DUAL-MIX FOR CROSS-MODAL RETRIEVAL WITH NOISY LABELS

Feng Ding *Xiu Liu* *Xinyi Wang* *Fangming Zhong**

School of Software Technology, Dalian University of Technology, Dalian, Liaoning 116620, China
fmzhong@dlut.edu.cn

ABSTRACT

Cross-modal retrieval with deep neural networks heavily relies on accurate annotation. However, existing methods may easily suffer from the scarcity and validity of annotations due to the expensive cost of manual labeling. In addition, it is inevitable that noisy labels are imposed during labeling. To this end, it is worthwhile to explore the potential of noisy labels in cross-modal retrieval. In this work, we propose a novel framework entitled Dual-Mix for Cross-Modal Retrieval with noisy labels (DMCM). It consists of two components, which are mixing the robust loss functions and mixing augmentation for noisy samples. In the first mixing stage, the normalized generalized cross entropy and mean absolute error are combined to boost each other. Then, after separating clean and noisy samples by Beta Mixture Model, we mix these samples via augmentation to further address the scarcity of labeled samples. Extensive experiments demonstrate the significant superiority of our DMCM.

***Index Terms*—** cross-modal retrieval, noisy labels, mixing, contrastive learning

1. INTRODUCTION

With the explosion in multimedia data on the Internet, cross-modal retrieval has received increasing attention, such as image-text [1, 2], and text-audio retrieval [3].

Many supervised methods have been proposed for cross-modal retrieval [4]. However, these methods heavily depend on the quality of the labels. Unfortunately, obtaining large-scale high-quality annotated labels through manual expert-labeling is extremely expensive. In addition, it inevitably introduces numerous mistakes or label noise. According to [5–7], Deep Neural Networks (DNN) can easily overfit to noisy labels within training. Although many unsupervised methods are proposed to avoid the interference of noisy labels, their performances are significantly inferior to supervised methods. Therefore, how to train a cross-modal retrieval model that is robust to noisy labels is crucial for improving the applicability and efficiency, which has not been well studied yet.

In the unimodal scenario, numerous studies [8–10] have been conducted to develop robust models which are capable of handling noisy labels and achieving promising performance, such as correction methods [11, 12] and Co-teaching [6]. However, in multimodal scenarios, the noisy labels can bring confusion in the connections between different modalities, leading to difficulties in bridging the heterogeneous gap.

Consequently, combating the impact of noisy labels and mitigating cross-modal semantic gap simultaneously become more challenging and complicated. Only a few studies have been conducted, which can be roughly categorized into two groups: robust algorithms and noise detection methods.

Robust algorithms are developed to mitigate the sensitivity to noisy labels, which involve constructing robust networks, employing robust loss functions, and applying robust regularization techniques. For instance, Xu et al. [13] employed an early learning regularization to punish overfitting. Recently, many studies have adopted the small-loss criterion [10, 14], which suggests that samples with smaller loss values are more likely to have clean labels. For instance, Multimodal Robust Learning (MRL) [15] introduces a modified cross-entropy loss, which assigns higher weights to clean samples with a small loss, aiming to guide the DNN to prioritize learning from clean labels. However, noise labels still remain in the training data, leaving a memorization effect on DNN, which adversely degrades the retrieval performance. As can be seen in Fig.1 MRL [15] tends to overfitting during training.

Noise detection methods aim at identifying noisy samples and devising strategies to alleviate the influence of noise samples, such as re-labeling with pseudo labels, and treating them as unlabeled samples in a semi-supervised manner. For example, Okamura et al. presented Label Correction based on Network Prediction (LCN) [16] to annotate the noisy samples with predicted labels. While Yang et al. proposed a Cross-Modal Mutual Quantization (CMMQ) [17] that exclusively uses clean samples for training, resulting in a significant reduction in sample size and degradation in cross-modal retrieval performance.

In this paper, by integrating the advantages of robust algorithms and noise detection methods, we propose a novel Dual-Mix for Cross-Modal Retrieval with Noisy Labels (DMCM).

*Corresponding author. This work is supported by the National Natural Science Foundation of China (62006035), Dalian Science and Technology Innovation Foundation (2023JJ13SN065), and the Fundamental Research Funds for the Central Universities (DUT22RC(3)011).

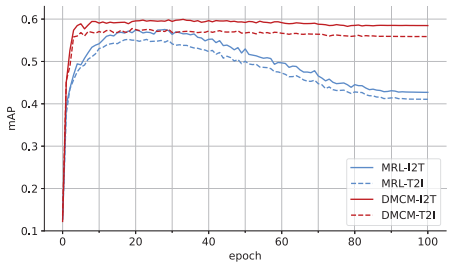

Fig. 1. Retrieval results by MRL and DMCM for the Wikipedia dataset under symmetric noise rates of 0.6.

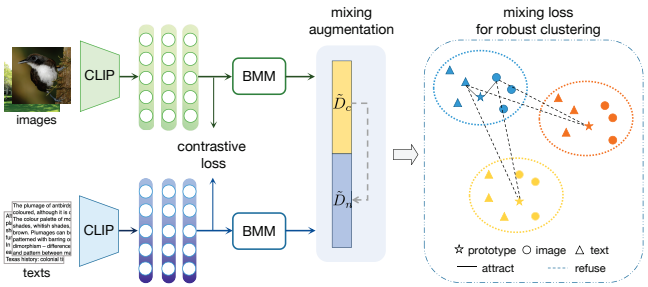

Fig. 2. Overall architecture of the proposed DMCM.

The framework of our method is shown in Fig. 2. Firstly, to narrow the heterogeneous gap, we take the pre-trained CLIP as backbone and stack a 3-layer fully connected network to project each modality into a shared embedding space. Then, in order to combat noisy labels, we propose to mix two robust loss functions, which are Normalized Generalized Cross Entropy (NGCE) [18] and Mean Absolute Error (MAE). Our new robust clustering loss ensures the robust learning of multimodal consistency. Moreover, we further mix clean and noisy samples by a data augmentation method. In doing so, our DMCM can reduce the influence of noisy labels during training and increase the training sample size simultaneously. Specifically, we discern clean samples from noisy ones by modeling the per-sample loss distribution of the dataset through a Beta Mixture Model (BMM). In addition, we employ multimodal contrastive learning to further improve the discrimination of comment embeddings. Extensive experiments are carried out on three benchmark datasets for cross-modal retrieval, our method demonstrates significant superiority over the state-of-the-art methods.

The contributions of this work are summarized as follows:

- A novel framework DMCM for cross-modal retrieval with noisy labels is proposed, where noise detection is incorporated to robust clustering loss and their advantages are seamlessly integrated.
- Dual mixing components are proposed, which are mixing loss for robust clustering and mixing augmentation for noisy samples. To the best of our knowledge, ours is the first attempt towards this end for cross-modal retrieval with noisy labels.

2. APPROACH

2.1. Preliminaries

Given a K-class dataset with noisy labels as $\mathcal{D} = \{\mathcal{M}_i\}_{i=1}^m$, where $\mathcal{M}_i = \left\{(\mathbf{x}_j^i, \mathbf{y}_j^i)\right\}_{j=1}^N$ is the i-th modality, $\mathbf{x}_j^i \in \mathbb{R}^d$ is the j-th sample from the i-th modality, $\mathbf{y}_j^i \in \{0,1\}^K$ is the corresponding one-hot label (possibly incorrect) for $\mathbf{x}_j^i$.

For cross-modal retrieval, multi-modal inputs are usually projected into a common semantic space through modality-specific functions $\{f_i : X_i \mapsto \mathbb{Z}\}_{i=1}^m$, f_i is the function for the i-th modality, which can be instantiated with a DNN parameterized with Θ_i, which can be formulated as:

$$\mathbf{z}_j^i = f_i(\mathbf{x}_j^i, \Theta_i) \in \mathbb{R}^c \tag{1}$$

where c indicates the dimension of common space.

Our method mainly consists of two mixing stages, which are described in detail as follows.

2.2. Mixing Loss for Robust Clustering

To learn discriminative common embeddings of samples from different modalities, we first leverage unified prototypes $\mathbf{C} = \{\mathbf{c}_1, \cdots, \mathbf{c}_K\}$ as anchors in the embedding space, where $\mathbf{c}_k$ represents the k-th class proxy. Then, the probability that a sample $\mathbf{x}_j^i$ belongs to the k-th class can be estimated by:

$$p\left(k \mid \mathbf{x}_j^i\right) = \frac{\exp\left(\frac{1}{\tau_1}\mathbf{c}_k^T \mathbf{z}_j^i\right)}{\sum_{t=1}^K \exp\left(\frac{1}{\tau_1}\mathbf{c}_t^T \mathbf{z}_j^i\right)} \tag{2}$$

where τ_1 is a temperature parameter.

Subsequently, we maximize the similarity between embeddings of all samples and their corresponding category prototypes. Thus, samples from different modalities are aligned to the anchors to bridge the semantic gap.

Usually, the modality-specific function f can be learned by minimizing robust loss functions such as GCE [19], FL [20], and RCE [21] in a dataset with noisy labels. Unfortunately, only using simple robust loss is still not enough for an excellent f. As mentioned in [18], several robust loss functions suffer from a problem of underfitting.

To address this challenge, we mix two different robust loss functions following [18]. Concretely, an "active" loss is used to only maximize the probability of being in the ground truth class, and a "passive" loss can further minimize the probabilities of being in other classes. In our work, Normalized Generalized Cross Entropy (NGCE) and Mean Absolute Error (MAE) are selected as the active and passive losses, respectively. This is the first mix of our model resulting in the

robust clustering loss function in Eq (3), where $\mathbf{y}_j^i(k)$ can be interpreted as the probability associated with the k-th prototype for the sample $\mathbf{x}_j^i$.

$$\begin{aligned} \mathcal{L}_m &= NGCE + MAE \\ NGCE &= \frac{1}{N}\sum_{i=1}^{m} \frac{-\sum_{j=1}^{N}\sum_{k=1}^{K}\mathbf{y}_j^i(k)\frac{1-p(k|\mathbf{x}_j^i)^\rho}{\rho}}{-\sum_{j=1}^{N}\sum_{c=1}^{K}\sum_{k=1}^{K}\mathbf{y}_j^i(c)\frac{1-p(k|\mathbf{x}_j^i)^\rho}{\rho}} \\ MAE &= \frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{N}\sum_{k=1}^{K}|p(k|\mathbf{x}_j^i) - \mathbf{y}_j^i(k)| \end{aligned} \tag{3}$$

2.3. Mixing Augmentation for Noisy Samples

In addition to the mixing loss for robust clustering, we further propose a mixing augmentation method for noisy samples, which is the second mix of our DMCM. After considerable investigations of existing works, we have a conclusion that discarding noisy samples directly leads to insufficient training samples while reserving noisy samples may result in overfitting. Motivated by such a fact, we attempt to augment the noisy samples by mixing clean samples. To the best of our knowledge, ours is the first attempt towards this end for cross-modal retrieval with noisy labels.

Therefore, it is necessary to separate noisy samples from clean samples. According to the small-loss criterion [10, 14], samples with smaller loss values are more likely to have clean labels. Thus, we can first compute the per-sample loss in Eq. (3). Then, we fit the per-sample loss of all training data from different modalities by using a Beta Mixture Model (BMM) to model the distribution of clean and noisy samples.

For each sample $\mathbf{x}_j^i$, its clean probability $\mathbf{w}_j^i$ is the posterior probability $p(g|\ell_i)$, where g is the Beta component with smaller loss, ℓ_i is the per-sample loss in Eq. (3). By introducing a threshold parameter denoted as δ, the training dataset is divided into a clean set $\tilde{D}_c$ and a noisy set $\tilde{D}_n$ as follows,

$$\tilde{D}_c = \{(\mathbf{x}_j^i, \mathbf{q}_j^i) \mid \mathbf{w}_j^i > \delta_i, \forall(\mathbf{x}_j^i, \mathbf{y}_j^i) \in \tilde{D}\} \tag{4}$$

$$\tilde{D}_n = \{(\mathbf{x}_j^i, \mathbf{q}_j^i) \mid \mathbf{w}_j^i \leq \delta_i, \forall(\mathbf{x}_j^i, \mathbf{y}_j^i) \in \tilde{D}\} \tag{5}$$

where $\tilde{D}$ indicates the mini-batch data and $\mathbf{q}_j^i$ is the one-hot vectors with floating-point values.

Drawing inspiration from the MixMatch [22], we augment the noisy samples by mixing them with randomly selected clean samples. Specifically, let $(\mathbf{x}_1, \mathbf{q}_1) \in \tilde{D}_n$ represent a noisy sample, and $(\mathbf{x}_2, \mathbf{q}_2) \in \tilde{D}_c$ denote a clean sample from $\tilde{D}_c$. The mixed sample $(\mathbf{x}', \mathbf{q}')$ is computed as follows:

$$\begin{aligned} \mathbf{x}' &= \lambda\mathbf{x}_1 + (1-\lambda)\,\mathbf{x}_2, \\ \mathbf{q}' &= \lambda\mathbf{q}_1 + (1-\lambda)\,\mathbf{q}_2. \end{aligned} \tag{6}$$

where λ is the tradeoff parameter. After mixing, noisy data is defined as $\tilde{D}'_n = \{(\mathbf{x}'^i_j, \mathbf{q}'^i_j)\}$. In the training stage, the clean data $\tilde{D}_c$ and the noisy data $\tilde{D}'_n$ are combined to train the robust clustering model by minimizing the loss in Eq. (3).

2.4. Multimodal Contrastive Learning

In the context of multimodal data, the inherent potential of contrastive learning can be harnessed to effectively enhance their mutual information. Building upon this principle, a multimodal contrastive loss is further constructed to mitigate semantic gap across different modalities [15]:

$$\mathcal{L}_c = -\frac{1}{N}\sum_{i=1}^{m}\sum_{j=1}^{N}\log\Big(\frac{\sum_{l=1}^{m}\exp\left(\frac{1}{\tau_2}\left(\mathbf{z}_j^l\right)^T\mathbf{z}_j^i\right)}{\sum_{l=1}^{m}\sum_{t=1}^{N}\exp\left(\frac{1}{\tau_2}\left(\mathbf{z}_t^l\right)^T\mathbf{z}_j^i\right)}\Big) \tag{7}$$

where τ_2 is a temperature parameter.

2.5. Overall Objective

Note that we first perform a warm-up process with the loss $\mathcal{L} = \mathcal{L}_m$ to ensure the network achieves initial convergence. After the warmup, the overall loss function is formulated as:

$$\mathcal{L} = \beta\mathcal{L}_m + (1-\beta)\mathcal{L}_c \tag{8}$$

where β is a hyper-parameter. By minimizing the overall loss Eq. (8), the network parameters $\{\Theta_i\}_{i=1}^m$ and prototypes $\mathbf{C}$ can be optimized using stochastic gradient descent.

Table 1. Statistics of three datasets used in our experiments.

Dataset	Training	Testing	Classes
Wikipedia	2157	462	10
Pascal-Sentence	8000	200	20
XmediaNet	32000	4000	200

3. EXPERIMENTS

3.1. Datasets and Features

Three benchmark datasets, i.e., Wikipedia [23], Pascal-Sentence [24], and XmediaNet [25] provided by [1] are used to validate our DMCM. Table 1 presents the details of training and testing. We adopt the pretrained CLIP as the backbones for images and texts on all datasets. Then, two 3-layer fully connected networks are stacked on the backbones respectively for learning the common representation of images and texts.

3.2. Implementation details

In this work, we employ ADAM [26] as our optimizer to train DMCM. The learning rate is 0.0001 and set τ_1 = 1, τ_2 = 1. For Wikipedia, Pascal-Sentences, and XMediaNet, we set the batch sizes as 100, 100, 200, the epochs as 100, 150, 250, and the β as 0.85, 0.7, 0.3. The common space dimension L is set to 512 on three datasets.

We initially warm up the model with 3 epochs for Wikipedia and XMediaNet, and 5 epochs for Pascal-Sentences

Table 2. Performance comparison in terms of mAP under the symmetric noise rates of 0.2, 0.4, 0.6, and 0.8 on three widely-used benchmark datasets. The best score is shown in bold. MRL* employs VGG19 and Doc2Vec as the backbone.

Method	Wikipedia								Pascal Sentences								XMediaNet							
	I2T				T2I				I2T				T2I				I2T				T2I			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
DCCA	0.467	0.467	0.467	0.467	0.453	0.453	0.453	0.453	0.610	0.610	0.610	0.610	0.614	0.614	0.614	0.614	0.430	0.430	0.430	0.430	0.413	0.413	0.413	0.413
DCCAE	0.468	0.468	0.468	0.468	0.455	0.455	0.455	0.455	0.620	0.620	0.620	0.620	0.618	0.618	0.618	0.618	0.443	0.443	0.443	0.443	0.428	0.428	0.428	0.428
SDML	0.576	0.559	0.484	0.307	0.579	0.546	0.476	0.278	0.639	0.591	0.093	0.265	0.646	0.594	0.123	0.290	0.690	0.627	0.471	0.114	0.692	0.624	0.452	0.072
DSCMR	0.599	0.572	0.515	0.304	0.569	0.551	0.501	0.362	0.678	0.615	0.518	0.315	0.680	0.637	0.552	0.381	0.711	0.660	0.554	0.009	0.721	0.673	0.585	0.014
LCN	0.614	0.601	0.594	0.545	0.579	0.571	0.568	0.523	0.705	0.703	0.667	0.636	0.712	0.705	0.675	0.635	0.626	0.627	0.632	0.634	0.631	0.644	0.638	0.641
MRL*	0.514	0.491	0.464	0.435	0.461	0.453	0.421	0.400	**0.724**	**0.719**	0.680	0.640	**0.727**	**0.724**	0.682	0.639	0.625	0.581	0.384	0.334	0.623	0.587	0.408	0.359
MRL	0.598	0.594	0.575	0.528	0.576	0.564	0.560	0.510	0.709	0.691	0.678	0.640	0.709	0.694	0.678	0.635	0.639	0.633	0.641	0.613	0.647	0.635	0.647	0.618
DMCM	**0.624**	**0.621**	**0.601**	**0.572**	**0.592**	**0.593**	**0.578**	**0.550**	0.717	0.715	**0.697**	**0.655**	0.716	0.720	**0.699**	**0.657**	**0.717**	**0.701**	**0.679**	**0.652**	**0.723**	**0.711**	**0.681**	**0.654**

to ensure they achieve initial convergence. We assume that the noise rate r is known. At each training epoch, we select the ratio of $(1-r)$ samples with a higher probability of being clean in Eq. (4) as clean samples. The remained data are regarded as noisy samples. In practice, if the noisy rate r is unknown in advance, it can be inferred by empirical analysis.

3.3. Comparison with the State-of-the-Arts

To validate the effectiveness of DMCM, we evaluate DMCM against several cross-modal retrieval baselines, including general methods (DCCA [23], DCCAE [27], SDML [28] and DSCMR [29])) and methods proposed to combat noise labels (MRL [15], LCN [16]). For fair comparisons, all baselines utilize the same backbones as our DMCM for feature extraction. The mean average precision (mAP) of two cross-modal retrieval tasks i.e., using image queries to retrieve text samples (I2T) and using text queries to retrieve image samples (T2I) are reported in Table 2.

From the results, we can see that DMCM outperforms baseline methods in most cases. Compared to MRL, DMCM can obtain an absolute increase of 3.43 % and 3.33 % in average mAP on three datasets for I2T and T2I. As the ratio of noise labels increases, retrieval performance slowly decreases. Moreover, the average mAP of our DMCM is 6.35% higher than MRL on XMediaNet, indicating that our method has excellent anti-interference ability even with more classes. It also can be seen from Fig. 1 that MRL [15] tends to overfit during training, while our method can produce stable output.

In addition, MRL has remarkable improvements to MRL* on Wikipedia and XmediaNet datasets. Our DMCM only slightly trails behind MRL* on Pascal-Sentence dataset when noise rate is 0.2 and 0.4. This is because the token length of CLIP's text extractor is limited to 70, impeding the feature extraction of long-length texts in this dataset. In summary, the results have demonstrated the effectiveness of our method.

3.4. Ablation Study

Table 3 displays the ablation study results of DMCM. DMCM ($\lambda = 1$) means without mixing augmentation. We can see the DMCM is superior to its three variants, indicating that both mixing components contribute to performance enhancement.

Table 3. Ablation study on Wikipedia.

Method	I2T				T2I			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
DMCM w/o MAE	0.605	0.592	0.582	0.551	0.571	0.568	0.558	0.530
DMCM w/o NGCE	0.620	0.620	0.593	0.560	0.588	0.587	0.571	0.540
DMCM ($\lambda = 1$)	0.615	0.607	0.587	0.549	0.583	0.580	0.567	0.532
DMCM	**0.624**	**0.621**	**0.601**	**0.572**	**0.592**	**0.593**	**0.578**	**0.550**

3.5. Parameter Analysis

Here, we analyze the impact of varied values of λ. By setting λ as 1, DMCM degenerates to a model without the mixing augmentation. When $\lambda = 0$ it transforms to the version that only with clean samples. As can be seen from Fig. 3, a smaller λ achieves better results, which validates that the mixing augmentation with clean samples indeed works.

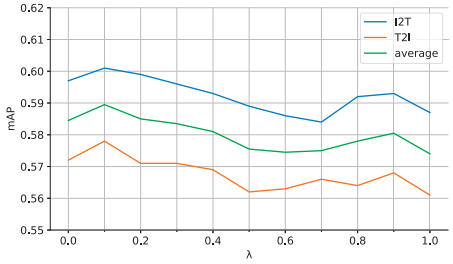

Fig. 3. mAP values under symmetric noise ratio 0.6 on Wikipedia for the two search tasks with different λ.

4. CONCLUSION

In this paper, we have proposed a novel DMCM for cross-modal retrieval with noisy labels. Two mixing components are designed to alleviate the impact of noisy labels. Extensive experiments have indicated the efficacy of DMCM. The future work shall include exploring more robust architectures and handling more types of noises.

5. REFERENCES

[1] Zhixiong Zeng and Wenji Mao, "A comprehensive empirical study of vision-language pre-trained model for supervised cross-modal retrieval," *arXiv preprint arXiv:2201.02772*, 2022.

[2] Georgii Mikriukov, Mahdyar Ravanbakhsh, and Begüm Demir, "Unsupervised contrastive hashing for cross-modal retrieval in remote sensing," in *ICASSP*. IEEE, 2022, pp. 4463–4467.

[3] Benno Weck and Xavier Serra, "Data leakage in cross-modal retrieval training: A case study," in *ICASSP*. IEEE, 2023, pp. 1–5.

[4] Fangming Zhong, Zhikui Chen, and Geyong Min, "Deep discrete cross-modal hashing for cross-media retrieval," *Pattern Recognition*, vol. 83, pp. 64–77, 2018.

[5] Devansh Arpit, Stanisław 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 *ICML*. PMLR, 2017, pp. 233–242.

[6] 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*, vol. 31, 2018.

[7] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, vol. 64, no. 3, pp. 107–115, 2021.

[8] Bin Huang, Ping Zhang, and Chaoyang Xu, "Combining layered label correction and mixup supervised contrastive learning to learn noisy labels," *Information Sciences*, vol. 642, pp. 119242, 2023.

[9] Shikun Li, Xiaobo Xia, Shiming Ge, and Tongliang Liu, "Selective-supervised contrastive learning with noisy labels," in *CVPR*, 2022, pp. 316–325.

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

[11] Bo Han, Jiangchao Yao, Niu Gang, Mingyuan Zhou, Ivor Tsang, Ya Zhang, and Masashi Sugiyama, "Masking: A new perspective of noisy supervision," in *NeurIPS*, 2018, pp. 5839–5849.

[12] Fuyan Ma, Bin Sun, and Shutao Li, "Transformer-augmented network with online label correction for facial expression recognition," *IEEE Transactions on Affective Computing*, 2023.

[13] 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 *ACM Multimedia*, 2022, pp. 629–637.

[14] Junnan Li, Richard Socher, and Steven C.H. Hoi, "Dividemix: Learning with noisy labels as semi-supervised learning," in *ICLR*, 2020.

[15] Peng Hu, Xi Peng, Hongyuan Zhu, Liangli Zhen, and Jie Lin, "Learning cross-modal retrieval with noisy labels," in *CVPR*, 2021, pp. 5403–5413.

[16] Daiki Okamura, Ryosuke Harakawa, and Masahiro Iwahashi, "Lcn: Label correction based on network prediction for cross-modal retrieval with noisy labels," in *APSIPA ASC*. IEEE, 2022, pp. 354–358.

[17] Erkun Yang, Dongren Yao, Tongliang Liu, and Cheng Deng, "Mutual quantization for cross-modal search with noisy labels," in *CVPR*, 2022, pp. 7551–7560.

[18] Xingjun Ma, Hanxun Huang, Yisen Wang, Simone Romano, Sarah Erfani, and James Bailey, "Normalized loss functions for deep learning with noisy labels," in *ICML*. PMLR, 2020, pp. 6543–6553.

[19] Zhilu Zhang and Mert Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.

[20] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár, "Focal loss for dense object detection," in *ICCV*, 2017, pp. 2980–2988.

[21] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey, "Symmetric cross entropy for robust learning with noisy labels," in *CVPR*, 2019, pp. 322–330.

[22] David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin A Raffel, "Mixmatch: A holistic approach to semi-supervised learning," *NeurIPS*, vol. 32, 2019.

[23] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos, "A new approach to cross-modal multimedia retrieval," in *ACM Multimedia*, 2010, pp. 251–260.

[24] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *NAACL*, 2010, pp. 139–147.

[25] Yuxin Peng, Jinwei Qi, and Yuxin Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Trans*, vol. 27, no. 11, pp. 5585–5599, 2018.

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

[27] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes, "On deep multi-view representation learning," in *ICML*. PMLR, 2015, pp. 1083–1092.

[28] Peng Hu, Liangli Zhen, Dezhong Peng, and Pei Liu, "Scalable deep multimodal learning for cross-modal retrieval," in *ACM SIGIR*, 2019, pp. 635–644.

[29] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng, "Deep supervised cross-modal retrieval," in *CVPR*, 2019, pp. 10394–10403.