Information Fusion 130 (2026) 104156

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full Length Article

Negative can be positive: A stable and noise-resistant complementary contrastive learning for cross-modal matching

Fangming Zhong [a,b], Xinyu He [a,b], Haiquan Yu [a,b], Xiu Liu [a,b], Suhua Zhang [a,b,*]

[a] School of Software, Dalian University of Technology, Dalian, China
[b] Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian, China

ARTICLE INFO

Keywords:
Cross-modal matching
Contrastive learning
Complementary learning

ABSTRACT

Cross-modal matching with noisy correspondence has drawn considerable interest recently, due to the mismatched data imposed inevitably when collecting data from the Internet. Training on such noisy data often leads to severe performance degradation, as conventional methods tend to overfit rapidly to wrongly mismatched pairs. Most of the existing methods focus on predicting more reliable soft correspondence, generating higher weights for the pairs that are more likely to be correct. However, there still remain two limitations: (1) they ignore the informative signals embedded in the negative pairs, and (2) the instability of existing methods due to their sensitivity to the noise ratio. To address these issues, we explicitly take the negatives into account and propose a stable and noise-resistant complementary learning method, named Dual Contrastive Learning (DCL), for cross-modal matching with noisy correspondence. DCL leverages both positive pairs and negative pairs to improve the robustness. With the complementary contrastive learning, the negative pairs also contribute positively to the model optimization. Specifically, to fully explore the potential of mismatched data, we first partition the training data into clean and noisy subsets based on the memorization effect of deep neural networks. Then, we employ vanilla contrastive learning for positive matched pairs in the clean subset. As for negative pairs including the noisy subsets, complementary contrastive learning is adopted. In such doing, whatever the level of noise ratio is, the proposed method is robust to balance the positive information and negative information. Extensive experiments indicate that DCL significantly outperforms the state-of-the-art methods and exhibits remarkable stability with an extremely low variance of R@1. Specifically, the R@1 scores of our DCL are 7% and 9.1% higher than NPC on image-to-text and text-to-image, respectively. The source code is released at https://github.com/hxy2969/dcl.

1. Introduction

Recently, multimodal retrieval has drawn much attention with the explosive growth of multimedia data, especially with the popularization of short-form video social platforms, such as TikTok and Instagram Reels. Cross-modal matching is one of the most significant techniques that tries to project different modalities into a shared feature space to match paired samples. It has powered various real-world applications, *e.g.*, image captioning [1,2], visual question answering [3], and the application in smart agriculture, i.e., cross-modal retrieval among text, images, and videos of plant diseases.

Although cross-modal matching has achieved promising performance, it heavily relies on large-scale high-quality labeled data. However, collecting such ideal data is time-consuming and extremely expensive. In practice, the most widely used datasets are harvested from the Internet by collecting the co-occurred image-text pairs to train cross-modal matching models, such as CLIP [4]. Nevertheless, it is hard to accurately annotate exactly matched samples because the textual description of images or videos is very subjective. Without meticulous manual annotation, it is inevitable that noise (i.e., mismatched pairs) will be introduced into the collected data, *a.k.a* noisy correspondence. Undoubtedly, the noisy pairs will wrongly guide the cross-modal matching models due to the memorization effect of deep neural networks and remarkably degrade the retrieval performance. Although the noisy correspondence issue is a widely existing problem [5], but has been rarely explored in cross-modal matching. The paradigm most similar to such an issue is the noisy label in classification [6,7]. Unfortunately, the noisy correspondence refers to wrongly aligned cross-modal pairs, which is more challenging than categorical annotation errors [8–10].

To date, many efforts have been made to address this issue. Most of the previous works [11–13] rectify the correspondence labels, and train the models with robust loss functions with a soft margin to counter

* Corresponding author.
E-mail addresses: fmzhong@dlut.edu.cn (F. Zhong), xinyuhe@mail.dlut.edu.cn (X. He), yhq@dlut.edu.cn (H. Yu), liuxiu912@gmail.com (X. Liu), suhua_zhang@mail.dlut.edu.cn (S. Zhang).

https://doi.org/10.1016/j.inffus.2026.104156
Received 3 September 2025; Received in revised form 10 January 2026; Accepted 14 January 2026
Available online 16 January 2026

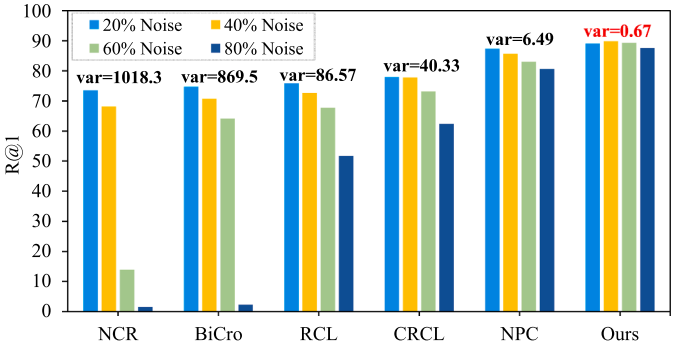


Fig. 1. Stability of different methods with the increase of noise ratio via R@1 of Image-to-Text.

the negative impact of noise. For example, Noisy Correspondence Rectifier (NCR) [11] is the first touch on this issue. NCR partitions the data into clean and noisy subsets, then employs an adaptive model to rectify the correspondence. Based on the assumption that similar images should correspond to similar textual descriptions, Bidirectional Cross-modal similarity consistency (BiCro) [12] estimates soft labels for noisy data pairs to reflect their true correspondence degree. Similarly, since the semantic variations caused by image changes should be proportional to those caused by text changes for any two matched samples, an Equivariant Similarity Consistency (ESC) [14] is presented to facilitate robust clean and noisy data separation. By viewing sample matching as classification tasks within the batch, Self-Reinforcing Errors Mitigation (SREM) [15] generates classification logits for the given sample. Then SREM introduces energy uncertainty to refine sample filtration and utilizes swapped classification entropy to estimate the model's sensitivity of selected clean samples. Unlike traditional paradigms that mainly focus on samples filtering or correction, Negative Pre-aware Cross-modal (NPC) matching [16] proposes a negative pre-aware paradigm, which adaptively estimates the potential negative impact of each sample before the model learning. Then, a small confidence weight will be assigned to high-negative samples. Though promising results have been achieved, the noise-rectify and re-weighting paradigms still have two limitations. Most of them ignore the positive contribution of negative pairs. In addition, the label correction may cause new noise resulting noisy accumulation due to the confirmation bias problem in the existence of severe noise thus cannot maintain the performance stability. The performance of these methods decreases dramatically as the noise ratio increases.

Different from previous methods, Robust Cross-modal Learning (RCL) [17] proposes a novel complementary contrastive learning paradigm that employs negative information exclusively. The negative information is more reliable compared to positive information, effectively mitigating overfitting risks. Evidently, complementary information demonstrates a significantly lower probability of containing false ground truth compared to positive information, thus avoiding overfitting to false supervision. However, RCL ignores the impact of noise ratio. With the noise ratio decreasing, the increasing informative and valuable clean pairs are ignored, leading the model to be underfitting. Therefore, the existing methods still exhibit unstable performance with a considerably larger variance than ours, as shown in Fig. 1 where we employ variance (var) of Recall@1 at different noise ratios to illustrate the performance stability.

To tackle the above issues, we propose a stable and noise-resistant complementary learning method, named Dual Contrastive Learning (DCL), for cross-modal matching with noisy correspondence. Different from previous methods, DCL aims at exploring the positive contribution of negative pairs and pursuing stable and consistent performance under various noise ratios with the combination of both positive and complementary learning. The framework is illustrated in Fig. 2. Firstly, we partition the data into clean and noisy subsets by modeling the per-sample loss distribution of the dataset through a Beta-Mixture-Model (BMM). The Beta Mixture Model (BMM) demonstrates superior modeling capability for the skewed loss distribution of clean data when compared to conventional Gaussian-Mixture-Model (GMM). Secondly, we adopt a vanilla cross-modal contrastive learning on the clean subset to learn common representations by maximizing the mutual information between well-matched data. Thirdly, complementary contrastive learning is adopted to the negative pairs. Since selecting incorrect complementary pairs has a much lower likelihood than selecting correct pairs in a noisy dataset, complementary learning could reduce the risk of providing incorrect supervision and smooth the losses. In addition, different from RCL [17], the negative pairs in this paper consist of both the selected noisy subset and the original mismatched pairs. Finally, our method can beyond the noisy correspondence that trains a robust cross-modal matching model on both the matched and mismatched set with positive and negative learning, respectively. Such paradigm contributes significantly to the consistent and stable performance under different levels of noise ratios. Extensive experiments affirm that the proposed DCL demonstrates notably superior performance against the existing methods.

The main contributions are summarized as follows:

- We propose a stable and noise-resistant Dual Contrastive Learning (DCL) paradigm for robust cross-modal matching with correspondence. Our DCL highlights the challenge of different levels of noise ratios and explores both positive and negative learning to leverage the contributions of matched and mismatched pairs to tackle the challenge.
- We adopt complementary contrastive learning to fully explore the potential positive contribution of negative pairs, which avoids the noisy accumulation of noise-rectify methods.
- Extensive experiments and analysis are conducted on MS-COCO, Flickr30K, and CC120K. The results indicate that DCL significantly outperforms the state-of-the-art methods and exhibits superior stability with extremely lower variances of R@1.

2. Related works

2.1. Image-text matching

Cross-modal matching [18–20] focuses on mapping data to a common feature space to measure the similarity scores of image-text pairs. For instance, SCAN [21] proposes stacked cross attention in order to align image regions and words. SGRAF [22] proposes SGR and SAF modules to compute the final similarity scores by graph inference and attention weighting respectively. Considering the effect of negatives in optimization, VSE++ [23] improves the triplet loss with hard negative pairs. Similarly, NAAF [24] improves the accuracy of the model by separating the distributions of matched and unmatched word-region similarities. To aggregate the local features, a generalized pooling operator (GPO) [25] is proposed to learn adaptive weights for different pooling strategies, which achieves promising pooling performance. Meanwhile, Pan et al. proposed to perform fine-grained image-text matching by Cross-modal Hard Aligning Network (CHAN) [26]. They identified a limitation in cross-attention mechanisms: the generation of redundant or irrelevant region-word alignments, which leads to declining retrieval accuracy and limiting efficiency. Therefore, CHAN exploits the most relevant region-word pairs and eliminates all other alignments. In addition to these matching models, cross-model hashing [27–30] has been widely explored to improve retrieval efficiency. Since pre-trained visual-language models [4,31,32] have demonstrated strong cross-modal learning and zero-shot learning performance in recent years, CLIP [4] is becoming increasingly popular in cross-modal matching. However, the above methods cannot perform well on datasets with noisy correspondence.

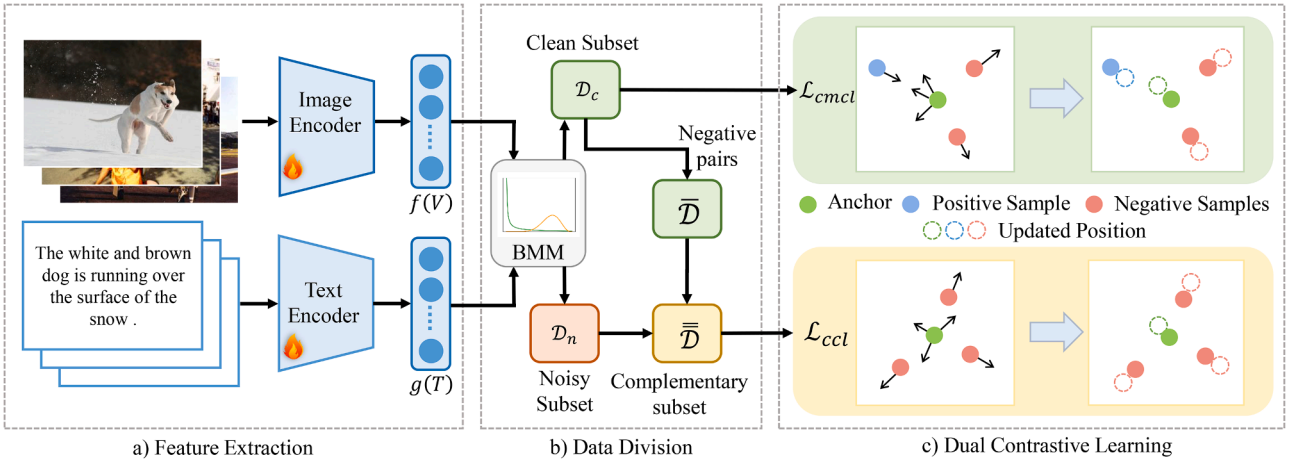


Fig. 2. The framework of our proposed method, which consists of three components: a) feature extraction, b) data division by BMM, and c) dual contrastive learning, including vanilla contrastive learning on positive pairs and complementary contrastive learning on negative pairs.

2.2. Noisy correspondence learning

Huang et al. [11] first introduced noisy correspondence to cross-modal retrieval, and proposed NCR to solve this problem. Inspired by the memorization effect of DNNs, NCR divides the dataset into clean and noisy subsets and predicts soft labels. Finally, it employs soft triplet loss to achieve robust cross-modal matching. Following NCR, BiCro [12] adopts BMM to fit the distribution of the loss, and computes the labels based on the bidirectional cross-modal consistency. CTPR [33] and CREAM [34] consider that the pairs with partial relevant semantic are difficult to be divided correctly, then split the dataset into three subsets. Different from these NCR-based methods, MSCN [35] proposes a meta leaning module to map similarity scores into soft labels. In addition, DECL [8] introduces a deep evidence learning paradigm to capture uncertainty induced by noise. Moreover, RCL [17] prevents the model from overfitting to the noise by exploiting complementary negative pairs, rather than matched pairs. Zha et al. [36] presented UCPM, which introduces Uncertainty Guided Division (UGD) strategy to divide the corrupted training data into confident matched (clean), easily-identifiable mismatched (noisy) and hardly-determined hard subsets.

To address the problems of excessive memorizing/overfitting and unreliable correction, Cross-modal Robust Complementary Learning (CRCL) [37], which combined a novel active complementary loss and an efficient self refining correspondence correction. In [38], Liu et al. introduced a novel Self-Drop and Dual-weight (SDD) approach, which firstly partitions data into four types: clean and significant, clean yet insignificant, vague, and noisy. Moreover, SDD employs self-drop to discard noisy samples to effectively mitigate the impact of noise. Duan et al. [39] introduced Pseudo-Classification based Pseudo-Captioning (PC^2), which includes threefold strategies. Firstly, PC^2 establishes an auxiliary "pseudo-classification" task that steers the model to learn image-text semantic similarity through a non-contrastive mechanism. Secondly, it generate pseudo-captions to provide more informative and tangible supervision for each mismatched pair. Then, correspondence correction is conducted with the assistance of the oscillation of pseudo-classification. Similarly, Relation Consistency (ReCon) [40] is proposed to discriminate the true correspondences and mitigate the adverse impact caused by mismatches, which leverages a novel relation consistency learning to ensure the dual-alignment. Meanwhile, a Geometrical Structure Consistency (GSC) method [13] is introduced, which maintains both intra-modal and cross-modal geometric relationships through structure-preserving constraints, allowing for the accurate discrimination of noisy samples based on structural differences. Then, GSC refines the learning of geometrical structures. To avoid the unstable performance of noise-rectify methods, NPC [16] builds a memory bank to measure the negative effect of data pairs and proposes a negative pre-aware and re-weighting paradigm.

However, the performance of these methods degrades significantly as the noise rate increases. In this work, we maintain the model stability by utilizing both positive and negative learning.

2.3. Noisy multimodal contrastive learning

Noisy multimodal contrastive learning is an effective paradigm for representation learning by discarding semantically irrelevant information and fusing multi-modal information. Ge et al. [41] proposed that CNNs depend on low-level features with insufficient semantics, which damages the robustness of the model. Texture-based and patch-based augmentations are applied to construct negative samples that only maintain redundant information. Therefore, the approach achieves a better classification accuracy and generalization ability. MMCL [42] proposes a contrastive learning framework which combines uni-modal contrastive coding and cross-modal contrastive prediction. It distracts robust uni-modal representations and capture inter-modal interactions. Furthermore, MMCL improves the performance of prediction network with instance-based and sentiment-aware tasks to maintain sentiment-related dynamics. Guo et.al[43]. proposed that different learning paces of positive and negative pairs can also negatively impact model performance. They introduced PN-NCE, a pace-adaptive multi-modal contrastive objective that addresses the problems of imbalanced paces and false labels. Then, it generates more modality-invariant fusion outputs by measuring the distance between the fused representation and the uni-modal representations. M^3ixup [44] employs multiple modal fusion modes to learn multi-modal semantics and enhances the robustness against missing modalities. Firstly, uni-modal and multi-modal mixup strategies are utilized to enhance the representations while improving robustness against missing modalities. Then, it extends the mixed strategies to the contrastive learning loss function, further aligning the multi-modal and original representations. Han et al. [45] proposed a two-step framework that is robust against noisy labels and noisy correspondence. A noise estimation component combined with category-level contrastive loss is introduced to mitigate consistency between different modals. The following hybrid-supervised component measures distance among features as refined labels and guides training progress. This method achieves superior performance on both synthetic and real datasets.

Different from the previous methods, our proposed DCL adopt complementary contrastive learning to fully explore the potential positive contribution of negative data.

3. Methodology

3.1. Preliminary

3.1.1. Problem definition

Cross-modal matching aims to increase the similarity of positive pairs, while decreasing the similarity of negative pairs. Given a image-text dataset $D = \{V, T, Y\}$, where $V = \{V_i\}_{i=1}^{N}$ is the visual training set with N samples, $T = \{T_i\}_{i=1}^{N}$ is the corresponding text set with N samples, and $Y \in \{0,1\}^{N\times N}$ is a matrix representing the correspondence labels, where $Y_{ij} = 1$ indicates (V_i, T_j) is matched, otherwise $Y_{ij} = 0$. Generally, the given dataset is regarded as well-matched, i.e., all diagonal elements of Y are 1, and other elements are 0. However, there inevitably exist some unknown mismatched pairs are annotated as matched, i.e., noise correspondence. Hence, for easy reading, we set the initial correspondence label matrix with noise as $\widetilde{Y}$, and the predicted true label matrix as Y. For computing the similarity score, We project the images and texts into a common feature space via image encoder $f(\cdot, \Theta_f)$ and text encoder $g(\cdot, \Theta_g)$, where Θ_f, Θ_g are the parameters of the two encoders. Therefore, the similarity between V_i and T_j could be calculated as follows,

$$S_{i,j} = \frac{f(V_i) \cdot g(T_j)}{\|f(V_i)\|\|g(T_j)\|}. \tag{1}$$

The goal of our method is to design a paradigm that enforces the model to learn high-quality image encoder $f(\cdot, \Theta_f)$ and text encoder $g(\cdot, \Theta_g)$ that are robust to the noisy correspondence, even under a high noise ratio.

3.2. Data division based on BMM

To learn a robust model against noisy correspondence, we first divide the dataset into a clean subset and a noisy subset as shown in Fig. 2. Previous work has proposed the memorization effect of DNNs which infers the phenomenon that DNNs tend to learn the meaningful patterns from clean data first, and then gradually fit the noisy data. In other words, during the early stages of training, losses are lower for clean data and higher for noisy data.

Inspired by contrastive learning, we consider the cross-modal matching as an N-way classification, i.e., each sample in the dataset is a category. The decision function from visual modality to text modality is defined as $h : V \xrightarrow{T} \mathbb{R}^N$. Similarly, the decision function from text to image is represented as $h : T \xrightarrow{V} \mathbb{R}^N$. Given a query sample V_i, the cross-modal matching probability of the textual sample T_j with respect to V_i is formulated as:

$$p_{ij}^{v2t} = p(Y_{ij} = 1 \mid V_i, T_j) = h(V_i, T_j) = \frac{\exp(S_{ij}/\tau)}{\sum_{k=1}^{N}\exp(S_{ik}/\tau)}, \tag{2}$$

where τ is a temperature parameter. Similarly, we compute the matching probability of V_i w.r.t T_j by:

$$p_{ij}^{t2v} = p(Y_{ji} = 1 \mid T_i, V_j) = h(T_i, V_j) == \frac{\exp(S_{ji}/\tau)}{\sum_{k=1}^{N}\exp(S_{ki}/\tau)}. \tag{3}$$

Since computing the matching probabilities over the entire training set is expensive, we randomly sample a mini-batch data $\mathcal{M}$ with M pairs with index $\{j_k\}_{k=1}^{M}$. Then, we can estimate $h(\cdot,\cdot)$ via Monte Carlo approximation:

$$h(V_i, T_j) = \frac{\exp(S_{ij}/\tau)}{\frac{N}{M}\sum_{k=1}^{M}\exp(S_{ij_k}/\tau)}, \tag{4}$$

$$h(T_i, V_j) = \frac{\exp(S_{ji}/\tau)}{\frac{N}{M}\sum_{k=1}^{M}\exp(S_{j_k i}/\tau)}, \tag{5}$$

The goal of cross-modal matching is to learn a model that minimizes the risk of decision function $h(\cdot,\cdot)$. Thus, given the data pair (V_i, T_i), the risk could be approximated by:

$$\ell_i = \mathcal{L}(h(V_i, T_i), \widetilde{Y}_{ii}) + \mathcal{L}(h(T_i, V_i), \widetilde{Y}_{ii}), \tag{6}$$

where $\mathcal{L}(\cdot,\cdot)$ is cross-entropy loss.

Then, we can divide the data into clean and noisy subsets according to the decision loss. Most of the previous studies employs Gaussian Mixture Model (GMM) to model the distribution of per-sample loss. However, since the loss of clean data is very close to 0, the Beta Mixture Model (BMM) can fit the loss distribution better than GMM [12]. Therefore, we adopt a two-component BMM to fit the distribution of per-sample loss of the training data. The overall probability density function is:

$$p(\ell) = \sum_{k=0}^{1}\lambda_k p(\ell \mid k), \tag{7}$$

where λ_k refers to the mixture coefficient, and $p(\ell \mid k)$ is the probability density function of k-th component:

$$p(\ell \mid \gamma, \beta) = \frac{\Gamma(\gamma+\beta)}{\Gamma(\gamma)\Gamma(\beta)}\ell^{\gamma-1}(1-\ell)^{\beta-1}, \tag{8}$$

where $\gamma, \beta > 0$, $\Gamma(\cdot,\cdot)$ is the Gamma function. We use an Expectation Maximization procedure to fit the distribution. Then, the probability for i-th pair to be clean or noisy is defined as:

$$p(k \mid \ell_i) = \frac{p(k)p(\ell_i \mid k)}{p(\ell_i)}, \tag{9}$$

where $k = 0/1$ denotes clean/noisy class. Given a threshold δ, we can select the clean subset as:

$$\mathcal{D}_c = \{(V_i, T_i, Y_{ii} = 1) \mid p(k = 0 \mid \ell_i) > \delta, \forall (V_i, T_i) \in \mathcal{M}\}, \tag{10}$$

and the noisy subset as:

$$\mathcal{D}_n = \{(V_i, T_i, Y_{ii} = 0) \mid p(k = 0 \mid \ell_i) \le \delta, \forall (V_i, T_i) \in \mathcal{M}\}, \tag{11}$$

where their sizes are denoted as N_1 and N_2, respectively.

3.3. Vanilla cross-modal contrastive learning

We propose to adopt vanilla cross-modal contrastive learning on the clean subsets, which encourages the model to pull together positive pairs and push away negative pairs.

Based on the decision functions in Eqs. (4) and (5), the purpose of cross-modal matching is to train a model minizing the matching risk stated as follows:

$$\begin{aligned}\mathcal{R}(h, \mathcal{L}) = {} & \mathbb{E}_{(V_i, Y_{i\cdot})\sim D}[\mathcal{L}(h(V_i, \cdot), Y_{i\cdot})] \\ & + \mathbb{E}_{(T_i, Y_{\cdot i})\sim D}[\mathcal{L}(h(T_i, \cdot), Y_{\cdot i})].\end{aligned} \tag{12}$$

Given the separated clean data $\mathcal{D}_c$, the risk can be approximated as:

$$\hat{\mathcal{R}}_c(h, \mathcal{L}) \simeq \frac{1}{N_1}\sum_{i=1}^{N_1}[\mathcal{L}(h(V_i, \cdot), Y_{i\cdot}) + \mathcal{L}(h(T_i, \cdot), Y_{\cdot i})]. \tag{13}$$

where the cross-entropy is employed as the loss function $\mathcal{L}$. Then, in each mini-batch, the loss for the clean subset is defined as:

$$\mathcal{L}_{cmcl} = -\frac{1}{N_1}\left(\sum_{p\in P_+^{v2t}}\log p + \sum_{p\in P_+^{t2v}}\log p\right) \tag{14}$$

where $P_+^{v2t} = \{p_{ii}^{v2t} \mid Y_{ii} = 1; i = 1, \ldots, N_1\}$ and $P_+^{t2v} = \{p_{ii}^{t2v} \mid Y_{ii} = 1; i = 1, \ldots, N_1\}$.

Obviously, $\mathcal{L}_{cmcl}$ is a cross-modal contrastive loss that can fully exploit the informative and valuable clean data to avoid the problem of underfitting.

3.4. Complementary contrastive learning

Most of the previous methods predict soft correspondence labels for both clean and noisy data pairs, and train models by minimising the triplet loss with soft margin. However, the triplet loss considers only one pair of mismatched pairs to push each other away, which makes the optimization unstable. To this end, we propose to adopt complementary contrastive learning with multiple mismatched pairs to improve the stability of model optimization.

Inspired by complementary learning [17,46,47], the complementary information is introduced to improve the robustness of the model against mismatched noisy pairs. First, we define a complementary set consisting of all original negative pairs and noisy pairs as follows:

$$\overline{\overline{D}} = \{(\mathcal{D}_n, \overline{D}), \overline{Y}\} = \{(V_i, T_j, \overline{Y}_{ij} = 1)\}. \tag{15}$$

where $\overline{Y}_{ij} = 1$ indicates (V_i, T_j) is unmatched pairs, $\overline{Y}$ is the complementary matrix of Y in a batch of size M, and $\overline{D}$ is the original complementary pairs in a batch $\mathcal{M}$. Since the mismatching probability of the complementary negative pairs of the selected noisy pairs is remarkably larger than the matching probability, we keep these complementary negative pairs. Hence, even the noisy pairs are wrongly selected, it has slight impact on complementary contrastive learning. Then, the size of the complementary set is $\widetilde{N} = M(M-1) + N_2$.

Afterward, based on the decision functions aforementioned, the matching risk of complementary learning on $\overline{\overline{D}}$ is defined as:

$$\hat{\mathcal{R}}_n(h, \overline{\mathcal{L}}) \simeq \frac{1}{\widetilde{N}} \sum_{i=1}^{\widetilde{N}} [\overline{\mathcal{L}}(h(V_i, \cdot), \overline{Y}_{i\cdot}) + \overline{\mathcal{L}}(h(T_i, \cdot), \overline{Y}_{\cdot i})], \tag{16}$$

where $\overline{\mathcal{L}}$ is a complementary loss. We adopt the negative learning loss to optimize Eq. (16), which is shown to be robust to noise [17]. Then the optimization can be transformed to minimize the loss function as follows:

$$\mathcal{L}_{ccl} = -\frac{1}{\widetilde{N}} \sum_{i=1}^{\widetilde{N}} \sum_{p \in P_-} \log(1-p) \tag{17}$$

where $P_- = P_-^{v2t} \cup P_-^{t2v}$, $P_-^{v2t} = \{p_{ij}^{v2t} \mid \overline{Y}_{ij} = 1; i, j = 1, \ldots, M\}$ and $P_-^{t2v} = \{p_{ij}^{t2v} \mid \overline{Y}_{ij} = 1; i, j = 1, \ldots, M\}$. It is noted that Eq. (17) is the instance-level complementary variant of negative learning loss with multiple negatives.

3.5. Objective function

Combining Eqs. (14) and (17), the overall objective function of dual contrastive learning can be formulated as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{cmcl} + \lambda_2 \mathcal{L}_{ccl}, \tag{18}$$

where λ_1 and λ_2 are hyper-parameters to balance the two losses. By minimising the Eq. (18), the model parameters can be updated. The detailed training pseudo-code is shown in Algorithm 1.

Algorithm 1 Dual contrastive learning.

Input: Training dataset D, pre-trained CLIP backbone as $f(\cdot, \Theta_f)$ and $g(\cdot, \Theta_g)$, threshold δ, balancing parameters λ_1, λ_2, batch size M, number of batches in an epoch *num_steps*, number of max epoch *epochs*
Output: The learned parameters Θ_f, Θ_g
for $i = 1 : epochs$ **do**
  Initialize the encoders parameters Θ_f, Θ_g;
  for $j = 1 : num\_steps$ **do**
    Fetch a mini-batch data $\mathcal{M}$ with size M;
    Extract visual and textual features via $f(\cdot, \Theta_f)$ and $g(\cdot, \Theta_g)$;
    Divide the data into clean subset $\mathcal{D}_c$ and noisy subset $\mathcal{D}_n$ by BMM via Eqs. (9)–(11);
    Construct the complementary data $\overline{\overline{D}}$ via Eq. (15);
    For positive data $\mathcal{D}_c$, calculate the vanilla contrastive loss via Eq. (14);
    For negative data $\overline{\overline{D}}$, calculate the complementary contrastive loss via Eq. (17);
    Calculate the overall loss $\mathcal{L}$ by Eq. (18);
    Update the parameters Θ_f, Θ_g by minimizing $\mathcal{L}$ with backpropagation;
  end for
end for

4. Experiments

4.1. Experimental setting

4.1.1. Datasets and evaluation metrics.

We conduct experiments on three widely-used benchmark datasets, MS-COCO [48], Flickr30K [49] ans Concaptual Captions[50]:

MS-COCO includes 123,287 images with 5 annotations per image. Following previous works [11], we use 5000 images for validation, 5000 images for testing, and 113,287 images for training.

Flickr30K includes 31,783 images with 5 annotations per image. Following previous works [11], we use 1000 images for validation, 1000 images for testing, and 29,783 images for training.

Conceptual Captions includes about 3M image-text pairs. Since the data are harvested from the Internet, there are about 3% ∼ 20% mismatched pairs. Following [16], we use a subset of Conceptual Captions, i.e. CC120K. This subset contains 118,851 images for training, 1000 images for validation, and 1000 images for testing.

4.1.2. Evaluation metrics.

We evaluate the cross-modal matching performance using the widely-used metric Recall@K (R@K). We report R@1, R@5, R@10 and their sum (i.e. rSum) to measures the overall performance. The variance (var) of R@1 at different noise ratios is also used to evaluate the stability of cross-modal matching, with lower var indicating higher stability.

4.1.3. Implementation details.

DCL can be used in many cross-modal matching frameworks. In this study, CLIP [4] with ViT-B/32 is used as a backbone. We initialize all parameters from the official pre-trained weights. Both image and text feature dimensions are 512. All experiments are conducted on an RTX 3090 GPU and AdamW [51] is used as the optimizer. The initial learning rate is 1e-6. We set the threshold δ to 0.5. The balance parameters of the loss function are $\lambda_1 = 0.2$ and $\lambda_2 = 128$, respectively. We train the model for 5 epochs with a mini-batch size of 128.

4.2. Comparison with state of the arts

To demonstrate the effectiveness of DCL, we compare it with a series of robust models against noisy correspondence. The baselines include NCR [11], DECL [8], RCL [17], L2RM [52], BiCro [12], CRCL [37], PC^2 [39] , ReCon [40] and NPC [16]. In addition, we fine-tune CLIP with ViT-B/32 as a baseline. The results are cited from the original papers and [52]. Since our method is motivated by RCL, we re-implement RCL with CLIP as backbone, i.e., RCL-CLIP, for fair comparison. To evaluate the performance of various methods at different levels of noise, the noise ratio increases from 20% to 80% at intervals 20%. Since Flickr30K and MS-COCO are well-annotated datasets, we inject noisy correspondence by randomly shuffling images for a specific percentage as the way in [11,12,16].

The results are reported in Table 1. Overall, we can observe that the proposed DCL achieves the best performance with different levels

Table 1
Image-text matching on MS-COCO 1K and Flickr30K.

		MS-COCO 1K							Flickr30K						
		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%	NCR	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
	BiCro	76.6	95.4	98.2	61.3	88.8	94.8	515.1	74.7	94.3	96.8	56.6	81.4	88.2	492.0
	DECL	77.5	95.9	98.4	61.7	89.3	95.4	518.2	77.5	93.8	97.0	56.1	81.8	88.5	494.7
	RCL	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
	L2RM	80.2	96.3	98.5	64.2	90.1	95.4	524.7	77.9	95.2	97.8	59.8	83.6	89.5	503.8
	PC^2	77.8	95.7	98.4	62.8	89.7	95.3	519.7	78.7	94.9	96.9	59.8	83.9	89.6	503.8
	CRCL	79.6	96.1	98.7	64.7	90.6	95.9	525.6	77.9	95.4	98.3	60.9	84.7	90.6	507.8
	ReCon	80.9	96.6	98.8	65.2	91.0	96.0	528.6	80.3	95.3	97.8	61.6	85.5	91.3	511.8
	CLIP	75.0	93.1	97.2	58.7	86.1	97.2	507.3	82.3	95.5	98.3	66.0	88.5	93.5	524.1
	RCL-CLIP	82.1	96.1	**99.0**	**68.2**	92.1	96.6	**534.1**	**89.4**	**98.3**	**99.2**	73.5	91.6	95.3	547.3
	NPC	79.9	95.9	98.4	66.3	90.8	**98.4**	529.7	87.3	97.5	98.8	72.9	92.1	**95.8**	544.4
	Ours	**82.2**	**96.8**	98.9	66.7	**92.2**	96.8	533.6	89.0	97.6	99.1	**74.2**	**92.2**	95.4	**547.5**
40%	NCR	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
	BiCro	75.2	95.3	98.1	60.0	87.8	94.3	510.7	70.7	92.0	95.5	51.9	77.7	85.4	473.2
	DECL	75.6	95.5	98.3	59.5	88.3	94.8	512.0	72.7	92.3	95.4	53.4	79.4	86.4	479.6
	RCL	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
	L2RM	77.5	95.8	98.4	62.0	89.1	94.9	517.7	75.8	93.2	96.9	56.3	81.0	87.3	490.5
	PC^2	77.4	95.8	98.4	62.1	89.4	95.1	518.2	75.8	93.5	96.9	57.5	81.9	88.2	493.8
	CRCL	78.2	95.7	98.3	63.3	90.3	95.7	521.5	77.8	95.2	98.0	60.0	84.0	90.2	505.2
	ReCon	79.9	96.2	98.6	63.5	90.5	95.9	524.5	79.4	94.3	97.6	59.9	83.9	90.1	505.2
	CLIP	70.7	91.7	96.2	54.7	83.4	96.2	492.9	76.2	93.3	96.5	59.4	85.0	90.9	501.3
	RCL-CLIP	82.1	**96.4**	98.6	66.9	91.1	95.8	530.9	88.8	**97.9**	**99.1**	73.1	91.5	94.8	545.2
	NPC	79.4	95.1	98.3	65.0	90.1	**98.3**	526.2	85.6	97.5	98.4	71.3	91.3	95.3	539.4
	Ours	**82.2**	96.3	**99.1**	**67.1**	**91.4**	96.6	**532.7**	**89.8**	**97.9**	99.0	**74.9**	**92.4**	**95.6**	**549.6**
60%	NCR	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
	BiCro	73.2	93.9	97.6	57.5	86.3	93.4	501.9	64.1	87.1	92.7	47.2	74.0	82.3	447.4
	DECL	73.0	94.2	97.9	57.0	86.6	93.8	502.5	65.2	88.4	94.0	46.8	74.0	82.2	450.6
	RCL	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
	L2RM	75.4	94.7	97.9	59.2	87.4	93.8	508.4	70.0	90.8	95.4	51.3	76.4	83.7	467.6
	PC^2	74.2	94.4	97.8	58.9	87.5	93.8	506.6	70.8	90.3	94.4	53.1	79.0	85.9	473.5
	CRCL	76.3	95.1	97.9	60.8	89.0	95.1	514.2	73.1	93.4	95.8	54.8	81.9	88.3	487.3
	ReCon	77.2	95.9	98.4	61.8	89.3	95.2	517.8	74.3	93.6	96.6	55.7	81.6	88.1	489.9
	CLIP	67.0	88.8	95.0	49.7	79.6	95.0	475.1	66.3	87.3	93.0	52.1	78.8	87.4	464.9
	RCL-CLIP	78.7	95.5	98.4	65.5	89.8	94.6	522.5	86.3	97.1	98.6	70.8	90.2	94.0	537.0
	NPC	78.2	94.4	97.7	63.1	89.0	**97.7**	520.1	83.0	95.9	98.6	68.1	89.6	94.2	529.4
	Ours	**81.5**	**96.3**	**98.9**	**66.9**	**91.3**	96.4	**531.3**	**89.3**	**98.0**	**99.0**	**73.2**	**92.1**	**95.6**	**547.2**
80%	NCR	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
	BiCro	62.2	88.6	94.6	47.4	79.2	88.5	460.5	2.3	9.2	17.2	2.6	10.2	16.8	58.3
	DECL	64.8	90.5	96.0	49.7	81.7	90.3	473.0	53.4	78.8	86.9	37.6	63.8	73.9	394.4
	RCL	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
	L2RM	69.0	91.9	96.4	52.6	82.4	90.3	482.6	55.7	80.8	87.8	39.4	65.4	74.9	404.0
	CRCL	72.7	93.5	97.6	57.5	86.8	93.7	501.8	62.3	86.8	92.8	46.0	73.6	82.2	443.7
	CLIP	62.9	86.4	92.8	43.6	75.1	86.2	447.0	61.9	86.4	91.9	45.0	73.1	82.3	440.6
	RCL-CLIP	52.2	80.2	88.1	40.6	68.6	78.9	408.6	62.5	83.5	89.7	46.5	70.2	77.3	429.7
	NPC	73.3	92.0	97.9	58.8	86.5	94.1	502.6	80.6	95.8	97.7	63.6	86.6	91.7	516.0
	Ours	**80.4**	**95.6**	**98.5**	**66.0**	**90.5**	**96.1**	**527.1**	**87.6**	**97.8**	**99.0**	**72.7**	**91.2**	**94.8**	**543.1**

of noise on both datasets. Notably, DCL outperforms the current state-of-the-art approach NPC with a large margin of R@1 on Flickr30K with 80% noise ratio. To be specific, R@1 of our DCL is 7% and 9.1% higher than NPC on image-to-text and text-to-image, respectively. Compared to the similar RCL that also adopts complementary learning, our method significantly outperforms it with remarkable margins, especially under high noise ratio. This is because RCL only utilizes the complementary negative pairs and ignores the informative clean pairs leading the model to be underfitting. In addition, the extended CRCL is also inferior to our method, which further indicates the effectiveness of our dual contrastive learning against noise correspondence.

Moreover, we can draw a conclusion that our DCL achieves stable and noise-resistant performance on Flickr30K and MS-COCO. Specifically, as the noise ratio increases from 20% to 80%, the rSums on Flickr30K and MS-COCO decrease by 4.4% and 6.5% respectively. On the contrary, the performance of other methods decreases dramatically, *e.g.*, the rSums of NPC decrease by 28.4% and 27.1% respectively. It is worth noting that the R@1 of our method with different datasets and tasks keeps steady with the increasing noise. This phenomenon can be easily found in Fig. 1. The stable results further validate the powerful ability of our method to deal with noisy correspondence under different noise levels, especially on the high noise ratio.

To evaluate the impact of CLIP on our method, we compare our DCL against CLIP, RCL-CLIP, and NPC which are also CLIP-based methods. Comparing the results of RCL-CLIP and the original RCL, we can conclude that the pre-trained CLIP encoders significantly improve the performance. However, even equipped with CLIP, RCL-CLIP is still inferior to NPC, which indicates that the negative pre-aware and re-weighting strategy performs better than only using complementary learning with negative pairs. In contrast, our method consistently achieves the highest

Table 2
Image-text matching on CC120K.

Methods	image-to-text			text-to-image			
	R@1	R@5	R@10	R@1	R@5	R@10	rSum
RCL*	34.2	58.0	71.2	34.2	61.2	71.7	330.5
CLIP	68.8	87.0	92.9	67.8	86.4	90.9	493.8
NPC*	71.9	90	**94.6**	69.3	89.1	94.2	509.1
Ours	**72.2**	**90.2**	94.2	**71.3**	**89.8**	**94.3**	**512.0**

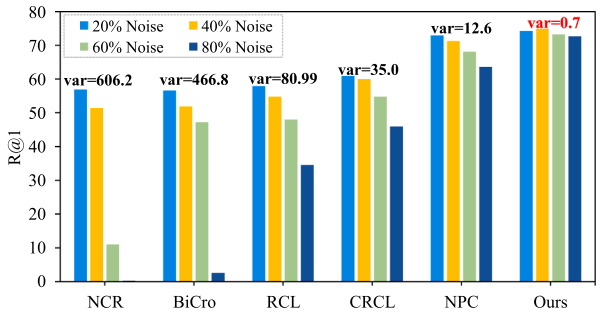


Fig. 3. Stability of different methods with the increase of noise ratio via R@1 of Text-to-Image on Flickr30K.

recall and rSum with the same backbone CLIP, and significantly outperforms NPC, especially for R@1 under high noise ratio. For example, our DCL outperforms NPC with R@1 gains of 7.1% and 7.2% at tasks of image-to-text and text-to-image on MS-COCO, respectively. The comparisons demonstrate that the pre-trained CLIP can benefit the cross-modal matching to some extent. However, the superior performance of our method is mainly attributed to the dual contrastive learning with both positives and negatives.

Since many source images in CC152K are unavailable, it stops us from validating on this dataset due to the utilization of CLIP. Following NPC [16], we conduct experiments on CC120K. The results are shown in Table 2 (* indicates our re-implementation results in the same setting for a fair comparison). Although NPC obtains higher R@10, our DCL still shows overall superiority on this large-scale dataset, which further demonstrates the effectiveness of our divide-and-conquer fashion and complementary learning.

To further evaluate the stability of various methods under different noise ratios, we analyze their R@1 performance on the task of text-to-image on Flickr30K dataset. As illustrated in Fig. 3, the noise-rectify methods, such as NCR and BiCro, have the largest variances indicating pronounced instability under high noise. Due to the consideration of complementary learning, RCL and CRCL can achieve satisfactory but still low performance with the increase of noise. Though NPC outperforms RCL and CRCL, it still suffers from significant performance fluctuations with a relatively large variance. This is because NPC cannot fully utilize the clean and noisy data due to the limited size of memory bank. In contrast, our method combines positive and negative learning together by dual contrastive learning, it can consistently achieve higher recalls with different levels of noise. Combined with Fig. 1, the results collectively demonstrate powerful stable and highly noise-resistant ability of our method under different noise ratios.

To provide explicit evaluation of the proposed DCL, we conduct a case study on the Flickr30K dataset under 20% noise. The visualization results are shown in Fig. 4. Given an image query, we present the top five most relevant textual descriptions ranked by similarity scores in Fig. 4(a). Moreover, given a text query, 5 associated images are presented in Fig. 4(b). The correctly matched items are highlighted with green boxes, while the mismatched samples are highlighted with red boxes for clear performance evaluation. The results reveal that our approach consistently identifies correct matches within the top five retrieved items.

4.3. Ablation study

We further conduct ablation studies on Flickr30K with different noise ratios to demonstrate the effectiveness of three components, i.e, BMM, $\mathcal{L}_{cmcl}$, and $\mathcal{L}_{ccl}$. Here, we employ CLIP as the baseline, i.e., without the three components. The results are reported in Table 3. As can be seen, each component is important for the performance improvement. Specifically, the vanilla cross-modal contrastive learning $\mathcal{L}_{cmcl}$ keeps steady with the increase of noise ratio. This my because we set the threshold as a fixed value. While the complementary contrastive learning $\mathcal{L}_{ccl}$ outperforms the baseline and $\mathcal{L}_{cmcl}$ with noise ratios of 20% and 40%. However, it is worth noting that $\mathcal{L}_{ccl}$ degrades the performance as the noise ratio increases. The reason is that, with the increasing of noise, the negative pairs will contain more and more wrongly mismatched pairs, which bring down the performance of complementary learning. The results also indicate the significant contribution of informative and valuable positive pairs, which is the basic motivation of our method. This is also the reason that our method remarkably outperforms RCL. From the ablation study we can conclude that it is the very combination of dual constrastive learning which leverages both positive and negative learning makes the stable and highly noise-resistant ability of our method.

Table 3
Ablation studies on Flickr30K.

Noise	Methods			image-to-text			text-to-image			
	BMM	$\mathcal{L}_{cmcl}$	$\mathcal{L}_{ccl}$	R@1	R@5	R@10	R@1	R@5	R@10	rSum
20%				82.3	95.5	98.3	66.0	88.5	93.5	524.1
	✓	✓		84.4	96.8	98.4	69.4	90.2	95.3	534.5
	✓		✓	88.2	97.4	99.0	74.1	91.9	**95.4**	546.0
	✓	✓	✓	**89.0**	**97.6**	**99.1**	**74.2**	**92.2**	**95.4**	**547.5**
40%				76.2	93.3	96.5	59.4	85.0	90.9	501.3
	✓	✓		85.6	96.7	98.4	70.0	90.1	94.6	535.4
	✓		✓	87.0	97.3	98.8	72.2	91.2	95.0	541.5
	✓	✓	✓	**89.8**	**97.9**	99.0	**74.9**	**92.4**	**95.6**	**549.7**
60%				66.3	87.3	93.0	52.1	78.8	87.4	464.9
	✓	✓		85.5	96.2	98.2	70.0	90.3	94.2	534.4
	✓		✓	82.7	97.0	98.6	69.7	88.4	92.8	529.2
	✓	✓	✓	**89.3**	**98.0**	**99.0**	**73.2**	**92.1**	**95.6**	**547.2**
80%				61.9	86.4	91.9	45.0	73.1	82.3	440.6
	✓	✓		85.5	96.7	98.4	69.8	90.4	94.1	534.9
	✓		✓	64.1	85.5	90.6	48.3	71.3	78.6	438.4
	✓	✓	✓	**87.6**	**97.8**	**99.0**	**72.7**	**91.2**	**94.8**	**543.1**

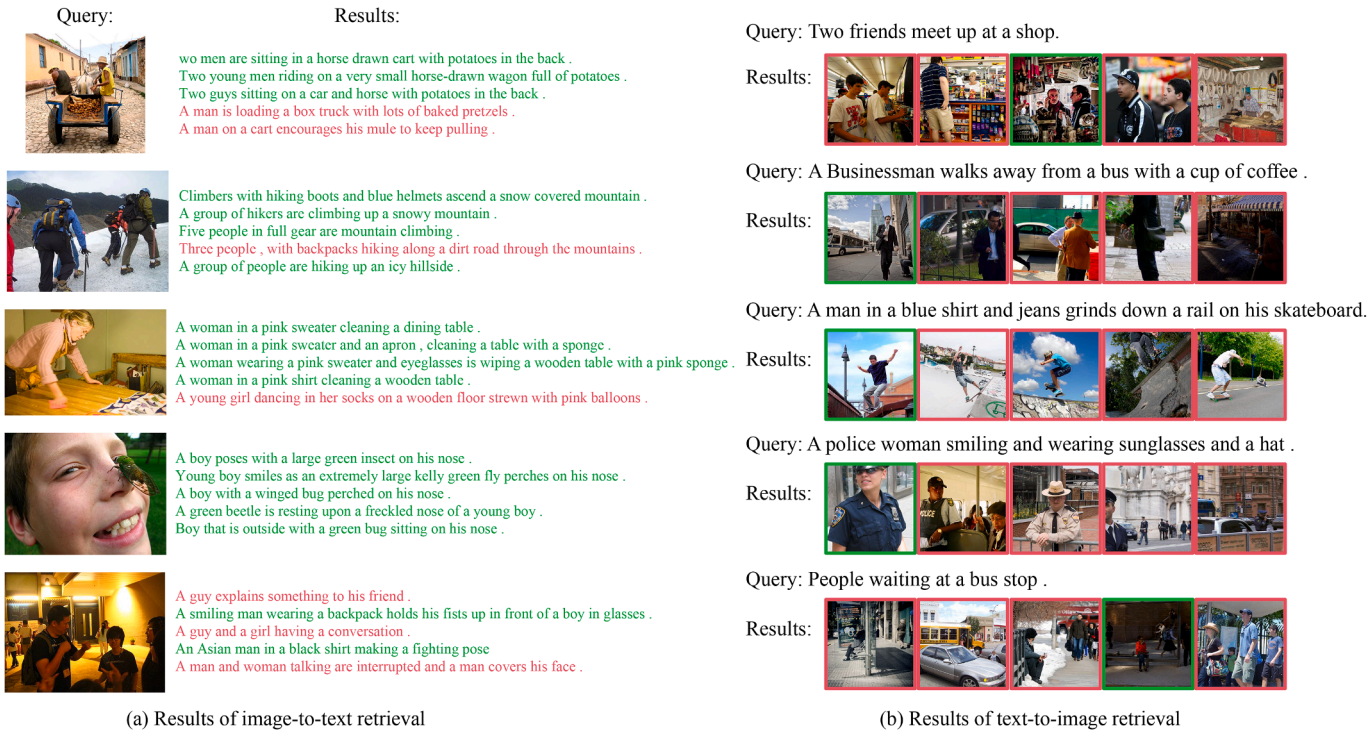
Table 4
Parameter analysis of δ in terms of recall and rSum scores on Flickr30K with only the vanilla contrastive loss.

Threshold	image-to-text			text-to-image			
	R@1	R@5	R@10	R@1	R@5	R@10	rSum
0.2	80.8	95.7	97.8	66.8	88.5	92.9	522.5
0.5	**85.6**	96.7	98.4	70.0	90.1	94.6	535.4
0.8	85.2	**96.9**	**98.5**	**70.9**	**91.1**	**94.8**	**537.4**

Furthermore, we analyze the parameter sensitivity of the selection threshold δ on Flickr30K with a 40% noise ratio. The recalls and rSum are plotted in Fig. 5. From the figure, we can see that when $\delta = 0.5$, DCL achieves the best performance. In addition, we can see that the performance is not very sensitive to the parameter δ. Since we propose to employ the negative data by complementary contrastive learning, it enhances the robustness of our method. Therefore, Fig. 5 shows only a slight fluctuation. To validate this argument, we conduct an external experiment with only the vanilla contrastive loss. The results are presented in Table 4. We can observe that different values (0.2, 0.5, 0.8) of the threshold affect the performance largely (522.5, 535.4, 537.4).

We also report the impacts of hyper-parameters λ_1 and λ_2 in Fig. 6. It is evident that values of λ_1 that are too small or too large result in decreased performance, with the best performance being achieved at a value of 0.1. While the trends corresponding to λ_2 are relatively steady. This may be because λ_1 balances the positive contrastive learning that with few training pairs.

In our method, the overall average inference time is 60.539ms, and the division of BMM costs 1.52ms. Though with CLIP, our method still exhibits high efficiency.



(a) Results of image-to-text retrieval (b) Results of text-to-image retrieval

Fig. 4. Visualization of a case study of the proposed method on Flickr30K with 20% noise ratio.

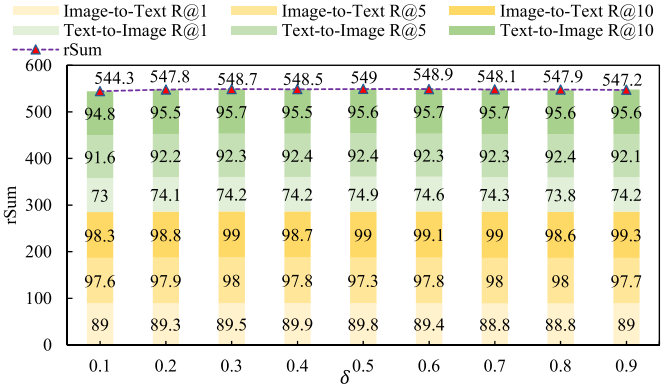


Fig. 5. Parameter analysis of δ in terms of recall and rSum scores on Flickr30K under 40% noise ratio.

Table 5
Impact of division models on Flickr30K.

Noise	BMM/GMM	image-to-text			text-to-image			
		R@1	R@5	R@10	R@1	R@5	R@10	rSum
0%	GMM	87.3	**97.7**	**99.1**	71.4	90.8	95.0	541.3
	BMM	**87.6**	97.2	98.6	**73.2**	**92.2**	**95.6**	**544.4**
20%	GMM	87.9	**97.8**	98.9	73.8	**92.3**	**95.6**	546.3
	BMM	**89.0**	97.6	**99.1**	**74.2**	92.2	95.4	**547.5**
40%	GMM	88.4	97.8	**99.2**	73.9	**92.4**	**95.6**	547.3
	BMM	**89.8**	**97.9**	99.0	**74.9**	**92.4**	**95.6**	**549.6**
60%	GMM	88.5	**98.3**	**98.9**	**73.9**	**92.1**	95.4	547.1
	BMM	**89.3**	98.2	98.8	73.2	**92.1**	**95.6**	**547.2**
80%	GMM	**88.5**	**98.1**	**99.6**	**73.2**	**91.6**	94.1	**545.1**
	BMM	87.6	97.8	99.0	72.7	91.2	**94.8**	543.1

5. Discussion and future work

In order to evaluate the impact of division models, we compare BMM with GMM on Flickr30K across varying noise ratios. The results are reported in Table 5. As can be seen, the overall trend with the increasing noise ratio is that BMM outperforms GMM. From this empirical analysis, it can be concluded that GMM may be appropriate for high noise ratio. The reason may be because GMM directly models the continuous distribution of similarity scores. When high noise causes the score distributions of correct and incorrect pairs to overlap significantly, BMM's binary input becomes unreliable. GMM, however, can use soft probabilistic assignments and the overall shape of the distributions to untangle the mixtures, making it more robust in such ambiguous scenarios. In addition, from the density function of BMM and GMM under 0 noise ratio in Fig. 7, we can observe that BMM fits better than GMM.

Moreover, we also conduct experiments with our method by using the most recent cross-modal alignment model SigCLIP [53] as a backbone. The results on Flickr30K dataset are reported in Table 6. Compared with the results in Table 1, it can be observed that our method using SigCLIP achieves higher performance than using CLIP, which indicates the scalability of our proposed method.

Since the noise ratio is unknown in practice, the threshold δ is set as a fixed value. However, under different noise ratios, the value of δ may potentially compromise the accuracy of data partitioning.

Table 6
Image-text matching on Flickr30K with SigLIP as baseline under 40% noise rate.

Noise	Models	image-to-text			text-to-image			
		R@1	R@5	R@10	R@1	R@5	R@10	rSum
0%	SigLIP	93.2	**99.1**	**99.7**	**82.3**	**96.0**	**98.4**	**568.7**
	Sig-DCL	**93.4**	98.9	**99.7**	81.2	95.4	97.7	566.3
20%	SigLIP	92.5	98.9	99.7	78.9	94.4	97.2	561.6
	Sig-DCL	**93.7**	**99.2**	**99.8**	**80.9**	**95.5**	**97.8**	**566.9**
40%	SigLIP	91.1	98.3	99.7	78.0	94.1	97.1	558.3
	Sig-DCL	**92.7**	**99.1**	**99.8**	**81.2**	**95.6**	**97.7**	**566.1**
60%	SigLIP	89.4	97.9	99.2	74.6	92.4	95.9	549.4
	Sig-DCL	**93.2**	**99.2**	**99.7**	**80.6**	**95.3**	**97.6**	**565.6**
80%	SigLIP	83.9	96.0	98.0	67.7	88.5	93.7	527.8
	Sig-DCL	**91.9**	**98.9**	**99.7**	**80.4**	**95.2**	**97.2**	**563.3**

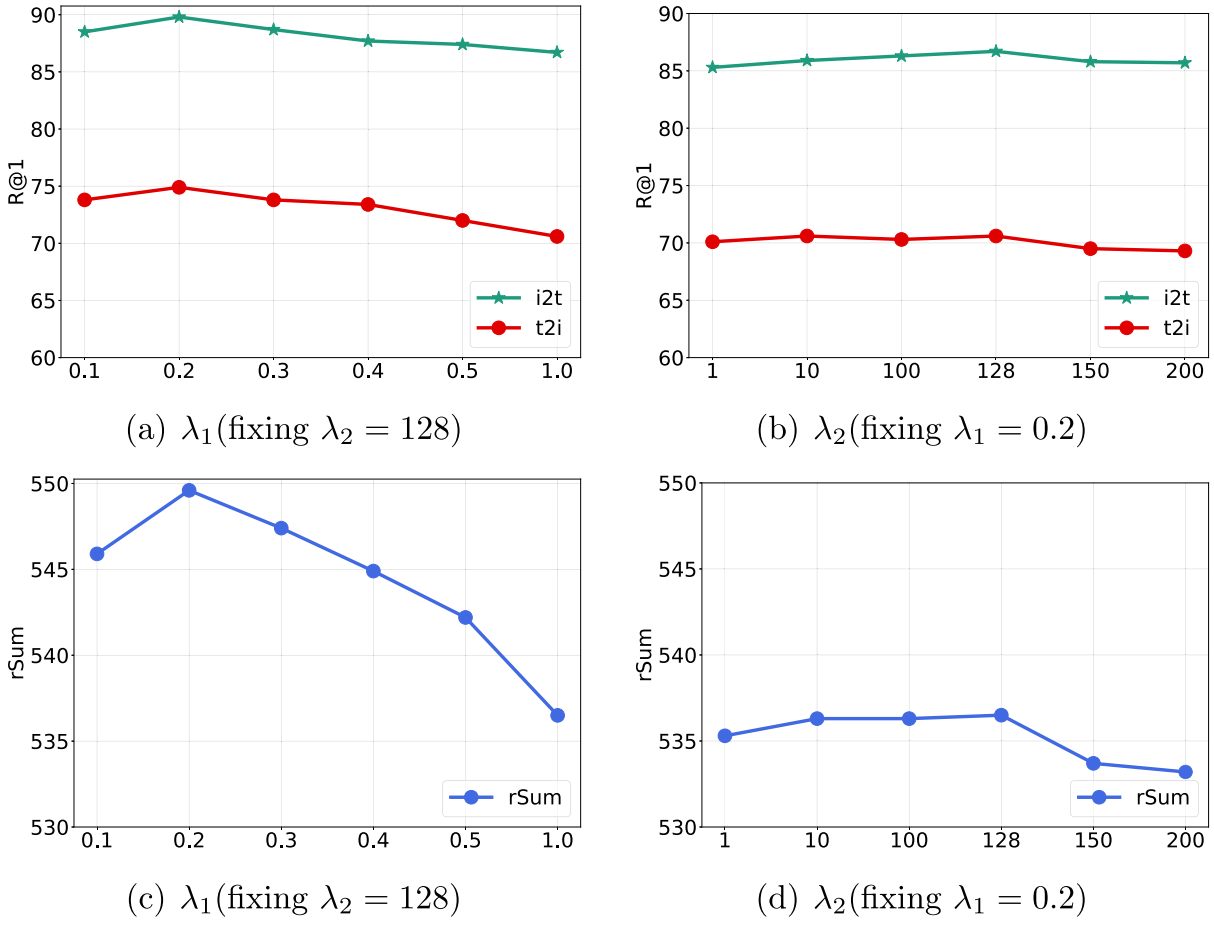


Fig. 6. Parameter analysis of λ_1 and λ_2 in terms of R@1 and rSum scores on Flickr30K under 40% noise ratio.

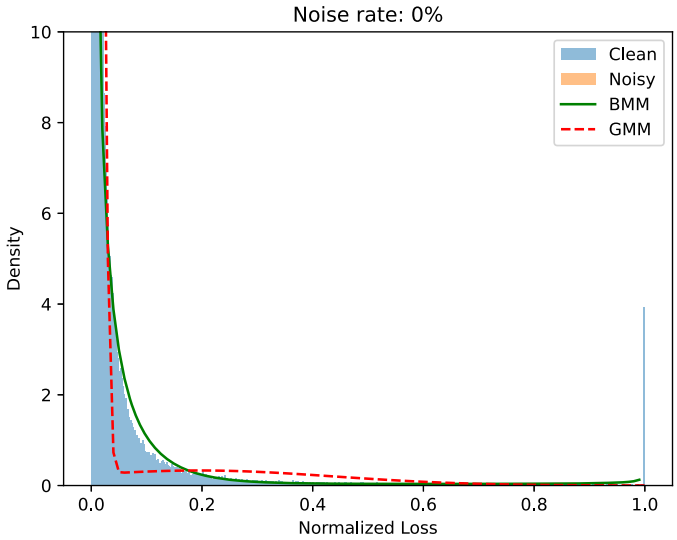


Fig. 7. The density function of BMM and GMM with 0 noise ratio.

Therefore, estimating the noise ratio and determining an adaptive threshold is our focus in the future.

It is worth noting that the noise ratio may be lower in the real-world scenario. Nevertheless, most of the existing methods, including ours, mainly concern the performance under high-noise scenarios. Although it is valuable to study algorithms that can perform well in different noise ratios, we believe that the approaches which are more robust to lower noise ratios will be more practical.

Furthermore, it is an indisputable fact that multimodal large language models (MLLMs) have great potential as backbones or prompts generation in image-text matching tasks. Only employing CLIP as a backbone in our proposed method is far from a full exploration of MLLMs.

6. Conclusion

In this paper, we have proposed a dual contrastive learning paradigm to leverage both matched and mismatched data to realize positive and negative learning in cross-modal matching with noisy correspondence. Combining the positive vanilla contrastive learning and negative complementary contrastive learning, our method can make the best use of both positive and negative information in the dataset. This also encourages our method to maintain promising and stable performance with different levels of noise ratios. Extensive experiments demonstrated the powerful stable and highly noise-resistant ability of our method. Especially, our method achieves small variances of 0.67 and 0.7 on Flickr30K for image-to-text and text-to-image, respectively.

CRediT authorship contribution statement

Fangming Zhong: Writing – review & editing, Visualization, Supervision, Resources, Methodology, Funding acquisition, Conceptualization; **Xinyu He:** Writing – original draft, Visualization, Software, Project administration, Data curation, Conceptualization; **Haiquan Yu:** Validation, Software, Methodology, Investigation; **Xiu Liu:** Software, Investigation, Conceptualization; **Suhua Zhang:** Writing – review & editing, Validation, Formal analysis.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the National Natural Science Foundation of China (Grant No. 62476036, 62006035) and Science and Technology Major Project on Artificial Intelligence of Liaoning Province (2023JH26/10100008).

References

[1] S. Kornblith, L. Li, Z. Wang, T. Nguyen, Guiding image captioning models toward more specific captions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 15259–15269.
[2] L. Li, H. Li, P. Ren, Underwater image captioning via attention mechanism based fusion of visual and textual information, Inf. Fus. 123 (2025) 103269.
[3] S. Li, C. Gong, Y. Zhu, C. Luo, Y. Hong, X. Lv, Context-aware multi-level question embedding fusion for visual question answering, Inf. Fus. 102 (2024) 102000.
[4] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38Th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, 139 of *Proceedings of Machine Learning Research*, PMLR, 2021, pp. 8748–8763.
[5] W. Qin, W. Yu, K. Zhang, H. Zhao, J. Xu, J.R. Wen, Uncertainty-aware evidential learning for legal case retrieval with noisy correspondence, Inf. Sci. (2025) 121915.
[6] J. Chen, R. Zhang, T. Yu, R. Sharma, Z. Xu, T. Sun, C. Chen, Label-retrieval-augmented diffusion models for learning from noisy labels, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10-16, 2023, 2023.
[7] Y. Li, H. Han, S. Shan, X. Chen, DISC: Learning from noisy labels via dynamic instance-specific selection and correction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 24070–24079.
[8] Y. Qin, D. Peng, X. Peng, X. Wang, P. Hu, Deep evidential learning with noisy correspondence for cross-modal retrieval, in: Proceedings of the 30th ACM International Conference on Multimedia, MM '22, 2022, p. 4948–4956. https://doi.org/10.1145/3503161.3547922
[9] Y. Qin, L. Huang, D. Peng, B. Jiang, J.T. Zhou, X. Peng, P. Hu, Trustworthy visual-textual retrieval, IEEE Trans. Image Process. 34 (2025) 4515–4526.
[10] Y. Sun, Y. Qin, Y. Li, D. Peng, X. Peng, P. Hu, Robust multi-view clustering with noisy correspondence, IEEE Trans. Knowl. Data Eng. 36 (12) (2024) 9150–9162.
[11] Z. Huang, G. Niu, X. Liu, W. Ding, X. Xiao, H. Wu, X. Peng, Learning with noisy correspondence for cross-modal matching, in: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, Virtual, 2021, pp. 29406–29419.
[12] S. Yang, Z. Xu, K. Wang, Y. You, H. Yao, T. Liu, M. Xu, BiCro: noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 19883–19892.
[13] Z. Zhao, M. Chen, T. Dai, J. Yao, B. Han, Y. Zhang, Y. Wang, Mitigating noisy correspondence by geometrical structure consistency learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, IEEE, 2024, pp. 27371–27380.
[14] Y. Yang, L. Wang, E. Yang, C. Deng, Robust noisy correspondence learning with equivariant similarity consistency, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, 2024, pp. 17700–17709.
[15] Z. Dang, M. Luo, C. Jia, G. Dai, X. Chang, J. Wang, Noisy correspondence learning with self-reinforcing errors mitigation, in: Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, February 20-27, 2024, Vancouver, Canada, 2024, pp. 1463–1471.
[16] X. Zhang, H. Li, M. Ye, Negative pre-aware for noisy cross-modal matching, in: Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, February 20-27, 2024, Vancouver, Canada, 2024, pp. 7341–7349.
[17] P. Hu, Z. Huang, D. Peng, X. Wang, X. Peng, Cross-modal retrieval with partially mismatched pairs, IEEE Trans. Pattern Anal. Mach. Intell. 45 (8) (2023) 9595–9610. https://doi.org/10.1109/TPAMI.2023.3247939
[18] Z. Luo, P. Zhao, C. Xu, X. Geng, T. Shen, C. Tao, J. Ma, Q. Lin, D. Jiang, LexLIP: Lexicon-Bottlenecked language-image pre-training for large-scale image-text sparse retrieval, in: 2023 IEEE/CVF International Conference on Computer Vision (ICCV), 2023.
[19] C. Huang, W. Liu, J. Wang, J. Cui, J. Wen, Dual-driven cross-modal contrastive hashing retrieval network via structural feature and semantic information, Inf. Fus. 123 (2025) 103252.
[20] Y. Xiu, X. Tong, Dual-layer cross-modal alignment recommendation based on the diffusion model, Inf. Fus. 125 (2026) 103472.
[21] K.H. Lee, X. Chen, G. Hua, H. Hu, X. He, Stacked cross attention for image-text matching, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 201–216.
[22] H. Diao, Y. Zhang, L. Ma, H. Lu, Similarity reasoning and filtration for image-text matching, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Event, February 2-9, 2021, 2021, pp. 1218–1226.
[23] F. Faghri, D.J. Fleet, J.R. Kiros, S. Fidler, VSE++: Improving visual-semantic embeddings with hard negatives, in: Proceedings of the British Machine Vision Conference (BMVC), 2018.
[24] K. Zhang, Z. Mao, Q. Wang, Y. Zhang, Negative-aware attention framework for image-text matching, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 15661–15670.
[25] J. Chen, H. Hu, H. Wu, Y. Jiang, C. Wang, Learning the best pooling strategy for visual semantic embedding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15789–15798.
[26] Z. Pan, F. Wu, B. Zhang, Fine-grained image-text matching by cross-modal hard aligning network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19275–19284.
[27] K. Liu, Y. Gong, Y. Cao, Z. Ren, D. Peng, Y. Sun, Dual semantic fusion hashing for multi-label cross-modal retrieval, in: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, 2024, pp. 4569–4577.
[28] Y. Sun, K. Liu, Y. Li, Z. Ren, J. Dai, D. Peng, Distribution consistency guided hashing for cross-modal retrieval, in: Proceedings of the 32nd ACM International Conference on Multimedia, MM '24, 2024, p. 5623–5632.
[29] Y. Sun, Z. Ren, P. Hu, D. Peng, X. Wang, Hierarchical consensus hashing for cross-modal retrieval, IEEE Trans. Multimedia 26 (2024) 824–836.
[30] D. Zhang, Z. Hu, X.J. Wu, J. Kittler, RSPH: Robust self-paced hashing for cross-modal retrieval, Pattern Recogn. 171 (2026) 112072.
[31] W. Kim, B. Son, I. Kim, ViLT: vision-and-language transformer without convolution or region supervision, in: Proceedings of the 38th International Conference on Machine Learning, 139, 2021, pp. 5583–5594.
[32] J. Li, D. Li, C. Xiong, S. Hoi, BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: Proceedings of the 39th International Conference on Machine Learning, 162, 2022, pp. 12888–12900.
[33] Z. Feng, Z. Zeng, C. Guo, Z. Li, L. Hu, Learning from noisy correspondence with tri-partition for cross-modal matching, IEEE Trans. Multimedia 26 (2024) 3884–3896.
[34] X. Ma, M. Yang, Y. Li, P. Hu, J. Lv, X. Peng, Cross-modal retrieval with noisy correspondence via consistency refining and mining, IEEE Trans. Image Process. 33 (2024) 2587–2598.
[35] H. Han, K. Miao, Q. Zheng, M. Luo, Noisy correspondence learning with meta similarity correction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7517–7526.
[36] Q. Zha, X. Liu, Y. Cheung, S. Peng, X. Xu, N. Wang, UCPM: uncertainty-guided cross-modal retrieval with partially mismatched pairs, IEEE Trans. Image Process. 34 (2025) 3622–3634.
[37] Y. Qin, Y. Sun, D. Peng, J.T. Zhou, X. Peng, P. Hu, Cross-modal active complementary learning with self-refining correspondence, in: Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10-16, 2023, 2023.
[38] F. Liu, C. Dong, C. Zhang, H. Zhou, J. Zhou, Robust noisy correspondence learning via self-drop and dual-weight, CoRR abs/2412.06172 (2024) 1–12.
[39] Y. Duan, Z. Gu, Z. Ying, L. Qi, C. Meng, Y. Shi, PC^2: pseudo-classification based pseudo-captioning for noisy correspondence learning in cross-modal retrieval, in: Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024, ACM, 2024, pp. 9397–9406.
[40] Q. Zha, X. Liu, S.J. Peng, Y.m. Cheung, X. Xu, N. Wang, Recon: enhancing true correspondence discrimination through relation consistency for robust noisy correspondence learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, 2025.
[41] S. Ge, S. Mishra, C.L. Li, H. Wang, D. Jacobs, Robust contrastive learning using negative samples with diminished semantics, in: Advances in Neural Information Processing Systems, 34, 2021, pp. 27356–27368.
[42] R. Lin, H. Hu, Multimodal contrastive learning via uni-modal coding and cross-modal prediction for multimodal sentiment analysis, in: Findings of the Association for Computational Linguistics: EMNLP 2022, 2022, pp. 511–523.
[43] X. Guo, A. Kot, A.W.K. Kong, Pace-adaptive and noise-resistant contrastive learning for multimodal feature fusion, IEEE Trans. Multimedia 25 (2023) 9437–9448.
[44] R. Lin, H. Hu, Adapt and explore: multimodal mixup for representation learning, Inf. Fus. 105 (2024) 102216.
[45] H. Han, Q. Zheng, M. Luo, K. Miao, F. Tian, Y. Chen, Noise-tolerant learning for audio-visual action recognition, IEEE Trans. Multimedia 26 (2024) 7761–7774.
[46] L. Feng, T. Kaneko, B. Han, G. Niu, B. An, M. Sugiyama, Learning with multiple complementary labels, in: International Conference on Machine Learning, PMLR, 2020, pp. 3072–3081.
[47] T. Ishida, G. Niu, W. Hu, M. Sugiyama, Learning from complementary labels, in: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, 2017, pp. 5639–5649.

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

[49] P. Young, A. Lai, M. Hodosh, J. Hockenmaier, From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions, Trans. Assoc. Comput. Linguist. 2 (2014) 67–78.

[50] P. Sharma, N. Ding, S. Goodman, 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, 2018, pp. 2556–2565.

[51] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, 2019.

[52] H. Han, Q. Zheng, G. Dai, M. Luo, J. Wang, Learning to rematch mismatched pairs for robust cross-modal retrieval, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024, 2024, pp. 26669–26678.

[53] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer, Sigmoid loss for language image pre-training, in: IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023, IEEE, 2023, pp. 11941–11952.