TSVC: Tripartite Learning with Semantic Variation Consistency for Robust Image-Text Retrieval

Shuai Lyu[1], Zijing Tian[2], Zhonghong Ou[3*], Yifan Zhu[1], Xiao Zhang[1], Qiankun Ha[1], Haoran Luo[1], Meina Song[1]

[1]School of Computer Science, Beijing University of Posts and Telecommunications, China
[2]School of Science, Beijing University of Posts and Telecommunications, China
[3]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, China
{Lxb_savior, tianzj, zhonghong.ou, yifan_zhu, xiao20010420}@bupt.edu.cn

Abstract

Cross-modal retrieval maps data under different modality via semantic relevance. Existing approaches implicitly assume that data pairs are well-aligned and ignore the widely existing annotation noise, i.e., noisy correspondence (NC). Consequently, it inevitably causes performance degradation. Despite attempts that employ the co-teaching paradigm with identical architectures to provide distinct data perspectives, the differences between these architectures are primarily stemmed from random initialization. Thus, the model becomes increasingly homogeneous along with the training process. Consequently, the additional information brought by this paradigm is severely limited. In order to resolve this problem, we introduce a Tripartite learning with Semantic Variation Consistency (TSVC) for robust image-text retrieval. We design a tripartite cooperative learning mechanism comprising a Coordinator, a Master, and an Assistant model. The Coordinator distributes data, and the Assistant model supports the Master model's noisy label prediction with diverse data. Moreover, we introduce a soft label estimation method based on mutual information variation, which quantifies the noise in new samples and assigns corresponding soft labels. We also present a new loss function to enhance robustness and optimize training effectiveness. Extensive experiments on three widely used datasets demonstrate that, even at increasing noise ratios, TSVC exhibits significant advantages in retrieval accuracy and maintains stable training performance.

Introduction

Cross-modal retrieval (Anderson et al. 2018; Li et al. 2019) aims to accurately associate and align data from different modalities, e.g., images and texts. As a key technology in the field of multi-modality, it has been widely applied in both industry and academia. In classification tasks, noise labels (Yan et al. 2023; Iscen et al. 2022; Yan et al. 2022) generally refer to labeling errors. In cross-modal matching tasks, however, noise labels pertain to alignment errors in paired data, also known as noise correspondence. Consequently, a large fraction of existing noise-robust learning methods (Zhong et al. 2024) designed for classification cannot be directly applied for cross-modal matching tasks.

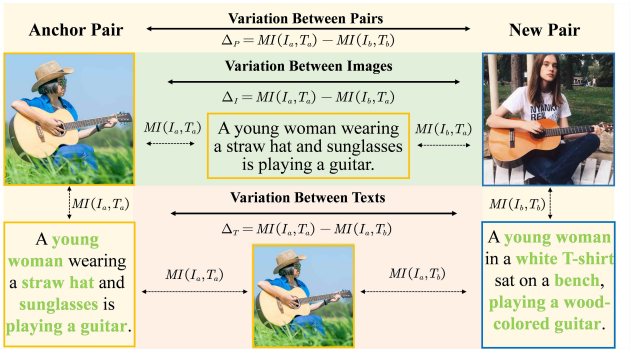


Figure 1: Illustration of semantic variance consistency guided sample filtering. We use Mutual Information (MI) to calculate the level of noise between anchor pair and new pair. We primarily consider the MI variation among sample pairs Δ_P, images Δ_I, and texts Δ_T. The smaller the variation is, the cleaner we consider the pair to be.

To address noisy correspondences in cross-modal matching, existing studies mainly focus on two aspects: noisy sample label estimation (Ma et al. 2024; Huang et al. 2024b) and loss function adjustment (Hu et al. 2021; Shi et al. 2024). Noisy sample label estimation primarily involves re-estimating and adjusting the labels of noisy samples by exploring the latent relationships within the dataset. The loss function adjustment aims to enhance model robustness by designing new loss functions. Recent studies have increasingly focused on the estimation of noisy sample labels.

Accurately estimating soft corresponding labels for noisy data has always been a big challenge for noise-robust cross-modal matching. Previous studies (Yang et al. 2024; Zhao et al. 2024b) have made efforts by employing the rate of similarity changes to quantify noise content. Yang et al. (Yang et al. 2023) leverage the inherent similarity between the two modalities to assign pseudo labels. Nevertheless, they fail to consider other crucial data relationships and features, thus limiting their ability to identify complex noise. On the other hand, some studies (Zheng, Awadallah, and Dumais 2021; Zhang, Li, and Ye 2024) mainly rely on the memory effect of Deep Neural Networks (DNN) (Zhang et al. 2021), which tends to prioritize learning simple patterns rather than noisy

*Corresponding author.

samples. They divide the training data into a clean set and a noisy set, and use the Co-Teaching training scheme (Han et al. 2018; Yan et al. 2023) to train them with different mechanisms.

Nevertheless, in the Co-Teaching paradigm, the differences between two networks with the same architecture primarily arise from random initialization. During the training process, both sides provide what they deem as important data to each other for training, resulting in limited additional information gain. Moreover, the predominant approaches (Yang et al. 2023; Qin et al. 2024; Yang et al. 2024) for noise-robust tasks typically used minimizing triplet loss with a soft margin, which fails to consider the differences and distribution characteristics between clean and noisy samples, causing suboptimal performance in distinguishing between them.

In order to resolve the problems mentioned above, we propose a **T**ripartite learning with **S**emantic **V**ariation **C**onsistency (**TSVC**) scheme for robust image-text retrieval, which mainly consists of three components.

First, we propose a **S**emantic **I**nformation **V**ariation **C**onsistency (**SIVC**) method for label estimation method, based on semantic variation of Mutual Information (MI) between new pairs and clean pairs. As shown in Fig. 1, we take into consideration three parts, i.e., pairs, images, and texts. The smaller the change, the closer the MI of new pairs is to that of clean pairs. It indicates that the new samples are cleaner, leveraging broader data relationships to better identify and quantify noise for label estimation. Second, to avoid model homogenization, we propose a novel **Tri**partite Cooperative **Learning** Mechanism (**Tri-learning**) that deviates from the conventional Co-Teaching paradigm. Specifically, Tri-Learning consists of three models: a Coordinator, a Master, and an Assistant. The Coordinator partitions data into clean and noisy sets. The Assistant selects clean samples from the noisy set to enhance the Master model. The Master trains on diverse data while preserving the ability to extract clean samples. The Coordinator adjusts partitioning for the next round based on feedback from the Assistant. Third, we train the segmented dataset using the newly introduced **D**istribution-**A**daptive **S**oft **M**argin (**DASM**) loss function, which takes into account the dynamic margin affected by rectified soft label and sample distribution deviation. Our main contributions are summarized as follows:

- We introduce a novel training paradigm Tri-learning to establish a collaborative relationship among three models. It mitigates the improvement limitation in traditional co-training paradigms, which is caused by the homogenization of models.
- We propose a soft label estimation method namely SIVC based on semantic Mutual Information variation, and present a loss function called DASM that corrects margins and distribution deviations, which enhances noise detection accuracy and robustness significantly.
- We conduct extensive experiments on three cross-modal noise datasets, including synthetic and real-world noises. Experimental results show that TSVC significantly outperforms state-of-the-art methods.

Related Work

Cross-modal Matching

Cross-modal matching (Yang et al. 2022; Deng et al. 2019; Wang et al. 2020) is an fundamental problem in the fields of information retrieval (Zhao et al. 2024a) and multi-modal analysis. It aims to retrieve one modality based on another modality as a query.

Image-text retrieval, as the most common task, can be roughly divided into two categories, i.e., coarse-grained alignment and fine-grained alignment. Coarse-grained alignment (Chen et al. 2021; Faghri et al. 2017; Diao et al. 2021) usually employs two separate networks to project the entire image and text into a unified embedding space. It then calculates the overall correlation using cosine similarity. Fine-grained alignment methods (Lee et al. 2018; Chen et al. 2020b; Zhang et al. 2022) often combine cross-modal interaction with local segment alignment. Afterwards, they accumulate the scores of local regions to acquire the overall similarity (Wu et al. 2019). It is worth noting that the superior performance of these methods relies on a large amount of accurately annotated training data, without considering the issue of noise correspondence.

Noisy Correspondence Learning

Noise correspondence, different from noise label learning, usually refers to mismatched sample pairs in multi-modal datasets. Some studies (Huang et al. 2021; Han et al. 2023; Ma et al. 2024) leverage the memory effect of DNN to construct error correction networks to enhance the data purification process or identify mismatching samples. Other works (Qin et al. 2022; Yang et al. 2023; Han et al. 2024; Li et al. 2024) improve the objective function with dynamic loss or bi-directional cross-modal similarity to identify the noise contained in the samples.

The methods mentioned above assume that the two networks can provide distinct data perspectives. Nevertheless, the additional information improvement remains limited, because the differences between the two networks with the same architecture primarily stem from random initialization (Qin et al. 2024). In contrast, we propose a Tripartite Cooperative Learning Method to augment network diversity, and utilize variations in Mutual Information across samples and modalities to estimate soft correspondence labels, which enhances the model's robustness against noise significantly.

The Proposed Method TSVC

Problem Formulation

Given the dataset $\mathcal{D} = (I_i, T_i, y_i)_{i=1}^{N}$ consisting of N samples for training. Each of them includes one pair of image and text, as well as a binary label y_i indicating the relevance of the pair as positive ($y_i = 1$) or negative ($y_i = 0$). In cross-modal matching, the primary objective is to maximize the similarity between positive samples (matched pairs) while minimizing the similarity between negative samples (unmatched pairs).

Considering that multi-modal datasets are frequently annotated or acquired from Internet using cost-effective methods, it is inevitable to encounter noisy data in cross-modal

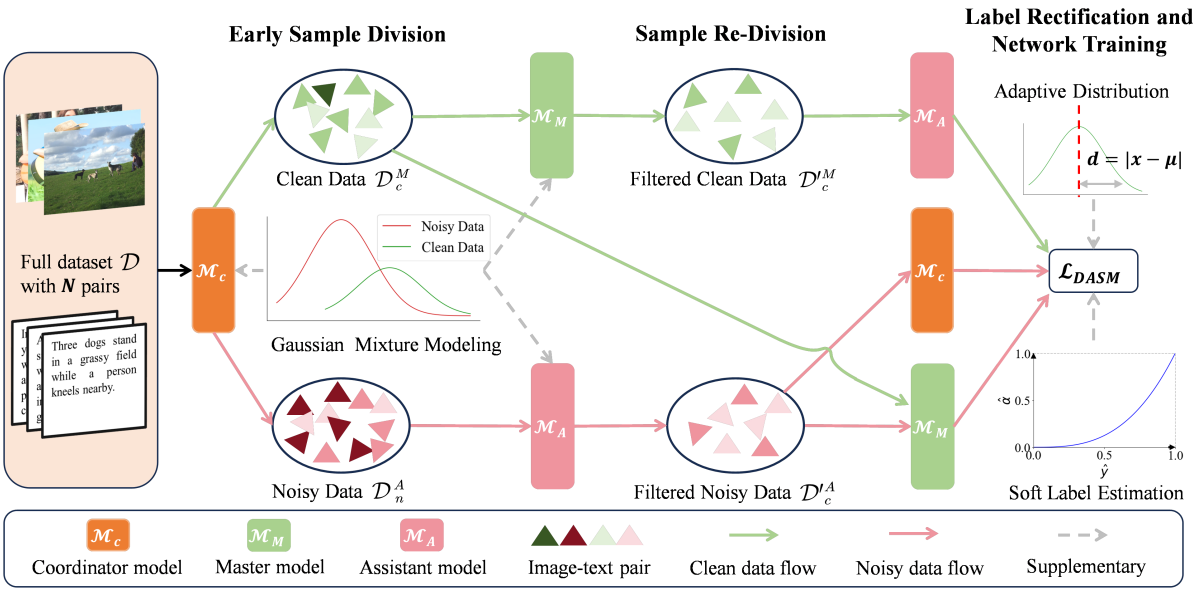


Figure 2: Illustration of Tripartite Cooperative Learning process. We divide the full data into two flows. We first utilize the noisy data and filtered clean data to train the Assistant model. We then employ the filtered noisy data to train the Coordinator and the Master model. Note that the cleaner samples are lighter in color. The input to the three models consists of image-text sample pairs. During the division stage, the output is the DASM loss used by GMM for classification. In the inference stage, it is the predicted probability.

matching tasks, which is known as the noisy correspondence problem. Specifically, it indicates that (I_i, T_i) is a mismatched pair with corresponding label $y_i = 1$, which makes the model overfit to the noisy dataset and pull the negative samples closer, causing misleading results. To address this problem, we propose a framework called **Tripartite Learning with Semantic Variation Consistency (TSVC)** to enhance its robustness to noise. The details are outlined below.

Semantic Information Variation Consistency

Mutual Information. Existing studies (Huang et al. 2024a; Wang et al. 2024) suggest that clean samples have a stronger correlation between images and texts. Consequently, Semantic Information Variation Consistency (SIVC) leverages Mutual Information (MI) (Shannon 1948) to estimate soft correspondence labels. MI quantifies the shared information between signals, providing a criterion to measure the dependency between image-text pairs and determine sample noise proportion. For a pair of discrete sample pairs (X, Y), MI is defined as:

$$MI(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right), \quad (1)$$

where $p(x, y)$ indicates the joint probability distribution, $p(x)$ and $p(y)$ represent the marginal probability distributions.

We assume that x and y are vectors with d dimension, we use the 1D histogram to approximate the marginal distributions. First, we divide the feature field into evenly spaced intervals, then count the frequency of each feature value x_i falling within different intervals $[x_j, x_{j+1})$ to obtain $p(x)$:

$$p(x)_j = \frac{1}{d} \sum_{i=1}^{d} \delta(x_i \in [x_j, x_{j+1})), \quad (2)$$

where $\delta(\cdot)$ is an indicator function. Same as $p(y)$. Then, we use the 2D histogram to approximate the joint probability distribution $p(x, y)$. We divide the feature fields into a 2D grid, then we count the frequency of each feature pair (x_i, y_i) falling within different cells to obtain $p(x, y)$:

$$p(x,y)_{j,k} = \frac{1}{d} \sum_{i=1}^{d} \delta(x_i \in [x_j, x_{j+1})) \times \delta(y_i \in [y_j, y_{j+1})). \quad (3)$$

Semantic Consistency Calculation. In practice, slight disparities exist in the similarity of clean subsets, and their labels may not exactly equal 1. When estimating soft correspondence labels, we prioritize variations in semantic consistency. Clean image-text pairs generally show balanced contributions from both modalities, avoiding excessive dominance by either.

First, we select a pair of samples (I_a, T_a) as anchor points in each batch, similar to previous studies, by choosing the image-text pair with the minimum loss within the batch. These anchor points are considered representative of clean samples. Subsequently, for the given sample (I_b, T_b), we perform calculations from three aspects to evaluate its consistency and noise characteristics.

- Change rate of MI between image-text pairs:

$$R_P = \frac{|MI\left(I_a, T_a\right) - MI\left(I_b, T_b\right)|}{MI\left(I_a, T_a\right)}. \tag{4}$$

- Change rate of MI between texts:

$$R_T = \frac{|MI\left(I_a, T_a\right) - MI\left(I_a, T_b\right)|}{MI\left(I_a, T_a\right)}. \tag{5}$$

- Change rate of MI between images:

$$R_I = \frac{|MI\left(I_a, T_a\right) - MI\left(I_b, T_a\right)|}{MI\left(I_a, T_a\right)}. \tag{6}$$

Soft Label Estimation. By calculating soft label estimation, we derive a standardized measure of the variations in consistency between given pairs and anchor pairs. Specifically, if the change rate between image-text pairs (R_P) is small, and the change rate between texts (R_T) is nearly equal to that between images (R_I), it indicates that texts and images contribute equally to the semantic information. In such cases, we infer that the new pairs are likely to be clean.

Eventually, we combine the inverse proportional function with the MI change rate to determine the soft correspondence label $y^* \in (0, 1]$ for the given image-text pair:

$$y^* = \frac{1}{1 + (R_P + |R_T - R_I|)}. \tag{7}$$

Tripartite Cooperative Learning Mechanism

We propose a framework named Tri-learning to overcome the limitations of traditional Co-training. The framework consists of three models: $\mathcal{M}_C$ (Coordinator) partitions the original dataset into a clean set and a noisy set, $\mathcal{M}_A$ (Assistant) selects low-loss samples and combines them with the clean samples to train the $\mathcal{M}_M$ (Master). $\mathcal{M}_M$ further refines $\mathcal{M}_A$, while $\mathcal{M}_C$ iteratively adjusts the partitioning scheme for the next round based on the clean samples, which are fed back by the Assistant model from noisy samples. The detailed training process is illustrated in Fig. 2 and is described as follows:

Step 1: Early Sample Division. For $\mathcal{M}_C$, we adopt the same network structure as $\mathcal{M}_M$ and $\mathcal{M}_A$, with only different initialization parameters. With the assistance of GMM, $\mathcal{M}_C$ divides the original dataset into clean samples $\mathcal{D}_c^M$ with losses less than threshold δ, and noisy samples $\mathcal{D}_n^A$ with losses larger than δ .

Step 2: Sample Re-Division. Through step 1, we obtain two separate data streams, one clean and the other noisy. For the clean data stream $\mathcal{D}_c^M$, we further mine the samples ${\mathcal{D}'}_c^M$ with losses less than δ by the Master model and GMM. It ensures that the training samples of $\mathcal{M}_A$ are as clean as possible. For the noisy data stream $\mathcal{D}_n^A$, we use the Assistant model and GMM to select relatively clean data ${\mathcal{D}'}_c^A$ with losses less than δ, to provide diverse data for training $\mathcal{M}_M$ and $\mathcal{M}_C$.

Step 3: Label Rectification and Network Training.

We first apply the **SIVC** to rectify noisy labels within each training batch. Then, we train $\mathcal{M}_A$ using the cleaned dataset ${\mathcal{D}'}_c^M$ from step 2, equipping $\mathcal{M}_A$ with the ability to extract clean samples from the noisy dataset $\mathcal{D}_n^A$. Similarly, we optimize $\mathcal{M}_M$ using a combination of ${\mathcal{D}'}_c^A$ and $\mathcal{D}_c^M$, enabling it to gain valuable insights from diverse samples while maintaining robustness and generalization. To prevent error propagation during data filtering, $\mathcal{M}_C$ is iteratively optimized using ${\mathcal{D}'}_c^A$, as $\mathcal{D}_n^A$ may contain clean samples. This allows $\mathcal{M}_C$ to detect potential errors and adjust the training and predictions of $\mathcal{M}_A$ and $\mathcal{M}_M$.

We repeat Step 1, 2, and 3 for a specific number of iterations, then use the $\mathcal{M}_A$ and $\mathcal{M}_M$ models as adaptation models for evaluation after the training is completed. A different training paradigm from regular Co-Teaching is adopted, which utilizes a different mechanism to filter and acquire low-loss samples. It enables the network to counter noisy correspondence from different perspectives and improves model robustness. Moreover, this approach prevents assimilation between the two models in Co-Teaching, avoiding negative impacts on prediction effectiveness.

Distribution-Adaptive Soft Margin Loss

We propose DASM loss, which adjusts the soft margin based on the similarity between sample pairs and their deviation from the center of the clean distribution. Specifically, for an image-text pair (I', T') in a mini-batch, we compute the distance from it to the center of the clean sample loss distribution $d = |L - L_{clean}|$, and then calculate the loss $\mathcal{L}_{DASM}$:

$$\begin{aligned}\mathcal{L}_{DASM}(I_i, T_i) = &\left[\tilde{a}_i - S\left(I_i, T_i\right) + S\left(I_i, \hat{T}_n\right)\right]_+ \\ &+ \left[\tilde{a}_i - S\left(I_i, T_i\right) + S\left(\hat{I}_n, T_i\right)\right]_+ .\end{aligned} \tag{8}$$

$$\tilde{a}_i = (2 + tanh(-d))\frac{m^{y_i^*} - 1}{m - 1}\alpha, \tag{9}$$

where $\hat{T}_n = argmax_{T_j \neq T_i} S(I_i, T_j)$ and $\hat{I}_n = argmax_{I_j \neq I_i} S(I_j, T_i)$ represent the hard negative text and image that are most similar to the aligned image-text pair (I_i, T_i) within a batch. As iterations progress, DASM gradually redirects samples that significantly deviate from the clean distribution center, strengthening the boundary between clean and noisy distributions. By slowing down the boundary growth deliberately, it avoids overly penalizing misclassification similarity scores, thus maintaining model balance and flexibility.

Experiments

Datasets

We utilize the following three widely used multi-modal datasets to evaluate our method:

Flickr30K This dataset contains 31,000 images collected from Flickr, with each image paired with five textual descriptions providing detailed annotations of the content. It covers a variety of scenes, objects, and activities, making it a common benchmark for cross-modal retrieval tasks. In our experiments, we use 29,000 images for training, 1,000 for validation, and 1,000 for testing to evaluate the model's performance.

Noise	Methods	Flick30K							MSCOCO						
		Image-Text			Text-Image			Rsum	Image-Text			Text-Image			Rsum
		R@1	R@5	R@10	R@1	R@5	R@10		R@1	R@5	R@10	R@1	R@5	R@10	
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
	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*	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
	MSCN*	77.4	94.9	97.6	59.6	83.2	89.2	501.9	78.1	97.2	98.8	64.3	90.4	95.8	524.6
	BiCro*	78.1	94.4	97.5	60.4	**84.4**	89.9	504.7	78.8	96.1	98.6	63.7	90.3	95.7	523.2
	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
	TSVC(Ours)	**79.6**	**95.7**	**98.6**	**60.9**	84.1	**90.9**	**509.8**	**79.9**	**97.5**	98.6	**65.3**	**91.8**	**97.3**	**530.4**
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
	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*	75.3	92.1	95.2	56.2	80.6	87.4	486.8	76.5	95.0	98.2	60.7	88.5	95.0	513.9
	MSCN*	74.4	94.4	96.9	57.2	81.7	**87.6**	492.2	74.8	94.9	98.0	60.3	88.5	94.4	510.9
	BiCro*	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
	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
	TSVC(Ours)	**77.7**	**95.3**	**98.3**	**58.8**	**83.1**	87.1	**500.3**	**79.0**	**97.3**	98.6	**65.5**	**91.3**	**96.2**	**527.9**
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
	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*	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
	MSCN*	70.4	91.0	94.9	53.4	77.8	84.1	471.6	74.4	95.1	97.9	59.2	87.1	92.8	506.5
	BiCro*	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
	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
	TSVC(Ours)	**73.2**	**92.0**	**95.1**	**54.8**	78.5	**86.2**	**479.8**	**77.4**	**96.8**	**99.5**	**61.7**	**89.0**	**94.9**	**519.3**

Table 1: Image-Text Retrieval on Flickr30K and MS-COCO 1K datasets under different noise ratios. * indicates the noise robust method. The best and sub-optimal indicators are represented in **bold** and underline respectively.

Methods	CC152K						
	Image-Text			Text-Image			Rsum
	R@1	R@5	R@10	R@1	R@5	R@10	
SCAN	30.5	55.3	65.3	26.9	53.0	64.7	295.7
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
MSCN*	40.1	65.7	76.6	40.6	67.4	76.3	366.7
BiCro*	40.8	67.2	76.1	42.1	67.6	76.4	370.2
ESC*	42.8	67.3	**76.9**	44.8	**68.2**	75.9	375.9
TSVC(Ours)	**44.7**	**69.4**	76.5	**45.1**	67.8	**76.6**	**380.1**

Table 2: Image-Text Retrieval on CC152K. * indicates the noise robust method. The best and sub-optimal indicators are represented in **bold** and underline respectively.

MSCOCO This dataset consists of 123,287 images, each annotated with five textual captions describing diverse visual content, including daily scenes and objects. It serves as a critical benchmark for multimodal tasks. For our experiments, 113,287 images are used for training, 5,000 for validation, and 5,000 for testing, ensuring a comprehensive performance evaluation.

Conceptual Captions. A large scale dataset mainly collected from the Internet, which comprises 3.3 million pictures, each accompanied by a corresponding caption. Around 3%-20% of the image-text pairs in the dataset exhibit inconsistencies (Sharma et al. 2018). Similar to previous studies (Huang et al. 2021), we use the CC152K subset from Conceptual Captions in our experiments, which contains 150k images for training, and 1000 images each for validation and testing.

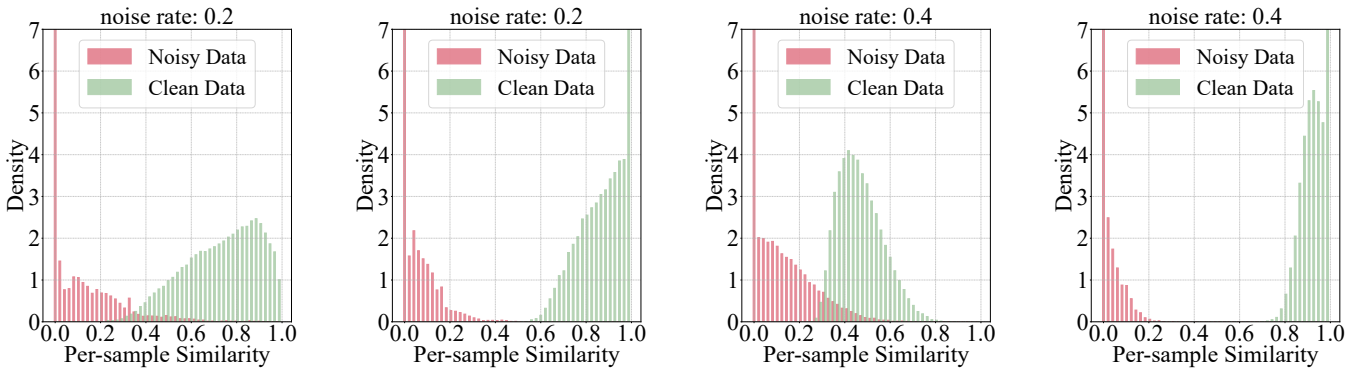
Evaluation Metrics

We employ Recall@K (R@K) to measure the retrieval performance, which quantifies the accuracy of identifying relevant items within the top K results. Additionally, to provide a more comprehensive analysis, we report R@1, R@5, and R@10 for both image-to-text and text-to-image retrieval tasks, and summarize these metrics as Rsum for a comprehensive performance evaluation.

Main Results

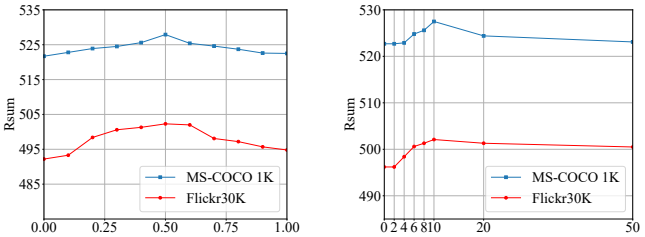
In this section, we validate performance of the TSVC network on the aforementioned three datasets. Specifically, we select 4 methods without noise learning, i.e., SCAN (Lee et al. 2018), IMRAM (Chen et al. 2020a), SAF and SGR (Diao et al. 2021), to illustrate the impact brought by noise. We select another 4 methods with noise learning, i.e., NCR (Huang et al. 2021), MSCN (Han et al. 2023), BiCro (Yang et al. 2023), and ESC (Yang et al. 2024)), to exhibit superior performance of TSVC.

Experiments on Simulated Noise. To evaluate performance under noise, we introduce simulated noise by randomly shuffling image-caption pairs in Flickr30K and MS-COCO,



(a) W/o DASM (b) W/ DASM (c) W/o DASM (d) W/ DASM

Figure 3: Distribution change on Flickr30K with noise ratios of 20% and 40%.



(a) δ (b) m

Figure 4: Analysis of hyper-parameters δ and m on Flickr30K with 40% noise. **Left:** δ is the filtering threshold. **Right:** m is the parameter of the soft-margin formula.

with noise ratios of 20%, 40%, and 60%. Results in Table 1 show that our proposed TSVC method consistently outperforms state-of-the-art methods. Compared to the best-performing baseline **ESC** (Yang et al. 2024), TSVC achieves improvements of 6.5, 10.7, and 4.8 on Flickr30K, and 3.6, 4.0, and 4.3 on MSCOCO-1k under the respective noise ratios.

Experiments on Real-world Noise. We evaluate our method on the real-world noisy scenes in the CC152K dataset. It includes mismatched pairs with unknown parts and serves as a challenging benchmark dataset. Detailed experimental results are listed in Table 2. From the table, we can observe that even under the condition of real noise, TSVC still demonstrates competitive performance. Specifically, TSVC outperforms the state-of-the-art baseline method ESC by 4.2 for overall retrieval performance. It indicates the effectiveness of TSVC in improving retrieval accuracy.

Ablation Study

We conduct an ablation study on three datasets to investigate the impact of individual components on performance. Due to space limit, we only present results on the Flickr30K dataset with 40% noise impact. The results are listed in Table 3. Similar trends are observed across the other datasets. We remove three key components of **TSVC**, i.e., SIVC, DASM, and Tri-learning, to observe their impacts.

Configuration	Image-Text			Text-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
TSVC	**77.7**	**95.3**	**98.3**	**58.8**	**83.1**	**87.1**
w/o SIVC	74.8	93.6	96.7	55.4	80.2	84.3
w/o DASM	75.6	94.1	97.0	57.2	81.5	85.8
w/o Tri-Learning	75.0	94.3	97.7	58.0	81.1	85.7

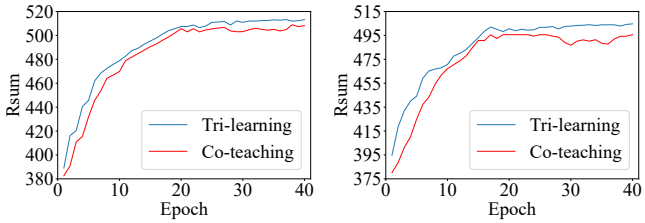
Table 3: Ablation studies on Flickr30K with 40% noise ratio.

To verify the effectiveness of SIVC, we employ hard labels for comparison, where clean samples are labeled as 1 and noisy samples as 0. To test the effectiveness of the distance metric DASM, we replace it with a regular soft-margin triplet loss function. To verify the effectiveness of the Tri-learning scheme, we compare it with the Co-teaching approach.

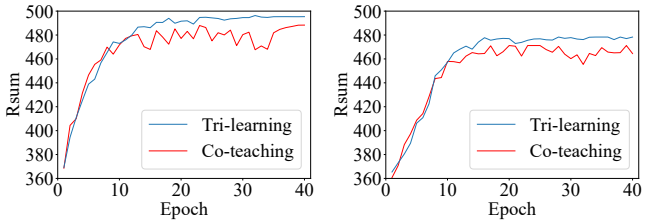
Results are presented in Table 3. From the table, we can make three observations. (1) Employing SIVC for soft label estimation outperforms relying solely on hard labels. It suggests that soft labels provide better measurement of matching degree between data pairs and help the models learn useful information. (2) The use of distance metric loss function is significantly superior to using ordinary triplet loss. Meanwhile, we visualize the sample similarity distribution before and after using DASM. As shown in Fig. 3, the improvement in data partitioning is significant. (3) The Tri-learning architecture demonstrates superior performance compared to the Co-training paradigm. This is because Tri-learning enables each model within the architecture to fulfill its designated role and collaborate with others. Thus, it effectively mitigates model homogenization and enhances network performance. (4) The complete TSVC model achieves the best overall performance, demonstrating the importance of all three components in terms of noise correspondence.

Analysis of the Tri-learning Method

Figure 5 illustrates the test Rsum values during the training process under different noise types. We can observe that when the noise ratio is 0.2, the learning curves of both



(a) noise ratio: 0 (Left) and 0.2 (Right).



(b) noise ratio: 0.4 (Left) and 0.6 (Right).

Figure 5: Results on Flickr30K dataset.

Co-teaching and Tri-learning are relatively stable. Nevertheless, as the noise ratio increases, the learning curve of Co-teaching exhibits greater fluctuations. Because of homogenization of the two models, the final predictive performance of Co-teaching is limited. In contrast, the learning curve of Tri-learning tends to fit the data more easily, exhibiting greater stability. This indicates that Tri-learning is more effective in handling higher levels of noise, achieving more robust performance in noisy environment.

Notably, as shown in Fig. 5 (b), under high noise ratios, Tri-learning initially performs worse than Co-teaching during training, because it may misclassify noisy samples as clean samples in the Re-Division stage. Nevertheless, as it is fully trained, its performance demonstrates greater stability than that of Co-teaching.

Hyperparameters Analysis

The hyperparameters δ and m represent the threshold used for dividing noisy samples and the parameter controlling the soft margin in DASM, respectively. These parameters play a important role in balancing noise filtering and optimizing margin flexibility. We conduct experiments on Flickr30K and MSCOCO datasets with a noise ratio of 40%. As depicted in Fig. 4, the value of Rsum increases continuously with an increase in δ, reaching its peak at $\delta = 0.5$ (between 0.5 to 0.6 for MSCOCO), and then gradually decreases. A larger mismatch threshold δ might classify weakly labeled samples as mislabeled ones, thus impairing generalization ability. The optimal value for m is 10, which significantly influences the performance of the model by affecting the size of the soft-margin. Therefore, it should be considered comprehensively during the model optimization process to ensure the best performance.

Visualization

To further illustrate effectiveness of TSVC, we present the retrieval process using textual and visual queries, as depicted in Fig. 6 and Fig. 7. The figures demonstrate how TSVC retrieves the top 3 similar images or texts. Even when the ground truth image or text does not rank first, the selected items still capture elements relevant to the query, maintaining consistency in scene or topic. This highlights TSVC's capability to capture the essential features with subtle comprehension across different modalities.

Figure 6: Image-To-Text matching results on Flickr30K.

Figure 7: Text-to-Image matching results on Flickr30K.

Conclusion

In this work, we proposed a Tripartite Learning with Semantic Variation Consistency (TSVC) framework to address noisy correspondences for cross-modal data. TSVC leverages information variation between clean image-text pairs to estimate soft correspondence noise labels. By distributing tasks among three models, TSVC effectively resolves the challenge of limited performance improvement due to model prediction homogenization. Extensive experiments on three datasets validate the effectiveness of our method.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62076035) and the SMP-Zhipu.AI Large Model Cross-Disciplinary Fund. Special thanks to the State Grid Hebei Electric Power Company (Project No. A2024160) and the China Unicom Beijing Branch Digitalization Department AI Capability Center (Project No. A2024054) for their support.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, 6077–6086. Computer Vision Foundation / IEEE Computer Society.

Chen, H.; Ding, G.; Liu, X.; Lin, Z.; Liu, J.; and Han, J. 2020a. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12655–12663.

Chen, J.; Hu, H.; Wu, H.; Jiang, Y.; and Wang, C. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15789–15798.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020b. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Deng, C.; Yang, E.; Liu, T.; and Tao, D. 2019. Two-stream deep hashing with class-specific centers for supervised image search. *IEEE transactions on neural networks and learning systems*, 31(6): 2189–2201.

Diao, H.; Zhang, Y.; Ma, L.; and Lu, H. 2021. Similarity reasoning and filtration for image-text matching. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 1218–1226.

Faghri, F.; Fleet, D. J.; Kiros, J. R.; and Fidler, S. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. *Advances in neural information processing systems*, 31.

Han, H.; Miao, K.; Zheng, Q.; and Luo, M. 2023. Noisy correspondence learning with meta similarity correction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7517–7526.

Han, H.; Zheng, Q.; Dai, G.; Luo, M.; and Wang, J. 2024. Learning to Rematch Mismatched Pairs for Robust Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26679–26688.

Hu, P.; Peng, X.; Zhu, H.; Zhen, L.; and Lin, J. 2021. Learning cross-modal retrieval with noisy labels. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5403–5413.

Huang, F.; Zhang, L.; Fu, X.; and Song, S. 2024a. Dynamic weighted combiner for mixed-modal image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 2303–2311.

Huang, H.; Nie, Z.; Wang, Z.; and Shang, Z. 2024b. Cross-Modal and Uni-Modal Soft-Label Alignment for Image-Text Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 18298–18306.

Huang, Z.; Niu, G.; Liu, X.; Ding, W.; Xiao, X.; Wu, H.; and Peng, X. 2021. Learning with noisy correspondence for cross-modal matching. *Advances in Neural Information Processing Systems*, 34: 29406–29419.

Iscen, A.; Valmadre, J.; Arnab, A.; and Schmid, C. 2022. Learning with neighbor consistency for noisy labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4672–4681.

Lee, K.-H.; Chen, X.; Hua, G.; Hu, H.; and He, X. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, 201–216.

Li, S.; Tao, Z.; Li, K.; and Fu, Y. 2019. Visual to Text: Survey of Image and Video Captioning. *IEEE Trans. Emerg. Top. Comput. Intell.*, 3(4): 297–312.

Li, Y.; Huang, H.; Xu, J.; and Huang, S.-L. 2024. NAC: Mitigating Noisy Correspondence in Cross-Modal Matching Via Neighbor Auxiliary Corrector. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6815–6819. IEEE.

Ma, X.; Yang, M.; Li, Y.; Hu, P.; Lv, J.; and Peng, X. 2024. Cross-modal Retrieval with Noisy Correspondence via Consistency Refining and Mining. *IEEE Transactions on Image Processing*.

Qin, Y.; Peng, D.; Peng, X.; Wang, X.; and Hu, P. 2022. Deep evidential learning with noisy correspondence for cross-modal retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4948–4956.

Qin, Y.; Sun, Y.; Peng, D.; Zhou, J. T.; Peng, X.; and Hu, P. 2024. Cross-modal active complementary learning with self-refining correspondence. *Advances in Neural Information Processing Systems*, 36.

Shannon, C. E. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3): 379–423.

Sharma, P.; Ding, N.; Goodman, S.; and Soricut, R. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2556–2565.

Shi, H.; Liu, M.; Mu, X.; Song, X.; Hu, Y.; and Nie, L. 2024. Breaking Through the Noisy Correspondence: A Robust Model for Image-Text Matching. *ACM Transactions on Information Systems*.

Wang, H.; Zhan, Y.; Liu, L.; Ding, L.; and Yu, J. 2024. Balanced Similarity with Auxiliary Prompts: Towards Alleviating Text-to-Image Retrieval Bias for CLIP in Zero-shot Learning. *arXiv preprint arXiv:2402.18400*.

Wang, S.; Wang, R.; Yao, Z.; Shan, S.; and Chen, X. 2020. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 1508–1517.

Wu, Y.; Wang, S.; Song, G.; and Huang, Q. 2019. Learning fragment self-attention embeddings for image-text matching. In *Proceedings of the 27th ACM international conference on multimedia*, 2088–2096.

Yan, J.; Luo, L.; Deng, C.; and Huang, H. 2023. Adaptive hierarchical similarity metric learning with noisy labels. *IEEE Transactions on Image Processing*, 32: 1245–1256.

Yan, J.; Luo, L.; Xu, C.; Deng, C.; and Huang, H. 2022. Noise is also useful: Negative correlation-steered latent contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 31–40.

Yang, S.; Li, Q.; Li, W.; Li, X.; and Liu, A.-A. 2022. Dual-level representation enhancement on characteristic and context for image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11): 8037–8050.

Yang, S.; Xu, Z.; Wang, K.; You, Y.; Yao, H.; Liu, T.; and Xu, M. 2023. Bicro: Noisy correspondence rectification for multi-modality data via bi-directional cross-modal similarity consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19883–19892.

Yang, Y.; Wang, L.; Yang, E.; and Deng, C. 2024. Robust Noisy Correspondence Learning with Equivariant Similarity Consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17700–17709.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2021. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3): 107–115.

Zhang, K.; Mao, Z.; Wang, Q.; and Zhang, Y. 2022. Negative-aware attention framework for image-text matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15661–15670.

Zhang, X.; Li, H.; and Ye, M. 2024. Negative Pre-aware for Noisy Cross-Modal Matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7341–7349.

Zhao, X.; Li, D.; Zhong, Y.; Hu, B.; Chen, Y.; Hu, B.; and Zhang, M. 2024a. SEER: Self-Aligned Evidence Extraction for Retrieval-Augmented Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, 3027–3041.

Zhao, Z.; Chen, M.; Dai, T.; Yao, J.; Han, B.; Zhang, Y.; and Wang, Y. 2024b. Mitigating Noisy Correspondence by Geometrical Structure Consistency Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27381–27390.

Zheng, G.; Awadallah, A. H.; and Dumais, S. 2021. Meta label correction for noisy label learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, 11053–11061.

Zhong, Y.; Wu, X.; Zhang, L.; Yang, C.; and Jiang, T. 2024. Causal-IQA: Towards the Generalization of Image Quality Assessment Based on Causal Inference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*.