

A Reconstruction-based Visual-Acoustic-Semantic Embedding Method for Speech-Image Retrieval

Wenlong Cheng, Wei Tang, Yan Huang, Yiwen Luo, and Liang Wang, *Fellow, IEEE*

Abstract—Speech-image retrieval aims at learning the relevance between image and speech¹. Prior approaches are mainly based on bi-modal contrastive learning, which can not alleviate the cross-modal heterogeneous issue between visual and acoustic modalities well. To address this issue, we propose a visual-acoustic-semantic embedding (VASE) method. First, we propose a tri-modal ranking loss by taking advantage of semantic information corresponding to the acoustic data, which introduces the auxiliary alignment to enhance the alignment between image and speech. Second, we introduce a cycle-consistency loss based on feature reconstruction. It can further alleviate the heterogeneous issue between different data modalities (*e.g.*, visual-acoustic, visual-textual and acoustic-textual). Extensive experiments have demonstrated the effectiveness of our proposed method. In addition, our VASE model achieves state-of-the-art performance on the speech-image retrieval task on the Flickr8K [4] and Places [2] datasets.

Index Terms—Speech-image retrieval, tri-modal ranking loss, cycle-consistency loss, visual-acoustic-semantic embedding.

I. INTRODUCTION

NOWADAYS, with the fast development of information technology and hardware devices, multimedia data (*e.g.*, image, audio, text and video) can be seen everywhere in our daily lives. For understanding and dealing with massive amount of multimedia data, multi-modal data processing technology appears and attracts a lot of attention. Cross-modal retrieval [42]–[45], visual question answering [46]–[49] and image captioning [50]–[53] are some of the representative tasks. Speech-image retrieval is a bidirectional cross-modal retrieval task, which is shown in Fig.1. It aims to retrieve relevant images given speeches (speech-to-image retrieval, S2I) or find relevant speeches given images (image-to-speech retrieval, I2S). It is more convenient and faster if we use speech instead of writing or typing in some scenarios [41], [59]. It

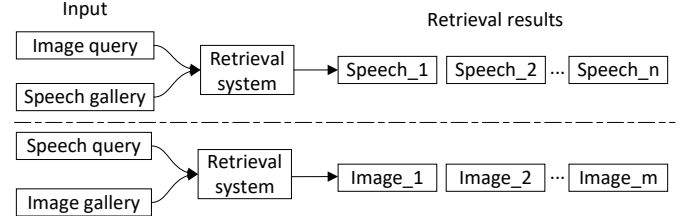


Fig. 1. The diagram of speech-image retrieval. It includes image-to-speech retrieval (top) and speech-to-image retrieval (bottom).

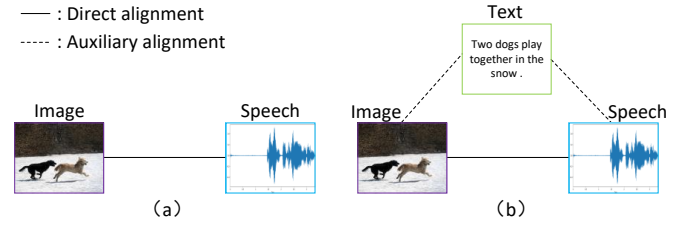


Fig. 2. The alignment between image and speech. The left figure describes the alignment before adding text, while the right figure describes the alignment after adding text. Before adding text, there is only direct alignment between image and speech (“speech-image”). After adding text, the new auxiliary alignment between image and speech is established through the path of “image-text-speech”.

can be applied to outdoor scenes that are inconvenient for writing or typing, or scenes that require real-time interactions, such as speech-based image search on mobile phones, human-computer interaction, *etc.* The challenge of speech-image retrieval lies in how to accurately measure the cross-modal similarities between images and speeches [1].

Recent studies are mainly based on bi-modal contrastive learning. These methods [2]–[4], [7], [8] are typically composed of two branches for generating visual and acoustic embedding features in a common feature space [31]. Then, a bi-modal similarity-based ranking loss (*e.g.*, triplet loss) is used to pull the matching image and speech pair closer and push the non-matching image and speech pair further away in a common feature space. Further, they can be divided into two major categories: global coarse-grained matching methods [7], [8] and local fine-grained matching methods [2]–[4]. The global coarse-grained matching methods are proposed to learn associations between the global natural images and free-form speeches. The local fine-grained matching methods are proposed to learn associations between the image fragments and speech segments, and then an aggregation algorithm is adopted to obtain the global similarities between images and speeches. Because of the usage of local fine-grained matching relationships, they can generally obtain better performance than global coarse-grained matching methods. However, both

Wenlong Cheng, Wei Tang and Yan Huang are with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China (e-mail: wenlong.cheng@nlpr.ia.ac.cn; tang-weirichard@163.com; yhuang@nlpr.ia.ac.cn).

Yiwen Luo is with the Institute of Artificial Intelligence and Robotics (IAIR), Xi'an Jiaotong University (XJTU), Xi'an 710049, China (email: luoyiwen@mail.nwpu.edu.cn).

Liang Wang is with the Center for Research on Intelligent Perception and Computing (CRIPAC), National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences (UCAS), Beijing 100049, China, and also with the CAS Center for Excellence in Brain Science and Intelligence Technology (e-mail: wangliang@nlpr.ia.ac.cn).

¹In this paper, the speech refers to the spoken caption.

these methods have two problems. First, they mainly focus on the direct alignment between image and speech, but ignore the auxiliary alignment between them. Second, the traditional bi-modal contrastive learning methods are easy to cause similar but non-matching samples to be far away from each other. Both of them will cause the modality gap between visual and acoustic modalities.

For the first problem, we introduce the auxiliary alignment by using the intermediate semantic information (text) corresponding to acoustic data, and accordingly propose a tri-modal ranking loss. Because the text is transcribed from the corresponding speech, it can also describe the matching image. The text is related to both image and speech, so it can be used as a bridge to alleviate the modality gap between image and speech. As shown in Fig 2, adding text will not change the original “image-speech” direct alignment, instead, the text is taken as a bridge to increase the new “image-text-speech” auxiliary alignment. The new auxiliary alignment is beneficial to the alignment between image and speech, thus alleviating the modality gap.

For the second problem, we propose a cycle-consistency loss based on feature reconstruction. Compared with previous feature reconstruction methods [31], [63], our proposed method not only uses matching data samples, but also uses non-matching data samples. This method can alleviate similar but non-matching data samples to be far away from each other. Besides, the non-matching data samples may contain some key semantic information not included in matching data samples. To balance the weights of matching data samples and non-matching data samples in the feature reconstruction process, we introduce dynamic weight factors to adjust the contribution of each data sample.

To demonstrate the effectiveness of our VASE model, we have conducted extensive experiments on the Flickr8K and Places datasets. Moreover, to further verify the robustness of our VASE model, we use two different sets of network structures (vgg16/DAVEnet and resnet50/ResDAVEnet) as the backbones of our proposed model.

The main contributions of this work can be summarized as follows.

- We propose to use the text as auxiliary supervision to bridge the modality gap between image and speech, and accordingly develop a tri-modal framework to enhance the alignment between image and speech.
- We propose a tri-modal reconstruction-based cycle-consistency loss to further alleviate the modality gap, which is quite new in the context of image-speech retrieval.
- Extensive experiments have demonstrated the effectiveness of our proposed method. Our VASE model has achieved state-of-the-art performance for the speech-image retrieval task on the Flickr8K [4] and Places [2] datasets.

The rest of this paper is organized as follows. Related work is introduced in Section II. Our proposed VASE method is described in detail in Section III. The experimental setups, results, visualization and discussion are shown in Section IV. Finally, conclusions are given in Section V.

II. RELATED WORK

A. Image-Text Retrieval

With the development of computer vision [10]–[15], image-text retrieval has made great progress. Some of the early work has been done to explore how to establish the global coarse-grained alignments between image and text. For instance, Kiros *et al.* [16] put forward the VSE method to associate whole images with whole sentences. On the basis of VSE, Faghri *et al.* [17] propose an improved VSE++ method to boost the performance of the model by introducing the hard negative samples, which can reduce the computing cost.

In addition to the global coarse-grained matching methods, there are a lot of local fine-grained matching methods. For instance, Frome *et al.* [9] propose a deep visual semantic embedding model to associate image regions with words by learning semantic relationships between labels. Karpathy *et al.* [18] present a deep multi-modal embedding model to establish the fine-grained matching relationships between image fragments and sentence segments. Huang *et al.* [20] propose a semantic-enhanced image and sentence matching model to achieve significant performance improvements by learning semantic concepts and organizing them in a correct semantic order. Lee *et al.* [19] put forward a novel stacked cross attention mechanism to softly align image fragments and words in a sentence. Li *et al.* [21] propose a simple and interpretable reasoning model VSRN to generate enhanced visual representations. Chen *et al.* [54] put forward an iterative matching with recurrent attention memory (IMRAM) method to capture the sophisticated correspondence between images and sentences. Liu *et al.* [55] present a novel graph structured matching network (GSMN) to learn the fine-grained correspondence by node-level matching and structure-level matching.

Other work tries to integrate additional information into the image-text retrieval. For example, Wang *et al.* [56] propose a consensus-aware visual-semantic embedding (CVSE) model to incorporate consensus information into the image-text matching. Castrejon *et al.* [23] put forward a method to regularize cross-modal convolutional neural networks, so that they have a shared representation that is agnostic of the modality. On the basis of [23], Aytar *et al.* [24] further expand this work.

Although these image-text retrieval methods have obtained impressive progress. However, these methods focus on discrete text and cannot deal with continuous speech well. Compared with text, speech is more difficult to be processed, but it is more convenient in most application scenarios. In this work, we use semantic information corresponding to speech to better deal with speech.

B. Speech-Image Retrieval

In recent years, the studies on speech-image retrieval have drawn much attention in the multimedia community. Some of these studies establish the global coarse-grained matching relationships between visual and acoustic modalities. For instance, Li *et al.* [61] introduce a novel cross-modal factor analysis (CFA) method for cross-modal associations. Harwath *et al.* [7] propose a deep neural network architecture to learn

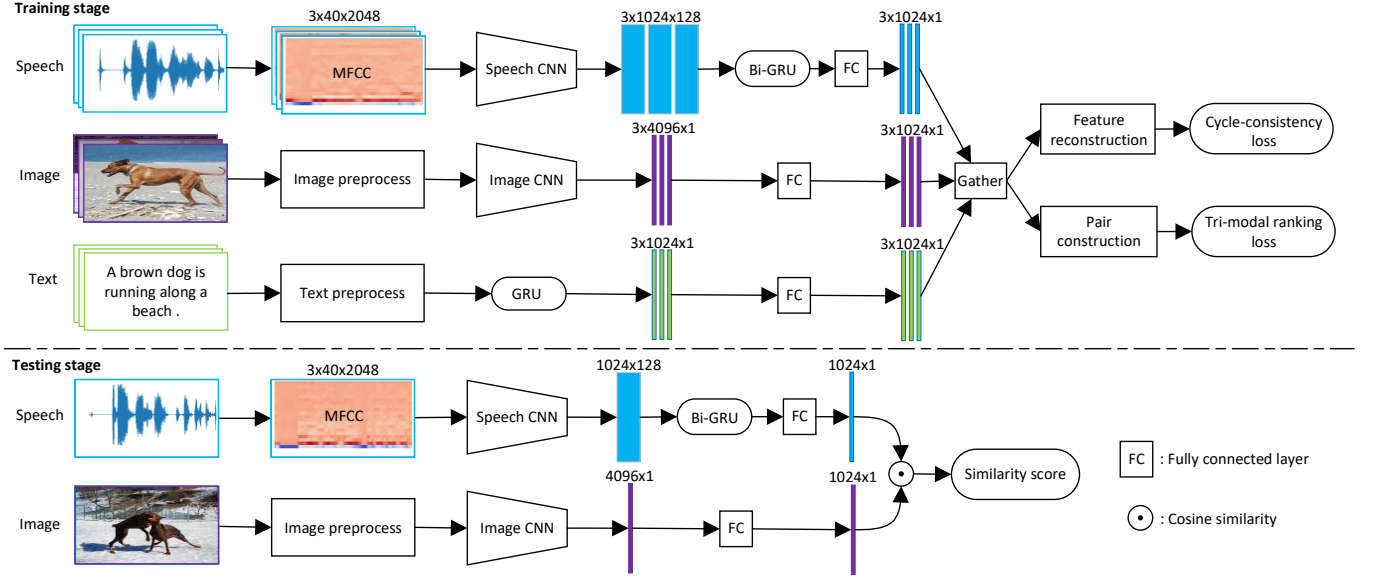


Fig. 3. The whole framework of our VASE model. The top figure demonstrates the training stage, while the bottom figure shows the testing stage. Our training stage is based on the batch data. In the top figure, the batch size of input data is assumed to be 3. “Gather” means merging three data streams into one data stream. (Best viewed in color.)

the global-level associations between the whole natural images and whole spoken captions. In a similar framework, Harwath and Glass [8] improve the encoder of spoken captions to obtain better acoustic feature. And Aytar *et al.* [22] present a deep convolutional network for learning discriminative representations by leveraging massive amounts of synchronized data. Guo *et al.* [58] propose a deep visual-audio network (DVAN) method to establish the correspondence of remote sensing images and spoken captions in a classification framework. On the basis of DVAN, Guo *et al.* [59] further improve the DVAN method by replacing the previous audio encoder with AudioNet. Chen and Lu [57] propose a deep triplet-based hashing (DTBH) method for remote sensing speech-image retrieval. And Zheng *et al.* [63] put forward a novel Adversarial-Metric Learning (AML) model for audio-visual matching.

Some of these studies establish the local fine-grained matching relationships between visual and acoustic modalities. For example, Harwath and Glass [4] propose a deep multi-modal embedding model to align image fragments with speech segments, and then map them into a common feature space. But this method uses RCNN [5] and a particular speech segmentation method [6] to obtain the features of image fragments and speech segments respectively. To address these problems, Harwath *et al.* [2] propose a local fine-grained matching method to associate spoken caption segments with relevant image regions by computing the 3-D density of spatio-temporal similarity. Different from [4], this method uses pixel features in the higher feature map instead of the image region features obtained by object detection. Based on [2], Harwath *et al.* [3] further improve the method by introducing the semi-hard negative mining strategy and enhancing the audio encoder.

However, these methods have two problems that are easy to cause modality gap between image and speech. First, they mainly focus on the direct alignment between image and

speech, but ignore the auxiliary alignment between them. Second, they are easy to cause similar but non-matching samples to be far away from each other. To alleviate the modality gap, we leverage semantic information of acoustic data to introduce the auxiliary alignment between image and speech, which enhances the alignment between them. In addition, we introduce a tri-modal reconstruction-based cycle-consistency loss to prevent similar but non-matching samples from staying away from each other in the common feature space.

III. METHODOLOGY

In this section, we introduce the reconstruction-based visual-acoustic-semantic embedding (VASE) model, which leverages semantic information corresponding to the acoustic data to enhance the alignment between image and speech and uses reconstruction-based cycle-consistency loss to alleviate the modality gap between visual and acoustic modalities. Our VASE model is shown in Fig. 3, and it consists of three parts: feature embedding, tri-modal ranking loss and cycle-consistency loss.

The pipeline of training procedure can be described as follows. First, we use convolutional neural network (CNN), speech CNN and recurrent neural network (RNN) to obtain visual features, acoustic features and textual features, respectively. To better model temporal characteristics of speeches, we use the bidirectional GRU [26] to further process the acoustic features. Then, we map them into a common feature space. Next we construct the positive and negative sample pairs in the same batch, and put them into the tri-modal ranking loss. Besides, we use the features of other modalities in the common feature space to reconstruct the features of current modality by utilizing the correlation among associated data of different modalities. After two feature reconstructions, we use original features in common feature space to constrain the second-order reconstructed features, and then obtain the cycle-consistency

loss. Finally, we use the tri-modal ranking loss and cycle-consistency loss to update our VASE model.

A. Feature Embedding

1) *Visual Embedding*: The feature of the last fully connected layer of CNN is selected as the image feature. In order to calculate the cosine similarities between different data modalities, then we map the image feature into the common feature space. The process of visual embedding can be formulated as follows.

$$v = W_I f_I(I) + b_I \quad (1)$$

where I is the given image, $f_I(\cdot)$ is the function of CNN, therefore, $f_I(I) \in R^{4096 \times 1}$ stands for the image feature after CNN. Besides, $W_I \in R^{1024 \times 4096}$ is the visual affinity matrix, $b_I \in R^{1024 \times 1}$ is the corresponding visual bias, and $v \in R^{1024 \times 1}$ is the final visual embedding in the common feature space. Here, we use two different CNN architectures: vgg16 [25] and resnet50 [30].

2) *Acoustic Embedding*: First of all, we use short-time fourier transform (STFT) to transform the speech into mel-frequency cepstral coefficients (MFCC), which is convenient for calculation. Because the generated spectrogram can be treated as a 1-channel image, we can use the speech CNN to deal with the generated spectrogram, and choose the feature map of the last convolution layer in speech CNN as the intermediate representation. To better model temporal characteristics of the speech, we regard the intermediate representation as a sequence from left to right, then further use the bidirectional GRU [26] to process the intermediate representation, and finally obtain the output of the bidirectional GRU as the acoustic feature. In order to calculate the cosine similarities between different data modalities, we map the acoustic feature into the common feature space. The process of acoustic embedding can be formulated as follows.

$$A^{MFCC} = f_{STFT}(A) \quad (2)$$

$$A^1 = f_A(A^{MFCC}) \quad (3)$$

$$A^2 = f_{BiGRU}(A^1) \quad (4)$$

$$a = W_A A^2 + b_A \quad (5)$$

where A is the given speech, $f_{STFT}(\cdot)$ represents STFT transform, $f_A(\cdot)$ stands for speech CNN, and $f_{BiGRU}(\cdot)$ stands for the bidirectional GRU. Therefore, $A^{MFCC} \in R^{40 \times 2048}$ stands for the MFCC feature of the speech, $A^1 \in R^{1024 \times 128}$ represents the intermediate acoustic feature obtained by speech CNN, and $A^2 \in R^{1024 \times 1}$ stands for the acoustic feature after the bidirectional GRU transform. Besides, $W_A \in R^{1024 \times 1024}$ is the acoustic affinity matrix, $b_A \in R^{1024 \times 1}$ is the corresponding acoustic bias, and $a \in R^{1024 \times 1}$ is the final acoustic embedding in the common feature space. We choose two different speech CNN architectures: DAVenet [2] and ResDAVENet [3].

3) *Semantic Embedding*: Considering that the textual sentence can be regarded as a word sequence, so we use GRU to deal with the given sentence, and choose the last output of GRU as the semantic feature. In order to calculate the

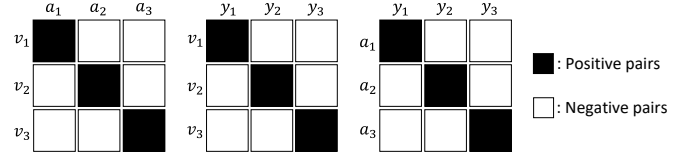


Fig. 4. The overview of three different types of positive and negative sample pairs. The diagrams from left to right are image-speech alignment, image-text alignment and speech-text alignment, respectively.

similarities between different data modalities, we map the semantic feature into the common feature space. The process of semantic embedding can be formulated as follows.

$$y = W_T f_T(T) + b_T \quad (6)$$

where T is the given textual sentence, $f_T(\cdot)$ is the function of GRU, therefore, $f_T(T) \in R^{1024 \times 1}$ stands for the semantic feature after GRU. $W_T \in R^{1024 \times 1024}$ is the semantic affinity matrix, $b_T \in R^{1024 \times 1}$ is the corresponding semantic bias, and $y \in R^{1024 \times 1}$ is the final semantic embedding in the common feature space. GRU has fewer parameters than LSTM [27], but it can achieve comparable results. Therefore, we choose the GRU to process the textual sentence.

B. Tri-modal Ranking Loss

Our proposed tri-modal ranking loss is based on batch data, and it contains three different types of positive and negative sample pairs, which represent three alignments, as shown in Fig. 4. The image-speech alignment can be formulated as follows.

$$L_{v2a} = \sum_v \sum_k \max\{0, \alpha_1 - s(v, a) + s(v, a_k)\} + \sum_a \sum_k \max\{0, \alpha_1 - s(a, v) + s(a, v_k)\} \quad (7)$$

where α_1 is the margin between visual and acoustic modalities, $s(\cdot)$ is the function of cosine similarity, a_k is a contrastive speech feature for image feature v , and vice-versa, v_k is a contrastive image feature for speech feature a . Besides, a is the corresponding speech feature for image feature v . The bi-modal ranking loss L_{v2a} describes the alignment between image and speech, and it is the direct alignment in the speech-image retrieval task. The image-text alignment can be formulated as follows.

$$L_{v2y} = \sum_v \sum_k \max\{0, \alpha_2 - s(v, y) + s(v, y_k)\} + \sum_y \sum_k \max\{0, \alpha_2 - s(y, v) + s(y, v_k)\} \quad (8)$$

where α_2 is the margin between visual and textual modalities, y_k is a contrastive text feature for image feature v , and vice-versa, v_k is a contrastive image feature for text feature y . Besides, y is the corresponding text feature for image feature v . The bi-modal ranking loss L_{v2y} describes the alignment between image and text, and it is an auxiliary alignment in the

speech-image retrieval task. The speech-text alignment can be formulated as follows.

$$L_{a2y} = \sum_a \sum_k \max\{0, \alpha_3 - s(a, y) + s(a, y_k)\} + \sum_a \sum_k \max\{0, \alpha_3 - s(y, a) + s(y, a_k)\} \quad (9)$$

where α_3 is the margin between acoustic and textual modalities, y_k is a contrastive text feature for speech feature a , and vice-versa, a_k is a contrastive speech feature for text feature y . Besides, y is the corresponding text feature for speech feature a . The bi-modal ranking loss L_{a2y} describes the alignment between speech and text, and it is also an auxiliary alignment in the speech-image retrieval task.

Here, we combine the direct alignment with two auxiliary alignments by using a simple summation method. The tri-modal ranking loss can be formulated as follows.

$$L_{tri} = L_{v2a} + L_{v2y} + L_{a2y} \quad (10)$$

where L_{tri} is our proposed tri-modal ranking loss, L_{v2a} is the loss of direct alignment between image and speech, and $L_{v2y} + L_{a2y}$ is the loss of auxiliary alignment between them. In order to save the adjustment time of hyperparameters, these three margins are set to be the same in our experiments, *i.e.*, $\alpha_1 = \alpha_2 = \alpha_3 = \alpha$.

C. Cycle-consistency Loss

To further alleviate the modality gap between image and speech, we propose a reconstruction-based cycle-consistency loss. The key insight of the cycle-consistency loss lies in the feature reconstruction between different data modalities. The mutual feature reconstruction between different data modalities can alleviate the modality gap between different modal samples. Compared with previous reconstructed-based methods, our proposed method not only uses matching data samples, but also uses non-matching data samples. There are two main reasons. First, the non-matching data samples can provide some key information that is not contained in the matching data samples. Second, there are some non-matching but similar data samples, which are easy to stay away from each other in the traditional bi-modal contrastive learning process. This method can alleviate non-matching but similar data samples from being far away from each other in the common feature space.

The data sample of one specific modality can be regarded as a combination of many related elements, such as image objects for image, words for text, and spoken words for speech. Although these elements have different data modalities, they can convey the same key information. For instance, the word “apple”, the spoken word “apple” and the image object “apple” convey the same semantic information. The semantic information of most key elements of the current data sample to be reconstructed can be found in matching data samples of other modalities², so it is reasonable to use the matching data





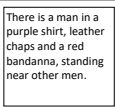
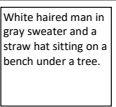
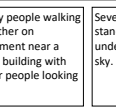
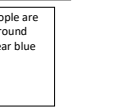
| Original instances | Matching instances | Non-matching instances | |
|---|---|--|---|
| Two dogs chase each other across the snowy ground. |  two dogs, chase, snowy ground |  two dogs, chase, ground |  two dogs, chase, ground |
|  There is a man in a purple shirt, leather chaps and a red bandanna, standing near other men. |  man, purple shirt, leather chaps, red bandanna, standing, other men |  White haired man in gray sweater and a straw hat sitting on a bench under a tree. |  Many people walking together on pavement near a brick building with other people looking on. |
| | |  Several people are standing around under a clear blue sky. | |

Fig. 5. An illustration of why non-matching data samples are also important in the reconstruction process. The first row shows that the non-matching images may include some key semantic information of the original textual caption. The second row shows that the non-matching textual captions may also contain some key semantic information of the original image. The words and phrases below instances represent the key semantic information of the original instance carried by the images and textual captions.

samples of other modalities to reconstruct the data sample of the current modality³.

However, there is still a big modality gap when only using the matching data samples of other modalities in the reconstruction process. Because some key elements may be lost when only using matching data samples of other modalities to reconstruct the data sample of current modality. Taking matching images and textual captions as an example. There are too many details in the original image, but some of them may not be described in matching textual captions. Vice-versa, a textual caption expresses the general idea that may describe multiple scenes, while matching images only present some of them. But non-matching data samples may contain some key elements, which are not included in matching data samples of other modalities. For example, the non-matching textual captions may provide some useful detailed information of current image, and the non-matching images may include some key semantic information of other scenes described by current textual caption, as shown in Fig. 5.

In order to reduce the interference of irrelevant semantic information of non-matching data samples, we introduce dynamic weight factors to adjust the contribution of each data sample in the reconstruction process. The dynamic weight factors are related to the similarities between original data samples and data samples participating in the reconstruction process. The greater the similarity, the greater the corresponding dynamic weight factor. In general, the dynamic weight factors of matching data samples are larger, while those of non-matching data samples are smaller.

The feature reconstruction between different data modalities plays a key role in cycle-consistency loss. After two feature reconstructions, the original data features are used to constrain the reconstructed data features, thus obtaining the cycle-consistency loss. According to whether semantic information is added or not, the cycle-consistency loss has two types: cycle-consistency loss between two data modalities and cycle-consistency loss among three data modalities.

1) *Cycle-consistency Loss Among Three Data Modalities:* The cycle-consistency loss among three data modalities is shown in Fig. 6. The reconstruction process is based on batch-

²The other modalities are the ones that participate in the reconstruction process.

³The current modality is the one that needs to be reconstructed.

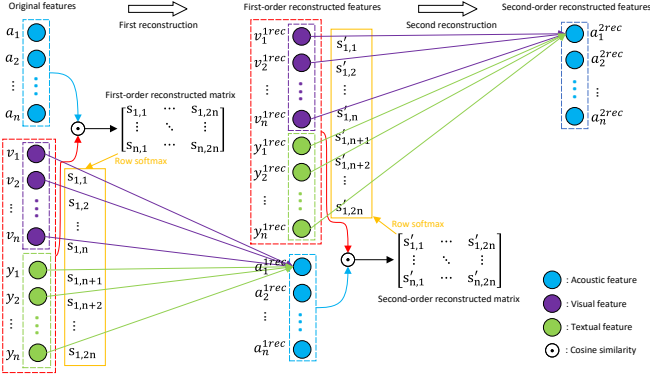


Fig. 6. Taking acoustic feature reconstruction as an example to describe the process of feature reconstruction among three data modalities. (Best viewed in color.)

based data samples. Taking the reconstruction of acoustic features as an example, the reconstruction process can be described as follows.

First of all, according to the modality of each data feature, the batch-based original input data features are divided into three queues, namely batch-based original visual features, batch-based original acoustic features and batch-based original textual features. Each queue is composed of data features of the same modality, and the number of data features in each queue is equal to the batch size. Besides, the data features with the same numerical index in three different queues are matching. In order to calculate the similarities between original acoustic features and original data features of other modalities, we need to concatenate the batch-based original visual and textual features. Then we can obtain the first-order reconstructed matrix of acoustic features by computing the cosine similarities between batch-based original acoustic features and batch-based concatenated original visual-textual features.

$$S = A_b(f_c(I_b, T_b))^T \quad (11)$$

where $I_b = [v_1, v_2, \dots, v_n]^T \in R^{n \times d}$ is the batch-based original visual features, $A_b = [a_1, a_2, \dots, a_n]^T \in R^{n \times d}$ is the batch-based original acoustic features, $T_b = [y_1, y_2, \dots, y_n]^T \in R^{n \times d}$ is the batch-based original textual features. $f_c(\cdot)$ stands for the concatenation operation, and $S \in R^{n \times 2n}$ is the first-order reconstructed matrix of acoustic features. n is the batch size, and d is the dimension of common feature space.

After that, we carry out the row softmax transformation on the first-order reconstructed matrix to obtain the dynamic weight factors of first-order reconstructed acoustic features. Then we use the dynamic weight factors to carry out the weighted summation on the batch-based concatenated original visual-textual features to obtain the first-order reconstructed acoustic features.

$$A_b^{1rec} = f_s(\beta \cdot S)f_c(I_b, T_b) \quad (12)$$

where $f_s(\cdot)$ stands for row softmax transformation, and β is a scaling approximator in the row softmax transformation. $A_b^{1rec} = [a_1^{1rec}, a_2^{1rec}, \dots, a_n^{1rec}]^T \in R^{n \times d}$ is the batch-based first-order reconstructed acoustic features.

Similarly, we can also obtain the first-order reconstructed visual and textual features.

$$S_1 = I_b(f_c(A_b, T_b))^T \quad (13)$$

$$I_b^{1rec} = f_s(\beta \cdot S_1)f_c(A_b, T_b) \quad (14)$$

$$S_2 = T_b(f_c(I_b, A_b))^T \quad (15)$$

$$T_b^{1rec} = f_s(\beta \cdot S_2)f_c(I_b, A_b) \quad (16)$$

where $S_1 \in R^{n \times 2n}$ is the first-order reconstructed matrix of visual features, and $S_2 \in R^{n \times 2n}$ is the first-order reconstructed matrix of textual features. $I_b^{1rec} = [v_1^{1rec}, v_2^{1rec}, \dots, v_n^{1rec}]^T \in R^{n \times d}$ is the batch-based first-order reconstructed visual features, and $T_b^{1rec} = [y_1^{1rec}, y_2^{1rec}, \dots, y_n^{1rec}]^T \in R^{n \times d}$ is the batch-based first-order reconstructed textual features.

The second feature reconstruction is similar to the first feature reconstruction. We concatenate batch-based first-order reconstructed visual and textual features to obtain batch-based concatenated first-order reconstructed visual-textual features. Then we can obtain the second-order reconstructed matrix of acoustic features by computing the cosine similarities between batch-based first-order reconstructed acoustic features and batch-based concatenated first-order reconstructed visual-textual features.

$$S' = A_b^{1rec}(f_c(I_b^{1rec}, T_b^{1rec}))^T \quad (17)$$

where $S' \in R^{n \times 2n}$ is the second-order reconstructed matrix of acoustic features.

After that, we carry out the row softmax transformation on the second-order reconstructed matrix to obtain the dynamic weight factors of second-order reconstructed acoustic features. Then we use the dynamic weight factors to carry out the weighted summation on the batch-based concatenated first-order reconstructed visual-textual features to obtain the second-order reconstructed acoustic features.

$$A_b^{2rec} = f_s(\beta \cdot S')f_c(I_b^{1rec}, T_b^{1rec}) \quad (18)$$

where $A_b^{2rec} = [a_1^{2rec}, a_2^{2rec}, \dots, a_n^{2rec}]^T \in R^{n \times d}$ is the batch-based second-order reconstructed acoustic features.

Similarly, we can also obtain the second-order reconstructed visual and textual features.

$$S'_1 = I_b^{1rec}(f_c(A_b^{1rec}, T_b^{1rec}))^T \quad (19)$$

$$I_b^{2rec} = f_s(\beta \cdot S'_1)f_c(A_b^{1rec}, T_b^{1rec}) \quad (20)$$

$$S'_2 = T_b^{1rec}(f_c(I_b^{1rec}, A_b^{1rec}))^T \quad (21)$$

$$T_b^{2rec} = f_s(\beta \cdot S'_2)f_c(I_b^{1rec}, A_b^{1rec}) \quad (22)$$

where $S'_1 \in R^{n \times 2n}$ is the second-order reconstructed matrix of visual features, and $S'_2 \in R^{n \times 2n}$ is the second-order reconstructed matrix of textual features. $I_b^{2rec} = [v_1^{2rec}, v_2^{2rec}, \dots, v_n^{2rec}]^T \in R^{n \times d}$ is the batch-based second-order reconstructed visual features, and $T_b^{2rec} = [y_1^{2rec}, y_2^{2rec}, \dots, y_n^{2rec}]^T \in R^{n \times d}$ is the batch-based second-order reconstructed textual features.

After completing the second reconstruction, we use the original features to constrain the second-order reconstructed

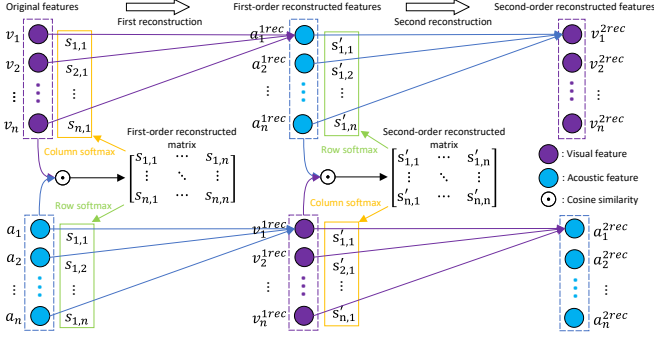


Fig. 7. Taking acoustic and visual feature reconstruction as examples to describe the process of feature reconstruction between two data modalities. (Best viewed in color.)

features. The cycle-consistency loss among three data modalities can be formulated as follows.

$$L_r = \sum_{i=1}^n \{ \|v_i^{2rec} - v_i\|_2^2 + \|a_i^{2rec} - a_i\|_2^2 + \|y_i^{2rec} - y_i\|_2^2 \} \quad (23)$$

where $v_i, a_i, y_i \in R^{d \times 1}$ are the original visual, acoustic and textual features, and $v_i^{2rec}, a_i^{2rec}, y_i^{2rec} \in R^{d \times 1}$ are the second-order reconstructed visual, acoustic and textual features. L_r is the cycle-consistency loss of our VASE model.

2) *Cycle-consistency Loss Between Two Data Modalities:* The cycle-consistency loss between two data modalities is shown in Fig. 7. Compared with the feature reconstruction among three data modalities, the feature reconstruction between two data modalities does not contain semantic information. The feature reconstruction between two data modalities is similar to the feature reconstruction among three data modalities, so we will not describe its process in detail. The first feature reconstruction process can be formulated as follows.

$$S_3 = A_b I_b^T \quad (24)$$

$$A_b^{1rec} = f_s(\beta \cdot S_3) I_b \quad (25)$$

$$I_b^{1rec} = f_s(\beta \cdot S_3^T) A_b \quad (26)$$

where $I_b = [v_1, v_2, \dots, v_n]^T \in R^{n \times d}$ is the batch-based original visual features, and $A_b = [a_1, a_2, \dots, a_n]^T \in R^{n \times d}$ is the batch-based original acoustic features. $S_3 \in R^{n \times n}$ is the first-order reconstructed matrix of acoustic and visual features. $f_s(\cdot)$ stands for row softmax transformation. β is a scaling approximator in the row softmax transformation. $I_b^{1rec} = [v_1^{1rec}, v_2^{1rec}, \dots, v_n^{1rec}]^T \in R^{n \times d}$ is the batch-based first-order reconstructed visual features, and $A_b^{1rec} = [a_1^{1rec}, a_2^{1rec}, \dots, a_n^{1rec}]^T \in R^{n \times d}$ is the batch-based first-order reconstructed acoustic features. n is the batch size, and d is the dimension of common feature space. The second reconstruction can be formulated as follows.

$$S'_3 = A_b^{1rec} (I_b^{1rec})^T \quad (27)$$

$$A_b^{2rec} = f_s(\beta \cdot S'_3) I_b^{1rec} \quad (28)$$

$$I_b^{2rec} = f_s(\beta \cdot (S'_3)^T) A_b^{1rec} \quad (29)$$

where $S'_3 \in R^{n \times n}$ is the second-order reconstructed matrix of acoustic and visual features, $I_b^{2rec} = [v_1^{2rec}, v_2^{2rec}, \dots, v_n^{2rec}]^T \in$

$R^{n \times d}$ is the batch-based second-order reconstructed visual features, and $A_b^{2rec} = [a_1^{2rec}, a_2^{2rec}, \dots, a_n^{2rec}]^T \in R^{n \times d}$ is the batch-based second-order reconstructed acoustic features.

After completing the second reconstruction, we also use the original features to constrain the second-order reconstructed features. The cycle-consistency loss between two data modalities can be formulated as follows.

$$L_{r_2s} = \sum_{i=1}^n \{ \|v_i^{2rec} - v_i\|_2^2 + \|a_i^{2rec} - a_i\|_2^2 \} \quad (30)$$

where $v_i, a_i \in R^{d \times 1}$ are the original visual and acoustic features, $v_i^{2rec}, a_i^{2rec} \in R^{d \times 1}$ are the second-order reconstructed visual and acoustic features, and L_{r_2s} is the cycle-consistency loss of VASE (2-stream) model, which remove the semantic information from VASE model.

D. Joint Learning

Combined with the tri-modal ranking loss, the total loss of our VASE model can be expressed as follows.

$$L_{total} = L_{tri} + \lambda_1 \cdot L_r \quad (31)$$

where L_{tri} is the tri-modal ranking loss, λ_1 is a balance weight factor, L_r is the cycle-consistency loss among three data modalities, and L_{total} is the total loss of our VASE model.

When no semantic information is added, our proposed VASE model degenerates into a 2-stream model, i.e., VASE (2-stream). Combined with the bi-modal ranking loss, the total loss of our VASE (2-stream) can be formulated as follows.

$$L_{total_2s} = L_{v2a} + \lambda_2 \cdot L_{r_2s} \quad (32)$$

where L_{v2a} is the bi-modal ranking loss in speech-image retrieval task, λ_2 is also a balance weight factor, L_{r_2s} is the cycle-consistency loss between two data modalities, and L_{total_2s} is the total loss of our VASE (2-stream) model. These two balance weight factors in our experiments are set to be same, i.e., $\lambda_1 = \lambda_2 = \lambda$.

IV. EXPERIMENTS

In order to demonstrate the effectiveness of our VASE model, we conduct the speech-image retrieval task on two related datasets: Flickr8K dataset and Places dataset. Conducted experiments include comparative experiments with the state-of-the-art methods and ablation studies. The datasets, evaluation metrics, parameter setups, quantitative results, visualization and related analysis are given in this section.

A. Datasets

1) *Flickr8K Dataset:* Built on the Flickr8K Audio Caption Corpus [4] and Flickr8K Text Caption Corpus [28], [65], the Flickr8K dataset consists of natural images, spoken captions and textual captions. These natural images are collected from the Flickr photo sharing website, and the collected images are trying to depict the actions of people or animals [4]. Each natural image is annotated with five textual captions and five spoken captions, and the five textual captions are transcribed from the corresponding five spoken captions. The

spoken and textual captions are collected on the Amazon’s Mechanical Turk platform. The dataset is divided into train, validation and test splits. The train split includes 6,000 natural images, 30,000 spoken captions and 30,000 textual captions. The validation split includes 1,000 natural images, 5,000 spoken captions and 5,000 textual captions. The test split includes 1,000 natural images, 5,000 spoken captions and 5,000 textual captions.

2) *Places Dataset*: Built on the Places Audio Caption Corpus [2], [7], [8] and Places 205 dataset [29], the Places dataset also consists of natural images, spoken captions and textual captions. The natural images are collected from the Places 205 dataset [29], and the spoken captions are collected from 2,683 unique speakers on the Amazon’s Mechanical Turk platform. Each natural image is annotated with one spoken caption and one textual caption. The textual caption is transcribed from the corresponding spoken caption by the Google ASR engine. And it includes 44,342 words vocabulary. The dataset is divided into train and validation splits. The train split contains 402,385 natural images, 402,385 spoken captions, and 402,385 textual captions. The validation split contains 1,000 natural images, 1,000 spoken captions, and 1,000 textual captions.

B. Evaluation Metrics

Following the existing work [2]–[4], [7], [8], the evaluation metrics used for image-to-speech (I2S) and speech-to-image (S2I) retrieval tasks are “r@1”, “r@5” and “r@10”, i.e., the recall rates [32]–[37] at top 1, 5, 10 results. To better measure the model’s overall performance for both I2S and S2I retrieval tasks, we also compute an additional evaluation metric “rsum” by summing all the 6 recall rates.

C. Implementation Details

The network details are summarized as follows. In speech encoder, the MFCC is obtained by transforming the speech, and its size is 40×2048 . Then the MFCC is put into speech CNN to obtain acoustic feature, which size is 1024×128 . Finally, the acoustic feature is put into a single-layer bi-directional GRU (Bi-GRU) to obtain final acoustic embedding, which size is 1024×1 . Here, we use DAVeNet or ResDAVeNet as our speech CNN. In image encoder, we use vgg16 or resnet50 as our image CNN, and both of them are pretrained on the imagenet dataset. The size of final visual embedding is 1024×1 . In text encoder, we use single-layer GRU to process the textual information. The dimension of the word embedding is 300, and the size of final semantic embedding is 1024×1 . In addition, the full connected layer (FC) is a single-layer network. The margin α in tri-modal ranking loss is 0.2. The balance weight factor λ is used to adjust the contributions of tri-modal ranking loss and cycle-consistency loss, and it is commonly set as 0.05. In feature reconstruction process, the scaling approximator β is an important hyperparameter to adjust contributions of matching and non-matching data samples, and it is commonly set as 2.0 or 4.0. In addition, the computation complexity of our VASE model is shown as Table I.

TABLE I
THE COMPUTATION COMPLEXITY OF OUR VASE MODEL.

| Model | Backbones | Params (M) | FLOPs (G) |
|-------|---------------------|------------|-----------|
| VASE | vgg16/DAVeNet | 164.92 | 23.09 |
| | resnet50/ResDAVeNet | 80.78 | 12.33 |

The hardware platform and training details are summarized as follows. All our experiments are conducted on ubuntu 16.04 system of dgx-1 server, and the GPU we use is Tesla P100 (16 G). Based on the Pytorch library, we implement the proposed model. We use Adam [64] to train our proposed model with an initial learning rate of 0.0002 (first stage) or 0.00002 (second stage). The learning rate needs to be divided by 10 in every 15 epochs in the training procedure. Because we use both matching and non-matching data samples in the feature reconstruction process, the batch size is also an important hyperparameter to affect final performance, and it is commonly set as 32. Besides, we use a two-stage training strategy to update the proposed model. In the first stage, we fix the weights of image CNN (pretrained on the imagenet dataset). The weights of other modules are randomly initialized. The first training procedure terminates after 100 epochs on the Flickr8K dataset, and terminates after 30 epochs on the Places dataset. In the second stage, the weights of our model are initialized by the best model in the first stage. Besides, we release the weights of image CNN (finetune the image CNN). The second training procedure terminates after 100 epochs on the Flickr8K dataset, and terminates after 40 epochs on the Places dataset.

The testing details are summarized as follows. First, we use the image encoder and speech encoder of the trained model to obtain visual and acoustic embeddings in common feature space, respectively. Then, we compute the cosine similarities between visual and acoustic embeddings. The retrieved instances are sorted according to the cosine similarities. The testing procedure are shown on the bottom figure of Fig. 3.

D. Comparison with the State-of-the-art Methods

In this section, we conduct speech-image retrieval experiments on the Flickr8K and Places datasets, and compare our results with existing related methods. On the Flickr8K dataset, we use the vgg16 [25] and DAVeNet [2] as the backbones⁴ of our VASE model, and compare our results with related state-of-the-art models, as shown in Table II. On the Places dataset, we not only use the vgg16 and DAVeNet as the backbones of our VASE model, but also use resnet50 and ResDAVeNet, and compare our results with related state-of-the-art methods. The results are shown in Table III and Table IV, respectively.

1) *Results on the Flickr8K Dataset*: The speech-image retrieval results on the Flickr8K dataset are shown in Table II. All the results are obtained on the test split, and their corresponding models are trained on the train split.

This table demonstrates the results of some local fine-grained matching speech-image retrieval models, such as Spectrogram CNN [4], SISA (P) [2], MISA (P) [2], SIMA

⁴On the Flickr8K dataset, the existing methods do not use resnet50 [30] and ResDAVeNet [3] as the backbones, so our VASE model only uses vgg16 and DAVeNet as the backbones.

TABLE II

COMPARISON RESULTS OF SPEECH-IMAGE RETRIEVAL ON THE FLICKR8K DATASET. (VGG16/DAVENET)

| Model | S2I (%) | | | I2S (%) | | | rsum (%) |
|---------------------|------------|-------------|-------------|------------|-------------|-------------|--------------|
| | r@1 | r@5 | r@10 | r@1 | r@5 | r@10 | |
| Spectrogram CNN [4] | — | — | 17.9 | — | — | 24.3 | — |
| SISA (P) [2] | 6.2 | 19.3 | 29.1 | 7.4 | 24.9 | 37.6 | 124.5 |
| MISA (P) [2] | 6.6 | 20.1 | 29.7 | 6.9 | 22.7 | 33.5 | 119.5 |
| SIMA (P) [2] | 4.4 | 15.1 | 23.8 | 6.9 | 20.2 | 32.3 | 102.7 |
| VASE (2-stream) | 4.6 | 16.1 | 25.6 | 8.1 | 21.8 | 32.6 | 108.8 |
| VASE | 6.0 | 20.4 | 31.2 | 8.0 | 26.1 | 38.9 | 130.6 |

TABLE III

COMPARISON RESULTS OF SPEECH-IMAGE RETRIEVAL ON THE PLACES DATASET. (VGG16/DAVENET)

| Model | S2I (%) | | | I2S (%) | | | rsum (%) |
|---------------------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|
| | r@1 | r@5 | r@10 | r@1 | r@5 | r@10 | |
| SISA (P) [2] | 16.5 | 43.1 | 55.9 | 12.0 | 36.3 | 50.6 | 214.4 |
| MISA (P) [2] | 20.0 | 46.9 | 60.4 | 12.7 | 37.5 | 52.8 | 230.3 |
| SIMA (P) [2] | 14.7 | 37.5 | 50.6 | 13.9 | 36.7 | 48.3 | 201.7 |
| Harwath <i>et al.</i> [8] | 16.1 | 40.4 | 56.4 | 13.0 | 37.8 | 54.2 | 217.9 |
| Harwath <i>et al.</i> [7] | 14.8 | 40.3 | 54.8 | 12.1 | 33.5 | 46.3 | 201.8 |
| VASE (2-stream) | 20.7 | 47.0 | 59.6 | 20.2 | 47.0 | 59.2 | 253.7 |
| VASE | 21.4 | 50.7 | 64.0 | 21.1 | 50.8 | 64.0 | 272.0 |

(P) [2]. And SISA (P), MISA (P) and SIMA(P) are our re-implemented local fine-grained matching models. The best performance of these three models is reported after hyperparameter finetunings. In this table, we use vgg16 and DAVEnet as the backbones of our VASE (2-stream), VASE, SISA (P), MISA (P) and SIMA (P). Besides, we use the symbol “—” to indicate some missing evaluation criteria.

From Table II, our VASE has achieved best overall performance. SISA (P) is the second best method among all compared methods. Compared with SISA (P), our VASE has achieved 6.1% improvement on overall performance *rsum*. Compared with VASE, the performance of VASE (2-stream) has dropped a lot, which means that semantic information corresponding to acoustic data is beneficial for the alignment between image and speech. However, it still has surpassed Spectrogram CNN⁵ and SIMA (P) in overall performance.

2) *Results on the Places Dataset*: The speech-image retrieval results on the Places dataset are shown in Table III and Table IV, respectively. The results in these two tables are obtained on the validation split, and their corresponding models are trained on the train split [39].

In Table III, we use vgg16 [25] and DAVEnet [2] as the backbones of our proposed models, and compare our results with several related state-of-the-art methods, including global coarse-grained matching methods and local fine-grained matching methods. The global coarse-grained matching methods include Harwath *et al.* [8] and Harwath *et al.* [7], and their results are directly copied from [2]. The local fine-grained matching methods contain SISA (P) [2], MISA (P) [2] and SIMA (P) [2].

From this table, our VASE and VASE (2-stream) have achieved top-2 performance in overall performance. MISA (P) is the third best method among all compared methods. Compared with MISA (P) [2], VASE (2-stream) outperforms it in five individual evaluation metrics except for *r@10* (S2I), and

⁵To the best of our knowledge, this is likely the only method conducted speech-image retrieval task on the Flickr8K dataset.

TABLE IV

COMPARISON RESULTS OF SPEECH-IMAGE RETRIEVAL ON THE PLACES DATASET. (RESNET50/RESDAVENET)

| Model | S2I (%) | | | I2S (%) | | | rsum (%) |
|-----------------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|
| | r@1 | r@5 | r@10 | r@1 | r@5 | r@10 | |
| Random [3] | 14.7 | 37.5 | 51.2 | 9.9 | 32.8 | 45.2 | 191.3 |
| Natural Sounds [3] | 26.8 | 54.5 | 68.4 | 21.1 | 52.8 | 66.0 | 289.6 |
| ImageNet/AudioSet [3] | 27.6 | 58.4 | 71.6 | 21.8 | 55.1 | 69.0 | 303.5 |
| VASE(2-stream) | 30.6 | 61.4 | 73.4 | 30.5 | 60.8 | 72.5 | 329.2 |
| VASE | 35.3 | 66.5 | 78.3 | 33.5 | 68.7 | 78.5 | 360.8 |

TABLE V

THE ABLATION STUDY ON THE FLICKR8K DATASET. (VGG16/DAVENET)

| Model | S2I (%) | | | I2S (%) | | | rsum (%) |
|-------------|------------|-------------|-------------|------------|-------------|-------------|--------------|
| | r@1 | r@5 | r@10 | r@1 | r@5 | r@10 | |
| base | 4.4 | 16.2 | 25.3 | 7.1 | 21.4 | 31.1 | 105.5 |
| base+cc | 4.6 | 16.1 | 25.6 | 8.1 | 21.8 | 32.6 | 108.8 |
| base+aux | 5.6 | 19.3 | 28.7 | 7.4 | 24.2 | 35.4 | 120.6 |
| base+cc+aux | 6.0 | 20.4 | 31.2 | 8.0 | 26.1 | 38.9 | 130.6 |

exceeds it by 23.4% in the overall evaluation metric *rsum*. In evaluation metric *r@10* (S2I), it is only 0.8% lower than MISA (P) [2]. Furthermore, VASE greatly outperforms MISA(P) [2] in all six individual evaluation metrics, and exceeds it by 41.7% in the overall evaluation metric *rsum*.

In Table IV, we use resnet50 [30] and ResDAVENet [3] as the backbones of our proposed models, and compare our results with some related state-of-the-art methods, such as Random [3], Natural Sounds [3] and ImageNet/AudioSet [3]. These three methods use the same SISA-SHN [3] framework, but their corresponding pre-trained ways are different. Random means that both image and speech branches of SISA-SHN are randomly initialized, Natural Sounds means that only the speech branch is pretrained on the AudioSet, and ImageNet/AudioSet means that the image branch is pretrained on the ImageNet and the speech branch is pretrained on the AudioSet.

From this table, we can observe that ImageNet/AudioSet is the best one among these three methods. Compared with ImageNet/AudioSet [3], VASE (2-stream) outperforms it in all six individual evaluation metrics, and exceeds it by 25.7% in the overall evaluation metric *rsum*. Furthermore, VASE can achieve better performance than VASE (2-stream), and it greatly outperforms ImageNet/AudioSet [3] in all six individual evaluation metrics, and exceeds it by 57.3% in the overall evaluation metric *rsum*.

Overall, both VASE and VASE (2-stream) have surpassed the existing methods in most of six individual evaluation metrics and the overall evaluation metric in both vgg16/DAVENet and resnet50/ResDAVENet backbones. In general, the local fine-grained matching methods are superior to the global coarse-grained matching methods. Although VASE and VASE (2-stream) are global coarse-grained matching methods, they have reached or even exceeded the existing local fine-grained matching methods.

Compared with VASE (2-stream), VASE obtains better performance. Because semantic information corresponding to acoustic data can further enhance the alignment between image and speech. In addition, it can be observed that the performance of speech-to-image (S2I) retrieval is overall higher than the performance of image-to-speech (I2S) retrieval for both existing global coarse-grained matching methods and

TABLE VI
THE ABLATION STUDY ON THE PLACES DATASET.

| Model | vgg16/DAVEnet | | | | | | | resnet50/ResDAVEnet | | | | | | |
|-------------|---------------|-------------|-------------|-------------|-------------|-------------|--------------|---------------------|-------------|-------------|-------------|-------------|-------------|--------------|
| | S2I (%) | | | I2S (%) | | | rsum (%) | S2I (%) | | | I2S (%) | | | 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 | |
| base | 18.1 | 44.1 | 56.8 | 18.8 | 44.2 | 57.3 | 239.3 | 27.7 | 59.9 | 72.7 | 29.5 | 61.0 | 71.7 | 322.5 |
| base+cc | 20.7 | 47.0 | 59.6 | 20.2 | 47.0 | 59.2 | 253.7 | 30.6 | 61.4 | 73.4 | 30.5 | 60.8 | 72.5 | 329.2 |
| base+aux | 22.5 | 49.9 | 61.8 | 22.5 | 49.1 | 62.0 | 267.8 | 33.9 | 65.3 | 75.5 | 32.1 | 64.8 | 76.1 | 347.7 |
| base+cc+aux | 21.4 | 50.7 | 64.0 | 21.1 | 50.8 | 64.0 | 272.0 | 35.3 | 66.5 | 78.3 | 33.5 | 68.7 | 78.5 | 360.8 |

TABLE VII
THE β HYPER-PARAMETER ANALYSIS OF OUR VASE (2-STREAM) MODEL ON THE PLACES DATASET.

| β | vgg16/DAVEnet | | | | | | | resnet50/ResDAVEnet | | | | | | |
|---------|---------------|-------------|-------------|-------------|-------------|-------------|--------------|---------------------|-------------|-------------|-------------|-------------|-------------|--------------|
| | S2I (%) | | | I2S (%) | | | rsum (%) | S2I (%) | | | I2S (%) | | | 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 | |
| 2.0 | 19.0 | 44.9 | 56.3 | 18.5 | 44.9 | 56.6 | 240.2 | 28.7 | 59.9 | 71.7 | 27.0 | 60.0 | 72.3 | 319.6 |
| 4.0 | 16.2 | 42.0 | 54.3 | 15.9 | 41.2 | 53.7 | 223.3 | 30.4 | 60.6 | 73.9 | 27.0 | 61.1 | 72.3 | 325.3 |
| 6.0 | 18.4 | 45.1 | 57.9 | 19.3 | 45.3 | 56.1 | 242.1 | 30.6 | 61.4 | 73.4 | 30.5 | 60.8 | 72.5 | 329.2 |
| 8.0 | 18.5 | 44.6 | 56.3 | 20.2 | 44.4 | 55.6 | 239.6 | 28.3 | 61.2 | 73.0 | 27.9 | 60.5 | 72.9 | 323.8 |
| 10.0 | 20.5 | 46.6 | 57.6 | 20.7 | 47.9 | 57.7 | 251.0 | 30.6 | 59.4 | 71.6 | 30.7 | 61.4 | 70.8 | 324.5 |
| 12.0 | 20.7 | 47.0 | 59.6 | 20.2 | 47.0 | 59.2 | 253.7 | 28.4 | 59.8 | 71.9 | 28.3 | 59.7 | 72.1 | 320.2 |

TABLE VIII
THE β HYPER-PARAMETER ANALYSIS OF OUR VASE MODEL ON THE PLACES DATASET.

| β | vgg16/DAVEnet | | | | | | | resnet50/ResDAVEnet | | | | | | |
|---------|---------------|-------------|-------------|-------------|-------------|-------------|--------------|---------------------|-------------|-------------|-------------|-------------|-------------|--------------|
| | S2I (%) | | | I2S (%) | | | rsum (%) | S2I (%) | | | I2S (%) | | | 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 | |
| 2.0 | 21.4 | 50.7 | 64.0 | 21.1 | 50.8 | 64.0 | 272.0 | 33.2 | 65.6 | 76.4 | 33.6 | 65.4 | 76.1 | 350.3 |
| 4.0 | 23.3 | 50.0 | 62.7 | 20.3 | 49.6 | 62.9 | 268.8 | 35.3 | 66.5 | 78.3 | 33.5 | 68.7 | 78.5 | 360.8 |
| 6.0 | 19.4 | 45.3 | 59.0 | 18.4 | 45.1 | 58.9 | 246.1 | 32.2 | 65.4 | 77.5 | 31.6 | 65.4 | 77.4 | 349.5 |
| 8.0 | 20.9 | 46.2 | 60.1 | 17.3 | 46.9 | 59.3 | 250.7 | 34.2 | 64.5 | 76.0 | 29.7 | 65.5 | 76.9 | 346.8 |
| 10.0 | 19.4 | 46.0 | 59.4 | 17.9 | 46.7 | 59.8 | 249.2 | 35.9 | 66.0 | 77.2 | 30.6 | 66.1 | 75.8 | 351.6 |
| 12.0 | 19.3 | 47.9 | 58.6 | 19.5 | 46.9 | 59.4 | 251.6 | 33.6 | 65.9 | 76.0 | 31.2 | 64.2 | 76.6 | 347.5 |

local fine-grained matching methods, which can lead to the imbalance problem of retrieval performance. From Table III and Table IV, our results are more balanced in speech-to-image and image-to-speech retrieval directions. Therefore, our VASE can solve the unbalanced performance problem to some extent.

E. Ablation Study

The auxiliary alignment loss and cycle-consistency loss are two essential modules of our proposed model. The auxiliary alignment loss is one part of tri-modal ranking loss. To demonstrate their effectiveness, we have conducted related ablation studies on the Flickr8K and Places datasets. We have evaluated multiple variants of our VASE model.

- **base**: We remove both auxiliary alignment loss and cycle-consistency loss from our full VASE model. This variant consists of visual and acoustic branches, and its total loss only includes direct alignment loss.

- **base+cc**: We only remove the auxiliary alignment loss from our full VASE model. This variant consists of visual and acoustic branches, and its total loss includes direct alignment loss and cycle-consistency loss between two data modalities. It is also called VASE (2-stream), and does not use semantic information corresponding to acoustic data.

- **base+aux**: We only remove the cycle-consistency loss from our full VASE model. This variant consists of visual, acoustic and semantic branches, and its total loss includes tri-modal ranking loss (direct alignment loss and auxiliary alignment loss). Here, we use semantic information corresponding to acoustic data.

- **base+cc+aux**: Our full VASE model, as shown in Fig. 3. This variant consists of visual, acoustic and semantic branches,

and its total loss includes tri-modal ranking loss and cycle-consistency loss among three data modalities.

We have evaluated these variants of our VASE model on the Flickr8K and Places datasets. In Table V, we choose vgg16 and DAVEnet as the backbones of these variants, and these results are obtained on the test split of Flickr8K dataset. In Table VI, we not only choose vgg16 and DAVEnet as the backbones, but also choose resnet50 and ResDAVEnet. The results of Table VI are obtained on the validation split of Places dataset.

From these two tables, both “base+cc” and “base+aux” have exceeded the “base”, which can demonstrate the effectiveness of our auxiliary alignment loss and cycle-consistency loss. The overall performance of “base+aux” is higher than “base+cc”, which means the auxiliary alignment loss is more effective than the cycle-consistency loss in our VASE model. Compared with “base+cc” and “base+aux”, the “base+cc+aux” can obtain further improvement. This means that our auxiliary alignment loss and cycle-consistency loss can promote each other, so it is more effective to combine them. In addition, our auxiliary alignment loss and cycle-consistency loss are not only applicable to vgg16/DAVEnet, but also applicable to resnet50/ResDAVEnet, so they have good generalization ability. Compared with the vgg16/DAVEnet backbones, the resnet50/ResDAVEnet backbones can obtain better performance.

F. Hyperparameter Analysis

We choose the scaling approximator β of row softmax transformation in the feature reconstruction as the hyperparameter to be analyzed on the Places dataset. The results

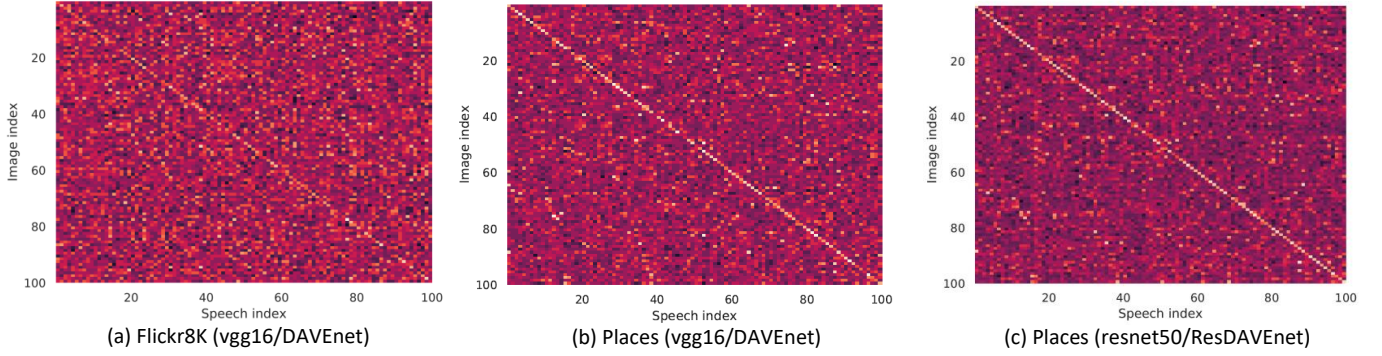


Fig. 8. Similarity matrices of top 100 image-speech pairs from the Flickr8K test set and Places validation set. The matched images and speeches have the same numerical indexes. Diagonal lines of three similarity matrices indicate that matched data samples have higher correlation than mismatched ones. (Best viewed in color.)

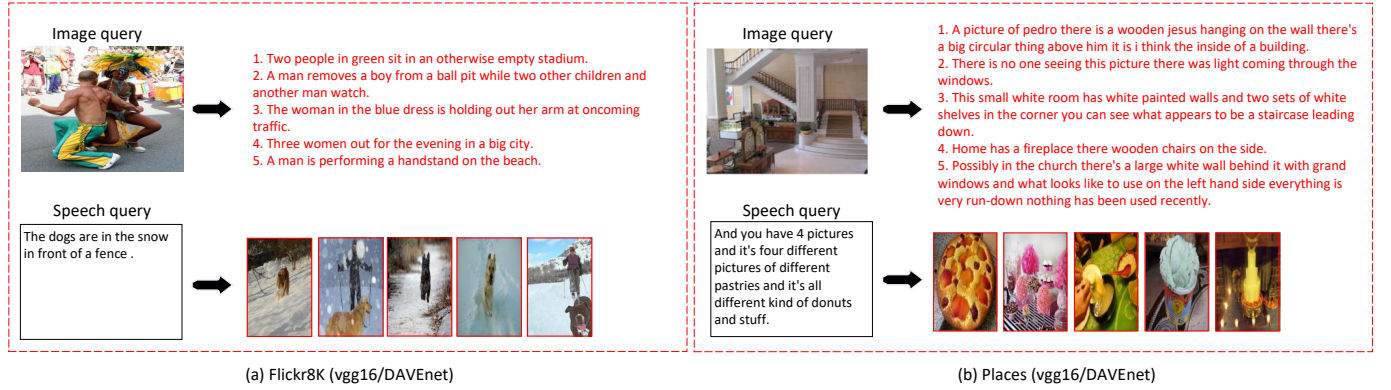


Fig. 9. Failure cases of our VASE model on Flickr8k and Places datasets. Here, we use transcribed texts to represent corresponding speeches. Given one query, the top-5 retrieved candidates are shown. The red fonts and boxes indicate the unmatched retrieval instances. Although these are failed examples, the retrieval results are reasonable and have great correlation with the query. (Best viewed in color.)

of VASE (2-stream) and VASE are shown in Table VII and Table VIII, respectively. By observing these two tables, we can find that the optimal β values of VASE (2-stream) are generally larger than that of VASE. Because there are fewer matching data samples of other modalities in the feature reconstruction process of VASE (2-stream). Increasing the value of β properly can increase the contributions of matching data samples and reduce the contributions of non-matching ones.

G. Qualitative Analysis

In addition to the quantitative results, inspired by [31] and [40], we further demonstrate the matching scores, which are obtained by computing the similarities between the visual and acoustic features. We select the top-100 speech-image pairs from the test split of the Flickr8K dataset and the validation split of the Places dataset, respectively, and calculate the similarities between images and speeches. The higher similarity between the image and speech pair, the brighter the corresponding position. As shown in Fig. 8, the matching speech-image pairs (with the same numerical index) have higher similarities than the non-matching ones. As can be seen from Fig. 8, there are some other bright positions except for the diagonal positions. One possible reason is that there are some similar speech-image pairs which are not marked as matching pairs. To demonstrate this reason, we also show some failure cases, which are shown in Fig. 9. From these failure cases, we can find that there are some data samples that are semantically

similar to the given queries but not marked as the matching pairs, which verifies our hypothesis.

In addition, to demonstrate our VASE model more intuitively, we compare it with MISA (P) [2], as shown in Fig. 10. The top example shows the speech-to-image retrieval task. From the top-6 retrieved images, both our VASE and MISA (P) can capture the key semantic information, such as “people” and “river”. The matching image ranks first in our VASE and sixth in the MISA (P). The bottom example shows the image-to-speech retrieval task. From the top-6 retrieved speeches, both our VASE and MISA (P) can grasp the key visual semantic information, such as “people”, “room” and “table”. The matching speech ranks first in our VASE and fourth in the MISA (P). Therefore, our VASE obtains better performance than MISA (P) in both image-to-speech and speech-to-image retrieval tasks.

H. Discussion

In this paper, we mainly focus on the cross-modal heterogeneous issue between image and speech for speech-image retrieval. Compared with global coarse-grained matching methods [7], [8] and local fine-grained matching methods [2]–[4], we leverage semantic information corresponding to acoustic data to introduce an auxiliary alignment, which bridges the modality gap between image and speech. In addition, we introduce a tri-modal reconstruction-based cycle-consistency loss to further alleviate the modality gap between visual and

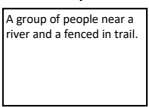













| Query | | Top-6 retrieved instances | | | | | |
|--|----------|---|---|---|--|---|---|
| Given speech  | MISA (P) |  |  |  |  |  |  |
| | VASE |  |  |  |  |  |  |
| Given image  | MISA (P) | <div> <div>Pictures of event center setup for wedding with tables and tablecloths.</div> <div>Is an image of a group of people seated at tables inside of a restaurant there are lots of people there and they appear to be having a good time.</div> <div>In this photograph you can see many people sitting down at a table looks like a restaurant.</div> <div>People are meeting in a large room there are large tables with food covered by tin foil there are people sitting along long rows of tables.</div> <div>This is an image of a cafeteria there are many tables with girl sitting and enjoying their meal.</div> <div>This is a group of people that sing at tables they are working on a project assembling stuff.</div> </div> | | | | | |
| | VASE | <div> <div>People are meeting in a large room there are large tables with food covered by tin foil there are people sitting along long rows of tables.</div> <div>Is an image of a group of people seated at tables inside of a restaurant there are lots of people there and they appear to be having a good time.</div> <div>A man is at the head of a class or conference room he appears to be teaching he is pointing to a display being projected against the wall there are students in laptop sitting at the table.</div> <div>Is a picture taking the inside of a restaurant you can see asian people sitting around at small tables i think they're reading books are playing board games i can't really tell it looks kind of fun.</div> <div>Large number of people in a conference room there's a large screen with the presentation.</div> <div>A large lobby with a mosaic of a greek god in a chariot with white horses there are people standing at the front of the counter in the lobby and there are people helping them behind the counter.</div> </div> | | | | | |

Fig. 10. Speech-image retrieval examples of our VASE model and MISA (P) [2] model on the Places dataset. Here, we use transcribed texts to represent corresponding speeches. The matching instances are surrounded by green boxes while the non-matching instances are surrounded by red boxes. A query only has one matched retrieval instance. (Best viewed in color.)

acoustic modalities, which is quite new in current vision and language area. Compared with previous feature reconstruction methods [31], [63], our proposed method not only uses matching data samples, but also uses non-matching data samples. It can alleviate similar but non-matching data samples being far away from each other in the common feature space. To balance the contributions of matching and non-matching data samples in the feature reconstruction process, we introduce dynamic weight factors, which are related to the similarities between original data samples and data samples participating in the reconstruction process. Extensive experiments have demonstrated the effectiveness of our proposed model.

However, compared with local fine-grained matching methods, the direct alignment in our proposed model only includes global coarse-grained matching relationship, which might not model the direct alignment very well. To address this issue, we will add local fine-grained matching relationship into the direct alignment between image and speech in future work. Besides, our proposed model needs semantic information in the training stage, so it is somewhat limited for some speech-image retrieval scenarios containing semantic information. To address this problem, we will combine the ASR technology into our model in future work.

V. CONCLUSION

In this paper, we have proposed a VASE model to deal with the cross-modal heterogeneous issue in speech-image retrieval. First, we propose a tri-modal ranking loss to bridge the modality gap between image and speech by taking advantage of semantic information corresponding to the acoustic data. Second, we introduce a reconstruction-based cycle-consistency loss to further alleviate the modality gap. Extensive experiments on the Flickr8K and Places datasets have demonstrated the effectiveness of our VASE model for the speech-image retrieval task. In future work, we will explore how to integrate

the ASR technology and local fine-grained matching relationship into our proposed model.

REFERENCES

- [1] Y. Huang and L. Wang, "ACMM: Aligned cross-modal memory for few-shot image and sentence matching," in *Proc. Int. Conf. Comput. Vision*, 2019, pp. 5774–5783.
- [2] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. R. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 649–665.
- [3] D. Harwath, A. Recasens, D. Surís, G. Chuang, A. Torralba, and J. R. Glass, "Jointly discovering visual objects and spoken words from raw sensory input," *Int. J. Comput. Vision*, vol. 128, no. 3, pp. 620–641, Mar. 2020.
- [4] D. Harwath and J. R. Glass, "Deep multimodal semantic embeddings for speech and images," in *Proc. Autom. Speech Recog. Understanding Workshop*, 2015, pp. 237–244.
- [5] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2014, pp. 580–587.
- [6] S. Bengio and G. Heigold, "Word embeddings for speech recognition," in *INTERSPEECH*, 2014, pp. 1053–1057.
- [7] D. Harwath, A. Torralba, and J. R. Glass, "Unsupervised learning of spoken language with visual context," in *Proc. Adv. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1858–1866.
- [8] D. Harwath and J. R. Glass, "Learning word-like units from joint audio-visual analysis," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 506–517.
- [9] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," in *Proc. Adv. Conf. Neural Inf. Process. Syst.*, 2013, pp. 2121–2129.
- [10] J. Shen and N. Robertson, "BBAS: Towards large scale effective ensemble adversarial attacks against deep neural network learning," *Inf. Sci.*, vol. 569, pp. 469–478, Aug. 2021.
- [11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Conf. Neural Inf. Process. Syst.*, 2014.
- [12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, and A. Rabinovich, "Going deeper with convolutions," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 1–9.
- [13] L. Xu, J. S. Ren, C. Liu and J. Jia, "Deep convolutional neural network for image deconvolution," in *Proc. Adv. Conf. Neural Inf. Process. Syst.*, 2014.
- [14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2818–2826.

- [15] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2017, pp. 4700–4708.
- [16] J. R. Kiros, R. Salakhutdinov, and R. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," in 2014, arXiv:1411.2539v1.
- [17] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives," in *Proc. Brit. Mach. Vision Conf.*, 2018.
- [18] A. Karpathy, A. Joulin, and F.-F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Proc. Adv. Conf. Neural Inf. Process. Syst.*, 2014, pp. 1889–1897.
- [19] K. H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vision*, 2018, pp. 201–216.
- [20] Y. Huang, Q. Wu, C. Song, and L. Wang, "Learning semantic concepts and order for image and sentence matching," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6163–6171.
- [21] K. Li, Y. Zhang, K. Li, Y. Li, and Y. Fu, "Visual semantic reasoning for image-text matching," in *Proc. Int. Conf. Comput. Vision*, 2019, pp. 4654–4662.
- [22] Y. Aytar, C. Vondrick, and A. Torralba, "See, hear, and read: Deep aligned representations," in 2017, arXiv:1706.00932v1.
- [23] L. Castrejón, Y. Aytar, C. Vondrick, H. Pirsiavash, and A. Torralba, "Learning aligned cross-modal representations from weakly aligned data," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2940–2949.
- [24] Y. Aytar, L. Castrejón, C. Vondrick, H. Pirsiavash, and A. Torralba, "Cross-Modal Scene Networks," *IEEE Trans. Pattern Anal. and Mach. Intell.*, vol. 40, no. 10, pp. 2303–2314, Oct. 2018.
- [25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015.
- [26] K. Cho, B. V. Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2014, pp. 1724–1734.
- [27] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.
- [28] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proc. NAACL HLT Workshop*, 2010, pp. 139–147.
- [29] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014.
- [30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 770–778.
- [31] Y. Liu, Y. Guo, L. Liu, E. M. Bakker, and M. S. Lew, "CycleMatch: A cycle-consistent embedding network for image-text matching," *Pattern Recognit.*, vol. 93, pp. 365–379, Sept. 2019.
- [32] A. T. Nguyen, T. T. Nguyen, T. N. Nguyen, D. Lo, and C. Sun, "Duplicate bug report detection with a combination of information retrieval and topic modeling," in *Proc. Int. Conf. Automated Softw. Eng.*, 2012, pp. 70–79.
- [33] J. N. och Dag, V. Gervasi, S. Brinkkemper, and B. Regnell, "Speeding up requirements management in a product software company: Linking customer wishes to product requirements through linguistic engineering," in *Proc. Int. Requirements Eng. Conf.*, 2004, pp. 283–294.
- [34] P. Runeson, M. Alexandersson, and O. Nyholm, "Detection of duplicate defect reports using natural language processing," in *Proc. Int. Conf. Softw. Eng.*, 2007, pp. 499–510.
- [35] C. Sun, D. Lo, S. C. Khoo, and J. Jiang, "Towards more accurate retrieval of duplicate bug reports," in *Proc. Int. Conf. Automated Softw. Eng.*, 2011, pp. 253–262.
- [36] X. Wang, L. Zhang, T. Xie, J. Anvik, and J. Sun, "An approach to detecting duplicate bug reports using natural language and execution information," in *Proc. Int. Conf. Softw. Eng.*, 2008, pp. 461–470.
- [37] F. Thung, D. Lo, and J. Lawall, "Automated library recommendation," in *Proc. Work. Conf. Reverse Eng.*, 2013, pp. 182–191.
- [38] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 207–218, Apr. 2014.
- [39] J. Yu, W. Zhang, Y. Lu, Z. Qin, Y. Hu, J. Tan, and Q. Wu, "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3196–3209, Dec. 2020.
- [40] Y. Huang, W. Wang, and L. Wang, "Unconstrained multimodal multi-label learning," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1923–1935, Nov. 2015.
- [41] W. Guo, Y. Zhang, X. Wu, J. Yang, X. Cai, and X. Yuan, "Re-attention for visual question answering," in *Proc. Assoc. Advancement Artif. Intell.*, 2020, pp. 91–98.
- [42] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.
- [43] D. Gao, L. Jin, B. Chen, M. Qiu, P. Li, Y. Wei, Y. Hu, and H. Wang, "Fashionbert: Text and image matching with adaptive loss for cross-modal retrieval," in *Proc. ACM SIGIR*, 2020, pp. 2251–2260.
- [44] T. Matsubara, "Target-oriented deformation of visual-semantic embedding space," *IEICE Trans. Inf. Syst.*, vol. 104, no. 1, pp. 24–33, Jan. 2021.
- [45] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y. D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM TOMM*, vol. 16, no. 2, pp. 1–23, June 2020.
- [46] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vision*, 2015, pp. 2425–2433.
- [47] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *Comput. Vision Image Understanding*, vol. 163, pp. 21–40, Oct. 2017.
- [48] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6077–6086.
- [49] W. Wang, Y. Huang, and L. Wang, "Long video question answering: A Matching-guided Attention Model," *Pattern Recognit.*, vol. 102, June 2020.
- [50] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 1–36, Feb. 2019.
- [51] L. Kaiser and S. Bengio, "Can active memory replace attention," in *Proc. Adv. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 3781–3789.
- [52] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2015, pp. 3156–3164.
- [53] J. Wang, W. Wang, Y. Huang, L. Wang, and T. Tan, "Hierarchical memory modelling for video captioning," in *Proc. ACM MM*, 2018, pp. 63–71.
- [54] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 12655–12663.
- [55] C. Liu, Z. Mao, T. Zhang, H. Xie, B. Wang, and Y. Zhang, "Graph structured network for image-text matching," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 10921–10930.
- [56] H. Wang, Y. Zhang, Z. Ji, Y. Pang, and L. Ma, "Consensus-aware visual-semantic embedding for image-text matching," in *Proc. Eur. Conf. Comput. Vision*, 2020, pp. 18–34.
- [57] Y. Chen and X. Lu, "A deep hashing technique for remote sensing image-sound retrieval," *Remote Sens.*, vol. 12, no. 1, pp. 84, Dec. 2019.
- [58] G. Mao, Y. Yuan, and L. Xiaoqiang, "Deep cross-modal retrieval for remote sensing image and audio," in *Proc. Pattern Recognit. Remote Sens. Workshops*, 2018, pp. 1–7.
- [59] M. Guo, C. Zhou, and J. Liu, "Jointly learning of visual and auditory: A new approach for RS image and audio cross-modal retrieval," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 11, pp. 4644–4654, Nov. 2019.
- [60] D. Suris, A. Duarte, A. Salvador, J. Torres and X. Giró-i-Nieto, "Cross-modal embeddings for video and audio retrieval," in *Proc. Eur. Conf. Comput. Vision Workshops*, 2018.
- [61] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. ACM MM*, 2003, pp. 604–611.
- [62] D. Dwibedi, Y. Aytar, J. Tompson, P. Sermanet, and A. Zisserman, "Temporal cycle-consistency learning," in *Proc. Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 1801–1810.
- [63] A. Zheng, M. Hu, B. Jiang, Y. Huang, Y. Yan, and B. Luo, "Adversarial-metric learning for audio-visual cross-modal matching," *IEEE Trans. Multimedia*, 2021.
- [64] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in 2014, arXiv:1412.6980.

- [65] M. Hodosh, P. Young and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *J. Artif. Intell. Res.*, vol. 47, pp. 853–899, 2013.



Wenlong Cheng received his BSc degree in Jilin University (JLU) in 2014. He is now a PhD candidate working in the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing, China. His research interests include visual question asking and cross-modal retrieval.



Liang Wang received the BEng and MEng degrees from Anhui University, in 1997 and 2000, respectively, and the PhD degree from the Institute of Automation, Chinese Academy of Sciences (CASIA), in 2004. From 2004 to 2010, he was a research assistant at Imperial College London, United Kingdom, and Monash University, Australia, a research fellow with the University of Melbourne, Australia, and a lecturer with the University of Bath, United Kingdom, respectively. Currently, he is a full professor of the Hundred Talents Program at the National Lab of Pattern Recognition, CASIA. His major research interests include machine learning, pattern recognition, and computer vision. He has widely published in highly ranked international journals such as the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and the *IEEE Transactions on Image Processing*, and leading international conferences such as CVPR, ICCV, and ECCV. He is a fellow of the IEEE and the IAPR.



internship of MSRA.

Wei Tang received the BSc degree from Harbin Engineering University (HEU) in 2013. Currently, he is a PhD candidate in the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA) under the guidance of professor Liang Wang. His research interests include deep learning and computer vision, model compression and acceleration. He has published papers in the conferences such as AAAI, CCCV. He has won the best student paper award in 2015 CCCV and the star of tomorrow award in



Yan Huang received the BSc degree from University of Electronic Science and Technology of China (UESTC) in 2012, and the PhD degree from University of Chinese Academy of Sciences (UCAS) in 2017. Since July 2017, He has joined the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA) as associate researcher. His research interests are computer vision and multimodal data analysis. He has published more than 50 papers in international journals and conferences in related fields.

Related papers have won the CVPR workshop best paper award, ICPR best student paper award, *etc.* He was selected in Beijing Science and Technology Star program and Microsoft Star Casting program. He was the Co-chair of multimodal symposiums on CVPR and ICCV. He has won the special award of president of Chinese Academy of Sciences, Excellent Doctoral Dissertation Award of Chinese society of artificial intelligence, baidu scholarship and NVIDIA Innovation Research Award.



Yiwen Luo received her bachelor's degree from Northwestern Polytechnical University (NWPU) in 2018. Currently, she is a master's student at the Institute of Artificial Intelligence and Robotics (IAIR), Xi'an Jiaotong University (XJTU). Her research interests include deep learning, computer vision, and medical image processing.