IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 26 December 2022, accepted 20 January 2023, date of publication 25 January 2023, date of current version 3 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3239858

RESEARCH ARTICLE

Cross-Modal Image Retrieval Considering Semantic Relationships With Many-to-Many Correspondence Loss

HUAYING ZHANG[1], (Student Member, IEEE), RINTARO YANAGI[1], (Student Member, IEEE), REN TOGO[2], (Member, IEEE), TAKAHIRO OGAWA[2], (Senior Member, IEEE), AND MIKI HASEYAMA[2], (Senior Member, IEEE)

[1]Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060-0814, Japan
[2]Faculty of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060-0814, Japan

Corresponding author: Miki Haseyama (mhaseyama@lmd.ist.hokudai.ac.jp)

This work was supported by the JSPS KAKENHI under Grant JP21H03456 and Grant JP20K19857.

ABSTRACT A cross-modal image retrieval that explicitly considers semantic relationships between images and texts is proposed. Most conventional cross-modal image retrieval methods retrieve the target images by directly measuring the similarities between the candidate images and query texts in a common semantic embedding space. However, such methods tend to focus on a one-to-one correspondence between a predefined image-text pair during the training phase, and other semantically similar images and texts are ignored. By considering the many-to-many correspondences between semantically similar images and texts, a common embedding space is constructed to assure semantic relationships, which allows users to accurately find more images that are related to the input query texts. Thus, in this paper, we propose a cross-modal image retrieval method that considers semantic relationships between images and texts. The proposed method calculates the similarities between texts as semantic similarities to acquire the relationships. Then, we introduce a loss function that explicitly constructs the many-to-many correspondences between semantically similar images and texts from their semantic relationships. We also propose an evaluation metric to assess whether each method can construct an embedding space considering the semantic relationships. Experimental results demonstrate that the proposed method outperforms conventional methods in terms of this newly proposed metric.

INDEX TERMS Cross-modal image retrieval, many-to-many correspondences, multimedia information retrieval, semantic similarity.

I. INTRODUCTION

With the recent spread of digital storage devices, the amount of images stored in personal databases, e.g., smartphones and personal computers, has increased [1], [2]. Therefore, a convenient and user-friendly image retrieval system is required to help users find their desired images from a huge number of images [3]. Among various image retrieval systems, image retrieval from query text (also referred to as text-to-image retrieval) is one of the most convenient retrieval methods for users. The development of text-to-image retrieval leads to various downstream applications, e.g., object retrieval using natural language [4], text-guided image manipulation [5], and visual question answering [6], [7].

Traditionally, text-to-image retrieval has been realized by labeling candidate images manually [8], referring to text-based image retrieval. Here, the candidate images in the database are assigned several text labels, and the retrieval process is performed by calculating the similarities between the input text query and the text labels [9], [10], [11]. Although such methods can realize image retrieval from a text query, a laborious labeling process is required. Recently, cross-modal image retrieval methods that can retrieve the

The associate editor coordinating the review of this manuscript and approving it for publication was Ramakrishnan Srinivasan.

target images from a database with unlabeled images have been proposed [12], [13], [14]. These methods embed the images and texts in a common semantic embedding space where the distances between the embedded features can be calculated directly [15]. These methods can achieve high accuracy from the detailed queries; however, users do not always clearly remember the specific details of the target images, which can result in ambiguous queries. For a more flexible retrieval, it is important to construct an embedding space that facilitates accurate retrieval even when ambiguous queries are input. Since an ambiguous query usually contains a wide range of meanings, it is helpful to leverage adequate information from a database. For this purpose, an embedding space where semantically similar images and texts are close is desired. By constructing such an embedding space, the wide range of meanings can be considered more accurately, and more images that are relevant to the ambiguous text query can be retrieved.

Since a query can be related to multiple images, semantically similar images utilized in the training phase are beneficial in terms of constructing the embedding space. However, most conventional cross-modal image retrieval methods ignore the distances between non-paired semantically similar images and texts, which should be close [16], [17]. These methods primarily focus on the one-to-one correspondences between images and texts predefined in general open datasets. Specifically, conventional methods follow the loss function that maximizes the similarity between predefined ground truth pairs than other samples in the embedding space, and evaluation metrics (e.g., Recall@k) that give a higher score to such an embedding space are used [18], [19], [20]. As a result, these methods do not focus on the many-to-many correspondences between semantically similar images and texts; thus, non-paired but semantically similar images and texts close will likely be distant in the embedding space [21]. In such an embedding space, even though images and texts representing an exact match may be located accurately, images that are similar to the query text are not located adequately [22], [23].

To realize image retrieval using an ambiguous query in the embedding space where semantically similar images and texts are close, the relationships between these images and text must be considered explicitly in the training phase. However, semantically similar images and texts are not predefined in general open datasets. Considering that there are many and various semantically similar images and texts, annotating all corresponding semantic relationships manually would be unreasonable. Therefore, a similarity calculation procedure that focuses on these text relationships is beneficial [24]. The key point of this procedure is to calculate the similarity between text captions and utilize the similarity as the proxy for the semantic similarity between samples. Following this procedure, the cross-modal retrieval that considers semantic relationships can be realized.

In this paper, we propose a cross-modal image retrieval method that can build an embedding space that preserves the semantic relationships among images and texts. Figure 1 illustrates the many-to-many correspondence concept compared to the one-to-one correspondence used in the conventional cross-modal retrieval method. We propose the sentence-based semantic loss function to achieve our objective. The proposed loss function utilizes the semantic relationship as a basis to construct the many-to-many correspondence in the embedding space. The semantic relationship is calculated from the similarity between text captions in the training phase. As a result, the proposed method attempts to mitigate the limitations of conventional one-to-one image retrieval methods. In addition, to evaluate the proposed method, we introduce the semantic relationship distance (SRD) metric, which evaluates whether semantic relationships are preserved.

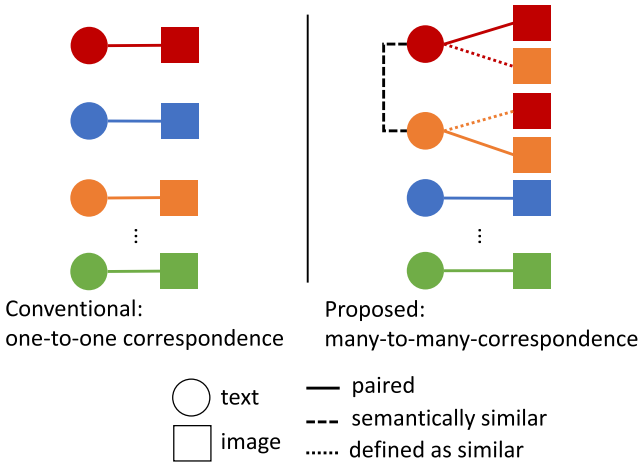


FIGURE 1. Construction of many-to-many correspondence compared to one-to-one correspondence.

Our primary contributions are summarized as follows.

- We propose a cross-modal image retrieval method that considers the semantic relationships between images and texts by minimizing the distances between semantically similar images and texts in the embedding space.
- We introduce the SRD metric to confirm whether a method constructs an embedding space in consideration of semantic relationships by comparing rankings calculated from image-text similarity and semantic similarity.

II. RELATED WORK

In the following, we review work related to cross-modal retrieval (Section II-A), semantic relationships (Section II-B), and evaluation metrics that consider semantic relationships (Section II-C).

A. CROSS-MODAL RETRIEVAL

The goal of cross-modal retrieval is to retrieve samples of one modality from a query of another modality. It is desirable for humans to retrieve images using a text query of natural language [25]; however, it is difficult to fill the semantic gap between images and texts [26]. To this end, a popular approach is to map images and texts into

a common embedding space. In early works, canonical correlation analysis was widely adopted to construct such embedding spaces [27]. With the rapid development of deep learning, convolutional neural networks (CNN) and recurrent neural networks (RNN) are frequently used to extract image and text features [28], [29]. Karpathy and Fei-Fei [30] combined CNN and RNN methods to map image and text features to a common embedding space for cross-modal retrieval. In addition, Faghri et al. [31] applied the hard negative mining strategy, which increased retrieval accuracy effectively.

Since a single global feature is not sufficiently representative for cross-modal retrieval, researchers began to match local features (e.g., objects, actions, and properties) and global features from images and texts to improve retrieval accuracy [32]. To correctly match images and texts, the attention mechanism was implemented in cross-modal retrieval to better capture semantically related local features [33], and the semantic consistency between images and texts was considered to improve the alignment between images and texts [17]. In addition to a single attention module, Song and Soleymani [34] utilized a multi-head self-attention network to exploit polysemous meanings. In addition, the graph convolutional network (GCN) has been employed in several methods to consider the relationship between local features, and these methods demonstrated good performance [35], [36].

The above methods have achieved impressive performance in retrieving a predefined ground truth image from large-scale public datasets. However, to the best of our knowledge, few existing methods consider the many-to-many correspondence between images and texts.

B. SEMANTIC RELATIONSHIP

Many conventional methods have focused on the one-to-one correspondence in the training phase by applying contrastive loss or triplet loss [37], [38]. Such methods can derive representative features and retrieve similar images; however, they do not exploit the many-to-many correspondence between images and texts. To this end, some uni-modal retrieval methods do consider the semantic relationship between samples. For example, Gordo and Larlus [16] indicated that a human-annotated text caption is semantically informative for images, and they selected images with text captions of high similarity as positive samples, and they mapped their features close in the embedding space. Despite the usage of captions, Gordo et al. considered all selected images as equally important to the loss calculation; thus, they failed to explicitly quantize the relationship between images. To consider the importance of the images with different similarities, Kim et al. [21] proposed a method that constrains the similarity between images to be consistent with the similarity between text labels. Even though the above methods performed well in unimodal image retrieval, they did not realize cross-modal image retrieval.

In the cross-modal retrieval field, the fact that text captions are not exclusively related to only a single image has also been considered recently. Li et al. [39] considered to reduce the loss caused by forcing the paired samples to be the same. However, this method did not consider to process the natural language query. Similar to our work, Yu et al. [40], Zhen et al. [41] attempted to bring semantically similar samples together in the embedding space. However, these methods required additional labels containing high-semantic information to find out the semantic relationships. To avoid the usage of the additional labels, Chun et al. [20] imported a probabilistic model to the cross-modal retrieval model, expecting queries to retrieve more semantically similar samples of another modality. The method proposed by Chun et al. [20] is somewhat similar to our proposed method; however, some differences should be highlighted. Chun et al. [20] expanded the range where samples are distributed in the embedding space. In contrast, in our method, the loss function is modified to quantize the relationship between semantically similar samples. Thus, we expect that our proposed method can construct many-to-many correspondences between samples more accurately.

C. EVALUATION METRICS CONSIDERING SEMANTIC RELATIONSHIPS

In most cross-modal retrieval works, retrieval performance is measured using the Recall@k, median rank, and mean rank evaluation metrics. These metrics can assess whether annotated ground truth image-text pairs are matched. However, these metrics focus on the one-to-one correspondence between images and texts and ignore the fact that text captions can describe multiple images in a given dataset. As a result, they do not exploit the many-to-many correspondence between images and texts. Thus, these conventional metrics cannot evaluate whether semantic relationships are preserved, and they cannot fairly assess methods when the retrieved targets are reasonably related to the query.

These metrics do not offer a fair evaluation of retrieval considering many-to-many correspondence; thus, Chun et al. [20] proposed the Plausible-Match R-Precision (PMRP) metric. The PMRP metric computes the ratio of plausibly positive samples ranked in the top-k, where plausibly positive samples are defined using pre-annotated object labels in the dataset. However, the object information is not sufficiently salient to reflect the semantic relationships between images and texts due to a lack of relationship representation between objects [16]. In addition, the PMRP metric requires a hyperparameter to compute the retrieval score, which makes it difficult to evaluate retrieval performance correctly. The evaluation metric we proposed in this paper is parameter-free and is more reliable in terms of reflecting semantic relationships between the images and texts.

In the video retrieval field, Wray et al. [24] proposed a semantic similarity calculation procedure using text captions. Inspired by Wray et al. [24], we propose a many-to-many evaluation metric based on the similarity between text

captions. Note that there are several ways to compute the similarity between sentences. In our work, we adopt the transformer-based Sentence-BERT [42] method to compute the similarity between text captions because it exhibits effective text representation abilities and fast calculation speeds.

III. PROPOSED CROSS-MODAL IMAGE RETRIEVAL METHOD

Here, we present the proposed cross-modal image retrieval method. The proposed method involves three main steps, i.e., STEP I: semantic similarity calculation; STEP II: cross-modal similarity calculation; and STEP III: loss calculation. Figure 2 shows an overview of the proposed method. The dataset used in the conventional method comprises images I_n ($n = 1, \ldots, N$, where N is the number of training samples) and texts T_m ($m = 1, \ldots, N$). Here, I_n and T_m for $n = m$ are the paired image and text in the dataset. First, we calculate the semantic similarities $s^{\mathrm{ss}}_{n,m}$ by computing the similarities between text captions T_n and T_m. We then calculate the embedded image and text features ($\boldsymbol{f}^{\mathrm{img}}_n, \boldsymbol{f}^{\mathrm{txt}}_m \in \mathcal{R}^{D_\mathrm{C}}$, where D_C is the dimension of the embedded features) and compute their cross-modal similarity $s_{n,m}$ following the conventional method. Finally, a many-to-many correspondence loss based on the semantic similarity feedback to each module.

A. STEP I: SEMANTIC SIMILARITY CALCULATION

In the first step, we calculate the semantic similarities using the text captions T_n to construct the many-to-many correspondences between semantically similar image and text samples. This process is shown as STEP I in Fig. 2. Inspired by [42], we extract the semantic features $\boldsymbol{f}^{\mathrm{ss}}_n \in \mathcal{R}^{D_\mathrm{S}}$ from T_n using a trained semantic encoder $\mathcal{E}^{\mathrm{ss}}(\cdot)$, where D_S represents the dimension of the semantic features. The extracted features $\boldsymbol{f}^{\mathrm{ss}}_n$ are used to calculate similarities $s^{\mathrm{ss}}_{n,m}$ between T_n and T_m. The above procedure can be expressed as follows:

$$s^{\mathrm{ss}}_{n,m} = \frac{\boldsymbol{f}^{\mathrm{ss}}_n \cdot \boldsymbol{f}^{\mathrm{ss}}_m}{|\boldsymbol{f}^{\mathrm{ss}}_n||\boldsymbol{f}^{\mathrm{ss}}_m|}, \tag{1}$$

$$\boldsymbol{f}^{\mathrm{ss}}_n = \mathcal{E}^{\mathrm{ss}}(\mathrm{T}_n). \tag{2}$$

By using the calculated semantic similarities $s^{\mathrm{ss}}_{n,m}$ in the training phase, the proposed method can keep the many-to-many correspondences between semantically similar images and texts. We construct the embedding space that can consider semantic relationships by adjusting the embedding space to follow $s^{\mathrm{ss}}_{n,m}$.

B. STEP II: CROSS-MODAL SIMILARITY CALCULATION

In the second step, I_n and T_m are embedded into the common semantic embedding space following the conventional method. This process is shown as STEP II in Fig. 2. Theoretically, an arbitrary cross-modal image retrieval method can be applied in this step; thus, we explain the proposed method in reference to the most basic cross-modal image retrieval architecture [44].

First, using the two embedding encoders $\mathcal{E}^{\mathrm{img}}(\cdot)$ and $\mathcal{E}^{\mathrm{txt}}(\cdot)$, which are provided by the conventional method, the proposed method calculates $\boldsymbol{f}^{\mathrm{img}}_n$ and $\boldsymbol{f}^{\mathrm{txt}}_m$ from I_n and T_m as follows:

$$\boldsymbol{f}^{\mathrm{img}}_n = \mathcal{E}^{\mathrm{img}}(\mathrm{I}_n), \tag{3}$$

$$\boldsymbol{f}^{\mathrm{txt}}_m = \mathcal{E}^{\mathrm{txt}}(\mathrm{T}_m). \tag{4}$$

The proposed method then calculates the similarities $s_{n,m}$ between $\boldsymbol{f}^{\mathrm{img}}_n$ and $\boldsymbol{f}^{\mathrm{txt}}_m$ as follows:

$$s_{n,m} = \frac{\boldsymbol{f}^{\mathrm{img}}_n \cdot \boldsymbol{f}^{\mathrm{txt}}_m}{|\boldsymbol{f}^{\mathrm{img}}_n||\boldsymbol{f}^{\mathrm{txt}}_m|}. \tag{5}$$

Conventional methods train the two embedding encoders $\mathcal{E}^{\mathrm{img}}(\cdot)$ and $\mathcal{E}^{\mathrm{txt}}(\cdot)$ to maximize $s_{n,m}$ for $n = m$ than $s_{n,m}$ for $n \neq m$. Although the training strategy in conventional methods allows the encoders to preserve the one-to-one correspondence, the many-to-many correspondence between the semantic similar samples is not guaranteed explicitly. To deal with them, the proposed method preserves both one-to-one and many-to-many correspondences using the semantic similarities $s^{\mathrm{ss}}_{n,m}$ and cross-modal similarities $s_{n,m}$. Specifically, the proposed method trains $\mathcal{E}^{\mathrm{img}}(\cdot)$ and $\mathcal{E}^{\mathrm{txt}}(\cdot)$ to follow $s^{\mathrm{ss}}_{n,m}$. With this procedure, the constructed embedding space is expected to preserve the semantic relationships between the images and texts.

C. STEP III: LOSS CALCULATION

In the third step, we calculate the proposed sentence-based semantic loss $\mathcal{L}_{\mathrm{sbs}}$ to fine-tune the embedding encoders. This process is shown as STEP III in Fig. 2. The sentence-based semantic loss $\mathcal{L}_{\mathrm{sbs}}$ is calculated by combining the text-to-image many-to-many correspondence loss $\mathcal{L}^{\mathrm{t2i}}_{\mathrm{sbs}}$ and the image-to-text many-to-many correspondence loss $\mathcal{L}^{\mathrm{i2t}}_{\mathrm{sbs}}$ as follows:

$$\mathcal{L}_{\mathrm{sbs}} = \mathcal{L}^{\mathrm{t2i}}_{\mathrm{sbs}} + \mathcal{L}^{\mathrm{i2t}}_{\mathrm{sbs}}. \tag{6}$$

Note that each loss focuses on preserving both the one-to-one and many-to-many correspondences from the text-to-image view and the image-to-text view, respectively. Although the goal of the proposed method is to retrieve desired images from a query text, we introduce both text-to-image and image-to-text directional losses following the conventional cross-modal image retrieval methods [45]. The introduced losses are constructed based on the combination of the triplet loss and log-ratio loss [21]. Generally, triplet loss is used in cross-modal image retrieval to preserve the one-to-one correspondence, and the log-ratio loss is used for assuring the many-to-many correspondence. By combining these two loss functions, the proposed text-to-image sentence-based semantic loss $\mathcal{L}^{\mathrm{t2i}}_{\mathrm{sbs}}$ is calculated as follows:

$$\mathcal{L}^{\mathrm{t2i}}_{\mathrm{sbs}} = \sum_n \sum_m \begin{cases} (\ln \frac{\nu_{n,m}}{s^{\mathrm{ss}}_{n,m}})^2 & (s^{\mathrm{ss}}_{n,m} \geq \lambda) \\ \max\{0, \delta - \hat{s}_n + s_{n,m}\} & (s^{\mathrm{ss}}_{n,m} < \lambda) \end{cases}, \tag{7}$$

$$\nu_{n,m} = \frac{s_{n,m}}{s_{n,n}}, \tag{8}$$

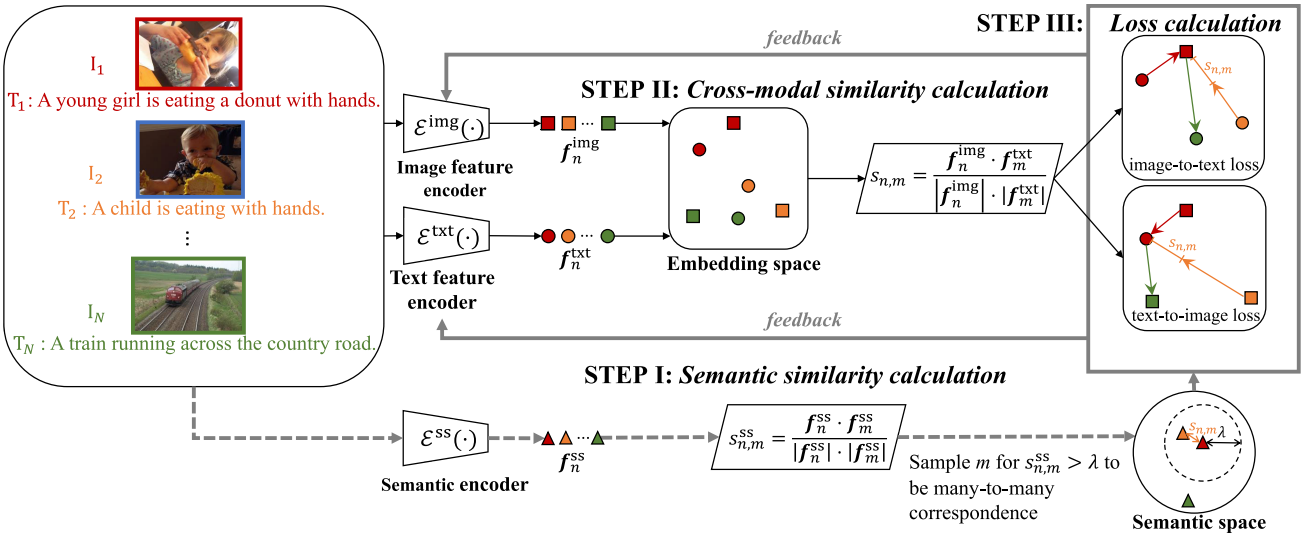


FIGURE 2. Overview of the proposed method.

TABLE 1. Properties of methods we used in the experiment.

Method	Image encoder	Text encoder	Strategy
VSE++ [31]	VGG19	GRU	Importing triplet loss
PVSE (K=1) [34]	ResNet152	Bi-GRU	Utilizing multi-head attention mechanism K=1 means the number of attention heads is 1
PVSE (K=2) [34]	ResNet152	Bi-GRU	Utilizing multi-head attention mechanism. K=2 means the number of attention heads is 2
SGR [35]	Faster-RCNN	Bi-GRU	Utilizing GCN for graph reasoning to infer the similarity between images and texts
SAF [35]	Faster-RCNN	Bi-GRU	Filtering irrelevant images and texts in addition to graph reasoning
CGMN [36]	Faster-RCNN	Bi-GRU	Utilizing GCN to achieve better intra-relation image-text reasoning
PCME [20]	ResNet152	Bi-GRU	Importing the hedged instance embeddings [43] to sample the images and texts as distributions

where λ, δ, and $\hat{s}_n$ are the threshold to determine similar text, the margin hyperparameter, and the minimum of $s^{\mathrm{ss}}_{n,m}$ for $s^{\mathrm{ss}}_{n,m} \geq \lambda$, respectively. The proposed text-to-image sentence-based semantic loss $\mathcal{L}^{\mathrm{t2i}}_{\mathrm{sbs}}$ is reduced as the cross-modal similarity between semantically similar samples is closer to the corresponding semantic similarity. In other words, by training the embedding encoders $\mathcal{E}^{\mathrm{txt}}(\cdot)$ and $\mathcal{E}^{\mathrm{img}}(\cdot)$ to minimize $\mathcal{L}^{\mathrm{t2i}}_{\mathrm{sbs}}$, the embedding space constructed by the embedding encoders preserves the semantic relationships between images and texts.

In addition, following the conventional cross-modal image retrieval procedure, we introduce the image-to-text sentence-based semantic loss $\mathcal{L}^{\mathrm{i2t}}_{\mathrm{sbs}}$ as follows:

$$\mathcal{L}^{\mathrm{i2t}}_{\mathrm{sbs}} = \sum_n \sum_m \begin{cases} (\ln \dfrac{\nu_{m,n}}{s^{\mathrm{ss}}_{m,n}})^2 & (s^{\mathrm{ss}}_{m,n} \geq \lambda) \\ \max\{0, \delta - \hat{s}_n + s_{m,n}\} & (s^{\mathrm{ss}}_{m,n} < \lambda) \end{cases}, \tag{9}$$

$$\nu_{n,m} = \frac{s_{n,m}}{s_{n,n}}. \tag{10}$$

As is known in the cross-modal image retrieval field, the overall loss can be constrained in the text-to-image and image-to-text directions by introducing both losses. These constraints treat images and texts fairly, which leads to the construction of the accurate embedding space [45].

Using the model trained by $\mathcal{L}_{\mathrm{sbs}}$, the retrieval task is performed by simply calculating the cross-modal similarity between the candidate images and query text, and then ranking the candidate images by the cross-modal similarity. Here, there is no need to calculate semantic similarity for the retrieval task.

IV. EXPERIMENTS

We conducted experiments on a frequently used open dataset to evaluate the effectiveness of the proposed method. The experimental settings and results are described in the following subsections.

A. EXPERIMENTAL SETTINGS

1) DATASETS

In our experiments, we used the large-scale MSCOCO dataset [46] and Flickr30K dataset [47], which are adopted

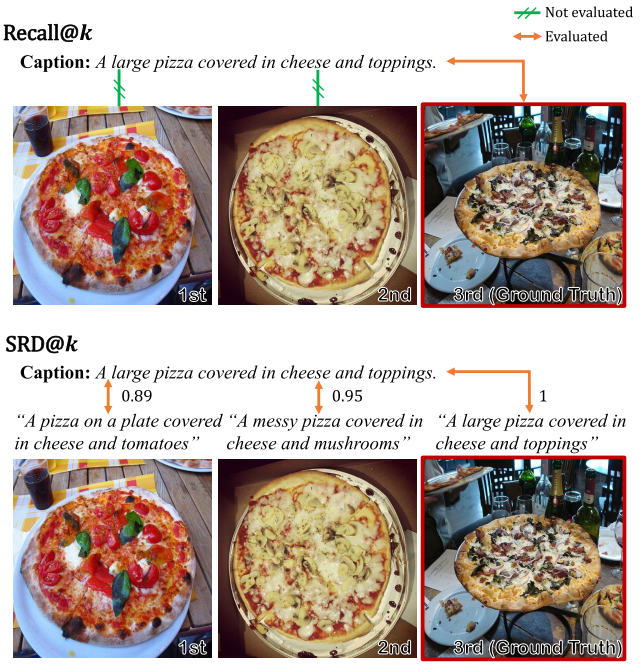


FIGURE 3. Comparison of Recall@k and SRD@k. Recall@k only evaluates the retrieval by the rank of the ground truth image, and SRD@k considers the semantic similarity between images and texts.

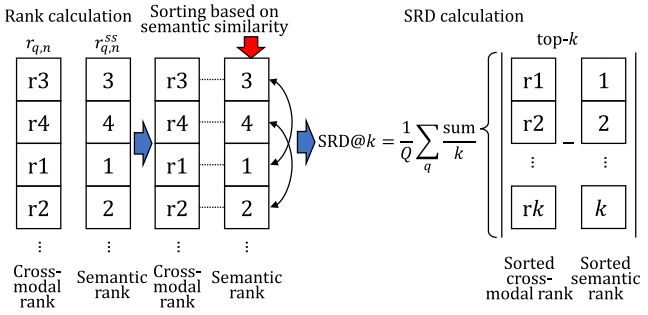


FIGURE 4. Calculation process of proposed SRD metric. $r_{q,n}$ and $r^{ss}_{q,n}$ are the ranks of candidate images calculated from $s_{q,n}$ and $s^{ss}_{q,n}$, respectively.

by most cross-modal image retrieval methods. The two datasets contain images and corresponding texts that describe the contents of the paired image. For MSCOCO, following the widely used data splits provided by [44], 123,287 and 5,000 images were used for the training and test sets, respectively. For Flickr30K, following the data splits provided by [31], 29,783, 1,000, and 1000 images were used for the training, validation, and test sets, respectively. After training, we evaluated the retrieval performance of the proposed by retrieving the target image from the test set using each correspondence text as a query.

2) IMPLEMENTATION DETAILS

For evaluating the effectiveness of our sentence-based semantic loss function, we introduce our loss to the training of recently proposed cross-modal image retrieval methods [31], [34], [35], [36]. We compared the cross-modal retrieval methods with our loss and the original ones. In addition, we compared the models fine-tuned with the proposed loss to PCME [20], which considered that images and texts are not exclusively related in a different way. Technical details of these methods are listed in Table 1. All comparative methods adopted the RNN to extract the text features, which were utilized in the form of a gated recurrent unit (GRU) [48]. For VSE++, PVSE, and PCME, CNN was adopted to extract image features. Specifically, VGG was utilized in VSE++, and ResNet was utilized in PVSE and PCME. For the SGRAF and CGMN methods, the Faster-RCNN [49] object detector was employed to calculate the image features, and then the GCN [50] was adopted to realize the image-text matching. These methods are implemented based on the open-source codes provided by each author. Note that the trained weights of all the models are also provided by each

TABLE 2. Experimental results for SRD@k on MSCOCO dataset. Bold indicates that each method w $\mathcal{L}_{\text{sbs}}$ outperforms the original one.

Method	SRD@1	SRD@5	SRD@10
PCME [20]	82.0	322.5	464.6
VSE++ [31]	176.0	640.0	929.2
VSE++ w $\mathcal{L}_{\text{sbs}}$	178.2	**448.3**	**604.1**
PVSE (K=1) [34]	109.2	387.2	552.2
PVSE (K=1) w $\mathcal{L}_{\text{sbs}}$	117.5	**293.5**	**398.1**
PVSE (K=2) [34]	92.7	469.2	694.5
PVSE (K=2) w $\mathcal{L}_{\text{sbs}}$	129.8	**348.5**	**482.8**
SGR [35]	80.5	397.6	581.7
SGR w $\mathcal{L}_{\text{sbs}}$	**78.7**	**239.8**	**329.5**
SAF [35]	68.3	339.8	496.2
SAF w $\mathcal{L}_{\text{sbs}}$	82.8	**251.5**	**345.1**
CGMN [36]	96.4	760.9	1146.6
CGMN w $\mathcal{L}_{\text{sbs}}$	112.7	**362.9**	**510.6**

TABLE 3. Experimental results for SRD@k on Flickr30K dataset. Bold indicates that each method w $\mathcal{L}_{\text{sbs}}$ outperforms the original one.

Method	SRD@1	SRD@5	SRD@10
PCME [20]	47.8	216.7	350.8
VSE++ [31]	108.5	322.0	501.8
VSE++ w $\mathcal{L}_{\text{sbs}}$	**100.5**	**271.4**	**415.7**
PVSE (K=1) [34]	62.2	229.6	365.1
PVSE (K=1) w $\mathcal{L}_{\text{sbs}}$	72.8	**216.7**	**334.3**
PVSE (K=2) [34]	97.1	331.5	507.8
PVSE (K=2) w $\mathcal{L}_{\text{sbs}}$	**84.4**	**267.0**	**409.5**
SGR [35]	80.3	305.8	472.8
SGR w $\mathcal{L}_{\text{sbs}}$	**40.9**	**162.5**	**263.8**
SAF [35]	77.6	557.1	922.6
SAF w $\mathcal{L}_{\text{sbs}}$	**44.3**	**162.6**	**262.0**
CGMN [36]	39.4	262.9	443.2
CGMN w $\mathcal{L}_{\text{sbs}}$	48.6	**231.8**	**376.1**

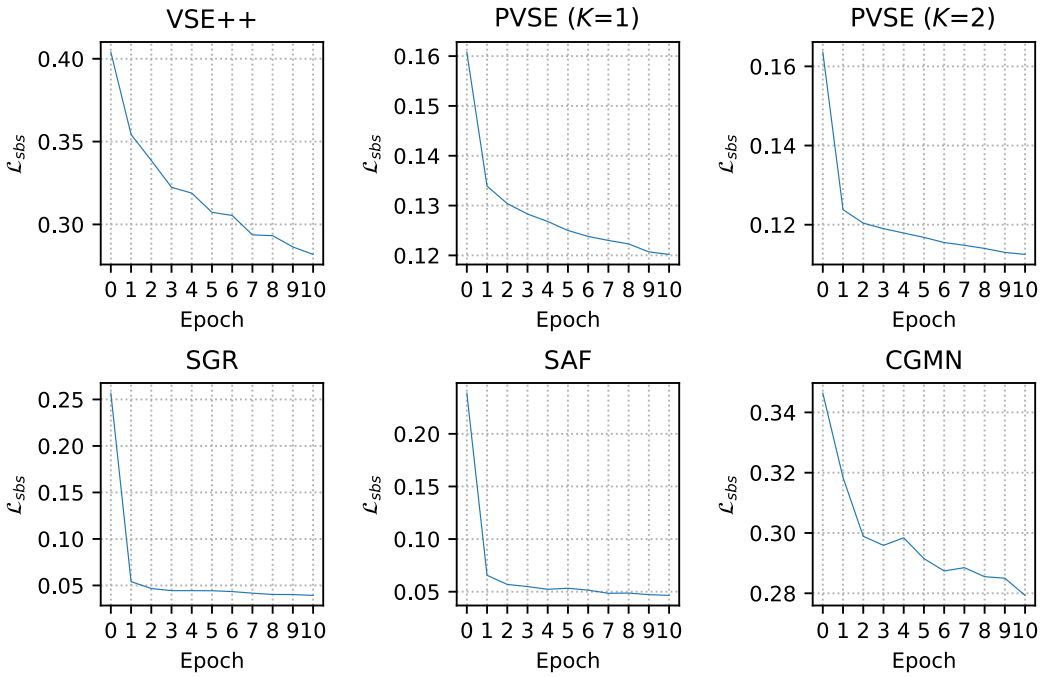


FIGURE 5. The curve of $\mathcal{L}_{\mathbf{sbs}}$ for each method with the number of epochs increasing.

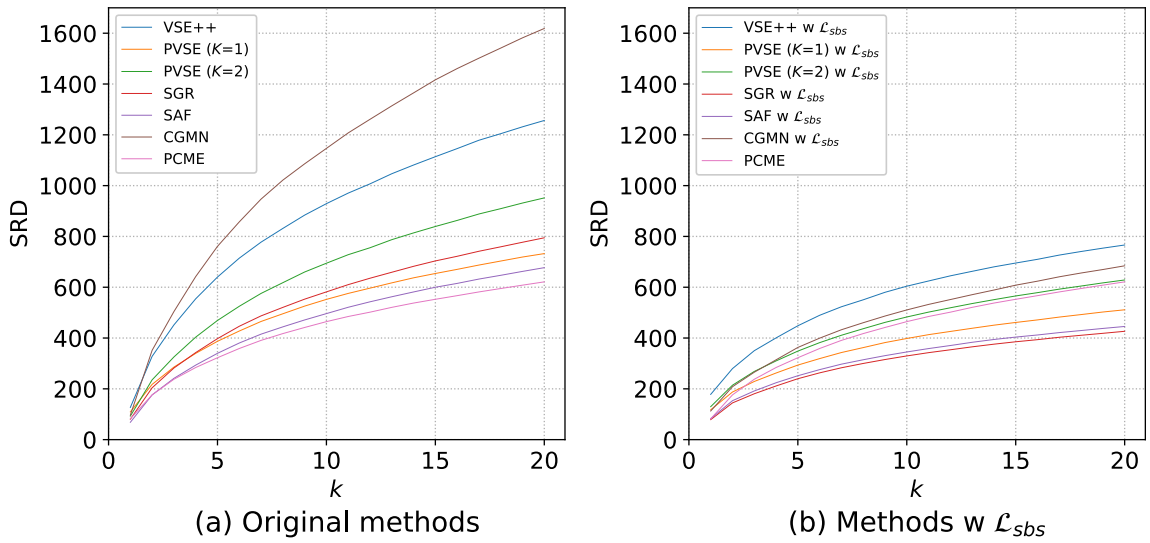


FIGURE 6. SRD for proposed and comparative methods at different k values, where (a) for curves of original methods and PCME and (b) for curves of methods w $\mathcal{L}_{\mathbf{sbs}}$ and PCME.

author, and we fine-tuned these models using our proposed loss function. In the fine-tuning process, we used Adam optimizer [51], and the models were fine-tuned for 10 epochs using our proposed loss function with an initial learning rate of 2e-5 and batch size of 64. For the hyperparameters, we experimentally set $\lambda = 0.75$ and $\delta = 0.1$, and the cosine similarity was normalized in the range $[0, 1]$. In our method, considering semantic information of the relationships between words is crucial for calculating semantic similarity. For this reason, we follow sentence-BERT [42] for constructing semantic feature encoder $\mathcal{E}^{\mathrm{ss}}(\cdot)$. Compared with the other sentence similarity calculation methods such as Bag-of-words, Word2Vec [52] and Sent2Vec [53], sentence-BERT can accurately calculate semantic similarity considering the relationships between words in the full sentence. This is because sentence-BERT is trained on datasets with huge amounts of annotated similar sentence pairs. By extracting the text features based on sentence-BERT, semantic information can be accurately considered in our method.

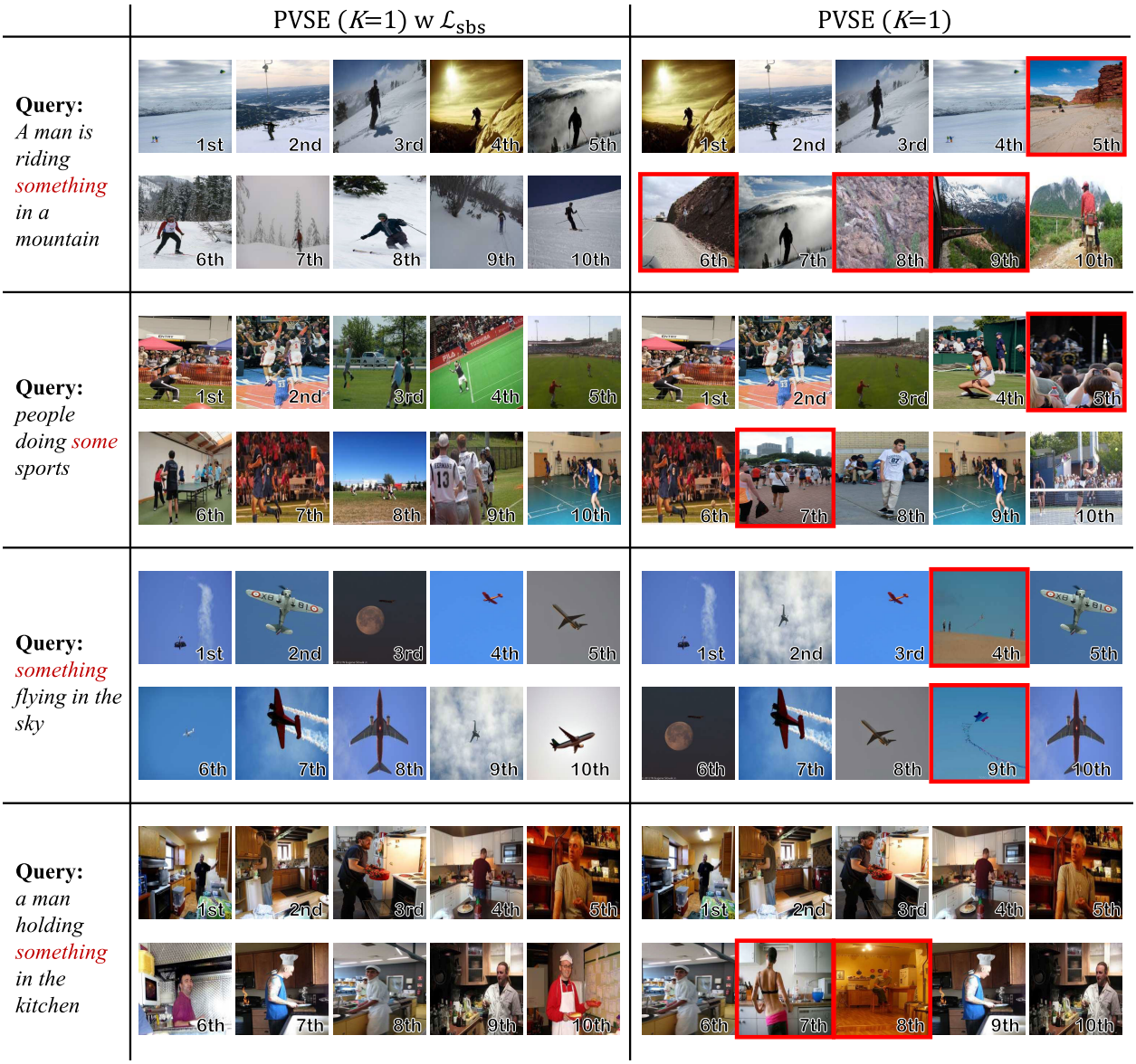


FIGURE 7. Top-10 retrieval results of the PVSE ($K = 1$) w $\mathcal{L}_{\text{sbs}}$ and the original PVSE ($K = 1$). The queries contain ambiguous words and phrases, e.g., *something* and *some sports*. Images with the red frame show that these images are less semantically consistent with the query.

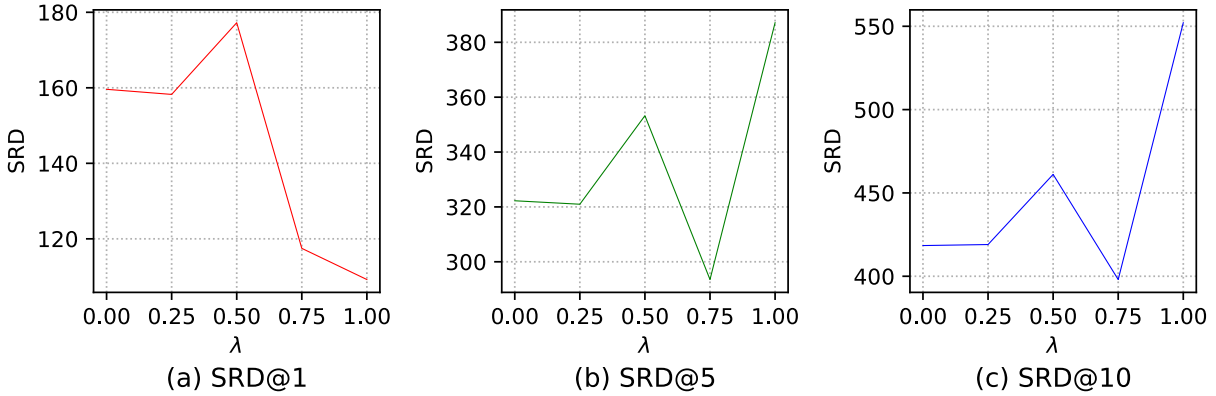


FIGURE 8. λ versus performance. SRD@1, SRD@5, and SRD@10 values of the PVSE ($K = 1$) w $\mathcal{L}_{\text{sbs}}$ against the value λ are shown in (a), (b), and (c), respectively. A lower SRD value indicates better retrieval performance.

B. EVALUATION CONSIDERING SEMANTICALLY SIMILAR SAMPLES

Here, we describe evaluations that focused on semantic relationships. Recall@k is used to evaluate the performance of cross-modal image retrieval; thus, evaluation metrics for semantic relationships have not been considered extensively. In addition, the MSCOCO and Flickr30K datasets do not provide multiple ground truth images that correspond to a single sentence. These may result in underestimation when evaluating the cross-modal retrieval method, as shown in Fig. 3. Thus, we introduce the SRD evaluation metrics to assess whether the semantic relationships are preserved. The calculation process is shown in Fig. 4. SRD@k simply calculates the distance between the ranking $r_{q,n}$ ($q = 1, \ldots, Q$) and ranking $r^{\mathrm{ss}}_{q,n}$ calculated from the cross-modal similarity and the semantic similarity, respectively, where $r_{q,n}$ reveals the rank of the n-th candidate image calculated from the q-th query. SRD@k is defined as follows:

$$\mathrm{SRD@}k = \frac{1}{Qk}\sum_{q}\sum_{n}\begin{cases}|r_{q,n} - r^{\mathrm{ss}}_{q,n}| & (r^{\mathrm{ss}}_{q,n} < k)\\ 0 & (\text{otherwise}).\end{cases} \tag{11}$$

The value of SRD@k becomes smaller as $r_{q,n}$ and $r^{\mathrm{ss}}_{q,n}$ are close. Considering that $r^{\mathrm{ss}}_{q,n}$ is calculated based on the semantic similarity, SRD@k can be used to evaluate whether the semantic relationships are preserved. Note that a small SRD value indicates the better many-to-many retrieval performance. As shown in Fig. 3, SRD considers the semantic relationships between images and texts in the evaluation procedure; thus, retrieval performance can be evaluated more reasonably.

C. EXPERIMENTAL RESULTS

1) CONVERGENCE ANALYSIS

We show the convergence curve of the proposed $\mathcal{L}_{\mathrm{sbs}}$ in each method on MSCOCO dataset in Fig. 5. As is shown in Fig. 5, in all methods, $\mathcal{L}_{\mathrm{sbs}}$ successfully converged.

2) QUANTITATIVE RESULTS

The experimental results obtained on the MSCOCO dataset are shown in Table 2 and Fig. 6. Note that a small SRD value indicates better many-to-many retrieval performance. As shown in the Table 2, each method with the proposed $\mathcal{L}_{\mathrm{sbs}}$ (noted as w $\mathcal{L}_{\mathrm{sbs}}$) outperforms the original method in terms of SRD@5 and SRD@10, respectively. In addition, Fig. 6 shows the SRD@k of the methods w $\mathcal{L}_{\mathrm{sbs}}$ and the comparative methods at different k values. As shown in Fig. 6, each method w $\mathcal{L}_{\mathrm{sbs}}$ achieves better SRD values than the original one when k>2. Especially, we can see that PVSE ($K = 1$) w $\mathcal{L}_{\mathrm{sbs}}$ outperforms PCME by 29.0 and 66.5 in terms of SRD@5 and SRD@10 in the MSCOCO dataset. Considering that the major difference between the two methods is that PVSE ($K = 1$) w $\mathcal{L}_{\mathrm{sbs}}$ considers the semantic relationships between samples, while PCME considers the distribution for a single sample, we confirmed that the usage of proposed $\mathcal{L}_{\mathrm{sbs}}$ enables the model to consider more semantically relative images. These results show the effectiveness of considering the semantic relationship between samples in the training phase. Also, it is notable that the PCME method exhibits a gentler upward trend than the other comparative methods shown in Fig. 6 (a). From this result, we consider that SRD@k is a reasonable metric to evaluate the many-to-many retrieval performance.

Here, since there are no completely identical sentences in this dataset, the similarity between a certain sentence and the other sentences in the dataset cannot achieve 1.0. Thus, SRD@1 only evaluates whether the one-to-one correspondences between images and texts are preserved. In addition, our proposed $\mathcal{L}_{\mathrm{sbs}}$ utilized semantically similar samples rather than one pair of samples for training. This somehow weakened the correspondence in the annotated pairs, but actually strengthened the semantic relationship between samples. On the other hand, despite the usage of distributions, PCME still considers one single pair. For the above reasons, it is reasonable that PVSE ($K = 1$) w $\mathcal{L}_{\mathrm{sbs}}$ performed poorer than PCME in terms of SRD@1, which is equivalent to the evaluation metric only for the one-to-one retrieval task. These reasons can also account for the decrease in SRD@1 for other methods.

Furthermore, the experimental results obtained on the Flickr30K dataset are shown in Table 3. As shown in Table 3, for the Flickr30K dataset, we obtained the same trend of SRD results as in MSCOCO dataset. In addition, some methods even obtained a gain in SRD@1. Considering that the training data of Flickr30K (29,783 images) is fewer than MSCOCO (123,287 images), we infer that the usage of semantically similar samples can boost the one-to-one retrieval performance when the training set is small. These results demonstrate the effectiveness of the proposed $\mathcal{L}_{\mathrm{sbs}}$ to consider the semantic relationships between images and texts.

3) QUALITATIVE RESULTS

To evaluate the influence of the proposed $\mathcal{L}_{\mathrm{sbs}}$ on ambiguous query retrieval performance, we conducted a qualitative experiment on PVSE($K = 1$) w $\mathcal{L}_{\mathrm{sbs}}$ and the original PVSE($K = 1$) trained while keeping the other conditions the same. Since PVSE ($K = 1$) is the most typical cross-modal retrieval method using CNN and RNN with the attention mechanism, we analyze the qualitative results of this method.

Here, we input queries including ambiguous pronouns instead of particular nouns. Figure 7 shows the retrieval results obtained by each version of PVSE ($K = 1$). As shown in Fig. 7, when the query "A man is riding something in a mountain" was used, PVSE ($K = 1$) w $\mathcal{L}_{\mathrm{sbs}}$ retrieved images containing information *people riding skis* and *mountains*. In comparison, PVSE ($K = 1$) paid more attention to the *mountain* information and ignored the *riding something* information. For the query "people doing some sports," PVSE ($K = 1$) w $\mathcal{L}_{\mathrm{sbs}}$ retrieved more images including *sports* information than the PVSE ($K = 1$). For the "something

is flying in the sky'' query, the retrieval results obtained by PVSE ($K = 1$) included some images of kites, and the results obtained by PVSE ($K = 1$) w $\mathcal{L}_{\text{sbs}}$ were all images of planes. When the query ''a man is holding something in the kitchen'' was input, PVSE ($K = 1$) retrieved some images that failed to include the information *holding something*. In comparison, PVSE ($K = 1$) w $\mathcal{L}_{\text{sbs}}$ successfully retrieved images related to all information given in the query. These results demonstrate that PVSE ($K = 1$) w $\mathcal{L}_{\text{sbs}}$ achieved better retrieval performance with ambiguous queries and constructed an embedding space that considers semantic relationships more effectively.

4) ABLATION STUDY

We also conducted an ablation study with different values for λ to analyze its complexity and the sensitivity. For the same reason mentioned in the qualitative result analysis, we conducted the study on PVSE (K=1). λ is the most important hyperparameter in $\mathcal{L}_{\text{sbs}}$ that determines how many samples should be considered semantically similar to the target sample. For a large λ value, fewer samples would be selected as being semantically similar to the anchor sample. Here, we set $\lambda = \{0.0, 0.25, 0.5, 0.75, 1.0\}$, and the SRD@$k$ results are shown in Fig. 8. For $\lambda = 0.0$, all samples were used In the log-ratio calculation, and the sentence-based semantic loss is considered as follows:

$$\mathcal{L}_{\text{sbs}} = \sum_n \sum_m (\ln \frac{r_{m,n}}{s^{\text{ss}}_{m,n}})^2, \tag{12}$$

$$r_{n,m} = \frac{s_{n,m}}{s_{n,n}}. \tag{13}$$

When $\lambda = 1.0$, the sentence-based semantic loss degrades to a triplet loss, which is expressed as follows:

$$\mathcal{L}_{\text{sbs}} = \max\{0, \delta - s_{n,n} + s_{n,m}\}. \tag{14}$$

We found that the best SRD@5 and SRD@10 results were obtained when λ was approximately 0.75. In addition, an acceptable SRD@1 was obtained at the same time. This means when λ was approximately 0.75, the retrieval performance considering many-to-many correspondence is guaranteed while maintaining fairly stable one-to-one retrieval performance. Thus, we selected $\lambda = 0.75$ for the proposed $\mathcal{L}_{\text{sbs}}$.

5) LIMITATIONS

Several limitations should be discussed. First, similar to the other machine learning methods, the performance of methods using $\mathcal{L}_{\text{sbs}}$ is sensitive to the threshold parameter λ (Fig. 8). It is difficult to define the extent to which two texts are truly similar from a human perspective, making it difficult to determine an appropriate value for λ. Improving the design of the semantic similarity calculation procedure may reduce such difficulty. Second, the performance of methods using $\mathcal{L}_{\text{sbs}}$ exhibited an undesired drop in both one-to-one retrieval and many-to-many retrieval performance when λ was set to approximately 0.5. This may result from the mixed-use ratio calculation and addition calculation in the loss function. A more carefully designed parameter-free sentence-based semantic loss function may reduce the impact of these limitations. In addition, the best SRD@10 value obtained by the proposed method was over 300 in the MSCOCO dataset (Table 2), which indicates that the semantically similar samples were still not ranked high enough in the retrieval process. Thus, in the future, we plan to construct a more reasonable architecture to better satisfy the many-to-many retrieval objective.

V. CONCLUSION

In this paper, we have newly proposed a cross-modal image retrieval method that can consider the many-to-many correspondence between images and texts. We achieved this objective by introducing a novel sentence-based semantic loss function that can be applied to an arbitrary cross-modal image retrieval method. The effectiveness of our proposed loss was evaluated experimentally, and the results showed that methods using our proposed loss function outperformed those without it in terms of the proposed SRD metric, which was designed to evaluate many-to-many correspondences. In addition, the results of the qualitative experiment indicate the ability of the introduction of our proposed loss function in retrieving semantically similar images using ambiguous queries. We expect that this work can trigger further research on semantic meanings in the embedding space. In the future, we plan to improve both our loss function and the design of model architecture that can better utilize the loss function.

REFERENCES

[1] J. Strauss, J. M. Paluska, C. Lesniewski-Laas, B. Ford, R. T. Morris, and M. F. Kaashoek, "Eyo: Device-transparent personal storage," in *Proc. USENIX Annu. Tech. Conf.*, 2011, pp. 1–14.

[2] T. M. Coughlin, "Development of digital storage for consumer electronics," in *Proc. IEEE Int. Symp. Consum. Electron.*, Jun. 2006, pp. 1–6.

[3] T. Mei, Y. Rui, S. Li, and Q. Tian, "Multimedia search reranking: A literature survey," *ACM Comput. Surv.*, vol. 46, no. 3, pp. 1–38, Jan. 2014.

[4] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell, "Natural language object retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4555–4564.

[5] B. Li, X. Qi, T. Lukasiewicz, and P. H. S. Torr, "ManiGAN: Text-guided image manipulation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7880–7889.

[6] 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. Vis. (ICCV)*, Dec. 2015, pp. 2425–2433.

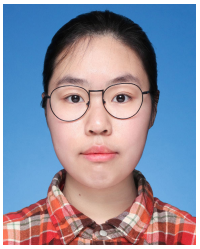
[7] R. Yanagi, R. Togo, T. Ogawa, and M. Haseyama, "Interactive re-ranking via object entropy-guided question answering for cross-modal image retrieval," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 3, pp. 1–17, Aug. 2022.

[8] L. Wu, R. Jin, and A. K. Jain, "Tag completion for image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 716–727, Mar. 2013.

[9] K. Ueki and T. Kobayashi, "Image retrieval under very noisy annotations," in *Proc. 24th Eur. Signal Process. Conf. (EUSIPCO)*, Aug. 2016, pp. 1277–1282.

[10] J. Ma, J. Fan, and W. Wang, "Multi-label classification for images with missing labels," in *Proc. IEEE 15th Int. Conf. Ind. Informat. (INDIN)*, Jul. 2017, pp. 1050–1055.

[11] D. Lu, X. Liu, and X. Qian, "Tag-based image search by social re-ranking," *IEEE Trans. Multimedia*, vol. 18, no. 8, pp. 1628–1639, Aug. 2016.
[12] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.
[13] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 154–162.
[14] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 686–701.
[15] K. Wang, Q. Yin, W. Wang, S. Wu, and L. Wang, "A comprehensive survey on cross-modal retrieval," 2016, *arXiv:1607.06215*.
[16] A. Gordo and D. Larlus, "Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6589–6598.
[17] H. Chen, G. Ding, Z. Lin, S. Zhao, and J. Han, "Cross-modal image-text retrieval with semantic consistency," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1749–1757.
[18] W. Zhou, H. Li, and Q. Tian, "Recent advance in content-based image retrieval: A literature survey," 2017, *arXiv:1706.06064*.
[19] Z. Wang, X. Liu, H. Li, L. Sheng, J. Yan, X. Wang, and J. Shao, "CAMP: Cross-modal adaptive message passing for text-image retrieval," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5764–5773.
[20] S. Chun, S. J. Oh, R. S. de Rezende, Y. Kalantidis, and D. Larlus, "Probabilistic embeddings for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8415–8424.
[21] S. Kim, M. Seo, I. Laptev, M. Cho, and S. Kwak, "Deep metric learning beyond binary supervision," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2288–2297.
[22] P. G. B. Enser, C. J. Sandom, J. S. Hare, and P. H. Lewis, "Facing the reality of semantic image retrieval," *J. Document.*, vol. 63, no. 4, pp. 465–481, Jul. 2007.
[23] B. Barz and J. Denzler, "Hierarchy-based image embeddings for semantic image retrieval," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 638–647.
[24] M. Wray, H. Doughty, and D. Damen, "On semantic similarity in video retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3650–3660.
[25] V. N. Gudivada and V. V. Raghavan, "Content based image retrieval systems," *Computer*, vol. 28, no. 9, pp. 18–22, Sep. 1995.
[26] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, "A survey of content-based image retrieval with high-level semantics," *Pattern Recognit.*, vol. 40, no. 1, pp. 262–282, 2007.
[27] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.
[28] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, and T. Liu, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, Jan. 2018.
[29] T. Mikolov, M. Karafiát, L. Burget, J. Černocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. 11th Annu. Conf. Int. Speech Commun. Assoc.*, vol. 2, 2010, pp. 1045–1048.
[30] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.
[31] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "VSE++: Improving visual-semantic embeddings with hard negatives," 2017, *arXiv:1707.05612*.
[32] L. Ma, W. Jiang, Z. Jie, and X. Wang, "Bidirectional image-sentence retrieval by local and global deep matching," *Neurocomputing*, vol. 345, pp. 36–44, Jun. 2019.
[33] K.-H. Lee, X. Chen, G. Hua, H. Hu, and X. He, "Stacked cross attention for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 201–216.
[34] Y. Song and M. Soleymani, "Polysemous visual-semantic embedding for cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1978–1979.
[35] H. Diao, Y. Zhang, L. Ma, and H. Lu, "Similarity reasoning and filtration for image-text matching," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2021, vol. 35, no. 2, pp. 1218–1226.
[36] Y. Cheng, X. Zhu, J. Qian, F. Wen, and P. Liu, "Cross-modal graph matching network for image-text retrieval," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 18, no. 4, pp. 1–23, Nov. 2022.
[37] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 2, Jun. 2006, pp. 1735–1742.
[38] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu, "Learning fine-grained image similarity with deep ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1386–1393.
[39] J. Li, E. Yu, J. Ma, X. Chang, H. Zhang, and J. Sun, "Discrete fusion adversarial hashing for cross-modal retrieval," *Knowl.-Based Syst.*, vol. 253, Oct. 2022, Art. no. 109503.
[40] E. Yu, J. Sun, J. Li, X. Chang, X.-H. Han, and A. G. Hauptmann, "Adaptive semi-supervised feature selection for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1276–1288, May 2018.
[41] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 10394–10403.
[42] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process.*, 2019, pp. 3982–3992.
[43] S. J. Oh, K. P. Murphy, J. Pan, J. Roth, F. Schroff, and A. C. Gallagher, "Modeling uncertainty with hedged instance embeddings," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–15.
[44] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," 2014, *arXiv:1411.2539*.
[45] F. Wu, X. Lu, Z. Zhang, S. Yan, Y. Rui, and Y. Zhuang, "Cross-media semantic representation via bi-directional learning to rank," in *Proc. 21st ACM Int. Conf. Multimedia*, Oct. 2013, pp. 877–886.
[46] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.
[47] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, Feb. 2014.
[48] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proc. 8th Workshop Syntax, Semantics Struct. Stat. Transl.*, 2014, pp. 103–111.
[49] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 1–15.
[50] S. Zhang, H. Tong, J. Xu, and R. Maciejewski, "Graph convolutional networks: A comprehensive review," *Comput. Social Netw.*, vol. 6, no. 1, pp. 1–23, Dec. 2019.
[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–13.
[52] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[53] M. Pagliardini, P. Gupta, and M. Jaggi, "Unsupervised learning of sentence embeddings using compositional n-gram features," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2018, pp. 528–540.



HUAYING ZHANG (Student Member, IEEE) received the B.S. degree in communication engineering from East China Normal University, China, in 2021. She is currently pursuing the M.S. degree with the Graduate School of Information Science and Technology, Hokkaido University. Her research interest includes machine learning and its applications.

RINTARO YANAGI (Student Member, IEEE) received the B.S. degree in electronics and information engineering from Hokkaido University, Japan, in 2019, and the M.S. degree from the Graduate School of Information Science and Technology, Hokkaido University, in 2021, where he is currently pursuing the Ph.D. degree. His research interest includes machine learning and its applications. He is a Student Member of ACM.

REN TOGO (Member, IEEE) received the B.S. degree in health sciences from Hokkaido University, Japan, in 2015, and the M.S. and Ph.D. degrees from the Graduate School of Information Science and Technology, Hokkaido University, in 2017 and 2019, respectively. He is currently a specially appointed Assistant Professor with the Laboratory of Media Dynamics, Faculty of Information Science and Technology, Hokkaido University. He is also a Radiological Technologist. His research interest includes machine learning and its applications. He is a member of ACM and IEICE.

TAKAHIRO OGAWA (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronics and information engineering from Hokkaido University, Japan, in 2003, 2005, and 2007, respectively. He joined the Graduate School of Information Science and Technology, Hokkaido University, in 2008. He is currently an Associate Professor with the Faculty of Information Science and Technology, Hokkaido University. His research interests include AI, the IoT, and big data analysis for multimedia signal processing and its applications. He is a member of ACM, IEICE, and ITE. He was the Special Session Chair of IEEE ISCE2009, the Doctoral Symposium Chair of ACM ICMR2018, the Organized Session Chair of IEEE GCCE2017-2020, the TPC Vice Chair of IEEE GCCE2018, and the Conference Chair of IEEE GCCE2019. He has been an Associate Editor of *ITE Transactions on Media Technology and Applications*.

MIKI HASEYAMA (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronics from Hokkaido University, Japan, in 1986, 1988, and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University, as an Associate Professor, in 1994. She was a Visiting Associate Professor at Washington University, Seattle, WA, USA, from 1995 to 1996. She is currently a Professor with the Faculty of Information Science and Technology, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She is a fellow of ITE and a member of IEICE and ASJ. She has been the Vice-President of the Institute of Image Information and Television Engineers (ITE), Japan, the Editor-in-Chief of *ITE Transactions on Media Technology and Applications*, and the Director of the International Coordination and Publicity, Institute of Electronics, Information and Communication Engineers (IEICE).