Learning Cross-Aligned Latent Embeddings for Zero-Shot Cross-Modal Retrieval

Kaiyi Lin,[1] **Xing Xu,**[1]* **Lianli Gao,**[1] **Zheng Wang,**[1] **Heng Tao Shen**[1]
[1]Center for Future Media and School of Computer Science and Engineering
University of Electronic Science and Technology of China, China

Abstract

Zero-Shot Cross-Modal Retrieval (ZS-CMR) is an emerging research hotspot that aims to retrieve data of new classes across different modality data. It is challenging for not only the heterogeneous distributions across different modalities, but also the inconsistent semantics across seen and unseen classes. A handful of recently proposed methods typically borrow the idea from zero-shot learning, *i.e.*, exploiting word embeddings of class labels (*i.e.*, class-embeddings) as common semantic space, and using generative adversarial network (GAN) to capture the underlying multimodal data structures, as well as strengthen relations between input data and semantic space to generalize across seen and unseen classes. In this paper, we propose a novel method termed Learning Cross-Aligned Latent Embeddings (LCALE) as an alternative to these GAN based methods for ZS-CMR. Unlike using the class-embeddings as the semantic space, our method seeks for a shared low-dimensional latent space of input multimodal features and class-embeddings by modality-specific variational autoencoders. Notably, we align the distributions learned from multimodal input features and from class-embeddings to construct latent embeddings that contain the essential cross-modal correlation associated with unseen classes. Effective cross-reconstruction and cross-alignment criterions are further developed to preserve class-discriminative information in latent space, which benefits the efficiency for retrieval and enable the knowledge transfer to unseen classes. We evaluate our model using four benchmark datasets on image-text retrieval tasks and one large-scale dataset on image-sketch retrieval tasks. The experimental results show that our method establishes the new state-of-the-art performance for both tasks on all datasets.

1 Introduction

With the explosive growth of multimedia data (*e.g.*, texts, images, videos and audios) in our daily life, cross-modal retrieval, which aims to support information retrieval across different modality data, has become an important research area. The typical retrieval scenarios include image-text retrieval (Deng et al. 2018; Peng et al. 2018) and image-sketch retrieval (Shen et al. 2018; Dey et al. 2019). The key challenge of cross-media retrieval is that the distributions and representations of different media types are inconsistent, making it hard to measure the similarity between different modalities. A majority of existing approaches utilize the training data from different modalities to learn a common subspace representation that enables the retrieval by using a suitable distance metric.

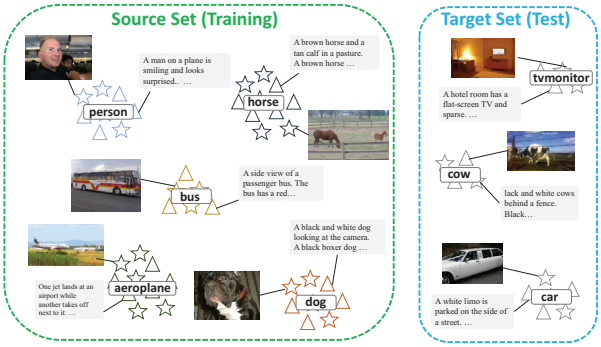


Figure 1: A brief illustration of ZS-CMR, where the multi-modal instances of the seen and unseen classes are disjoint.

Most of existing cross-modal algorithms (Rasiwasia et al. 2010; Wang et al. 2017) consider that data from the same set of classes are provided during training and testing (retrieval). However, this assumption cannot always hold in real-world dynamic scenario where the database is enlarging with data of new classes continuously. In the current pipeline for cross-modal retrieval, to achieve a good retrieval performance, every time a new class is added to the database, the current algorithms need to be re-trained from scratch, which is computationally inefficient. Therefore, we feel that the generalizability of a model is crucial to unseen new classes. To this end, this paper focuses on the *zero-shot cross-modal retrieval* (ZS-CMR) problem. It aims to perform retrieval across multiple modality data in the zero-shot setting, where new categories are supported to retrieve only with limited categories for training. A brief illustration of ZS-CMR is shown in Fig. 1. Compared with traditional cross-modal retrieval, ZS-CMR is more challenging since it needs to handle not only the heterogeneous distributions of different modal-

*Corresponding author

ity data, but also the inconsistent semantics across seen and unseen classes.

Since ZS-CMR is a new research hotspot, only a few studies have been proposed very recently on typical scenarios of image-text retrieval (Xu et al. 2018; Chi and Peng 2018; 2019) and image-sketch retrieval (Dey et al. 2019; Dutta and Akata 2019). These methods are typically inspired by zero-shot learning (ZSL), *i.e.*, exploiting class-embeddings that are easily extracted from linguist corpus (*e.g.*, Wikipedia), to build semantic space and enabling the knowledge transfer across seen and unseen classes. Though it has shown promising results to utilize the semantic space of class-embeddings in traditional ZSL approaches for the classification task, it may not be optimal in the retrieval scenario like ZS-CMR. Besides, they commonly adopt the popular generative adversarial network (GAN) (Goodfellow et al. 2014) as a basic module to generate common embeddings in the adversarial training process, so that the common embeddings can capture the heterogeneous distributions of different modality data. Unfortunately, the GAN-based loss functions in these methods suffer from instability in training.

In this paper, to address the above issues, we propose a novel framework dubbed Learning Cross-Aligned Latent Embeddings (LCALE), which is an alternative to existing GAN based ZS-CMR methods (Chi and Peng 2019; Dutta and Akata 2019; Dey et al. 2019). A brief overview of our proposed LCALE is shown in Fig. 2. Specifically, rather than directly using the class-embeddings as the semantic space, we train a multimodal variational autoencoder (VAE) that consisting of three modality-specific VAEs (Kingma and Welling 2014; Schönfeld et al. 2019) to encode the features of different modalities as well as the class-embeddings in a low-dimensional latent embedding space. The learned latent embeddings are aligned by matching their parameterized distributions and by enforcing a cross-reconstruction criterion via the decoders in each VAE. Consequently, by explicitly enforcing alignment both in the latent embeddings learned using different modalities with two parallel cross-alignment schemes, the VAEs enable knowledge transfer to unseen classes. Meanwhile, a cycle consistency constraint is employed to the class-embeddings reconstruction to further enhance the semantic consistency between each of image and text modalities with the class-embeddings. These essential components jointly ensure our LCALE method to obtain an effective latent embedding space for ZS-CMR.

Our contributions can be summarized as follows:

- We propose an alternative to existing GAN-based approaches for ZS-CMR. Instead of generating multimodal input features, we generate latent embeddings in a low-dimensional latent space via correlated variational autoencoders, which achieves both stable training and superior retrieval performance.
- We develop cross-reconstruction and cross-alignment schemes for matching the learned latent embeddings for different modality data and the class-embeddings, which effectively enhance the latent embeddings space and enable the knowledge transfer to unseen classes in the space.
- We conduct sufficient experiments on five widely-used datasets in total for two cross-modal retrieval scenarios: image-text retrieval and image-sketch retrieval, showing the effectiveness of our proposed method on both tasks and its new state-of-the-art performance.

2 Related Work

Cross-Modal Retrieval. The main effort on cross-modal retrieval is to represent data of different modalities with common representations, to eliminate the "modality gap". Existing methods can be grouped into *traditional methods* and *DNN-based methods* according to the difference of basic models. The traditional methods (Rasiwasia et al. 2010; Wang et al. 2013; 2016) mainly learn linear projections to maximize the correlation between the pairwise data of different modalities, which project the features of different modalities into one common space to generate common representations. Recent work (Wang et al. 2015; Yan and Mikolajczyk 2015; Peng et al. 2018) have made great efforts to apply deep neural network (DNN) to standard cross-modal retrieval. The DNN based methods are usually more advantaged to model nonlinearity in multimodal data distributions and can incorporate various information into common representation learning. Recently, several cross-modal GAN studies (Wang et al. 2017; Deng et al. 2018; Zhu et al. 2019; Wang et al. 2019) have also been proposed for cross-modal retrieval. They utilize the strong ability of GAN to model heterogeneous data distribution with inconsistent semantics and learn discriminative common representations of different modalities via the adversarial training process.

Zero-Shot Cross-Modal Retrieval. As discussed in Sec. 1, existing cross-modal retrieval methods may not generalize well on the ZS-CMR task, where the test data belongs to unseen classes that have no overlap with the training data. Recently, a few work (Xu et al. 2018; Chi and Peng 2018; 2019) focus on image-text retrieval scenario have made initial steps on ZS-CMR. Following the pipeline in existing zero-shot learning (ZSL) studies (Felix et al. 2018; Schönfeld et al. 2019) for unimodal data, they also exploit the class-embeddings as external knowledge, and perform zero-shot learning and correlation learning at the same time for zero-shot cross-modal retrieval. The difference in these approaches is how to utilize the class-embeddings: in (Xu et al. 2018), class-embeddings are used as condition information for auto-encoder reconstruction, while in (Chi and Peng 2018), they are directly treated as to be learned semantic space for retrieval.

On the other hand, three recent studies explore another task of image-sketch retrieval, which is also known as sketch-based image retrieval (SBIR) (Yelamarthi et al. 2018). These methods mainly address the unique characteristics such as intrinsic visual sparsity and large intra-class variance in this task. Nevertheless, they still adopt the similar learning architecture as in the above image-text retrieval task, *i.e.*, a common semantic space of class-embeddings is chosen in which both the image or sketch features are projected by GAN model or autoencoders. Additional constraints such as cycle-consistency (Dutta and Akata 2019)

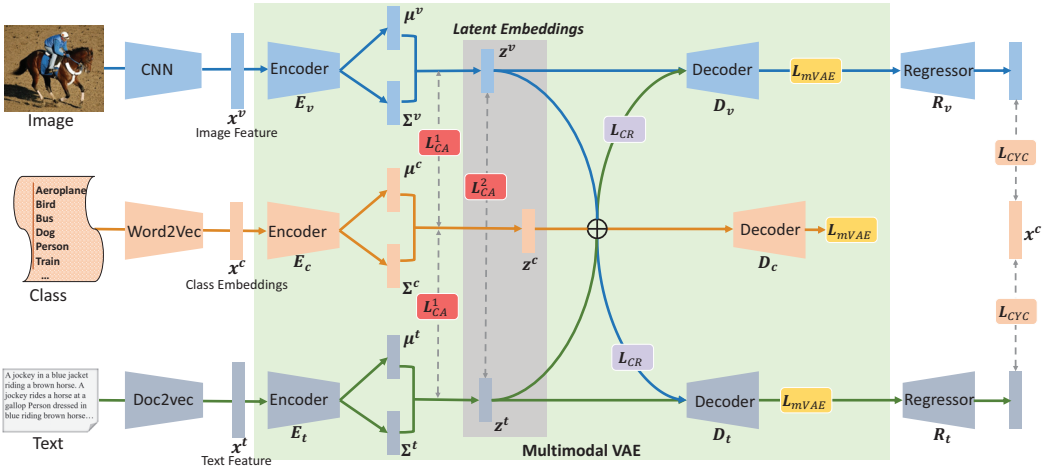


Figure 2: The flowchart of our proposed LCALE method.

and domain disentanglement (Dey et al. 2019) are developed to capture the cross-modal association and mitigate the modality gap between different modalities.

In this work, instead of using GAN, we choose VAE as the backbone and design a multimodal VAE structure of three modality-specific VAE to generate aligned latent embeddings of different modalities in the latent space. Effective cross-reconstruction and cross-alignment criterions are developed to ensure that more effective latent embeddings are learned, as well as the knowledge transfer is also accomplished. Compared to the existing approaches, our proposed method is advanced to be applied to both image-text retrieval and image-sketch retrieval tasks.

3 Proposed Method

3.1 Problem Formulation

We first describe a formal definition of the cross-modal retrieval task. Our goal is to learn a common subspace from the training multimodal data of seen classes and then apply the learned subspace to generate common representations for the test data of unseen classes. Without loss of generality, we consider bimodal data (*i.e.*, image and text) and assume that we have a set of seen classes data consisting of N multimodal instances, *i.e.*, $\mathcal{O}_s = \{o_i\}_{i=1}^{N_s}$, $o_i = (x_i^{(v)}, x_i^{(t)}, x_i^{(c)}, y_i)$, where $x_i^{(v)}$, $x_i^{(t)}$, $x_i^{(c)}$ and y_i denote its image feature vector, text feature vector, class-embeddings and the class label, respectively. Similarly, we also have a set of unseen classes data $\mathcal{O}_u = \{o_j\}_{j=1}^{N_u}$, $o_j = (x_j^{(v)}, x_j^{(t)}, x_j^{(c)}, y_j)$ for testing, where the notations are consistently defined as those in the seen classes set. Note that under the zero-shot setting, the class labels $\{y_i\}_{i=1}^{N_s} \in \mathcal{Y}_s$ in the seen class set and in the unseen class set $\{y_j\}_{j=1}^{N_u} \in \mathcal{Y}_u$ are disjoint, *i.e.*, $\mathcal{Y}_s \bigcap \mathcal{Y}_u = \phi$.

3.2 Our LCALE Approach

Model Architecture. The overall framework of our LCALE method is illustrated in Fig. 2. It is constructed by integrating three modality-specific VAEs (respectively for modalities of image, text, and class-embeddings) into one multimodal VAE (mVAE) through a shared latent embedding space. In particular, each VAE contains an encoder and a decoder, where the encoder maps the original input features into the latent embedding space, and the decoder decodes the latent embeddings to any other modality data of the same class to achieve cross-reconstruction. Then, we employ the class-embeddings as a bridge and develop two cross-alignment schemes to achieve the robust knowledge transfer of the latent embeddings across pairwise modalities. Furthermore, class-embeddings reconstruction with cycle consistency constraint is performed on two regressors of image and text modalities, to further enhance the semantic consistency between each of two modalities with the class-embeddings. The detailed procedure of our LCALE method is depicted as follows.

Multimodal Variational Autoencoder (mVAE). We employ VAE, an effective generative prototype, as the basic encoder-decoder module for each modality. A standard VAE (Kingma and Welling 2014) is decomposed into an encoder that obtains low-dimensional latent variable z from input data x and a decoder that obtains output x' close to x from z. Typically, the variational inference is adopted in VAE to find the true conditional probability distribution $p(z|x)$ over the latent variable z. Due to the intractability of $p(z|x)$, its closest proxy posterior $q(z|x)$ is used as the approximation, by minimizing the distance of $q(z|x)$ and $p(z|x)$ using a variational lower bound. Thus, the objective function of a VAE is the variational lower bound on the marginal likelihood of input data x, as:

$$\mathcal{L}_{VAE} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{KL}(q(z|x)||p(z)). \quad (1)$$

The former term denotes the reconstruction error and the latter is a prior regularizer term that measures the degree of Kullback-Leibler (KL) divergence (Higgins et al. 2017) as D_{KL}, respectively. The conditional probability distribution $q(z|x)$, $p(x|z)$ are the forms of the encoder and decoder. Besides, $p(z)$ is the prior distribution of z modeled as the mul-

tivariate Gaussian distribution; μ and Σ are the mean and variance of the posterior distribution $q(z|x) = \mathcal{N}(\mu, \Sigma)$.

In our LCALE method, we combine three individual VAEs to a multimodal VAE (mVAE) structure to learn a shared latent embedding space of different modalities. Each modality-specific VAE is expected to encode the input modality data to the latent space and then reconstruct the original data through the decoder with minimal information loss. Formally, our mVAE sums the losses in the three modality-specific VAEs as:

$$\mathcal{L}_{mVAE} = \sum_{m} \mathbb{E}_{q(z|x^{(m)})}[\log p(x^{(m)}|z)] - D_{KL}(q(z|x^{(m)})||p(z)), \quad (2)$$

where $m = \{v, t, c\}$ denotes the three modality data, *i.e.*, images, text and class-embeddings, respectively.

Cross-Reconstruction with Latent Embeddings. Underlying the shared latent embedding space, the reconstruction can not only be accomplished in individual modality (as in Eq. 2), but also can be established across different modalities. Notably, our mVAE allow reconstructing the modality data of an instance $x^{(m)}$, by decoding the latent embeddings of a different instance $x^{(n)}$ from another modality of the same class. It is intuitive that the latent embeddings of the instances of the same class are expected to semantically consistent even though they come from different modalities. Then, the cross-reconstruction loss of three modalities can be derived as:

$$\mathcal{L}_{CR} = \sum_{m} \sum_{n \neq m} \|x^{(n)} - D_n(E_m(x^{(m)}))\|_2^2, \quad (3)$$

where $m, n = \{v, t, c\}$, $E_m(\cdot)$ denotes the encoder of the m-th modality, and $D_n(\cdot)$ is the decoder for the n-th modality.

Cross-Alignment in Latent Embedding Space. Moreover, to ensure the consistency of the latent embeddings of different modalities in the shared latent space, we develop two parallel cross-alignment schemes.

1) For the first scheme, we take the class-embeddings as a bridge to align the multivariate Gaussian distributions of the latent embeddings across pairwise modalities. Specifically, the 2-Wasserstein distance (Givens and Shortt 1984) is adopted as the alignment criterion between two modality distributions as:

$$W_{mn} = [||\mu_m - \mu_n||_2^2 + Tr(\Sigma_m) + Tr(\Sigma_n) - 2(\Sigma_m^{\frac{1}{2}} \Sigma_m \Sigma_n^{\frac{1}{2}})^{\frac{1}{2}}]^{\frac{1}{2}}, \quad (4)$$

where m and n represent different modality type, *i.e.*, $\{v, t, c\}$. As the diagonal covariance matrices predicted by an encoder is commutative, the Eq. 4 can be simplified to

$$W_{mn} = (||\mu_m - \mu_n||_2^2 + ||\Sigma_m^{\frac{1}{2}} - \Sigma_n^{\frac{1}{2}}||_F^2))^{\frac{1}{2}}, \quad (5)$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Finally, we apply Eq. 5 to our special case of minimizing the 2-Wasserstein distance between pairwise modalities, *i.e.*, image *vs.* class-embeddings, and text *vs.* class-embeddings. Then the cross-alignment loss can be derived as:

$$\mathcal{L}_{CA}^1 = W_{vc} + W_{tc}. \quad (6)$$

We can see that by explicitly enforcing the cross alignment of the latent embeddings across pairwise modalities, the mVAE enables knowledge transfer to unseen classes with the connection of the seen classes.

2) As the associations between image and text modalities are implicitly built through class-embeddings, here we additionally consider another scheme to explicitly enhance the semantic correlation of the two modalities. In particular, we measure the maximum mean discrepancy (MMD) (Huang, Peng, and Yuan 2018; Felix et al. 2018) of the pairwise latent embeddings of image and text modalities. The MMD criterion is a kernel-based distance function measuring the relevance of pairwise instances that have recently been explored in cross-modal analysis. Finally, The loss for our second scheme is formulated as:

$$\mathcal{L}_{CA}^2 = ||\mathbb{E}_p[\kappa(z^{(v)})] - \mathbb{E}_q[\kappa(z^{(t)})]||_{\mathcal{H}_k}^2. \quad (7)$$

Here p and q denote the distributions of latent embeddings $z^{(v)}$ and $z^{(t)}$ in image and text modalities. κ is the feature map with canonical form $\kappa(x) = k(x, \cdot)$, and $\mathcal{H}_k$ is the reproducing kernel Hilbert space (RKHS) endowed with a Gaussian kernel $k(x, x') = \exp(-\sigma||x - x'||)^2$. Eq. 7 can be regarded as shrinking the gap between two modalities.

Class-embeddings Reconstruction with Cycle Consistency. In order to further enhance the semantic consistency of different modalities features of the same class in the latent embedding space, we adopt the cycle consistency constraint (Felix et al. 2018) to ensure the robust reconstruction of the class-embeddings, which is derived as:

$$\mathcal{L}_{CYC} = \mathbb{E}[\| x^{(c)} - R_v(D_v(z^{(v)})) \|_2^2] + \mathbb{E}[\| x^{(c)} - R_t(D_t(z^{(t)})) \|_2^2], \quad (8)$$

where $R_v(\cdot)$, $R_t(\cdot)$ denotes the regressors following the two VAEs of image and text modalities, which respectively map the outputs of the decoders from two modalities to the space of class-embeddings.

Overall Objective and Optimization. According to the above definitions, the final objective function of our proposed method can be formulated by a linear combination of the above five loss terms as:

$$\mathcal{L} = \mathcal{L}_{mVAE} + \alpha \mathcal{L}_{CR} + \beta \mathcal{L}_{CA}^1 + \lambda \mathcal{L}_{CA}^2 + \gamma \mathcal{L}_{CYC}. \quad (9)$$

where α, β, λ and γ are hyper-parameters that balance the contribution of each loss terms. For simplicity, we use θ_E, θ_D and θ_R to denote the parameters in all encoders, decoders and regressors in the mVAE. The stochastic gradient descent algorithm (*i.e.*, Adam optimizer (Kingma and Ba 2014)) is utilized to optimize the three types of parameters alternatively. The detailed procedure is shown in Algorithm 1.

4 Experiment

4.1 Experimental Setup

Datasets and Features. To verify the effectiveness of our proposed method, we conduct experiments under two cross-modal retrieval scenarios: *image-text retrieval* and *image-sketch* retrieval. The *image-text* retrieval is evaluated on four

Algorithm 1 Training procedure of the proposed LCALE.

Input: $\mathcal{O}_s = \{(x_i^{(v)}, x_i^{(t)}, x_i^{(c)}, y_i)\}_{i=1}^{N_s}$, batch size B, hyper-parameters α, β, λ, γ, learning rate μ.
Output: Model parameters θ_E, θ_D, θ_R

1: **repeat**
2: Sample multimodal pairs $\{(x_b^{(v)}, x_b^{(t)}, x_b^{(c)}, y_b)\}_{b=1}^{B}$ in $\mathcal{O}_s$ with batch.
3: Update θ_E by $\theta_E \leftarrow \theta_E - \mu\nabla_{\theta_E}(\mathcal{L}_{mVAE} + \alpha\mathcal{L}_{CR} + \beta\mathcal{L}_{CA}^1 + \lambda\mathcal{L}_{CA}^2 + \gamma\mathcal{L}_{CYC})$.
4: Update θ_D by $\theta_D \leftarrow \theta_D - \mu\nabla_{\theta_D}(\mathcal{L}_{mVAE} + \alpha\mathcal{L}_{CR} + \beta\mathcal{L}_{CA}^1 + \lambda\mathcal{L}_{CA}^2 + \gamma\mathcal{L}_{CYC})$.
5: Update θ_R by $\theta_R \leftarrow \theta_R - \mu\nabla_{\theta_R}(\gamma\mathcal{L}_{CYC})$.
6: **until** The Eq. 9 converges or reach maximum iterations.
7: Use the encoders E_v and E_t to map instances in $\mathcal{O}_u$ to the latent embedding space for ZS-CMR.

widely-used cross-modal datasets, named *Wikipedia* (Rasiwasia et al. 2010), *Pascal Sentence* (Rashtchian, Young, and Hockenmaier 2010), *NUS-WIDE* (Chua et al. 2009) and *PKU-XMediaNet* (Huang, Peng, and Yuan 2018). We follow the dataset split and feature exaction strategies from (Chi and Peng 2018). Specifically, we adopt a 19-layer VGGNet (Simonyan and Zisserman 2014) to extract 4,096D feature vector by the fc7 layer of VGGNet on all datasets. The feature vector for each text is extracted from Doc2Vec (Le and Mikolov 2014) model pre-trained on Wikipedia with 300D, and the 300D word-embeddings for classes are extracted by Word2Vec (Mikolov et al. 2013) model pre-trained on Google News.

For *image-sketch* retrieval task, we follow the dataset split and feature extraction settings in (Dutta and Akata 2019) to perform experiments on the *Sketchy* dataset. Specifically, we extract sketch/image features from the VGG-16 (Simonyan and Zisserman 2014) network model pre-trained on ImageNet and apply an attention mechanism inspired by (Song et al. 2017) to obtain 512D feature vector. The statistical information of the five datasets are summarized in Table 1.

Table 1: The general statistics of all datasets. Here "*/*" denotes the number of seen/unseen "Classes", and the number of image/text (or sketch) in "Train" and "Test", respectively.

Datasets	Classes	Train	Test
Wikipedia	5/5	2,173/2,173	693/693
Pascal Sentences	10/10	800/800	200/200
NUS-WIDE	5/5	42,941/42,941	28,661/28,661
PKU-XMediaNet	100/100	32,000/32,000	8,000/8,000
Sketchy	100/25	58,376/61,060	14,626/14,421

Evaluation Metric. For *image-text* retrieval, following (Xu et al. 2018; Chi and Peng 2019) we evaluate cross-modal retrieval on two scenarios: image-to-text (Img2Txt) and text-to-image (Txt2Img) that take one modality data as query, *i.e.*, images (text), to retrieve related items in the other modality, *i.e.*, texts (images). The widely-used mean average precision (MAP) score computed from all returned results are used as the evaluation measure, as it jointly considers the overall ranking information and precision. As for *image-sketch* retrieval task, one retrieval scenario of sketch-to-image that using sketch as query to retrieve relevant images, is considered. For the evaluation criterion, besides the MAP score, we also include precision on top-100 (P@100) retrievals to keep the same as (Dutta and Akata 2019).

Details of Network. We implement our LCALE method using the popular PyTorch toolkit. For our network architecture, all encoders contains three fully connected layers with dimensions [4096,2048,64] and activated by ReLU active function. Similarly, all decoders contain three fully connected layers with dimensions [4096,2048,K_*] with the layer is activated by ReLU, where $* = v, t, c$, K_* represents the dimensions of the original image, text, and classes feature, respectively. In addition, we build the regressors of both image and text modalities with three fully connected layers of $[4096, 4096, 300]$ for class-embedding reconstruction with each layer following a ReLU layer.

The hyper-parameters α, β, λ and γ are set to 1, 0.1, 0.1, 0.01 respectively and the latent embedding size is set to 64. The learning rate μ is initially set at 0.0001 with weighted decay every 10 epoch. A sensitivity analysis of the hyper-parameters is provided in Fig. 6 in the latter experiment.

4.2 Comparisons on Image-Text Retrieval

Compared Methods. We compare our LCALE with six state-of-the-art methods on image-txt retrieval task. The DCCA (Yan and Mikolajczyk 2015), Deep-SM (Wei et al. 2017), and ACMR (Wang et al. 2017) are DNN-based standard retrieval approaches. Here we directly apply them on ZS-CMR. While MASLN (Xu et al. 2018), DANZCR (Chi and Peng 2018) and DADN (Chi and Peng 2019) are three latest approaches designed for ZS-CMR. We implement the compared methods DANZCR and DADN according to the instructions in the papers, and use the source codes released by the authors of the other compared methods. All the experiments are conducted ten times under the same configurations to make fair comparison.

Overall Results. Table 2 presents the overall comparison of our LCALE method and the compared ones on four datasets. We can observe that the traditional methods DCCA, DeepSM and ACMR perform worse than the other methods as they fail to consider the challenges in the ZS-CMR task. Nevertheless, our LCALE approach achieves the best retrieval accuracy among all methods and gains a large margin compared with the three ZS-CMR approaches MASLN, DANZCR and DADN. Specifically, on Wikipedia dataset, the performance of our approach increased the average MAP score from 0.298 to 0.362 compared to the best counterparts DADN; on the Pascal Sentences dataset, it consistently performs the best, and beats DADN by 15.3% and 13.5% in Img2Txt and Txt2Img tasks, respectively. Moreover, on the large-scale datasets such as PKU-XMediaNet and NUS-WIDE, our LCALE still gains a remarkable improvement compared to DADN. In general, the best performance of our LCALE method attributes to the advanced mVAE structure to correlate multimodal input features in the latent embeddings space, as well as the effectiveness of the

Table 2: The MAP scores of image-text retrieval for our LCALE approach and the compared methods on four datasets.

Methods	Wikipedia			Pascal Sentences			NUS-WIDE			PKU-XMediaNet		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
DCCA (ICML 2015)	0.282	0.266	0.274	0.297	0.264	0.281	0.406	0.407	0.407	0.039	0.043	0.041
DeepSM (TCYB 2017)	0.265	0.258	0.262	0.276	0.251	0.264	0.401	0.414	0.408	0.040	0.096	0.068
ACMR (ACM-MM 2017)	0.276	0.262	0.269	0.306	0.291	0.299	0.407	0.425	0.416	0.036	0.043	0.040
MASLN (ICMR 2018)	0.284	0.264	0.274	0.307	0.294	0.301	0.411	0.426	0.419	0.040	0.045	0.043
DANZCR (IJCAI 2018)	0.297	0.287	0.292	0.334	0.338	0.336	0.416	0.469	0.443	0.106	0.117	0.112
DADN (TCSVT 2019)	0.305	0.291	0.298	0.359	0.353	0.356	0.423	0.472	0.448	0.112	0.130	0.121
LCALE (Ours)	**0.367**	**0.357**	**0.362**	**0.414**	**0.394**	**0.404**	**0.566**	**0.567**	**0.567**	**0.135**	**0.164**	**0.150**

cross-reconstruction and cross-alignment schemes that enable the knowledge transfer to unseen classes.

4.3 Comparisons on Image-Sketch Retrieval

Compared Methods. Furthermore, on the image-sketch retrieval task, we consider three state-of-the-art ZS-CMR methods: ZSIH (Shen et al. 2018), ZS-SBIR (Yelamarthi et al. 2018), and SEM-PCYC (Dutta and Akata 2019). Besides, three traditional retrieval approaches GN Triplet (Sangkloy et al. 2016), FRWGAN (Felix et al. 2018) and GDH (Zhang et al. 2018) designed for sketch-based image retrieval (SBIR) are also directly extended for ZS-CMR. We utilize the released codes of the compared methods and run the experiments ten times under the same configurations to make a fair comparison.

Overall Results. The comparison results on image-sketch retrieval task is shown Table 3. We can see that the SBIR based approaches perform worse than the ZS-CMR method since they cannot generalize well to produce effective common embeddings for unseen classes. Nevertheless, the performance of our method is superior to all the comparison methods. In particular, it improves the MAP score from 0.349 to 0.476 compared to the best counterpart SEM-PCYC. The results show the advantage of our method on well handling the large intra-class variance in both images and sketches to boost the retrieval performance. Furthermore, we also plot the precision-recall (PR) curves of our method and several counterparts in Fig. 3. Besides, Fig. 4 shows some successfully retrieved examples on the Sketchy dataset of our LCALE approach, where the only one failure case (in red box) is caused by the incorrect annotation ("mushroom") in the groundtruth.

Table 3: Image-sketch retrieval results of our LCALE and the compared methods on Sketchy dataset. Here "(binary)" denotes the method produce *binary* semantic features for retrieval via quantization.

Types	Methods	Sketchy	
		MAP	P@100
SBIR	GN Triplet (ACM-TOG 2016)	0.204	0.296
	FRWGAN (ECCV 2018)	0.127	0.169
	GDH (binary) (ECCV 2018)	0.187	0.259
ZS-CMR	ZSIH (binary) (CVPR 2018)	0.258	0.342
	ZS-SBIR (ECCV 2018)	0.196	0.284
	SEM-PCYC (binary) (CVPR 2019)	0.344	0.399
	SEM-PCYC (CVPR 2019)	0.349	0.463
	LCALE (binary) (Ours)	**0.397**	**0.485**
	LCALE (Ours)	**0.476**	**0.583**

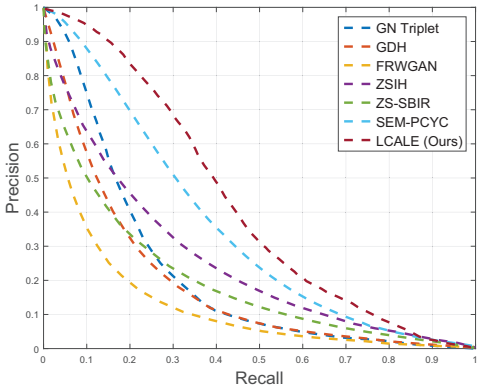


Figure 3: The PR curves of our LCALE and six compared methods on Sketchy dataset.

Figure 4: Top 10 retrieved images given a query sketch for image-sketch retrieval on Sketchy dataset.

4.4 Further Analysis on LCALE

Baseline Experiments. To investigate the impact of each loss term in Eq. 9, we design four variants as the baselines of LCALE by excluding one loss term in Eq. 9 during the training procedure. The comparison of the LCALE model and its four baselines on Wikipedia and Sketchy datasets is shown in Table 4, where "wo" indicates the exclude loss term.

From the comparison results in Table 4, we can observe that: 1) The baseline (wo $\mathcal{L}_{CR}$) performs the worst, showing that cross-reconstruction scheme can align the distribution of the shared latent representation to incorporate the cross-modal correlation. 2) The result of baseline (wo $\mathcal{L}^1_{CA}$) indicates that the distribution matching is also significant to implicitly enhance the correlation of latent embeddings from pairwise modalities. 3) The result of baseline (wo $\mathcal{L}^2_{CA}$)

shows that the MMD loss is also effective to explicitly alleviate the discrepancy of the pairwise modalities in the latent embedding space. 4) When excluding the cycle-consistent constraint, the performance of the baseline (wo $\mathcal{L}_{CYC}$) is reduced, showing that the cycle-consistent constraint is also important to enhance the semantic consistency between latent embeddings and the class-embeddings.

Visualization of the Learned Latent Embeddings. We further use the t-SNE (Maaten and Hinton 2008) tool to visualize the distribution of different features on the Wikipedia dataset. Fig. 5(a) shows the distribution of the image and text features in their original feature space. It can be observed that the image and text features have a large diversity in distribution, and hard to be classified. Fig. 5(b) displays the distribution of the image/text in the latent embedding space of compared method ACMR (Wang et al. 2017). It shows that the latent embeddings of two modalities are likely to mix together with some degree of semantic distinction, while some classes are still indistinguishable. On the contrary, in Fig. 5(c) of our LCALE, the latent embeddings of images and texts are fully mixed in the latent embedding space. Therefore, it indicates that our LCALE method can not only ensure the alignment of the distribution from different modalities, but also effectively divide the instances into several semantic clusters according to their classes.

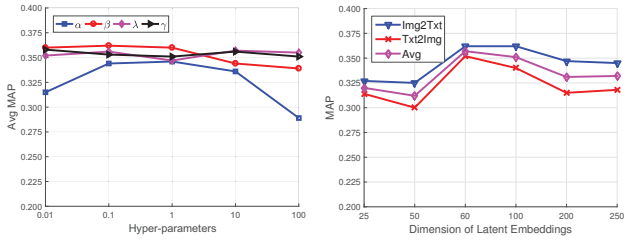
(a) Original (b) ACMR (c) LCALE (Ours)

Figure 5: The t-SNE visualization for the chosen data in Wikipedia datasets. The circle denotes the visual features and the pentagon represents the text features. Clusters with different color belong to different classes.

Table 4: Baseline experiments for image-text retrieval on Wikipedia and image-sketch retrieval on Sketchy dataset.

Baselines	Wikipedia			Sketchy	
	Img2Txt	Txt2Img	Avg.	MAP	P@100
LCALE (wo $\mathcal{L}_{CR}$)	0.261	0.240	0.250	0.422	0.441
LCALE (wo $\mathcal{L}_{CA}^1$)	0.337	0.329	0.333	0.454	0.561
LCALE (wo $\mathcal{L}_{CA}^2$)	0.353	0.341	0.347	0.395	0.536
LCALE (wo $\mathcal{L}_{CYC}$)	0.344	0.334	0.339	0.458	0.567
LCALE (All)	**0.367**	**0.357**	**0.362**	**0.476**	**0.583**

Analysis on Parameter Sensitivity. Furthermore, we study the effect of hyper-parameters α, β, λ, γ in Eq. 9 of our LCALE method on Wikipedia dataset. Specifically, the value of each hyper-parameter is set in the range of [0.01, 100]. The performance of our LCALE with different values of the hyper-parameters are shown in Fig. 6. We can see that the results vary with different values of the hyper-parameters, indicating that the loss terms in Eq. 9 have different importance. In practice, we can efficiently tune the optimal hyper-parameters through the validation process for different datasets.



(a) (b)

Figure 6: The effect of the hyper-parameters and the dimension of latent embeddings in our LCALE method on Wikipedia dataset.

Analysis on the Dimension of Latent Embeddings. Finally, we investigate the effect of different dimensions of the latent embeddings on retrieval performance. We again use the Wikipedia dataset as testbed, and the results of our method with different dimensions are shown in Fig. 6(b). It can be observed that when the dimension is low, *e.g.*, [25, 50], the performance is limited as information loss exists. The performance reaches the peak when the value is around [60, 100] and decreases with higher dimensions in [100, 250] with redundant information. Therefore, it shows that finding a low-dimensional latent embedding space is beneficial for retrieval than using the high-dimensional semantic space of class-embeddings in existing ZS-CMR approaches. The accuracy initially increases with the increase of the dimension, and decreases after reaching the peak.

5 Conclusion

We have proposed a novel method dubbed Learning Cross-Aligned Latent Embeddings (LCALE) for the ZS-CMR task. It is a promising alternative to existing GAN-based approaches for ZS-CMR. An effective multimodal VAE structure in LCALE finds a low-dimensional latent embedding space of input multimodal features and class-embeddings. Effective cross-reconstruction and cross-alignment schemes are designed to match the learned latent embeddings of different modality data and the class-embeddings, which effectively enhance the latent embeddings space and enable the knowledge transfer to unseen classes. Comprehensive experiments and ablation studies on five datasets in total for both image-text retrieval image-sketch retrieval tasks have demonstrated the superiority of the LCALE method.

6 Acknowledgments

Kaiyi Lin also belongs to School of Software and Microelectronics, Peking University, China. This work was supported in part by the National Natural Science Foundation of China under grants No. 61976049, 61872064, 61632007 and the Sichuan Science and Technology Program, China, under Grants No. 2019ZDZX0008 and 2018GZDZX0032.

References

Chi, J., and Peng, Y. 2018. Dual adversarial networks for zero-shot cross-media retrieval. In *IJCAI*, 256–262.

Chi, J., and Peng, Y. 2019. Zero-shot cross-media embedding learning with dual adversarial distribution network. *IEEE Transactions on Circuits and Systems for Video Technology* 1–1.

Chua, T.-S.; Tang, J.; Hong, R.; Li, H.; Luo, Z.; and Zheng, Y.-T. 2009. Nus-wide: A real-world web image database from national university of singapore. In *ACM CVIR*.

Deng, C.; Chen, Z.; Liu, X.; Gao, X.; and Tao, D. 2018. Triplet-based deep hashing network for cross-modal retrieval. *IEEE Trans. Image Processing* 27(8):3893–3903.

Dey, S.; Riba, P.; Dutta, A.; Lladós, J.; and Song, Y. 2019. Doodle to search: Practical zero-shot sketch-based image retrieval. In *CVPR*.

Dutta, A., and Akata, Z. 2019. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *CVPR*.

Felix, R.; Kumar, B. G. V.; Reid, I. D.; and Carneiro, G. 2018. Multi-modal cycle-consistent generalized zero-shot learning. In *ECCV*, 21–37.

Givens, C. R., and Shortt, R. M. 1984. A class of wasserstein metrics for probability distributions. *The Michigan Mathematical Journal* 31(2):231–240.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*, 2672–2680.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*.

Huang, X.; Peng, Y.; and Yuan, M. 2018. Mhtn: Modal-adversarial hybrid transfer network for cross-modal retrieval. *IEEE Trans. Cybernetics* 14(6):143–156.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. In *ICLR*.

Kingma, D. P., and Welling, M. 2014. Auto-encoding variational bayes. In *ICLR*.

Le, Q. V., and Mikolov, T. 2014. Distributed representations of sentences and documents. In *ICML*, 1188–1196.

Maaten, L., and Hinton, G. 2008. Visualizing data using t-sne. *JMLR* 9(Nov):2579–2605.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Peng, Y.; Qi, J.; Huang, X.; and Yuan, Y. 2018. Ccl: Cross-modal correlation learning with multigrained fusion by hierarchical network. *IEEE Transactions on Multimedia* 20(2):405–420.

Rashtchian, C.; Young, P. Hodosh, M.; and Hockenmaier, J. 2010. Collecting image annotations using amazon's mechanical turk. In *NAACL HLT 2010 Workshop*, 674–686.

Rasiwasia, N.; Costa Pereira, J.; Coviello, E.; Doyle, G.; Lanckriet, G.; Levy, R.; and Vasconcelos, N. 2010. A new approach to cross-modal multimedia retrieval. In *ACM MM*, 251–260.

Sangkloy, P.; Burnell, N.; Ham, C.; and Hays, J. 2016. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Trans. Graph.* 35(4):119:1–119:12.

Schönfeld, E.; Ebrahimi, S.; Sinha, S.; Darrell, T.; and Akata, Z. 2019. Generalized zero- and few-shot learning via aligned variational autoencoders. In *CVPR*, 8247–8255.

Shen, Y.; Liu, L.; Shen, F.; and Shao, L. 2018. Zero-shot sketch-image hashing. In *CVPR*, 3598–3607.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR* abs/1409.1556.

Song, J.; Yu, Q.; Song, Y.; Xiang, T.; and Hospedales, T. M. 2017. Deep spatial-semantic attention for fine-grained sketch-based image retrieval. In *ICCV*, 5552–5561.

Wang, K.; He, R.; Wang, W.; Wang, L.; and Tan, T. 2013. Learning coupled feature spaces for cross-modal matching. In *ICCV*, 2088–2095.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015. On deep multi-view representation learning. In *ICML*, 1083–1092.

Wang, K.; He, R.; Wang, L.; Wang, W.; and Tan, T. 2016. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38(10):2010–2023.

Wang, B.; Yang, Y.; Xu, X.; Hanjalic, A.; and Shen, H. T. 2017. Adversarial cross-modal retrieval. In *ACM MM*, 154–162.

Wang, T.; Xu, X.; Yang, Y.; Hanjalic, A.; Shen, H. T.; and Song, J. 2019. Matching images and text with multi-modal tensor fusion and re-ranking. In *ACM MM*, 12–20.

Wei, Y.; Zhao, Y.; Lu, C.; Wei, S.; Liu, L.; Zhu, Z.; and Yan, S. 2017. Cross-modal retrieval with CNN visual features: A new baseline. *IEEE Trans. Cybernetics* 47(2):449–460.

Xu, X.; Song, J.; Lu, H.; Yang, Y.; Shen, F.; and Huang, Z. 2018. Modal-adversarial semantic learning network for extendable cross-modal retrieval. In *ACM ICMR*, 46–54.

Yan, F., and Mikolajczyk, K. 2015. Deep correlation for matching images and text. In *CVPR*, 3441–3450.

Yelamarthi, S. K.; Reddy, M. S. K.; Mishra, A.; and Mittal, A. 2018. A zero-shot framework for sketch based image retrieval. In *ECCV*, 316–333.

Zhang, J.; Shen, F.; Liu, L.; Zhu, F.; Yu, M.; Shao, L.; Shen, H. T.; and Gool, L. V. 2018. Generative domain-migration hashing for sketch-to-image retrieval. In *ECCV*, 304–321.

Zhu, B.; Ngo, C.-W.; Chen, J.; and Hao, Y. 2019. R2gan: Cross-modal recipe retrieval with generative adversarial network. In *CVPR*.