Robust Noisy Correspondence Learning with Equivariant Similarity Consistency

Yuchen Yang, Likai Wang, Erkun Yang*, Cheng Deng*
School of Electronic Engineering, Xidian University, Xi'an 710071, China
yuchenyanggm@gmail.com, lkwang@stu.xidian.edu.cn, erkunyang@gmail.com, chdeng.xd@gmail.com

Abstract

The surge in multi-modal data has propelled cross-modal matching to the forefront of research interest. However, the challenge lies in the laborious and expensive process of curating a large and accurately matched multi-modal dataset. Commonly sourced from the Internet, these datasets often suffer from a significant presence of mismatched data, impairing the performance of matching models. To address this problem, we introduce a novel regularization approach named ***Equivariant Similarity Consistency*** *(ESC), which can facilitate robust clean and noisy data separation and improve the training for cross-modal matching. Intuitively, our method posits that the semantic variations caused by image changes should be proportional to those caused by text changes for any two matched samples. Accordingly, we first calculate the ESC by comparing image and text semantic variations between a set of elaborated anchor points and other undivided training data. Then, pairs with high ESC are filtered out as noisy correspondence pairs. We implement our method by combining the ESC with a traditional hinge-based triplet loss. Extensive experiments on three widely used datasets, including Flickr30K, MS-COCO, and Conceptual Captions, verify the effectiveness of our method.*

1. Introduction

Cross-modal learning aims to extract and understand meaningful correspondences between data from different modalities, which contains various downstream tasks, such as audio-visual recognition [1, 2], multi-modal fusion [23, 27], and cross-modal generation [31, 32, 34, 52]. One of the most essential tasks in cross-modal learning is cross-modal matching, which aims to capture the semantic similarities between instances from multiple modalities and retrieve relevant instances correctly. With the rapid development of computility, cross-modal matching methods [25, 44, 45, 53] have achieved remarkable progress with massive and correct correspondence data. However, collecting such high-quality labeled training data is significantly hard due to its economic cost. Current multi-modal datasets unavoidably contain noisy correspondence so that the performance of cross-modal tasks almost reaches a bottleneck. Therefore, learning with noisy correspondence has become an essential and indispensable research field.

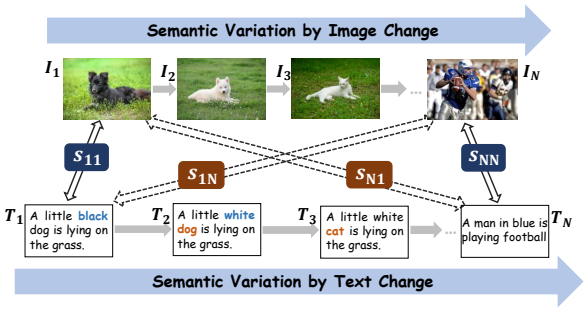

Figure 1. Illustration of the key idea in ESC. We intentionally create image and text spaces with different lengths to emphasize that semantic changes in two spaces are not necessarily equal.

Similar to noisy label problems [3, 16–18, 42, 43, 47], noisy correspondence learning focuses on dividing training data into clean set and noisy set, and then training different sets with different strategies. Relevant methods like [15, 29, 49] are almost based on the memorization effect of DNNs [51] observed in [13], which means that DNNs often prioritize learning simple patterns over fitting noisy samples. Inspired by this empirical observation, training data can be separated into two partitions (clean and noisy sets) based on their loss difference, i.e., small-loss samples are more likely to be clean data. Most existing methods for addressing noisy correspondence utilize triplet loss to filter out noisy samples and train matching models. Nevertheless, the sensitivity of margin α in triplet loss to different pairs of noisy correspondence data varies, causing triplet loss to easily fail in noisy datasets and lower the robustness in division and training. Therefore, how to improve the robustness remains an urgent problem to be addressed in noisy correspondence.

*Corresponding authors.

In this work, we propose a regularization called **Equivariant Similarity Consistency** (ESC) to constrain the division and training of cross-modal matching additionally. Take the image-text matching task as a proxy, the core idea in our method is *the semantic variations caused by image changes should be proportional to the semantic variations caused by text changes*, which has been proven to exist in paired data [39]. As illustrated in Fig. 1, the only semantic variation between I_1 and I_2 is the color of the little dog, so T_1 and T_2 should only focus on the color change in text (**black** $\rightarrow$ **white**). At the same time, the semantic difference between I_1 and I_i is the object in image (**dog** $\rightarrow$ **cat**), so the rest of the information in text should not change. With the continuous variation of pixels, for any two matched sample pairs, such as (I_1, T_1) and (I_N, T_N), we no longer require an extra intermediate pair to achieve the ESC idea. In practice, we first select some clean pairs as anchor points from the training data, and then calculate a regularization between the anchor points and the undivided pairs based on ESC. Using this regularization and triplet function, we divide training data into clean and noisy sets. Finally, the divided sets are trained by soft-margin triplet loss and our ESC regularization in a co-teaching manner [13].

Our main contributions can be summarized as follows:

- We explore a significant and challenging problem in cross-modal retrieval tasks, i.e., Noisy Correspondence Learning. We find that the most commonly used triplet loss in cross-modal matching is not robust to noise correspondence.
- We propose a simple yet effective regularization called *Equivariant Similarity Consistency* (ESC), which is intuitively explained as the semantic variations caused by image changes should be proportional to the semantic variations caused by text changes for any different matched samples. It's easy-pluggable for robust division and equivariant training in noisy correspondence problem.
- We conduct experiments on both synthetic and real-world noisy datasets and demonstrate the outstanding performance of our method.

2. Related Works

In this section, we provide a brief introduction to recent advancements in cross-modal matching, noisy correspondence learning, and equivariance learning.

2.1. Cross-Modal Matching

Cross-modal matching [8, 9, 37, 46, 48] is a challenging task in the field of multimedia analysis and information retrieval. The goal of this task is to bridge the semantic gap between different modalities and enable effective searching, browsing, and organizing of large-scale multimodal data. In recent, VSE++ [11] utilized hard negatives to optimize the triplet loss. SCAN [20] incorporated a stacked cross-attention mechanism to discover the full latent alignments using both image regions and words. VSRN [22] built up connections between image regions and performed reasoning with Graph Convolutional Networks [5]. SGRAF [10] established a graph structure for multimodal data to facilitate the learning of detailed correspondence. Unfortunately, the superior performance of all the above methods is based on a large amount of correct-matched training data, and they do not take into account the problem of noisy correspondence.

2.2. Noisy Correspondence Learning

Different from the traditional noisy labels, noisy correspondence refers to the alignment errors in paired data rather than the errors in category annotations. As a widely-exist but rarely-explored problem, Huang et al.[15] raised this issue for the first time and proposed a solution NCR simultaneously. Motivated by the memorization effect of DNNs, NCR divided the training data into clean and noisy subsets based on small-loss criterion, and then recast the rectified label as the soft margin of a hinge-based triplet loss [11]. After that, DECL [29] integrated a cross-modal evidential learning paradigm and a dynamic hinge triplet loss with positive and negative learning. Recently, MSCN [14] utilized meta-process to optimize a triplet ranking loss in order to learn discrimination from positive and negative meta-data. BiCro [49] estimated soft correspondence labels as a triplet loss's soft margins. CRCL [30] proposes an active complementary loss to directly replace the triplet loss for model training, focusing on addressing severe-noise scenarios. Nonetheless, most previous works only concentrate on how to access soft margins of the triplet loss accurately. The inherent non-robustness of this triplet loss to noise correspondence issues is overlooked. In contrast, our work proposes a regularization ESC to remediate the vulnerability of the triplet loss and enhance the robustness of cross-modal division and training.

2.3. Equivariance Learning

In deep learning, invariance learning and equivariance learning are both important research fields. Invariance learning refers to the network's ability to maintain stability in its output despite variations in the input. Differently, equivariance learning refers to the network's capability to adjust its output in response to changes in the input. It is challenging to practically implement strict group equivariance [6, 40], particularly with image data. However, equivariance learning remains crucial across a wide range of fields, including language understanding [12], representation learning [28], and self-supervised learning [38, 41]. In this work, we introduce a regularization loss ESC based on equivariance learning to robust division and training.

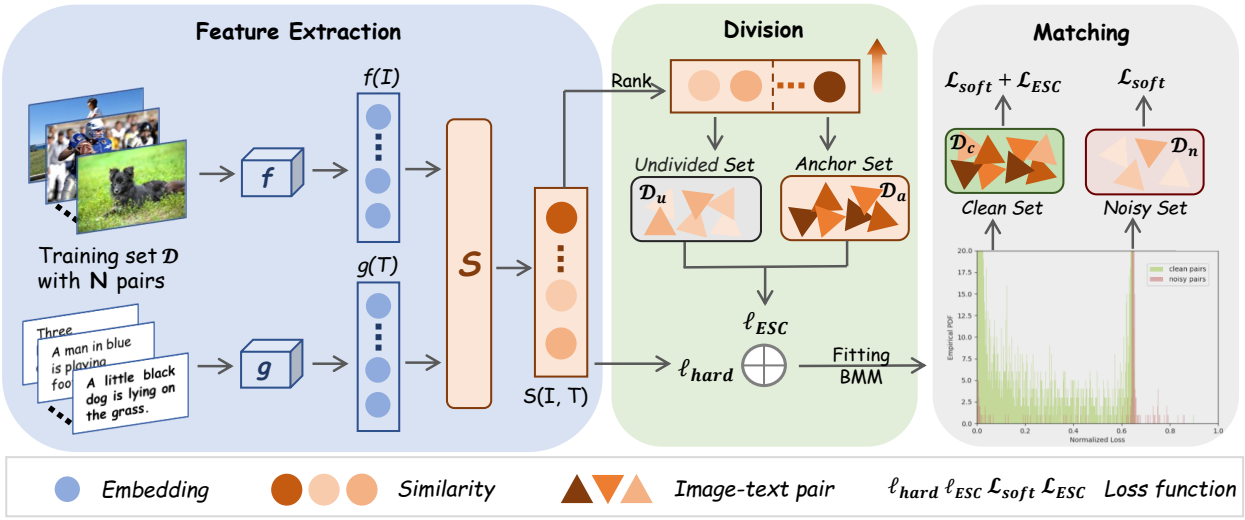

Figure 2. Overview of the proposed method. We can broadly divide the approach into three parts: 1) Feature Extraction: The matching model projects the image and text into a joint embedding space by the modal-specific networks f and g, respectively. Then, the similarity $S(I,T)$ is computed on the extracted features $f(I)$ and $g(T)$. 2) Division: The regularization l_{ESC} calculated by anchor points and remaining pairs selected via sorted similarity $S(I,T)$ strengthens the discriminability of the triplet loss l_{hard} for noisy correspondence. 3) Matching: According to different sets, the matching loss is also different.

3. Methodology

In this section, we first elaborate on the problem formulation of Noisy Correspondence Learning in Section Sec. 3.1. Then, we introduce the data division strategy and combine the triplet loss and ESC regularization to split training data in Section Sec. 3.2. Finally, we detail how to constrain the divided data above with ESC to achieve robust cross-modal matching in Sec. 3.3.

3.1. Problem Formulation

Without loss of generality, we take image-text matching as an example to discuss the noisy correspondence problem in cross-modal matching. Given a training set $\mathcal{D} = \{(I_i, T_i, y_i)\}_{i=1}^{N}$, where (I_i, T_i) is the i-th image-text pair and $y_i \in \{0, 1\}$ represents the hard correspondence label that (I_i, T_i) is the correct correspondence as $y_i = 1$ otherwise $y_i = 0$, and N is the total number of training pairs. In our implementation, we project a cross-modal pair (I_i, T_i) into a shared embedding space via two modal-specific networks f and g, and then compute their similarity by a similarity function S. The aim of cross-modal matching is that positive data pairs have higher embedding similarities and negative data pairs have lower embedding similarities. In the following paragraphs, we abbreviate the similarity $S(f(I_i), g(T_i))$ to $S(I_i, T_i)$. The noisy correspondence problem in this case manifests that (I_i, T_i) is a mismatched pair but its correspondence label $y_i = 1$. To tackle this issue, we propose our ESC regularization to achieve more robust cross-modal matching.

3.2. Data Division Based on Memorization Effect

Despite the powerful pattern recognition and feature extraction capabilities of deep neural networks (DNNs), they are prone to overfitting on noisy correspondence pairs during training, resulting in a significant decline in the performance of cross-modal matching. To tackle this problem, recent research has discovered a memorization effect in DNNs [13], wherein they prioritize learning training data with clean labels before tackling noisy labels. This observation enables us to identify clean correspondences by selecting pairs with small-loss criterion, thus necessitating the formulation of an appropriate loss function as a division loss for pair samples to partition the training data.

Hinge-based Triplet Loss. Inspired by the use of a triplet loss for image-text retrieval, we compute a hinge-based triplet loss with a hard margin α [11] for each image-text pair (I_i, T_i) by:

$$l_{hard}(I_i, T_i) = [\alpha - S(I_i, T_i) + S(I_i, \hat{T})]_+ + [\alpha - S(I_i, T_i) + S(\hat{I}, T_i)]_+, \quad (1)$$

where $\alpha > 0$ serves as a given hard margin parameter, and $[x]_+ = \max(x, 0)$. In this loss, the first term treats I_i as

queries taking over all negative text $\hat{T}$, while the second term treats T_i as queries taking over all negative images $\hat{I}$.

We can consider the triplet loss above as a division loss function, which can divide the training pairs into clean and noisy correspondences in rough by small-loss criterion. In Eq. (1), if I_i and T_i are closer to one another in the joint embedding space than to any negative, by the margin α, the triplet loss is zero. However, both image and text data are sampled from a latent manifold space, indicating the existence of a latent manifold for this dataset where variations between different pairs of data occur continuously. Depending solely on similarity distances being smaller than a threshold α to determine if a sample pair represents a clean correspondence can lead to erroneous judgments for samples that are near the threshold boundary. As a result, the small-loss criterion becomes ineffective in distinguishing between clean and noisy sample pairs, leading to a significant reduction in the efficacy of model training. In this work, we propose to impose an additional regularization on the division loss function for robust noisy correspondence learning.

Equivariant Similarity Consistency Regularization. Motivated by [39], we put forward a novel regularization called ESC, which leverages the continuous variations of data in the manifold space to filter out noisy samples.

First of all, we introduce the core idea in our method: *for two clean pairs, the semantic variation caused by image change is proportional to the semantic variation caused by text change*. Based on this idea, we can represent the semantic variations caused by two modalities. Given two correctly-matched pairs (I_1, T_1) and (I_2, T_2), we can obtain four similarity scores s_{11}, s_{12}, s_{22}, s_{21}, where s_{11} and s_{22} are self-instance similarities which are determined by I and T from one pair, s_{12} and s_{21} are cross-instance similarities which are determined by two different pairs. Using these similarities, we define the semantic variation by text change as:

$$s_{11} - s_{12} = S(I_1, T_1) - S(I_1, T_2). \tag{2}$$

Accordingly, we define the semantic variation by image change as:

$$s_{11} - s_{21} = S(I_1, T_1) - S(I_2, T_1). \tag{3}$$

The semantic variations computed from image-text similarities are equivariant. By combining Eq. (2) and Eq. (3), we deduct a ratio equality as Equivariant Similarity Consistency:

$$\frac{s_{11} - s_{12}}{s_{11} - s_{21}} = \frac{s_{22} - s_{21}}{s_{22} - s_{12}} = C = 1. \tag{4}$$

For two correct-matched samples, $C = 1$ can be derived easily. Furthermore, the ESC can be simplified as the following regularization:

$$s_{12} = s_{21}. \tag{5}$$

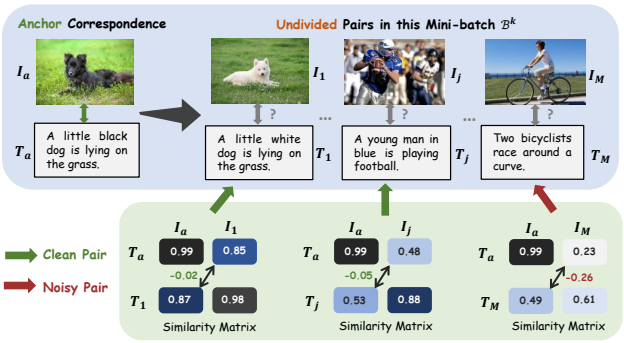

Figure 3. Illustration of ESC computation process.

Based on the conclusion of ESC in Eq. (5), we have an assumption that *any two clean pairs are constrained by their cross-instance similarities equality*. Therefore, we can collect some anchor points from the training data, considered as the 100% correct correspondence samples, and then calculate the cross-instance similarities between the anchor points and the undivided pairs. If the cross-instance similarities are close, this undivided pair will be considered as the clean one.

In detail, for the given training dataset $\mathcal{D} = (I_i, T_i, y_i)_{i=1}^{N}$, we separate it into many mini-batches $\{\mathcal{B}^k = (I_i, T_i, y_i)_{i=1}^{M}\}_{k=1}^{\frac{N}{M}}$, where M is the batch-size and k denotes the k-th mini-batch. For each mini-batch $\mathcal{B}^k$, the sample with the highest similarity score is collected as the anchor correspondence, the anchor point (I_a, T_a) and the anchor set $\mathcal{D}_a$ can be represented as follows:

$$\begin{aligned} (I_a, T_a) &= \text{argmax}_{(I_i, T_i)} S(I_i, T_i), \\ \mathcal{D}_a &= \{(I_a, T_a), \forall (I_i, T_i) \in \mathcal{B}^k, k = \{1, \cdots, \frac{N}{M}\}\}. \end{aligned} \tag{6}$$

The other undivided pairs in this mini-batch are represented as $\mathcal{B}_u = \mathcal{B}^k / (I_a, T_a)$, and the undivided set is denoted by $\mathcal{D}_u = \sum_{k=1}^{\frac{N}{M}} \mathcal{B}_u^k$. Then, we compute the cross-instance similarities between this anchor point and other pairs (I_j, T_j) in undivided set $\mathcal{B}_u$:

$$s_{aj} = S(I_a, T_j), \quad s_{ja} = S(I_j, T_a), \tag{7}$$

where s_{aj} is the cross-instance similarity between the image of the anchor correspondence and the j-th text and s_{ja} is the cross-instance similarity between the j-th image and the text of the anchor correspondence. In our implementation, we adopt Mean Square Error loss to regularize the equation

of ESC. In practice, ESC regularization can be written as:

$$l_{ESC}(I_i, T_i) = [||s_{ai} - s_{ia}||_2^2 - \alpha_1]_+, \tag{8}$$

where $[x]_+ = \max(x, 0)$, $||\cdot||_2$ denotes the L2 norm and α_1 denotes the division margin. The anchor point's l_{ESC} must be equal to zero with common sense.

Final Division Loss. For robust division, the final division loss function can be written as:

$$l_{div} = l_{hard} + \beta l_{ESC}, \tag{9}$$

where β is a hyperparameter to control the strength of regularization.

Then, we fit the division loss l_{div} of all training pairs by using a two-component Beta Mixture Model (BMM) [26]:

$$p(l_i) = \sum_{k=1}^{K} \lambda_k \phi(l_i|\gamma_k, \beta_k), \tag{10}$$

where $K = 2$, λ_k is the mixture coefficient, and $\phi(l_i|\gamma_k, \beta_k)$ indicates the probability density function with parameters $\gamma_k, \beta_k > 0$. We choose BMM over GMM because of its better performance in modeling symmetric and skewed distributions, as demonstrated in [49] and [14]. We use an Expectation Maximization [7] procedure to optimize this BMM and then compute the posterior probability $p(k|l_i)$ as the clean probability of i-th sample:

$$p(k|l_i) = p(k)p(l_i|k)/p(l_i), \tag{11}$$

where $k \in \{0, 1\}$ represents this pair is noisy or clean and $l_i \in (0, 1)$ denotes the normalized l_{div} for (I_i, T_i). Based on the aforementioned memorization effect of DNNs, the final clean set $\mathcal{D}_c$ contains the anchor correspondence samples and the clean data filtered out from the undivided set, and the noise set $\mathcal{D}_n$ contains the remaining pairs in the undivided set:

$$\begin{aligned} \mathcal{D}_c &= \mathcal{D}_a \cup \{(I_i, T_i)|p(k=0|l_i) > \delta, \forall (I_i, T_i) \in \mathcal{D}_u\}, \\ \mathcal{D}_n &= \{(I_i, T_i)|p(k=0|l_i) \leq \delta, \forall (I_i, T_i) \in \mathcal{D}_u\}. \end{aligned} \tag{12}$$

3.3. Robust Matching with Equivariant Similarity Consistency

Following [15, 21, 49], we also adopt the co-teaching manner[13] to avoid error accumulation. The detailed training pipeline is illustrated in the *supplementary material*. In practice, we maintain two matching models $\theta^A = \{f^A, g^A, S^A\}$ and $\theta^B = \{f^B, g^B, S^B\}$ with different initializations and batch sequences, respectively. Specifically, one matching model filters out the noisy pairs from training data and estimates the soft correspondence label $\hat{y} \in [0, 1]$ for each noisy data. Simultaneously, these divided pairs with their soft labels are trained by another model. Note that the soft correspondence label $\hat{y}$ is expected to be able to describe the correspondence degree between I and T (i.e., the clean pair's soft label is equal to 1, and the noisy one is close to 0). The computational detail of the soft correspondence label refers to [49]. The pairs in $\mathcal{D}_c = \{(I_i, T_i, y_i = 1)\}$ and $\mathcal{D}_n = \{(I_i, T_i, y_i = \hat{y}_i)\}$ are trained by minimizing the following triplet loss with a soft margin $\hat{\alpha}$:

$$\begin{aligned} \mathcal{L}_{soft}(I_i, T_i) = &[\hat{\alpha}_i - S(I_i, T_i) + S(I_i, \hat{T}_h)]_+ \\ &+ [\hat{\alpha}_i - S(I_i, T_i) + S(\hat{I}_h, T_i)]_+, \end{aligned} \tag{13}$$

where $\hat{\alpha}_i$ is a soft margin which is adaptively determined by the i-th sample's soft correspondence label $\hat{y}_i$. Like [15, 49], $\hat{\alpha}_i = \frac{m^{\hat{y}_i} - 1}{m - 1}\alpha$, where m is a hyperparameter.

This soft triplet loss can assign large margins to the true positive pairs and small ones to the false positive pairs. Thus, it can be utilized to learn a better shared embedding space in the noisy correspondence problem. In fact, the correspondence degrees in the clean subset $\mathcal{D}_c$ inevitably have a slight difference, but all soft correspondence labels y_i in $\mathcal{D}_c$ are equal to 1. Therefore, we can also impose the ESC mentioned above regularization on the soft triplet loss while the model is trained by clean pairs.

Similar to the aforementioned method, we select the pseudo negative sample $(\hat{I}_p, \hat{T}_p)$ of each clean pair in $\mathcal{D}_c$ at first, where $\hat{I}_p = \text{argmax}_{I_i \neq I_j} S(I_j, T_i)$ and $\hat{T}_p$ is the correspondent text of $\hat{I}_p$. The major difference between hard negative samples $\hat{I}_h$ / $\hat{T}_h$ and pseudo negative samples $\hat{I}_p$ / $\hat{T}_p$ is that pseudo negative sample $(\hat{I}_p, \hat{T}_p)$ is a correspondent pair in the clean dataset. In contrast, hard negative samples may not come from one correspondent pair. Using Eq. (7), we compute the cross-instance similarities between one pair (I_i, T_i) and its pseudo negative pair $(\hat{I}_p, \hat{T}_p)$ in the clean set. Like Eq. (8), ESC regularization in training can be written as:

$$\mathcal{L}_{ESC}(I_i, T_i) = [||s_{pi} - s_{ip}||_2^2 - \alpha_2]_+, \tag{14}$$

where s_{pi} is the cross-instance similarity between $\hat{I}_p$ and T_i, s_{ip} is the cross-instance similarity between I_i and $\hat{T}_p$, and α_2 is also a margin hyperparameter. In the end, the final matching loss function for $\mathcal{D}_n$ is also $\mathcal{L}_{soft}$ and the loss for $\mathcal{D}_c$ is determined by:

$$\mathcal{L} = \mathcal{L}_{soft} + \mathcal{L}_{ESC}. \tag{15}$$

4. Experiments

4.1. Experimental Settings

Datasets. The following three datasets are used to evaluate our method and baselines, where *Flickr30K* [50] and *MS-COCO* [24] contain the synthetic noise and *Conceptual*

Captions [36] contains the real-world noise. The details in these datasets are delineated as follows:

- **Flickr30K:** The Flickr30K dataset contains 31,014 images with five captions each, collected from the Flickr website. In our experiments, we use 1,014 images for model validation, 1,000 for model testing, and 29,000 for model training.
- **MS-COCO:** This dataset is widely used in cross-modal learning, which contains 123,287 images, and each image is associated with five captions. Following the split in [20], 5,000 images are used for modal validation, 5,000 for model testing, and 113,287 for model training.
- **Conceptual Captions:** Conceptual Captions is a large-scale real-world dataset with noisy correspondence containing about 3% $\sim$ 20% mismatched image-text pairs. It comprises 3,334,173 images with a single caption each. Following [15], we use a smaller version of the Conceptual Captions dataset in terms of the number of pairs, i.e., CC152K. 1,000 images are used for model validation, 1,000 images are used for model testing, and 150,000 images are used for model training in CC152K.

Evaluation Metrics. We evaluate the retrieval performance with the recall at K (R@K) metric. R@K measures the proportion of relevant items successfully retrieved from the top K items. In our experiments, we report the results of R@1, R@5, R@10, and the sum of three recalls for image-to-text and text-to-image matching. Among them, due to the demands of user experience in practical applications, the most critical evaluation metric for retrieval tasks is R@1.

Implementation Details. Our method can be easily implemented with nearly all cross-modal matching techniques to enhance robustness. Like [20], we first take the Faster-RCNN [33] to extract the top 36 regions for every image as a preprocess. Following previous noisy correspondence works [15], a full-connected layer serves as the image embedding extractor f, while a Bi-GRU [35] serves as the text embedding extractor g. The similarity function S is computed by combining local and global features using graph reasoning techniques proposed in [10]. Before training, we warmup the matching models θ^A and θ^B for 10 epochs on the original training data to achieve initial convergence with l_{hard}. Then, we train two models using the Adam optimizer [19] with the default parameters and a batch size 128 for 40 epochs. Following [49], clean samples train the models during the first 20 epochs, and all samples train the subsequent 20 epochs. At each training epoch, we choose the pair whose similarity is ranked at the top 1 in every mini-batch as anchor correspondence, which computes the l_{ESC} with the remaining pairs in this mini-batch to determine whether these undivided samples are clean or not. Moreover, we set the margin α as 0.2 and $m = 10$ to calculate the soft margin. In division loss, we set the hyperparameters $\alpha_1 = 0$ and $\beta = 0.5$. In matching loss, we set $\alpha_2 = 0$ as well. The choice of these hyperparameters will be discussed in *supplementary material*. Finally, we average the similarity scores from two matching models at the inference phase.

Methods	Image $\longrightarrow$ Text			Text $\longrightarrow$ Image			
	R@1	R@5	R@10	R@1	R@5	R@10	rSum
SCAN	30.5	55.3	65.3	26.9	53.0	64.7	295.7
VSRN	32.6	61.3	70.5	32.5	59.4	70.4	326.7
IMRAM	33.1	57.6	68.1	29.0	56.8	67.4	312.0
SAF	31.7	59.3	68.2	31.9	59.0	67.9	318.0
SGR	11.3	29.7	39.6	13.1	30.1	41.6	165.4
NCR	39.5	64.5	73.5	40.3	64.6	73.2	355.6
DECL	39.0	66.1	75.5	40.7	66.3	76.7	364.3
BiCro	40.8	67.2	76.1	42.1	67.6	76.4	368.9
MSCN	40.1	65.7	76.6	40.6	67.4	76.3	366.7
CRCL	41.8	**67.4**	76.5	41.6	68.0	**78.4**	373.7
ESC	**42.8**	67.3	**76.9**	**44.8**	**68.2**	75.9	**375.9**

Table 1. Image-text matching performance on CC152K. **Best** and second-best results are highlighted in each column.

Noise Ratio	Method	Image $\longrightarrow$ Text			Text $\longrightarrow$ Image			
		R@1	R@5	R@10	R@1	R@5	R@10	rSum
40%	NCR	55.5	82.4	90.2	39.7	68.5	79.2	415.5
	BiCro	**56.3**	83.0	**90.8**	40.1	69.0	79.5	418.7
	MSCN	49.7	78.9	88.0	36.9	66.1	77.1	396.7
	CRCL	55.8	83.1	90.1	40.9	67.8	**80.6**	418.3
	ESC	56.2	**83.2**	90.7	**41.0**	**69.5**	79.8	**420.4**
60%	NCR	49.6	78.1	87.3	35.5	64.2	75.7	390.4
	BiCro	52.5	80.0	88.4	37.8	66.2	77.1	402.0
	MSCN	48.1	76.0	85.5	34.5	63.5	75.1	382.7
	CRCL	53.1	**81.2**	89.0	37.6	66.3	77.4	404.6
	ESC	**53.4**	81.1	**89.2**	**38.2**	**66.7**	**77.5**	**406.1**

Table 2. Image-text matching performance on MS-COCO 5K with 40% and 60% noises. **Best** and second-best results are highlighted in each column.

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

In this section, we carry out experiments to present the performance of ESC on the three datasets above. Since the data in Flickr30K and MS-COCO is correctly matched, We generated noisy correspondences by randomly shuffling the captions of training images, with the percentage denoted as the noise ratio. Specifically, we conduct comparison experiments under 20%, 40%, and 60% correspondence noise scenarios. Although the latest published work [30] presented experimental results with 80% noise, we hold the opinion

Noise Ratio	Methods	Flickr30K							MS-COCO						
		Image ⟶ Text			Text ⟶ Image				Image ⟶ Text			Text ⟶ Image			
		R@1	R@5	R@10	R@1	R@5	R@10	rSum	R@1	R@5	R@10	R@1	R@5	R@10	rSum
20%	SCAN	58.5	81.0	90.8	35.5	65.0	75.2	406.0	62.2	90.0	96.1	46.2	80.8	89.2	464.5
	VSRN	33.4	59.5	71.3	25.0	47.6	58.6	295.4	61.8	87.3	92.9	50.0	80.3	88.3	460.6
	IMRAM	22.7	54.0	67.8	16.6	41.8	54.1	257.0	69.9	93.6	97.4	55.9	84.4	89.6	490.8
	SAF	62.8	88.7	93.9	49.7	73.6	78.0	446.7	71.5	94.0	97.5	57.8	86.4	91.9	499.1
	SGR	55.9	81.5	88.9	40.2	66.8	75.3	408.6	25.7	58.8	75.1	23.5	58.9	75.1	317.1
	NCR*	75.0	93.9	97.5	58.3	83.0	89.0	496.7	77.7	95.6	98.2	62.6	89.3	95.3	518.7
	DECL	77.5	93.8	97.0	56.1	81.8	88.5	494.7	77.5	95.9	98.4	61.7	89.3	95.4	518.2
	BiCro*	76.5	93.1	97.4	58.1	82.3	88.5	495.9	78.8	96.1	98.6	63.7	90.3	95.7	523.2
	MSCN*	76.4	94.5	97.6	58.8	83.5	**89.2**	500.0	78.1	**97.2**	98.8	64.3	90.4	95.8	524.6
	CRCL*	78.9	**94.8**	**97.9**	58.7	83.0	**89.2**	502.5	77.8	96.1	98.5	63.4	90.3	95.9	522.0
	ESC	**79.0**	**94.8**	97.5	**59.1**	**83.8**	89.1	**503.3**	**79.2**	97.0	**99.1**	**64.8**	**90.7**	**96.0**	**526.8**
40%	SCAN	26.0	57.4	71.8	17.8	40.5	51.4	264.9	42.9	74.6	85.1	24.2	52.6	63.8	343.2
	VSRN	2.6	10.3	14.8	3.0	9.3	15.0	55.0	29.8	62.1	76.6	17.1	46.1	60.3	292.0
	IMRAM	5.3	25.4	37.6	5.0	13.5	19.6	106.4	51.8	82.4	90.9	38.4	70.3	78.9	412.7
	SAF	7.4	19.6	26.7	4.4	12.2	17.0	87.3	13.5	43.8	48.2	16.0	39.0	50.8	211.3
	SGR	4.1	16.6	24.1	4.1	13.2	19.7	81.8	1.3	3.7	6.3	0.5	2.5	4.1	18.4
	NCR*	73.5	92.6	95.8	55.7	80.3	86.9	484.8	76.6	95.6	98.2	61.0	88.9	94.9	515.2
	DECL	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
	BiCro*	72.5	91.7	95.3	53.6	79.0	86.4	478.5	75.1	95.9	98.3	59.8	89.1	94.9	513.1
	MSCN*	69.5	90.8	95.7	53.2	79.9	86.4	475.5	74.5	96.0	98.1	60.8	89.0	95.0	513.4
	CRCL*	74.1	92.6	**96.9**	55.5	**80.9**	**87.6**	487.6	76.6	95.6	98.5	62.3	89.7	95.4	518.1
	ESC	**76.1**	**93.1**	96.4	**56.0**	80.8	87.2	**489.6**	**78.6**	**96.6**	**99.0**	**63.2**	**90.6**	**95.9**	**523.9**
60%	SCAN	13.6	36.5	50.3	4.8	13.6	19.8	138.6	29.9	60.9	74.8	0.9	2.4	4.1	173.0
	VSRN	0.8	2.5	5.3	1.2	4.2	6.9	20.9	11.6	34.0	47.5	4.6	16.4	25.9	140.0
	IMRAM	1.5	8.9	17.4	1.9	5.0	7.8	42.5	18.2	51.6	68.0	17.9	43.6	54.6	253.9
	SAF	0.1	1.5	2.8	0.4	1.2	2.3	8.3	0.1	0.5	0.7	0.8	3.5	6.3	11.9
	SGR	1.5	6.6	9.6	0.3	2.3	4.2	24.5	0.1	0.6	1.0	0.1	0.5	1.1	3.4
	NCR*	70.0	**91.0**	94.4	52.3	76.9	84.0	468.6	72.6	93.8	97.4	57.0	86.4	93.6	500.8
	DECL	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
	BiCro*	68.5	89.1	93.1	48.2	74.8	82.7	456.4	73.9	94.7	97.9	58.7	87.0	93.8	506.0
	MSCN*	68.8	88.6	93.1	48.8	76.4	84.0	459.7	73.7	**95.1**	**98.5**	57.0	86.9	94.0	505.2
	CRCL*	70.4	90.4	**94.9**	52.6	78.1	85.1	471.5	75.2	94.9	98.0	60.1	88.5	94.8	511.5
	ESC	**72.6**	90.9	94.6	**53.0**	**78.6**	**85.3**	**475.0**	**77.2**	**95.1**	98.1	**61.1**	**88.6**	**94.9**	**515.0**

Table 3. Image-text matching performance under synthetic noise ratios of 20%, 40%, and 60% on Flickr30K and MS-COCO 1K. **Best** and second-best results are highlighted in each column. (*) indicates that we run the algorithm.

that extremely severe noise does not have practical application scenarios. The CC152K is collected from the Internet and contains a large portion of mismatched pairs in real world. The baselines of our method include general matching methods (SCAN [20], VSRN [22], IMRAM [4], SGRAF, SGR, and SAF [10]) and specific methods with noisy correspondence (NCR [15], DECL [29], BiCro [49], MSCN [14], and CRCL [30]). For all methods, we select the best checkpoint on the validation dataset and report its performance on the test dataset.

Experiments on Real-world Noise. Tab. 1 presents the performance comparison of ESC and other state-of-the-art methods on CC152K. CC152K remains a challenging benchmark dataset due to its real-world noisy correspondence. Our ESC demonstrates outstanding performance from the results, especially R@1 for text-to-image matching, which is 2.7% higher than the best baseline. Our method exhibits a larger improvement on R@1 compared to R@5 and R@10, indicating that our ESC can better enhance retrieval accuracy.

Top1: Three men are working on a roof .
Top2: Two men sitting on the roof of a house while another one stands on a ladder .

Top1: A woman reads a book while sitting in a row of red chairs .
Ground Truth: A man wearing a reflective vest sits on the sidewalk and holds up pamphlets with bicycles on the cover .

Figure 4. Some image-to-text retrieval results on Flickr30K.

Experiments on Synthetic Noise. In Tab. 3, we present our experimental results on Flickr30k and MS-COCO 1K, respectively. For experiments, we consider synthetic noise ratios of 20%, 40%, and 60%. ESC performs consistently better from moderate to severe noise on these two datasets than baseline methods. In the case of Flick30K, ESC improves the sum score of recalls by 0.8%, 2.0%, and 3.5% under different noise ratios. For MS-COCO, the sum scores of ESC are 2.2%, 5.8%, and 3.5% higher than the best baseline. Note that a method like [15] fails for high noise ratio. This situation may be due to excessive false positive samples, which renders the triplet loss ineffective. In Tab. 2, we also demonstrate the superior performance of ESC on the MSCOCO full 5K dataset. And we present some results of image-to-text retrieval on the Flickr30K in Fig. 4. The left column displays correct retrieval results, where Top1 and Top2 indicate the two text descriptions with the highest similarity match to the image. The right column shows a failed case where multiple objects in the image make it difficult for the sparse text to fully capture the image context, resulting in incorrect matches for similar texts.

4.3. Ablation Study

In this section, we conduct an ablation study of the proposed components to evaluate their effectiveness in Tab. 4, which is carried out on the Flickr30K with 40% noise.

- **Warmup:** Before training, we use all available data to train for several epochs, instead of segregating it into clean and noisy subsets as Warmup process for a rapid convergence of the matching models.
- **Division Regularization l_{ESC}:** Although the division regularization loss l_{ESC} used to separate noisy correspondence does not participate in gradient backpropagation, i.e., it is not involved in model optimization, it plays a crucial role in improving the model's performance. To some extent, ESC regularization compensates for the robustness of the triplet loss, resulting in a more vital constraint needed to classify training data as clean samples.

Methods			Image $\longrightarrow$ Text		
Warmup	l_{ESC}	$\mathcal{L}_{ESC}$	R@1	R@5	R@10
✓	✓	✓	**76.1**	**93.1**	**96.4**
✓	✓		73.0	92.3	95.8
✓		✓	74.4	92.5	95.5
	✓	✓	5.8	19.8	26.8
Methods			Text $\longrightarrow$ Image		
Warmup	l_{ESC}	$\mathcal{L}_{ESC}$	R@1	R@5	R@10
✓	✓	✓	**56.0**	**80.8**	**87.2**
✓	✓		55.4	80.5	86.9
✓		✓	55.6	80.0	86.7
	✓	✓	4.6	15.7	23.8

Table 4. Ablation studies on Flickr30K with 40% noise ratio.

- **Matching Regularization $\mathcal{L}_{ESC}$:** The final matching regularization function $\mathcal{L}_{ESC}$ needs to undergo gradient backpropagation and directly affects the optimization performance of models. Note that this regularization is only applied to clean samples, constraining the embedding distribution of clean samples to a more reasonable position in the embedding space. This constraint will be beneficial for noise separation in the subsequent training steps.

5. Conclusion

In this work, we explore a simple yet effective method ESC to address a significant and challenging problem of learning with noisy correspondence. The key idea in our proposed method is *the semantic variations caused by image changes should be proportional to those caused by text changes for any different matched samples*. Based on this grounded theory, we propose a regularization called ESC to achieve robust division and training in cross-modal matching. Meanwhile, we conduct experiments on three widely used datasets to verify the effectiveness of our method in both synthetic and real-world noise correspondences.

6. Acknowledgement

Our work was supported in part by the National Key R&D Program of China (No.2023YFC3305600), Joint Fund of Ministry of Education of China (8091B022149, 8091B02072404), National Natural Science Foundation of China (62202365, 62132016, 62171343, and 62071361), and Fundamental Research Funds for the Central Universities (ZDRC2102), Guangdong Basic and Applied Basic Research Foundation (2021A1515110026), Natural Science Basic Research Program of Shaanxi (No.2022JQ-608), and Young Elite Scientists Sponsorship Program by CAST (2023QNRC001).

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):8717–8727, 2018. 1

[2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, pages 609–617, 2017. 1

[3] Yingbin Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *NeurIPS*, 34:24392–24403, 2021. 1

[4] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *CVPR*, pages 12655–12663, 2020. 7

[5] Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In *ICML*, pages 1725–1735. PMLR, 2020. 2

[6] Taco Cohen and Max Welling. Group equivariant convolutional networks. In *ICML*, pages 2990–2999. PMLR, 2016. 2

[7] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *J. R. Stat. Soc., B: Stat. Methodol.*, 39(1):1–22, 1977. 5

[8] Cheng Deng, Erkun Yang, Tongliang Liu, Jie Li, Wei Liu, and Dacheng Tao. Unsupervised semantic-preserving adversarial hashing for image search. *IEEE Trans. Image Process.*, 28(8):4032–4044, 2019. 2

[9] Cheng Deng, Erkun Yang, Tongliang Liu, and Dacheng Tao. Two-stream deep hashing with class-specific centers for supervised image search. *IEEE Trans. Neural Netw. Learn. Syst.*, 31(6):2189–2201, 2019. 2

[10] Haiwen Diao, Ying Zhang, Lin Ma, and Huchuan Lu. Similarity reasoning and filtration for image-text matching. In *AAAI*, pages 1218–1226, 2021. 2, 6, 7

[11] 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, 3

[12] Jonathan Gordon, David Lopez-Paz, Marco Baroni, and Diane Bouchacourt. Permutation equivariant models for compositional generalization in language. In *ICLR*, 2019. 2

[13] Bo Han, Quanming Yao, Xingrui Yu, Gang Niu, Miao Xu, Weihua Hu, Ivor Tsang, and Masashi Sugiyama. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *NeurIPS*, 31, 2018. 1, 2, 3, 5

[14] Haochen Han, Kaiyao Miao, Qinghua Zheng, and Minnan Luo. Noisy correspondence learning with meta similarity correction. In *CVPR*, pages 7517–7526, 2023. 2, 5, 7

[15] Zhenyu Huang, Guocheng Niu, Xiao Liu, Wenbiao Ding, Xinyan Xiao, Hua Wu, and Xi Peng. Learning with noisy correspondence for cross-modal matching. *NeurIPS*, 34: 29406–29419, 2021. 1, 2, 5, 6, 7, 8

[16] Ahmet Iscen, Jack Valmadre, Anurag Arnab, and Cordelia Schmid. Learning with neighbor consistency for noisy labels. In *CVPR*, pages 4672–4681, 2022. 1

[17] Lu Jiang, Di Huang, Mason Liu, and Weilong Yang. Beyond synthetic noise: Deep learning on controlled noisy labels. In *ICML*, pages 4804–4815. PMLR, 2020.

[18] Nazmul Karim, Mamshad Nayeem Rizve, Nazanin Rahnavard, Ajmal Mian, and Mubarak Shah. Unicon: Combating label noise through uniform selection and contrastive learning. In *CVPR*, pages 9676–9686, 2022. 1

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

[20] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *ECCV*, pages 201–216, 2018. 2, 6, 7

[21] Junnan Li, Richard Socher, and Steven CH Hoi. Dividemix: Learning with noisy labels as semi-supervised learning. *arXiv preprint arXiv:2002.07394*, 2020. 5

[22] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *ICCV*, pages 4654–4662, 2019. 2, 7

[23] Bin Liang, Chenwei Lou, Xiang Li, Min Yang, Lin Gui, Yulan He, Wenjie Pei, and Ruifeng Xu. Multi-modal sarcasm detection via cross-modal graph convolutional network. In *Trans. Assoc. Comput. Linguist.*, pages 1767–1777. Association for Computational Linguistics, 2022. 1

[24] TY Lin, M Maire, S Belongie, J Hays, P Perona, D Ramanan, P Dollár, CL Zitnick, et al. Microsoft coco: Common objects in context, 2014. 5

[25] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. Graph structured network for image-text matching. In *CVPR*, pages 10921–10930, 2020. 1

[26] Zhanyu Ma and Arne Leijon. Bayesian estimation of beta mixture models with variational inference. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(11):2160–2173, 2011. 5

[27] Aditya Prakash, Kashyap Chitta, and Andreas Geiger. Multimodal fusion transformer for end-to-end autonomous driving. In *CVPR*, pages 7077–7087, 2021. 1

[28] Guo-Jun Qi, Liheng Zhang, Feng Lin, and Xiao Wang. Learning generalized transformation equivariant representations via autoencoding transformations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(4):2045–2057, 2020. 2

[29] Yang Qin, Dezhong Peng, Xi Peng, Xu Wang, and Peng Hu. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *ACM MM*, pages 4948–4956, 2022. 1, 2, 7

[30] Yang Qin, Yuan Sun, Dezhong Peng, Joey Tianyi Zhou, Xi Peng, and Peng Hu. Cross-modal active complementary learning with self-refining correspondence. In *NeurIPS*, 2023. 2, 6, 7

[31] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, pages 8821–8831. PMLR, 2021. 1

[32] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *NeurIPS*, 28, 2015. 6

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1

[35] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process*, 45(11):2673–2681, 1997. 6

[36] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *ACL*, pages 2556–2565, 2018. 6

[37] Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *WACV*, pages 1508–1517, 2020. 2

[38] Tan Wang, Zhongqi Yue, Jianqiang Huang, Qianru Sun, and Hanwang Zhang. Self-supervised learning disentangled group representation as feature. *NeurIPS*, 34:18225–18240, 2021. 2

[39] Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Equivariant similarity for vision-language foundation models. In *ICCV*, pages 11998–12008, 2023. 2, 4

[40] Maurice Weiler and Gabriele Cesa. General e (2)-equivariant steerable cnns. *NeurIPS*, 32, 2019. 2

[41] Yuyang Xie, Jianhong Wen, Kin Wai Lau, Yasar Abbas Ur Rehman, and Jiajun Shen. What should be equivariant in self-supervised learning. In *CVPR*, pages 4111–4120, 2022. 2

[42] Jiexi Yan, Lei Luo, Chenghao Xu, Cheng Deng, and Heng Huang. Noise is also useful: Negative correlation-steered latent contrastive learning. In *CVPR*, pages 31–40, 2022. 1

[43] Jiexi Yan, Lei Luo, Cheng Deng, and Heng Huang. Adaptive hierarchical similarity metric learning with noisy labels. *IEEE Trans. Image Process.*, 32:1245–1256, 2023. 1

[44] Erkun Yang, Cheng Deng, Wei Liu, Xianglong Liu, Dacheng Tao, and Xinbo Gao. Pairwise relationship guided deep hashing for cross-modal retrieval. In *AAAI*, 2017. 1

[45] Erkun Yang, Cheng Deng, Chao Li, Wei Liu, Jie Li, and Dacheng Tao. Shared predictive cross-modal deep quantization. *IEEE Trans. Neural Netw. Learn. Syst.*, 29(11):5292–5303, 2018. 1

[46] Erkun Yang, Tongliang Liu, Cheng Deng, Wei Liu, and Dacheng Tao. Distillhash: Unsupervised deep hashing by distilling data pairs. In *CVPR*, pages 2946–2955, 2019. 2

[47] Erkun Yang, Dongren Yao, Tongliang Liu, and Cheng Deng. Mutual quantization for cross-modal search with noisy labels. In *CVPR*, pages 7551–7560, 2022. 1

[48] Song Yang, Qiang Li, Wenhui Li, Xuanya Li, and An-An Liu. Dual-level representation enhancement on characteristic and context for image-text retrieval. *IEEE Trans. Circuits Syst. Video Technol.*, 32(11):8037–8050, 2022. 2

[49] 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 *CVPR*, pages 19883–19892, 2023. 1, 2, 5, 6, 7

[50] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Linguist.*, 2:67–78, 2014. 5

[51] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Commun. ACM*, 64(3):107–115, 2021. 1

[52] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *CVPR*, pages 833–842, 2021. 1

[53] Qi Zhang, Zhen Lei, Zhaoxiang Zhang, and Stan Z Li. Context-aware attention network for image-text retrieval. In *CVPR*, pages 3536–3545, 2020. 1