88.1	94.7	509.7	68.1	89.6	94.8	51.4	78.4	84.8	467.1
	RCL-SGRAF	**77.0**	**95.5**	**98.3**	**61.2**	**88.5**	**94.8**	**515.3**	**72.7**	**92.7**	**96.1**	**54.8**	**80.0**	**87.1**	**483.4**
0.6	NCR* [43]	0.1	0.3	0.4	0.1	0.5	1.0	2.4	13.7	34.7	46.9	10.1	27.4	38.4	171.2
	RCL-SAF	**70.1**	**93.1**	**96.8**	**54.5**	**84.4**	**91.9**	**490.8**	**63.9**	**84.8**	**91.7**	**43.0**	**71.2**	**79.4**	**434.0**
	RCL-SGR	**71.4**	**93.2**	**97.1**	**55.4**	**84.7**	**92.3**	**494.1**	**62.3**	**86.3**	**92.9**	**45.1**	**71.3**	**80.2**	**438.1**
	NCR [43]	0.1	0.3	0.4	0.1	0.5	1.0	2.4	13.9	37.7	50.5	11.0	30.1	41.4	184.6
	RCL-SGRAF	**74.0**	**94.3**	**97.5**	**57.6**	**86.4**	**93.5**	**503.3**	**67.7**	**89.1**	**93.6**	**48.0**	**74.9**	**83.3**	**456.6**
0.8	NCR* [43]	0.1	0.3	0.4	0.1	0.5	1.0	2.4	0.9	2.7	4.7	0.2	0.8	1.6	10.9
	RCL-SAF	**62.9**	**89.3**	**94.9**	**47.1**	**77.9**	**87.4**	**459.5**	**45.0**	**72.8**	**80.8**	**30.7**	**56.5**	**67.3**	**353.1**
	RCL-SGR	**63.2**	**89.3**	**95.2**	**47.6**	**78.7**	**88.0**	**462.0**	**47.1**	**70.5**	**79.4**	**30.3**	**56.1**	**66.3**	**349.7**
	NCR [43]	0.1	0.3	0.4	0.1	0.5	1.0	2.4	1.5	6.2	9.9	0.3	1.0	2.1	21.0
	RCL-SGRAF	**67.4**	**90.8**	**96.0**	**50.6**	**81.0**	**90.1**	**475.9**	**51.7**	**75.8**	**84.4**	**34.5**	**61.2**	**70.7**	**378.3**

* denotes the results of one single model for NCR.

4.5 Image-Text Matching with Different Upper Bounds

In this section, we investigate the effectiveness of the variants of our framework with different upper bounds, *i.e.*, different loss functions. From the experimental results, one could see that the vanilla $\mathcal{L}_{\mathrm{mae}}$ cannot achieve satisfactory performance, due to the underfitting issue faced by complementary learning. Thanks to the proposed strategy of multiple negatives, $\mathcal{L}_{\mathrm{mae}}$ achieves comparable results. However, MAE treats each point equally and ignores hard samples, thus leading to performance degradation. To address such a problem, we optimize different upper bounds of MAE to improve the performance while preserving robustness. From the experimental results, one could find that all upper bounds could improve $\mathcal{L}_{\mathrm{mae}}$ by $1.7 \sim 3.5$ in terms of the overall scores (*i.e.*, rSum).

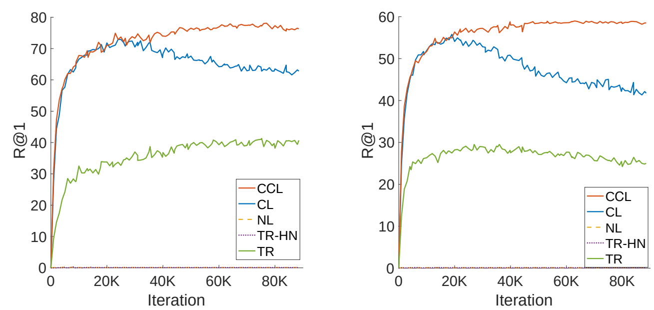
TABLE 7: Comparison with different presented loss functions under the mismatching rates (MRate) of 0.6 on MS-COCO.

Loss	Image-to-Text R@1	R@5	R@10	Text-to-Image R@1	R@5	R@10	rSum
$\mathcal{L}^{*}_{\mathrm{mae}}$	0.1	0.5	1.0	0.1	0.5	1.0	3.2
$\mathcal{L}_{\mathrm{mae}}$	67.8	93.3	97.2	55.4	**85.8**	**92.9**	492.4
$\mathcal{L}_{\mathrm{exp}}$	72.0	92.9	97.2	54.9	85.0	92.7	494.7
$\mathcal{L}_{\mathrm{log}}$	71.4	93.2	97.1	55.4	84.7	92.3	494.1
$\mathcal{L}_{\mathrm{gce}}$	**72.6**	**93.7**	**97.3**	55.4	84.6	92.1	495.7
$\mathcal{L}_{\mathrm{tan}}$	72.2	**93.7**	**97.3**	**55.5**	84.7	92.5	**495.9**

4.6 Ablation Study

To comprehensively investigate the effectiveness of our CCL, we carry out some ablation studies by using the following five loss functions:

- **TR** [45] is the hinge-based triplet ranking loss.
- **TR-HN** [29] is the widely-used hinge-based triplet ranking loss with hard negatives.
- **CL** [23] is the contrastive learning loss, *i.e.*, Equation (8).
- **NL** [17] is the negative learning (aka complementary learning) loss.



(a) Image-to-Text (b) Text-to-Image

Fig. 5: Performance of different loss functions in SGR in terms of R@1 scores. The evaluation is conducted on the validation set of MS-COCO with MRate=0.6.

Besides the choice in the loss function, the experiments are conducted by training the same settings including but not limited to network structure, hyper-parameters, and optimizer. The ablation study is carried out on MS-COCO and Flickr30K in terms of image-text matching. As demonstrated in Table 5, Figs. 4 and 5, one could see that TR-HN overfits the false positives because it focuses on the hardest pairs. With the soft relaxation, TR has achieved better performance than TR-HN because it could avoid overfitting PMPs. Different from TR-HN, NL only utilizes negative labels. However, as the negative labels are less informative, NL will encounter the underfitting issue as elaborated in Sections 1 and 3. In a contrastive learning manner, although CL could be immune to the false positive in the early training stage, it also overfits the uncorrected supervision with further training, thus leading to performance degradation. Fortunately, our CCL could simultaneously address the overfitting and underfitting issues as claimed and achieve the best performance.

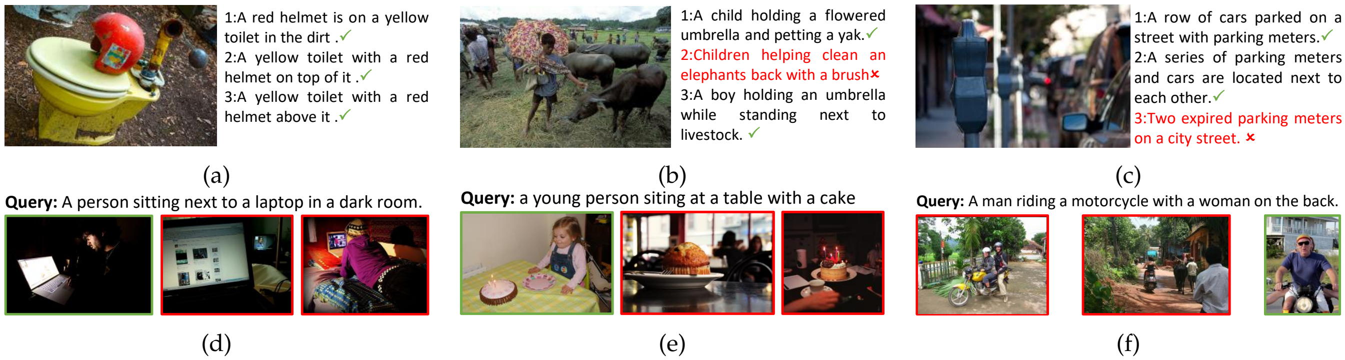


Fig. 6: The ability of our RCL to capture latent semantics for cross-modal retrieval with MRate=0.6. The figure shows some retrieved examples of the image-to-text (as shown in (a)–(c)) and text-to-image (as shown in (d)–(f)) for RCL-SGR on the validation set of MS-COCO dataset. We show the top-3 retrieved texts and images for each given image and text query, respectively. The correctly matched ones are marked in green, and incorrectly matched in red. Specifically, the correctly matched sentences are with green check marks, and the incorrectly matched ones are with red words and X marks. The ground-truth matched images are outlined in green boxes and unmatched in red boxes.

4.7 Parameter Analysis

In this section, we investigate the influence of the hyper-parameter τ in Fig. 7 by plotting the average scores of image-text matching (R@1, R@5, and R@10) with different τ on the Flickr30K dataset. From the figure, one could observe that our method performs stably in a large range of τ, *i.e.*, from 0.01 to 0.1.

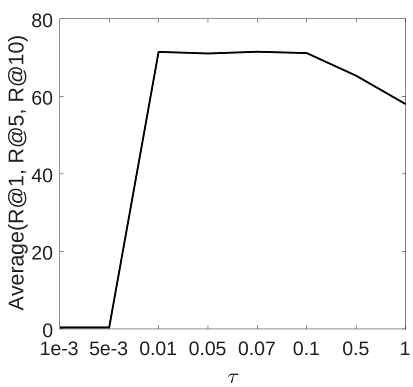


Fig. 7: Parameter analysis of RCL-SAF in terms of average scores ($R@1$, $R@5$, and $R@10$) for image-text matching with MRate=0.6 on the validation set of Flickr30K.

4.8 Benefit Study on PMPs

To comprehensively investigate the effectiveness of our RCL, we conduct some comparison experiments with two competitive baselines by filtering out the mismatched pairs from the noisy data:

- **SAF-C** and **SGR-C**: The variants are strong baselines, which are trained on the clean pairs by discarding all the mismatched pairs.
- **SAF [5]+CLIP [9]** and **SGR [5]+CLIP [9]**: The pre-trained CLIP (ViT-L/14@336px) [9] is applied to filter out the predicted mismatched pairs, and the remaining pairs with high cross-modal similarities are utilized to train SGR and SAF.

The comparison results are shown in Table 8. From the table, one could see that filtering out mismatched pairs could alleviate the adverse impact of PMPs. Even if pretrained CLIP can filter out some mismatched pairs to improve the robustness of SGR and SAF against PMPs, they still perform worse than SGR-C and SAF-C, indicating that there are still some residual mismatched pairs. Additionally, under high noise rates, many pairs will be filtered out, leading to poor performance on Flick30K. This indicates that although filtering out mismatched pairs can improve performance, it also discards a large number of pairs that contain semantic information. Our approach not only reduces the negative impact of PMPs, but also leverages PMPs to improve performance, embracing the best performance.

4.9 Visualization and Analysis

In this section, we visually verify the robustness of RCL and conduct the case study.

4.9.1 *Robustness Analysis against PMPs*

To intuitively show the robustness performance of our method, we illustrate pairwise similarity distributions of TP-FP (*i.e.*, true positive pairs versus false positive pairs), TN-FN (*i.e.*, true negative pairs versus false negative pairs), and PP-NP (*i.e.*, positive pairs versus negative pairs) of our RCL-SGR and its variants (see Section 4.8) on MS-COCO. Specifically, Figs. 4(a)–4(e), Figs. 4(f)–4(j), and Figs. 4(k)–4(o) show the distributions of TP-FP, TN-FN, and PP-NP on all training positive pairs, the training set, and the validation set of MS-COCO, respectively. From Figs. 4(a)–4(e), one could see that TR and CL could not separate the true and false positive pairs apart enough, which degrades their performance since the existence of PMPs. However, our CCL could correctly separate the true and false positive pairs well because our CCL only focuses on the negative information resulting in robustness against the false positive pairs as shown Fig. 4(e). From Figs. 4(f)–4(j), one could see that true and false negative pairs are more difficult to separate than true and false positive ones. Although our method only focuses on negative information, it also could discriminate the true and false negative pairs better because of a low proportion of false negative pairs in the training set. For the positive learning methods (TR and CL), they will

TABLE 8: Comparison with filtering-based baselines under different mismatching rates (MRate) on MS-COCO 1K and Flickr30K.

Noise	Methods	MS-COCO							Flickr30K						
		Image-to-Text			Text-to-Image			rSum	Image-to-Text			Text-to-Image			rSum
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
0.2	SAF [5]	71.5	94.0	97.5	57.8	86.4	91.9	499.1	62.8	88.7	93.9	49.7	73.6	78.0	446.7
	SGR [5]	25.7	58.8	75.1	23.5	58.9	75.1	317.1	55.9	81.5	88.9	40.2	66.8	75.3	408.6
	SAF [5]+CLIP [9]	74.0	95.2	98.0	58.7	88.0	94.4	508.3	68.1	90.1	94.0	49.6	76.6	83.6	462.0
	SGR [5]+CLIP [9]	74.7	94.9	98.1	58.9	87.8	94.3	508.7	69.1	90.1	94.2	50.3	76.1	83.8	463.6
	SAF-C	74.9	94.8	98.0	58.7	88.0	94.4	508.8	68.3	90.6	95.0	51.1	77.5	84.7	467.2
	SGR-C	74.4	95.1	98.1	58.6	87.6	94.2	508.0	72.2	91.3	95.5	51.5	76.3	82.0	468.8
	RCL-SAF	**77.1**	**95.5**	**98.2**	**61.0**	**88.8**	**94.6**	**515.2**	**72.0**	**91.7**	**95.8**	**53.6**	**79.9**	**86.7**	**479.7**
	RCL-SGR	**77.0**	**95.5**	**98.1**	**61.3**	**88.8**	**94.8**	**515.5**	**74.2**	**91.8**	**96.9**	**55.6**	**81.2**	**87.5**	**487.2**
0.4	SAF [5]	13.5	43.8	48.2	16.0	39.0	50.8	211.3	7.4	19.6	26.7	4.4	12.0	17.0	87.1
	SGR [5]	1.3	3.7	6.3	0.5	2.5	4.1	18.4	4.1	16.6	24.1	4.1	13.2	19.7	81.8
	SAF [5]+CLIP [9]	71.4	94.3	97.9	57.1	86.8	94.0	501.5	61.5	85.6	92.1	44.5	72.0	81.1	436.8
	SGR [5]+CLIP [9]	72.7	94.3	97.9	56.8	86.5	93.2	501.4	62.2	86.0	92.1	44.6	71.4	78.6	434.9
	SAF-C	72.4	94.3	97.8	57.5	86.9	93.8	502.7	63.9	88.7	93.2	46.7	73.5	81.4	447.4
	SGR-C	72.7	94.2	97.9	57.5	87.0	93.8	503.1	67.1	89.6	93.7	47.6	73.5	81.1	452.6
	RCL-SAF	**74.8**	**94.8**	**97.8**	**59.0**	**87.1**	**93.9**	**507.4**	**68.8**	**89.8**	**95.0**	**51.0**	**76.7**	**84.8**	**466.1**
	RCL-SGR	**73.9**	**94.9**	**97.9**	**59.0**	**87.4**	**93.9**	**507.0**	**71.3**	**91.1**	**95.3**	**51.4**	**78.0**	**85.2**	**472.3**
0.6	SAF [5]	0.1	0.5	0.7	0.8	3.5	6.3	11.9	0.1	1.5	2.8	0.4	1.2	2.3	8.3
	SGR [5]	0.1	0.6	1.0	0.1	0.5	1.1	3.4	1.5	6.6	9.6	0.3	2.3	4.2	24.5
	SAF [5]+CLIP [9]	68.4	93.0	96.8	54.3	85.0	92.7	490.2	21.9	53.8	69.1	16.2	40.3	53.3	254.6
	SGR [5]+CLIP [9]	56.5	85.6	93.6	42.9	77.1	87.4	443.1	2.3	7.7	12.2	1.9	6.9	11.1	42.1
	SAF-C	69.1	92.6	96.9	54.0	84.9	**92.8**	490.3	45.3	74.2	84.1	32.8	59.8	69.5	365.7
	SGR-C	66.9	92.0	96.6	52.3	83.6	91.8	483.2	47.1	72.2	82.1	31.8	57.4	66.6	357.2
	RCL-SAF	**70.1**	**93.1**	96.8	**54.5**	84.4	91.9	**490.8**	**63.9**	**84.8**	**91.7**	**43.0**	**71.2**	**79.4**	**434.0**
	RCL-SGR	**71.4**	**93.2**	**97.1**	**55.4**	**84.7**	92.3	**494.1**	**62.3**	**86.3**	**92.9**	**45.1**	**71.3**	**80.2**	**438.1**
0.8	SAF [5]	0.2	0.8	1.4	0.1	0.5	1.0	4.0	0.0	0.8	1.2	0.1	0.5	1.1	3.7
	SGR [5]	0.2	0.6	1.0	0.1	0.5	1.0	3.4	0.2	0.3	0.5	0.1	0.6	1.0	2.7
	SAF [5]+CLIP [9]	24.1	37.2	40.4	20.0	34.0	38.2	193.9	3.1	8.6	13.8	0.5	1.8	3.0	30.8
	SGR [5]+CLIP [9]	22.0	54.6	69.8	17.0	47.5	64.8	275.7	0.5	1.1	2.1	0.2	0.9	1.7	6.5
	SAF-C	60.3	88.7	94.4	47.1	**80.4**	**89.9**	460.8	3.8	12.2	18.2	0.9	3.9	6.8	45.8
	SGR-C	50.1	81.3	90.2	39.0	72.5	84.5	417.6	0.2	1.4	3.2	0.4	1.6	2.9	9.7
	RCL-SAF	**62.9**	**89.3**	**94.9**	**47.1**	77.9	87.4	459.5	**45.0**	**72.8**	**80.8**	**30.7**	**56.5**	**67.3**	**353.1**
	RCL-SGR	**63.2**	**89.3**	**95.2**	**47.6**	78.7	88.0	**462.0**	**47.1**	**70.5**	**79.4**	**30.3**	**56.1**	**66.3**	**349.7**

overfit the false positive pairs, and degrade the performance of discriminating true or false negative pairs. Figs. 4(k)–4(o) illustrate the similarity distributions of different methods on the validation set of MS-COCO, which is consistent with their retrieval performance. Furthermore, TR-HN and NL will face very serious overfitting and underfitting problems, thus leading to an optimization inability and the worst performance for PMPs.

In conclusion, by paying more attention to positive ones, TR will push more positive pairs to the high similarities. However, this radical learning paradigm will easily overfit the false positive and negative pairs, thus a considerable number of negative ones are pushed to the high similarity as shown in Figs. 4(b), 4(g) and 4(l). With a more soft learning paradigm, CL could not extremely separate the positive and negative pairs like TR, it could achieve more correct separation. However, these positive learning paradigms will face the overfitting problem. Thanks to our CCL, the negative information could be fully leveraged to alleviate the interference brought by PMPs, thus embracing better separation and robustness against PMPs, which also is demonstrated in Sections 4.3 and 4.8.

4.9.2 *Retrieved Examples*

To visually illustrate the ranking performance of our RCL, we show some retrieved text and image samples using image queries and text queries on the validation set of MS-COCO in Fig. 6 like [53], respectively. Specifically, each figure of Figs. 6(a)–6(c) shows one given image query (left) and its top 3 ranked sentences (right). Similarly, each figure

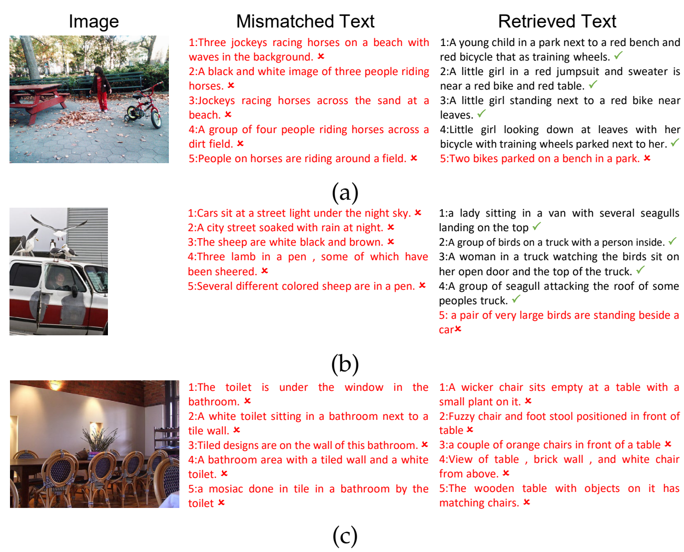


Fig. 8: The robustness of our RCL against PMPs with MRate=0.6. This figure shows some mismatched and retrieved examples for our RCL-SGR on the training set of the MS-COCO dataset. (a)– (c) illustrate the mismatched (middle) and top-5 retrieved (right) textual examples for each given image (left). The correctly matched samples are marked by green check marks, and incorrectly matched ones are marked by red X marks.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3247939

of Figs. 6(d)–6(f) shows one given sentence query (top) and its top 3 ranked images (bottom). Noted that in MS-COCO, one image has five relevant sentences, but one sentence has only one paired image. From these retrieved results, one could see that most of the relevant samples could be correctly retrieved across different modalities by our approach. Although some retrieved examples are not correct based on the ground truth, they also are semantically close to the given queries. For example, one could see that all retrieved images share the same semantic concept in Fig. 6(f), *i.e.*, "riding a motorcycle with a woman" in the given text query. Similar observations also could be found in other retrieved results. In summary, although the inference model is trained from PMPs, our RCL also could endow it with the ability to learn correct semantics, and make the model robust against mismatching information.

4.9.3 Mismatched Examples

To visually investigate the performance of our RCL against PMPs, we also illustrate some mismatched examples and the corresponding retrieved textual examples by our RCL-SGR on the training set of MS-COCO as shown in Fig. 8. Like Fig. 6, each figure of Figs. 8(a)–8(c) shows five mismatched textual examples (middle) for a given image (left), and top-5 retrieved results by our RCL. From the given examples, one could see that our method could still capture the semantics for cross-modal retrieval despite the presence of mismatched pairs. Thanks to our CCL, our method could not overfit the mismatched pairs. More specifically, although the training set gives the wrong ground truths as shown in the middle column, our method still could obtain the correctly matched pairs as shown in the right column, which indicates that our RCL is robust against PMPs and alleviates overfitting on PMPs of the training data. Even for the wrongly retrieved results as shown in the right column, they also are semantically close to the given image. For example, these sentences have captured the semantic concept of "chair" and "table" in Fig. 8(c). In other words, our method could excavate the semantics from PMPs, and semantically correlate cross-modal pairs, thus resulting in alleviating the interference of mismatched pairs to improve retrieval performance.

5 Conclusion

In this paper, we study a less-touched problem in the community, namely, cross-modal retrieval with partially mismatched pairs. To tackle this challenging problem, we propose RCL which consists of CMCL and CCL. The former is used to compute the matching probability across modalities, and the latter is a novel complementary learning paradigm that is specifically designed to overcome the overfitting issue faced by CMCL and the underfitting issue faced by complementary learning. Extensive experiments are conducted on five benchmark cross-modal datasets to verify the effectiveness, robustness, and generalization of our method.

References

[1] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3893–3903, 2018.

[2] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 201–216.

[3] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 4654–4662.

[4] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 921–10 930.

[5] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," Technical Report, Tech. Rep., 2021.

[6] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman, "Use what you have: Video retrieval using representations from collaborative experts," in *Proceedings of the British Machine Vision Conference (BMVC)*, September 2019.

[7] J. Dong, X. Li, C. Xu, S. Ji, Y. He, G. Yang, and X. Wang, "Dual encoding for zero-example video retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 9346–9355.

[8] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning (ICML)*, 2021.

[9] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," *arXiv preprint arXiv:2103.00020*, 2021.

[10] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "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)*, 2018, pp. 2556–2565.

[11] A. Veit, N. Alldrin, G. Chechik, I. Krasin, A. Gupta, and S. Belongie, "Learning from noisy large-scale datasets with minimal supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 839–847.

[12] B. Han, J. Yao, G. Niu, M. Zhou, I. Tsang, Y. Zhang, and M. Sugiyama, "Masking: A new perspective of noisy supervision," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 5836–5846.

[13] B. Han, Q. Yao, X. Yu, G. Niu, M. Xu, W. Hu, I. Tsang, and M. Sugiyama, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018, pp. 8527–8537.

[14] X. Yu, B. Han, J. Yao, G. Niu, I. W. Tsang, and M. Sugiyama, "How does disagreement help generalization against label corruption?" in *International Conference on Machine Learning (ICML)*, 2019.

[15] X. Xia, T. Liu, B. Han, C. Gong, N. Wang, Z. Ge, and Y. Chang, "Robust early-learning: Hindering the memorization of noisy labels," in *International Conference on Learning Representations (ICLR)*, 2020.

[16] Y. Bai, E. Yang, B. Han, Y. Yang, J. Li, Y. Mao, G. Niu, and T. Liu, "Understanding and improving early stopping for learning with noisy labels," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[17] Y. Kim, J. Yim, J. Yun, and J. Kim, "NLNL: Negative learning for noisy labels," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 101–110.

[18] D. T. Nguyen, C. K. Mummadi, T. P. N. Ngo, T. H. P. Nguyen, L. Beggel, and T. Brox, "Self: Learning to filter noisy labels with self-ensembling," in *International Conference on Learning Representations (ICLR)*, 2020.

[19] J. Li, R. Socher, and S. C. Hoi, "DivideMix: Learning with noisy labels as semi-supervised learning," in *International Conference on Learning Representations (ICLR)*, 2020.

[20] A. Ghosh, H. Kumar, and P. Sastry, "Robust loss functions under label noise for deep neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 31, no. 1, 2017.

[21] X. Ma, H. Huang, Y. Wang, S. Romano, S. Erfani, and J. Bailey, "Normalized loss functions for deep learning with noisy labels," in *International Conference on Machine Learning (ICML)*, 2020.

[22] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 3733–3742.

This article has been accepted for publication in IEEE Transactions on Pattern Analysis and Machine Intelligence. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2023.3247939

[23] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International Conference on Machine Learning (ICML)*, 2020.

[24] T. Ishida, G. Niu, W. Hu, and M. Sugiyama, "Learning from complementary labels," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017, pp. 5644–5654.

[25] L. Feng, T. Kaneko, B. Han, G. Niu, B. An, and M. Sugiyama, "Learning with multiple complementary labels," in *International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 3072–3081.

[26] Y. Kim, J. Yun, H. Shon, and J. Kim, "Joint negative and positive learning for noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 9442–9451.

[27] R. Chhabra and M. G. Basavaraj, "Chapter 6 - motion of particles in a fluid," pp. 281–334, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978008101098300007X

[28] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9879–9889.

[29] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018.

[30] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 7181–7189.

[31] H. Zhang, A. Sun, W. Jing, G. Nan, L. Zhen, J. T. Zhou, and R. S. M. Goh, "Video corpus moment retrieval with contrastive learning," in *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021, pp. 685–695.

[32] S. Chen, Y. Zhao, Q. Jin, and Q. Wu, "Fine-grained video-text retrieval with hierarchical graph reasoning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10 638–10 647.

[33] Y. Li, D. Zhang, and Y. Mu, "Visual-semantic matching by exploring high-order attention and distraction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 12 786–12 795.

[34] H. Song, M. Kim, D. Park, and J.-G. Lee, "Learning from noisy labels with deep neural networks: A survey," *arXiv preprint arXiv:2007.08199*, 2020.

[35] C. Gong, Q. Wang, T. Liu, B. Han, J. J. You, J. Yang, and D. Tao, "Instance-dependent positive and unlabeled learning with labeling bias estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[36] Y. Li, J. Yang, Y. Song, L. Cao, J. Luo, and L.-J. Li, "Learning from noisy labels with distillation," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 1910–1918.

[37] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu, "Making deep neural networks robust to label noise: A loss correction approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1944–1952.

[38] D. Hendrycks, M. Mazeika, D. Wilson, and K. Gimpel, "Using trusted data to train deep networks on labels corrupted by severe noise," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[39] A. Vahdat, "Toward robustness against label noise in training deep discriminative neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[40] L. Jiang, Z. Zhou, T. Leung, L.-J. Li, and L. Fei-Fei, "MentorNet: Learning data-driven curriculum for very deep neural networks on corrupted labels," in *International Conference on Machine Learning (ICML)*, 2018, pp. 2304–2313.

[41] Y. Yan, Z. Xu, I. W. Tsang, G. Long, and Y. Yang, "Robust semi-supervised learning through label aggregation," in *Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, 2016.

[42] Z. Zhang and M. R. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[43] Z. Huang, G. Niu, X. Liu, W. Ding, X. Xiao, X. Peng *et al.*, "Learning with noisy correspondence for cross-modal matching," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[44] P. Hu, X. Peng, H. Zhu, L. Zhen, and J. Lin, "Learning cross-modal retrieval with noisy labels," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 5403–5413.

[45] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *arXiv preprint arXiv:1411.2539*, 2014.

[46] J. Wei, X. Xu, Y. Yang, Y. Ji, Z. Wang, and H. T. Shen, "Universal weighting metric learning for cross-modal matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 005–13 014.

[47] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "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 (CVPR)*, 2020, pp. 12 655–12 663.

[48] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 740–755.

[49] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[50] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 190–200.

[51] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VT: A large video description dataset for bridging video and language," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5288–5296.

[52] A. Miech, I. Laptev, and J. Sivic, "Learning a text-video embedding from incomplete and heterogeneous data," *arXiv preprint arXiv:1804.02516*, 2018.

[53] Y. Wu, S. Wang, G. Song, and Q. Huang, "Learning fragment self-attention embeddings for image-text matching," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 2088–2096.