A GENERAL FRAMEWORK FOR INCOMPLETE CROSS-MODAL RETRIEVAL WITH MISSING LABELS AND MISSING MODALITIES

Mingyang Li, Shao-Lun Huang, Lin Zhang*

Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University

ABSTRACT

Among various cross-modal retrieval methods, the supervised methods achieve the best performance by exploiting the semantic labels. However, in realistic applications, the data are not always complete with labels and full multi-modal data, which makes these methods hard to be used. In this paper, we propose a general framework for handling cross-modal retrieval tasks with both missing labels and missing modalities. To be more specific, in our framework we embed the data in each modality and labels all into a common feature space and maximize their correlation altogether. When labels or data in some modalities are missing, we can still maximize the correlation between the remaining data or labels. Combined with the label prediction and data reconstruction modules, our model can effectively extract useful information from the incomplete data for cross-modal retrieval tasks. In the extensive experiments, our model outperforms many other methods on different datasets, which proves the effectiveness and flexibility for handling incomplete data of our model.

***Index Terms*—** Cross-modal Retrieval, Multi-modal Learning, Missing Modality

1. INTRODUCTION

Nowadays, multi-modal data are widespread in our life and a huge amount of them are generated every day. For instance, people usually share data on the Internet in different forms such as images, texts, and videos, etc. In reality, we also collect data in different modalities using various sensors, such as cameras and microphones. It is of great significance to better analyze and exploit these multi-modal data. Among assorted applications of multi-modal data [1, 2], cross-modal retrieval aims to search for the most semantically related data to the query from different modalities, which attracts increasing interests in both research and industry recently. However, multi-modal data are in varied forms and distributions, making them hard to be compared with directly for cross-modal retrieval tasks. Such a phenomenon is called the heterogeneity gap [3].

Recently, to bridge the heterogeneity gap among multi-modal data, many supervised cross-modal retrieval methods [4, 5, 6] are proposed using the shared semantic labels to guide the learning of the correspondence between data in different modalities, which outperforms traditional unsupervised methods [7, 8, 9, 10, 11]. However, the supervised cross-modal retrieval methods require not only the multi-modal data are completely paired or matched but also their shared semantic labels are given. The rigorous requirement for the data is not close to realistic situations and leaves difficulty in multi-modal data collection and labeling. For example, the lots of multi-modal data, such as images, texts, and videos, uploaded by users do not have labels or tags. And a huge amount of data are uni-modal without correspondence among multi-modal data. It is of pivotal importance to develop algorithms that can handle complex situations where data and labels are not completely available. Here, we consider a more complex and realistic scenario for cross-modal retrieval tasks, where the dataset can be incomplete with missing labels and missing modalities. To be more specific, only a small portion of the dataset is in the complete forms with full data in each modality and their shared semantic labels. For the rest of the data, some of them do not have the shared semantic labels, and some of them have incomplete multi-modal data. Such a setting can better satisfy the complex scenarios for data collection in the real world.

In this paper, we propose a novel and general cross-modal retrieval framework for handling such complex settings with missing labels and missing modalities. Our model embeds all the data in each modality and the shared semantic labels to a common feature space, in which the correlation among cross-modal data and correspondence between data and semantic labels can be learned jointly. When partial multi-modal data or labels are missing, we can still exploit the remaining incomplete data in our model to maximize the correlation of multi-modal data and labels, which is not able for the common supervised methods. Our method also incorporates cross-modal reconstruction for bridging the heterogeneity gap and label prediction for strengthening discriminative feature extraction, which can further improve the cross-modal retrieval performance. In extensive experiments, our method outperforms many other cross-modal retrieval methods in different settings, which shows the superior performance and generality of our proposed methods.

2. METHOD

2.1. Notations

Without loss of generality, in this paper we focus on the bi-modal cross-modal retrieval tasks with images and texts. Consider the settings with missing labels and missing modalities, the full dataset are consist of three parts, which denoted as $\mathcal{D} = \{\mathcal{D}_{comp}, \mathcal{D}_{un}, \mathcal{D}_{miss}\}$, where $\mathcal{D}_{comp}$ represents the complete dataset, $\mathcal{D}_{un}$ represents the unlabeled dataset, and $\mathcal{D}_{miss}$ represents the missing modality dataset. The complete dataset $\mathcal{D}_{comp}$ is a collection of n_1 instances of complete image-text-label triplets, denoted as $\mathcal{D}_{comp} = \{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i)\}_{i=1}^{n_1}$, where $\mathbf{x}_i \in \mathbb{R}^{d_x}, \mathbf{y}_i \in \mathbb{R}^{d_y}, \mathbf{z}_i \in \mathbb{R}^C$ are the image sample, text sample and their shared semantic label vector of the ith instance, d_x and d_y are the di-

*Corresponding Author.

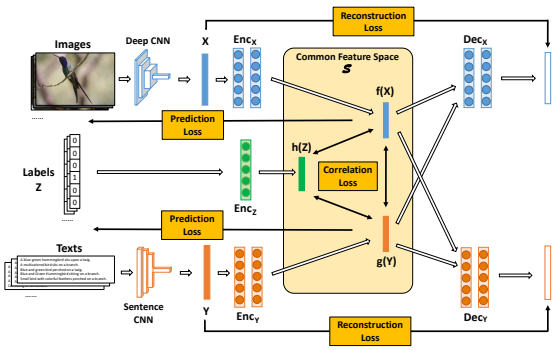


Fig. 1. The general framework of our proposed model. (Zoom in for a better view.)

mensions of image and text data and C is the number of semantic categories. The label vector is represented as $\mathbf{z}_i = [z_{i1}, z_{i2}, ..., z_{iC}]$, where $z_{ij} = 1$ if the ith instance belongs to jth semantic category, otherwise $z_{ij} = 0$. The unlabeled dataset $\mathcal{D}_{un}$ is a collection of n_2 instances of only image-text pairs without labels, denoted as $\mathcal{D}_{un} = \{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^{n_2}$. The missing modality dataset $\mathcal{D}_{miss}$ is a collection of n_3+n_4 instances of only either image or text data with labels, denoted as $\mathcal{D}_{miss} = \{(\mathbf{x}_k, \mathbf{z}_k)\}_{k=1}^{n_3} \bigcup \{(\mathbf{y}_l, \mathbf{z}_l)\}_{l=1}^{n_4}$. We set the number of instance of each modality the same in the missing modality dataset. We use uppercase letters to denote the image, text and label matrices for all instances in the dataset as $\mathbf{X} \in \mathbb{R}^{n \times d_x}$, $\mathbf{Y} \in \mathbb{R}^{n \times d_y}$ and $\mathbf{Z} = \{\mathbf{z}_1^T; ...; \mathbf{z}_n^T\} \in \mathbb{R}^{n \times C}$ respectively.

2.2. The Proposed Method

The general framework of our proposed method is illustrated in Fig 1. We take as input the pre-processed features of image and text data. And we use two encoders $f(\cdot, \theta_f)$ and $g(\cdot, \theta_g)$ to embed input features from different modalities onto a common feature space $\mathcal{S}$. Unlike previous cross-modal retrieval methods that only use labels for prediction, in our method we also use another encoder $h(\cdot, \theta_h)$ to embed label vectors onto the same common feature space $\mathcal{S}$ to maximize the correlation between multi-modal data and labels. To extract discriminative features, we use a one-layer fully connected layer for the prediction of the semantic labels. To keep less information loss in the embedding process onto the common feature space, we further involve two decoders $\mathbf{Dec}_X$ and $\mathbf{Dec}_Y$ to reconstruct the original input features of each modality.

Now we introduce the losses used in the training process of different kinds of datasets. Consider the settings with complete data, unlabeled data, and data with missing modalities ($\mathcal{D} = \{\mathcal{D}_{comp}, \mathcal{D}_{un}, \mathcal{D}_{miss}\}$), the total loss is the combination of the loss on each dataset, which is formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{comp} + \mathcal{L}_{un} + \mathcal{L}_{miss} \tag{1}$$

For the loss term on each dataset in the above equation, it consists of three parts, which are correlation loss, prediction loss, and reconstruction loss. We formulate the loss function on each dataset as follows:

$$\mathcal{L}_{\mathcal{D}} = \mathcal{L}_{\mathcal{D}}^{pred} + \alpha * \mathcal{L}_{\mathcal{D}}^{corr} + \beta * \mathcal{L}_{\mathcal{D}}^{recon} \tag{2}$$

where $\mathcal{D} \in \{comp, un, miss\}$ and α, β are the hyper-parameters to control the weight of each loss term. The hyper-parameters α, β are kept the same for each dataset. At first, we introduce these training losses used for the complete dataset. For the datasets with missing labels or missing modalities, these loss terms can be viewed as the variants of the loss terms used for the complete dataset.

2.2.1. *Training of the Complete Data*

Correlation Loss The core of our proposed framework is maximizing the correlation among paired multi-modal data and their shared semantic labels altogether. For the image-text-label triplets in the complete dataset, we embed them into the common feature space respectively and train the model to extract features with maximal correlation altogether. Other methods usually only maximize the correlation between the features of multi-modal data [7, 9]. In contrast, our method involves features of labels and maximizes the correlation among multi-modal data and labels altogether. In this way, the features of data in each modality contain more information of the shared semantic labels, which is beneficial for mitigating the heterogeneous characteristics of multi-modal data and improving cross-modal retrieval performance. We train our model with each pair in the triplet (i.e. image-text, image-label, and text-label) using soft-HGR loss [11], which is based on solving a more general HGR maximal correlation problem and achieves excellent performance. The formulation of the correlation loss is as follows:

$$\begin{aligned}\mathcal{L}_{comp}^{corr} =& - corr(\mathbf{X}, \mathbf{Y}) - corr(\mathbf{X}, \mathbf{Z}) - corr(\mathbf{Y}, \mathbf{Z}) \\ =& \frac{1}{2} tr(\Lambda_f \Lambda_g) - \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}_i)^{\mathrm{T}} g(\mathbf{y}_i) + \\ & \frac{1}{2} tr(\Lambda_f \Lambda_h) - \frac{1}{n} \sum_{i=1}^{n} f(\mathbf{x}_i)^{\mathrm{T}} h(\mathbf{z}_i) + \\ & \frac{1}{2} tr(\Lambda_g \Lambda_h) - \frac{1}{n} \sum_{i=1}^{n} g(\mathbf{y}_i)^{\mathrm{T}} h(\mathbf{z}_i)\end{aligned} \tag{3}$$

where, $corr$ represents the correlation loss using soft-HGR, Λ represents the covariance matrix of features, tr represents the trace of matrix. When the dataset has missing labels or missing modalities, we can still maximize the correlation among the available data or labels in the datasets, exploiting the utilizable information of them effectively. Details are given in the following sections.

Reconstruction Loss To make the extracted features also keep the information of the original data, we adopt the auto-encoder framework to reconstruct the original data of each modality. To be specific, we use decoders to map the features in the common feature space back to the original input space and use Mean Squared Error (MSE) loss for training. Not only do we reconstruct the input data with the features extracted from the data itself, but also we use cross-modal features for reconstruction. Such a cross-modal reconstruction process also helps to shrink the heterogeneity gap between multi-modal data. The formulation is given as follows:

$$\begin{aligned}\mathcal{L}_{comp}^{recon} =& \frac{1}{n} (\sum_{i=1}^{n} \|Dec_X(f(\mathbf{x}_i)) - \mathbf{x}_i\|_2^2 + \sum_{i=1}^{n} \|Dec_Y(g(\mathbf{y}_i)) - \mathbf{y}_i\|_2^2) + \\ & \frac{1}{n} (\sum_{i=1}^{n} \|Dec_X(g(\mathbf{y}_i)) - \mathbf{x}_i\|_2^2 + \sum_{i=1}^{n} \|Dec_Y(f(\mathbf{x}_i)) - \mathbf{y}_i\|_2^2)\end{aligned} \tag{4}$$

where $\mathbf{Dec}_X$ and $\mathbf{Dec}_Y$ represent the decoders for image and text modality respectively.

Prediction Loss To fully exploit the semantic label information, we use a shared prediction layer to classify the features from

different modalities in the common feature space. The label prediction can make the extracted features to be discriminative so that features with different semantic labels will lie distant in the common feature space. We adopt the MSE loss for training as it performs better than other loss function such as cross-entropy in our experiments, the formulation of which is as follows:

$$\mathcal{L}_{comp}^{pred} = \frac{1}{n}(\sum_{i=1}^{n} \|Pred(f(\mathbf{x}_i)) - \mathbf{z}_i\|_2^2 + \sum_{i=1}^{n} \|Pred(g(\mathbf{y}_i)) - \mathbf{z}_i\|_2^2) \tag{5}$$

where $Pred$ represents the prediction layer.

To summary, the final loss for the complete dataset is a weighted summation of the three losses introduced above. For the dataset with missing labels or missing modalities, we use the variants of these losses to fit the different settings in each dataset.

2.2.2. Training of the Data with Missing Labels

For the dataset $\mathcal{D}_{un} = \{(\mathbf{x}_j, \mathbf{y}_j)\}_{j=1}^{n_2}$, we only have multi-modal data pairs without common labels. In our framework under such setting, we modify the correlation loss to only maximize the correlation between available data pairs as follows:

$$\mathcal{L}_{corr}^{un} = -corr(\mathbf{X}, \mathbf{Y}) = \frac{1}{2}tr(\Lambda_f \Lambda_g) - \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{x}_i)^{\mathrm{T}} g(\mathbf{y}_i) \tag{6}$$

Since no labels can be used in this dataset, we adopt the Entropy Minimization (EM) loss in our framework for the prediction loss term, which is widely used in semi-supervised learning [12]. The EM loss can make the prediction of unlabeled data to be more confident on some classes based on the prediction layer learned from the labeled data, which helps to extract discriminative features from these unlabeled data. The EM loss in our framework is as follows:

$$\mathcal{L}_{pred}^{un} = \frac{1}{n}\sum_{i=1}^{n}(Ent(Pred(f(\mathbf{x}_i))) + Ent(Pred(g(\mathbf{y}_i)))) \tag{7}$$

where Ent represents the entropy of the prediction, which is calculated as $Ent(pred) = -\sum_{k=1}^{C}(pred_k * log(pred_k))$. The outputs of the prediction layer are normalized by softmax operation.

For the reconstruction loss, it is kept the same for the unlabeled dataset as the loss used for the complete dataset since it does not get influenced by missing labels.

2.2.3. Training of the Data with Missing Modalities

For the dataset with missing modalities $\mathcal{D}_{miss} = \{(\mathbf{x}_k, \mathbf{z}_k)\}_{k=1}^{n_3} \bigcup \{(\mathbf{y}_l, \mathbf{z}_l)\}_{l=1}^{n_4}$, we have two separate uni-modal datasets with labels for each modalities. For the correlation loss, we can still utilize the uni-modal dataset to maximize the correlation between data in each modality and labels as follows:

$$\begin{aligned}\mathcal{L}_{miss}^{corr} =& - corr(\mathbf{X}, \mathbf{Z}) - corr(\mathbf{Y}, \mathbf{Z}) \\ =& \frac{1}{2}tr(\Lambda_f \Lambda_h) - \frac{1}{n_X}\sum_{i=1}^{n_X} f(\mathbf{x}_i)^{\mathrm{T}} h(\mathbf{z}_i) + \\ & \frac{1}{2}tr(\Lambda_g \Lambda_h) - \frac{1}{n_Y}\sum_{j=1}^{n_Y} g(\mathbf{y}_j)^{\mathrm{T}} h(\mathbf{z}_j)\end{aligned} \tag{8}$$

where n_X, n_Y are the size of each uni-modal dataset.

For the prediction loss, we can just use MSE loss on each uni-modal dataset as follows:

$$\mathcal{L}_{miss}^{pred} = \frac{1}{n_X + n_Y}(\sum_{i=1}^{n_X} \|Pred(f(\mathbf{x}_i)) - \mathbf{z}_i\|_2^2 + \sum_{j=1}^{n_Y} \|Pred(g(\mathbf{y}_j)) - \mathbf{z}_j\|_2^2) \tag{9}$$

For the reconstruction loss, though we do not have the multi-modal data pairs, we can still reconstruct data using features extracted from the data itself to train our model as follows:

$$\mathcal{L}_{miss}^{recon} = \frac{1}{n_X}\sum_{i=1}^{n_X} \|Dec_X(f(\mathbf{x}_i)) - \mathbf{x}_i\|_2^2 + \frac{1}{n_Y}\sum_{i=1}^{n_Y} \|Dec_Y(g(\mathbf{y}_i)) - \mathbf{y}_i\|_2^2 \tag{10}$$

In conclusion, the total loss (2) proposed in our framework incorporates the correlation between multi-modal data and semantic labels, data reconstruction, and label prediction altogether. The features extracted by our model are self-expressive and discriminative, and the heterogeneity gap between different modalities is also effectively reduced. With simple modifications on the three loss terms, our framework can handle datasets with missing labels and missing modalities, fully exploiting the data in all settings to learn a better model for cross-modal retrieval tasks.

3. EXPERIMENTS

3.1. Setup

In the experiments, we choose three widely-used datasets Wikipedia [13], Pascal Sentence [14] and XmediaNet [15] with different numbers of classes and samples. For Wikipedia and Pascal Sentence dataset, following [6], we use pre-trained VGGNet [16] and Sentence CNN [17] to extract features from the raw image and text data respectively. For XmediaNet, we directly use the features provided in the dataset, which are deep image features extracted by VGGNet and BoW text features.

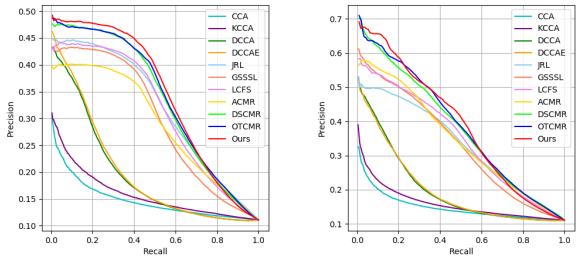
In our experiments, we split each training dataset randomly into three parts: $\mathcal{D} = \{\mathcal{D}_{comp}, \mathcal{D}_{un}, \mathcal{D}_{miss}\}$. We set three settings with different degrees of difficulty: Easy, Medium, and Hard, which contain different proportions of complete data. In the Easy setting, the proportions of $\mathcal{D}_{comp}, \mathcal{D}_{un}, \mathcal{D}_{miss}$ are $(30\%, 35\%, 35\%)$. Similarly, the data proportions in the Medium and Hard settings are $(20\%, 40\%, 40\%)$ and $(10\%, 45\%, 45\%)$ respectively. In our model, the feature dimension of the common feature space is set to 1024. We train our model using Adam optimizer [18] with a learning rate of 1e-4 for 200 epochs. We set the hyper-parameter $\alpha = 0.1$ and $\beta = 0.01$ used in equation (2) by grid search on each dataset.

3.2. Comparison with Other Methods

In our experiments, we compare our method with 9 other cross-modal retrieval methods, including unsupervised (CCA [7], KCCA [8], DCCA [9], DCCAE [10]), semi-supervised (JRL [19], GSSSL [20]), and supervised (LCFS [4], ACMR [5], DSCMR [6], OTCMR [21]) methods. For the unsupervised and semi-supervised methods, we train the models using both complete data and unlabeled data. For the supervised methods, we can only train them with complete data. All the methods use the same data for training. For evaluating each method, we retrieve text with image (I2T) and retrieve image with text (T2I) in the test dataset. The average results of mean average precision (mAP) on I2T and T2I tasks of each method under different settings on different datasets are shown in Table 1. And we

Method	Wikipedia			Pascal Sentence			XmediaNet		
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
CCA	0.1673	0.1695	0.1713	0.4539	0.4525	0.4533	0.4667	0.4489	0.4311
KCCA	0.1814	0.1841	0.1858	0.5036	0.5103	0.5028	0.3764	0.3538	0.3266
DCCA	0.2409	0.2426	0.2439	0.5680	0.5658	0.5560	0.5133	0.5115	0.5072
DCCAE	0.2465	0.2439	0.2427	0.5608	0.5526	0.5570	0.5232	0.5029	0.5024
JRL	0.4227	0.4134	0.4000	0.6482	0.6284	0.5495	0.6389	0.6061	0.5040
GSSSL	0.4220	0.4114	0.3757	0.6544	0.6025	0.4463	0.4901	0.4577	0.3931
LCFS	0.4303	0.4231	0.4009	0.6690	0.6582	0.5765	0.5055	0.4316	0.3112
ACMR	0.4269	0.4087	0.3817	0.6283	0.6162	0.5306	0.4470	0.4137	0.3081
DSCMR	0.4692	0.4586	0.4295	0.6699	0.6612	0.5526	0.6247	0.5928	0.4910
OTCMR	0.4775	0.4687	0.4324	0.6757	0.6526	0.5589	0.6715	0.6457	0.5699
Ours	**0.4826**	**0.4747**	**0.4679**	**0.7021**	**0.7018**	**0.6863**	**0.7044**	**0.6938**	**0.6090**

Table 1. The comparison of the average mAP results of I2T and T2I retrieval tasks on each dataset under each setting.



(a) Image → Text (b) Text → Image

Fig. 2. The comparison of P-R curves.

Loss Terms	I2T	T2I	Avg
$\mathcal{L}_{corr}$	0.4180	0.3937	0.4059
$\mathcal{L}_{recon}$	0.4299	0.4098	0.4198
$\mathcal{L}_{pred}$	0.4798	0.4290	0.4544
$\mathcal{L}_{pred}$ & $\mathcal{L}_{corr}$	0.5016	0.4417	0.4717
$\mathcal{L}_{pred}$ & $\mathcal{L}_{corr}$ & $\mathcal{L}_{recon}$	**0.5029**	**0.4464**	**0.4747**

Table 2. The ablation analysis of different loss terms.

plot the P-R curves of each method on the Wikipedia dataset under medium setting in Fig 2.

From the results in Table 1, our method outperforms all the other methods on each task, each setting, and each dataset. It is due to the fact that our method can effectively extract the correlation information from all the complete data, unlabeled data, and data with missing modalities, while other methods can not. The improvement of our method compared with the best average result of other methods is greater in the Hard setting with the fewest labeled data. The improvements are about 8%, 23%, and 7% on each dataset respectively. The performance of supervised and semi-supervised methods usually degrade obviously when complete data are scarce while our method degrades much slower for it can still take advantage of incomplete data. The P-R curves of our method are above the others in both tasks. These results show that our method can fully exploit the incomplete data for learning the correlation to improve cross-modal retrieval performance.

3.3. Ablation Study

Here, we give an ablation analysis of the different loss terms used in our loss function and the datasets with different forms used in the training process to give a clear understanding of how these elements contribute in our method. We take Wikipedia dataset on medium setting for experiments of ablation study.

For the loss terms, we train several variants of our model using all the data in the dataset with each loss term in the loss function and adding them one by one. The results of each variant are shown in Table 2. The variants trained with each loss term only can all get fairly good cross-modal retrieval performance. By adding prediction loss, correlation loss, and reconstruction loss one by one in the loss function, we can see the performance will continually increase, which shows the loss terms proposed in our method can all contribute to our total loss function and they are compatible with each other.

Datasets	I2T	T2I	Avg
$\mathcal{D}_{comp}$	0.4799	0.4257	0.4528
$\mathcal{D}_{comp}$ & $\mathcal{D}_{un}$	0.4905	0.4397	0.4651
$\mathcal{D}_{comp}$ & $\mathcal{D}_{miss}$	0.4981	0.4446	0.4714
$\mathcal{D}_{comp}$ & $\mathcal{D}_{un}$ & $\mathcal{D}_{miss}$	**0.5029**	**0.4464**	**0.4747**

Table 3. The ablation analysis of different dataset forms.

For the datasets used in our model, we train one variant model with the complete dataset only as baseline and then train two more variant models using the unlabeled dataset or missing modality dataset added with the complete dataset. The results are shown in Table 3. Adding the unlabeled dataset and missing modality dataset both further improves the performance than baseline and the missing modality dataset improves more, which shows the semantic labels in the missing modality dataset is more helpful for the cross-modal retrieval task. Training with all three datasets, our method achieves the best cross-modal retrieval performance.

4. CONCLUSION

In this paper, we propose a novel and general framework for handling cross-modal retrieval tasks on the dataset with missing labels and missing modalities, which is more suitable for realistic scenarios. The strategy of learning the correlation among the image-text-label triplet in the common feature space guarantees the effectiveness and flexibility for extracting useful information when any part is missing. By combining correlation, prediction, and reconstruction loss terms from different aspects altogether, our method outperforms many other methods on different datasets and different settings. The ablation study shows the contributions of each loss term and each form of the datasets clearly. All these results prove the efficacy of our method for cross-modal retrieval tasks with incomplete data.

ACKNOWLEDGEMENT The research of Shao-Lun Huang is supported in part by the Shenzhen Science and Technology Program under Grant KQTD20170810150821146, National Key R&D Program of China under Grant 2021YFA0715202 and High-end Foreign Expert Talent Introduction Plan under Grant G2021032013L.

5. REFERENCES

[1] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[2] Wenzhong Guo, Jianwen Wang, and Shiping Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[3] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang, "A comprehensive survey on cross-modal retrieval," *arXiv preprint arXiv:1607.06215*, 2016.

[4] Kaiye Wang, Ran He, Wei Wang, Liang Wang, and Tieniu Tan, "Learning coupled feature spaces for cross-modal matching," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2088–2095.

[5] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen, "Adversarial cross-modal retrieval," in *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 2017, pp. 154–162.

[6] Liangli Zhen, Peng Hu, Xu Wang, and Dezhong Peng, "Deep supervised cross-modal retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10394–10403.

[7] Harold Hotelling, "Relations between two sets of variates," in *Breakthroughs in statistics*, pp. 162–190. Springer, 1992.

[8] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639–2664, 2004.

[9] Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu, "Deep canonical correlation analysis," in *International conference on machine learning*, 2013, pp. 1247–1255.

[10] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes, "On deep multi-view representation learning," in *International Conference on Machine Learning*, 2015, pp. 1083–1092.

[11] Lichen Wang, Jiaxiang Wu, Shao-Lun Huang, Lizhong Zheng, Xiangxiang Xu, Lin Zhang, and Junzhou Huang, "An efficient approach to informative feature extraction from multimodal data," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 5281–5288.

[12] Yves Grandvalet, Yoshua Bengio, et al., "Semi-supervised learning by entropy minimization.," *CAP*, vol. 367, pp. 281–296, 2005.

[13] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, and Nuno Vasconcelos, "On the role of correlation and abstraction in cross-modal multimedia retrieval," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 3, pp. 521–535, 2013.

[14] Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 139–147.

[15] Yuxin Peng, Xin Huang, and Yunzhen Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Transactions on circuits and systems for video technology*, vol. 28, no. 9, pp. 2372–2385, 2017.

[16] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[17] Yoon Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, Oct. 2014, pp. 1746–1751, Association for Computational Linguistics.

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

[19] Xiaohua Zhai, Yuxin Peng, and Jianguo Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 6, pp. 965–978, 2013.

[20] Liang Zhang, Bingpeng Ma, Guorong Li, Qingming Huang, and Qi Tian, "Generalized semi-supervised and structured subspace learning for cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 128–141, 2017.

[21] Mingyang Li, Shao-Lun Huang, and Lin Zhang, "Otcmr: Bridging heterogeneity gap with optimal transport for cross-modal retrieval," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021, pp. 3216–3220.