

# Zero-shot Cross-modal Retrieval by Assembling AutoEncoder and Generative Adversarial Network

XING XU, JIALIN TIAN, and KAIYI LIN, University of Electronic Science and Technology of China, China

HUIMIN LU, Kyushu Institute of Technology, Japan

JIE SHAO and HENG TAO SHEN, University of Electronic Science and Technology of China, China and Sichuan Artificial Intelligence Research Institute, Yibin, China

Conventional cross-modal retrieval models mainly assume the same scope of the classes for both the training set and the testing set. This assumption limits their extensibility on **zero-shot cross-modal retrieval** (ZS-CMR), where the testing set consists of unseen classes that are disjoint with seen classes in the training set. The ZS-CMR task is more challenging due to the heterogeneous distributions of different modalities and the semantic inconsistency between seen and unseen classes. A few of recently proposed approaches are inspired by zero-shot learning to estimate the distribution underlying multimodal data by generative models and make the knowledge transfer from seen classes to unseen classes by leveraging class embeddings. However, directly borrowing the idea from zero-shot learning (ZSL) is not fully adaptive to the retrieval task, since the core of the retrieval task is learning the common space. To address the above issues, we propose a novel approach named **Assembling AutoEncoder and Generative Adversarial Network** (AAEGAN), which combines the strength of AutoEncoder (AE) and Generative Adversarial Network (GAN), to jointly incorporate common latent space learning, knowledge transfer, and feature synthesis for ZS-CMR. Besides, instead of utilizing class embeddings as common space, the AAEGAN approach maps all multimodal data into a learned latent space with the distribution alignment via three coupled AEs. We empirically show the remarkable improvement for ZS-CMR task and establish the state-of-the-art or competitive performance on four image-text retrieval datasets.

CCS Concepts: • **Information systems** → **Information retrieval**; • **Computing methodologies** → **Visual content-based indexing and retrieval**;

Additional Key Words and Phrases: Cross-modal retrieval, zero-shot learning, feature synthesis

This work is partially supported by the National Natural Science Foundation of China (No. 61976049 and 61632007); the Fundamental Research Funds for the Central Universities (No. ZYGX2019Z015); and the Sichuan Science and Technology Program, China (No. 2019ZDZX0008, 2019YFG0003, 2019YFG0533, 2019YFG0535, 2020YFS0057, and 2020YJ0038).

Authors' addresses: X. Xu, J. Tian, and K. Lin, University of Electronic Science and Technology of China, No. 2006, Xiyuan Ave, Chengdu, 611731, China; emails: xing.xu@uestc.edu.cn, {tian.garin, lky.linkaiyi}@gmail.com; H. Lu (corresponding author), Kyushu Institute of Technology, 1-1 Sensui, Tobata, Kitakyushu, 804-8550, Japan; email: dr.huimin.lu@ieee.org; J. Shao and H. T. Shen, University of Electronic Science and Technology of China, No. 2006, Xiyuan Ave, Chengdu, 611731, China and Sichuan Artificial Intelligence Research Institute, No. 430, 2nd Subsubsection, West Section, Changjiang North Road, Yibin, 644004, China; emails: shaojie@uestc.edu.cn, shenhengtao@hotmail.com.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

© 2021 Association for Computing Machinery.

1551-6857/2021/03-ART3 \$15.00

<https://doi.org/10.1145/3424341>

**ACM Reference format:**

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 Trans. Multimedia Comput. Commun. Appl.* 17, 1s, Article 3 (March 2021), 17 pages. <https://doi.org/10.1145/3424341>

**1 INTRODUCTION**

As Internet technology has made great progress in recent years, people are used to sending on-line messages to describe the same events through text descriptions and corresponding pictures or videos. As a result, we are witnessing an explosive growth of multimodal data (e.g., images, texts, videos, and audio) in our daily life [30, 36, 47], and searching valuable information effectively among different multimedia data has arisen urgent need in the research field of information retrieval [48, 50]. Cross-modal retrieval (CMR) [33] becomes a highlighted research area due to the cross-modal correlation existing among different modalities, which is manifested as the co-occurrence of different modality data [16]. The fundamental function of CMR is to use a query of any modality data to retrieve data of other modalities, e.g., text-image retrieval, image-sketch retrieval, and recipe retrieval.

However, the research on CMR faces a key challenge that the “modality gap” from heterogeneous distributions across data of different media types makes the similarity between different modalities difficult to measure. To bridge the modality gap, a majority of existing methods [29, 41, 44, 45, 54, 58] consider common latent representation with a suitable distance metric as a solution. However, the training of these methods requires that the training set and testing set share the same scope of classes. It limits the ability of these models to be extended to more realistic scenarios, which means retrieving data of classes that are unseen in the training set. Therefore, cross-modal retrieval extending to the zero-shot learning setting has recently received some attention.

Zero-shot learning (ZSL) [19, 20] is a paradigm that learns to classify objects when labeled instances of previous unseen classes are not available during training. Even though it is a more realistic scenario in the real world and reduces the cost of data collection and human annotation, existing ZSL approaches only take the single-media data into consideration and ignore the multimedia data with heterogeneous distribution across different modalities. Therefore, this article deals with the problem of cross-modal retrieval in zero-shot settings that the query instances and database instances belong to unseen classes. When data of unseen classes become available, the overall model for zero-shot cross-modal retrieval (ZS-CMR) does not need to be retrained from scratch to achieve promising performance, which is computationally efficient.

The research on ZS-CMR [7, 54, 55, 57] is a more challenging task due to insufficient training data and semantic inconsistency across seen and unseen classes, because all data are split into seen and unseen classes and the data of unseen classes are absent during training, which leads to the risk of overfitting and inferior performance in the zero-shot setting. A schematic diagram of ZS-CMR is shown in Figure 1. However, several recent studies [7, 8, 25, 56, 57] have been proposed to meet these challenges for the image-text retrieval task and the image-sketch retrieval task in the zero-shot setting. Inspired by zero-shot learning, all these methods exploit the class-level word-embeddings (*class embeddings*), which can be easily extracted from pre-trained natural language processing (NLP) models, as side information to guide the learning of common latent space. Interestingly, all of them are utilizing generative models as the basic structure to learn the common latent space, where generative adversarial network (GAN) [7, 8, 56], variational autoencoder (VAE) [25], and autoencoder (AE) [57] are respectively utilized. Although they have achieved promising results, they still have the following defects. First, these models are too complex for that

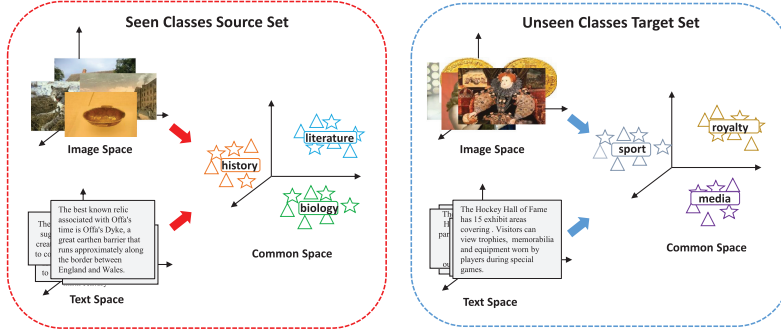


Fig. 1. An illustration of zero-shot cross-modal retrieval (ZS-CMR).

many components are not directly related to the core task of learning the latent space. Second, the class embedding space is not the most optimal choice for latent space in ZS-CMR, although it has shown its effectiveness in the zero-shot learning area. Third, a single kind of generative model cannot fully capture the correlation among modalities. The above three defects are probably to make these models have poor generalization ability in zero-shot setting and relatively inferior performance.

In this article, to tackle the above issues of existing ZS-CMR models, we devise a novel and effective framework, namely Assembling AutoEncoder and Generative Adversarial Network (AAEGAN). Figure 2 shows the overall framework of our proposed AAEGAN. Our proposed AAEGAN is an end-to-end learning framework consisting of three coupled AEs and two coupled GANs that decoders and generators share the same parameters. The coupled encoders and decoders aim to learn common representations for each modality and reconstruct them back to the original modality by optimizing the reconstruction loss. Additionally, the coupled generators and discriminators learn the distribution of original modality data by optimizing Wasserstein-GAN (WGAN) loss. During the testing stage, AAEGAN maps features of image and text to common space using encoders, and the cross-modal retrieval is conducted in this space. Thus, it jointly incorporates encoding for each modality data, the synthesis of multimodal features and knowledge transfer. To verify the true ability of AAEGAN structure, the training flow and objective function of the model are simplified. Notably, we choose AE instead of VAE as the basic structure for its suitability for the ZS-CMR task. More specific reasons will be explained later.

We highlight our contributions in this article as follows:

- We propose a novel AAEGAN model that assembles AEs and GANs to combine their advantages and promote the performance with each other and is able to learn the common latent space, synthesize multimodal features, and transfer knowledge simultaneously.
- To enhance the learning of common latent space, we develop an effective constraint of distribution alignment to preserve the semantic compatibility between modalities. This constraint is beneficial to learn more robust common space and capture the cross-modal correlation of different modalities that is compatible with ZS-CMR scenarios.
- We conduct extensive experiments on four widely used cross-modal retrieval datasets, i.e., Wikipedia [33], Pascal Sentences [32], NUS-WIDE [9], and PKU XMediaNet [16] and the results clearly demonstrate the superior retrieval performance of the AAEGAN approach compared with a bundle of state-of-the-art approaches.

The remainder of this article is organized as follows. We briefly review the related work on CMR problem, ZSL problem, and related generative models in Section 2. In Section 3, our

proposed AAEGAN method is depicted in detail. Then we present the experimental results and the comprehensive analysis in Section 4. Finally, Section 5 gives a conclusion of our work.

## 2 RELATED WORK

### 2.1 Cross-Modal Retrieval

The primary problem of CMR is to bridge the “modality gap” by establishing the cross-modal correlation of heterogeneous representations of different modalities. During the past decade, learning common latent space for retrieval has become the mainstream idea [33, 44, 46, 51], which aims at learning various linear or nonlinear transformations to project the inconsistent representations of different modalities to the shared space. As a result, the content similarity of the multimodal data can be computationally measured.

The CMR models can be categorized into *shallow learning* models and *deep neural network (DNN)-based* models for the different transformation methods. For the shallow learning models [23, 33, 45, 46, 59], the transformations are usually assumed to be linear. For example, in the pioneering work of [33], the canonical correlation analysis (CCA) is utilized to learn a linear transformation with the maximization of the cross-modal pairwise correlation such that it can project original modality data into a common space. Several subsequent works incorporate additional information to CCA and extend it to be several variants, e.g., multi-view CCA [12] and multi-label CCA [31]. Except for CCA, there are also other alternative methods that develop advanced schemes such as factor analysis [23], dictionary learning [62], graph regularization [59], and feature selection [45, 46] to enhance to learn more effective linear transformations.

Due to the great advance achieved by DNN on fundamental vision problems like image classification and object detection, recent studies on cross-modal retrieval also exploit DNN to learn common representation, expecting to capture the nonlinear cross-modal correlation. The DNN-based approaches usually follow the pipeline of the shallow learning methods, i.e., modeling specific subnetworks for different modalities and connecting them via a joint layer as the common space to model the cross-modal correlation [47]. The earlier works [40, 41] adopt Bimodal Autoencoders, Correspondence Auto-Encoder [11], and Restricted Boltzmann Machine. Notably, CCA is also extended to its DNN-based version, i.e., Deep Canonical Correlation Analysis (DCCA) [58], where DNN is used as the nonlinear feature projections. Besides, several more advanced schemes such as hierarchical network stacking [28] and multi-task learning [30] are developed to jointly explore the inter-media differences and the intra-media correlations.

However, conventional CMR methods mostly make the assumption that the training set and the testing set have the same scope of classes, which are unable to retrieve data of unseen classes. In the pipeline of current cross-modal retrieval models, every time a new class appears, the entire model has to be retrained for promising retrieval performance, which makes no efficiency in computation. Since the collection and annotation of cross-media data takes a lot of labor and time, it is becoming more and more necessary to explore cross-modal retrieval in zero-shot setting for the numerous and dynamically emerging new class in the real world. Xu et al. [57] first assessed several latest standard cross-modal retrieval approaches on the ZS-CMR task and found that their performance drastically decreases. Several latter works [7, 8, 25, 56] follow the mainstream pipeline in the unimodal ZSL problem and leverage the class-level word embeddings as side information for knowledge transfer, which can be easily extracted by pre-trained NLP models on a large corpus (e.g., Wikipedia web pages or Google News). The studies [7, 8, 56] directly adopt the class embedding space as the common space, while Lin et al. [25] learns a lower-dimensional space to encode all data of modalities.

Similarly to Reference [25], our proposed AAEGAN uses the learned latent space instead of class embedding space as the common space. However, we adopt AE instead of VAE as the basic structure for encoders and decoders, because AE is more appropriate for retrieval. The Kullback-Leibler divergence term existing in VAE measures the differences of distributions of prior centered isotropic Gaussian noise and the Gaussian distribution constructed by learned means and variances via encoders while minimizing this term conflicts with the ultimate goal of the ZS-CMR task, i.e., learning more class-specific latent features for retrieval.

## 2.2 Generative Models

Autoencoder, which is first proposed in Reference [15], is an unsupervised method that encodes input images into low-dimensional latent embedding space through the encoder and reconstructs them back to original images via the decoder. After that, many variants of AE were proposed, including denoising AE [43], sparse AE [10], and VAE [18]. The difference between VAE and the original AE is that the outputs of the encoder are the means and standard deviations for each instance by optimizing the lower bound of the log-likelihood. GAN is also one of the most promising generative models that is originally proposed for image synthesis, which consists of a generator and a discriminator that work against each other. However, as investigated in References [3, 4, 13], GAN also suffers the instability issue in the training procedure and the mode collapse problem. To mitigate these problems and improve the quality of synthetic samples, Arjovsky et al. [4] proposes Wasserstein-GAN (WGAN), which optimizes GAN on an approximated Wasserstein distance by enforcing 1-Lipschitz constraint. Although WGAN is more stable than the original GAN, it still suffers from gradient vanishing and gradient explosion problems. Thus, Gulrajani et al. [13] proposes WGAN with a gradient penalty (GP) as the improved version of WGAN. Considering the stability of training and the adaptability of ZS-CMR scenarios, we adopted WGAN with GP and AE as our basic model structure.

## 2.3 Zero-shot Learning

Zero-shot learning aims to learn more generalized models to identify instances of unseen classes, which has positive implications for the cost mitigation of data collection and human annotation. Pioneering works for ZSL [19, 20] leverage class-level attributes as side information to transfer knowledge from seen classes to unseen classes, which makes it possible to recognize instances of unseen classes. Subsequent works have evolved into the embedding-based framework, and the core of such methods is learning to project from the visual space to the semantic space [34, 39] or vice versa [37, 60], or jointly learn a compatibility function between visual and semantic features through an embedding space [1, 2, 42]. More recently, several synthesizing methods [6, 35, 52, 53, 61] make use of generative models to synthesize features for unseen classes for their advantage in data generation, which convert the ZSL problem to traditional supervised classification problem. Compared with the projection model, the synthesizing models are superior in performance. Similarly, some generative models for ZS-CMR [7, 8, 25, 56] are also proposed. The key difference between ZS-CMR methods and other ZSL methods is that ZS-CMR methods exploit the correlation of heterogeneous distribution of multi-modal data rather than unimodal data. Hence, in this article, we propose AAEGAN to perform correlation learning in the zero-shot setting, which aims to overcome the “modality gap” and semantic inconsistency in ZS-CMR.

# 3 PROPOSED METHOD

## 3.1 Problem Formulation

In the ZS-CMR problem, we consider bimodal data (i.e., images, texts) as example to formulate our proposed AAEGAN. Assuming a source set consists of  $N_s$  pairwise image-text instances, i.e.,

$O_s = \{o_i\}_{i=1}^{N_s}$ ,  $o_i = (\mathbf{v}_i, \mathbf{t}_i, \mathbf{c}_i, y_i)$ , where corresponding image feature, text feature, class embedding, and the class label for the  $i$ th instance are denoted as  $o_i$ ,  $\mathbf{v}_i$ ,  $\mathbf{t}_i$ ,  $\mathbf{c}_i$ , and  $y_i$ . Again, the target set is represented as  $O_t = \{o_j\}_{j=1}^{N_t}$ ,  $o_j = (\mathbf{v}_j, \mathbf{t}_j, \mathbf{c}_j, y_j)$ , and these notations are defined in accordance with the source set.

For the settings of ZS-CMR task, we first split the source set and the target set into seen and unseen subsets, which are respectively represented as  $O_s = \{O_s^s, O_s^u\}$  and  $O_t = \{O_t^s, O_t^u\}$ , where  $O_s^s$ ,  $O_s^u$ ,  $O_t^s$ , and  $O_t^u$  denote seen class source set, unseen class source set, seen class target set, and unseen class target set, respectively.

Notably, the seen class set and the unseen class set are denoted by  $\mathcal{Y}_s$  and  $\mathcal{Y}_u$  and the overall class set are denoted by  $\mathcal{Y} = \{1, \dots, C\}$  for  $\mathcal{Y}_s \cup \mathcal{Y}_u = \mathcal{Y}$ ,  $\mathcal{Y}_s \cap \mathcal{Y}_u = \emptyset$ . The ultimate goal of our AAEGAN is to learn a common latent space from the data of seen classes and make knowledge transfer to unseen classes. Thus, only the seen class source set (i.e.,  $O_s^s$ ) is used in the training phase. In the test phase, for the ZS-CMR task, we use  $O_t^u$  as queries to retrieve relevant cross-model instances in  $O_s^u$  denoted by  $O_t^u \rightarrow O_s^u$ .

### 3.2 Our AAEGAN Approach

**3.2.1 Network Architecture.** As the overall framework of our AAEGAN method shown in Figure 2, three coupled AEs form three parallels respectively for image and text modalities to generate common representations and synthetic features, and two discriminators follow the two decoders/generators to determinate the reality of synthetic features, which form two coupled GANs. The common representations of class embeddings of seen classes are utilized as a guide to learn the correlations between paired images and texts, which is realized via domain alignment, i.e., MMD losses and Minimum Squared Error (MSE) losses. Then in testing time, the two encoders for image and text modalities are used to project the original unseen features to common representations. Thus ZS-CMR is successively performed in the common latent space. We will elaborately depict the three key components in our AAEGAN model as follows.

**3.2.2 Basic AE and GAN Models.** In the coupled autoencoders, the encoders  $E^v$ ,  $E^t$ , and  $E^c$  generates common latent representations  $e^v$ ,  $e^t$ , and  $e^c$  in common space from original modality data  $v$ ,  $c$ , and  $t$ , respectively, and the decoders  $G^v$ ,  $G^t$ , and  $G^c$  reconstruct the common representations back to original modality space  $\tilde{v}$ ,  $\tilde{t}$ , and  $\tilde{c}$ . It is formulated as follows:

$$\mathcal{L}_{Rec}(\theta_E, \theta_G) = \frac{1}{N_s^s} \sum_{i=1}^{N_s^s} \|\mathbf{v}_i - \tilde{\mathbf{v}}_i\|_2^2 + \frac{1}{N_s^s} \sum_{i=1}^{N_s^s} \|\mathbf{t}_i - \tilde{\mathbf{t}}_i\|_2^2 + \frac{1}{N_s^s} \sum_{i=1}^{N_s^s} \|\mathbf{c}_i - \tilde{\mathbf{c}}_i\|_2^2, \quad (1)$$

where  $\theta_E, \theta_G$  denote the network parameters for the encoders and the decoders, and  $N_s^s$  denotes the instance number of  $O_s^s$ . The dimension of common latent space not reflected in this formula is marked as  $K_s$ .

In the coupled GANs, the generators  $G^v$  and  $G^t$ , which share the same networks with decoders in coupled AEs, synthesize features  $\tilde{v}$  and  $\tilde{t}$  from the encoded latent representations  $e^v$  and  $e^t$ , and discriminators  $D^v$  and  $D^t$  take synthetic features and original modality features concatenated with class embedding as input and output a real value. Specifically, the structure of Wasserstein GAN with a gradient penalty, which is one of the most stable GAN structures for the robust training process, is chosen for these two GANs. For the image pathway, the objective function is optimizing:

$$\begin{aligned} \mathcal{L}_{WGAN}^v(\theta_{G^v}, \theta_{D^v}) = & \mathbb{E}_{v \sim P_v} [D(v, c)] - \mathbb{E}_{\tilde{v} \sim P_{G^v}} [D(\tilde{v}, c)] \\ & - \lambda \mathbb{E} \left[ (\|\nabla_{\hat{v}} D(\hat{v}, c)\|_2 - 1)^2 \right], \end{aligned} \quad (2)$$



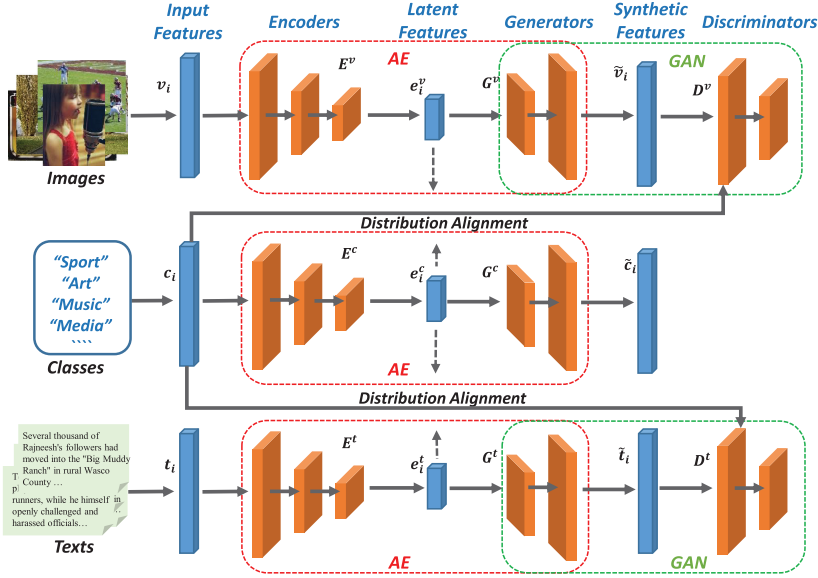


Fig. 2. The overall framework of our AAEGAN, which consists of three encoders and decoders/generators for images, classes, and texts, as well as two discriminators for images and texts. For each modality, encoders generate common representations from original features, decoders/generators reconstruct them back to original features, and discriminators determine whether the synthetic features are real or not.

where  $P_v$  and  $P_{G^v}$  are distribution of original image features and synthetic image feature;  $\theta_{G_v}$  and  $\theta_{D_v}$  are the parameters of the generator and the discriminator for image;  $\lambda$  is the coefficient of gradient penalty, which is empirically set as 10; and  $\hat{v} = \alpha v + (1 - \alpha)\tilde{v}$  is the linear interpolations between real and synthetic feature with  $\alpha \sim U(0, 1)$ . The first two terms approximate Wasserstein distance of the distribution of original and synthetic image features, and the third term in Equation (2) is the gradient penalty that enforces the gradient of discriminator  $D_v$  to have a unit norm for the linear interpolation of paired  $v$  and  $\tilde{v}$  [52].

Similarly, the objective function for  $G^v$  and  $G_t$  is optimizing:

$$\begin{aligned} \mathcal{L}_{WGAN}^t(\theta_{G_t}, \theta_{D_t}) = & \mathbb{E}_{t \sim P_t} [D(t, c)] - \mathbb{E}_{\tilde{t} \sim P_{G_t}} [D(\tilde{t}, c)] \\ & - \lambda \mathbb{E} \left[ \left( \|\nabla_{\hat{t}} D(\hat{t}, c)\|_2 - 1 \right)^2 \right], \end{aligned} \quad (3)$$

where  $\theta_{G_t}$  and  $\theta_{D_t}$  are the parameters of the generator and the discriminator for text.

Finally, the baseline GAN model combines above two losses for two modalities together as

$$\mathcal{L}_{WGAN}(\theta_G, \theta_D) = \mathcal{L}_{WGAN}^v(\theta_{G_v}, \theta_{D_v}) + \mathcal{L}_{WGAN}^t(\theta_{G_t}, \theta_{D_t}). \quad (4)$$

**3.2.3 Our AAEGAN Model.** Integrating a VAE and a GAN has shown a promising result on generating image and image feature in References [21, 53]. Because VAE and GAN can capture different patterns of data, the VAEGAN argues that these two generative models are complementary and can promote themselves for each other by end-to-end learning, especially when the training data are from a multimodal distribution that is difficult to model. In this article, our ultimate goal is to learn a common latent space as a retrieval space. But the Kullback–Leibler divergence term in the objective function of VAE tries to minimize the differences of distributions of prior centered isotropic Gaussian noise and the Gaussian distribution constructed by learned

means and variances via encoders, which conflicts with the ultimate goal of the ZS-CMR task. So we integrate GAN with AE instead of VAE, which is more compatible with the ZS-CMR task.

As described earlier, we introduce encoders  $E^v : \mathcal{V} \rightarrow \mathcal{S}$  and  $E^t : \mathcal{T} \rightarrow \mathcal{S}$  to encode paired images and texts to common representations, and discriminators, and discriminators  $D^v : \mathcal{S} \rightarrow \mathbb{R}$  and  $D^t : \mathcal{S} \rightarrow \mathbb{R}$  to determine the reality of the synthetic feature, optimizing

$$\mathcal{L}_{AAEGAN}(\theta_E, \theta_G, \theta_D) = \mathcal{L}_{Rec}(\theta_E, \theta_G) + \beta_1 \mathcal{L}_{WGAN}(\theta_G, \theta_D), \quad (5)$$

where the generators of GANs and decoders of AEs share the same networks, which denote as  $G^v$  and  $G^t$ , and  $\beta_1$  is the hyperparameter used to control the weighting coefficients of the two losses.

Furthermore, it is important to keep paired common representations close in latent space, because they are projections of pairs of images, text, and class embedding. Therefore, we compare statistics the common representations of different modalities and expect to transfer knowledge from these modalities to modal class-level semantic information across seen classes. Specifically, we adopt the maximum mean discrepancy (MMD) to measure the difference of the pairwise common representations of image, text, and class embedding modalities, as it is utilized in the CMR task recently. Using kernel trick, the MMD criterion can be formulated by a simple function as

$$\mathcal{L}_{MMD} = \left\| \mathbb{E}_{s^v \sim P_{E^v}} [\kappa(s^v)] - \mathbb{E}_{s^t \sim P_{E^t}} [\kappa(s^t)] \right\|_{\mathcal{H}_k}^2, \quad (6)$$

where  $\mathcal{H}_k$  is the reproducing kernel Hilbert space constructed by  $\kappa(x)$ , which represents the linear combination of multiple radial basis function kernels as follows:

$$\kappa(x_i, x_j) = \sum_n \eta_n \exp \left\{ -\frac{1}{2\sigma_n} \|x_i - x_j\|^2 \right\}. \quad (7)$$

In addition, we also adopt MSE to measure instance-level differences between the common representations of three modalities, optimizing:

$$\mathcal{L}_{MSE} = \frac{1}{N_s^s} \sum_{i=1}^{N_s^s} \|s_i^v - s_i^t\|_2^2 + \frac{1}{N_s^s} \sum_{j=1}^{N_s^s} \|s_j^v - s_j^c\|_2^2 + \frac{1}{N_s^s} \sum_{k=1}^{N_s^s} \|s_k^t - s_k^c\|_2^2. \quad (8)$$

$$\mathcal{L}_{DA} = \mathcal{L}_{MMD} + \mathcal{L}_{MSE}. \quad (9)$$

Unlike MMD loss, the MSE loss also measures the difference between class embedding and other two modalities, which guide the learning process of common representation so that clusters of common representations of images and texts are embedded around that of class embedding.

### 3.3 Overall Objective and Optimization

As defined above, the full objective of our AAEGAN is as follows:

$$\mathcal{L} = \mathcal{L}_{Rec} + \beta_1 \mathcal{L}_{WGAN} + \beta_2 \mathcal{L}_{DA}, \quad (10)$$

where  $\beta_1$  and  $\beta_2$  are the hyperparameters that balance the contribution of baseline AAEGAN and distribution alignment.

According to the above objective involved GAN losses, the optimization algorithm for Equation (10) plays a mini-max game under adversarial training style, which optimizes the parameters of encoders, generators, and discriminator alternatively. When optimizing all parameters of AAEGAN, we follow the training procedure proposed in WGAN [13]. The training procedure of our AAEGAN model is summarized as Algorithm 1.

#### 1) Optimizing the generative models:

$$\theta_G = \arg \min_{\theta_G} \mathcal{L}_{Rec} + \beta_1 \mathcal{L}_{WGAN}. \quad (11)$$














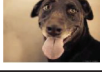
Dataset	Image			Text		
Wikipedia				In 1775, Fort Ticonderoga, in disrepair, was still manned by a token force. On May ...	Sarah Hare died in 1692 and was buried in Westminster Abbey, and Hare in 1708, to be ...	Chalukyan temples fall into two categories — the first being temples with a ...
Pascal Sentence				Several people riding dirt bikes with number plates on the bikes.	A yellow motorcycle is parked on the street	A motocross racer wearing blue and red protective gear
NUS-WIDE				blue light sky Sun golden	sky clouds glow red Aerial	blue cloud white Color clouds
PKU-XMediaNet				An attack dog is any dog trained by a human to defend or attack a territory, proper ...	Early in recorded history there are records of dogs being trained for the purpose of ...	Napoleon utilized dogs for their superior senses, putting them to work in roles ...

Fig. 3. Illustrations of image and text examples in the datasets used in our experiments.

**ALGORITHM 1:** Training of the proposed AAEGAN.

**Require:** Seen class source set  $O_s^s = \{(v_i, t_i, c_i, y_i)\}_{i=1}^{N_s}$ , batch size  $B$ , weighting coefficients  $\beta_1, \beta_2$ , learning rate  $\mu$ .

**Ensure:** Modal parameters  $\theta_E, \theta_G, \theta_D$ .

- 1: Initialize hyper-parameters  $\beta_1$  and  $\beta_2$ , learning rate  $\mu$  and batch size  $B$ .
- 2: **repeat**
- 3:   Sample paired instances  $\{v_i, t_i, c_i, y_i\}_{b=1}^B$  with batch size  $B$ .
- 4:   Update  $\theta_D$  by  $\theta_D \leftarrow \theta_D - \mu \nabla_{\theta_D} (\beta_1 \mathcal{L}_{WGAN}(\theta_G, \theta_D))$ .
- 5:   Update  $\theta_G$  by  $\theta_G \leftarrow \theta_G - \mu \nabla_{\theta_G} (\mathcal{L}_{Rec} + \beta_1 \mathcal{L}_{WGAN}(\theta_G, \theta_D))$ .
- 6:   Update  $\theta_E$  by  $\theta_E \leftarrow \theta_E - \mu \nabla_{\theta_E} (\mathcal{L}_{Rec} + \beta_1 \mathcal{L}_{WGAN}(\theta_G, \theta_D) + \beta_2 \mathcal{L}_{DA})$ .
- 7: **until** AAEGAN model in Equation (10) converges.
- 8: The encoders  $E^v$  and  $E^t$  generate latent embeddings for original data of images and texts for testing, respectively.

## 2) Optimizing the discriminative models:

$$\theta_D = \arg \min_{\theta_D} \mathcal{L}_{WGAN}. \quad (12)$$

## 3) Optimizing the encoders:

$$\theta_E = \arg \min_{\theta_E} \mathcal{L}_{Rec} + \beta_1 \mathcal{L}_{WGAN} + \beta_2 \mathcal{L}_{DA}. \quad (13)$$

**4 EXPERIMENT****4.1 Experimental Setup**

**4.1.1 Datasets and Features.** In this section, we prove the superiority of our method by comparing it with a few state-of-the-art approaches. The experimental results are obtained from four widely-used benchmark datasets for cross-modal retrieval, i.e., Wikipedia [33], Pascal Sentence [32], NUS-WIDE [9], and PKU-XMediaNet [16]. Figure 3 shows typical image-text pairs in the datasets, indicating the various formats of the association between images and texts.

We keep the same training and testing split and feature extraction methods as in References [8, 25]. Due to the strong representative ability and wide application scope of convolutional neural network (CNN) features, we adopt the CNN feature as the representation of images for all datasets in all experiments. In particular, the 4,096-dimensional CNN feature vectors extracted

Table 1. The General Statistics of Four Datasets and Their Default Train/Test Split

Datasets	Pairs	Labels	F (I)	F (T)	F (L)	Train	Test
Wikipedia [33]	2,866	10	VGG	DV	WV	2,173	693
Pascal Sentences [32]	1,000	20	VGG	DV	WV	800	200
NUS-WIDE [9]	71,602	10	VGG	DV	WV	42,941	28,661
PKU-XMediaNet [16]	40,000	200	VGG	DV	WV	32,000	8,000

Here “F (I),” “F (T),” and “F (L)” denote features for images, texts and labels, respectively.

from the fc7 layer in VGGNet-19 [38] are used for all compared methods for a fair comparison. Moreover, the feature vectors for text modality are extracted by the Doc2Vec (DV) [22] with 300 dimensions, and the 300-dimensions class embedding for each class is extracted by Word2Vec (WV) [27] model pre-trained on Google News. The statistical information of all datasets is shown in Table 1.

**4.1.2 Retrieval Tasks.** As briefly described in Section 3.1, following the protocol proposed in References [7, 26], we conduct a new dataset split settings that the source set and target set are further split as two subsets respectively, which are denoted as seen class source set, unseen class source set for source set, and seen class target set and unseen class target set for target set. Note that each set consists of a half of classes and the seen classes are disjoint with the unseen classes. In the ZS-CMR, we use seen class source set for training and use one modality data in the unseen class target set as queries to retrieve the instances of the other modality in the unseen class source set during testing. Therefore, two retrieval scenarios are included: image-to-text and text-to-image, which use image (or text) modality data to retrieve text (or image) modality data.

**Compared Methods and Evaluation Metric.** The proposed AAEGAN is compared with 14 state-of-the-art methods developed for ZS-CMR. The methods CCA [33], CFA [23], KCCA [14], JRL [59], and LGCFL [17] are shallow methods designed for conventional CMR task, while DCCA [58], Deep-SM [49], CMDH [28], and ACMR [44] are DNN-based methods. Here they are directly applied to the ZS-CMR task. In addition, the DEMZSL [60], MASLN [57], DANZCR [7], DADN [8], TANSS [56], and LCALE [25] are the recently proposed CMR models under the zero-shot setting. All models were fairly compared with others under the same settings. We first learn the projections or deep models for all methods, which are further utilized to convert the original features of the testing multimodal data to the latent features in the common space. Last, the ZS-CMR is performed by measuring the cosine distance of the pairwise testing data based on their latent features.

Moreover, the ranking results obtained in the previous step are used to calculate the mean average precision (MAP) score, which is commonly used for evaluation in CMR problem for its comprehensive consideration of precision and ranking information. The performance of both directions of image-text retrieval is evaluated, and their mean value is the final evaluation result.

**4.1.3 Details of Network.** Our proposed AAEGAN approach is implemented by the Tensorflow framework with a single GEFORCE RTX 2080TI GPU. For the network structure, each encoder is a three-layer fully connected network with hidden dimensions  $[4096, 2048, K_S]$ , where  $K_S$  is the dataset-specific dimension of the common latent feature. Every fully connected layer is activated by LeakyRelu, except for the last layer that is activated by Relu. The generators are two-layer fully connected networks with hidden dimensions  $[4096, K_M]$ , where  $K_M$  is 300 for text and class embedding modalities and 4096 for image modality. Similarly, the first layer is activated by LeakyRelu and the last layer is activated by Relu. As for discriminators, they are also two-layer fully connected networks with hidden dimensions  $[4096, 1]$ , and LeakyRelu is also the activation for the first layer while no activation exists in the last layer. The Adam optimizer is adopted for each component

Table 2. The MAP Scores of Zero-shot Retrieval for Our AAEGAN Approach and Other Compared Methods on Wikipedia and Pascal Sentences Datasets

Methods	Wikipedia			Pascal Sentences		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA [33] (2010)	0.238	0.236	0.237	0.207	0.183	0.195
CFA [24] (2003)	0.275	0.285	0.280	0.270	0.294	0.282
KCCA [5] (2014)	0.279	0.288	0.284	0.310	0.321	0.316
JRL [59] (2014)	0.264	0.266	0.265	0.298	0.283	0.291
LGCFL [17] (2015)	0.261	0.258	0.260	0.273	0.258	0.266
DCCA [58] (2015)	0.282	0.266	0.274	0.297	0.264	0.281
DeepSM [49] (2017)	0.265	0.258	0.262	0.276	0.251	0.264
ACMR [44] (2017)	0.276	0.262	0.269	0.306	0.291	0.299
DEMZSL [60] (2017)	0.310	0.239	0.275	0.308	0.318	0.313
MASLN [57] (2018)	0.284	0.264	0.274	0.307	0.294	0.301
DANZCR [7] (2018)	0.297	0.287	0.292	0.334	0.338	0.336
DADN [8] (2019)	0.305	0.291	0.298	0.359	0.353	0.356
TANSS [56] (2019)	0.314	0.303	0.309	0.362	0.355	0.359
LCALE [25] (2020)	0.367	<b>0.357</b>	0.362	0.414	0.394	0.404
<b>AAEGAN (Ours)</b>	<b>0.395</b>	0.346	<b>0.370</b>	<b>0.437</b>	<b>0.412</b>	<b>0.425</b>

network with the learning rate  $\mu$  as 0.0001 and the mini-batch size as 64. And the learning rate decay is decayed to be 1/10th of that after a certain number of epochs.

#### 4.2 Overall Results

The ZS-CMR MAP scores of our proposed AAEGAN and the compared methods on four datasets are shown in Table 2 and Table 3. Based on the experimental observations, the results of DNN-based methods are not significantly different from those of conventional shallow methods, and some traditional methods even have better performance than DNN-based methods. The reason may be that the DNN-based method requires a large amount of training data while there are not sufficient data in the zero-shot retrieval task. On the one hand, the division of the original dataset into two subsets results in the reduction of training data. On the other hand, no data of unseen classes are available during the training process, which limits the nonlinear mapping capability of DNN. Nevertheless, our AAEGAN approach still achieves the best retrieval performance on three of four datasets, with the only exception being PKU-XMediaNet. Even though LCALE [25], which is the most latest compared method, achieves a very large performance improvement on all datasets, our AAEGAN still outperforms it by a lot on Pascal Sentences and NUS-WIDE, and slightly beats it on Wikipedia. Specifically, on the Pascal dataset, the performance of our approach has been significantly improved, outperforming the counterparts LCALE on average (i.e., 0.425 vs. 0.404); on the NUS-WIDE dataset, our AAEGAN achieves the highest average MAP score of 0.586 compared with 0.567 obtained by LCALE; on the Wikipedia dataset, our approach still gains a slight improvement over LCALE (i.e., 0.370 vs. 0.362). On the another large-scale PKU-XMediaNet dataset, although our AAEGAN only gets the second place, we still improve a lot compared to third-best TANSS [56] method (i.e., 0.140 vs. 0.124). It is worth noting that the design of our method tends to be simplified, which may be the reason why we do not get the best results on this data set. However, this does not affect the effectiveness of the AAEGAN method shown in the results. The AAEGAN method is capable learn representative common feature for all modalities' data, and

Table 3. The MAP Scores of Zero-shot Retrieval for Our AAEGAN Approach and Other Compared Method on NUS-WIDE and PKU-XMediaNet Datasets

Methods	NUS-WIDE			PKU-XMediaNet		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA [33] (2010)	0.400	0.397	0.399	0.031	0.044	0.038
CFA [24] (2003)	0.410	0.355	0.383	0.058	0.071	0.065
KCCA [5] (2014)	0.402	0.413	0.408	0.040	0.057	0.049
JRL [59] (2014)	0.401	0.449	0.425	0.083	0.055	0.069
LGCFL [17] (2015)	0.396	0.422	0.409	0.062	0.064	0.063
DCCA [58] (2015)	0.406	0.407	0.407	0.039	0.043	0.041
DeepSM [49] (2017)	0.401	0.414	0.408	0.040	0.096	0.068
ACMR [44] (2017)	0.407	0.425	0.416	0.036	0.043	0.040
DEMZSL [60] (2017)	0.396	0.466	0.431	0.104	0.122	0.113
MASLN [57] (2018)	0.411	0.426	0.419	0.040	0.045	0.043
DANZCR [7] (2018)	0.416	0.469	0.443	0.106	0.117	0.112
DADN [8] (2019)	0.423	0.472	0.448	0.112	0.130	0.121
TANSS [56] (2019)	0.446	0.483	0.465	0.110	0.137	0.124
LCALE [25] (2020)	0.566	0.567	0.567	<b>0.135</b>	<b>0.164</b>	<b>0.150</b>
<b>AAEGAN (Ours)</b>	<b>0.584</b>	<b>0.587</b>	<b>0.586</b>	0.126	0.154	0.140

Table 4. Baseline Experiments for ZS-CMR on Wikipedia Dataset and Pascal Sentence Dataset

Baselines	Wikipedia			Pascal Sentences		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
AAEGAN ( $\mathcal{L}_{WGAN}, \mathcal{L}_{DA}$ )	0.260	0.250	0.255	0.152	0.119	0.135
AAEGAN ( $\mathcal{L}_{WGAN}$ )	0.320	0.305	0.312	0.379	0.353	0.366
AAEGAN ( $\mathcal{L}_{DA}$ )	0.270	0.259	0.265	0.155	0.145	0.150
AAEGAN ( $\mathcal{L}_{Rec}$ )	0.360	0.332	0.346	0.391	0.377	0.384
AAEGAN ( $\mathcal{L}_{MMD}$ )	0.372	0.342	0.357	0.396	0.383	0.390
AAEGAN ( $\mathcal{L}_{MSE}$ )	0.274	0.229	0.251	0.153	0.173	0.163
AAEGAN (All)	<b>0.382</b>	<b>0.353</b>	<b>0.367</b>	<b>0.437</b>	<b>0.412</b>	<b>0.425</b>

transfer knowledge to unseen classes effectively from class embeddings via distribution alignment, which leads to superiority and generalization under the ZSL setting.

### 4.3 Further Analysis on AAEGAN

**4.3.1 Baseline Experiments.** According to the final objective function in Equation (10), the proposed AAEGAN contains three kinds of loss terms that are the WGAN loss, the reconstruction loss, and the distribution alignment loss. Six variants of AAEGAN act as baselines and are conducted by excluding corresponding loss term in Equation (10) in the training pipeline. The influence of each loss item can be reflected by comparing performance with the full AAEGAN. Table 4 shows the retrieval results of the seven models on Wikipedia dataset and Pascal Sentences dataset. Here  $\mathcal{L}_{\neq}$  indicates the specific loss term that is excluded from Equation (10) when training the baselines.

By observing the data in Table 4, we can draw the following conclusions: (1) The baseline AAEGAN ( $\mathcal{L}_{WGAN}, \mathcal{L}_{DA}$ ) gains the worst performance, because it is constructed by the simple AE structure and is guided only by the reconstruction loss. (2) AAEGAN ( $\mathcal{L}_{WGAN}$ ) improves

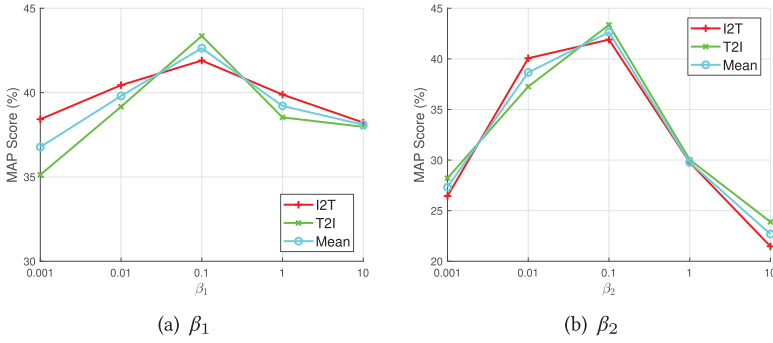


Fig. 4. A sensitivity analysis of the weighting coefficient  $\beta_1$  and  $\beta_2$  of our AAEGAN method on Pascal Sentence dataset.

performance greatly on the baseline AAEGAN ( $\mathcal{L}_{WGAN}, \mathcal{L}_{DA}$ ), proving that the distribution alignment is critical to our AAEGAN to encode more compact latent features across modalities via knowledge transfer from class embeddings. (3) AAEGAN ( $\mathcal{L}_{DA}$ ) also performs better than the baseline AAEGAN ( $\mathcal{L}_{WGAN}, \mathcal{L}_{DA}$ ) but does not improve as much as AAEGAN ( $\mathcal{L}_{WGAN}$ ). This shows that the WGAN loss term is not as important as distribution alignment, but is still an effective means of improving performance in practice. (4) AAEGAN ( $\mathcal{L}_{REC}$ ) performs worse than AAEGAN (ALL), indicating that the reconstruction loss term is also effective for retrieval. (5) We further analyze the effect of the MMD and MSE loss term in distribution alignment, as shown in AAEGAN ( $\mathcal{L}_{MMD}$ ) and AAEGAN ( $\mathcal{L}_{MSE}$ ). AAEGAN ( $\mathcal{L}_{MMD}$ ) is significantly better than AAEGAN ( $\mathcal{L}_{MSE}$ ), but results are higher when the MSE and MMD are combined. (6) The whole of our AAEGAN achieves the highest result among all the comparison algorithms, which proves once again that each component in Equation (10) is valid.

**4.3.2 Analysis on the Weighting Coefficient.** In the experiment, we further analyze how valid each component is by varying the value of  $\beta_1$  and  $\beta_2$  in Equation (10). The numerical range of both hyperparameters is set as  $[0.001, 10]$ . When a hyperparameter is changed at each time, another hyperparameter is fixed for fair comparison. As the experimental results shown in Figure 4, the MAP score varies with  $\beta_1$  and  $\beta_2$  and draws a bow shape, showing the optimal value for  $\beta_1$  and  $\beta_2$  are both 0.1. When  $\beta_1$  and  $\beta_2$  are too large (e.g., in  $[1, 10]$ ) or too small (e.g., in  $[0.001, 0.01]$ ), the final retrieval performance will be negatively affected, which indicates the contributions of loss term behind them.

**4.3.3 Analysis on the Dimension of Latent Features.** The dimension of latent features  $K_S$  plays a key role in retrieval performance. In the experiment, we investigate the effect of different dimensions  $K_S$  for each dataset, and the results are shown in Figure 5. We simply set  $K_S$  to the exponents of 2, i.e.,  $[64, 128, 256, 512, 1024, 2048]$ . The experimental results show that the optimal  $K_S$  for each dataset is different, and the best choices for Pascal Sentences, Wikipedia, Nus-Wide, PKU-XMediaNet are 256, 256, 512, and 1024. The performances reach the peaks under these choices and decrease when the lower dimension is not enough to encode all the valid information and the higher dimension will bring more noise. The reason for the difference in optimal dimensions for each dataset may be that each dataset contains different amounts of information and different modality gap for modalities. Hence, an optimal dimension of latent feature space is vital for achieving promising retrieval performance.

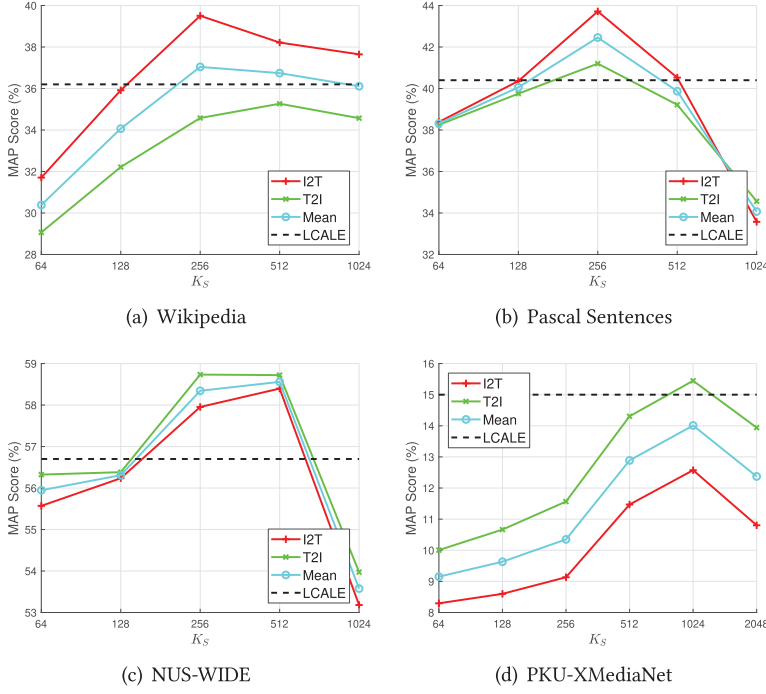


Fig. 5. A sensitivity analysis of the latent dimension  $K_S$  of our AAEGAN method on all datasets.

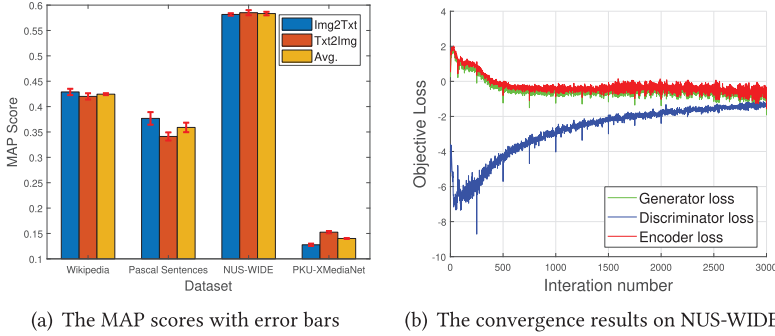


Fig. 6. A model robustness and convergence analysis of our AAEGAN method.

**4.3.4 Analysis on Model Robustness and Convergence.** Finally, we plot the error bar of MAP score for ZS-CMR tasks on all datasets and show it in Figure 6(a). We can observe that our AAEGAN approach is not sensitive to network initialization and obtain stable retrieval performance on all datasets. The reason is that the novel AAEGAN method with two coupled AEs for images and texts can generate effective and compact latent features with the guidance of class embeddings for robust training.

In addition, we also show the convergence experiment for our AAEGAN approach on the NUS-WIDE dataset to assess the process of adversarial training and evaluate its training efficiency. The curves on the loss value of the generators, discriminators, and encoders are shown in Figure 6(b). It is worth noting that the loss value are calculated by the objective function in Equations (11), (12), and (13) for the three components. Since the objective functions of the generators and the encoders



are similar, the trends of the generator loss and the encoder loss in Figure 6(b) look very similar. At the very beginning, the losses of the generators rise rapidly, because the synthetic features are very unreal. On the contrary, the losses of discriminators decrease very quickly, because the real features and the synthetic features are easy to distinguish. But the turning point comes very quickly. The generator progressively estimates the distribution of real features, which makes the discriminator difficult to distinguish. As a result, the losses of the generator and the discriminator change alternatively until they reach an equilibrium state in which the network converges. As mentioned above, the curve of the encoder loss shows a similar trend with that of the generator loss, showing that our AAEGAN can effectively learn the common latent space across different modalities with the supervision of class embeddings.

## 5 CONCLUSION

We propose a novel method named AAEGAN for the more realistic ZS-CMR scenario in this article. Our AAEGAN model is designed as an end-to-end framework that jointly incorporates common latent space learning, feature synthesis, and knowledge transfer. The three coupled AEs with distribution alignment collaboratively learn the latent space and establish correlations among modalities, while two coupled GANs capture the distribution under each modality data and facilitate the final retrieval task. Extensive experiments demonstrate the superiority of our AAEGAN method on ZS-CMR tasks, and the sensitivity analysis shows that each component is effective. For the future work, we plan to explore the extendability of our AAEGAN method on other modality data, e.g., sketch images, videos, and sensor data.

## REFERENCES

- [1] Zeynep Akata, Mateusz Malinowski, Mario Fritz, and Bernt Schiele. 2016. Multi-cue zero-shot learning with strong supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 59–68.
- [2] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. 2015. Evaluation of output embeddings for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2927–2936.
- [3] Martin Arjovsky and Léon Bottou. 2017. Towards principled methods for training generative adversarial networks. *arXiv:1701.04862*. Retrieved from <https://arxiv.org/abs/1701.04862>.
- [4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein gan. *arXiv:1701.07875*. Retrieved from <https://arxiv.org/abs/1701.07875>.
- [5] Lamberto Ballan, Tiberio Uricchio, Lorenzo Seidenari, and Alberto Del Bimbo. 2014. A cross-media model for automatic image annotation. In *Proceedings of the Annual ACM International Conference on Multimedia Retrieval (ICMR'14)*. 73:73–73:80.
- [6] Long Chen, Hanwang Zhang, Jun Xiao, Wei Liu, and Shih-Fu Chang. 2018. Zero-shot visual recognition using semantics-preserving adversarial embedding networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1043–1052.
- [7] 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 (IJCAI'18)*. 256–262.
- [8] J. Chi and Y. Peng. 2020. Zero-shot cross-media embedding learning with dual adversarial distribution network. *IEEE Trans. Circ. Syst. Vid. Technol.* 30, 4 (2020), 1173–1187.
- [9] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Ziping Luo, and Yan-Tao. Zheng. 2009. NUS-WIDE: A real-world web image database from national university of singapore. In *Proceedings of the ACM International Conference on Content-based Image and Video Retrieval (CIVR'09)*.
- [10] Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*. 215–223.
- [11] F. Feng, X. Wang, and R. Li. 2014. Cross-modal retrieval with correspondence autoencoder. In *Proceedings of the ACM Multimedia Conference*. 7–16.
- [12] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *Int. J. Comput. Vis.* 106, 2 (2014), 210–233.
- [13] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C. Courville. 2017. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*. 5767–5777.

- [14] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. 2004. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.* 16, 12 (2004), 2639–2664.
- [15] Geoffrey E. Hinton and Ruslan R. Salakhutdinov. 2006. Reducing the dimensionality of data with neural networks. *Science* 313, 5786 (2006), 504–507.
- [16] Xin Huang, Yuxin Peng, and Mingkuan Yuan. 2018. MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval. *IEEE Trans. Cybernet.* 48, 6 (2018), 143–156.
- [17] Cuicui Kang, Shiming Xiang, Shengcai Liao, Changsheng Xu, and Chunhong Pan. 2015. Learning consistent feature representation for cross-modal multimedia retrieval. *IEEE Trans. Multimedia* 17, 3 (2015), 370–381.
- [18] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv:1312.6114*. Retrieved from <https://arxiv.org/abs/1312.6114>.
- [19] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2009. Learning to detect unseen object classes by between-class attribute transfer. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 951–958.
- [20] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-based classification for zero-shot visual object categorization. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 3 (2014), 453–465.
- [21] A. B. L. Larsen, S. K. Sønderby, and O. Winther. 2015. Autoencoding beyond pixels using a learned similarity metric. In *Proceedings of the International Conference on Machine Learning*. 1558–1566.
- [22] Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the International Conference on Machine Learning (ICML'14)*. 1188–1196.
- [23] D. Li, N. Dimitrova, M. Li, and I. K. Sethi. 2003. Multimedia content processing through cross-modal association. In *Proceedings of the ACM Multimedia Conference*. 604–611.
- [24] Dongge Li, Nevenka Dimitrova, Mingkun Li, and Ishwar K. Sethi. 2003. Multimedia content processing through cross-modal association. In *Proceedings of the ACM International Conference on Multimedia*. 604–611.
- [25] 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 (AAAI'20)*. 11515–11522.
- [26] Ruoyu Liu, Yao Zhao, Liang Zheng, Shikui Wei, and Yi Yang. 2017. A new evaluation protocol and benchmarking results for extendable cross-media retrieval. *arXiv:1703.03567*. Retrieved from <https://arxiv.org/abs/1703.03567>.
- [27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv:1301.3781*. Retrieved from <https://arxiv.org/abs/1301.3781>.
- [28] Y. Peng, X. Huang, and J. Qi. 2016. Cross-media shared representation by hierarchical learning with multiple deep networks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'16)*. 3846–3853.
- [29] Yuxin Peng and Jinwei Qi. 2019. CM-GANs: Cross-modal generative adversarial networks for common representation learning. *Trans. Multimedia Comput. Commu. Appl.* 15, 1 (2019), 22:1–22:24.
- [30] Yuxin Peng, Jinwei Qi, Xin Huang, and Yuxin Yuan. 2018. CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network. *IEEE Trans. Multimedia* 20, 2 (2018), 405–420.
- [31] Viresh Ranjan, Nikhil Rasiwasia, and C. V. Jawahar. 2015. Multi-label cross-modal retrieval. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV'15)*. 4094–4102.
- [32] C. Rashtchian, M. Young, P. Hodosh, and J. Hockenmaier. 2010. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. 674–686.
- [33] N. Rasiwasia, J. Costa Pereira, E. Coviello, G. Doyle, G. Lanckriet, R. Levy, and N. Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the ACM International Conference on Multimedia (MM'10)*. 251–260.
- [34] Bernardino Romera-Paredes and Philip Torr. 2015. An embarrassingly simple approach to zero-shot learning. In *Proceedings of the International Conference on Machine Learning*. 2152–2161.
- [35] Mert Bulent Sariyildiz and Ramazan Gokberk Cinbis. 2019. Gradient matching generative networks for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2168–2178.
- [36] Heng Tao Shen, Luchen Liu, Yang Yang, Xing Xu, Zi Huang, Fumin Shen, and Richang Hong. 2020. Exploiting subspace relation in semantic labels for cross-modal hashing. *IEEE Trans. Knowl. Data Eng.* (2020). 1–16. <https://ieeexplore.ieee.org/document/8974240/>.
- [37] Yutaro Shigeto, Ikumi Suzuki, Kazuo Hara, Masashi Shimbo, and Yuji Matsumoto. 2015. Ridge regression, hubness, and zero-shot learning. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 135–151.
- [38] K. Simonyan and A. Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*. Retrieved from <https://arxiv.org/abs/1409.1556>.

- [39] Richard Socher, Milind Ganjoo, Christopher D. Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*. 935–943.
- [40] N. Srivastava and R. Salakhutdinov. 2012. Learning representations for multimodal data with deep belief nets. In *Proceedings of the International Conference on Machine Learning Workshop*.
- [41] N. Srivastava and R. Salakhutdinov. 2012. Multimodal learning with deep boltzmann machines. In *Advances in Neural Information Processing Systems*. 2222–2230.
- [42] Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H. S. Torr, and Timothy M. Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1199–1208.
- [43] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*. 1096–1103.
- [44] Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. 2017. Adversarial cross-modal retrieval. In *Proceedings of the ACM International Conference on Multimedia (MM'17)*. 154–162.
- [45] K. Wang, R. He, L. Wang, W. Wang, and T. Tan. 2011. Joint feature selection and subspace learning for cross-modal retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* 38, 10 (2011), 2010–2023.
- [46] K. Wang, R. He, W. Wang, L. Wang, and T. Tan. 2013. Learning coupled feature spaces for cross-modal matching. In *Proceedings of the IEEE International Conference on Computer Vision*. 2088–2095.
- [47] Yang Wang. 2020. Survey on deep multi-modal data analytics: Collaboration, rivalry and fusion. *arXiv:2006.08159*. Retrieved from <https://arxiv.org/abs/2006.08159>.
- [48] Yang Wang, Xuemin Lin, Lin Wu, and Wenjie Zhang. 2017. Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval. *IEEE Trans. Image Process.* 26, 3 (2017), 1393–1404.
- [49] Yunchao Wei, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan. 2017. Cross-modal retrieval with CNN visual features: A new baseline. *IEEE Trans. Cybernet.* 47, 2 (2017), 449–460.
- [50] Lin Wu, Yang Wang, Junbin Gao, and Xue Li. 2018. Where-and-when to look: Deep siamese attention networks for video-based person re-identification. *IEEE Trans. Multimedia* 21, 6 (2018), 1412–1424.
- [51] Lin Wu, Yang Wang, and Ling Shao. 2019. Cycle-consistent deep generative hashing for cross-modal retrieval. *IEEE Trans. Image Process.* 28, 4 (2019), 1602–1612.
- [52] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. 2018. Feature generating networks for zero-shot learning. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'18)*. 5542–5551.
- [53] Yongqin Xian, Saurabh Sharma, Bernt Schiele, and Zeynep Akata. 2019. f-VAEGAN-D2: A feature generating framework for any-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10275–10284.
- [54] Xing Xu, Li He, Huimin Lu, Lianli Gao, and Yanli Ji. 2019. Deep adversarial metric learning for cross-modal retrieval. *World Wide Web* 22, 2 (2019), 657–672.
- [55] Xing Xu, Kaiyi Lin, Lianli Gao, Huimin Lu, Heng Tao Shen, and Xuelong Li. 2020. Cross-modal common representations by private-shared subspaces separation. *IEEE Trans. Cybernet.* (2020), 1–14. <https://ieeexplore.ieee.org/document/9165187>.
- [56] Xing Xu, Huimin Lu, Jingkuan Song, Yang Yang, Heng Tao Shen, and Xuelong Li. 2020. Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval. *IEEE Trans. Cybernet.* 50, 6 (2020), 2400–2413.
- [57] 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 ACM Annual International Conference on Multimedia Retrieval (ICMR'18)*. 46–54.
- [58] F. Yan and K. Mikolajczyk. 2015. Deep correlation for matching images and text. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR'15)*. 3441–3450.
- [59] X. Zhai, Y. Peng, and J. Xiao. 2014. Learning cross-media joint representation with sparse and semisupervised regularization. *IEEE Trans. Circuits Syst. Vid. Technol.* 24, 6 (2014), 965–978.
- [60] Li Zhang, Tao Xiang, and Shaogang Gong. 2017. Learning a deep embedding model for zero-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2021–2030.
- [61] Yizhe Zhu, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal. 2018. A generative adversarial approach for zero-shot learning from noisy texts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1004–1013.
- [62] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu. 2013. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 1070–1076.

Received April 2020; revised August 2020; accepted September 2020