Joint Feature Synthesis and Embedding
Adversarial Cross-Modal Retrieval Revisited

Xu, Xing; Lin , Kaiyi ; Yang, Yang; Hanjalic, Alan; Shen, Heng Tao

DOI
10.1109/TPAMI.2020.3045530

Publication date
2022

Document Version
Final published version

Published in
IEEE Transactions on Pattern Analysis and Machine Intelligence

Joint Feature Synthesis and Embedding: Adversarial Cross-Modal Retrieval Revisited

Xing Xu, Kaiyi Lin, Yang Yang, Alan Hanjalic, *Fellow, IEEE*, and Heng Tao Shen, *Senior Member, IEEE*

Abstract—Recently, generative adversarial network (GAN) has shown its strong ability on modeling data distribution via adversarial learning. Cross-modal GAN, which attempts to utilize the power of GAN to model the cross-modal joint distribution and to learn compatible cross-modal features, is becoming the research hotspot. However, the existing cross-modal GAN approaches typically 1) require labeled multimodal data of massive labor cost to establish cross-modal correlation; 2) utilize the vanilla GAN model that results in unstable training procedure and meaningless synthetic features; and 3) lack of extensibility for retrieving cross-modal data of new classes. In this article, we revisit the adversarial learning in existing cross-modal GAN methods and propose Joint Feature Synthesis and Embedding (JFSE), a novel method that jointly performs multimodal feature synthesis and common embedding space learning to overcome the above three shortcomings. Specifically, JFSE deploys two coupled conditional Wassertein GAN modules for the input data of two modalities, to synthesize meaningful and correlated multimodal features under the guidance of the word embeddings of class labels. Moreover, three advanced distribution alignment schemes with advanced cycle-consistency constraints are proposed to preserve the semantic compatibility and enable the knowledge transfer in the common embedding space for both the true and synthetic cross-modal features. All these add-ons in JFSE not only help to learn more effective common embedding space that effectively captures the cross-modal correlation but also facilitate to transfer knowledge to multimodal data of new classes. Extensive experiments are conducted on four widely used cross-modal datasets, and the comparisons with more than ten state-of-the-art approaches show that our JFSE method achieves remarkably accuracy improvement on both standard retrieval and the newly explored zero-shot and generalized zero-shot retrieval tasks.

Index Terms—Cross-modal retrieval, embedding features, adversarial learning, zero-shot learning, knowledge transfer

1 Introduction

In our daily life, multimedia content usually consists of different modalities with associations to represent comprehensive semantics and information. For example, a piece of online news commonly consists of texts with corresponding images or videos describing the same event; a recipe for a special dish may also depict the cooking steps in terms of textual descriptions, photos and demo videos. Therefore, effective retrieval methods have become the primary issue for multimodal information acquisition. Under this situation, *cross-modal retrieval* [1], [2], [3], which aims to find the related results from one modality data given a query from another modality, has become a highlighted research topic in recent years. The related research has been concentrated on various retrieval scenarios, e.g., image-text retrieval [1], [2], recipe retrieval [4], [5], sketch-based image retrieval [6], [7], and scene image retrieval [8]. Unlike the unimodal retrieval scenarios such as image/video search that return retrieved results of the same modality with the query, cross-modal retrieval can provide search results of different modalities with a query of any modality. However, the challenge of "*heterogeneity gap*" [9] in cross-modal retrieval, i.e., the inconsistent distribution and representation of different modality data, makes it difficult to establish the cross-modal correlation and to directly measure the cross-modal similarity.

During the last decade, a large number of research studies have been devoted to cross-modal retrieval focusing on the *standard retrieval* scenario and following the paradigm of *common embedding space learning*. Using image and text modalities as an example, as the illustration of the standard retrieval shown in Fig. 1a, a collection of annotated cross-modal instances with each one consisting of a labeled image-text pair, is divided into a *source set* and a *target set*. The labels of the instances in the two sets belong to *the same scope of pre-defined classes*, (i.e., 10 classes in total). With the assumption that an intermediate common embedding space exists in the cross-modal data and can represent instances of relevant semantics as similar real-valued "embedding features", the typical pipeline for textual-visual cross-modal embedding is to encode the source set instances of individual modalities into their original features and then maps them into a common embedding space in the training phase. Later, the embedding features of the target set instances can be generated in the learned space, which can be finally utilized to measure the cross-modal similarities of

- *Xing Xu, Kaiyi Lin, Yang Yang, and Heng Tao Shen are with the Center for Future Media and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China. E-mail: xing.xu@uestc.edu.cn, {lky.linkaiyi, dlyyang}@gmail.com, shenhengtao@hotmail.com.*
- *Alan Hanjalic is with the Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, 2628 Delft, The Netherlands. E-mail: A.Hanjalic@tudelft.nl.*

Manuscript received 10 Nov. 2019; revised 9 Nov. 2020; accepted 6 Dec. 2020. Date of publication 17 Dec. 2020; date of current version 5 May 2022. (Corresponding author: Heng Tao Shen.) Recommended for acceptance by X. Bai. Digital Object Identifier no. 10.1109/TPAMI.2020.3045530

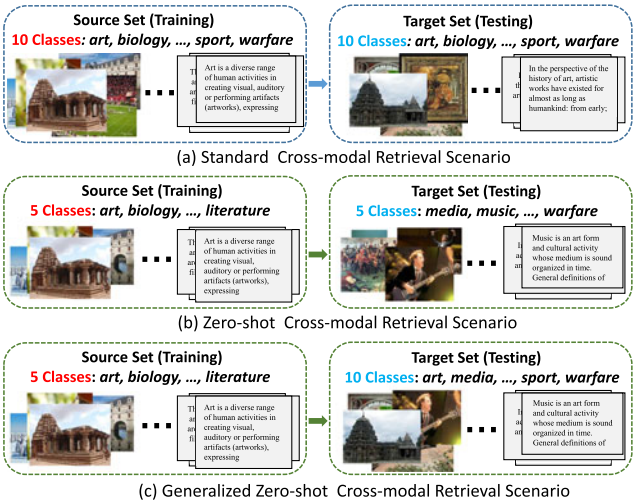


Fig. 1. Illustration of the three different retrieval scenarios: (a) standard cross-modal retrieval, where the data in the source set and the target set share the same 10 classes; (b) zero-shot cross-modal retrieval, where the data in each set are from 5 disjoint classes; and (c) generalized zero-shot cross-modal retrieval, where the target set has additional 5 classes that are absent in the source set.

pairwise instances and accomplish the cross-modal retrieval task.

Existing methods for common embedding space learning range from the early shallow methods [1], [2], [10] and the recent deep learning methods [11], [12], [13], [14], [15]. Notably, another group of hashing-based approaches [16], [17], [18], [19] have also been proposed to learn *binary codes* as embedding features in the common embedding subspace. These methods generally aim to tackle the efficiency issue for retrieval as binary embedding features boost the speed for Hamming distance computation and require much less storage consumption. Nevertheless, in this work, we mainly discuss the cross-modal retrieval problem that aims at learning real-valued embedding features. Specifically, the shallow learning methods usually obtain the embedding features of different modalities by learning a linear projection function for each modality. Due to the powerful ability of the deep neural network (DNN) on feature extraction and modeling the nonlinear cross-modal correlation, the DNN-based methods have been proposed and gained accuracy improvement. Especially, generative adversarial network (GAN) [20] (named "vanilla GAN" in this work) that consisting a generative module and a discriminative module has shown its remarkable ability to capture the observed data distribution. Several cross-modal GAN approaches [13], [15], [21], [22] have been proposed to improve the embedding learning in existing DNN-based approaches. Notably, the adversarial training scheme is widely used in these approaches to either boost the common embedding learning effectiveness or to enable the synthetic feature generation in individual modalities. Though significantly improved accuracy has gained for cross-modal retrieval, these cross-modal GAN approaches indeed still have the following shortcomings:

First, *they commonly require elaborately labeled multimodal data for supervised training*. To establish accurate cross-modal correlation, massive labeled multimodal data of high-quality is expected to provide informative correlation cues. However, different from the unimodal data annotation such as image tagging, collecting and labeling multimodal data are extremely label-consuming, as it is inevitable to browse relevant instances of multiple media types by seeing images, reading documents, watching videos and listening to audios. Though a few semi-supervised cross-modal GAN approaches recently [23], [24] attempt to exploit informative unlabeled data to complement the limited labeled cross-modal data, without label information, the unlabeled data can only be utilized in each modality separately, hence limitedly enriches the cross-modal correlation.

Second, *the GAN structure in existing cross-modal GAN approaches is less effective, where the vanilla GAN [20] model is typically adopted*. The vanilla GAN is initially designed for unimodal data synthesis. It may neglect the cross-modal correlation when it is applied to generate multimodal data, as the data synthesis in each modality is weakly associated. Besides, the training procedure of the vanilla GAN is usually unstable and hard to converge, thus, learning an effective common embedding space in existing cross-modal GAN approaches is even harder due to the influence of the probably incorrect synthetic multimodal data.

Third, *existing cross-modal GAN approaches lack of expansibility for retrieving data of new categories*. To address this issue, recent studies explore another two practical retrieval scenario of *zero-shot cross-modal retrieval* [25] and *generalized zero-shot cross-modal retrieval* [26]. The general settings of the two retrieval scenarios are illustrated in Figs. 1b and 1c, respectively. These two zero-shot retrieval scenarios aim to search the data of unseen classes in the target set only given the training data of limited seen classes in the source set. Compared with the standard retrieval scenario, these two retrieval scenarios are more difficult, because they require the learned retrieval model to not only cope with the heterogeneous distributions of multimodal data but also bridge the inconsistent semantics across the seen and unseen classes. A few recent works [22], [25], [26], [27], [28] adopt various knowledge transfer schemes from the research field of zero-shot learning (ZSL) [29], [30] for zero-shot cross-modal retrieval. These method have shown promising results compared with the traditional cross-modal retrieval approaches, however, they still adopt the conventional embedding learning schemes such as triplet ranking loss [13], [26], [31] or correlation loss [15], [32] for common embedding space learning. These loss terms inevitably ignore the semantic gap between seen and unseen classes, hence they are not appropriate to accomplish the knowledge transfer in zero-shot retrieval.

In this paper, we propose a novel method termed *Joint Feature Synthesis and Embedding* (JFSE) to jointly overcome the three shortcomings of the existing cross-modal GAN approaches. The proposed JFSE method is designed as an end-to-end neural network that fully incorporates multimodal feature synthesis, knowledge transfer, and common embedding space learning, to accomplish more accurate standard retrieval and zero-shot retrieval at once. The general flowchart of the proposed JFSE is illustrated in Fig. 2. Taking two modalities (i.e., images and texts) for example, each modality has an improved GAN structure, i.e., conditional Wassertein-GAN (cWGAN) and the two cWGAN are coupled for multimodal feature synthesis. During training, the proposed JFSE consists of two successive procedures:

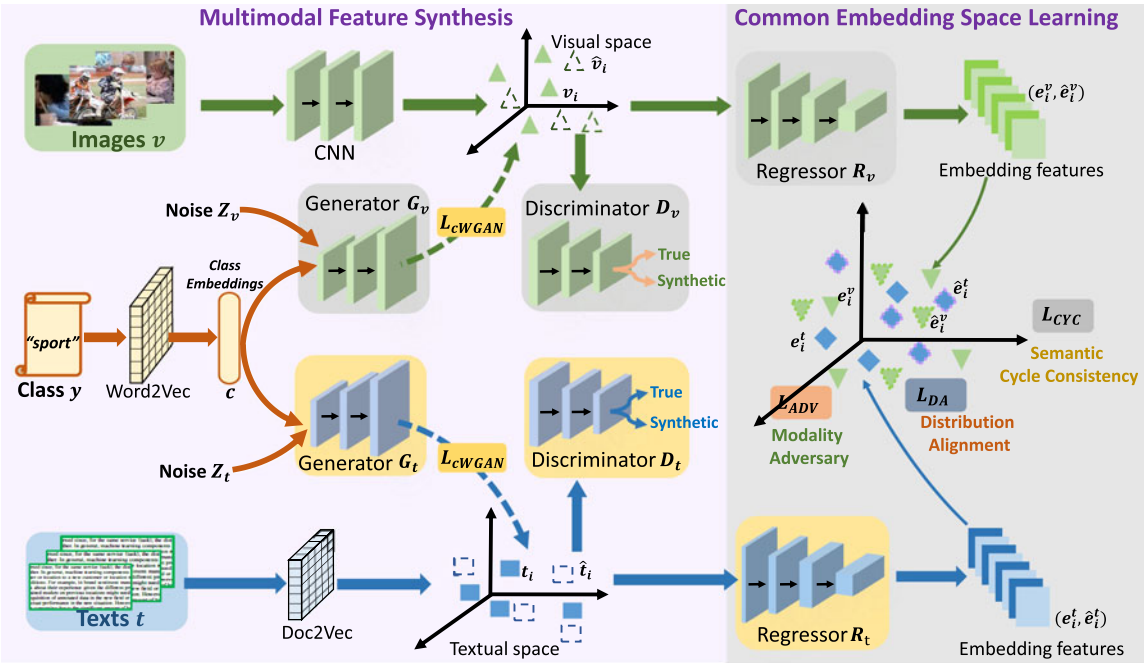


Fig. 2. The flowchart of our proposed JFSE method. It includes two coupled cWGANs that take the class embeddings as guidance to produce meaningful synthetic multimodal (e.g., image and text) features for robust training, as well as capture cross-modal correlation via distribution alignment under the common embedding space. The advanced cycle-consistency constraints further enhances the knowledge transfer between heterogeneous data with different classes.

multimodal feature synthesis and *common embedding space learning*. In the former procedure, the class embeddings of source set are utilized as side information in each cWGAN to synthesize modality-specific features via adversarial learning, while the shared class embeddings ensure the synthetic features in each modality incorporate the semantic of the class labels. In the latter process, both the synthetic and true features of two modalities are projected into the common embedding space via three effective distribution alignment strategies, where the cross-modal correlation is modeled and the knowledge transfer is conducted in the space. Meanwhile, a modality discriminator is added to distinguish the modality type of the mapped the true and synthetic embedding features. Moreover, an effective constraint of cycle-consistency is further deployed to minimize the error between both the true embedding features and its reconstructed ones, making the semantic consistency between the true input features and the embedding features within each modality are well preserved. During the test stage, for either standard retrieval or (generalized) zero-shot retrieval scenarios, the embedding features of the target set instances can be directly obtained by mapping their true input features into the learned common embedding space. As a result, the embedding features of different modalities can be directly used to compute the cross-modal similarity of pairwise instances.

We summarize our contributions in this paper as follows:

- *Effective multimodal feature synthesis with improved GAN structure*. We improve the vanilla GAN model widely used in existing cross-modal GAN approaches, and design an improved GAN structure, i.e., cWGAN, to simultaneously synthesize correlated multimodal features with the guidance of the class embeddings. Besides, the advantage of a stable training procedure in the cWGAN model benefits to produce meaningful synthetic features and to learn an effective common embedding space.
- *Advanced common embedding space learning*. To support the standard retrieval, zero-shot retrieval, and generalized zero-shot retrieval tasks, we develop three advanced distribution alignment schemes to capture cross-modal correlation and enable the knowledge transfer during common embedding space learning. All schemes perform distribution matching between the embedding features and the class embeddings of both the true and synthetic data, by considering the correlation of pairwise instances as well as the overall statistical properties of the heterogeneous data. Besides, to enable the knowledge transfer between classes, we introduce the advanced cycle-consistency constraints that preserve the semantic compatibility between the input features and the mapped embedding features of both true and synthetic data.

Difference With Our Conference Work. This paper is an extension and improvement of our previous work Adversarial Cross-Modal Retrieval (ACMR) [13]. The ACMR method is a pioneering work that utilizes adversarial learning to obtain more effective common embedding space. It has inspired the later studies on cross-modal GAN. Compared with ACMR, our JFSE approach in this work achieves three newly-added contributions: (1) Different from ACMR that only uses adversarial learning in common embedding space learning without vanilla GAN model, JFSE additionally leverages it for multimodal feature synthesis with the more advanced cWGAN model. This is beneficial to enrich the quantity of meaningful multimodal training data that

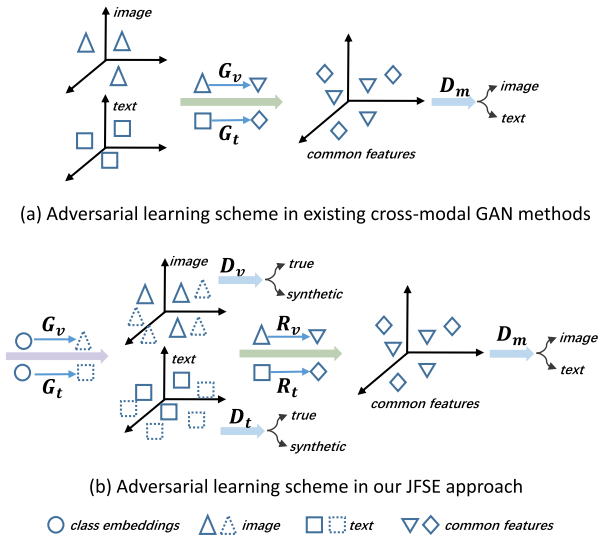


Fig. 3. Comparison of the adversarial learning used in existing cross-modal GAN methods and our JFSE approach.

facilitates to learn more effective common embedding space. (2) The JFSE approach adopts three advanced distribution alignment scheme with semantic consistency constraints for learning cross-modal correlation, which are more effective than the triplet ranking constraint used in ACMR. (3) The JFSE approach fully considers the expansibility of cross-modal retrieval and can be smoothly applied to both standard, zero-shot and generalized zero-shot retrieval scenarios, while ACMR is originally designed for standard retrieval and cannot be extended to the other two practical retrieval scenarios. Notably, the JFSE approach also provides more comprehensive analysis and different coverage of the distribution alignment schemes and the various retrieval scenarios than our related study in [33]. Extensive experiments with comprehensive ablation studies on four widely-used cross-modal retrieval datasets demonstrate the effectiveness of our proposed JFSE approach compared with a bundle of state-of-the-art methods on three different retrieval tasks.

Comparison With Other Cross-Modal GAN Approaches. As the comparison illustrated in Fig. 3, the early work of ACMR takes the feature projection as the "implicit" generator to generate the embedding features, which is not the true meaning of the GAN structure. Instead, it leverages a discriminator to distinguish the source of the projected features from images or text captions, which helps to learn a modality-invariant embedding space. The later works of GXN [21], CM-GANS [15], R2GAN [5], DADN [28] and TANSS [22] all have two pairs of generater-discriminator for individual modalities, where a generator is commonly a vanilla GAN model for independent image-image and text-text generation on the feature level or pixel level. Note that these approaches also have a discriminator to discriminate against the modality of an embedding feature. Different from the existing cross-modal GAN approaches, our JFSE approach takes two coupled cWGANs that consider the class embeddings as side information for cross-modal data synthesis on the feature level. Meanwhile, the class embeddings are treated as the common embedding space, which is more effective to correlate the feature synthesis for each modality and encapsulate a rich set of loss functions for effective distribution alignment. The proposed three distribution alignment schemes are more general and comprehensive that cover the diverse strategies used in CM-GANS [15], DADN [28] and TANSS [22]. Moreover, our JFSE method enables the knowledge transfer between seen and unseen classes for the practical scenarios of zero-shot and generalized zero-shot retrieval, which have not been investigated in previous cross-modal GAN approaches.

2 Related Work

In this section, we first present a briefly introduction of cross-modal retrieval, then we discuss the related techniques of zero-shot learning, adversarial learning, and data synthesis used in cross-modal retrieval.

Cross-Modal Retrieval. The main effort on cross-modal retrieval is to bridge the "heterogeneity gap", i.e., establishing the correlation of heterogeneous data with inconsistent distribution and different representation. As aforementioned in Section 1, the mainstream idea is learning a common embedding space where embedding features of data in different modalities can be obtained for cross-modal similarity measurement. Regarding the variations of the basic models, existing methods can be grouped into *shallow learning methods* and *DNN-based methods*. The core issue in the shallow learning methods [1], [2], [34], [35] is to learn effective linear projection functions that can map the input features of different modalities into a common space. Various constraints are proposed to ensure the learned projection functions to maximize the correlation between the pairwise data of different modalities. More discussion of the shallow learning methods can be found in the survey [3], [36].

With the powerful ability of DNN models on feature representation, recent studies [12], [14], [31], [32], [37], [38] based on DNNs now become mainstream solutions for cross-modal retrieval. These methods focus on using DNN to capture the nonlinear cross-modal correlation and to generate more compact embedding features. Sufficient high-quality and large-scale labeled cross-modal data are usually required to train these methods, but collecting and annotating cross-modal data is labor-consuming. Even though several semi-supervised learning strategies [23], [24] can be adopted in cross-modal retrieval by taking unlabeled unimodal data into account as the complementary source, without label information, the unlabeled data can only be utilized in each modality separately and has limited impact on enriching the cross-modal correlation. Differently, in our JFSE method, we exploit the data synthesis strategy to synthesize correlated multimodal data rather than explicitly use additionally unlabeled data.

Adversarial Learning for Cross-Modal Retrieval. The adversarial learning strategy is first proposed in the generative adversarial network [20]. The vallina GAN is composed of a generator module and a discriminator module. Specifically, the former module tries to produce fake data that can mimic the empirical distribution of true data, while the latter one aims to distinguish the true data and the generated fake data. The parameters in the two modules are learned via the adversarial learning strategy. GAN has also been extended to conditional GAN [39], which utilizes additional side

information such as class labels or sentence descriptions for both generative and discriminative models to incorporate semantical cues in the generated data. Compared with the vanilla GAN, conditional GAN is superior to producing more meaningful synthetic data related to the given side information.

Existing GAN based methods mostly focus on the data generation problem with unimodal data. Recently, several cross-modal GAN studies [13], [15], [15] have been proposed for cross-modal retrieval task with multimodal data. For example, Wang *et al.* [13] proposed adversarial cross-modal retrieval that first utilizes adversarial learning to learn a modality-invariant common embedding space. Peng *et al.* [15] designed a cross-modal GAN (CM-GAN) structure to jointly capture the correlation within inter-modality and intra-modality, where the adversarial learning mechanism involves two kinds of discriminative modules to preserve both the intra-modality and inter-modality discrimination. Gu *et al.* [21] incorporated two generative models of pixel-level image-text generation and text-image generation into the textual-visual feature embedding. Zhu *et al.* [5] further designed a special GAN structure with one generator module and two discriminator modules in their method recipe retrieval GAN (R2GAN), which makes the generation of a pixel-level image from textual recipe feasible. Generally, these methods usually introduce a modality discriminator to judge the modality type of the semantic features, while a cross-modal generator reduces the difference of cross-modal embedding features to confuse the modality discriminator. Besides, they commonly adopt the vanilla GAN as the basic structure for generator-discriminator pair, which may not be stable to train in practice. Though our JFSE approach is also a GAN based model, it leverages an improved GAN structure of cWGAN, which not only inherits the merit of WGAN on more stable training procedure but also takes the class embeddings as guidance for multimodal feature synthesis with semantic correlation preserved.

Zero-Shot Learning for Cross-Modal Retrieval. The research on zero-shot learning aims to imitate the ability of humans to recognize new (unseen) classes which are not observed in the training stage. As visual knowledge of new classes is unavailable during training, ZSL requires additional side information to complement the unknown visual knowledge. Most existing ZSL methods [29], [30], [40], [41], [42] focus on unimodal scenario, and utilize side information such as attributes and word vectors of class as the intermediate semantics for recognizing unseen classes.

For cross-modal retrieval, existing standard retrieval approaches may not well adapt to the *zero-shot* retrieval scenario. To this end, a few recent works [22], [25], [27], [28], [33] have been proposed for zero-shot cross-modal retrieval. They make use of the word embeddings of class labels (i.e., class embeddings) from pre-trained natural language processing models as auxiliary knowledge, to perform cross-modal correlation learning and knowledge transfer at the same time. These approaches differ in how the class embeddings are leveraged for knowledge transfer. In [27], class embeddings are used as conditional input signal for the data reconstruction in the auto-encoder structure, while in [22], [25], [28], [33], the class embeddings are directly treated as embedding features in the to-be-learned common embedding space. Notably, in this work, we develop an improved GAN structure of cWGAN that takes the class embeddings as the conditional signal to guide multimodal feature synthesis procedure, which enhances the correlation of synthetic features of different modalities.

Data Synthesis. Data synthesis is an effective strategy to deal with the lack of training data, such as in the learning from imbalanced data and few-shot learning problem [41]. In the zero-shot scenario, the recent study in [43] shows that using visual feature space for zero-shot learning is intrinsically effective to alleviate the hubness problem. Later, several other works [42], [44], [45] leverage various data augmentation schemes based on GANs or variational auto-encoder (VAE) [46] modules to estimate the empirical data distributions and synthesize visual features for unseen classes. However, these synthesis strategies are limited to unimodal data such as images and videos. Recently, a more complex task of text-to-image synthesis [47], [48] that requires to generate photo-realistic images according to a short text description has been explored. This task is commonly treated as a cross-modal data synthesis problem with constrained direction (i.e., from text modality to image modality), and extend the GANs module with the text conditioned auxiliary information for pixel-level image generation.

Similarly, several existing cross-modal GAN approaches such as GXN [21] and R2GAN [5] also explore feasible schemes for cross-modal synthesis in the vanilla GAN module. However, the feature synthesis procedure of the individual modalities in these methods is implicitly correlated or even without association. Differently, in our JFSE approach, the class embeddings are leveraged as shared cues for feature synthesis in each modality. Therefore it can be considered as the more advanced task of *multimodal data synthesis* since correlated multimodal features can be jointly generated and well preserve the semantics according to the auxiliary information of the class embeddings.

3 Proposed Method

3.1 Problem Formulation

Suppose we have a cross-modal dataset $\mathcal{O} = \{\mathcal{O}_S, \mathcal{O}_T\}$ consisting of multiple instances (i.e., labeled image-text pairs), where $\mathcal{O}_S$ and $\mathcal{O}_T$ denote the source set for training and the target set for testing, respectively. Here $\mathcal{O}_S = \{\mathbf{v}_i, \mathbf{t}_i, \mathbf{c}_i, y_i\}_{i=1}^{N_S}$ and $\mathcal{O}_T = \{\mathbf{v}_j, \mathbf{t}_j, \mathbf{c}_j, y_j\}_{j=1}^{N_T}$ have N_S and N_T instances, respectively. For each instance in the two sets, $\mathbf{v}_i, \mathbf{v}_j \in \mathbb{R}^{d_v}$; $\mathbf{t}_i, \mathbf{t}_j \in \mathbb{R}^{d_t}$; $\mathbf{c}_i, \mathbf{c}_j \in \mathbb{R}^{s}$; $y_i, y_j \in \mathbb{R}^1$ denote the d_v-dim image feature vector, d_t-dim text feature vector, s-dim class embeddings and the index of class label, respectively. In addition, the class labels of all the instances in $\mathcal{O}_S$ are from a pre-defined class set $\mathcal{Y}_S = \{1, \ldots, C\}$, where C denotes the total number of classes. Notably, depending on the specific retrieval scenario, the class set $\mathcal{Y}_T$ of $\mathcal{O}_T$ may be variant compared with $\mathcal{Y}_S$. According to the illustration of the three retrieval scenarios in Fig. 1, for the standard retrieval, $\mathcal{Y}_S = \mathcal{Y}_T$ as the scope of classes in the two sets $\mathcal{O}_S$ and $\mathcal{O}_T$ are identical; for the zero-shot retrieval, $\mathcal{Y}_S \bigcap \mathcal{Y}_T = \phi$ due to the disjoint classes in the two sets; for the generalized zero-shot retrieval, $\mathcal{Y}_S \subset \mathcal{Y}_T$ as the classes in $\mathcal{O}_T$ have unseen classes that are not contained in $\mathcal{O}_S$.

Due to the potential semantic inconsistency of the two class sets $\mathcal{Y}_S$ and $\mathcal{Y}_T$, the intrinsic properties of the two sets

$\mathcal{O}_S$ and $\mathcal{O}_T$ may also be different. In this work, our goal is to learn an effective common embedding space from $\mathcal{O}_S$, which can be well adapted to represent the data in $\mathcal{O}_S$ for all three retrieval scenarios.

3.2 Key Components of Our JFSE Method

As the overall framework of our JFSE method shown in Fig. 2, we first describe its core components on the standard retrieval task. Then we will discuss its extension to the zero-shot retrieval and generalized zero-shot retrieval scenarios.

3.2.1 Multimodal Feature Synthesis

Different from existing studies [45], [49] that perform unimodal data synthesis on pixel-level, we put forward an improved GAN structure based on two coupled conditional Wasserstein GAN [42] to synthesize multimodal input data (i.e., images and texts) on the feature-level. This is clearly different with the existing cross-modal GAN approaches [13], [15], [22], [28] that take the vanilla GAN model as the basic model for each pathway. Notably, the cWGAN is superior to the vanilla GAN since it is more stable to achieve robust training procedure. Besides, it can naturally take the class embeddings as the input side information for feature synthesis, which thereby provides semantic supervision to produce more informative synthetic features. Specifically, each cWGAN contains a generator $G_*(\cdot)$ and a discriminative module $D_*(\cdot)$ that are built with several fully-connected layers in the network. Here $* = v, t$ denotes the image and text modalities, respectively.

In the image pathway, given the feature $\mathbf{v}_i$ of an image with its class label y_i, the class embedding $\mathbf{c}_i$ of y_i and a noise vector $\mathbf{z}_v$ are first concatenated as the input for the generator G_v. Here the class embedding $\mathbf{c}_i \in \mathbb{R}^s$, $\mathbf{z}_v \in \mathbb{R}^m$ is randomly sampled from Gaussian distribution, and G_v is expected to generate synthetic image feature $\tilde{\mathbf{v}}_i$ according to the semantic information of $\mathbf{c}_i$ as $\tilde{\mathbf{v}}_i = G_v(\mathbf{z}_v, \mathbf{c}_i)$. Subsequently, the discriminative module D_v distinguishes the true image feature $\mathbf{v}_i$ and its class embeddings $\mathbf{c}_i$ against the synthetic image feature $\tilde{\mathbf{v}}_i$ and $\mathbf{c}_i$. Finally, the feature synthesis for all the images in $\mathcal{O}_s$ can be formulated by the cWGAN loss as

$$\begin{aligned} L^v_{cWGAN}(\theta_{G_v}, \theta_{D_v}) &= \mathbb{E}[D_v(\mathbf{v}, \mathbf{c}; \theta_{D_v})] - \mathbb{E}[D_v(\tilde{\mathbf{v}}, \mathbf{c}; \theta_{D_v})] \\ &- \alpha \mathbb{E}[(\| \nabla_{\hat{\mathbf{v}}} D_v(\hat{\mathbf{v}}, c; \theta_{D_v}) \|_2 - 1)^2]. \end{aligned} \tag{1}$$

Here $\mathbb{E}[\cdot]$ represents the expected value for all images, the penalty coefficient λ is a preset constant value, θ_{G_v} and θ_{D_v} denote the parameters in the two modules. Notably, the third term $\nabla_{\hat{\mathbf{v}}}$ denotes the gradient penalty as defined in [44]. It requires the gradient of D_v to have unit norm along the straight line between the pairs of true and synthetic features, which ensures the numeric stability when updating the parameters. In addition, $\hat{\mathbf{v}} = a\mathbf{v} + (1-a)\tilde{\mathbf{v}}$ with a is a weighted constant value sampled from the uniform distribution $U(0, 1)$.

Likewise, in the text pathway, the feature synthesis for all the texts follows the same procedure as the above image pathway, which can be formulated as:

$$\begin{aligned} L^t_{cWGAN}(\theta_{G_t}, \theta_{D_t}) &= \mathbb{E}[D_t(\mathbf{t}, \mathbf{c}; \theta_{D_t})] - \mathbb{E}[D_t(\tilde{\mathbf{t}}, \mathbf{c}; \theta_{D_t})] \\ &- \alpha \mathbb{E}[(\| \nabla_{\hat{\mathbf{t}}} D_t(\hat{\mathbf{t}}, \mathbf{c}; \theta_{D_t}) \|_2 - 1)^2], \end{aligned} \tag{2}$$

where $\mathbb{E}[\cdot]$ denotes the expected value for all texts, θ_{G_t} and θ_{D_t} are the model parameters in the two modules in the text pathway. Finally, the multimodal feature synthesis procedure integrates the two losses Eqs. (1) and (2) of the two pathways as

$$L_{cWGAN} = L^v_{cWGAN}(\theta_{G_v}, \theta_{D_v}) + L^t_{cWGAN}(\theta_{G_t}, \theta_{D_t}). \tag{3}$$

3.2.2 Common Embedding Space Learning With Distribution Alignment

With the help of the two cWGANs, we can generate sufficient synthetic instances of multimodal features $\{\tilde{\mathbf{v}}_j, \tilde{\mathbf{t}}_j\}_{j=1}^{\tilde{N}}$. Note that the value of $\tilde{N}$ can be different from N of the true instances $\{\mathbf{v}_i, \mathbf{t}_i\}_{i=1}^{N}$ as one true instance can generate multiple synthetic embedding features in our cWGAN model. We further equip two regressors $R_v(\cdot)$ and $R_t(\cdot)$ (with several fully-connected layers) after the two cWGANs in our JFSE model to utilize class embeddings to map both the true and synthetic features of two modalities into a common embedding space. Given the true feature $\mathbf{v}_i$ and $\mathbf{t}_i$ for the ith image-text pair in $\mathcal{O}_S$, its corresponding embedding features $\mathbf{e}^v_i$ and $\mathbf{e}^t_i$ can be respectively obtained by the two regressors R_v and R_t as $\mathbf{e}^v_i = R_v(\mathbf{v}_i)$ and $\mathbf{e}^t_i = R_t(\mathbf{t}_i)$. Likewise, for the jth synthetic image-text pair, its embedding features of two modalities can also be obtained by the two regressors as $\tilde{\mathbf{e}}^v_i = R_v(\tilde{\mathbf{v}}_i)$ and $\tilde{\mathbf{e}}^t_i = R_t(\tilde{\mathbf{t}}_i)$, respectively.

The widely used schemes like correlation loss [11], [15], [32] and triplet ranking loss [13], [31] consider to learn the cross-modal correlation on the pairwise instance-level. One serious problem of these schemes is that they pay more attention to each image-text pair, which is trivial to model the discrepancy between the data distributions of two modalities. Moreover, when considering the zero-shot and generalized zero-shot retrieval scenarios, the semantic gap between the heterogeneous data of both seen and unseen classes is also necessary to be considered. Unfortunately, these schemes are weak to cope with these two scenarios. Therefore, in this work, we develop three advanced schemes to model the cross-modal correlation on the statistical data distribution-level. These proposed schemes can align the distributions of both true and synthetic features of two modalities in the common embedding space with transferable knowledge according to their semantics. In the following parts, we give detailed descriptions of the three proposed schemes.

Cross-Modal Maximum Mean Discrepancy (CMMD). The target of the CMMD scheme is to maximize the mean discrepancy of both the true and synthetic embedding features of pairwise instances of different modalities. It differs from the feature adaptation method [50] that processing the unimodal data. Considering the true embedding features $\{\mathbf{e}^v_i\}_{i=1}^{N_S}$ and $\{\mathbf{e}^t_j\}_{j=1}^{N_S}$ of pairwise instances in $\mathcal{O}_S$, the CMMD scheme measures the cross-modal similarity for all instances via a kernel-based distance function, which is deriving as

$$\begin{aligned} L^{true}_{CMMD}(\theta_R) = &\frac{1}{N_S^2}\sum_{i,j=1}^{N_S} \kappa\left(\mathbf{e}^v_i, \mathbf{e}^v_j\right) - \frac{2}{N_S^2}\sum_{i,j=1}^{N_S} \kappa\left(\mathbf{e}^v_i, \mathbf{e}^t_j\right) \\ &+ \frac{1}{N_S^2}\sum_{i,j=1}^{N_S} \kappa\left(\mathbf{e}^t_i, \mathbf{e}^t_j\right). \end{aligned} \tag{4}$$

where i, j denote the indices of a pair. Notably, the kernel function $\kappa(x_i, x_j) = \sum_n \eta_n \exp\{-\frac{1}{2\sigma_n}\|x_i - x_j\|^2\}$ linearly combines multiple Radial basis function (RBF) kernels, where η_n and σ_n denote the weight value and standard deviation of the nth RBF kernel, respectively. In our experiment, we empirically use 19 RBF kernels and set the values of η_n and σ_n all these kernels according to the official implementation of the MMD loss in Tensorflow toolkit.

Meanwhile, the CMMD scheme also applies to measure the cross-modal similarity of the synthetic embedding features, where in this case the formulation of the loss term is consistent with the case of the true embedding features in Eq. (4). Intuitively, the final loss term for the CMMD scheme is a combination of the loss terms of both the true and synthetic embedding features as

$$L_{DA}^{CMMD} = L_{CMMD}^{true}(\theta_R) + L_{CMMD}^{syn}(\theta_R). \tag{5}$$

By minimizing Eq. (5), the cross-modal correlation of both the true and synthetic embedding features can be captured and the modality discrepancy can also be narrowed.

Cross-Modal Correlation Alignment (CMCA). The CMCA scheme is another scheme the measures the cross-modal correlation by exploring the overall data distribution of all instances in different modalities. Unlike the CMMD scheme that models the overall cross-modal correlation based on pairwise instances, the CMCA scheme treats the embedding features of all the true and synthetic instances as matrix forms, and measure the cross-modal distance with the covariance of a matrix. The covariance reflects the second-order statistics of the data distributions [51], which may be more informative than the first-order modeling in the above CMMD scheme.

First, let $\mathbf{U} = \{\mathbf{e}_i^v\}_{i=1}^{N_S} \in \mathbb{R}^{N_S \times s}$ and $\mathbf{Z} = \{\mathbf{e}_i^t\}_{i=1}^{N_S} \in \mathbb{R}^{N_S \times s}$ be the matrices consisting the embedding features of the true image and text instances. Then define the covariance term of $\mathbf{U}$ and $\mathbf{Z}$ as

$$\mathbf{C}^v = \frac{1}{N_S - 1}(\mathbf{U}^\top \mathbf{U} - \frac{1}{N_S}(\mathbf{1}^\top \mathbf{U})^\top (\mathbf{1}^\top \mathbf{U})), \tag{6}$$

$$\mathbf{C}^t = \frac{1}{N_S - 1}(\mathbf{Z}^\top \mathbf{Z} - \frac{1}{N_S}(\mathbf{1}^\top \mathbf{Z})^\top (\mathbf{1}^\top \mathbf{Z})), \tag{7}$$

where $\mathbf{1}$ is an N_S dimensional column vector with all elements equal to 1. Similarly, we can also form the matrices $\tilde{\mathbf{U}} = \{\tilde{\mathbf{e}}_j\}_{j=1}^{\tilde{N}_S}$ and $\tilde{\mathbf{Z}} = \{\tilde{\mathbf{e}}_j\}_{j=1}^{\tilde{N}_S}$ according to the embedding features of the synthetic image and text instances. Then, the covariances of $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{Z}}$ can be derived as

$$\tilde{\mathbf{C}}^v = \frac{1}{\tilde{N}_S - 1}(\tilde{\mathbf{U}}^\top \tilde{\mathbf{U}} - \frac{1}{\tilde{N}_S}(\mathbf{1}^\top \tilde{\mathbf{U}})^\top (\mathbf{1}^\top \tilde{\mathbf{U}})), \tag{8}$$

$$\tilde{\mathbf{C}}^t = \frac{1}{\tilde{N}_S - 1}(\tilde{\mathbf{Z}}^\top \tilde{\mathbf{Z}} - \frac{1}{\tilde{N}_S}(\mathbf{1}^\top \tilde{\mathbf{Z}})^\top (\mathbf{1}^\top \tilde{\mathbf{Z}})), \tag{9}$$

Then the distribution alignment of the CMMD scheme is accomplished by reducing the error between the covariances of both the true and synthetic instances in each modality. The loss term can be written as

$$\mathcal{L}_{DA}^{CMCA}(\theta_R) = \frac{1}{4m^2}\|\mathbf{C}^v - \mathbf{C}^t\|_F^2 + \frac{1}{4m^2}\|\tilde{\mathbf{C}}^v - \tilde{\mathbf{C}}^t\|_F^2, \tag{10}$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm for matrices.

Cross-Modal Projection Matching (CMPM). The CMPM scheme is designed to model the cross-modal correlation by minimizing the Kullback-Leibler (KL) divergence between the normalized matching distributions and projection compatibility distributions of different modalities [52]. Specifically, considering B true image and text embedding features in a mini-batch during training, the probability of matching $\mathbf{e}_i^v$ to $\xi(\mathbf{e}_j^t)$ is derived as

$$p_{i,j} = \frac{\exp(\mathbf{e}_i^{v\top} \xi(\mathbf{e}_j^t))}{\sum_{k=1}^n \exp(\mathbf{e}_i^{v\top} \xi(\mathbf{e}_k^t))}, \tag{11}$$

where $\xi(\mathbf{e}_j^t) = \frac{\mathbf{e}_j^t}{\|\mathbf{e}_j^t\|}$ denotes the normalized text embedding features, $\mathbf{e}_i^{v\top}\xi(\mathbf{e}_j^t)$ reflects the scalar projection between image embedding features $\mathbf{e}_i^v$ onto text embedding feature $\xi(\mathbf{e}_j^t)$ and the probability $p_{i,j}$ denotes the percentage of this scalar projection between pairs in a mini-batch. If the image embedding feature is more similar to the text embedding feature, the scalar projection will be larger. Considering in each mini-batch there might be more than one matched image-text pair that has the same class, the true matching probability is normalized as follows: $q_{i,j} = \frac{y_{i,j}}{\sum_{k=1}^B y_{i,k}}$. Where $y_{i,j} = 1$, $y_{i,j} = 0$ respectively represents the matched image-text pair and the unmatched pair. For each mini-batch, the cross-modal projection matching loss of associating each image embedding feature with correctly matched text embedding feature is derived as

$$\mathcal{L}_{i2t}^{true}(\theta_R) = -\frac{1}{B}\sum_{i=1}^n \sum_{j=1}^n p_{i,j} \log\left(\frac{p_{i,j}}{q_{i,j} + \epsilon}\right), \tag{12}$$

where ϵ is a small number for preventing numerical problems and the Eq. (12) actually denotes the KL divergence from the true matching distribution q_i to the probability of matching p_i. The same procedure is followed to compute text-to-image matching loss by replacing $\mathbf{e}_i^v$ with $\mathbf{e}_j^t$ in Eq. (11). Then, the bi-directional cross-modal projection matching loss $\mathcal{L}_{CMPM}^{true} = \mathcal{L}_{i2t}^{true} + \mathcal{L}_{t2i}^{true}$.

Similarly, we also adopt bi-directional cross-modal projection matching loss in the synthetic image and text features. Finally, the CMPM scheme combines the loss terms of both the true and synthetic common embedding features as

$$\mathcal{L}_{DA}^{CMPM}(\theta_R) = \mathcal{L}_{CMPM}^{true}(\theta_R) + \mathcal{L}_{CMPM}^{syn}(\theta_R). \tag{13}$$

3.2.3 Modal-Adversarial Consistency Learning

The above distribution alignment schemes consider the statistical property of the true and synthetic embedding features of two modalities. Here, we further import the modal-adversarial consistency learning, which focuses on reducing the instance-level difference of the true and synthetic embedding features in each modality. Note that this scheme is different from our previous ACMR method and the other cross-modal GAN approaches that only focus on the true embedding features. In particular, we introduce a modality discriminator $D_m(\cdot)$ to distinguish the modality type of the embedding features that are produced by the two regressors R_v and R_t, while the two regressors attempt to produce confusing embedding features with less difference between

modalities, as an adversarial training style. The D_m consists of a gradient reversal layer (GRL) [53] and several fully-connected layers parameterized by θ_M, where the GRL is an identity transformation that makes the modality discriminator and the two regressors follow an adversarial training style.

First, let $\{l_i^*\}_{i=1}^{N_S}$ and $\{\tilde{l}_j^*\}_{j=1}^{\tilde{N}_S}$, where l_i^*, $\tilde{l}_j^* \in \{0,1\}$, $* = v,t$ represent the modality labels that are assigned to the true and synthetic embedding features $\{\mathbf{e}_i^v\}_{i=1}^{N_S}, \{\mathbf{e}_i^t\}_{i=1}^{N_S}$ and $\{\tilde{\mathbf{e}}_i^v\}_{i=1}^{\tilde{N}_S}, \{\tilde{\mathbf{e}}_i^t\}_{j=1}^{\tilde{N}_S}$ of two modalities, respectively. Then the modal-adversarial consistency loss on the true and synthetic embedding features can be written by maximizing the predicted score of D_m and the modality labels, as

$$\begin{aligned} L_{ADV}(\theta_M) &= \frac{1}{N_S}\sum_{*=v,t}\sum_{i=1}^{N_S} f_{sigmoid}(D_M(\mathbf{e}_i^*), l_i^*; \theta_M) \\ &+ \frac{1}{\tilde{N}_S}\sum_{*=v,t}\sum_{j=1}^{\tilde{N}_S} f_{sigmoid}(D_M(\tilde{\mathbf{e}}_i^*), \tilde{l}_i^*; \theta_M)), \end{aligned} \tag{14}$$

where $f_{sigmoid}(x, \hat{x}) = -(x\log\hat{x} + (1-x)\log(1-\hat{x}))$ is the sigmoid cross entropy loss function. By maximizing L_{ADV}, the heterogeneity gap between two modalities in the common embedding space can be explicitly reduced. Besides, the consistency of both the true and synthetic embedding features across different modalities can also be enhanced.

3.2.4 Reconstructing Embedding Features With Cycle Consistency

Recent studies [42], [54], [55] have shown that combining the cycle-consistency constraints in data reconstruction is surprisingly effective at learning the mapping across domains. These works can effectively accomplish pixel-level image-to-image translation between different domains, even without explicitly labeled image pairs. Inspired by these works, to further enhance the mutual associations between the multimodal representations in the raw feature space and the embedding features in the common embedding space, we further extend the cycle consistency constraints to ensure the embedding feature reconstruction procedure in a meaningful way. In particular, it is expected that both the true and synthetic embedding features to reconstruct back to their true class embeddings to form a cycle-loop, where the class embeddings incorporate the underlying relationship between two modalities. As a result, the cycle consistency constraints in this work are formulated by minimizing the reconstruction error as

$$\begin{aligned} L_{CYC} &= \sum_{*=v,t}\mathbb{E}[\| \mathbf{c} - R_*(G_*(\mathbf{c}, \mathbf{z}; \theta_G); \theta_R) \|_2^2] \\ &+ \mathbb{E}[\|\mathbf{c} - R_v(\mathbf{v}; \theta_R)\|_2^2] + \mathbb{E}[\|\mathbf{c} - R_t(\mathbf{t}; \theta_R)\|_2^2]. \end{aligned} \tag{15}$$

Notably, the true and synthetic features of two modalities can be implicitly paired, since the regressors R_* effectively preserve the discriminability of the class embeddings in these features. At this point, our JFSE approach is different from our previous ACMR and other methods [14], [15], [38] that require additional classification module to predict the semantic labels of the embedding features. Moreover, another advantage of using the cycle consistency constraints in Eq. (15) is that the knowledge transfer across different classes (e.g., seen and unseen classes) can be effectively accomplished by the intermediate of class embeddings. Therefore, it facilitates to learn more compact embedding features for the target set multimodal data and improves the performance on the zero-shot and generalized zero-shot retrieval scenarios.

3.3 Optimization

Our JFSE is an end-to-end architecture that combines the above four key components in the two procedures of multimodal feature synthesis and common embedding space learning. Finally, the full objective function can be formulated by integrating the related loss terms as

$$\begin{aligned} &\theta_G, \theta_D, \theta_R, \theta_M = arg\min_{\theta_G,\theta_R}\max_{\theta_D,\theta_M} L_{cWGAN}(\theta_G, \theta_D) \\ &+ \beta L_{DA}(\theta_G, \theta_R) + \gamma L_{ADV}(\theta_M) + \lambda L_{CYC}(\theta_G, \theta_R). \end{aligned} \tag{16}$$

Here L_{DA} denotes the three alternative schemes L_{DA}^{CMMD}, L_{DA}^{CMCA}, and L_{DA}^{CMPM}. The optimization algorithm for the loss terms in Eq. (16) follows a mini-max game under the adversarial training style. In practice, we utilize the stochastic gradient descent optimization algorithms (e.g., Adam [56] optimizer) to joint accomplish the above optimization procedure. The detailed training procedure of our JFSE method is summarized in Algorithm 1.

Algorithm 1. Training of the Proposed JFSE Approach

Input: Source set $\mathcal{O}_s = \{(\mathbf{v}_i, \mathbf{t}_i, y_i)\}_{i=1}^{N_S}$, class-embeddings $\{\mathbf{c}_i\}_{j=1}^{C}$, batch size B, the number of iterations T, hyper-parameters β, γ, λ, and learning rate μ.

1: **repeat**
2: Sample image-text pair $\{(\mathbf{v}_i, \mathbf{t}_i, y_i)\}_{b=1}^{B}$ and class embeddings $\{c_j\}_{b=1}^{B}$ with batch.
3: **while** $i < T$ **do**
4: Update $\theta_D \leftarrow \theta_D - \mu\nabla_{\theta_D}(-L_{cWGAN})$.
5: **end while**
6: Update θ_G by $\theta_G \leftarrow \theta_G - \mu\nabla_{\theta_G}(L_{cWGAN} + \beta L_{DA} + \gamma L_{ADV} + \lambda L_{CYC})$.
7: Update θ_R by $\theta_R \leftarrow \theta_R - \mu\nabla_{\theta_R}(\beta L_{DA} + \gamma L_{ADV} + \lambda L_{CYC})$.
8: Update θ_M by $\theta_M \leftarrow \theta_M - \mu\nabla_{\theta_M}\gamma L_{ADV}$.
9: **until** The loss function in Eq. (16) converges or reaches the maximum iterations.

Output: The regressors $R_v(\cdot)$ and $R_t(\cdot)$ that can map the true input features of images and texts to the common embedding space, respectively.

3.4 Extension to Zero-Shot Retrieval

Since the classes in the source set and the target set have no overlap for the zero-shot retrieval scenario, to narrow the semantic gap between the classes in the two sets, we further employ the distribution alignment schemes on both the synthetic embedding features of the seen and unseen classes to transfer knowledge via the class embeddings. Since the generalized zero-shot retrieval just differs from the zero-shot retrieval on the testing stage, i.e., the test data of both seen and unseen classes are incorporated in the target set, therefore, we can directly follow the same training procedure as for the zero-shot retrieval.

Intuitively, though the true features of two modalities for unseen classes are unavailable during training, we can still

generate their synthetic features $\{\tilde{\mathbf{e}}_k^v, \tilde{\mathbf{e}}_k^t\}_{k=1}^{\tilde{N}^T}$ according to the class embeddings of unseen classes. Take the distribution alignment scheme CMMD (denoted in Eq. (5)) as instance, we can adopt one or more of the three schemes CMMD, CMCA, and CMPM to align the synthetic features of both seen and unseen classes of two modalities in the common embedding space, which can be formulated as the following loss:

$$L_{DA}^{zero} = L_{DA}^{seen}(\theta_G, \theta_R) + L_{DA}^{unseen}(\theta_G, \theta_R), \tag{17}$$

where $L_{DA}^{seen}(\theta_G, \theta_R)$ denotes the domain adaptation loss in seen classes defined in Eq. (5). The term $L_{DA}^{unseen}(\theta_G, \theta_R)$ for unseen classes has the similar form as in Eq. (4), where the $\{\tilde{\mathbf{e}}_k^v, \tilde{\mathbf{e}}_k^t\}_{k=1}^{\tilde{N}^T}$ are used.

Moreover, we also address the modal-adversarial consistency learning to generate a modality-invariant embedding features for both the seen and unseen classes, as

$$L_{ADV}^{zero}(\theta_M) = L_{ADV}^{seen}(\theta_M) + L_{ADV}^{unseen}(\theta_M), \tag{18}$$

where $L_{ADV}^{unseen}(\theta_M)$ denotes the modal-adversarial loss in unseen classes.

Similar to Eq. (15), now the cycle-consistency constraints force the synthetic embedding features of both seen classes and unseen classes to reconstruct their true class-embeddings. Therefore, it would be beneficial for zero-shot and generalized zero-shot retrieval scenarios since the knowledge can be transferred from seen classes to unseen classes via the shared semantics of class embeddings. The cycle-consistency constraints in this case is formulated as

$$L_{CYC}^{zero} = L_{CYC}^{seen}(\theta_R) + L_{CYC}^{unseen}(\theta_R). \tag{19}$$

Note that the final objective function for both zero-shot and generalized zero-shot retrieval scenarios is to replace Eqs. (5), (14) and (15) by Eqs. (17), (18) and (19) in Eq. (16), respectively. The optimization procedure generally follows the steps in Algorithm 1. Once the model parameters of the regressors $R^v(\cdot)$ and $R^t(\cdot)$ are obtained, we can use the two regressors to map the unseen cross-modal instances to generate their embedding features to perform both zero-shot and generalize zero-shot retrieval.

4 Experiment

4.1 Experimental Setup

Datasets and Features. Following the existing cross-modal GAN approaches, we adopt four widely-used cross-modal datasets (i.e., Wikipedia [1], Pascal Sentences [57], NUS-WIDE [58] and PKU-XMediaNet [24]) in the experiment. All datasets consist of image-text pairs with manually annotated class labels. Table 1 summarizes the general information of the four datasets and the train/test data splits on both standard and (generalized) zero-shot retrieval tasks.

Following the recent works [13], [14], [22], [28], we utilize the deep neural networks (DNNs) models pre-trained on large corpus of images and texts to extract compact feature representations for the image and text modality data. Specifically, each image on all datasets is represented by a 4,096 dimensions convolutional neural network (CNN) feature vector extracted from the 7th fully-connected (FC) layer in the pre-trained 19-layer VGGNet [59] for all the compared methods for fair comparison. In the standard retrieval, we extract the 300 dimensional feature vector for each text using the WordCNN (WCNN) [60] as in [15], [24]. In the zero-shot retrieval, similar as the previous work [22], [25], the text representations are the 300-dimensional feature vector extracted from the Doc2Vec [61] (DV) model pre-trained on Wikipedia, and the 300-dimensional word-embeddings for classes are extracted by Word2Vec [62] (WV) model pre-trained on Google News. For a class label that has multiple words, we take the averaged vector based on the word vectors extracted for all words. In Table 1, we also list the different types of features used for different retrieval scenarios, where "F (I)", "F (T)" and "F (C)" denote features for images, texts, and classes respectively. Besides, for "Class" column, "*/*" denotes the number of seen and unseen classes for zero-shot retrieval.

TABLE 1
The General Statistics of the Four Datasets Under the Standard (top panel) and Zero-Shot Retrieval (bottom panel) Scenarios

Datasets	Train	Test	Class	F (I)	F (T)	F (C)
Wikipedia	2,173	462	10	VGG	WCNN	-
Pascal Sentences	800	100	20	VGG	WCNN	-
NUS-WIDE	8,000	1,000	10	VGG	WCNN	-
PKU-XMediaNet	32,000	4,000	200	VGG	WCNN	-
Wikipedia	2,173	693	5/5	VGG	DV	WV
Pascal Sentences	800	200	10/10	VGG	DV	WV
NUS-WIDE	42,941	28,661	5/5	VGG	DV	WV
PKU-XMediaNet	32,000	8,000	100/100	VGG	DV	WV

Retrieval Tasks and Evaluation Metrics. To comprehensively evaluate the effectiveness of the proposed method, we conduct three different cross-modal retrieval tasks on all datasets, including standard retrieval, zero-shot retrieval, and generalized zero-shot retrieval. Each task has two subtasks: image-to-text (Img2Txt) and text-to-image (Txt2Img), where one modality data is used as a query to match the most related items in the other modality. The data split of the three retrieval tasks is briefly illustrated in Fig. 4, and their detailed settings are as follows:

Standard Retrieval. Like the traditional protocol adopted in [1], [13], [14], [15], we the default data split provided by

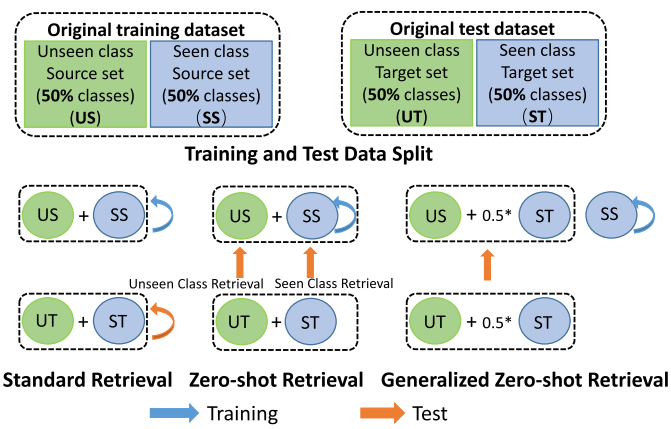


Fig. 4. The training/test data split of each dataset on the three retrieval scenarios, where the arrows with different colors represents the training and test procedures for each scenario.

each dataset for the standard retrieval task. Here both the training and test sets contain all the classes in one dataset. Note that the training set is used to learn the retrieval model and the evaluation is conducted on the test set.

Zero-Shot Retrieval. Following the protocol in [25], [28], [63], for each dataset, we integrate its original test set and the validation set to form a test set in this retrieval task. To ensure the disjoint and balanced classes between the seen and unseen set, for each dataset, we take the image-text pairs of one half (50 percent) classes as the seen set and those of the remaining classes as the unseen set. As a result, the original training set and test set in each dataset are respectively divided into four subsets: *seen class source set* (SS), *unseen class source set* (US), *seen class target set* (ST), and *unseen class target set* (UT). Based on the new data split for zero-shot retrieval, we conduct two retrieval modes: *seen class retrieval* and *unseen class retrieval*. Specifically, for the seen class retrieval, the seen class source set is used for training and the seen class target set is considered as queries to retrieve on the seen class source set during testing. For the unseen class retrieval, it also uses the seen class source set for training but takes the unseen class target set as queries to retrieve on the unseen class source set during testing.

Generalized Zero-Shot Retrieval. For this retrieval task, the seen class source set is still utilized for training. Differently, as an extension of the above zero-shot retrieval task, we further divide the seen class target set into two equal parts (i.e., 50 percent each). Then each part is added to the unseen class source set and unseen class target set, respectively, to form the database set and the query set during testing. This data split scheme ensures that both seen and unseen classes are involved during retrieval, which is coherent with the protocol advocated in [26].

We adopt the standard metric of *mean average precision* (MAP) score to evaluate the retrieval performance of all methods on the two subtasks of Img2Txt and Txt2Img. MAP is the mean value of Average Precision (AP) of each query, where a larger MAP score indicates better retrieval performance. Following [25], [28], we calculate the results of MAP score on all returned instances on all the datasets. Besides, following [2], [13], [14], we also use the *precision-recall curve* (PR curve) as a complementary metric for the MAP score. The PR curve can elaborately reflect the retrieval performance with different cases of the returned instances. All the experiments are performed on a workstation with two NVIDIA GeForce GTX 1080 Ti GPUs.

Details of Network. The proposed JFSE approach is implemented by the popular deep learning toolkit Tensorflow. For the two cWGAN modules, each generator of the image and text pathways are built with two fully connected layers, where the first layer is followed by a LeakyReLU activation function layer and the second layer is followed by a ReLU activation function layer. The numbers of hidden units of the two FC layers are [4096, K_*], where $* = v, t$, and $K_v = 4096$ and $K_t = 300$ are the dimensions of the input image and text features, respectively. Each discriminator module is formed by two FC layers with dimensions $[4096, 1]$, and it maps the embedding features into a real-valued confidence score to distinguish the true or synthetic categories. Each regressor for the image and text modalities contains three FC layers with the number of hidden units as $[4096, 4096, 300]$. Each FC layer in the regressor following a ReLU layer. The modality classifier is built with three FC layers with the number of hidden units as $[300, 150, 1]$.

TABLE 2
Comparison of Different Distribution Alignment Schemes on All Datasets

Dataset	Different schemes	Task		
		Img2Txt	Txt2Img	Avg.
Wikipedia	COR	0.519	0.473	0.496
	TRIP	0.522	0.476	0.499
	CMMD	0.524	0.477	0.501
	CMCA	0.524	0.478	0.501
	CMPM	**0.527**	**0.484**	**0.506**
	CMPM + CMMD	0.524	0.481	0.503
	CMPM + CMCA	0.523	0.483	0.503
	CMPM + CMMD + CMCA	0.520	0.477	0.499
Pascal Sentences	COR	0.613	0.597	0.605
	TRIP	0.614	0.601	0.608
	MMD	0.623	0.608	0.616
	CMCA	0.625	0.610	0.618
	CMPM	**0.632**	**0.610**	**0.621**
	CMPM + CMMD	0.622	0.609	0.616
	CMPM + CMCA	0.628	0.607	0.618
	CMPM + CMMD + CMCA	0.623	0.604	0.614
NUS-WIDE	COR	0.554	0.561	0.558
	TRIP	0.557	0.564	0.561
	CMMD	0.561	0.573	0.567
	CMCA	0.560	0.575	0.568
	CMPM	**0.564**	**0.583**	**0.574**
	CMPM + CMMD	0.562	0.576	0.569
	CMPM + CMCA	0.561	0.578	0.570
	CMPM + CMMD + CMCA	0.560	0.575	0.568
PKU-XMediaNet	COR	0.691	0.683	0.687
	TRIP	0.694	0.678	0.686
	CMMD	0.693	0.687	0.690
	CMCA	0.696	0.685	0.691
	CMPM	**0.701**	**0.691**	**0.696**
	CMPM + CMMD	0.696	0.685	0.691
	CMPM + CMCA	0.697	0.685	0.691
	CMPM + CMMD + CMCA	0.689	0.683	0.685

The Adam optimizer with an initial learning rate of $\mu = 0.0001$ is used to train our proposed JFSE network, and the mini-batch size is set to 64. We tune the hyper-parameters β, γ, and λ on all datasets and provide the sensitivity analysis of the three parameters in the later experiment (Fig. 10).

4.2 Baseline Experiments

First, we conduct two baseline experiments to investigate the effectiveness of each component in our JFSE method.

4.2.1 Effect of Different Distribution Alignment Schemes

In Section 3, we have developed three distribution alignment schemes CMMD, CMCA, and CMPM to capture the cross-modal correlation on the statistical data distribution-level. To comprehensively evaluate the effect of the three schemes, we additionally take into account another two widely used schemes on the pairwise instance-level, i.e., correlation similarity (COR) [15], [32] and triplet ranking loss (TRIP) [13], [31]. In this experiment, we train our JFSE model using each scheme individually or combinative in the loss term L_{DA}. The overall retrieval performance of our JFSE method with the different distribution alignment schemes on four datasets are shown in Table 2.

When considering the retrieval results of each scheme separately, we can observe that the three proposed schemes CMMD, CMCA, and CMPM obtain much better retrieval performance than those of the COR and TRIP schemes. It indicates that modeling the cross-modal correlation on the data distribution-level is more effective for common embedding

TABLE 3
Baseline Experiments for Standard Retrieval on All Datasets

Dataset	Baseline	Task		
		Img2Txt	Txt2Img	Avg.
Wikipedia	JFSE ($\cancel{L_{cWGAN}}$)	0.512	0.461	0.487
	JFSE ($\cancel{L_{CYC}}$)	0.484	0.437	0.461
	JFSE ($\cancel{L_{DA}}$)	0.518	0.471	0.495
	JFSE ($\cancel{L_{ADV}}$)	0.516	0.475	0.496
	JFSE (All)	**0.527**	**0.484**	**0.506**
Pascal Sentences	JFSE ($\cancel{L_{cWGAN}}$)	0.602	0.594	0.598
	JFSE ($\cancel{L_{CYC}}$)	0.532	0.527	0.530
	JFSE ($\cancel{L_{DA}}$)	0.612	0.596	0.604
	JFSE ($\cancel{L_{ADV}}$)	0.615	0.597	0.606
	JFSE (All)	**0.632**	**0.610**	**0.621**
NUS-WIDE	JFSE ($\cancel{L_{cWGAN}}$)	0.547	0.562	0.555
	JFSE ($\cancel{L_{CYC}}$)	0.521	0.527	0.524
	JFSE ($\cancel{L_{DA}}$)	0.554	0.562	0.558
	JFSE ($\cancel{L_{ADV}}$)	0.557	0.568	0.563
	JFSE (All)	**0.564**	**0.583**	**0.574**
PKU-XMediaNet	JFSE ($\cancel{L_{cWGAN}}$)	0.681	0.672	0.677
	JFSE ($\cancel{L_{CYC}}$)	0.537	0.511	0.524
	JFSE ($\cancel{L_{DA}}$)	0.688	0.679	0.684
	JFSE ($\cancel{L_{ADV}}$)	0.686	0.678	0.682
	JFSE (All)	**0.701**	**0.691**	**0.696**

space learning than on the pairwise instance-level. Among the three proposed schemes, as the CMPM scheme shows superior overall performance to the other two schemes, we further evaluate the performance of its combinations with the other two schemes. The results of different combinative schemes are shown in the last three rows in each panel of Table 2. It can be seen that the results of the three combinative counterparts are inferior to the results of the single CMPM scheme. Due to the similar purpose of modeling the cross-modal correlation on the data distribution-level, the three schemes may have potential incompatibility and negative influence on each other during model training. Based on the above observations, we choose the best individual scheme of CMPM in our JFSE method to compare with the other approaches for different retrieval scenarios in all the latter experiments.

4.2.2 Effect of Each Key Components

To assess the impact of different components of JFSE on the retrieval performance, we further design a baseline experiment by disabling each loss term in Eq. (16) during training.

Table 3 shows the retrieval results of the original JFSE and its six baseline experiments on all datasets, where $\cancel{L_*}$ represents a loss term that is excluded from the L_{all} during training. We can observe that: 1) When excluding the cycle-consistency constraints in JFSE, the baseline $\cancel{L_{CYC}}$ obtains the worst performance. The reason is that cycle-consistency constraints can force the generated synthetic features to reconstruct their class embeddings back by minimizing the error between the class embeddings and their reconstructed ones. As aforementioned in Eq. (15), it plays a similar role as the classification loss adopted in existing cross-modal GAN methods, which preserves the discriminability of the embedding features in the common embedding space. 2) The baseline $\cancel{L_{cWGAN}}$ that excludes the key components of the two cWGANs obtains the second-worst performance among all comparison baselines, because multimodal feature synthesis network with two coupled cWGANs can generate meaningful multimodal features for robust training and improving the common embedding space learning. 3) The performance of the baseline $\cancel{L_{DA}}$ that excludes the CMPM loss is inferior to the full model, showing that an advanced alignment scheme can effectively capture the correlation between different modalities, as well as ensuring the smoothness of the knowledge transfer in the common embedding space. 4) When excluding the component of modal-adversarial semantic learning (baseline $\cancel{L_{ADV}}$), the retrieval performance drops significantly, showing the importance of the modal-adversarial semantic learning on making the embedding features discriminative for classes while invariant for different modalities.

4.3 Comparison on Standard Retrieval

Compared Methods. We choose 15 state-of-the-art methods for a comprehensive comparison with our proposed JFSE approach. Specifically, CCA [1], CFA [34], KCCA [69], JRL [66] and LGCFL [10] are traditional methods, while DNN-based methods include Corr-AE [32], DCCA [12], CMDH [37] Deep-SM [67], ACMR [13], CCL [14], CBT [68], MCSM [31], CM-GANS [15] and DSCMR [38]. To ensure the fair comparison, we use the source code published by the authors of the comparison methods and use the same image and text features as our JFSE approach. As the source code of CBT [68] is unavailable, we directly report its results on three other datasets except for the NUS-WIDE dataset. Since different compared approaches are derived from different hypotheses and built with diverse network architectures, the properties of the learned common space (including the dimension of the common space) in each approach are different. Nevertheless, we tune the optimal dimension of the learned common embedding space for each method to ensure that the best performance of each method is obtained. Additionally, each comparison method performs ten runs and the averaged MAP scores are reported.

Analysis on the Results. The MAP scores of our JFSE method with the compared methods on the four datasets are shown in Table 4. Due to the strong nonlinear mapping ability of the deep network, most DNN-based methods have better performance than the traditional methods, and the CM-GAN method achieves the best accuracy in all compared methods. Nevertheless, the shallow method JRL performs better than some DNN-based methods, because JRL uses a semi-supervised learning scheme with sparse regularization constraint. It is notable that our proposed JFSE significantly outperforms all the compared methods on all datasets in terms of the average MAP scores. Specifically, on the PKU-XMediaNet dataset, our method improves the average MAP score from 0.627 to 0.696 compared to the compared method DSCMR. The similar remarkable improvement can be observed on the other three datasets, our JFSE approach still maintains its advantage and achieves the best retrieval performance.

Based on the results reported in Table 4, our proposed JFSE approach achieves the best retrieval performance attribute to the following key factors: 1) a novel feature synthesis network architecture that uses cWGAN model with semantic correlation for robust training; 2) effective cycle-consistency constraints that preserve the semantic compatibility to learn more robust common embedding space; 3) an advanced distribution alignment scheme that fully captures the cross-modal correlation of different modalities in the common embedding space; 4) modality-adversarial learning that makes the embedding features discriminative for classes but invariant for different modalities to strengthen the cross-modal semantic consistency.

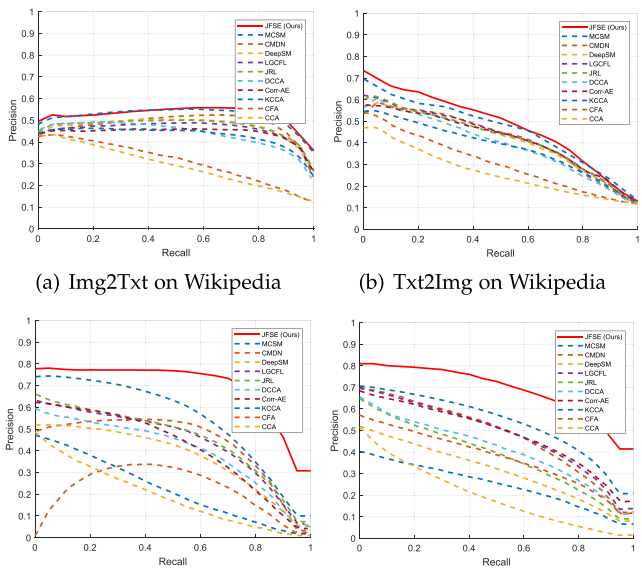
TABLE 4
The MAP Scores of Standard Retrieval for Our JFSE Approach and Other Compared Methods on All Datasets

Methods	Wikipedia			Pascal Sentencess			NUS-WIDE			PKU-XMediaNet		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA [1] (2010)	0.298	0.273	0.286	0.203	0.208	0.206	0.167	0.181	0.174	0.212	0.217	0.215
CFA [64] (2003)	0.319	0.316	0.318	0.476	0.470	0.473	0.406	0.435	0.421	0.252	0.400	0.326
KCCA [65] (2014)	0.438	0.389	0.414	0.488	0.446	0.467	0.351	0.356	0.354	0.252	0.270	0.261
JRL [66] (2014)	0.479	0.428	0.454	0.563	0.505	0.534	0.466	0.499	0.483	0.488	0.405	0.447
LGCFL [10] (2015)	0.466	0.431	0.449	0.539	0.503	0.521	0.453	0.485	0.469	0.441	0.509	0.475
DCCA [12] (2015)	0.445	0.399	0.422	0.568	0.509	0.539	0.452	0.465	0.459	0.425	0.433	0.429
Corr-AE [32] (2014)	0.442	0.429	0.436	0.532	0.521	0.527	0.441	0.494	0.468	0.469	0.507	0.488
CMDN [37] (2016)	0.487	0.427	0.457	0.544	0.526	0.535	0.492	0.542	0.517	0.485	0.516	0.501
Deep-SM [67] (2017)	0.478	0.422	0.450	0.560	0.539	0.550	0.497	0.478	0.488	0.399	0.342	0.371
ACMR [13] (2017)	0.468	0.412	0.440	0.538	0.544	0.541	0.519	0.542	0.531	0.536	0.519	0.528
CCL [14] (2018)	0.505	0.457	0.481	0.576	0.561	0.569	0.481	0.520	0.501	0.537	0.528	0.533
CBT [68] (2018)	0.516	0.464	0.490	0.602	0.583	0.592	-	-	-	0.577	0.575	0.576
MCSM [31] (2018)	0.516	0.458	0.487	0.598	0.598	0.598	0.533	0.561	0.547	0.540	0.550	0.545
DSCMR [38] (2019)	0.515	0.479	0.497	0.595	0.598	0.597	0.554	0.563	0.559	0.622	0.632	0.627
CM-GANS [15] (2019)	0.521	0.466	0.494	0.603	0.604	0.604	0.544	0.562	0.553	0.567	0.551	0.559
JFSE (Ours)	**0.527**	**0.484**	**0.506**	**0.632**	**0.610**	**0.621**	**0.564**	**0.583**	**0.574**	**0.701**	**0.691**	**0.696**

Furthermore, the PR curves of our JFSE method and several counterparts on Wikipedia and PKU-XMediaNet datasets are demonstrated in Fig. 5. It can be observed that our JFSE approach consistently maintains the highest accuracy of all recall levels in the PR curves, which demonstrates the effectiveness of our approach. Especially on the PKU-XMediaNet datasets, the curve height greatly exceeds other comparison methods with a large margin. Besides, Fig. 6 further demonstrates typical Img2Txt and Txt2Img retrieval exemplars obtained by our JFSE approach and two latest DNN-based compared methods ACMR [13] and DSCMR [38] on the PKU-XMediaNet dataset.

4.4 Comparison on Zero-Shot Retrieval

*Compared Methods.*We further compare our JFSE approach with 14 state-of-the-art methods for zero-shot retrieval by evaluating the unseen class retrieval performance. There



(a) Img2Txt on Wikipedia (b) Txt2Img on Wikipedia

(c) Img2Txt on PKU-XMediaNet (d) Txt2Img on PKU-XMediaNet

Fig. 5. The precision-recall (PR) curves of standard retrieval results on the Wikipedia and PKU-XMediaNet datasets.

compared methods contains 5 latest approaches designed for zero-shot retrieval task: DEMZSL [43], MASLN [27], DANZCR [25], DADN [28] and TANSS [22]. Besides, we also take 8 methods designed for the standard retrieval into consideration, which are directly evaluated under the zero-shot retrieval scenario. The overall comparisons for both seen class retrieval and unseen class retrieval are shown in Tables 5 and 6, respectively.

Results on Seen Class Retrieval. From Table 5, it can be seen that the experimental results of the seen class retrieval show similar trends compared to the standard retrieval. As the seen classes of test data are included in the training set, the MAP scores of the seen class retrieval are significantly higher than the results of the unseen class retrieval. We can clearly see that our JFSE approach consistently outperforms all the counterparts, including both the traditional methods and the DNN-based methods. The reason is that our JFSE can synthesize effective multimodal features through two coupled cWGANs with robustness training procedure, and learns a common embedding space through distribution adaptation to capture the correlation of cross-modal data. Notably, on the large-scale PKU-XMediaNet dataset, our JFSE obtains the highest average MAP score of 0.773 compared with 0.761 from the best counterpart DSCMR.

Results on Unseen Class Retrieval. The overall comparison of the unseen class retrieval is shown in Table 6. Unlike the results on seen class retrieval scenario in Table 5, all the methods obtain remarkably lower MAP scores on the unseen class retrieval for all datasets. The reason is that the data of unseen classes is unavailable during training, hence degrades the extendability of the learned model on test data of unseen classes. We can see that some traditional methods achieve better results than DNN-based methods. For example, on the Pascal Sentences dataset, the traditional method KCCA achieves good performance and is superior to most DNN-based methods. We also notice that TANSS obtains the best retrieval accuracy on all datasets among the DNN-based compared methods. Comparing with TANSS, our proposed JFSE approach gains remarkable improvement accuracy on all datasets. In particular, on the Wikipedia

Query	Method	Top 5 Results (Img2Txt)				
 Horse	JFSE (Ours)	Horse (Equus ferus caballus) is one of two extant subspecies of Equus ferus. It is an odd-toed …	A non-castrated male horse four years old and older. The term "horse" is sometimes used colloquially …	In the Australian registry, there are various levels of crossbred horses. A first cross Andalusian …	Horses were venerated in Norse mythology, a custom brought to Iceland by the country's …	The Andalusian, also known as the Pure Spanish Horse or PRE (Pura Raza Española), is a horse breed …
	DSCMR	Andalusian horses are elegant and strongly built. Members of the breed have heads of medium …	The traditional standard for height of a horse or a pony at maturity is 14.2 hands (58 inches, 147 cm). …	"Hot blooded" breeds included "oriental horses" such as the Akhal-Teke, Arabian horse, ..	The Modern English word goat comes from Old English gāt "she-goat, goat in general", …	All horses move naturally with four basic gaits: the four-beat walk, which averages 6-8 kilometres …
	ACMR	Domesticated horses may face greater mental challenges than wild horses, because ,,,,,,,,	Domestication is also studied by using the genetic material of present-day horses and comparing …	It occurs in dry savannah in Africa south of the Sahara although in nests in somewhat wetter habitats. …	Although machinery has replaced horses in many parts of the world, an estimated 100 …	Turkey, Argentina, the United States, and South Africa are the top producers of mohair.
 Airplane	JFSE (Ours)	An airplane or aeroplane(A (informally plane) is a powered, fixed-wing aircraft that is propelled i …	A fighter aircraft is a military aircraft designed primarily for air-to-air combat against …	Fighter aircraft may be fielded anywhere in a continuum ranging on one end from light, …	Airbus SAS is a division of the multinational Airbus Group SE that manufactures civil aircraft, …	On manned aircraft, cockpit instruments provide information to the pilots, including flight data, engine …
	DSCMR	In the face of lukewarm support from airlines for a 300+ seat Airbus A300, the partners submitted …	When risk is measured by deaths per passenger kilometer, air travel is approximately 10 times safer …	The asteroid belt formed from the primordial solar nebula as a group of planetesimals. Planetesimals …	Fighter aircraft may be fielded anywhere in a continuum ranging on one end from light, …	A fighter aircraft is a military aircraft designed primarily for air-to-air combat against other …
	ACMR	A fighter aircraft is a military aircraft designed primarily for air-to-air combat against other aircraft, as …	Sometimes a short section of deer antler (an "antler sleeve") was used, which prevented the splitting …	Airplanes have complex flight control systems. The main controls allow the pilot to direct…	On 9 October 2006, Christian Streiff, Humbert's successor, resigned due to differences with parent …	To function properly, a balance bicycle must be small enough that the rider can walk the…

(a) Examples of Img2Txt retrieval results on PKU-XMediaNet dataset.

Query	Method	Top 5 Results (Txt2Img)				
Throwing axe: Any of a number of ranged weapons designed to strike with a similar splitting action as their melee counterparts. These are often small in profile and usable with one hand. Axe	JFSE (Ours)					
	DSCMR					
	ACMR					
American football is the third-most popular sport in the United States, behind professional baseball and professional football. The NCAA, the largest collegiate organization, is divided into three Divisions: Division I, Division II and Division III. Division I football is further divided into two subdivisions: American Football	JFSE (Ours)					
	DSCMR					
	ACMR					

(b) Examples of Txt2Img retrieval results on PKU-XMediaNet dataset.

Fig. 6. Examples of the Img2Txt and Txt2Img retrieval results on PKU-XMediaNet dataset by our JFSE approach as well as compared methods DSCMR [38] and ACMR [13]. In these examples the groundtruth class label for each query is presented for instruction. Besides, the true matches and the incorrect retrieval results are marked in green and red rectangles, respectively.

dataset, our JFSE method improves the average MAP score from 0.309 to 0.329 of TANSS; on the Pascal Sentencess dataset, we can see that our JFSE consistently outperforms TANSS with a clear margin; when turning to the large-scale NUS-WIDE and PKU-XMediaNet dataset, our JFSE still achieves the highest MAP scores in both Img2Txt and Txt2Img subtasks. The best performance of our proposed JFSE approach can be attributed to the jointly multimodal feature synthesis and common embedding space learning with an adversary can fully capture the cross-modal correlation. Furthermore, the advanced distribution alignment process effectively transfers the knowledge to unseen classes and captures the correlations across modalities via cycle-consistency constraints.

4.5 Comparison on Generalized Zero-Shot Retrieval

Finally, Table 7 shows the overall comparison results of our JFSE method and the compared methods for generalized zero-shot retrieval on Wikipedia and Pascal Sentences datasets. We can see that the generalized zero-shot retrieval task is more challenging than the above zero-shot retrieval task, since the MAP scores of all the methods in Table 7 are lower than those in Tables 5 and 6. The reason is that the query and database sets in this task contain instances from both seen and unseen classes, making the learned models difficult to well distinguish the seen and unseen classes. We can also observe that the DNN based approaches generally have much better retrieval performance than the shallow methods. Notably, among the DNN based approaches, the MASLN, DADN, and TANSS that are specifically proposed for the zero-shot retrieval task consistently outperform the traditional approaches such as DCCA, ACMR, and DSCMR. Nevertheless, our proposed JFSE method still obtains the highest average MAP scores on the two datasets compared with all the counterparts. It again validates the promising generalization ability of JFSE for facilitating the knowledge transfer among classes.

4.6 Further Analysis

Effect of Multimodal Feature Synthesis. In this experiment, we utilize the t-SNE [70] tool to jointly visualize the distribution of the true image and text features and the synthetic features generated by two coupled cWGANs in common embedding space. For small datasets such as Wikipedia and Pascal Sentences, we utilize all the true image and text features in the training set and generate the same quantity of synthetic features. For the NUS-WIDE dataset, we randomly select the true and synthetic features of 2,000 instances to form 4,000 samples containing both images and texts. It is notable that since the PKU-XMediaNet dataset has 200 classes, which is hardly informative to visualization, we randomly select 10 classes, and 200 samples per class, and finally form 4,000 samples in total for visualization. The visualization results of the mixture of true and synthetic cross-modal features on the four datasets are shown in Fig. 7, where the circle and the cross symbols represent the image and text modalities, respectively. We can observe that the true and synthetic multimodal features form compact clusters in the common embedding space on each dataset, which verifies that our JFSE can not only effectively generate semantically related multimodal features, but also captured the cross-modal correlations of different modalities.

Effect of Adversarial Learning. In order to investigate the effectiveness of adversarial learning applied in multimodal feature synthesis (i.e., in two cWGANs) and in common embedding space learning, we visualize the distributions of the embedding features of the NUS-WIDE test set in the learned common embedding space. It can be observed from Fig. 8b that the embedding features of two different modalities of image and text form several discriminative and compact clusters in the common embedding space. Furthermore, the comparison between Figs. 8a and 8b indicates that feature synthesis network with cWGANs can generate enough multimodal features for robust training to promote learning the representation of different modalities.

Analysis on Class Embeddings. Since the class embeddings are extracted from the pre-trained Word2Vec model and they are fixed as input for modeling the training process, it

TABLE 5
The MAP Scores of Zero-Shot Retrieval on Seen Classes for Our JFSE Approach and Other Compared Methods on All Datasets

Methods	Wikipedia			Pascal Sentencess			NUS-WIDE			PKU-XMediaNet		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA [1] (2010)	0.261	0.267	0.264	0.214	0.183	0.199	0.432	0.438	0.435	0.201	0.246	0.224
CFA [64] (2003)	0.464	0.457	0.461	0.594	0.590	0.592	0.466	0.475	0.471	0.344	0.346	0.345
KCCA [65] (2014)	0.421	0.520	0.471	0.493	0.497	0.495	0.423	0.482	0.453	0.289	0.328	0.309
JRL [66] (2014)	0.522	0.604	0.563	0.636	0.677	0.657	0.480	0.616	0.548	0.330	0.235	0.283
LGCFL [10] (2015)	0.510	0.586	0.548	0.592	0.638	0.615	0.459	0.529	0.494	0.314	0.315	0.315
DCCA [12] (2015)	0.448	0.446	0.447	0.511	0.507	0.509	0.428	0.430	0.429	0.203	0.206	0.205
Deep-SM [67] (2017)	0.674	0.872	0.773	0.728	0.841	0.785	0.680	0.667	0.674	0.731	0.650	0.691
ACMR [13] (2017)	0.674	0.863	0.769	0.726	0.756	0.741	0.604	0.702	0.653	0.704	0.651	0.678
DEMZSL [43] (2017)	0.614	0.855	0.735	0.713	0.811	0.762	0.594	0.630	0.612	0.692	0.651	0.672
MASLN [27] (2018)	0.678	0.872	0.775	0.728	0.759	0.744	0.624	0.712	0.668	0.712	0.657	0.685
DANZCR [25] (2018)	0.672	0.887	0.780	0.737	0.868	0.803	0.727	0.709	0.718	0.739	0.736	0.738
DSCMR [38] (2019)	0.697	0.902	0.800	0.766	0.874	0.820	0.769	0.778	0.772	0.757	0.765	0.761
DADN [28] (2020)	0.677	0.892	0.785	0.748	0.878	0.813	0.732	0.712	0.722	0.753	0.743	0.748
TANSS [22] (2020)	0.686	0.897	0.792	0.761	0.884	0.823	0.737	0.755	0.746	0.751	0.760	0.756
JFSE (Ours)	**0.708**	**0.906**	**0.807**	**0.791**	**0.896**	**0.844**	**0.786**	**0.792**	**0.789**	**0.760**	**0.785**	**0.773**

TABLE 6
The MAP Scores of Zero-Shot Retrieval on Unseen Classes for Our JFSE Approach and Other Compared Methods on All Datasets

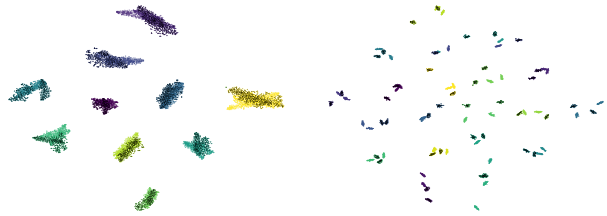
Methods	Wikipedia			Pascal Sentencess			NUS-WIDE			PKU-XMediaNet		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA [1] (2010)	0.238	0.236	0.237	0.207	0.183	0.195	0.400	0.397	0.399	0.031	0.044	0.038
CFA [64] (2003)	0.275	0.285	0.280	0.270	0.294	0.282	0.410	0.355	0.383	0.058	0.071	0.065
KCCA [65] (2014)	0.279	0.288	0.284	0.310	0.321	0.316	0.402	0.413	0.408	0.040	0.057	0.049
JRL [66] (2014)	0.264	0.266	0.265	0.298	0.283	0.291	0.401	0.449	0.425	0.083	0.055	0.069
LGCFL [10] (2015)	0.261	0.258	0.260	0.273	0.258	0.266	0.396	0.422	0.409	0.062	0.064	0.063
DCCA [12] (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
Deep-SM [67] (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 [13] (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
DEMZSL [43] (2017)	0.310	0.239	0.275	0.308	0.318	0.313	0.396	0.466	0.431	0.104	0.122	0.113
MASLN [27] (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 [25] (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
DSCMR [38] (2019)	0.312	0.280	0.296	0.327	0.319	0.323	0.473	0.482	0.478	0.096	0.051	0.074
DADN [28] (2020)	0.305	0.291	0.298	0.359	0.353	0.356	0.423	0.472	0.448	0.112	0.130	0.121
TANSS [22] (2020)	0.314	0.303	0.309	0.362	0.355	0.359	0.446	0.483	0.465	0.110	0.137	0.124
JFSE (Ours)	**0.343**	**0.314**	**0.329**	**0.382**	**0.371**	**0.377**	**0.508**	**0.514**	**0.511**	**0.123**	**0.153**	**0.138**

is necessary to explore the impact of their dimension on the retrieval performance. To this end, we conducted an experiment by changing the dimension of the class embeddings before the model training procedure. As the initial dimension of the class embeddings is 300, here we utilize the dimension reduction tool of principal component analysis (PCA) to reduce the dimension in the range of $[10, 20, 50, 100, 200, 300]$. Fig. 9a shows the standard and zero-shot retrieval results (average MAP scores) of our JFSE method with different dimensions of class embeddings on the Wikipedia dataset. It can be seen that when the dimension is reduced in the range of $[200, 100]$, the retrieval results of both two tasks reduce drastically, while the dimension is in $[50, 10]$, the results of both two tasks slightly increases. It is notable that the best retrieval performances of two retrieval tasks are obtained with the original dimension 300 without dimension reduction. Therefore, it indicates that using the original class embeddings for model training is beneficial and can avoid information loss when reducing the dimension of the class embeddings.

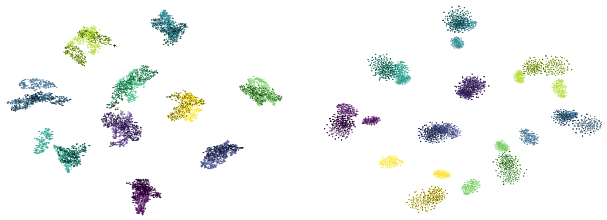
TABLE 7
The MAP Scores of Generalized Zero-Shot Retrieval for Our JFSE Approach and Other Compared Methods on Wikipedia and Pascal Sentences Datasets

Methods	Wikipedia			Pascal Sentencess		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA [1]	0.134	0.118	0.126	0.129	0.141	0.135
KCCA [65]	0.187	0.173	0.180	0.166	0.182	0.174
JRL [66]	0.175	0.169	0.172	0.161	0.178	0.170
DCCA [12]	0.210	0.186	0.198	0.205	0.227	0.216
ACMR [13]	0.251	0.199	0.225	0.250	0.284	0.267
MASLN [27]	0.255	0.209	0.232	0.252	0.287	0.269
DSCMR [38]	0.247	0.215	0.231	0.268	0.255	0.277
DADN [28]	0.251	**0.239**	0.245	0.265	0.296	0.281
TANSS [22]	0.246	0.227	0.236	0.261	0.288	0.274
JFSE (Ours)	**0.283**	0.237	**0.260**	**0.310**	**0.316**	**0.313**

Analysis on Unbalanced Data Split. As aforementioned in Section 4.1, we use the balanced data split for all the dataset

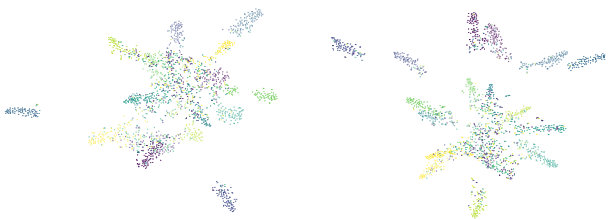


(a) Wikipedia (b) Pascal Sentences



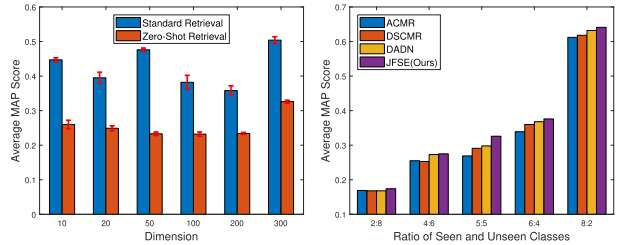
(c) NUS-WIDE (d) PKU-XMediaNet

Fig. 7. The t-SNE results for the chosen data on four datasets. Clusters with different colors are from different classes.



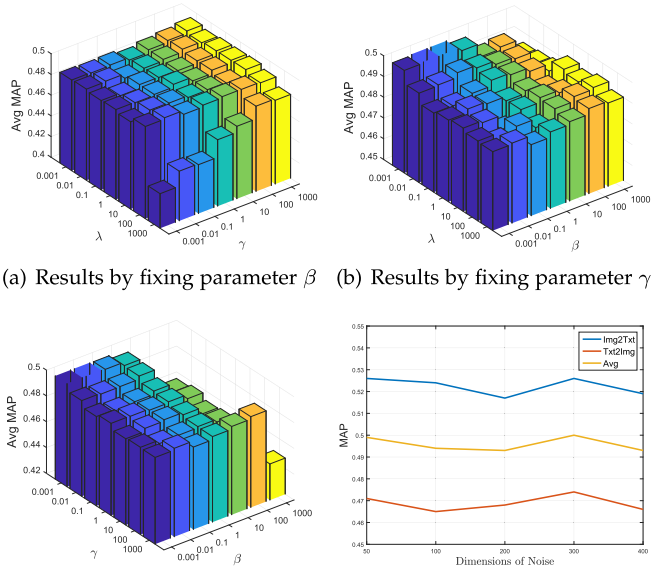
(a) Without cWGANs. (b) Full model of JFSE.

Fig. 8. The t-SNE visualization for with/without adversarial learning of JFSE on NUS-WIDE dataset.



(a) Results of different dimensions of the class embeddings (b) Results of different ratios of the seen and unseen classes

Fig. 9. Experiments of our JFSE method with (a) different dimensions of the class embeddings and (b) different ratios of the seen and unseen classes on Wikipedia dataset.



(a) Results by fixing parameter β (b) Results by fixing parameter γ

(c) Results by fixing parameter λ (d) Results with different noise dimension m

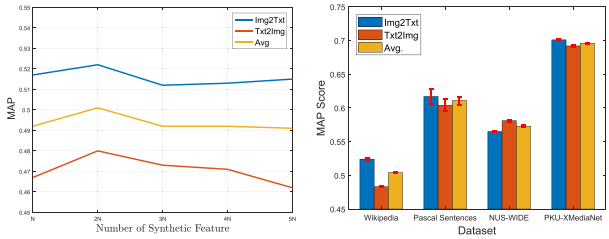
Fig. 10. A sensitivity analysis of the hyper-parameters of our JFSE approach on Wikipedia dataset.

for the zero-shot retrieval task, i.e., the numbers of the seen and unseen classes are identical. Actually, we can vary the ratio (denoted by r) of the seen and unseen classes to form unbalanced data split for the training and test procedures. To further investigate the effect of our proposed JFSE method with unbalanced data split, we take the Wikipedia dataset as a testbed, and set different values of r as $[2:8, 4:6, 5:5, 6:4, 8:2]$ with total 10 classes. Fig. 9b demonstrates the unseen class retrieval results (average MAP scores) of our JFSE method and several compared methods on different ratios r of unbalanced data split. It can be observed that the retrieval results of all the methods continue to increase with larger r, as more data of seen classes can be used for model training and benefit the knowledge transfer to the unseen classes during testing. Nevertheless, the proposed JFSE approach consistently outperforms the four compared approaches for all cases of r, showing its advance on handing the unbalanced data split for model training and improving the performance in generalized zero-shot retrieval scenario.

Analysis on Parameter Sensitivity. In this experiment, we explore the impact of the key parameters on our JFSE approach. Specifically, we still take the Wikipedia dataset as the testbed to assess the effects of the hyper-parameters β, γ, and λ in Equ. (16) for stand retrieval task. Specifically, we set the value of each hyper-parameter in the range of [0.001, 1000], and we keep the value of one parameter unchanged and tune the value of the other parameters each time. Besides, we also investigate the effect of the dimension m of the noise inputs (z_v and z_t) in the two cWGAN modules by setting its range in $[50, 400]$.

The first three subfigures in Fig. 10 shows the sensitivity analysis of the three parameters of our JFSE method, respectively. We can see that the average MAP scores are sensitive to the values of the three parameters and the degree of the sensitivity varies with different parameters. For β, γ, when the value of the hyper-parameter is too large (e.g., nearby 10), the performance is seriously degraded, while the value of the hyper-parameter λ is too small, the performance also dramatically drops. The best average MAP scores are obtained by our JFSE when β, γ, and λ are in the range of $[0.01, 1]$ on the Wikipedia dataset. Therefore, we can effectively tune the optimal values of the parameters via validation of different datasets in practice. Moreover, the dimension m of noise input in each cWGAN will directly affect the quality of the generated features. From the results shown in Fig. 10d, we can see that our JFSE approach achieves the best accuracy at the dimension of noise around the class embeddings (i.e., 300) on both the Wikipedia and Pascal Sentences datasets.

Furthermore, as the number of the synthetic features $\tilde{N}_S$ is flexible in our method, we further vary $\tilde{N}_S =$

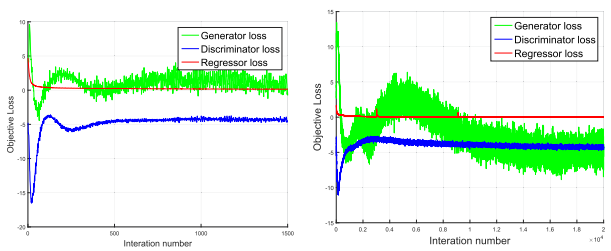


(a) Results of various number of synthetic features (b) MAP scores with standard deviation for standard retrieval

Fig. 11. Experiments of our JFSE method about the impact of (a) the number of synthetic features on Wikipedia dataset and (b) the MAP scores with standard deviation for standard retrieval on all datasets.

$N_S, 2N_S, \ldots, 5N_S$ to investigate its impact on the final performance of JFSE, where N_S is the number of the true image/text features. Fig. 11a shows the retrieval result of various $\tilde{N}_S$ on Wikipedia dataset. We can observe that with a considerable number of synthetic features, the performance of JFSE increases, and when $\tilde{N}_S = 2N_S$, it obtains the best MAP scores. Nevertheless, when adding more synthetic features (e.g., $\tilde{N}_S = 4N_S, 5N_S$), the performance decreases as too many synthetic features may deteriorate the model on modeling the true features. Indeed, the valid number of synthetic features is highly dependent on the statistics of the specific dataset, thereby it is flexible for us to control the quantities of the synthetic features to learn a more effective model. Finally, in Fig. 11b), we also demonstrate the MAP scores of the error bar for standard retrieval on all datasets. We can observe that our JFSE approach is stable and insensitive to the random initialization of the network parameters.

Analysis on Model Convergence and Complexity. In this experiment, we also conduct convergence experiments for our JFSE approach to assess its training efficiency. The curve on the loss values for the different modules of the generators, discriminators, and regressors of JFSE on the Wikipedia and NUS-WIDE datasets are shown in Fig. 12. In the initial training phase, the synthetic features produced by the generator are very different from the true features, so the discriminator can distinguish them well, which shows the unstable vibration of the generator, while the discriminator get a smooth downtrend. With more training iterations, the capability of the generator is improved, so the discrimination task for the discriminator becomes difficult, and the performance of the discriminator declines. Finally, both the generator and the discriminator reach a balanced state, and the network converges. Furthermore, the loss curve of the regressor loss has always shown an overall downward trend, which proves that our JFSE can effectively learn the common embedding space across different modalities with the supervision of class embeddings.



(a) Wikipedia (b) NUS-WIDE

Fig. 12. Convergence experiments of our JFSE method on Wikipedia and NUS-WIDE datasets.

Notably, we have also recorded the time consumption (both training and test processes) of the proposed JFSE methods and other compared methods. In particular, our JFSE approach needs around 80 and 840 seconds to reach model convergence during training on the small Wikipedia dataset and large PKU-XMediaNet dataset, respectively. It is more efficient than the other cross-modal GAN-based approaches DADN and TANSS that generally require over 100 and 1,000 seconds on the two datasets, respectively. The reason is that our JFSE method has less GAN modules and less complicated network structure than the DADN and TANSS methods. Regarding the retrieval efficiency during the test phase, these cross-modal GAN approaches as well as our JFSE method have similar time consumption since the optimal dimension of the common features are identical (i.e., 300 in our experiment). In practice, the retrieval process of these cross-modal GAN approaches are considerably efficient, i.e., less than one second for hundreds of test instances on Wikipedia and less than 100 seconds for thousands of test instances on PKU-XMediaNet, respectively.

5 Conclusion

In this paper, we revisited the adversarial learning in existing cross-modal GAN approaches for cross-modal retrieval. We further proposed a novel method termed Joint Feature Synthesis and Embedding, which is an essential extension and improvement of our previous ACMR method of the conference version. It comprehensively inherits the merit of ACMR and other cross-modal GAN approaches and makes sufficient twists that boost the effectiveness for cross-modal retrieval. We conducted extensive experiments on four widely-used cross-modal datasets and provided insightful analysis of our JFSE method compared with a bundle of existing approaches on both the standard, zero-shot, and generalized zero-shot retrieval scenarios. The experimental results and the ablation study fully demonstrate the superiority of our JFSE method on all retrieval scenarios, where the effectiveness of the key components in JFSE (e.g., multimodal feature synthesis, three distribution alignment schemes, and cycle-consistency constraints) are validated.

For future work, we will further leverage the multimodal feature synthesis scheme to select relevant unlabeled true images or texts for boosting the common embedding space learning. Besides, alternative side information such as knowledge graph will be leveraged to improve the knowledge transfer scheme for zero-shot and generalized zero-shot retrieval scenarios.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Project 61976049 and 61632007; and the Fundamental Research Funds for the Central Universities under Project (ZYGX2019Z015).

References

[1] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 251–260.

[2] K. Wang, R. He, L. Wang, W. Wang, and T. Tan, "Joint feature selection and subspace learning for cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2010–2023, Oct. 2016.

[3] Y. Peng, X. Huang, and Y. Zhao, "An overview of cross-media retrieval: Concepts, methodologies, benchmarks, and challenges," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2372–2385, Sep. 2018.

[4] A. Salvador *et al.*, "Learning cross-modal embeddings for cooking recipes and food images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3068–3076.

[5] B. Zhu, C.-W. Ngo, J. Chen, and Y. Hao, "R2GAN: Cross-modal recipe retrieval with generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11469–11478.

[6] A. Dutta and Z. Akata, "Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5089–5098.

[7] S. Dey, P. Riba, A. Dutta, J. Lladós, and Y. Song, "Doodle to search: Practical zero-shot sketch-based image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2179–2188.

[8] Y. Aytar, L. Castrejón, C. Vondrick, H. Pirsiavash, and A. Torralba, "Cross-modal scene networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2303–2314, Oct. 2018.

[9] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[10] C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Trans. Multimedia*, vol. 17, no. 3, pp. 370–381, Mar. 2015.

[11] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[12] F. Yan and K. Mikolajczyk, "Deep correlation for matching images and text," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3441–3450.

[13] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. 25th ACM Int. Conf. Multimedia*, 2017, pp. 154–162.

[14] Y. Peng, J. Qi, X. Huang, and Y. Yuan, "CCL: Cross-modal correlation learning with multigrained fusion by hierarchical network," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 405–420, Feb. 2018.

[15] Y. Peng and J. Qi, "CM-GANs: Cross-modal generative adversarial networks for common representation learning," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 15, no. 1, pp. 22:1–22:24, 2019.

[16] X. Xu, F. Shen, Y. Yang, H. T. Shen, and X. Li, "Learning discriminative binary codes for large-scale cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2494–2507, May 2017.

[17] D. Wang, X. Gao, X. Wang, and L. He, "Label consistent matrix factorization hashing for large-scale cross-modal similarity search," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 10, pp. 2466–2479, Oct. 2019.

[18] F. Zheng, Y. Tang, and L. Shao, "Hetero-manifold regularisation for cross-modal hashing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 5, pp. 1059–1071, May 2018.

[19] M. Hu, Y. Yang, F. Shen, N. Xie, R. Hong, and H. T. Shen, "Collective reconstructive embeddings for cross-modal hashing," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2770–2784, Jun. 2019.

[20] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[21] J. Gu, J. Cai, S. R. Joty, L. Niu, and G. Wang, "Look, imagine and match: Improving textual-visual cross-modal retrieval with generative models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7181–7189.

[22] X. Xu, H. Lu, J. Song, Y. Yang, H. T. Shen, and X. Li, "Ternary adversarial networks with self-supervision for zero-shot cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 6, pp. 2400–2413, Jun. 2020.

[23] X. Huang and Y. Peng, "Deep cross-media knowledge transfer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8837–8846.

[24] X. Huang, Y. Peng, and M. Yuan, "MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval," *IEEE Trans. Cybern.*, vol. 50, no. 3, pp. 1047–1059, Mar. 2020.

[25] J. Chi and Y. Peng, "Dual adversarial networks for zero-shot cross-media retrieval," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 256–262.

[26] T. Dutta and S. Biswas, "Generalized zero-shot cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5953–5962, Dec. 2019.

[27] X. Xu, J. Song, H. Lu, Y. Yang, F. Shen, and Z. Huang, "Modal-adversarial semantic learning network for extendable cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia Retrieval*, 2018, pp. 46–54.

[28] J. Chi and Y. Peng, "Zero-shot cross-media embedding learning with dual adversarial distribution network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 1173–1187, Apr. 2020.

[29] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 453–465, Mar. 2014.

[30] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning - A comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, 2019.

[31] Y. Peng, J. Qi, and Y. Yuan, "Modality-specific cross-modal similarity measurement with recurrent attention network," *IEEE Trans. Image Process.*, vol. 27, no. 11, pp. 5585–5599, Nov. 2018.

[32] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 7–16.

[33] X. Xu, K. Lin, H. Lu, L. Gao, and H. T. Shen, "Correlated features synthesis and alignment for zero-shot cross-modal retrieval," in *Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2020, pp. 1419–1428.

[34] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 604–611.

[35] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 965–978, Jun. 2014.

[36] T. Baltrusaitis, C. Ahuja, and L. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[37] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks," in *Proc. 25th Int. Joint Conf. Artif. Intell.*, 2016, pp. 3846–3853.

[38] L. Zhen, P. Hu, X. Wang, and D. Peng, "Deep supervised cross-modal retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 394–10 403.

[39] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[40] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid, "Label-embedding for attribute-based classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 819–826.

[41] Y. Xian, B. Schiele, and Z. Akata, "Zero-shot learning - The good, the bad and the ugly," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3077–3086.

[42] R. Felix, B. G. V. Kumar, I. D. Reid, and G. Carneiro, "Multi-modal cycle-consistent generalized zero-shot learning," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 21–37.

[43] L. Zhang, T. Xiang, and S. Gong, "Learning a deep embedding model for zero-shot learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3010–3019.

[44] Y. Xian, T. Lorenz, B. Schiele, and Z. Akata, "Feature generating networks for zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5542–5551.

[45] V. K. Verma, G. Arora, A. Mishra, and P. Rai, "Generalized zero-shot learning via synthesized examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4281–4289.

[46] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019.

[47] H. Zhang, T. Xu, and H. Li, "StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5908–5916.

[48] L. Gao, D. Chen, J. Song, X. Xu, D. Zhang, and H. T. Shen, "Perceptual pyramid adversarial networks for text-to-image synthesis," in *Proc. 33rd AAAI Conf. Artif. Intell.*, 2019, pp. 8312–8319.

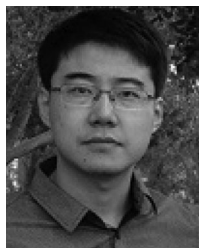
[49] A. Mishra, M. S. K. Reddy, A. Mittal, and H. A. Murthy, "A generative model for zero shot learning using conditional variational autoencoders," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 2188–2196.

[50] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. J. Smola, "A kernel two-sample test," *J. Mach. Learn. Res.*, vol. 13, pp. 723–773, 2012.

[51] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2058–2065.
[52] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," in *Proc. Eur. Conf. Comput. Vis.*, 2018, vol. 11205, pp. 707–723.
[53] Y. Ganin and V. S. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
[54] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2242–2251.
[55] J. Hoffman *et al.*, "CyCADA: Cycle-consistent adversarial domain adaptation," in *Proc. 35th Int. Conf. Mach. Learn.*, 2018, pp. 1994–2003.
[56] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–15.
[57] C. Rashtchian, M. Young, P. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's mechanical turk," in *Proc. Workshop Creating Speech Lang. Data Amazon's Mech. Turk*, 2010, pp. 674–686.
[58] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng, "NUS-WIDE: A real-world web image database from National University of Singapore," in *Proc. ACM Int. Conf. Image Video Retrieval*, 2009, Art. no. 48.
[59] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–14.
[60] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. Conf. Empir. Methods Nat. Lang. Process.*, 2014, pp. 1746–1751.
[61] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.
[62] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013, *arXiv:1301.3781*.
[63] R. Liu, Y. Zhao, L. Zheng, S. Wei, and Y. Yang, "A new evaluation protocol and benchmarking results for extendable cross-media retrieval," 2017, *arXiv:1703.03567*.
[64] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. 11th ACM Int. Conf. Multimedia*, 2003, pp. 604–611.
[65] L. Ballan, T. Uricchio, L. Seidenari, and A. Del Bimbo, "A cross-media model for automatic image annotation," in *Proc. Int. Conf. Multimedia Retrieval*, 2014, pp. 73:73–73:80.
[66] X. Zhai, Y. Peng, and J. Xiao, "Learning cross-media joint representation with sparse and semisupervised regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 6, pp. 965–978, Jun. 2014.
[67] Y. Wei *et al.*, "Cross-modal retrieval with CNN visual features: A new baseline," *IEEE Trans. Cybern.*, vol. 47, no. 2, pp. 449–460, Feb. 2017.
[68] J. Qi and Y. Peng, "Cross-modal bidirectional translation via reinforcement learning," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, 2018, pp. 2630–2636.
[69] D. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, 2004.
[70] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

Xing Xu received the PhD degree from Kyushu University, Japan, in 2015. He is currently with the School of Computer Science and Engineering, University of Electronic of Science and Technology of China, China. He is the recipient of six academic awards, including the IEEE Multimedia Prize Paper 2020, Best Paper Award from ACM Multimedia 2017, and the World's FIRST 10K Best Paper Award-Platinum Award from IEEE ICME 2017. His current research interests mainly focus on multimedia information retrieval and computer vision.

Kaiyi Lin received the BS degree in software engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2018. He is currently working toward the master's degree at the School of Software and Microelectronics, Peking University, Beijing, China. His main research interests include multimedia content analysis and computer vision.



Yang Yang received the PhD degree in computer science from the University of Queensland, Brisbane, QLD, Australia, 2012. He is currently with the University of Electronic Science and Technology of China, Chengdu, China. He was a research fellow with the National University of Singapore, Singapore, from 2012 to 2014. His current research interests include multimedia content analysis, computer vision, and social media analytics.

Alan Hanjalic (Fellow, IEEE) is currently a professor of computer science and head of the Intelligent Systems Department at the Delft University of Technology (TU Delft), The Netherlands. His research interests are in the fields of multimedia information retrieval and recommender systems, in which he (co-)authored more than 250 publications. He is co-recipient of the Best Paper Award at the ACM Conference on Recommender Systems (ACM RecSys) 2012, the ACM International Conference on Multimedia (ACM Multimedia) 2017 and the IEEE International Conference on Multimedia Big Data (IEEE BigMM) 2019. He served as the chair of the Steering Committee of the *IEEE Transactions on Multimedia*, the associate editor-in-chief of the *IEEE MultiMedia Magazine*, and an associate editor of many scientific journals, including the *IEEE Transactions in Multimedia* and the *IEEE Transactions on Affective Computing*. He also served as the general and program (co-)chair in the organizing committees of all major conferences in the multimedia domain, including ACM Multimedia, ACM CIVR/ICMR, and IEEE ICME.

Heng Tao Shen (Senior Member, IEEE) received the BSc (1st class Honours) degree and PhD degree from Department of Computer Science, National University of Singapore, in 2000 and 2004 respectively. He is currently professor and Dean of School of Computer Science and Engineering, Executive Dean of AI Research Institute at University of Electronic Science and Technology of China (UESTC). Then he joined the University of Queensland and became a professor in late 2011. His research interests include multimedia search, computer vision, and artificial intelligence. He has published more than 300 peer-reviewed papers, including more than 100 IEEE/ACM Transactions, and received 8 Best Paper Awards from international conferences and journals, including Best Paper Award from ACM Multimedia 2017 and Best Paper Award from IEEE Transactions on Multimedia 2020. He is/was an Associate Editor of *ACM Transactions of Data Science*, *IEEE Transactions on Image Processing*, and *IEEE Transactions on Knowledge and Data Engineering*. He is an OSA Fellow and ACM Distinguished Member.