Mitigating Noisy Correspondence by Geometrical Structure Consistency Learning

Anonymous CVPR submission

Paper ID *****

Abstract

Noisy correspondence that refers to mismatches in cross-modal data pairs, are prevalent on human-annotated or web-crawled datasets. Prior approaches to leverage such data mainly consider the application of uni-modal noisy label learning without amending the impact on both cross-modal and intra-modal geometrical structures in multimodal learning. Actually, we find that both structures are effective to discriminate noisy correspondence through structural differences when being well-established. Inspired by this observation, we introduce a Geometrical Structure Consistency (GSC) method to infer the true correspondence. Specifically, GSC ensures the preservation of geometrical structures within and between modalities, allowing for the accurate discrimination of noisy samples based on structural differences. Utilizing these inferred true correspondence labels, GSC refines the learning of geometrical structures by filtering out the noisy samples. Our experiments across three well-known cross-modal datasets confirm that GSC effectively identifies noisy samples under various conditions of noisy correspondence, and significantly outperforms the current leading methods.

1. Introduction

Cross-modal retrieval [22, 30, 31] that focuses on querying the most relevant samples across modalities, has garnered considerable interest in multimodal scenarios [3, 39]. Most current methods presuppose that a large quantity of well-annotated data is available. However, real-world datasets [4, 16, 37], often own non-expert annotation or are collected by web crawling, which are prone to noisy correspondence. Such discrepancy can cause severe degradation to retrieval models if without proper handling [15, 34].

Recently, learning with noisy correspondence has gathered increasing attention. The majority of these efforts share a common target of accurately learning the true soft correspondence labels that can reliably indicate the matching degree between data pairs. For example, NCR [15] pioneered this area by employing a co-teaching approach to classify samples with higher losses as noisy. Such method, which has been further refined by [11, 41] by introducing meta-learning and leveraging clean data subsets, however fundamentally remains a variation of the uni-modal sample selection philosophy [25]. They may not be very effective to identify the accurate correspondence and finally overfit on noise in face of the intricacies of multimodal learning.

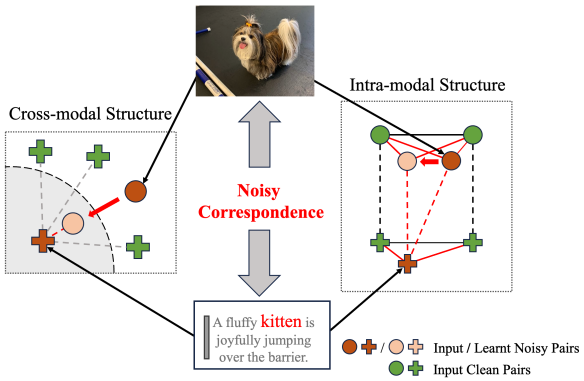

Figure 1. Noisy correspondence impacts both cross-modal and intra-modal geometrical structures. **Left:** Cross-modal distance between mismatched text and image is initially distant but wrongly reduced. **Right:** Intra-modal structures of mismatched image(above) and text(below) are initially distinct but wrongly aligned, thus similar samples within the modality are pulled apart.

Noisy correspondence affects multimodal representations in a more complicated way than uni-modal noisy label problem. As illustrated in Fig. 1, in the perspective of the cross-modal geometrical structure that refers to the similarities between representations across modalities, the presence of noisy correspondence can disrupt this structure by erroneously reducing the distance between mismatched data pairs. On the other hand, in the perspective of the intra-modal geometrical structure that refers to the similarities within the modality, as different samples exhibit asymmetric intra-modal structures, attempting to align the distinct

structures across mismatched modalities can lead to intra-model collapse. Fortunately, we have observed significant differences between clean and noisy samples within well-established cross-modal and intra-modal geometrical structures where clean samples tend to show better alignments, as detailed in Sec. 3.2. These differences conversely offer a promising strategy for discriminating noisy samples and accurately predicting true correspondence labels.

Inspired by the distinct structural characteristics, we introduce the Geometrical Structure Consistency (GSC) method to mitigate the issue of noisy correspondence. Specifically, GSC maintains the integrity of geometrical structures by optimizing a contrastive loss that aligns the intra-modal geometrical structures along with the traditional loss for cross-modal alignment. Benefiting from the memorization effect of DNNs [1, 38], geometrical structure can be well-established in the early stage. During this phase, GSC assesses the true soft correspondence labels based on the differences in geometrical structure within and across modalities. For the cross-modal aspect, GSC identifies potential noise by recognizing data pairs with low cross-modal similarities, while in the case of intra-modal aspect, the similarity between the queried sample and other samples are calculated to determine the intra-modal geometrical structure, and data pairs with inconsistent structures across modalities are considered as noisy. These assessed labels are then used to clarify the learning of the geometrical structure, creating a positive feedback loop. To conclude, the contributions of this work are summarized as follows:

- We identify the impact of noisy correspondence on both cross-modal and intra-modal geometrical structures, and find the significant difference between clean samples and noisy samples within a well-established overall structure.
- We introduce the novel Geometrical Structure Consistency (GSC) approach, which utilizes the structural differences to accurately predict true correspondence labels and counteract the adverse effects of noisy correspondences.
- Our proposed GSC method is compatible with existing approaches for handling noisy correspondences. Through extensive experiments on three benchmark datasets, we have proven the consistent superiority of GSC over current state-of-the-art methods.

2. Related Works

2.1. Cross-modal Retrieval

Cross-modal retrieval [8, 14, 29, 33], which focuses on using information from one modality to query the most relevant data in other modalities, has been a key area of research in multi-modal learning [2, 9, 13, 36]. Due to the restriction of large-scale annotated multi-modal corpus [18, 21, 32], the unsupervised guided framework that directly aligns representations across modalities has always been the mainstream. VSE++ [7] strategically incorporates hard negatives in their approach to enhance retrieval efficacy. SCAN [24] implements a stacked cross-attention framework which facilitates a dual-context attention mechanism to effectively dissect the intricate interplay between different modalities. SGRAF [6] introduces Graph Neural Network to establish graph correspondences and an attention mechanism to select the most representative alignments, enhancing the precision of cross-modal similarity assessments. However, these cross-modal retrieval methods highly rely on well-aligned data while amending the existence of noisy correspondence. In recent works, direct alignment between modalities has been argued as sub-optimal for inconsistent downstream predictions. To combat, CYCLIP [10] explicitly optimizes representations to maintain geometric consistency, incorporating two additional cycle consistency constraints. Jiang et al. [17] further improves the method by introducing additional loss functions for inter-modal and intra-modal regularization. This emphasis on geometric consistency not only directly benefits cross-modal retrieval learning, but also aids in distinguishing noisy correspondences in our own work.

2.2. Noisy Correspondence Learning

Noisy correspondence (NC) is first aroused in [15], which is a novel paradigm in the field of noise learning [12, 19, 28, 40], referring to the mismatched pairs within the multi-modal dataset. To tackle this problem, NCR [15] utilizes DivideMix [25] to distinguish clean pairs from noisy ones and rectify correspondence based on the memorization effect of DNNs. Yang et al. [41] further improves NCR by switching to Beta Mixture Model and estimates soft correspondence labels by sample-wise comparison. Han et al. [11] proposes a meta-similarity correction network that reinterprets binary classification of correspondence as a meta-process, enhancing the process of data purification. Despite NCR-based models, other attempts like robust loss based methods have also been undertaken. Qin et al. [34] combines the idea of evidential learning with NC and puts forward a confidence-based method. Chuang et al. [5] introduces one effective robust symmetric contrastive loss. Although these models have showcased promising performances, they only cast their spotlight on interactions across modalities, which is insufficient in utilizing the semantic-abundant cross-modal data, further motivating us to take not only cross-modal but also intra-modal together into consideration to help mitigate NC.

3. Proposed Method

3.1. Preliminary

We start by defining notations for cross-modal retrieval in the presence of noisy correspondences, employing the

widely studied image-text retrieval task as a generalized exemplar. Consider a multimodal dataset $\mathcal{D} = \{I_i, T_i, y_i\}_{i=1}^{N}$, where each I_i, T_i represents the i-th image-text pair, standard retrieval models project these data pairs into a shared representation space using separate encoders f for images and g for texts. Then similarity scores between the representations are computed through cosine similarity or an inference model, denoting as $S(f(I), g(T))$, or $\langle I, T \rangle$ in brief. The associated label y_i indicates whether the pair is positively correlated ($y_i = 1$) or not ($y_i = 0$). Note that these labels may contain noise, as pairs in the dataset are often presumed to be matched.

3.2. Geometrical Structure Consistency Learning

To address the above issue, we introduce the *Geometrical Structure Consistency* (GSC) learning method to identify and correct noisy correspondences, which is detailed in the following sections.

Motivation. The core concept of GSC is to preserve the consistency of geometrical structures and distinguish samples with noisy correspondence through structural differences. Initially, we demonstrate these differences through a straightforward experiment. As shown in Fig. 2, a retrieval model is optimized on a clean dataset to maintain consistent cross-modal and intra-modal structures, which is then assessed on a dataset with simulated noise. During the experiment, significant discrepancy between clean and noisy samples can be observed in both cross-modal and intra-modal structures. Specifically, for cross-modal structure, clean samples typically possess higher similarity scores than those noisy ones, exhibiting a disparity in distribution. For intra-modal structure, the calculated similarities between intra-modal structures of clean and noisy samples manifest a bimodal distribution with most values of noisy samples lower than 0.5 threshold, suggesting noisy samples tend to have asymmetric intra-modal structures.

Geometrical Structure Consistency. Here, we give definitions to both cross-modal and intra-modal geometrical structures and the corresponding training objectives. From the cross-modal aspect, the geometrical structure is defined by the similarities between representations across different modalities. Considering an example of a given query image I_i, the cross-modal geometrical structure can be represented as $\mathcal{G}_{\text{CM}}^{i} = \{\langle I_i, T_j \rangle\}_{j=1}^{N}$. GSC preserves the consistency of this structure by minimizing the expected risk for cross-modal objective, as expressed in the following equation,

$$\mathcal{R}_{\mathcal{L}_{\text{CM}}}(f, g) = \min_{f,g} \mathbb{E}_{(I,T,y)\sim\mathcal{D}} \left[\mathcal{L}_{\text{CM}}(\langle I, T \rangle, y)\right] \quad (1)$$

where $\mathcal{L}_{\text{CM}}$ is the cross-modal loss function, typically a contrastive or triplet loss in line with conventional retrieval models. The goal is to align the cross-modal representations according to the correspondence label y, thus the similarity between matching data pairs can be maximized, contrasting to other data pairs.

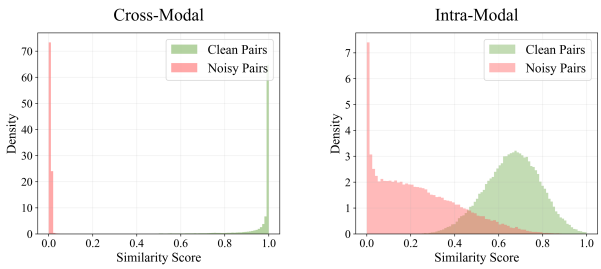

Figure 2. Geometrical Structure Consistency helps discriminate samples with noisy correspondence. The model is first trained on clean Flickr30K dataset, then evaluated on the same dataset with 40% simulated noise. **Left:** Calculated cross-modal similarity scores of both clean and noisy samples. **Right:** Calculated intra-modal similarity scores of both clean and noisy samples.

From the intra-modal aspect, the geometrical structure refers to the similarities within the modality. The intra-model structure for the i-th sample then denotes as $\mathcal{G}_{\text{IM}}^{i} = \{\langle I_i, I_j \rangle, \langle T_i, T_j \rangle\}_{j=1}^{N}$, where $\langle I_i, I_j \rangle$ and $\langle T_i, T_j \rangle$ represent the pairwise similarities among images and texts, respectively. To uphold the consistency of such structure, GSC incorporates the intra-modal objective as following,

$$\mathcal{R}_{\mathcal{L}_{\text{IM}}}(f, g) = \min_{f,g} \mathbb{E}_{(I,T,y)\sim\mathcal{D}} \left[\mathcal{L}_{\text{IM}}(\langle I, I \rangle, \langle T, T \rangle, y)\right] \quad (2)$$

In this equation, $\mathcal{L}_{\text{IM}}$ denotes the intra-modal loss function. The intra-modal objective ensures that the intra-modal structures of matching samples are constrained to be similar across modalities. Notably, optimizing without maintaining intra-modal structure is considered sub-optimal for inconsistent reasoning [10, 17]. Thus the introduction of intra-modal structure consistency can also directly benefit multimodal representation learning. As illustrated in Fig. 3(a), GSC simultaneously optimizes both objectives to establish stable cross-modal and intra-modal structures.

3.3. Noise Discrimination & Purification

Deep neural networks typically exhibit the memorization effect that tends to initially learn the clean patterns within the dataset before over-fitting on noise. Leveraging this, GSC is able to learn a well-established structure in the early stage, which can be further utilized to predict accurate correspondence indicator.

Cross-modal Discrimination. As illustrated in Fig. 3(b), based on the well-established cross-modal structure in the early stage, clean data pairs are expected to exhibit more closely aligned cross-modal representations compared to noisy pairs. GSC leverages this structural discrepancy and introduces a function to signify a cross-modal bidirectional

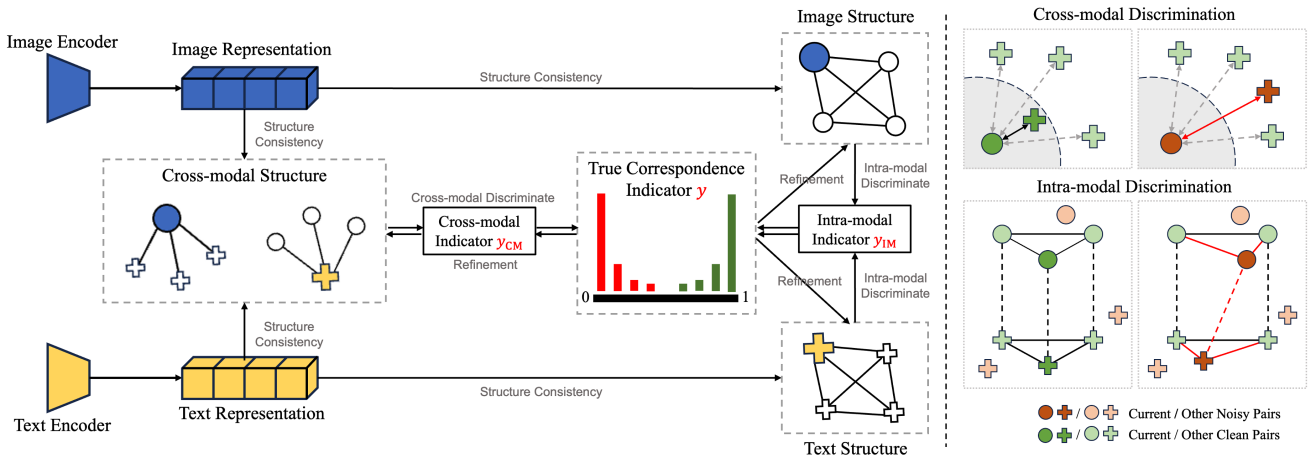

Figure 3. An overview of GSC. **Left:** The framework of GSC. GSC first extracts image and text representations through separate encoders, then simultaneously optimizes cross-modal and intra-modal objectives to preserve geometrical structure consistency. GSC leverages both structures to discriminate noisy samples and estimate the true correspondence indicator y, which can be further utilized to purify the overall learning. **Right:** GSC discriminates noisy samples by structural differences from both cross-modal and intra-modal aspects.

correspondence indicator.

$$y^i_{\text{CM}} = \frac{1}{2}\Big[\frac{\exp\left(\langle I_i, T_i\rangle/\tau_1\right)}{\sum_j^N \exp\left(\langle I_i, T_j\rangle/\tau_1\right)} + \frac{\exp\left(\langle I_i, T_i\rangle/\tau_1\right)}{\sum_j^N \exp\left(\langle I_j, T_i\rangle/\tau_1\right)}\Big] \tag{3}$$

where τ_1 is the temperature coefficient. Take the former half as an example, it measures the proportion of similarity between the current data pairs I_i and T_i against the sumation of similarity between the I_i and all text data. For a clean sample, the similarity between current I_i and T_i should dominate the proportion, thus the value of indicator should approach 1. Conversely, for a noisy sample, the similarity between current I_i and T_i would be close to 0, resulting in the indicator's value trending toward 0.

Intra-modal Discrimination. For the intra-modal aspect, as illustrated in Fig. 3(c), matched image-text pairs should have similar intra-modal structures that mirror each other, whereas mismatched pairs would possess distinct structures that reflect their divergent positions. Specifically, we employ cosine similarity to measure the resemblance between intra-modal structures as $\mathcal{S}^i_{\text{IM}} = \cos(\{\langle I_i, I_j\rangle, \langle T_i, T_j\rangle\}_{j=1}^N)$. During experiments, the cosine similarity scores of clean samples are observed consistently higher than those of noisy samples during experiments, presenting a bimodal distribution of scores across the dataset (demonstrated in Fig. 5(b)). This distribution can be accurately modeled using a two-component Gaussian Mixture Model (GMM) [25], described by the equation:

$$p(\mathcal{S}_{\text{IM}}) = \sum_{k=1}^{K} \alpha_k \phi(\mathcal{S}_{\text{IM}}|k),\ y^i_{\text{IM}} = \frac{\alpha_{k_l}\phi(\mathcal{S}^i_{\text{IM}}|k_l)}{\sum_{k=1}^K \alpha_k \phi(\mathcal{S}^i_{\text{IM}}|k)} \tag{4}$$

Specifically, α_k denotes the k-th coefficient and $\phi(\mathbf{s}_{\text{IM}}|k)$ is the probability density for that component. The second equation is the estimation of the intra-modal correspondence indicator y^i_{IM}. It calculates the probability of an observed sample belonging to the cleaner component, denoted by $k = k_l$. This probability approaches 1 for clean samples and 0 for noisy samples, thereby enabling the distinction of samples with noisy correspondence.

So far, we have estimated the true correspondence labels, y_{CM} and y_{IM}, by leveraging the structural differences in both cross-modal and intra-modal contexts. Our objective is to optimally utilize these two labels to surmount the respective challenges and accurately identify all samples with noisy correspondence. Therefore, we define the final correspondence label for each sample as the minimum of the two labels, which can expressed as below,

$$y^i = \min\{y^i_{\text{CM}}, y^i_{\text{IM}}\} \tag{5}$$

Noise purification. We address the issue of noisy correspondence by refining both cross-modal and intra-modal objectives. Since the estimated label is a soft label with values in the range of [0, 1] which can directly reflect the degree of true correspondence, we can seamlessly apply it to the loss functions on a sample-wise basis. For the cross-modal objective, we choose the widely-applied contrastive loss as the loss function. The purified cross-modal loss can be denoted as follows,

$$\begin{aligned}\mathcal{L}_{\text{CM}} = &-\frac{1}{2N}\sum_{i=1}^{N} y^i \log \frac{\exp\left(\langle I_i, T_i\rangle/\tau_1\right)}{\sum_{j=1}^N \exp\left(\langle I_i, T_j\rangle/\tau_1\right)} \\ &-\frac{1}{2N}\sum_{j=1}^{N} y^j \log \frac{\exp\left(\langle I_j, T_j\rangle/\tau_1\right)}{\sum_{i=1}^N \exp\left(\langle I_i, T_j\rangle/\tau_1\right)}\end{aligned} \tag{6}$$

where y is directly applied before the sample-wise loss. For the intra-modal side, in addition to sample-wise purification

Algorithm 1 Pipeline of learning with our GSC method.

Input: Multi-modal dataset $\mathcal{D} = \{I_i, T_i, y_i\}_{i=1}^{N}$
Initialize parameters for networks A and B separately
for *each epoch* $t = 1, 2, \ldots, T$ **do**
  for *network* $k = A, B$ **do**
    for *each minibatch* $\mathcal{B}$ *from* $\mathcal{D}$ **do**
      Compute modality representations: $f(I)$, $g(T)$
      Estimate *cross-modal* indicator $y_{\text{CM}}[t]_k$ by Eq. 3
      Estimate *intra-modal* indicator $y_{\text{IM}}[t]_k$ by Eq. 4
      Calaulate *cross-modal* loss by Eq. 6
      Calculate *intra-modal* loss by Eq. 7
      Train Net_k by optimizing the combination of two losses using Eq. 9
    Update $y_{\text{CM}}[t]$ and $y_{\text{IM}}[t]$ by Eq. 10
    Update the final $y[t]$ for the other network.
Output: Refined networks Net_A, Net_B

for the intra-modal loss, it is crucial to prevent the impact of noisy correspondences from distorting the calculation of the geometrical structure. Specifically, the distances from the queried sample to those samples with noisy correspondences should be excluded. The purified intra-model loss denotes as,

$$\mathcal{L}_{\text{IM}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp(\sum_{k=1}^{N} y^k\langle I_i, I_k\rangle y^k\langle T_i, T_k\rangle/\tau_2)}{\sum_{j=1}^{N}\exp\left(\sum_{k=1}^{N} y^k\langle I_i, I_k\rangle y^k\langle T_j, T_k\rangle/\tau_2\right)} \tag{7}$$

where τ_2 is also a temperature coefficient and y^k is multiplied to both $\langle I_i, I_k\rangle$ and $\langle T_i, T_k\rangle$ to precisely filter out the noise. Furthermore, noisy correspondences can also interfere with the discrimination of noisy samples during the computation of cosine similarity scores for intra-modal discrimination. Similar modifications are employed as follows,

$$\mathcal{S}_{\text{IM}}^{i} = \frac{\sum_{j=1}^{N} y^j\langle I_i, I_j\rangle y^j\langle T_i, T_j\rangle}{\sqrt{\sum_{j=1}^{N}(y^j\langle I_i, I_j\rangle)^2}\sqrt{\sum_{j=1}^{N}(y^j\langle T_i, T_j\rangle)^2}} \tag{8}$$

where y^j is similarly multiplied. Considering the high computational cost for the entire dataset, we employ a Monte Carlo sampling approach to relax size N to size B of a mini-batch. The overall loss function of GSC is formulated as a weighted sum of two loss functions, as expressed in the following equation,

$$\mathcal{L} = \mathcal{L}_{\text{CM}} + \gamma\mathcal{L}_{\text{IM}} \tag{9}$$

where γ is the hyper-parameter keeping the balance between two losses to reach the best optimization.

3.4. Training Schedule

To integrate the estimation of true correspondence labels with the enhancement of cross-modal retrieval learning, we adopt the temporal ensembling technique, drawing inspiration from Liu et al. [27], to iteratively update the estimated correspondence labels. Specifically, both y_{CM} and y_{IM} are updated through a momentum-based combination of the estimates from the current epoch t and the previous epoch $t-1$ before taking minimum as shown below,

$$\begin{aligned} y_{\text{CM}}^{i}[t] &= \beta_1 y_{\text{CM}}^{i}[t] + (1-\beta_1)y_{\text{CM}}^{i}[t-1], \\ y_{\text{IM}}^{i}[t] &= \beta_2 y_{\text{IM}}^{i}[t] + (1-\beta_2)y_{\text{IM}}^{i}[t-1] \end{aligned} \tag{10}$$

where β_1 and β_2 are separate momentum. The estimation of true correspondence labels can be further improved by utilizing two separate neural networks, where the true correspondence labels for each network are computed from the output of the other network. The ablation in Section 4 shows that both strategies can significantly improve the performance. In conclusion, the overall procedure of our Geometrical Structure Consistency (GSC) method is depicted in the pseudo-algorithm 1.

3.5. Discussion

The improvement of our proposed GSC mainly comes from preserving the geometrical structure consistency and better optimizing strategies, which is compatible with most existing methods. In terms of computational complexity, GSC does not introduce additional computational costs when integrated with a backbone that calculates similarity scores directly from representations, while it necessitates two extra forward passes when the backbone computes similarity scores using a similarity module. This requirement is significantly less demanding compared to MSCN, which involves computations for an additional meta-learning model, or BiCro, which compares each sample against a clean subset. While RINCE shares similar computational cost as GSC, the robust loss function without explicitly excluding noisy samples underperforms at higher noise levels. Furthermore, methods based on NCR framework, *i.e.* NCR, MSCN and BiCro, rely on an inseparable dual-network structure, while GSC is effective with a single model. This attribute makes GSC more adaptable to larger models.

4. Experiment

4.1. Datasets and Evaluation Metrics

Datasets. Following the experimental settings and dataset splits in Huang et al. [15], three widely-used image-text retrieval datasets are introduced to evaluate our method:

- Flickr30K [42] contains 31,000 images with five captions each, collected from the Flickr website. We assign 1,000 image-text pairs for validation, 1,000 image-text pairs for testing and the rest for training.
- MS-COCO [26] includes 123,287 images with five captions each. We assign 5,000 image-text pairs for validation, 5,000 image-text pairs for testing and the rest for

Table 1. The retrieval performance on Flickr30K and MS-COCO datasets under 20%, 40% and 60% noise rates separately. The best results and the second best results are respectively marked by **bold** and underline.

Noise	Methods	Flickr30K							MS-COCO						
		Image → Text			Text → Image				Image → Text			Text → Image			
		R@1	R@5	R@10	R@1	R@5	R@10	Sum	R@1	R@5	R@10	R@1	R@5	R@10	Sum
20%	SGR	55.9	81.5	88.9	40.2	66.8	75.3	408.6	25.7	58.8	75.1	23.5	58.9	75.1	317.1
	SGRAF	72.8	90.8	95.4	56.4	82.1	88.6	486.1	75.4	95.2	97.9	60.1	88.5	94.8	511.9
	NCR-SGR	73.5	93.2	96.6	56.9	82.4	88.5	491.1	76.6	95.6	98.2	60.8	88.8	95.0	515.0
	DECL-SGRAF	77.5	93.8	97.0	56.1	81.8	88.5	494.7	77.5	95.9	98.4	61.7	89.3	95.4	518.2
	RINCE-SGR	72.1	92.2	95.7	54.9	79.8	85.3	480.0	73.8	95.6	98.5	61.7	89.2	94.7	513.5
	MSCN-SGR	77.4	**94.9**	97.6	59.6	83.2	89.2	501.9	78.1	**97.2**	98.8	64.3	90.4	95.8	524.6
	BiCro-SGRAF	78.1	94.4	97.5	**60.4**	84.4	89.9	504.7	78.8	96.1	98.6	63.7	90.3	95.7	523.2
	GSC-SGR	**78.3**	94.6	**97.8**	60.1	**84.5**	**90.5**	**505.8**	**79.5**	96.4	**98.9**	**64.4**	**90.6**	**95.9**	**525.7**
40%	SGR	4.1	16.6	24.1	4.1	13.2	19.7	81.8	1.3	3.7	6.3	0.5	2.5	4.1	18.4
	SGRAF	8.3	18.1	31.4	5.3	16.7	21.3	101.1	15.8	23.4	54.6	17.8	43.6	54.1	209.3
	NCR-SGR	68.1	89.6	94.8	51.4	78.4	84.8	467.1	74.7	94.6	98.0	59.6	88.1	94.7	509.7
	DECL-SGRAF	72.7	92.3	95.4	53.4	79.4	86.4	479.6	75.6	95.5	**98.3**	59.5	88.3	94.8	512.0
	RINCE-SGR	71.2	90.7	95.6	52.7	78.5	85.6	474.3	71.2	95.8	97.9	59.1	88.6	94.3	506.9
	MSCN-SGR	71.6	92.8	96.2	54.8	80.7	87.4	483.5	75.3	95.4	98.2	60.3	88.6	94.8	512.6
	BiCro-SGRAF	74.6	92.7	96.2	55.5	81.1	87.4	487.5	77.0	**95.9**	**98.3**	61.8	89.2	94.9	517.1
	GSC-SGR	**76.5**	**94.1**	**97.6**	**57.5**	**82.7**	**88.9**	**497.3**	**78.2**	**95.9**	98.2	**62.5**	**89.7**	**95.4**	**519.9**
60%	SGR	1.5	6.6	9.6	0.3	2.3	4.2	24.5	0.1	0.6	1.0	0.1	0.5	1.1	3.4
	SGRAF	2.3	5.8	10.9	1.9	6.1	8.2	35.2	0.2	3.6	7.9	1.5	5.9	12.6	31.7
	NCR-SGR	13.9	37.7	50.5	11.0	30.1	41.4	184.6	0.1	0.3	0.4	0.1	0.5	1.0	2.4
	DECL-SGRAF	65.2	88.4	94.0	46.8	74.0	82.2	450.6	73.0	94.2	97.9	57.0	86.6	93.8	502.5
	RINCE-SGR	64.5	86.8	92.9	46.5	72.8	79.7	443.2	72.3	94.0	97.9	58.4	86.6	92.5	501.7
	MSCN-SGR	68.8	90.3	94.4	50.8	77.4	84.4	466.1	72.5	93.6	97.1	57.7	87.0	93.9	501.8
	BiCro-SGRAF	67.6	90.8	94.4	51.2	77.6	84.7	466.3	73.9	94.4	97.8	58.3	87.2	93.9	505.5
	GSC-SGR	**70.8**	**91.1**	**95.9**	**53.6**	**79.8**	**86.8**	**478.0**	**75.6**	**95.1**	**98.0**	**60.0**	**88.3**	**94.6**	**511.7**

training. Notably, MS-COCO can be either evaluated by using the whole 5,000 test set or the average of 5-fold 1,000 test sets [20].

- Conceptual Captions [37] is a large-scale dataset with real-world noisy correspondence problem. It contains 3.3M images with one caption each. All the data pairs in the Conceptual Captions dataset are automatically harvested from the Internet, therefore about 3%-20% image-text pairs in the dataset are mismatched or weakly-matched [37]. We use a subset of the Conceptual Captions dataset named CC152K in our experiments, in which we assign 1,000 image-text pairs for validation, 1,000 image-text pairs for testing and 150,000 image-text pairs for training.

Evaluation Protocol. We evaluate the retrieval performance with the recall rate at K (R@K) metric. In a nutshell, R@K measures the proportion of relevant items retrieved within the top K items closest to the query. In our experiments, we take image and text as queries, respectively, and report R@1, R@5, R@10 results and their sum for a comprehensive evaluation.

4.2. Implementation Details

For all experiments, we apply the Adam optimizer [23] with the initial learning rate of 2×10^{-4} which decays by 0.2 in 15 epochs. We train the model on one NVIDIA A100 GPU and select the model that performs best on the validation set for testing. The dimension of the common representation is set to 1024. For experiments besides ablation study, we set the batch size B as 128. The two temperature coefficients τ_1 and τ_2 are set to 0.07 and 1 in default. The hyperparameter λ serving as the balancing ratio between $\mathcal{L}_{\text{CM}}$ and $\mathcal{L}_{\text{IM}}$ is set to 0.01, and the two momentum β_1 and β_2 are set to 0.7 respectively.

4.3. Comparison with the State-of-the-Art

We conducted extensive evaluations against seven contemporary state-of-the-art methods on three benchmark datasets to validate the effectiveness of our proposed GSC model. These comparisons include two baseline models, *i.e.*, SGR and SGRAF [6], and five robust learning methods designed to handle noisy correspondences, *i.e.*, NCR [15], DECL [34], RINCE [5], MSCN [11] and BiCro [41]. No-

Table 2. The retrieval performance on CC152K dataset. The best results and the second best results are respectively marked by **bold** and underline.

Methods	CC152K						
	Image → Text			Text → Image			
	R@1	R@5	R@10	R@1	R@5	R@10	Sum
SGR	11.3	29.7	39.6	13.1	30.1	41.6	165.4
SGRAF	32.5	59.5	70.0	32.5	60.7	68.7	323.9
NCR-SGR	39.5	64.5	73.5	40.3	64.6	73.2	355.6
DECL-SGRAF	39.0	66.1	75.5	40.7	66.3	76.7	364.3
RINCE-SGR	35.9	63.0	73.8	37.6	65.0	73.4	348.7
MSCN-SGR	40.1	65.7	76.6	40.6	67.4	76.3	366.7
BiCro-SGRAF	40.8	67.2	76.1	42.1	**67.6**	76.4	370.2
GSC-SGR	**42.1**	**68.4**	**77.7**	**42.2**	**67.6**	**77.1**	**375.1**

Table 3. Comparison with CLIP and NCR on MS-COCO 5K. CLIP-L and CLIP-B are abbreviations for CLIP (ViT-L/14) and CLIP (ViT-B/32). The best results are marked by **bold**.

Noise	Methods	Image → Text			Text → Image			
		R@1	R@5	R@10	R@1	R@5	R@10	Sum
0% Zeroshot	CLIP-L	58.4	81.5	88.1	37.8	62.4	72.2	400.4
	CLIP-B	50.2	74.6	83.6	30.4	56.0	66.8	361.6
20% Finetune	CLIP-B	21.4	49.6	63.3	14.8	37.6	49.6	236.3
	NCR	56.9	83.6	91.0	40.6	69.8	80.1	422.0
	GSC	**58.9**	**84.9**	**91.7**	**42.0**	**71.4**	**81.8**	**430.8**
50% Finetune	CLIP-B	10.9	27.8	38.3	7.8	19.5	26.8	131.1
	NCR	53.1	80.7	88.5	37.9	66.6	77.8	404.6
	GSC	**55.5**	**81.8**	**90.1**	**40.0**	**69.1**	**79.7**	**416.3**

tably, both DECL and BiCro are built upon a stronger SGRAF backbone, which is an ensemble of SGR and SAF. To thoroughly assess the robustness of GSC, we simulate various levels of noisy correspondences, namely 20%, 40%, and 60%, by randomly shuffling the captions on MSCOCO and Flickr30K like [15]. In addition, we extend our experiments to real-world noisy conditions using the CC152K dataset. Comprehensive comparison results are detailed in the supplementary material for fully demonstration of GSC.

Results on Flickr30K and MS-COCO. To evaluate the robustness of all methods under different extents of noise, we quantify the noise rate to 20%, 40% and 60% on both well-annotated MS-COCO and Flickr30K datasets, as recorded in Tab. 1. The results demonstrate that GSC significantly outperforms established noisy correspondence methods such as NCR, DECL, RINCE, MSCN, and BiCro, achieving an average increase in recall sum score of 7.5% on Flickr30K and 3.4% on MS-COCO to the second best results, which indicates the better robustness of GSC. Notably, GSC also excels over BiCro and DECL under various conditions, even though they are implemented on the enhanced SGRAF backbone. Moreover, GSC can carry about more enhancement at higher noise rates, especially under 60% noise level, proving that our method remains stable and reliable even in severely noisy conditions.

Results on CC152K. To further validate GSC in handling with noisy correspondence in real-world scenarios, we additionally conduct tests on CC152K dataset, detailed in Tab. 2. According to the results, GSC achieves the best performance with an overall score of 375.1%, surpassing the second best method BiCro by 4.9%. Moreover, GSC brings about a larger gain of 209.7% to its backbone SGR, which is significantly higher than the improvement of 46.3% brought about by BiCro to its backbone SGRAF. The results affirm GSC's capability to manage not only simply simulated but also complex, real-world noisy correspondences.

Table 4. Ablation study on Flicker30K with 40% noise with different components in GSC. The best results are marked by **bold**.

Momen.	Dual	$\mathcal{L}_{\text{IM}}$	Image → Text			Text → Image			
			R@1	R@5	R@10	R@1	R@5	R@10	Sum
✓	✓		72.3	92.8	96.5	56.5	82.1	88.8	489.0
✓		✓	73.0	91.5	95.9	54.5	80.4	87.6	482.9
	✓	✓	74.5	92.5	96.9	57.0	82.0	88.3	491.1
✓	✓	✓	**76.5**	**94.1**	**97.6**	**57.5**	**82.7**	**88.9**	**497.3**

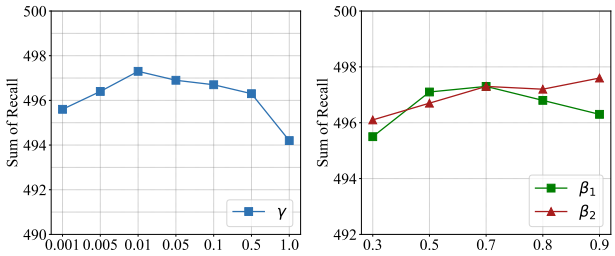

Figure 4. Analysis of different hyper-parameter combinations on Flicker30K with 40% noise. **Left:** γ is the balancing parameter between $\mathcal{L}_{\text{CM}}$ and $\mathcal{L}_{\text{IM}}$. **Right:** β_1 and β_2 are separate momentums for the cross-modal and intra-modal temporal ensembling.

Comparison to pre-trained model. In line with Huang et al. [15], we compare GSC to the pre-trained CLIP model [35] on the MS-COCO dataset. CLIP is a well-known large pre-trained model trained on a massive 400 million image-text pair dataset harvested from the Internet, which can inevitably include samples with noisy correspondence. Here, we report the zero-shot and fine-tuning performances of different CLIP models, together with NCR. Results indicate a notable performance decline in CLIP models when fine-tuning with 20% and 50% noise levels. On the contrary, GSC not only withstands but excels over zero-shot CLIP under 50% noise, emphasizing the importance of addressing data mismatches and the robustness of GSC.

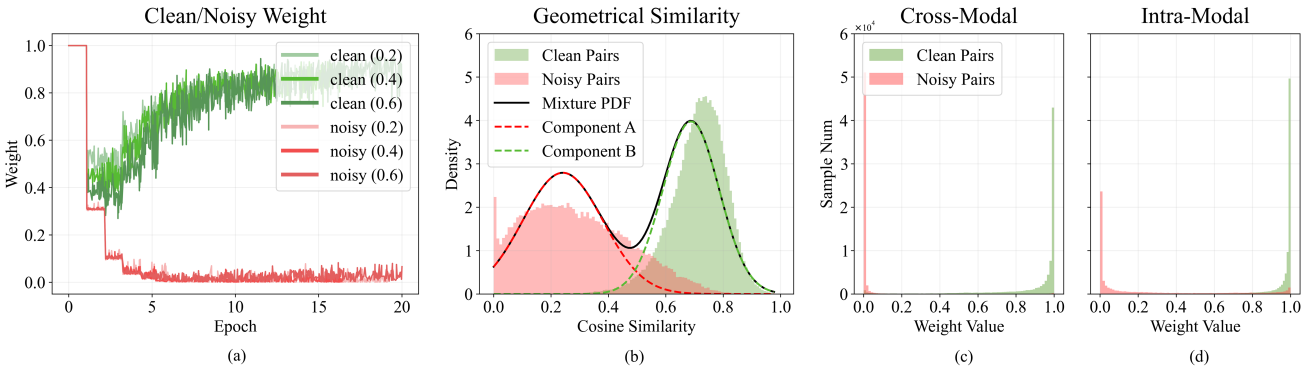

Figure 5. (a) The changing values of clean and noisy sample weight when the noise rate is 20%, 40%, and 60%. (b) Distribution of intra-modal geometrical similarity, including PDFs of clean and noisy pair similarities and estimated Gaussian distribution components. (c) Cross-modal weight distributions of GSC on clean and noisy pairs. (d) Intra-modal weight distributions of GSC on clean and noisy pairs. Experiments from (b) to (d) are conducted on Flickr30K with the noise rate of 0.4.

Table 5. Analysis of different batch sizes on Flicker30K with 40% noise. The best results are marked by **bold**.

Batch Size	Image → Text			Text → Image			
	R@1	R@5	R@10	R@1	R@5	R@10	Sum
32	73.6	92.1	95.4	54.4	81.0	87.4	483.9
64	75.3	94.0	96.7	55.9	81.7	88.2	491.8
128	76.5	**94.1**	97.6	57.5	**82.7**	88.9	497.3
192	**76.7**	94.0	**97.7**	**57.6**	**82.7**	**89.1**	**497.8**

4.4. Experimental Analysis

Ablation study. We show the effect of each component of GSC in Tab. 4. The ablation studies are conducted without temporal ensembling (Momen. in the table), dual networks or intra-modal learning separately. Specifically for the ablation study for temporal ensembling, we use a 5-epoch warm-up stage to replace the technique. According to the results, all components are important to achieve advantageous results. Notably, the performance of GSC with single model still outperforms most robust methods, including methods with dual networks like NCR and DECL, which further proves the effectiveness and high efficiency.

Impacts of hyper-parameters γ, β_1 and β_2. The GSC method incorporates three main hyper-parameters including γ, β_1 and β_2 with their effects detailed in Fig. 4. γ strikes a balance between cross-modal and intra-modal optimization. According to the results, GSC shows stability for $\gamma \in [0.005, 0.1]$, while $\gamma = 0.01$ is chosen for optimal performance. β_1 and β_2 are the momentum coefficients ensuring steady updates for temporal ensembling. The results indicate stable performance with parameter values higher than 0.5. Lower beta values may laten timely updates, potentially causing the model to overfit on noise.

Impact of batch size. We also explore the model performance under different batch sizes during training. As shown in Tab. 5, as the batch size increases, the retrieval performance of the model steadily improves. Specifically, when the batch size is increased from 32 to 128, there is a significant enhancement from 483.9% to 497.3% as to the sum of recall, and further expanding the batch size from 128 to 192 results in only marginal growth, which indicates that larger batch size helps in consolidating the stableness of model structure until reaching a proper point.

Experimental visualization. To further offer insights of GSC against noisy correspondence, we visually present the value curves and weight distributions of predicted correspondence labels in Fig. 5. Figure (a) shows the stability of predicted labels across varying noise levels, with minimal fluctuation for clean samples and consistently low values for noisy ones. Figure (b) depicts a bimodal distribution of intra-modal cosine similarities, which can be well-fitted by a two-component GMM. Figures (c) and (d) confirm that both cross-modal and intra-modal structures effectively distinguish noisy samples, with discrimination accuracy of approximately 0.96 for cross-modal and about 0.91 for intra-modal, culminating in an overall accuracy of about 0.98. Such visualization explains the reason for the steady and reliable performance of GSC under different noise rates.

5. Conclusion

In this paper, we propose Geometrical Structure Consistency (GSC) learning framework to mitigate the problem brought by noisy correspondence. Specifically, we identify the impact of noisy correspondence on both cross-modal and intra-modal geometrical structures. Leveraging the structural differences between noisy and clean pairs within a well-established structure, our approach infers accurate correspondence labels for each data pair. The inferred labels are further utilized to refine the consistent learning of structures. GSC can seamlessly integrate with most existing retrieval methods. Extensive experiments across various cross-modal benchmark datasets showcase the robustness and effectiveness of our proposed GSC method across diverse settings.

References

[1] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 2

[2] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *European Conference on Computer Vision*, pages 348–367. Springer, 2022. 2

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 1

[4] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009. 1

[5] Ching-Yao Chuang, R Devon Hjelm, Xin Wang, Vibhav Vineet, Neel Joshi, Antonio Torralba, Stefanie Jegelka, and Yale Song. Robust contrastive learning against noisy views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16670–16681, 2022. 2, 6

[6] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1218–1226, 2021. 2, 6

[7] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 2

[8] Zheren Fu, Zhendong Mao, Yan Song, and Yongdong Zhang. Learning semantic relationship among instances for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15159–15168, 2023. 2

[9] Yunhao Ge, Jie Ren, Andrew Gallagher, Yuxiao Wang, Ming-Hsuan Yang, Hartwig Adam, Laurent Itti, Balaji Lakshminarayanan, and Jiaping Zhao. Improving zero-shot generalization and robustness of multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11093–11101, 2023. 2

[10] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *Advances in Neural Information Processing Systems*, 35:6704–6719, 2022. 2, 3

[11] Haochen Han, Kaiyao Miao, Qinghua Zheng, and Minnan Luo. Noisy correspondence learning with meta similarity correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7517–7526, 2023. 1, 2, 6

[12] Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5403–5413, 2021. 2

[13] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International Conference on Machine Learning*, pages 9226–9259. PMLR, 2022. 2

[14] Yan Huang, Yuming Wang, Yunan Zeng, and Liang Wang. Mack: multimodal aligned conceptual knowledge for unpaired image-text matching. *Advances in Neural Information Processing Systems*, 35:7892–7904, 2022. 2

[15] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34:29406–29419, 2021. 1, 2, 5, 6, 7

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 1

[17] Qian Jiang, Changyou Chen, Han Zhao, Liqun Chen, Qing Ping, Son Dinh Tran, Yi Xu, Belinda Zeng, and Trishul Chilimbi. Understanding and constructing latent modality structures in multi-modal representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7661–7671, 2023. 2, 3

[18] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 2

[19] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9676–9686, 2022. 2

[20] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 6

[21] Douwe Kiela, Edouard Grave, Armand Joulin, and Tomas Mikolov. Efficient large-scale multi-modal classification. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[22] Jae Myung Kim, A Koepke, Cordelia Schmid, and Zeynep Akata. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2584–2594, 2023. 1

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[24] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018. 2

[25] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020. 1, 2, 4

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 5

[27] Sheng Liu, Jonathan Niles-Weed, Narges Razavian, and Carlos Fernandez-Granda. Early-learning regularization prevents memorization of noisy labels, 2020. 5

[28] Sheng Liu, Zhihui Zhu, Qing Qu, and Chong You. Robust training under label noise by over-parameterization. In *International Conference on Machine Learning*, pages 14153–14172. PMLR, 2022. 2

[29] Yang Liu, Hong Liu, Huaqiu Wang, and Mengyuan Liu. Regularizing visual semantic embedding with contrastive learning for image-text matching. *IEEE Signal Processing Letters*, 29:1332–1336, 2022. 2

[30] Yaxin Liu, Jianlong Wu, Leigang Qu, Tian Gan, Jianhua Yin, and Liqiang Nie. Self-supervised correlation learning for cross-modal retrieval. *IEEE Transactions on Multimedia*, 2022. 1

[31] Haoyu Lu, Nanyi Fei, Yuqi Huo, Yizhao Gao, Zhiwu Lu, and Ji-Rong Wen. Cots: Collaborative two-stream vision-language pre-training model for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15692–15701, 2022. 1

[32] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):187–203, 2019. 2

[33] Zhengxin Pan, Fangyu Wu, and Bailing 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*, pages 19275–19284, 2023. 2

[34] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4948–4956, 2022. 1, 2, 6

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[36] Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3677–3685, 2023. 2

[37] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu 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)*, pages 2556–2565, 2018. 1, 6

[38] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*, 2020. 2

[39] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1

[40] Tianyuan Xu, Xueliang Liu, Zhen Huang, Dan Guo, Richang Hong, and Meng Wang. Early-learning regularized contrastive learning for cross-modal retrieval with noisy labels. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 629–637, 2022. 2

[41] Shuo Yang, Zhaopan Xu, Kai Wang, Yang You, Hongxun Yao, Tongliang Liu, and Min Xu. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19883–19892, 2023. 1, 2, 6

[42] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 5