Abstract
Most existing unsupervised hashing methods learn binary hash codes through similarity structure preserving or contrastive learning of hash codes. However, these methods usually use the visual similarity of images to guide hash learning, which does not fully utilize the high-level semantic concept information contained in images, resulting in limited retrieval performance. To tackle this problem, we propose a novel high-level deep unsupervised hashing method called Language Guidance Hashing (LGH). Specifically, LGH utilizes a language model to mine high-level semantic concept information in images and construct a language-based similarity structure, which is used to guide hash learning. By introducing features of textual modality, higher information gain can be brought. In addition, we also propose a language-guided contrastive learning method for learning high-quality binary hash codes. Extensive experimental results show that LGH significantly outperforms state-of-the-art unsupervised hashing methods on three benchmark image datasets.

Introduction
On the other hand, hashing methods based on contrastive learning usually learn the hash code by maximizing the similarity of augmented versions of the same image. Although somewhat effective, contrastive learning ignores a large amount of useful similarity information between different images because it assumes that all other images are considered negative samples. To solve the above problems, we propose a novel unsupervised hashing method named Language Guidance Hashing (LGH), as shown in Figure 1(b), which utilizes the large-scale pre-trained natural language model to mine high-level semantic concept information as a guide for hash learning. Prior studies [9, 10] have shown that the introduction of additional modal features can provide higher information gain and thus substantially improve model performance. Therefore, we add language guidance into contrastive learning to model features into hash learning, thus constructing a high-quality semantic similarity structure of the original data space. Specifically, previous unsupervised hashing methods generally utilize solely the deep visual features extracted by ImageNet pre-trained model to construct similarity structure. Further, we use the classification head of the ImageNet pre-trained model to generate a set of text pseudo-labels for the samples, which represent the high-level semantic concepts contained in the images. Then we feed the text pseudo-labels of the samples into a large-scale pre-trained language model [11] to provide language guidance for hash learning. After that, with the help of distillation, we preserve as much of the high-quality semantic structure constructed by text features into the hash space as possible. In addition, for better contrastive learning, instead of using a data augmentation strategy to provide positive and negative sample pairs, we use the language model to find conceptually similar or dissimilar image pairs. To this end, we leverage the previously obtained language-based similarity structure to generate pairwise pseudo label information between sample pairs. The pairwise pseudo-labels indicate whether two samples have the same semantics, which is a condition to guide contrastive learning for generating hash codes.

Method
Figure 2. Overview of LGH. First, LGH obtains Top-k ImageNet pseudo-labels for the samples with the help of a pre-trained model and its classification head. Pseudo-labels and prompt templates are combined to generate the corresponding texts, which are fed to the language model to construct a language-based similarity structure. During training, LGH transforms generated by the language model and the semantic similarity structure in the Hamming space into probability distributions, and then uses KL divergence to constrain the consistency of the distributions. Meanwhile, LGH guides contrastive learning with the help of the language-based similarity structure.

Results
Table 1 shows Mean Average Precision (MAP) results for different number of bits on CIFAR-10, NUS-WIDE and MSCOCO. Figure 3 shows the t-SNE visualizations of 64 bits hash codes generated by CIBHash, SGD and LGH on CIFAR-10. Figure 4 shows examples of the top 10 retrieved images and Precision@10 on MSCOCO with 64 bits hash codes.

References