1 INTRODUCTION

Cross-modal retrieval plays an important role in multi-agent environment perception. For instance, in the process of earthquake rescue, multiple smart robots are often used to sense the rescue environment. Each robot usually carries a variety of different sensors, including visible light sensors, infrared sensors, thermal imaging cameras, and so on. When multiple robots perceive and model the environment in a distributed and collaborative manner, we first need to determine whether the environmental information obtained from different sensors of different robots originates from the same target. However, traditional retrieval methods [55] can not meet the needs of users since they can only do retrieval in the same modality. Therefore, cross-modal retrieval system [8, 40, 46] has emerged as a promising method to solve the above problem.

Unlike traditional retrieval methods, cross-modal retrieval methods aim to utilize massive multi-modal data to achieve flexible retrieval between documents in different modalities from queries. Take image-text retrieval as an example, methods take images as the query to retrieve relevant texts. The results obtained across multi-modalities can be helpful for users to gain a full range of data information about the target objects or topics.

The main problem of performing cross-modal retrieval is how to measure the semantic similarities between different modalities of data, which is referred to as the heterogeneity gap. A common approach to minimize the gap between multi-modal data with the same semantic information is representation learning [49]. It attempts to map data from different modalities into a common space where the semantic similarities among multi-modal data can be calculated directly. A variety of cross-modal retrieval methods [19, 28, 48] have been proposed, which develop different approaches for embedding multi-modal data into a common space. The traditional statistical correlation analysis methods [10, 13, 34] learn linear projections by optimising target statistical values. Canonical Correlation Analysis (CCA) [10] is one of the most popular traditional unsupervised representation learning methods, which adopts linear projection methods to establish inter-modal relationships between multi-modal data. However, the complex correlation between multimedia data in the real world cannot be effectively modeled only by using linear projection methods.

Inspired by the widespread popularity of deep neural networks [16], a large number of deep learning-based representation learning approaches [1, 7, 27, 44] have been produced to learn the common space for multi-modal data. For instance, Ngiam [26] proposes a cross-modal learning method based on deep neural networks. The model considers multi-modal fusion learning, cross-modal learning and shared representation learning. The effectiveness of this method has been verified by video and speech recognition. Considering semantic information comprehensively, Li [17] uses multi-class supervised information to learn a common semantic space to achieve cross-modal retrieval. Specifically, for a single modality, the method uses a deep network to learn corresponding features, and the common semantic vectors of different modalities are used as optimization targets for different modal correlations to achieve semantic association of data from different modalities. Nevertheless, some existing cross-modal retrieval methods [6, 9, 10, 23, 31] involve the data embeddings in the common space, but do not pay attention to the content information that the embeddings contain. Other methods that consider preserving information only preserve semantic information or original feature information.

In this paper, we propose a novel deep supervised cross-modal retrieval method, which learns projections that map multi-modal data to the embeddings which fully preserve the original feature information as well as semantic information of each modality. For the sake of generality, we use image and text modality as an example to demonstrate the method. As illustrated in Figure 1, the whole process can be ensured by three steps. First of all, we preliminarily embed image-

text sample pairs to a common representation space through two neural networks. Secondly, multi-modal conditional principal label space transformation (CPLST) [47] is used to learn the content-aware feature vector. Thirdly, we learn projections via two pairs of encoders and decoders, one for image modality and the other for text modality. It is worth noting that the decoders use additional constraint to decompose content-aware feature vector into feature representations, which should be consistent with the feature vectors in common representation space obtained in the first step.

The main contributions of our work can be summarised as follows:

- A deep supervised cross-modal learning architecture using multi-modal content autoencoder (MMCA-CMR) is proposed to eliminate the heterogeneity gap between different modalities. In this way, MMCA-CMR can improve the precision of cross-modal retrieval by learning feature vectors from different modalities and content information simultaneously.
- The MMCA-CMR method utilizes the feature extraction ability of deep neural networks to nonlinearly map multi-modal data to a common representation space that makes multi-modal data from different feature spaces share the same content information as similar as possible.
- Extensive experiments on four widely-used benchmark datasets (the Wikipedia dataset, the Pascal Sentence dataset, the MIRFLICKR dataset and the NUS-WIDE dataset) have been conducted to indicate the effectiveness of our proposed method. The results demonstrate that our method outperforms state-of-the-art methods for cross-modal retrieval.

The remainder of this paper is organized as follows. Section 2 reviews the related work in cross-modal retrieval and autoencoder. Section 3 demonstrates our proposed approach. Section 4 provides experimental results and analysis. Finally, we conclude our method in Section 5.