Fine-Grained Alignment Network for Zero-Shot Cross-Modal Retrieval

SHIPING GE, ZHIWEI JIANG, YAFENG YIN, CONG WANG, ZIFENG CHENG, and QING GU, State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

Zero-Shot Cross-Modal Retrieval (ZS-CMR) aims to perform cross-modal retrieval on data of unseen classes, where a key challenge is how to address the modality-gap and domain-shift problems simultaneously. Existing methods tackle this challenge mainly by embracing a sample-label alignment paradigm, which aligns samples of different modalities but of the same class with the word embedding of their class label. However, these methods only focus on the class-level alignment and overlook the alignment of rich fine-grained semantic information in samples, incurring coarse understanding of sample matching and poor generalization on unseen classes. In this article, we propose a novel Fine-Grained Alignment Network, an end-to-end framework that learns representation with two fine-grained alignment strategies, yielding representation space that can be better generalized to unseen classes. Specifically, we extract two kinds of fine-grained representations, region embedding and label distribution, respectively, from aspects of both feature and label. To optimize the region embedding, we propose a Fine-Grained Contrastive Learning (FGCL) strategy to simultaneously conduct class-level alignment and model the intra-class discrepancy. To optimize the label distribution, we propose a Fine-Grained Label Alignment (FGLA) strategy to align diverse fine-grained semantic information among samples, rather than merely label information. Finally, both region embedding and label distribution are utilized together to perform ZS-CMR at a finer granularity. Experimental results on three widely used datasets demonstrate that our method outperforms the state-of-the-art methods by a large margin. Detailed ablation studies have also been carried out, which provably affirm the advantage of each component we propose. Our code will be available at https://github.com/ShipingGe/FGAN.

CCS Concepts: • **Information systems → Image search**; *Combination, fusion and federated search*; • **Applied computing → Document searching**;

Additional Key Words and Phrases: Zero-Shot Cross-Modal Retrieval, Contrastive Learning, Knowledge Distillation, Multimodal Learning

This work was supported by National Natural Science Foundation of China (Grant Nos. 61972192, 62172208, 61906085), Collaborative Innovation Center of Novel Software Technology and Industrialization, and Fundamental Research Funds for the Central Universities (Grant No. 14380001).

Authors' Contact Information: Shiping Ge, State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China; e-mail: shipingge@smail.nju.edu.cn; Zhiwei Jiang (corresponding author), State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China; e-mail: jzw@nju.edu.cn; Yafeng Yin, State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China; e-mail: yafeng@nju.edu.cn; Cong Wang, State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China; e-mail: cw@smail.nju.edu.cn; Zifeng Cheng, State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China; e-mail: chengzf@smail.nju.edu.cn; Qing Gu, State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China; e-mail: guq@nju.edu.cn.

ACM 1551-6865/2025/10-ART275
https://doi.org/10.1145/3722223

ACM Reference format:
Shiping Ge, Zhiwei Jiang, Yafeng Yin, Cong Wang, Zifeng Cheng, and Qing Gu. 2025. Fine-Grained Alignment Network for Zero-Shot Cross-Modal Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 21, 10, Article 275 (October 2025), 24 pages.
https://doi.org/10.1145/3722223

1 Introduction

Cross-Modal Retrieval (CMR) aims to retrieve semantically similar data across different modalities, such as images and text. Traditional CMR methods operate under the closed-world assumption, where the classes in the test data must have been present during training. However, labeled data for new classes are often unavailable, and collecting and annotating these data can be time-consuming. To tackle this challenge, **Zero-Shot Cross-Modal Retrieval (ZS-CMR)** has been introduced, which trains a model on seen classes while enabling retrieval for unseen (new) classes [6, 54]. ZS-CMR faces significant challenges, including (1) modality gap, the inconsistency between the representation spaces of different modalities, and (2) domain shift, which occurs when representations learned from training data do not generalize well to unseen classes. These two issues adversely affect the model generalizability in ZS-CMR.

To tackle the modality gap and domain shift issues, most existing methods [6, 27, 54, 57] typically align data from different modalities into a shared space. They also use word embeddings of class labels as an auxiliary semantic space to transfer knowledge between seen and unseen classes. However, these methods perform alignment only at the class level, neglecting the rich semantic nuances and intra-class discrepancies carried by individual samples, resulting in coarse-grained alignment. As shown in the left part of Figure 1, while this method ensures that samples within the same class (e.g., “Bicycle”) are clustered together, it overlooks important details of each individual samples. The class-level cross-modal alignment can ensure that samples of different modalities, but within the same class are closely distributed, based on cross-entropy loss [27] or triplet ranking loss [57]. However, it treats the same-class samples with no difference, ignoring the intra-class discrepancy. Additionally, while label-sample alignment aims to connect each sample to its corresponding label embedding, it fails to account for the semantic diversity of samples. As shown in the right part of Figure 1, although the query text and Hard Semi-Positive image belong to the same class, their semantic match is minimal, necessitating that they be distanced in the embedding space, which shows that each sample may contain more information than just its label. This coarse-grained alignment, focusing only on label information from seen classes, overlooks the alignment of fine-grained information, leading to suboptimal generalization to unseen classes.

To achieve better generalization on unseen classes, we propose a fine-grained alignment approach that simultaneously models intra-class discrepancies and the semantic diversity of samples. As shown in the right part of Figure 1, we incorporate intra-class alignment to distinguish between exact and partial matching relationships among same-class samples, allowing distances in the embedding space to more accurately reflect their fine-grained semantic differences. Furthermore, to capture the semantic diversity of samples, we extract a fine-grained label distribution defined in a large label space, such as ImageNet labels, for each sample. By aligning this distribution with the pseudo ImageNet label distribution obtained from a pre-trained model, we ensure that the label distribution effectively represents the fine-grained semantic information carried by each sample (bottom right).

To address the challenges in ZS-CMR, we propose the **Fine-Grained Alignment Network (FGAN)**, an end-to-end framework designed to learn representations with fine-grained alignment by jointly modeling intra-class discrepancies and the semantic diversity of samples. FGAN consists of

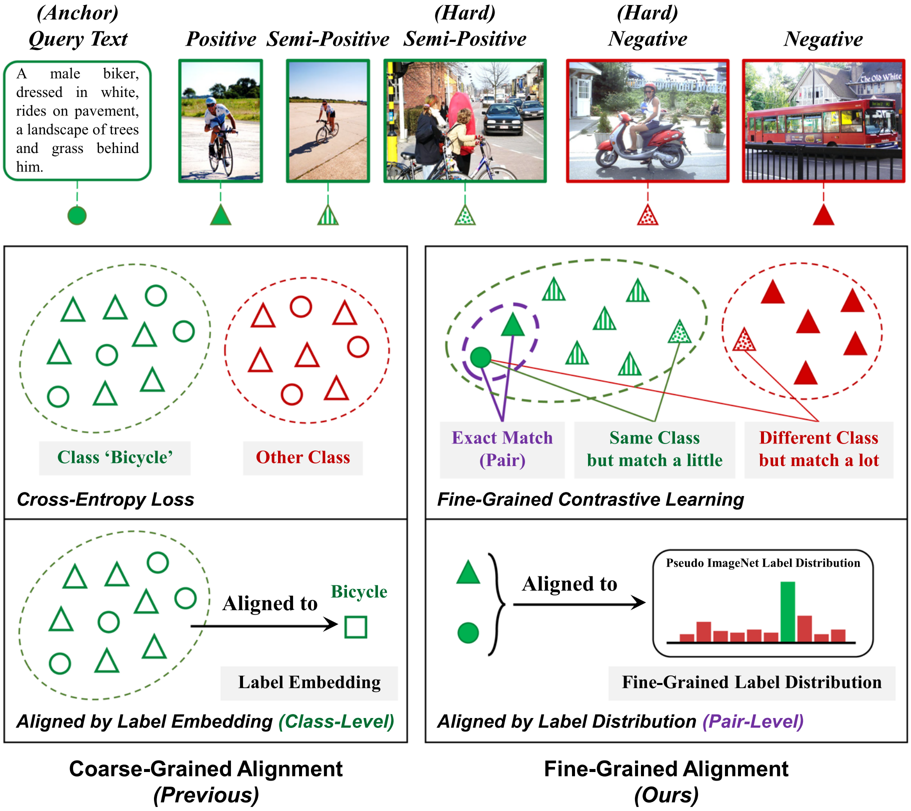


Fig. 1. Illustration of coarse-grained vs. our proposed fine-grained alignment strategies in ZSCMR. The left panel illustrates the traditional coarse-grained alignment approach used in existing methods. In this scenario, samples from different modalities (e.g., images and text) are aligned based solely on their class labels, which overlooks the rich semantic diversity and intra-class discrepancies present in the data. The right panel presents our proposed fine-grained alignment strategy. Unlike the coarse-grained approach, our method distinguishes between two types of relationships among same-class samples: exact matches (positive relations) and partial matches (semi-positive relations). The figure demonstrates how our method aligns positive pairs closely in the embedding space while appropriately distancing semi-positive pairs to reflect their partial semantic overlap.

three main components: the **Fine-Grained Encoder (FGE)**, **Fine-Grained Contrastive Learning (FGCL)**, and **Fine-Grained Label Alignment (FGLA)**. FGE extracts two types of fine-grained features from each input sample: region embeddings and label distributions, in an end-to-end manner. For learning region embeddings, the FGCL strategy aims to bring samples with positive relations (exact matches) closer together while pushing samples with semi-positive (partial matches) or negative relations (different classes) further apart. For learning label distributions, the FGLA strategy aligns a sample's label distribution with a pseudo ImageNet label distribution generated by a pre-trained model. Ultimately, both region embeddings and label distributions can be jointly utilized for retrieval on unseen classes.

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

- We propose a novel end-to-end ZS-CMR framework FGAN, which can learn fine-grained alignment-aware representation for data of different modalities.

- We develop an FGE along with two fine-grained alignment strategies to jointly model the intra-class discrepancy and semantic diversity of samples.
- Extensive experiments conducted on three benchmark datasets demonstrate that our FGAN outperforms the state-of-the-art methods by a significant margin.

2 Related Work

2.1 CMR

CMR aims to retrieve similar instances in one modality using a query instance from another modality [49, 52, 65]. Therefore, the main challenge of CMR is how to measure the content similarity between different modalities of data, which is referred to as the *heterogeneity gap*. Compared with the traditional retrieval tasks, CMR requires cross-modal modeling, so that the users can retrieve what they want by submitting what they have [3, 45, 49].

Traditional methods usually utilize the visual Bag-of-Words model (e.g., SIFT BoVW) to represent image features and match them with text features generated by the language model (e.g., Word2Vec, Doc2Vec). Rasiwasia et al. [34] employ **Canonical Correlation Analysis (CCA)** to learn correlations between different features. Sharma et al. [35] present Generalized Multiview Analysis, which is a supervised extension of CCA and has the potential to replace CCA whenever classification or retrieval is the purpose and label information is available. Wang et al. [47] propose a coupled linear regression framework to deal with the problems of relevance measurement and feature selection in CMR.

In recent years, deep learning methods become widely adopted in CMR which utilize deep visual features and deep networks to model the share embedding space. Wang et al. [46] propose a regularized deep neural network for semantic mapping across modalities to capture both intra-modal and inter-modal relationships. Wei et al. [52] implement several classic methods with CNN visual features and achieve better results than previous traditional methods. Zhen et al. [65] minimize the discrimination loss in both the label space and the common representation space to supervise the model learning discriminative features. More recently, researchers increasingly concentrate on more realistic scenarios for CMR. Fang et al. [14] propose the first method to apply prompt tuning for **Universal Cross-Domain Retrieval (UCDR)**, which employs a two-step process to simulate content-aware dynamic prompts to produce generalized features for UCDR. Li et al. [25] propose a novel prototype-based aleatoric uncertainty quantification framework to provide trustworthy predictions by quantifying the uncertainty arisen from the inherent data ambiguity for CMR. Zhang et al. [64] introduce a Unified Prompt Generation module to dynamically produce modality-aware prompt tokens, enabling the perception of prior semantic information on both video and text inputs for text–video retrieval. Shen et al. [36] present an end-to-end pre-training network with Hierarchical Matching and Momentum Contrast to explore the hierarchical semantic information in videos via multilevel semantic matching between videos and texts for text–video retrieval.

2.2 ZS-CMR

ZS-CMR is first proposed by Chi and Peng [6] to achieve retrieval across multiple media types in zero-shot scenario where there are no overlaps between categories of training and testing data. The key challenge of ZS-CMR is how to mitigate the modality gap between images and texts of the same classes and handle the domain-shift problem between seen and unseen classes [6, 27, 54, 55, 57].

Most existing ZS-CMR methods follow the zero-shot learning paradigm and utilize generative models (e.g., **Generative Adversarial Network (GAN)** [17], **Variational Autoencoder (VAE)** [22]) and pre-defined class-prototypes to learn common representations for retrieval.

Chi and Peng [6] propose the first ZS-CMR method, which learns common representations by mitigating the difference between image representations, text representations, and category word embeddings using two GANs. Xu et al. [55] use three paralleled sub-networks to capture the intrinsic data structures of different modalities and leverage word vectors of both seen and unseen labels as guidance to enhance the knowledge transfer to unseen labels. Lin et al. [27] employ modality-specific VAEs to learn a shared low-dimensional latent space of input multi-modal features. Similarly, Xu et al. [54] utilize category label embeddings to guide WGANs to synthesize multi-modal features for stable training. Yang et al. [59] propose a bidirectional random walk scheme to mining reliable relationships between images by traversing heterogeneous manifolds in the feature space of each modality. Xu et al. [57] combine the strength of AE and GAN to jointly incorporate common latent space learning, knowledge transfer, and feature synthesis for ZS-CMR. Wang et al. [48] propose an **Instance-Level Semantic Alignment (ILSA)** method to make full use of the instance-level information to mitigate the problem of the ignorance of intra-modal variance. Tian et al. [40] use two coupled disentanglement VAEs and a fusion exchange VAE to disentangle the original representations of each modality into modality-invariant and modality-specific features.

Different from most previous methods that utilize label embedding as auxiliary space to learn a common representation for different modalities, we do not make use of label embedding but learn both fine-grained region embedding and label distribution for each sample instead. To enhance the generalization of region embedding on unseen classes, we propose a supervised contrastive loss to align region embedding at a finer granularity than the class level. Besides, we also use the label distribution on ImageNet labels to describe the semantic information carried by each sample, providing more information for retrieval on unseen classes.

2.3 Contrastive Learning

Contrastive learning is a machine learning technique used to learn the general representations of samples by grouping similar samples closer and dissimilar samples far from each other [21, 41, 61]. Contrastive learning has recently received interest due to its success in self-supervised representation learning in the computer vision, natural language processing, and other domains [24, 29, 51]. Given an anchor data point, the goal of contrastive learning is to pull the anchor close to positive data points and push it away from negative data points in the latent space [1, 5, 16, 19].

Based on the simple idea of contrastive learning, many contrastive losses and their variants have been designed. Wang and Gupta [51] design a Siamese-triplet network with a contrastive ranking loss function to train a CNN for image representation without semantic labels. Sohn [38] proposes multi-class N-pair loss to address the slow convergence problem of traditional pairwise contrastive loss and triplet loss. Hermans et al. [18] show that using a variant of the triplet loss to perform end-to-end deep contrastive learning outperforms most other methods for the Person Re-Identification task.

Contrastive learning has shown its success in learning visual representations, text representations, and multi-modal representations. Chen et al. [4] learn representations by maximizing agreement between differently augmented views of the same data example via a contrastive loss in the latent space. Wang et al. [50] explore supervised contrastive learning strategies to learn better visual representations to boost long-tailed image classification performance. Li et al. [26] propose a novel contrastive learning-based zero-shot text classification framework to capture the semantic relations between classes and build more discriminative embeddings. Yu et al. [62] apply a captioning loss and a contrastive loss between uni-modal image and text embeddings to pre-train an image–text encoder–decoder foundation model.

3 Task Definition

In this section, we introduce the notations and provide a formal definition of the ZS-CMR task. Without loss of generality, we consider two modalities, i.e., image and text, as an example to illustrate the task. Let $\mathcal{D}_s = \{(\mathbf{V}_i, \mathbf{T}_i, y_i)\}_{i=1}^{N_s}$ denote the seen dataset with N_s image–text pairs (i.e., image and its corresponding text description), where $\mathbf{V}_i$, $\mathbf{T}_i$, and y_i refer to image, text, and class label of the ith image–text pair, respectively. Similarly, the unseen dataset is denoted as $\mathcal{D}_u = \{(\mathbf{V}_j, \mathbf{T}_j, y_j)\}_{j=1}^{N_u}$. For ZS-CMR, we assume that the class label sets of the seen dataset and unseen dataset are disjoint. Thus, we denote $\mathcal{Y}_s = \{y_i\}_{i=1}^{N_s}$ and $\mathcal{Y}_u = \{y_j\}_{j=1}^{N_u}$ as the class labels sets of $\mathcal{D}_s$ and $\mathcal{D}_u$, respectively, and ensure $\mathcal{Y}_{seen} \cap \mathcal{Y}_{unseen} = \emptyset$. The objective of ZS-CMR is to construct a CMR model using the seen dataset $\mathcal{D}_s$ and subsequently apply this model for CMR on the unseen dataset $\mathcal{D}_u$.

4 Proposed Method

In this section, we first provide an overview of our proposed FGAN. Subsequently, we introduce the main components of FGAN. Lastly, we describe how to train FGAN and perform retrieval based on FGAN.

4.1 Overview

The basic idea of FGAN is to learn a common representation space for samples from different modalities, while at the same time, align the salient fine-grained semantic information among samples. The advantage of aligning more fine-grained information beyond the class label is that it can alleviate the domain-shift problem of transferring the retrieval model from seen classes to unseen classes, so as to achieve better generalization. We consider performing fine-grained alignment from two aspects. First, regarding sample features, we can align features at a finer granularity than the class level to capture the discrepancies among samples within the same class, ensuring that information beyond class labels is effectively aligned. Second, in terms of sample labels, we assume that each sample may carry more diverse information than merely the class label. We can align the label distribution of samples with different modalities in a large label space (e.g., ImageNet labels) to reflect the information carried by each sample.

Specifically, as shown in Figure 2, we employ an FGE, comprising a text encoder and an image encoder, to encode input text or images into two types of representations: region embedding and label distribution. For the training of region embedding, we propose an FGCL strategy. In this strategy, samples of different modalities can be categorized into three types of relations: positive, semi-positive, and negative. By properly adjusting the margins among these three types of relations in the space, the distance among samples in the space can reflect the correlation of samples at a finer granularity than the class level. For the training of label distribution, we propose an FGLA strategy. In this strategy, each image and its corresponding text description are considered to carry the same semantic information. By aligning the label distribution of text and image in such a pair with the pseudo label distribution of the image obtained from the pre-trained model, the final FGAN will know how to effectively estimate the label distribution of input text and image. After FGAN is well trained, region embedding and label distribution can be used for retrieval on unseen classes.

4.2 FGE

Different from previous methods that simply use the off-the-shelf features extracted from the VGG-19 model and the Doc2Vec model pre-trained on Wikipedia corpus, we process the data in an end-to-end manner. As shown in Figure 3, we first use two base encoders E_{base}^{v} and E_{base}^{t} to encode

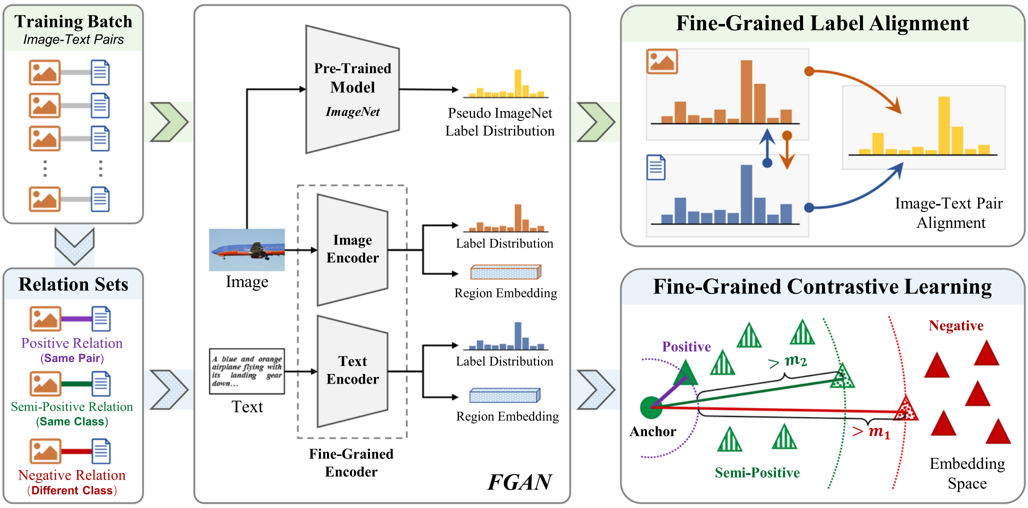


Fig. 2. Overview of the proposed FGAN. The architecture consists of three main components: the FGE, FGCL, and FGLA. The FGE extracts two types of representations: region embeddings from images and label distributions from text. In FGCL, the model utilizes a selective triplet loss to ensure that positive samples are closely aligned while semi-positive and negative samples are appropriately distanced, capturing fine-grained semantic relationships. The FGLA aligns the label distributions of samples with a pseudo ImageNet label distribution obtained from a pre-trained model, preserving the semantic diversity of the data. Together, these components enable FGAN to effectively perform ZS-CMR by leveraging fine-grained semantic information.

the raw images and texts, respectively. Then, we generate region embedding and label distributions based on the region embedding encoders E^v_{re} and E^t_{ld} and label distribution encoders E^v_{ld} and E^t_{ld}, respectively.

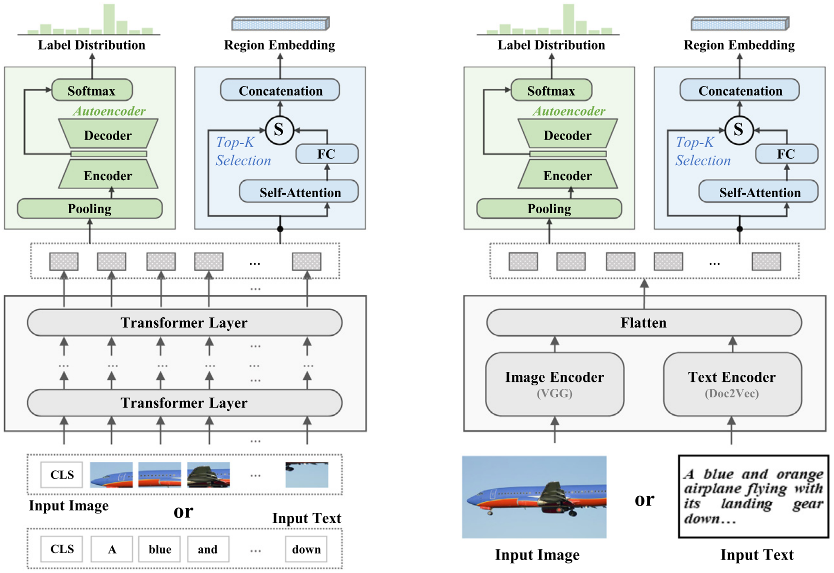
4.2.1 Base Encoder. As shown in Figure 3, to make FGE more generalizable and compatible with different types of network structures, we design two encoding schemes for the base encoder: Transformer-based (e.g., ViT [12] and BERT [11], left part in Figure 3), and non-Transformer-based (e.g., VGG [37] and Doc2Vec [23], right part in Figure 3).

For the non-Transformer-based base encoder, we use VGG to encode image and Doc2Vec to encode text, similar to previous methods. To be consistent with the Transformer-based encoder, we set a flatten operation after VGG and Doc2Vec to generate a sequence of vectors as representation for the input image or text. Specifically, given an input image $\mathbf{V} \in \mathbb{R}^{c \times h \times w}$, we first send it into the VGG model and obtain a feature map $\mathbf{F} \in \mathbb{R}^{d'_v \times h' \times w'}$ from a specified intermediate layer of the VGG model. Then, we flatten the feature map $\mathbf{F}$ into $h' \times w'$ contextualized vectors $\boldsymbol{v}'_i \in \mathbb{R}^{d'_v}$, so that each image can be represented as a sequence of contextual representation $\mathcal{V}'$:

$$\mathcal{V}' = \{\boldsymbol{v}'_1, \boldsymbol{v}'_2, \dots, \boldsymbol{v}'_i, \dots, \boldsymbol{v}'_{h' \times w'}\}, \boldsymbol{v}_i \in \mathbb{R}^{d'_v}, \tag{1}$$

where d'_v, h', and w' are the number of channels, height, and width of the feature map $\mathbf{F}$, respectively. Similarly, given an input text $\mathbf{T}$ with n words, apart from the whole text's feature vector $\boldsymbol{t}'_0$ directly extracted from the Doc2Vec model, we also extract a feature vector $\boldsymbol{t}'_i$ for each word in the text where the feature vector is the word embedding learned in the Doc2Vec model. Then, each text can be represented as a sequence of feature vectors $\mathcal{T}'$:

$$\mathcal{T}' = \{\boldsymbol{t}'_0, \boldsymbol{t}'_1, \dots, \boldsymbol{t}'_i, \dots, \boldsymbol{t}'_n\}, \boldsymbol{t}'_i \in \mathbb{R}^{d'_t}. \tag{2}$$



(a) Transformer-based Model **(b) Non Transformer-based Model**

Fig. 3. Illustration of our FGE. To accommodate different basic encoders, we design two encoding schemes. For the Transformer-based model (left), we first flatten the raw image into sequential image patches before inputting them into the model. In contrast, for the non-Transformer-based model (right), we first extract the feature map and then flatten it into sequential features.

For the Transformer-based base encoder, we use ViT to encode image and BERT to encode text. Both of them can generate a sequence of vectors as representation for the input image or text. In all, both kinds of base encoder can generate a sequence of representation for their input image or text. These sequences of representations (i.e., $\mathcal{V}'$, and $\mathcal{T}'$) can then be used for the following transformations, which ultimately lead to our desired region embedding and label distribution.

4.2.2 Region Embedding. To extract more fine-grained information, we aim to identify the most informative regions in the input image (text) and then combine the representation of these regions to form the final representation of the input image (text), referred to as region embedding. To achieve this goal, we can first evaluate a saliency score for each region to reflect the importance of the fine-grained information contained in each region. Then, we can aggregate the top-K regions with high saliency scores to form the final region embedding. Given the sequence of contextual vectors generated by the base encoder, we can assume that each vector in the sequence corresponds to a region (e.g., an image patch) in the input image (text). As shown in the top-right part of Figure 3(a), the region embedding encoders (i.e., E_{re}^{t} for image and E_{re}^{v} for text) contain two components: region saliency evaluation and top-K region selection.

Region Saliency Evaluation. For the evaluation of region saliency, we use a saliency evaluation network ϕ to predict a saliency score for each region of the image or text, respectively. Specifically, we construct the saliency evaluation network with a **Multi-Head Self-Attention (MSA)** [44] module and a linear layer with a sigmoid function. Given the sequence of contextual representation such as $\mathcal{V}$, $\mathcal{T}$, $\mathcal{V}'$, or $\mathcal{T}'$, which can be uniformly denoted as $\mathbf{X} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{N_r}\}$, the MSA is computed as follows:

$$\mathrm{MSA}(\mathbf{X}) = \mathrm{Concat}(\mathrm{head}_1, \ldots, \mathrm{head}_m)\mathbf{W}_O, \quad \mathrm{head}_i = \mathrm{SA}_i(\mathbf{X}), \tag{3}$$

$$\mathrm{SA}_i(\mathbf{X}) = \mathrm{Softmax}\left(\frac{(\mathbf{X}\mathbf{W}_i^Q)(\mathbf{X}\mathbf{W}_i^K)^\top}{\sqrt{d/m}}\right)(\mathbf{X}\mathbf{W}_i^U), \tag{4}$$

where $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^U \in \mathbb{R}^{d\times\frac{d}{m}}$ and $\mathbf{W}_O \in \mathbb{R}^{d\times d}$ are the transformation matrices, N_r is the number of regions in the sequential representation, d is the dimension of each region, and m is the number of the self-attention heads. Then, we can obtain the saliency scores $\boldsymbol{\epsilon}$ of all regions by mapping the output of the MSA to scalars using a linear layer with the sigmoid function:

$$\boldsymbol{\epsilon} = \mathrm{Sigmoid}(\mathrm{MSA}(\mathbf{X})\mathbf{W}_P), \tag{5}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^{N_r}$ and $\mathbf{W}_P \in \mathbb{R}^{d\times 1}$. The use of MSA ensures that the encoder can capture contextual dependencies between different parts of the input. By formally defining the attention mechanism as a weighted sum of input features, the encoder emphasizes regions with higher saliency scores, thereby optimizing the representation for ZS-CMR.

Top-K Regions Selection. For the selection of top-K regions, given the saliency scores calculated by the saliency evaluation network, we select the K regions with top-K highest scores and then concatenate the contextual representations of these selected regions as the region embedding. To train the whole model using back-propagation, we need to make the top-K selection operation differentiable. We employ the perturbed maximum method [2, 9] for this purpose. Specifically, the differentiable top-K selection operation is equivalent to the following linear program:

$$\begin{gathered}\max_{\mathrm{Y}\in C}\langle \mathrm{Y}, \boldsymbol{\epsilon}\mathbf{1}^\top\rangle \\ \text{where } C = \left\{\mathrm{Y} \in \mathbb{R}^{N_r\times K} : \mathrm{Y}_{n,k} \geq \mathbf{0}, \mathbf{1}^\top\mathrm{Y} = 1, \mathrm{Y}\mathbf{1} \leq \mathbf{1}, \sum_{i\in[N]} i\mathrm{Y}_{i,k} < \sum_{j\in[N]} j\mathrm{Y}_{\mathrm{j,k'}}, \forall k < k'\right\}\end{gathered} \tag{6}$$

where Y are the indicator vectors to be optimized, $\boldsymbol{\epsilon}\mathbf{1}^\top \in \mathbb{R}^{N\times K}$ are the scores that are replicated K times, and C is the convex polytope constraint set. Then, the forward operations can be defined as follows:

$$\mathrm{Y}_\sigma = \mathbb{E}_{\mathbf{Z}}\left[\arg\max_{\mathrm{Y}\in C}\langle \mathrm{Y}, \boldsymbol{\epsilon}\mathbf{1}^\top + \sigma\mathbf{Z}\rangle\right], \tag{7}$$

and the Jacobian for the backward computation during training is defined as follows:

$$J_{\boldsymbol{\epsilon}}\mathrm{Y} = \mathbb{E}_{\mathbf{Z}}\left[\arg\max_{\mathrm{Y}\in C}\langle \mathrm{Y}, \boldsymbol{\epsilon}\mathbf{1}^\top + \sigma\mathbf{Z}\rangle\mathbf{Z}^\top/\sigma\right], \tag{8}$$

where σ is a hyper-parameter, $\mathbf{Z}$ is the noise sampled from the normal distribution. By using the differentiable top-K operation, the model can be optimized using back propagation algorithm during training. The differentiable top-K selection is proved to maintain differentiability and allow for gradient-based optimization, ensuring that the most informative regions are consistently selected during training.

Region Embedding Generation. Finally, the selected regions are concatenated and mapped into the cross-modal embedding space as follows:

$$\boldsymbol{x} = \mathrm{Concat}(\{\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_{s_i}\})\mathbf{W}_r, \tag{9}$$

where s_i is the region index of the ith selected image (text) region and $\mathbf{W}_r \in \mathbb{R}^{(d*K)\times d_x}$ is a learnable mapping matrix.

4.2.3 Label Distribution. To reflect the diverse semantic information carried by the input image (text), we expect to estimate its label distribution in a large label space based on the sequence of contextual representation from base encoder. To achieve this goal, we can first pool the sequence of contextual representation into a unit-length representation which is then embedded into a large label space of ImageNet labels, so that the semantic similarity of two samples (i.e., an image and a text) can be well reflected by the similarity of their label distributions in the label space. As shown in the top-left part of Figure 3(a), the label distribution encoders (i.e., E^v_{ld} for image and E^t_{ld} for text) contain two components: global average pooling and label distribution generation.

Global Average Pooling. We first employ a global average pooling layer to pool the sequence of contextual representations $\mathbf{X} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, \dots, \boldsymbol{r}_{N_r}\}$ into a unit-length representation $\hat{\boldsymbol{r}}$:

$$\hat{\boldsymbol{r}} = \text{GAP}(\{\boldsymbol{r}_1, \boldsymbol{r}_2, \dots, \boldsymbol{r}_{N_r}\}) = \frac{1}{N_r}\sum_{i=1}^{N_r} \boldsymbol{r}_i, \tag{10}$$

where GAP denotes the global average pooling layer, $\boldsymbol{r}_i$ represents the contextual representation of the ith region of the input image or text, and N_r is the total number of regions.

Label Distribution Generation. Then, we incorporate a VAE [22] $A_{\boldsymbol{v}} = \{A^v_E, A^v_D\}$ (or $A_{\boldsymbol{t}} = \{A^t_E, A^t_D\}$ for text) along with a linear layer ζ to compress the pooled representation $\hat{\boldsymbol{r}}_v$ (or $\hat{\boldsymbol{r}}_t$) into a label distribution $\boldsymbol{g}_{\boldsymbol{v}}$ (or $\boldsymbol{g}_{\boldsymbol{t}}$), while preserving the salient semantic information:

$$\boldsymbol{g}^v = \zeta^v(A^v_E(\hat{\boldsymbol{r}}^v)), \quad \boldsymbol{g}^t = \zeta^t(A^t_E(\hat{\boldsymbol{r}}^t)), \tag{11}$$

where A^v_E, A^t_E are the encoders of the VAEs. The FGE is designed to capture detailed semantic information by extracting region embeddings and label distributions from input images and texts. The encoder utilizes attention mechanisms and differentiable top-K selection to focus on the most informative regions, thus preserving fine-grained details that are crucial for distinguishing between similar samples.

4.3 FGCL

To effectively capture the intra-class discrepancy, we propose the FGCL strategy to fine-tune the region embeddings, thereby optimizing the cross-modal embedding space. A meaningful embedding space for retrieval is one in which relevant embeddings are close to each other while irrelevant embeddings are separated, rather than simply putting the same-class embedding into a cluster. Therefore, we design a special *cross-modal selective triplet loss* to adjust the embedding space of region embedding at a finer granularity than the class level.

Specifically, given a batch of M golden pairs of image and text $\{(\boldsymbol{x}^v_i, \boldsymbol{x}^t_i), y_i\}^M_{i=1}$ (i.e., an image and its corresponding text description), we can construct a total of M^2 image–text pairs $\{(\boldsymbol{x}^v_i, \boldsymbol{x}^t_j) | i \in [1, M], j \in [1, M]\}$. These image–text pairs can be divided into three types of relations (i.e., *positive*, *semi-positive*, and *negative*) based on whether the image and text in the pair are from the same golden pairs and whether they belong to the same class:

(1) *Positive Relation*: $i = j$;
(2) *Semi-Positive Relation*: $i \neq j$, but $y_i = y_j$;
(3) *Negative Relation*: $i \neq j$, $y_i \neq y_j$.

Based on these relations given above, for an anchor image embedding $\boldsymbol{x}^v_i$, we can define three sets of text embeddings $\mathrm{S}_P(\boldsymbol{x}^v_i)$, $\mathrm{S}_S(\boldsymbol{x}^v_i)$, and $\mathrm{S}_N(\boldsymbol{x}^v_i)$ that hold *Positive Relation*, *Semi-Positive Relation*, and *Negative Relation* with the anchor image, respectively. Then, we define two kinds of cross-modal triplets for contrastive learning. Taking an image embedding $\boldsymbol{x}^v_i$ as an example anchor, these two kinds of cross-modal triplets can be defined as follows:

$$\langle \boldsymbol{x}_i^v; \boldsymbol{x}_p^t \in \mathrm{S}_P(\boldsymbol{x}_i^v); \boldsymbol{x}_s^t \in \mathrm{S}_S(\boldsymbol{x}_i^v)\rangle, \tag{12}$$

$$\langle \boldsymbol{x}_i^v; \boldsymbol{x}_p^t \in \mathrm{S}_P(\boldsymbol{x}_i^v); \boldsymbol{x}_n^t \in \mathrm{S}_N(\boldsymbol{x}_i^v)\rangle. \tag{13}$$

Based on the above triplets, we formulate our proposed FGCL as a two-fold loss function. First, we aim to ensure that the distance between $\boldsymbol{x}_i^v$ (*anchor*) and $\boldsymbol{x}_p^t$ (*positive*) is smaller than the distance between $\boldsymbol{x}_i^v$ and the *closest* $\boldsymbol{x}_n^t$ (*negative*) by a specific margin:

$$\mathrm{dist}(\boldsymbol{x}_i^v, \boldsymbol{x}_p^t) + m_1 < \min_{\boldsymbol{x}_n^t \in \mathrm{S}_N(\boldsymbol{x}_i^v)} \mathrm{dist}(\boldsymbol{x}_i^v, \boldsymbol{x}_n^t), \tag{14}$$

where m_1 is a margin that is enforced between positive and negative pairs. Then, the ***Anchor-Positive-Negative* (*A-P-N*)** loss function can be defined as follows:

$$\mathcal{L}_{APN} = \frac{1}{M}\sum_{i=1}^{M} \max(m_1 + \mathrm{dist}(\boldsymbol{x}_i^v, \boldsymbol{x}_p^t) - \min_{\boldsymbol{x}_n^t \in \mathrm{S}_N(\boldsymbol{x}_i^v)} \mathrm{dist}(\boldsymbol{x}_i^v, \boldsymbol{x}_n^t), 0). \tag{15}$$

Second, as shown in Figure 1, to model the intra-class discrepancy, we aim to ensure that the distance between $\boldsymbol{x}_i^v$ and $\boldsymbol{x}_p^t$ is smaller than the distance between $\boldsymbol{x}_i^v$ and the *farthest* $\boldsymbol{x}_s^t$ (*semi-positive*) by another specific margin m_2:

$$\mathrm{dist}(\boldsymbol{x}_i^v, \boldsymbol{x}_p^t) + m_2 < \max_{\boldsymbol{x}_s^t \in \mathrm{S}_S(\boldsymbol{x}_i^v)} \mathrm{dist}(\boldsymbol{x}_i^v, \boldsymbol{x}_s^t). \tag{16}$$

In this way, the semantically dissimilar image and text in *Semi-Positive* pairs can be pushed away from each other. Then, the ***Anchor-Positive-Semipositive* (*A-P-S*)** loss function is defined as follows:

$$\mathcal{L}_{APS} = \frac{1}{M}\sum_{i=1}^{M} \max(m_2 + \mathrm{dist}(\boldsymbol{x}_i^v, \boldsymbol{x}_p^t) - \max_{\boldsymbol{x}_s^t \in \mathrm{S}_S(\boldsymbol{x}_i^v)} \mathrm{dist}(\boldsymbol{x}_i^v, \boldsymbol{x}_s^t), 0). \tag{17}$$

Finally, the total loss of FGCL can be calculated as follows:

$$\mathcal{L}_{FGCL} = \mathcal{L}_{APN} + \alpha\mathcal{L}_{APS}, \tag{18}$$

where α is the hyper-parameter that controls the weight of $\mathcal{L}_{APS}$ in the total loss. This FGCL strategy allows for a more nuanced optimization of the cross-modal embedding space, leading to improved performance. FGCL refines the embedding space by explicitly modeling the relationships between positive, semi-positive, and negative samples. This approach allows the model to capture intra-class variability and ensures that the learned embeddings are both discriminative and semantically meaningful. This loss $\mathcal{L}_{FGCL}$ ensures that the embedding space is optimized such that positive samples are closer to each other than negative or semi-positive samples, with the margins m_1 and m_2 controlling the separation.

4.4 FGLA

To effectively capture the diverse semantic information carried by samples, we propose the FGLA strategy that utilizes VAEs and a pre-trained image classification model to optimize the fine-grained label distribution. The VAE framework, with its encoder–decoder architecture, allows for the reconstruction of input features while also regularizing the latent space to approximate a prior distribution. Therefore, we employ VAE to generate compact and informative label distributions for images and texts. VAEs are particularly suited for this task because they can learn a probabilistic latent space that captures the underlying semantic structure of the data. This regularization is crucial for generating meaningful label distributions that align with the pseudo labels. While two image–text pairs may belong to the same class, the objects they carry can be very different. Previous methods often simply align the representation of image and text to the label embedding of their belonging class, neglecting the diverse and fine-grained semantic information in image and text. To address this, we propose to align the label distribution of an image or text to a meaningful pseudo ImageNet label distribution generated by the pre-trained model, maintaining the semantic diversity of samples.

4.4.1 Pseudo Label Generation. First, we use a pre-trained ImageNet [10] classification model $\mathcal{I}$ to generate a pseudo soft label τ for each image:

$$\tau = \mathcal{I}(\mathbf{V}), \quad \tau \in \mathbb{R}^{1000}. \tag{19}$$

The pseudo label is a 1000D probability distribution reflecting the likelihood of the image containing objects from the 1,000 different ImageNet classes. This label serves as a reference for aligning the label distributions of image and text, reflecting the semantic diversity of real-world scenes. Although objects in the retrieval images may not appear in the pre-defined 1,000 ImageNet classes, most objects in real-world scenes are semantically related to one or more ImageNet classes (i.e., having similar appearance or co-occurrence). Therefore, the pseudo label τ provides useful semantic information for matching image and text on unseen classes.

4.4.2 Label Distribution Calculation. Given the pooled image feature $\hat{\boldsymbol{r}}_i^v$ of the image $\mathbf{V}$ and pooled text feature $\hat{\boldsymbol{r}}^t$ of the text $\mathbf{T}$, we utilize VAEs to calculate the label distributions $\boldsymbol{g}^v$ and $\boldsymbol{g}^t$ (as described in Section 4.2.3). These label distributions, generated by the VAEs, provide a compact and comprehensive representation of the images and texts:

$$\boldsymbol{g}^v = \zeta^v(A_E^v(\hat{\boldsymbol{r}}^v)), \quad \boldsymbol{g}^t = \zeta^t(A_E^t(\hat{\boldsymbol{r}}^t)), \tag{20}$$

where ζ^v and ζ^t are two linear layers that map the latent embeddings to the 1000D label embeddings.

4.4.3 FGLA Loss. The FGLA loss is defined as the sum of the VAE loss and the alignment loss.

VAE Loss. We first calculate the VAE loss, which is the sum of the reconstruction loss and the **Kullback–Leibler (KL)** divergence loss for both image and text features. The VAE loss encourages the VAEs to learn a compact and informative latent representation of the input features while maintaining the overall structure of the data. The VAE loss is defined as follows:

$$\mathcal{L}_{VAE} = \mathcal{L}_{recon}^v + \mathcal{L}_{recon}^t + \mathcal{L}_{KL}^v + \mathcal{L}_{KL}^t, \tag{21}$$

where $\mathcal{L}_{recon}^v$ and $\mathcal{L}_{recon}^t$ are the reconstruction losses for image and text features, respectively, defined as follows:

$$\mathcal{L}_{recon}^v = \left\|A_v^D(\boldsymbol{g}^v) - \hat{\boldsymbol{r}}_i^v\right\|_2, \quad \mathcal{L}_{recon}^t = \left\|A_t^D(\boldsymbol{g}^t) - \hat{\boldsymbol{r}}_i^t\right\|_2, \tag{22}$$

and $\mathcal{L}_{KL}^v$ and $\mathcal{L}_{KL}^t$ are the KL divergence losses for image and text features, respectively, defined as follows:

$$\mathcal{L}_{KL}^v = \mathrm{D_{KL}}(q^v(z|\hat{\boldsymbol{r}}^v)\|p(z)), \quad \mathcal{L}_{KL}^t = \mathrm{D_{KL}}(q^t(z|\hat{\boldsymbol{r}}^t)\|p(z)), \tag{23}$$

where $\mathrm{D_{KL}}$ is the KL divergence, $q^v(z|\hat{\boldsymbol{r}}_i^v)$ and $q^t(z|\hat{\boldsymbol{r}}_i^t)$ are the approximate posterior distributions of the latent variables $\boldsymbol{z}$ for image and text features, respectively, and $p(z)$ is the prior distribution of the latent variables.

Alignment Loss. We then align the label distribution of image $\mathbf{V}$ and text $\mathbf{T}_i$ in the same golden pair (i.e., image and its corresponding text description) to the image's pseudo soft label by minimizing the following alignment loss:

$$\mathcal{L}_{align} = \mathcal{L}_{align}^v + \mathcal{L}_{align}^t, \tag{24}$$

where $\mathcal{L}^v_{align}$ and $\mathcal{L}^t_{align}$ represent the alignment loss for the image and text, respectively. These losses are defined as follows:

$$\mathcal{L}^v_{align} = \frac{1}{2}\left(\mathrm{KL}(\tau \| \boldsymbol{g}^v) + \mathrm{KL}(\boldsymbol{g}^v \| \boldsymbol{g}^t)\right), \quad \mathcal{L}^t_{align} = \frac{1}{2}\left(\mathrm{KL}(\tau \| \boldsymbol{g}^t) + \mathrm{KL}(\boldsymbol{g}^t \| \boldsymbol{g}^v)\right). \tag{25}$$

The point-wise KL divergence loss is defined as follows:

$$\mathrm{KL}(\tau_i \| \boldsymbol{g}_i) = \tau_i \log \frac{\tau_i}{\boldsymbol{g}_i}. \tag{26}$$

Finally, the total loss of FGLA can be calculated as follows:

$$\mathcal{L}_{FGLA} = \mathcal{L}_{VAE} + \mathcal{L}_{align}. \tag{27}$$

This loss function guides the VAEs to generate meaningful label distributions that align with the pseudo ImageNet label distribution. This loss ensures that the label distributions of images and texts are aligned with the pseudo label distribution, preserving semantic diversity and improving generalization to unseen classes.

4.5 Model Training

In summary, we can jointly train the FGEs corresponding to image and text by simultaneously minimizing the loss functions of two fine-grained alignment strategies FGCL and FGLA. The full objective function of our FGAN model is:

$$\mathcal{L} = \mathcal{L}_{FGCL} + \mu \mathcal{L}_{FGLA}, \tag{28}$$

where μ is the hyper-parameter set to balance the optimization of $\mathcal{L}_{FGCL}$ and $\mathcal{L}_{FGLA}$. Given a batch of training data, our model first generates the region embeddings and label distributions using FGE. Then, the loss of FGCL and FGLA is calculated. Finally, we sum the losses together and optimize the model with the back-propagation algorithm.

4.6 Joint Retrieval

At the retrieval stage, the similarity between an image and a text can be measured by jointly using their region embeddings and label distributions. Specifically, given an unseen query text $\mathbf{T}_i$ and an unseen image set $\{\mathbf{V}_j\}_{j=1}^N$, the FGE can output their region embeddings $\boldsymbol{x}^t_i$, $\{\boldsymbol{x}^v_j\}_{j=1}^N$ and label distributions $\boldsymbol{g}^t_i$, $\{\boldsymbol{g}^v_j\}_{j=1}^N$, respectively. Then, for each image $\mathbf{V}_j$, the final similarity between $\mathbf{T}_i$ and $\mathbf{V}_j$ is:

$$\mathrm{Sim}(\mathbf{T}_i, \mathbf{V}_j) = \mathrm{Sim}_{\mathrm{RE}}(\mathbf{T}_i, \mathbf{V}_j) + \mathrm{Sim}_{\mathrm{LD}}(\mathbf{T}_i, \mathbf{V}_j), \tag{29}$$

where $\mathrm{Sim}_{\mathrm{RE}}$ and $\mathrm{Sim}_{\mathrm{LD}}$ are the similarity functions measured on region embedding and label distribution, respectively:

$$\mathrm{Sim}_{\mathrm{RE}}(\mathbf{T}_i, \mathbf{V}_j) = \frac{\boldsymbol{x}^t_i \cdot \boldsymbol{x}^{v\top}_j}{\|\boldsymbol{x}^t_i\| \|\boldsymbol{x}^v_j\|}, \tag{30}$$

$$\mathrm{Sim}_{\mathrm{LD}}(\mathbf{T}_i, \mathbf{V}_j) = \tanh(-(\mathrm{KL}(\boldsymbol{g}^t_i \| \boldsymbol{g}^v_j) + \mathrm{KL}(\boldsymbol{g}^v_j \| \boldsymbol{g}^t_i))), \tag{31}$$

where tanh is the hyperbolic tangent function that maps the results to $(-1, 1)$. Finally, we can perform joint retrieval according to the final similarity between the query text and the support images; the images sorted in descending order according to similarities will be returned as retrieval results.

5 Experiments

In this section, we first introduce the datasets, the evaluation metric used in the experiments and the implementation details of our method. Then, to validate the effectiveness of our proposed method, we make extensive comparisons with existing methods. Finally, we conduct ablation studies and display some visualization results to thoroughly investigate the effectiveness of each component in our model.

5.1 Experimental Setup

5.1.1 Datasets. To verify the effectiveness of our method, we evaluate our model on four widely used CMR datasets: Pascal Sentence [15], Wikipedia [34], PKU-XMediaNet [32], and NUS-WIDE [8], and two sketch-based image retrieval datasets: Sketchy [28] and TU-Berlin [13]. We briefly introduce the general information of these datasets as follows:

- *Pascal Sentence dataset* contains 1,000 image/text pairs organized into 20 categories which are selected from 2008 PASCAL development kit. Each image is described with five textual sentences. The dataset is randomly split into two parts: 800 image/text pairs selected as training set and 200 image/text pairs as the testing set.
- *Wikipedia dataset* contains 2,866 image/text pairs crawled from the Wikipedia web site with 10 high-level semantic categories (e.g., art, biology, sport). Of them, 2,173 image/text pairs are selected as training set and 693 image/text pairs are selected as testing set.
- *PKU-XMediaNet dataset* is a large-scale dataset which contains 40,000 image/text pairs with 200 categories. The categories are selected from WordNet to ensure the semantic hierarchy structure. The dataset is randomly split into two parts: 32,000 image/text pairs for training and 8,000 image/text pairs for testing.
- *NUS-WIDE dataset* contains about 270,000 images categorized into 81 tags. Only images from the 10 largest categories were selected, resulting in around 70,000 image/text pairs, each with a unique category label.
- *Sketchy dataset* is composed of 75,471 sketches and 73,002 images from 125 classes.
- *TU-Berlin dataset* consists of 20,000 sketches and 204,489 images, which shows a severe imbalance between the numbers of sketches and images.

Following the dataset setting in the previous works [6, 27, 55], for each dataset, the training set and testing set in the original split setting are further divided into two sets: seen category set and unseen category set, respectively, and each set includes 50% categories. The unseen category set in the testing set is used as the query set and the unseen category set in the training set is used as the support set to perform the zero-shot retrieval task. Two zero-shot retrieval scenarios are considered: **Image-to-Text (I2T)** and **Text-to-Image (T2I)**. While I2T takes images as queries to retrieve texts, T2I takes texts as queries to retrieve images. Finally, we report the performance on I2T and T2T, and their averaged performance for overall evaluation.

5.1.2 Evaluation Metric. Following previous work, we use the **Mean Average Precision (MAP)** as the evaluation metric, which is calculated as the mean of the **Average Precision (AP)** for each query. Take T2I retrieval as an example, given a query text $\mathbf{T}_i$, the AP is calculated as follows:

$$AP = \frac{1}{N}\sum_{k=1}^{K} \mathrm{P}(k) \times \mathrm{RL}(k), \tag{32}$$

where N is the number of query text's relevant images (i.e., images in the same classes as the query text) in the support set, K is the size of the support set, $\mathrm{P}(k)$ is the precision at cut-off k

Table 1. Comparison of Our Model with the Baseline and Existing CMR and ZS-CMR Methods

Method	Image Encoder	Text Encoder	Pascal Sentence			Wikipedia			PKU-XMediaNet		
			I2T	T2I	Avg	I2T	T2I	Avg	I2T	T2I	Avg
DCCA [58]	TF-IDF	LDA	0.207	0.183	0.195	0.238	0.237	0.236	0.031	0.044	0.038
DeepSM [52]	SIFT	LDA	0.276	0.251	0.264	0.265	0.258	0.262	0.040	0.096	0.068
ACMR [45]	VGG	BoW	0.306	0.291	0.299	0.276	0.262	0.269	0.036	0.043	0.040
MASLN [56]	VGG	BoW	0.307	0.294	0.301	0.284	0.264	0.274	0.040	0.045	0.043
DSCMR [65]	VGG	Doc2Vec	0.327	0.319	0.323	0.312	0.280	0.296	0.096	0.051	0.074
DANZCR [6]	VGG	Doc2Vec	0.334	0.338	0.336	0.297	0.287	0.292	0.106	0.117	0.112
DADN [7]	VGG	Doc2Vec	0.359	0.353	0.356	0.305	0.291	0.298	0.112	0.130	0.121
TANSS [55]	VGG	Doc2Vec	0.351	0.365	0.358	0.313	0.289	0.301	0.108	0.120	0.114
LCALE [27]	VGG	Doc2Vec	0.414	0.394	0.404	0.367	0.357	0.362	0.135	0.154	0.150
CFSA [54]	VGG	Doc2Vec	0.378	0.368	0.373	0.341	0.311	0.326	0.121	0.152	0.137
AAEGAN [57]	VGG	Doc2Vec	0.437	0.412	0.425	0.395	0.346	0.370	0.126	0.154	0.140
ILSA [48]	VGG	Doc2Vec	0.440	0.418	0.429	0.397	0.364	0.381	0.135	0.165	0.150
MDVAE [40]	VGG	Doc2Vec	0.440	0.418	0.429	0.404	0.362	0.383	0.135	**0.167**	0.151
FGAN$_{RE}$	VGG	Doc2Vec	0.459	0.425	0.442	0.392	0.362	0.377	0.141	0.132	0.137
FGAN$_{LD}$	VGG	Doc2Vec	0.430	0.400	0.415	0.390	0.355	0.372	0.138	0.123	0.131
FGAN$_{JR}$	VGG	Doc2Vec	**0.493**	**0.457**	**0.475**	**0.411**	**0.383**	**0.397**	**0.161**	0.147	**0.154**

The best results are marked in bold. FGAN$_{RE}$ represents the retrieval results using only the region embeddings. FGAN$_{LD}$ represents the retrieval results using only the label distributions. FGAN$_{JR}$ represents the retrieval results using our proposed joint retrieval method.

in the list, and RL(k) is an indicator function equaling 1 if the item at rank k is a relevant image, 0 otherwise [42].

5.1.3 Implementation Details. Our models are implemented using the Pytorch [31] and Huggingface [53] libraries. The model is optimized using the AdamW Optimizer [30] with a learning rate of 1e-4. Additionally, the cosine annealing scheduler is adopted to adjust the learning rates and the warm-up strategy is performed at the first 10 iterations. We fine-tune the model with a batch size of 16 on small-scale datasets as follows: 40 epochs on Wikipedia and Pascal Sentence, and 20 epochs each on PKU-XMediaNet, NUS-WIDE, Sketchy, and TU-Berlin. The training process for FGAN takes approximately 16.3 minutes for Wikipedia, 4.7 minutes for Pascal Sentence, 38.3 minutes for PKU-XMediaNet, 32.1 minutes for NUS-WIDE, 81.3 minutes for Sketchy, and 24.7 minutes for TU-Berlin datasets. We set hyper-parameters $\sigma = 0.05$, $\alpha = 0.1$, and $\mu = 1$ for experiments on all dataset. The setting for $\sigma = 0.05$ is consistent with the same setting in [2], while $\alpha = 0.1$ balances the different parts of $\mathcal{L}_{FGCL}$ and $\mathcal{L}$ to maintain the same order of magnitude. For the m_1 and m_2 and the number of regions in $\mathcal{L}_{FGCL}$, we investigate their influence on the model performance in Section 5.3.3. Following the setting in [27], all the experiments are conducted 10 times under the same configurations to make a fair comparison. The experiments are conducted on a machine equipped with an NVIDIA Tesla V100 GPU featuring 32 GB of VRAM, utilizing mixed precision training with FP16.

5.2 Comparison with Existing Methods

To ensure a fair comparison with previous methods, we employ VGG and Doc2Vec as our base encoders. Additionally, for our FGAN, we consider three retrieval approaches: using only the region embeddings (i.e., FGAN$_{RE}$), using only the label distributions (i.e., FGAN$_{LD}$), and using both (i.e., FGAN$_{JR}$). Table 1 presents the following observations:

Table 2. Retrieval Results of the Image and Text Encoders of the FGAN Model Replaced with More Advanced Encoders ViT, BERT, and CLIP on the Pascal Sentence Dataset

Image Encoder	Text Encoder	Region Embedding			Label Distribution			Joint Retrieval		
		I2T	**T2I**	**Avg**	**I2T**	**T2I**	**Avg**	**I2T**	**T2I**	**Avg**
VGG	Doc2Vec	0.459	0.425	0.442	0.430	0.400	0.415	0.493	0.457	0.475
ViT	Doc2Vec	0.530	0.500	0.515	0.407	0.425	0.416	0.533	0.521	0.527
VGG	BERT	0.492	0.454	0.473	0.418	0.403	0.411	0.511	0.487	0.499
ViT	BERT	0.553	0.547	0.550	0.429	0.469	0.449	0.592	0.568	0.580
$\text{CLIP}^{\text{visual}}_{\text{ViT-B/16}}$	$\text{CLIP}^{\text{text}}_{\text{ViT-B/16}}$	**0.601**	**0.589**	**0.595**	**0.451**	**0.488**	**0.470**	**0.639**	**0.682**	**0.660**

Bold values indicate the best performance for each metric across the different encoder combinations.

(1) The MAP scores of the ZS-CMR methods are significantly higher than those of the traditional CMR methods, indicating that traditional CMR methods are not suitable for the ZS-CMR task due to the domain-shift problem between seen and unseen data.
(2) Retrieval using either the region embeddings or label distributions of our model separately (i.e., FGAN_{RE} and FGAN_{LD} in Table 1) achieves comparable performance with previous methods, validating the effectiveness of our proposed FGE, FGCL, and FGLA.
(3) Our method with joint retrieval (i.e., FGAN_{JR} in Table 1) attains the best MAP scores and outperforms all baselines on all three datasets. Specifically, on the Pascal Sentence dataset, our method surpasses the previous state-of-the-art MDVAE [40] by 12.0% and 9.3% on the I2T and T2I tasks, respectively. On the Wikipedia dataset, our method outperforms MDVAE by 1.7% and 5.8% on the I2T and T2I tasks, respectively. On the large-scale PKU-XMediaNet dataset, our method still outperforms MDVAE on average performance. These results demonstrate that both region embeddings and label distributions learned by our method are effective for retrieval, and joint retrieval with region embeddings and label distributions is more discriminative and generalizable for retrieval across seen and unseen classes.

We further investigate whether our method is compatible with more advanced Transformer-based encoders such as ViT and BERT. To accomplish this, we replace the VGG and Doc2Vec encoders with ViT and BERT encoders in our model, respectively. The results are shown in Table 2. We observe that by replacing the encoders with more advanced CLIP [33] visual and text encoders, our method's performance can be further enhanced. By utilizing these advanced encoders, our method can better capture the semantic meaning of the input data, leading to improved performance. To demonstrate the novelty and effectiveness of our proposed approach, we further compare our method with recent ZS-CMR and **Zero-Shot Sketch-Based Image Retrieval (ZS-SBIR)** methods. For our experiments on the ZS-SBIR task, we modified our method by replacing the text encoder with an image encoder to process sketch images. As shown in Table 3, our experimental findings indicate that our approach achieves state-of-the-art performance when compared to the latest methods in the field. Specifically, we surpass several recent models in key metrics such as mAP and precision, further demonstrating the effectiveness and novelty of our proposed approach.

5.3 Model Analysis

We further analyze the effectiveness of each component in our proposed FGAN by removing or replacing one or more of them. Table 4 presents the averaged results of I2T and T2I tasks on three datasets, where three retrieval schemes are tested: retrieval using only region embeddings, retrieval using only label distributions, and joint retrieval (i.e., using both region embeddings and label distributions together).

Table 3. Comparison of Our Model with Recent ZS-CMR on the NUS-WIDE Dataset and ZS-SBIR Methods on the Sketchy and TU-Berlin Datasets

Model	**Sketchy**		**TU-Berlin**		**NUS-WIDE**		
	mAP@all	Prec@100	mAP@all	Prec@100	mAP (I2T)	mAP (T2I)	mAP (Avg)
MDVAE (Tian et al. [40])	0.613	0.723	0.486	0.612	0.604	0.595	0.599
ILSA (Wang et al. [48])	0.667	0.754	0.483	0.593	0.609	0.599	0.604
SMF (Ho et al. [20])	0.642	0.770	0.507	0.620	-	-	
OAN (Zhang et al. [63])	0.617	0.737	0.505	625	-	-	
URHNL (Yong et al. [60])	-	-	-	-	0.642	0.644	0.643
CMAAM (Su et al. [39])	0.730	0.809	0.560	0.665	-	-	-
Ours	**0.737**	0.797	**0.573**	**0.675**	**0.661**	**0.655**	**0.658**

ZS-SBIR, Zero-Shot Sketch-Based Image Retrieval. Bold values indicate the best performance for each metric across the compared models.

Table 4. Model Analysis of FGAN

Model Setting	**Pascal Sentence**			**Wikipedia**			**PKU-XMediaNet**		
	RE	**LD**	**JR**	**RE**	**LD**	**JR**	**RE**	**LD**	**JR**
Full Model	**0.442**	**0.415**	**0.475**	**0.377**	**0.372**	**0.397**	**0.137**	**0.131**	**0.154**
— Loss of FGLA	0.437	0.115	0.436	0.364	0.233	0.364	0.133	0.011	0.133
— Loss of FGCL	0.125	391	0.397	0.264	0.325	0.318	0.014	0.132	0.123
— E_{re}^{v} and E_{re}^{v} (using global average pooling)	0.429	0.397	0.448	0.356	0.361	0.374	0.125	0.129	0.137
— ϕ (using random selection)	0.336	0.402	0.419	0.319	0.361	0.366	0.053	0.125	0.125
— A_v and A_t (using linear layer)	0.437	0.393	0.457	0.375	0.361	0.385	0.135	0.117	0.147
— A-P-S loss (using only A-P-N loss)	0.437	0.401	0.463	0.364	0.370	0.388	0.133	0.131	0.152
— Selective triplets (using all triplets)	0.371	0.400	0.409	0.328	0.359	0.357	0.033	0.121	0.118
— Triplet loss (using pairwise loss)	0.417	0.399	0.448	0.341	0.359	0.371	0.081	0.127	0.132
— Contrastive learning (using classification loss)	0.343	0.401	0.377	0.291	0.347	0.341	0.131	0.117	0.141
— Pseudo soft label (using random label)	0.441	0.115	0.437	0.377	0.225	0.373	0.107	0.013	0.087
— Pseudo soft label (using one-hot label)	0.442	0.121	0.439	0.376	0.220	0.381	0.094	0.025	0.097
— ImageNet label (using word embedding)	0.442	0.216	0.436	0.371	0.243	0.371	0.099	0.031	0.093

The average MAP scores are reported. RE, LD, and JR represent performing retrieval using region embeddings, label distributions, and both of them, respectively. The setting "— Region Selection (using global average pooling)" means that the region selection strategy is ablated form the framework and global average pooling is adopted. Bold values indicate the best performance for each metric across different settings.

5.3.1 Ablation of Loss Function. As demonstrated in the first three rows of Table 4, both our proposed FGLA and FGCL contribute to the performance improvement of our FGAN method. For instance, without $\mathcal{L}_{FGLA}$, retrieval using region embeddings still performs well on the datasets. However, using label distributions that are not optimized by FGLA almost performs like a random selection, and the joint retrieval results using region embeddings and label distributions slightly degrade the performance compared to retrieval using only region embeddings. These results indicate that the proposed FGLA and FGCL strategies can work independently, with the FGCL strategy successfully optimizing the region embeddings and the FGLA strategy optimizing the label distributions. Furthermore, using them simultaneously can further improve the MAP scores, suggesting that the two strategies are effective and complementary.

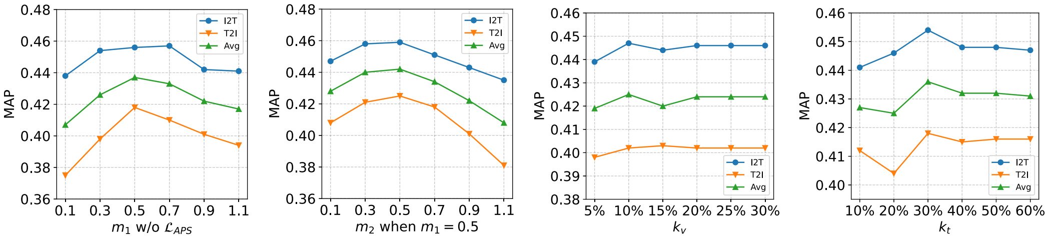


Fig. 4. The I2T and T2I retrieval results using region embeddings with different m_1 and m_2 and using different numbers of regions in the FGCL on the Pascal Sentence dataset.

5.3.2 Effect of Designments in FGE. To study the effectiveness of FGE, we compare our method with the following settings: (1) global average pooling over all regions instead of E^v_{re} and E^t_{re} for the generation of region embeddings; (2) random region selection instead of according to the saliency evaluation network ϕ in E^v_{re} and E^t_{re}; and (3) removing the AEs A_v and A_t in E^v_{ld} and E^v_{ld}, respectively, and using a linear layer to map the representation after GAP to 1000D as the label distributions.

As shown in Table 4, the retrieval performance gets worse after replacing E^v_{re} and E^t_{re} with the global average pooling operation over all regions, and the model fails to retrieve meaningful samples after replacing the saliency evaluation network ϵ with random selection in E^v_{re} and E^t_{re} on all three datasets. The retrieval results indicate the effectiveness of E^v_{re} and E^t_{re} in the generation of region embeddings and the necessity of mining the semantically meaningful regions to form the region embeddings. Moreover, removing the AEs in E^v_{ld} and E^t_{ld} and using the representation generated by a linear layer as the label distributions also lead to performance degradation, which verifies the effectiveness of incorporating A_v and A_t in E^v_{ld} and E^t_{ld} for the generation of label distributions.

5.3.3 Effect of Designments in FGCL. In this part, we conduct ablation studies on the proposed FGCL strategy by comparing the FGCL strategy with four variants: (1) removing the $\mathcal{L}_{APS}$ in $\mathcal{L}_{FGCL}$; (2) using all triplets instead of the closest and farthest triplets; (3) replacing the triplets contrastive loss with the pairwise contrastive loss; and (4) replacing $\mathcal{L}_{FGCL}$ with the cross entropy classification loss.

As shown in Table 4, the FGCL strategy achieves the best results among all the settings. Specifically, without the $\mathcal{L}_{APS}$ in $\mathcal{L}_{FGCL}$, the retrieval performance decreases on all three datasets. Using all triplets for $\mathcal{L}_{FGCL}$ degrades retrieval performance significantly. This is mainly because the semantic similarity between samples in different triplets is also different; simply maximizing the margin in all *A-P-S* or *A-P-N* triplets will affect the learning of the common embedding space. Furthermore, replacing the triplets contrastive loss with the pairwise contrastive loss leads to a decrease on the retrieval using region embeddings, and replacing $\mathcal{L}_{FGCL}$ with the cross entropy classification loss results in the most significant performance degradation. In addition, we further investigate the influence of different m_1 and m_2 in the FGCL strategy, and the ration of selected image regions k_v and text regions k_t for the learning of region embeddings. To avoid the effect of number of text regions when conducting experiments on various image regions, we replace E^t_{re} with global average pooling, and *vice versa*. The results are shown in Figure 4. It can be seen that using $m_1 = 0.5$, $m_2 = 0.5$, 10% image regions, and 30% text regions could achieve the best performance.

5.3.4 Effect of Designments in FGLA. In this part, we validate the effectiveness of the FGLA strategy. We compare our FGLA method with three settings: (1) replacing the pseudo soft label with random data sampled from normal distribution; (2) replacing the pseudo soft label with one-hot

Table 5. Performance with Different Image and Text Base Encoder in the FGE

Image Encoder	Text Encoder	Trainable Params (M)	FLOPs (G)	Memory (GB)	Speed	mAP$_{\text{Avg}}$
ResNet-18	Doc2Vec	9.5	6.0	0.72	28.8	0.461
VGG-19	Doc2Vec	9.5	23.8	2.4	18.3	0.475
ViT	Doc2Vec	9.2	21.5	20.7	16.9	0.527
VGG-19	BERT	10.2	25.2	2.7	16.4	0.499
ViT	BERT	10.0	22.9	29.7	15.8	0.580
CLIP$^{\text{visual}}_{\text{ViT-B/16}}$	CLIP$^{\text{text}}_{\text{ViT-B/16}}$	7.7	22.1	25.6	15.9	0.660

"FLOPs" denotes the floating point operations per second during inference, "Memory" denotes the required GPU memory of inference with the batch size set to 16. "Speed" denotes the inference samples per second. "mAP$_{\text{Avg}}$" denotes the average mAP score.

label predicted by the pre-trained ImageNet model; and (3) replacing the FGCL with pre-defined label embedding like previous methods.

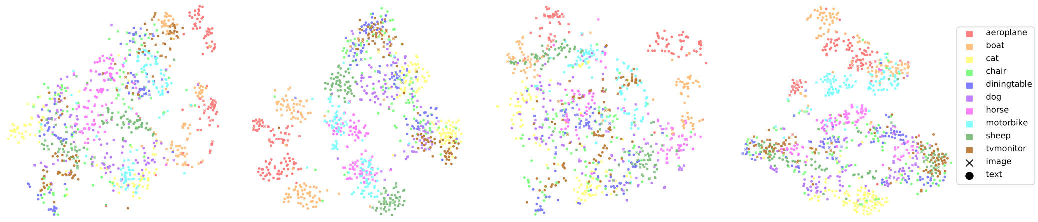
As shown in the last three rows of Table 4, the model fails to learn meaningful label distribution with randomly generated labels, and the soft label distributions outperform one-hot label distributions on all three datasets. Compared to the pre-defined label embedding method that previous methods commonly adopted, our FGLA strategy achieves better results. These results indicate that the pseudo soft labels can provide useful semantic information for label distribution learning and thus is beneficial to the ZS-CMR task.

5.3.5 Model Complexity and Scalability. In this part, we investigate the scalability and applicability in real-time or resource-constrained environments of our model. The primary sources of model parameters and computational complexity stem from the backbones used for visual and text encoding within the FGE. To address varying resource scenarios, we can utilize different-sized backbones, allowing for flexibility based on available computational resources. As shown in Table 5, we conducted experiments comparing different backbones and ablation settings. The table summarizes the model's computational load, parameter count, required GPU memory, inference speed, and performance metrics (mAP$_{\text{Avg}}$) across various configurations. This demonstrates that our model can efficiently adapt to real-time and resource-constrained environments while maintaining competitive performance. By selecting appropriate backbones, we can optimize our model for various applications, ensuring it remains practical under different operational constraints.

5.4 Qualitative Analysis

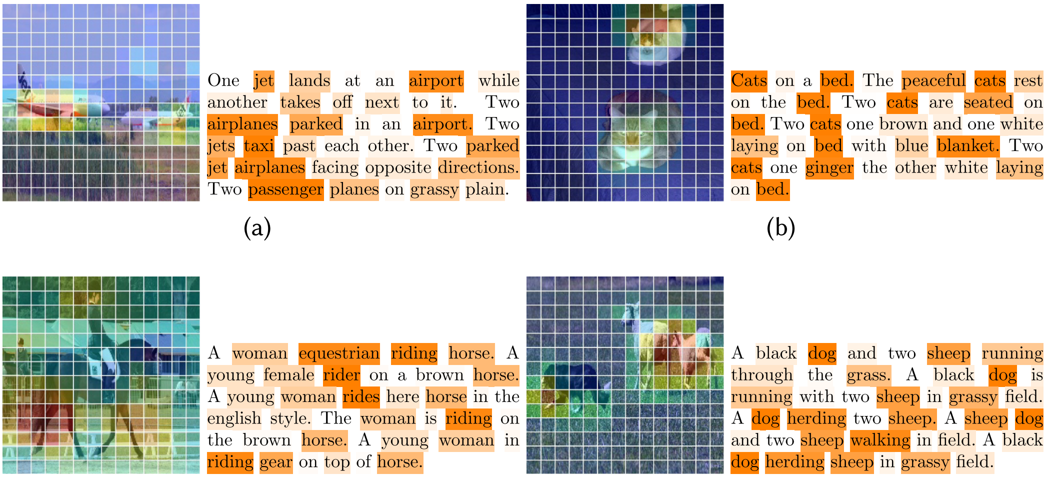
Finally, we conduct a qualitative analysis for our proposed method. We visualize the learned embeddings of both region embeddings and label distributions and the weights produced by the saliency evaluation network.

5.4.1 Visualizing Learned Embeddings with t-SNE. Figure 5 presents the visualization of region embeddings and label distributions before and after training, using the t-SNE tool [43]. From the figure, it is evident that the untrained image and text region embeddings, as well as label distributions, are far apart in the embedding space. In contrast, the trained image and text embeddings from the same classes are closely situated in the learned common embedding space. Furthermore, after training, most samples are accurately separated into different semantic clusters based on their classes. Additionally, semantically similar classes are also positioned closely in the learned embedding space, for example, *dog* and *cat*, and *sheep* and *horse*. This visualization result demonstrates that our FGAN method effectively aligns data from different modalities and constructs a meaningful semantic embedding space.



(a) untrained region embeddings (b) trained region embeddings (c) untrained label distributions (d) trained label distributions

Fig. 5. The t-SNE visualization results of the unseen region embeddings and label distributions before and after training under a random data split. The ● symbol denotes the text features and the × symbol denotes the image features. Clusters of different color belong to different classes.



One jet lands at an airport while another takes off next to it. Two airplanes parked in an airport. Two jets taxi past each other. Two parked jet airplanes facing opposite directions. Two passenger planes on grassy plain.

(a)

Cats on a bed. The peaceful cats rest on the bed. Two cats are seated on bed. Two cats one brown and one white laying on bed with blue blanket. Two cats one ginger the other white laying on bed.

(b)

A woman equestrian riding horse. A young female rider on a brown horse. A young woman rides here horse in the english style. The woman is riding on the brown horse. A young woman in riding gear on top of horse.

(c)

A black dog and two sheep running through the grass. A black dog is running with two sheep in grassy field. A dog herding two sheep. A sheep dog and two sheep walking in field. A black dog herding sheep in grassy field.

(d)

Fig. 6. The visualization results of the weights produced by the scorner net. For each sample, the left is the image split into 196 patches and the right is the sentences describing the image. Patches or words with higher scores are marked with deeper color.

5.4.2 Visualizing Weights from the Saliency Evaluation Network. Figure 6 displays the weights of each region in the texts and images generated by the saliency evaluation network for four test image–text pairs from different classes. This visualization is used to verify if our model can effectively identify salient semantic regions. As evident in the images, the salient objects receive higher scores, while the surrounding backgrounds have comparatively lower scores. Likewise, in the texts, words describing objects, appearance, actions, and states tend to obtain high scores as well. These visualizations suggest that the proposed FGE and FGCL methods can effectively capture fine-grained meaningful regions while discarding irrelevant noise and backgrounds in both images and texts.

5.4.3 Case Study. In this subsection, we provide an analysis of specific instances where our model did not perform well. We selected three examples from the test sets and returned their top five retrieval results, as shown in Figure 7. It can be observed that failures in these cases

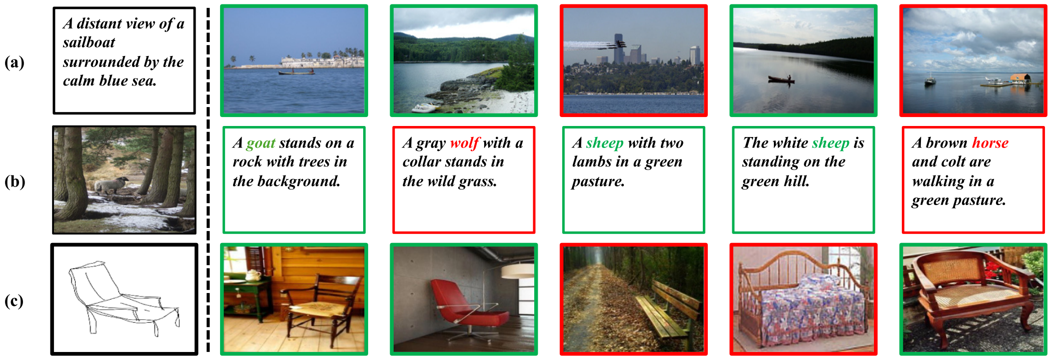

(a)
A distant view of a sailboat surrounded by the calm blue sea.
(b)
A goat stands on a rock with trees in the background.
A gray wolf with a collar stands in the wild grass.
A sheep with two lambs in a green pasture.
The white sheep is standing on the green hill.
A brown horse and colt are walking in a green pasture.
(c)

Fig. 7. Retrieval results for three test cases. Each case shows the top five retrieval results for a given query. Results marked with a green rim indicate correct matches, while those with red rims indicate incorrect matches.

stem from the inherent similarities between the concepts in the query and the retrieval results. Additionally, this phenomenon is reflected in the t-SNE visualization (Figure 5), where samples from similar categories are positioned close together in the embedding space. These insights suggest that enhancing the model's ability to differentiate between closely related categories could improve performance in future work.

6 Conclusion

In this article, we propose an FGAN to learn representations with fine-grained alignment for ZS-CMR. Specifically, we propose an FGE to encode the images and texts into their region embeddings and label distributions. Then, we propose an FGCL strategy to fine-tune the region embeddings, which can better model intra-class discrepancy and measure the similarity of samples within the same class. Moreover, we propose an FGLA strategy to fine-tune the label distributions, which can further capture the rich semantic information carried by single sample. Finally, both region embeddings and label distributions are utilized together to perform ZS-CMR at a fine granularity. Experimental results on three widely used datasets demonstrate that our method outperforms the state-of-the-art methods by a large margin.

References

[1] Philip Bachman, R. Devon Hjelm, and William Buchwalter. 2019. Learning representations by maximizing mutual information across views. *Advances in Neural Information Processing Systems* 32 (2019), 15535–15545.

[2] Quentin Berthet, Mathieu Blondel, Olivier Teboul, Marco Cuturi, Jean-Philippe Vert, and Francis Bach. 2020. Learning with differentiable perturbed optimizers. *Advances in Neural Information Processing Systems* 33 (2020), 9508–9519.

[3] Yue Cao, Mingsheng Long, Jianmin Wang, Qiang Yang, and Philip S. Yu. 2016. Deep visual-semantic hashing for cross-modal retrieval. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1445–1454.

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1597–1607.

[5] Xinlei Chen and Kaiming He. 2021. Exploring simple Siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15750–15758.

[6] Jingze Chi and Yuxin Peng. 2018. Dual adversarial networks for zero-shot cross-media retrieval. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 663–669.

[7] Jingze Chi and Yuxin Peng. 2019. Zero-shot cross-media embedding learning with dual adversarial distribution network. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 4 (2019), 1173–1187.

[8] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. Nus-wide: A real-world web image database from national university of Singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, 1–9.

[9] Jean-Baptiste Cordonnier, Aravindh Mahendran, Alexey Dosovitskiy, Dirk Weissenborn, Jakob Uszkoreit, and Thomas Unterthiner. 2021. Differentiable patch selection for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2351–2360.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 248–255.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805. Retrieved from https://arxiv.org/abs/1810.04805

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929. Retrieved from https://arxiv.org/abs/2010.11929

[13] Mathias Eitz, James Hays, and Marc Alexa. 2012. How do humans sketch objects? *ACM Transactions on Graphics* 31, 4 (2012), 1–10.

[14] Kaipeng Fang, Jingkuan Song, Lianli Gao, Pengpeng Zeng, Zhi-Qi Cheng, Xiyao Li, and Heng Tao Shen. 2024. Pros: Prompting-to-simulate generalized knowledge for universal cross-domain retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17292–17301.

[15] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1778–1785.

[16] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. arXiv:2104.08821. Retrieved from https://arxiv.org/abs/2104.08821

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM* 63, 11 (2020), 139–144.

[18] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In defense of the triplet loss for person re-identification. arXiv:1703.07737. Retrieved from https://arxiv.org/abs/1703.07737

[19] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2018. Learning deep representations by mutual information estimation and maximization. arXiv:1808.06670. Retrieved from https://arxiv.org/abs/1808.06670

[20] Yi-Hsuan Ho, Der-Lor Way, and Zen-Chung Shih. 2023. Sharing model framework for zero-shot sketch-based image retrieval. In *Computer Graphics Forum*, Vol. 42. Wiley Online Library, e14947.

[21] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. arXiv:2004.11362. Retrieved from https://arxiv.org/abs/2004.11362

[22] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. arXiv:1312.6114. Retrieved from https://arxiv.org/abs/1312.6114

[23] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning*. PMLR, 1188–1196.

[24] Jason D. Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. 2021. Predicting what you already know helps: Provable self-supervised learning. *Advances in Neural Information Processing Systems* 34 (2021), 309–323.

[25] Hao Li, Jingkuan Song, Lianli Gao, Xiaosu Zhu, and Hengtao Shen. 2024. Prototype-based aleatoric uncertainty quantification for cross-modal retrieval. *Advances in Neural Information Processing Systems* 36 (2024), 24564–24585.

[26] Kun Li, Meng Lin, Songlin Hu, and Ruixuan Li. 2022. CLZT: A contrastive learning based framework for zero-shot text classification. In *Proceedings of the International Conference on Database Systems for Advanced Applications*. Springer, 623–630.

[27] Kaiyi Lin, Xing Xu, Lianli Gao, Zheng Wang, and Heng Tao Shen. 2020. Learning cross-aligned latent embeddings for zero-shot cross-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 11515–11522.

[28] Li Liu, Fumin Shen, Yuming Shen, Xianglong Liu, and Ling Shao. 2017. Deep sketch hashing: Fast free-hand sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2862–2871.

[29] Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. arXiv:1803.02893. Retrieved from https://arxiv.org/abs/1803.02893

[30] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. arXiv:1711.05101. Retrieved from https://arxiv.org/abs/1711.05101

[31] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. arXiv:1912.01703. Retrieved from https://arxiv.org/abs/arXiv:1912.01703

[32] Yuxin Peng, Xin Huang, and Yunzhen Zhao. 2017. An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 9 (2017), 2372–2385.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*. PMLR, 8748–8763.

[34] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*, 251–260.

[35] Abhishek Sharma, Abhishek Kumar, Hal Daume, and David W. Jacobs. 2012. Generalized multiview analysis: A discriminative latent space. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2160–2167.

[36] Wenxue Shen, Jingkuan Song, Xiaosu Zhu, Gongfu Li, and Heng Tao Shen. 2023. End-to-end pre-training with hierarchical matching and momentum contrast for text-video retrieval. *IEEE Transactions on Image Processing* 32 (2023), 5017–5030.

[37] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556. Retrieved from https://arxiv.org/abs/1409.1556

[38] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. *Advances in Neural Information Processing Systems* 29 (2016), 1857–1865.

[39] Hanwen Su, Ge Song, Kai Huang, Jiyan Wang, and Ming Yang. 2024. Cross-modal attention alignment network with auxiliary text description for zero-shot sketch-based image retrieval. In *Proceedings of the International Conference on Artificial Neural Networks*. Springer, 52–65.

[40] Jialin Tian, Kai Wang, Xing Xu, Zuo Cao, Fumin Shen, and Heng Tao Shen. 2022. Multimodal disentanglement variational AutoEncoders for zero-shot cross-modal retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 960–969.

[41] Yonglong Tian, Chen Sun, Ben Poole, Dilip Krishnan, Cordelia Schmid, and Phillip Isola. 2020. What makes for good views for contrastive learning? *Advances in Neural Information Processing Systems* 33, 6827–6839.

[42] Andrew Turpin and Falk Scholer. 2006. User performance versus precision measures for simple search tasks. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 11–18.

[43] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 11.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017), 6000–6010.

[45] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM International Conference on Multimedia*, 154–162.

[46] Cheng Wang, Haojin Yang, and Christoph Meinel. 2015. Deep semantic mapping for cross-modal retrieval. In *Proceedings of the 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 234–241.

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

[48] Kai Wang, Yifan Wang, Xing Xu, Zuo Cao, and Xunliang Cai. 2022. Instance-level semantic alignment for zero-shot cross-modal retrieval. In *Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 1–6.

[49] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. 2016. A comprehensive survey on cross-modal retrieval. arXiv:1607.06215. Retrieved from https://arxiv.org/abs/1607.06215

[50] Peng Wang, Kai Han, Xiu-Shen Wei, Lei Zhang, and Lei Wang. 2021. Contrastive learning based hybrid networks for long-tailed image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 943–952.

[51] Xiaolong Wang and Abhinav Gupta. 2015. Unsupervised learning of visual representations using videos. In *Proceedings of the IEEE International Conference on Computer Vision*, 2794–2802.

[52] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. 2016. Cross-modal retrieval with CNN visual features: A new baseline. *IEEE Transactions on Cybernetics* 47, 2 (2016), 449–460.

[53] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. arXiv:1910.03771. Retrieved from https://arxiv.org/abs/1910.03771

[54] Xing Xu, Kaiyi Lin, Huimin Lu, Lianli Gao, and Heng Tao Shen. 2020. Correlated features synthesis and alignment for zero-shot cross-modal retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1419–1428.

[55] Xing Xu, Huimin Lu, Jingkuan Song, Yang Yang, Heng Tao Shen, and Xuelong Li. 2019. Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval. *IEEE Transactions on Cybernetics* 50, 6 (2019), 2400–2413.

[56] Xing Xu, Jingkuan Song, Huimin Lu, Yang Yang, Fumin Shen, and Zi Huang. 2018. Modal-adversarial semantic learning network for extendable cross-modal retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 46–54.

[57] Xing Xu, Jialin Tian, Kaiyi Lin, Huimin Lu, Jie Shao, and Heng Tao Shen. 2021. Zero-shot cross-modal retrieval by assembling AutoEncoder and generative adversarial network. *ACM Transactions on Multimedia Computing, Communications, and Applications* 17, 1s (2021), 1–17.

[58] Fei Yan and Krystian Mikolajczyk. 2015. Deep correlation for matching images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3441–3450.

[59] Fan Yang, Zheng Wang, Jing Xiao, and Shin'ichi Satoh. 2020. Mining on heterogeneous manifolds for zero-shot cross-modal image retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34, 12589–12596.

[60] Kailing Yong, Zhenqiu Shu, and Zhengtao Yu. 2024. Unpaired robust hashing with noisy labels for zero-shot cross-modal retrieval. *Engineering Applications of Artificial Intelligence* 133 (2024), 108197.

[61] Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. 2020. Graph contrastive learning with augmentations. *Advances in Neural Information Processing Systems* 33 (2020), 5812–5823.

[62] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. arXiv:2205.01917. Retrieved from https://arxiv.org/abs/2205.01917

[63] Haoxiang Zhang, He Jiang, Ziqiang Wang, and Deqiang Cheng. 2023. Ontology-aware network for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '23)*. IEEE, 1–5.

[64] Haonan Zhang, Pengpeng Zeng, Lianli Gao, Jingkuan Song, and Heng Tao Shen. 2024. UMP: Unified modality-aware prompt tuning for text-video retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 11 (Nov. 2024), 11954–11964.

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

Received 1 July 2024; revised 7 December 2024; accepted 27 February 2025