Ternary Adversarial Networks With Self-Supervision for Zero-Shot Cross-Modal Retrieval

Xing Xu, Huimin Lu, Jingkuan Song, Yang Yang, *Member, IEEE*, Heng Tao Shen, *Senior Member, IEEE*, and Xuelong Li, *Fellow, IEEE*

***Abstract*—Given a query instance from one modality (e.g., image), cross-modal retrieval aims to find semantically similar instances from another modality (e.g., text). To perform cross-modal retrieval, existing approaches typically learn a common semantic space from a labeled source set and directly produce common representations in the learned space for the instances in a target set. These methods commonly require that the instances of both two sets share the same classes. Intuitively, they may not generalize well on a more practical scenario of *zero-shot cross-modal retrieval*, that is, the instances of the target set contain *unseen* classes that have inconsistent semantics with the seen classes in the source set. Inspired by zero-shot learning, we propose a novel model called ternary adversarial networks with self-supervision (TANSS) in this paper, to overcome the limitation of the existing methods on this challenging task. Our TANSS approach consists of three paralleled subnetworks: 1) two semantic feature learning subnetworks that capture the intrinsic data structures of different modalities and preserve the modality relationships via semantic features in the common semantic space; 2) a self-supervised semantic subnetwork that leverages the word vectors of both seen and unseen labels as guidance to supervise the semantic feature learning and enhances the knowledge transfer to unseen labels; and 3) we also utilize the adversarial learning scheme in our TANSS to maximize the consistency and correlation of the semantic features between different modalities. The three subnetworks are integrated in our TANSS to formulate an end-to-end network architecture which enables efficient iterative parameter optimization. Comprehensive experiments on three cross-modal datasets show the effectiveness of our TANSS approach compared with the state-of-the-art methods for zero-shot cross-modal retrieval.**

***Index Terms*—Adversarial learning, cross-modal retrieval, self-supervision, zero-shot learning (ZSL).**

Manuscript received January 17, 2019; revised April 3, 2019 and May 20, 2019; accepted June 28, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61602089, Grant 61632007, Grant 61772116, and Grant 61572108, in part by the Leading Initiative for Excellent Young Researcher of Ministry of Education, Culture, Sports, Science, and Technology, Japan, under Grant 16809746, in part by the Research Fund of the Telecommunications Advancement Foundation, and in part by the Sichuan Science and Technology Program of China under Grant 2018GZDZX0032. This paper was recommended by Associate Editor L. Rutkowski. *(Xing Xu and Huimin Lu contributed equally to this work.) (Corresponding author: Heng Tao Shen.)*

X. Xu, J. Song, Y. Yang, and H. T. Shen are with the Center for Future Multimedia, University of Electronic Science and Technology of China, Chengdu 610051, China, and also with the School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China (e-mail: xing.xu@uestc.edu.cn; jingkuan.song@gmail.com; dlyyang@gmail.com; shenhengtao@hotmail.com).

H. Lu is with Shanghai Jiaotong University, Shanghai 200240, China, and also with the Department of Mechanical and Control Engineering, Kyushu Institute of Technology, Kitakyushu 804-8550, Japan (e-mail: dr.huimin.lu@ieee.org).

X. Li is with the School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China, and also with the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: xuelong_li@nwpu.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2019.2928180

I. Introduction

With the ever-increasing multimodal data, such as images, text, audios, and videos in our daily life, cross-modal retrieval has drawn increasing interest for searching related data across different modalities. Different from typical unimodal retrieval methods [1]–[3] that process data from one single modality, cross-modal retrieval [4]–[7] can cope with the modalities of input queries and output retrieval results with different format. The major challenge of cross-modal retrieval is the "modality gap" [8], that is, the different modalities have inconsistent data distributions and feature representations, making the similarity measurement among cross-modal data infeasible. The mainstream cross-modal retrieval methods rely on the assumption that there exists a commonly shared semantic space among different modality data. In this space, the inconsistent representations of the heterogeneous data can be transformed to consistent semantic features with identical dimension; meanwhile, the cross-modal correlation is captured. Thereby, the similarities of the heterogeneous data are directly measured via their semantic features in the common semantic space. To learn effective common subspace, traditional methods [6], [7], [9], [10] mainly take linear projections by optimizing the statistical properties underlying the heterogeneous data, while the latest methods [11]–[15] utilize multilayer deep neural networks (DNNs) to model the nonlinearity of cross-modal correlation which have shown superior retrieval accuracy.

Existing cross-modal retrieval approaches commonly consider the so-called *standard* retrieval scenario. Taking the benchmark cross-modal dataset Wikipedia [16] for instance, Fig. 1(a) demonstrates the standard retrieval scenario. Specifically, a *source set* and a *target set* both containing image–text pairs are given, where each instance in both sets

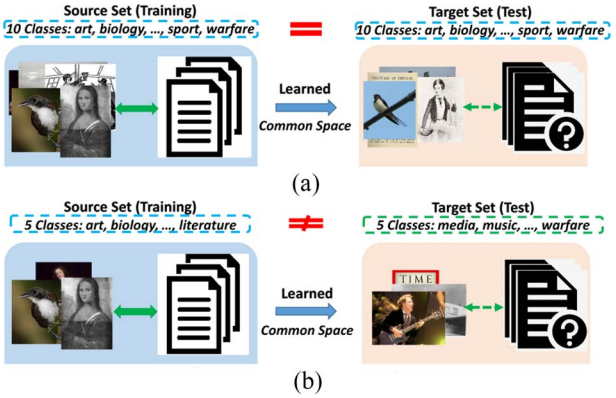


Fig. 1. Illustration of different cross-modal retrieval settings. (a) *Standard retrieval* that has been widely studied in most existing methods. (b) *Zero-shot retrieval* that is newly explored in this paper.

is assigned with a class label from one of the predefined ten classes. During training, the source set data is utilized to learn a common semantic space, which is then used to generate common semantic features for the target set instances to conduct cross-modal retrieval. Nevertheless, the standard retrieval is an ideal scenario that is incompatible with the real situation, as in practice, the target set probably have multimodal of new classes that are absent in the source set in the training process. Therefore, the existing cross-modal retrieval approaches the lack of extensibility to retrieve multimodal data of unseen classes.

In this paper, we agree with the viewpoint in [17] and [18] that a model designed for cross-modal retrieval is ought to be extendable, that is, once it is learned from the source set, it can also accurately measure the cross-modal similarities of the target set instances of the unseen classes. Therefore, we consider a new but more practical scenario of *zero-shot* cross-modal retrieval. This scenario is inspired by the zero-shot learning (ZSL) that has been studied to recognize objects of new classes based on limited training data of seen classes. Note that ZSL aims to enable the knowledge transfer in unimodal data, such as images [19]–[22] or videos [23], [24], while we focus on the extensibility of the cross-modal retrieval task on multimodal data with disjoint classes. As shown in Fig. 1(b), in this new scenario, we denote the five classes in the source set as seen classes, and the other five classes in the target set as unseen classes. It is notable that the seen and the unseen classes are disjoint without overlap. In the scenario, a common subspace is learned based on the source set of seen classes and then directly applied to the target set of unseen classes, assessing the extensibility of the learned subspace. Compared to the standard scenario, here, it is very challenging to handle the heterogeneous data distributions and the incompatible semantics between the seen and unseen classes. In the latter experiment in Section IV, we show the performance decline of the existing approaches on the zero-shot cross-modal retrieval scenario.

To tackle the limitations of the existing approaches on the zero-shot cross-modal retrieval, we propose a novel framework called ternary adversarial networks with self-supervision (TANSS). Unlike the supervised learning that learns feature representations of data from the annotated labels, the self-supervision is a learning framework that mines it from the nature of the data [25], [26]. It is rational for us to employ self-supervised learning to zero-shot cross-modal retrieval problem, that is, learning the common semantic space with some additional supervision, rather than directly relying on the labels. Specifically, as the flowchart demonstrated in Fig. 2, during training, the image–text pairs of seen classes in the source set are used as input data, and employ two semantic feature learning subnetworks (called ImgNet and TxtNet) to capture the intrinsic data structures of different modalities and preserve the modality correlations in the common semantic space. Meanwhile, we deploy a self-supervised semantic subnetwork (called LabNet) that exploits the word vectors of seen and unseen classes as supervision to guide the semantic feature learning in the modality-specific subnetworks ImgNet and TxtNet. The LabNet helps to preserve the semantic relevance and allow the effective knowledge transfer underlying the common semantic space for retrieval on the target data of unseen labels. In addition, the adversarial learning scheme is utilized to introduce two discriminators that, respectively, distinguish the semantic features generated by LabNet and the modality-specific semantic features produced by ImgNet (or TxtNet). Thereby the distribution consistency of the semantic features of different modalities in the learned common semantic space can further be enhanced. For testing, the images and texts in the target set of unseen classes can be directly fed into the ImgNet and TxtNet to obtain their semantic features; hence, the zero-shot retrieval can be performed by measuring the pairwise similarities of these semantic features.

We summarize our main contributions as follows.

1) We propose a novel network architecture called TANSS that consists of two modality-specific semantic learning subnetworks and a self-supervised semantic learning subnetwork for the challenging task of zero-shot cross-modal retrieval. The three subnetworks in TANSS form an end-to-end network structure and mutually boost each other, ensuring to learn more effective common semantic space that is capable to generalize across source set data of seen classes and target set data of unseen classes.
2) The self-supervision scheme is employed to exploit word vectors of seen and unseen labels to supervise the semantic feature learning process with cycle-consistence, which preserves the semantic relevance and enhances the knowledge transfer to the target set of unseen labels.
3) We also integrate adversarial learning to distinguish the modality type of the generated semantic features of different modalities in the common semantic space with the guidance of the word vectors of classes. It helps to alleviate the modality gap and to preserve the semantic consistency across different modalities.

In our previous work presented in [27], we have developed a framework called modal-adversarial semantic learning network (MASLN) that is also for zero-shot cross-modal retrieval. The MASLN approach is formed by two subnetworks. First, the cross-modal reconstruction subnetwork is to minimize the distribution discrepancy between different modalities through conditional autoencoders, where the word vectors of seen

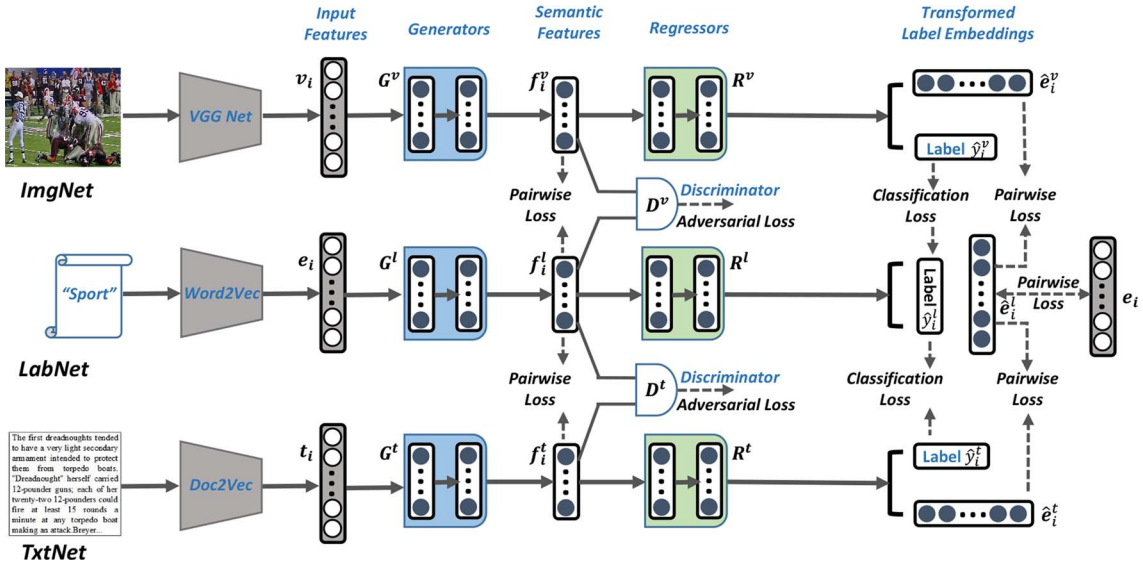


Fig. 2. General flowchart of the proposed TANSS that includes three subnetworks: ImgNet and TxtNet for modality-specific feature learning, and LabNet for self-supervised semantic learning. Each subnetwork consists of a generator for semantic feature generation in the common semantic space and a regressor for label embeddings transformation. Note that the layers with white nodes represent the input features, while the layers with gray nodes are the hidden layers. The pretrained VGG [28], Word2Vec (WV) [29], and Doc2Vec (DV) [30] models are utilized to extract input features of images, labels, and texts, respectively.

classes are utilized as condition for reconstruction. Second, the modal-adversarial semantic learning subnetwork generates indiscriminative semantic representations for modalities through adversarial learning manner. Indeed, there are sufficient improvements in the proposed TANSS compared with our previous MASLN method, which lie in the following aspects.

1) The three-branch adversarial networks in TANSS are more powerful than the previous two-branch autoencoder networks, to capture the semantics of labels in common subspace learning and generalization across seen and unseen labels.
2) Compared with the previous conditional autoencoder, the self-supervised semantic learning subnetwork in TANSS can explicitly exploit the word vectors of labels to enable knowledge transfer across seen and unseen labels.
3) The two discriminators in TANSS are more advanced to reduce the modality gap than the previous single discriminator in MASLN.

Extensive experiments compared with our previous MASLN approach as well as a bundle of the state-of-the-art methods on three cross-modal datasets show the superiority of our TANSS approach.

The remainder of this paper is organized as follows. Section II briefly introduces the literature on cross-modal retrieval and the related research topics, such as transfer learning and ZSL. Section III presents the detailed procedures of our TANSS approach. Section IV discusses the experiments and the ablation study. Finally, we make the conclusion in Section V.

II. Related Work

In this section, we first review the literature of cross-modal retrieval, then we present the recent studies on ZSL and adversarial learning for cross-modal retrieval.

1) Cross-Modal Retrieval: The primary issue on cross-modal retrieval is to seek effective scheme to represent the heterogeneous data by common representations in a shared subspace, where the modality gap can be eliminated. According to the basic models used for subspace learning, the existing methods can be grouped into two categories: 1) the *traditional methods* and 2) the *DNN-based methods*.

The traditional methods typically learn linear projection functions for different modalities to map their features into the common representations in a common subspace, where the cross-modal correlation is expected to be maximized. The pioneering works are canonical correlation analysis (CCA) [16], cross-modal factor analysis (CFA) [9], etc. Besides, more advanced schemes, such as feature selection [6], [7]; graph regularization [10]; and dictionary learning [31] are leveraged to improve retrieval accuracy. Recent works [11]–[15], [32]–[34] have put effort to import DNN to the cross-modal retrieval problem, typical approaches are correspondence autoencoder (Corr-AE) [11], deep Boltzman machine (DBM) [35], deep model version of CCA (DCCA) [13], cross-media multiple deep networks (CMDNs) [12], cross-modal correlation learning (CCL) [15], etc. Comparing to the traditional methods, these DNN-based methods are more effective for common subspace learning as they are capable to capture nonlinearity underlying the heterogeneous data distributions.

As the common representations learned in the above methods are real-valued feature vectors, which is less efficient for distance computation on large-scale multimodal data. To improve the retrieval efficiency, a bundle of hashing-based methods have been developed to learn *binary codes* as common representations in a common subspace, which are generally more efficient for Hamming distance computation and require much less storage for cross-modal retrieval. Similar as the above methods that learn real-valued common representations, the existing hashing-based methods also

include traditional methods [36]–[40] and DNN-based methods [41]–[45] according to the linear or nonlinear architectures used for modeling the cross-modal similarities. However, these cross-modal hashing methods usually employ additional binary constraints in the objective function and utilize iterative quantization or discrete optimization algorithms to learn binary codes.

2) Transfer Learning for Cross-Modal Retrieval: Transfer learning [46], which exploits general knowledge from available source-domain data as guidance to promote the model learning in the target domain, has been widely studied for more than a decade and applied in various scenarios, such as image classification [47] and image translation [48], where the source and target domains contain features are from the same unimodal data. When considering the cross-modal retrieval task that involves multimodal data, several recent works apply transfer learning to multimodal scenarios. Huang and Peng [49] took a large cross-modal dataset as knowledge base to enhance the effect of model training on another small cross-modal dataset, targeting to reduce the discrepancy of different datasets. Huang *et al.* [50] addressed the problem of knowledge transfer from unimodal data to multimodal data, where they use the knowledge from the large-scale unimodal image data to enrich the insufficient cross-modal training data. Different from these works, this paper considers the knowledge transfer between two cross-modal datasets that have disjoint label spaces and, thus, it is more challenging due to the absence of the explicit semantic relatedness between the two cross-modal datasets.

3) Zero-Shot Learning for Cross-Modal Retrieval: Recently, a few works [18], [27], [51] have made initial steps on the zero shot cross-modal retrieval task. They exploit the word vectors of labels from trained NLP models as external knowledge, and jointly perform subspace learning, correlation learning, and ZSL for zero-shot cross-modal retrieval. The difference in these approaches lie in that how to utilize the word vectors. In [27], word vectors are used as condition information for autoencoder reconstruction, while in [18], the word vectors are directly treated as semantic feature space for retrieval. In our TANSS, we introduce a self-supervised learning subnetwork that takes the word vectors as supervised information to guide the semantic space learning procedure, which is more effective than the above methods and further enhances the correlation between different modalities.

4) Adversarial Learning for Cross-Modal Retrieval: Adversarial learning scheme recently has been used in cross-modal learning to seek for more effective cross-modal semantic features in common subspace. Generative adversarial network (GAN) [52], which contains a generator and a discriminator, forms the basic architecture in adversarial learning. On the one hand, the generator can generate fake data for estimating the distribution of observed real data, on the other hand, the discriminator tries to distinguish the observed real data and the fake data produced by the generator. The model parameters in the generator and the discriminator are learned in an adversarial training style. GAN has been extended and applied to various unimodal application areas, for example, image translation [48], image segmentation [53], saliency detection [54], etc. Recently, several studies [8], [14], [33], [50] utilize GAN for cross-modal retrieval task. Specifically, a modality discriminator is usually introduced in these methods to distinguish the semantic features across different modalities; meanwhile, a cross-modal generator is devised to reduce the difference of the cross-modal semantic features for confusing the modality discriminator. Follow a similar idea, the latest works [18], [27] also apply GAN model to zero-shot cross-modal retrieval. They also use a single discriminator to discriminate the modality type of the generated semantic features, where the word vectors of labels are leveraged as external knowledge to help the discriminator for the discrimination process. Our TANSS is also a GAN-based model, differently, it has two discriminators that distinguish the generated semantic features among images, texts, and label embeddings. Therefore, it is advantageous to associate the image and text modalities under the supervision of discriminative label information, resulting in more effective semantic space.

III. Proposed Method

A. Problem Formulation

1) Preliminaries: Assume that there is a source set consists of n_s instances formed by image–text pairs, that is, $\mathcal{O}_s = \{o_i\}_{i=1}^{n_s}$, $o_i = (\mathbf{v}_i, \mathbf{t}_i)$, where $\mathbf{v}_i \in \mathbb{R}^{d_v}$ and $\mathbf{t}_i \in \mathbb{R}^{d_t}$ correspondingly denote the visual and textual feature vector for the ith instance o_i, d_v, and d_t are the dimensionality of the two feature spaces. Besides, for each instance o_i, it has a label y_i, where $y_i \in \mathcal{Y}_s$, that is, the seen label set of c_s unique labels for $\mathcal{O}_s$. There is also a target set of n_t instances, $\mathcal{O}_t = \{o_j\}_{j=1}^{n_t}$, where each instance $o_j = (\mathbf{v}_j, \mathbf{t}_j)$ is represented by the same features as those in the source set. The primary difference is that each instance o_j in $\mathcal{O}_t$ has a label $y_q \in \mathcal{Y}_t$ of the unseen class set (containing c_t unseen labels), where $\mathcal{Y}_t$ and $\mathcal{Y}_s$ share no overlap, that is, $\mathcal{Y}_t \bigcap \mathcal{Y}_s = 0$.

2) Zero-Shot Cross-Modal Retrieval: We intend to learn a common semantic subspace from the instances in the source set $\mathcal{O}_s$ of seen labels during training, and then in testing stage, we can generate *semantic features* for the instances in the target set $\mathcal{O}_t$ of unseen labels in the learned common subspace to perform zero-shot cross-modal retrieval. Motivated by several existing ZSL approaches [19]–[21], [55] that utilize embedding features (i.e., attributes or word vectors) of labels as side information for knowledge transfer from the source set to the target set. In this paper, we consider to leverage the word vectors of classes (called *label embeddings*) as embeddings features for each label in $\mathcal{Y}_s$ as $\mathcal{E}_s = \{\mathbf{e}_s\}_{s=1}^{c_s}$ and $\mathcal{Y}_t$ as $\mathcal{E}_t = \{\mathbf{e}_t\}_{t=1}^{c_t}$, where the items in both $\mathcal{E}_s$ and $\mathcal{E}_t$ are m-dimensional vectors. In particular, we extract the label embeddings for each class label via pretrained natural language processing models on large-scale linguistic corpus (i.e., Wikipedia [30]), which need much lower labor cost than manually annotated attributes. As the instances in the same class share the label embeddings, each instance o_i in $\mathcal{O}_s$ has the same label embeddings $e_i \in \mathcal{E}_s$ according to y_i.

B. Architecture of TANSS

As the flowchart demonstrated in Fig. 2, the proposed TANSS includes three paralleled subnetworks: LabNet that is a self-supervised semantic reconstruction network; ImgNet and TxtNet that are two modality-specific feature learning networks, respectively, for image and text modalities. A generator (i.e., $G^*(\cdot)$, $* = v, t, l$) and a regressor ($R^*(\cdot)$, $* = v, t, l$) are equipped on each subnetwork, where the generator first generates semantic features in the common semantic space (i.e., $\mathbf{f}^*$, $* = v, t, l$) given the original representations of specific data (i.e., images, texts, and labels); and the regressor transforms the generated semantic features to label embeddings. Besides, a discriminator is, respectively, assigned for each of ImgNet and TxtNet to distinguish the semantic features generated between (ImgNet and LabNet) and (TxtNet and LabNet). The semantic features learned from the label embeddings in LabNet are leveraged to supervise the modality-specific feature learning in both ImgNet and TxtNet. Note that each generator, regressor, or discriminator is built with several fully connected layers. We introduce the detailed procedure of each subnetwork in the following sections.

C. Self-Supervised Semantic Learning in LabNet

As the label embeddings are commonly utilized as a conduciveness to enhance the semantic relevance of different modalities; here, the LabNet learns semantic features from the label embeddings to further associate modality-specific features of ImgNet and TxtNet in a common semantic space where the associations between different modalities are enriched. Given the label embeddings $\mathbf{e}_i$ for a label y_i, the LabNet generates its semantic features $\mathbf{f}_i^l$ via generator G_l as $\mathbf{f}_i^l = G_l(\mathbf{e}_i)$. It is expected that the similarity relationships between the label embeddings and their semantic features in the common subspace are well preserved. In particular, we use the inner product to quantify the similarity of pairwise semantic features. Given two semantic feature vectors $\mathbf{f}_i^l$ and $\mathbf{f}_j^l$ of two labels y_i and y_j, the pairwise similarity S_{ij} can be formulated as

$$s_{ij}\left(\mathbf{f}_i^l, \mathbf{f}_j^l\right) = \begin{cases} \delta(i,j), & y_i = y_j \\ 1 - \delta(i,j), & y_i \neq y_j \end{cases} \tag{1}$$

where $\delta(i,j) = 1 + \exp(-(1/2)(\mathbf{f}_i^l \odot \mathbf{f}_j^l))$, and $\odot$ denotes the inner product operation. Equation (1) indicates that two vectors with a larger inner product are ought to be similar more probably. Therefore, the similarity between semantic features in the common subspace can be thereby cast to computing the inner product of the embeddings of two labels. Formally, the similarity loss for all pairwise semantic features is derived as a negative log-likelihood function

$$\mathcal{J}_1^l = -\sum_{i,j=1}^{n_s}\left(s_{ij}\Delta_{ij}^l - \log\left(1 + \exp\left(\Delta_{ij}^l\right)\right)\right) \tag{2}$$

where $\Delta_{ij}^l = (1/2)(\mathbf{f}_{*i}^l)^\top(\mathbf{f}_{*j}^l)$. To comprehensively understand the effect of the above loss term, we denote Δ_{ij}^l as x to simplify the notation, then the similarity loss value $f(x)$ of each pair of semantic features in (2) can be further derived according to the similarity indicator s_{ij} as

$$f(x) = \begin{cases} f_1(x) = \log(1 + \exp(x)) - x, & s_{ij} = 1 \\ f_2(x) = \log(1 + \exp(x)) & s_{ij} = 0. \end{cases} \tag{3}$$

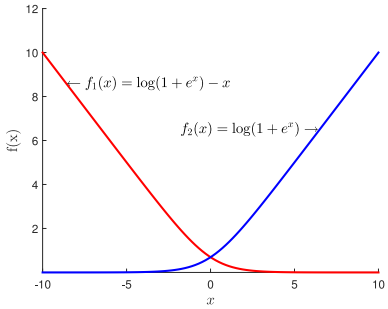


Fig. 3. Curves of two derivatives of $f(x)$, that is, $f_1(x)$ when $s_{ij} = 1$ and $f_2(x)$ when $s_{ij} = 0$.

Fig. 3 plots the two derivatives of $f(x)$ in (3) according to the value of s_{ij}. On the one hand, the function $f_1(x)$ encourages two semantic features with the same label (i.e., $s_{ij} = 1$) to be similar (i.e., $x > 0$) with small loss value, while penalizing the dissimilar ones (i.e., $x < 0$) with large loss value. On the other hand, the function $f_2(x)$ allocates two semantic features of different labels (i.e., $s_{ij} = 0$) with small loss value when they are dissimilar (i.e., $x < 0$), while giving large penalty when they are similar (i.e., $x > 0$). Therefore, the loss term in (2) can effectively measure the similarity of pairwise semantic features.

In LabNet, the regressor $R^l(\cdot)$ further transforms the semantic features $\mathbf{f}_i^l$ back to the reconstructed label embeddings $\hat{\mathbf{e}}_i^l$, which are expected to be similar to the original label embeddings $\mathbf{e}_i^l$ as a *cycle-consistence*. In particular, we argue to ensure the cycle-consistence for both seen labels in the source set and the unseen labels in the target set. This constraint is intrinsically different from the previous self-supervised learning mechanism used in [45] that limitedly considers the seen labels. Actually, considering the cycle-consistence in the regressor for unseen labels allows the knowledge transfer to unseen labels and alleviates the inconsistent semantics between source and target set. Therefore, it is beneficial for zero-shot cross-modal retrieval task. In particular, we formulate the cycle-consistence loss for both seen and unseen labels as

$$\mathcal{J}_2^l = \sum_{i=1}^{c_s}\left\|\hat{\mathbf{e}}_i^l - \mathbf{e}_i^l\right\|_2^2 + \sum_{j=1}^{c_t}\left\|\hat{\mathbf{e}}_j^l - \mathbf{e}_j^l\right\|_2^2 \tag{4}$$

where $\|\cdot\|_2^2$ is the L_2-norm that ensures the robust elementwise reconstructions of vectors, and $\{\mathbf{e}_i^l\}_{i=1}^{c_s}$ and $\{\mathbf{e}_j^l\}_{j=1}^{c_t}$ denote the label embeddings for seen labels and unseen labels, respectively. Indeed, other optional vector norm, for example, the L_1-norm ($\|\cdot\|_1$), can also be used in (4). The latter experiment, in Section IV, demonstrates the robustness of the L_2-norm used here. In addition, the similar semantic features should have the same label to preserve the discrimination of labels in the common subspace, which is cast to a classification loss as

$$\mathcal{J}_3^l = \sum_{i=1}^{c_s}\left\|\hat{y}_i^l - y_i\right\|_2^2 \tag{5}$$

where y_i represents the groundtruth label, and $\hat{y}_i^l$ is the predicted label from LabNet.

Accordingly, the final objective function for LabNet is the combination of the pairwise similarity loss, the cycle-consistence loss, and the classification loss, which is derived as

$$\mathcal{L}^l = \alpha^l \mathcal{J}_1^l + \beta^l \mathcal{J}_2^l + \gamma^l \mathcal{J}_3^l \tag{6}$$

where α^l, β^l, and γ^l are three hyperparameters that balance the three loss terms.

D. Modality-Specific Feature Learning in ImgNet and TxtNet

To keep the semantic relevance of instances across image and text modalities, we take the LabNet as the guidance in both ImgNet and TxtNet as the explicity supervision for their modality-specific feature learning process, where the cues of supervision are from the semantic features and the label embeddings in the LabNet. In other words, the LabNet acts as the self-supervision information when used on both ImgNet and TxtNet. For the supervision of the semantic features on ImgNet, given an image sample $\mathbf{v}_i$ with label y_i and label embeddings $\mathbf{e}_i$, the semantic features $\mathbf{f}_i^v = G^v(\mathbf{v}_i)$ produced by generator $G^v(\cdot)$ are associated with the relevant semantic features $\mathbf{f}_i^l = G^l(\mathbf{e}_i)$. The association is measured by the pairwise similarity of semantic features. Likewise, the supervision of the semantic features can also be applied to the TxtNet. The self-supervision process on the two subnetworks can be formulated as the pairwise similarity loss involving in the semantic features generated by the three generators $G^v(\cdot)$, $G^t(\cdot)$, and $G^l(\cdot)$, as

$$\begin{aligned}\mathcal{J}_1^{v,t} = &\sum_{i,j=1}^{n_s}\left(s_{ij}\Delta_{ij}^v - \log\left(1+\exp\left(\Delta_{ij}^v\right)\right)\right)\\ &+\sum_{i,j=1}^{n_s}\left(s_{ij}\Delta_{ij}^t - \log\left(1+\exp\left(\Delta_{ij}^t\right)\right)\right)\end{aligned} \tag{7}$$

where $\Delta_{ij}^v = (1/2)(\mathbf{f}_i^v)^\top(\mathbf{f}_j^l)$ and $\Delta_{ij}^t = (1/2)(\mathbf{f}_i^t)^\top(\mathbf{f}_j^l)$ account for the similarity of semantic features between images (or text) and the label embeddings.

As the label embeddings intrinsically incorporate the semantic information, the supervision of label embeddings also helps to strengthen the semantic relevance of different modalities. It is expected that on ImgNet the transformed label embeddings $\hat{\mathbf{e}}_i^v$ from regressors $R^v(\cdot)$ are associated to the relevant reconstructed label embeddings $\hat{\mathbf{e}}_i^l$ from $R_l(\cdot)$. Similarly, on the TxtNet, the LabNet can be utilized to supervise the embedding learning process similarly. Like (7), the supervision of the label embeddings on ImgNet and TxtNet can be formulated as the pairwise similarity loss

$$\begin{aligned}\mathcal{J}_2^{v,t} = &\sum_{i,j=1}^{n_s}\left(s_{ij}\Phi_{ij}^v - \log\left(1+\exp\left(\Delta_{ij}^v\right)\right)\right)\\ &+\sum_{i,j=1}^{n_s}\left(s_{ij}\Phi_{ij}^t - \log\left(1+\exp\left(\Delta_{ij}^t\right)\right)\right)\end{aligned} \tag{8}$$

where $\Phi_{ij}^v = (1/2)(\mathbf{e}_i^v)^\top(\mathbf{e}_j^l)$ and $\Phi_{ij}^t = (1/2)(\mathbf{e}_i^t)^\top(\mathbf{e}_j^l)$ represent the similarity of semantic features between images (or texts) and the label embeddings.

Furthermore, on both ImgNet and TxtNet, we also intent to capture the discrimination of labels in the generated semantic features, that is, similar semantic features should have the same label. Similar as (5), the classification loss for both ImgNet and TxtNet can be jointly formulated as

$$\mathcal{J}_3^{v,t} = \sum_{i=1}^{n_s}\left\|\hat{y}_i^v - y_i\right\|_2^2 + \sum_{i=1}^{n_s}\left\|\hat{y}_i^t - y_i\right\|_2^2 \tag{9}$$

where y_i denotes the groundtruth label, and $\hat{y}_i^v$ and $\hat{y}_i^t$ are the predicted label from ImgNet and TxtNet, respectively. Accordingly, the modality-specific feature learning on both ImgNet and TxtNet jointly combines the two pairwise similarity losses and the classification loss, which is formulated as

$$\mathcal{L}^{v,t} = \alpha^{v,t}\mathcal{J}_1^{v,t} + \beta^{v,t}\mathcal{J}_2^{v,t} + \gamma^{v,t}\mathcal{J}_3^{v,t} \tag{10}$$

where $\alpha^{v,t}$, $\beta^{v,t}$, and $\gamma^{v,t}$ are hyperparameters that balance the three loss terms in the two subnetworks.

E. Adversarial Learning in TANSS

To alleviate the heterogeneous data distribution of different modalities and obtain modality-invariant semantic features in the common subspace, in our TANSS approach, we further adopt the adversarial learning scheme. Specifically, for the two modalities, we introduce two discriminators (one of each modality) to distinguish the difference between their data distributions. It is worth noting that the difference of distributions is measured through the LabNet, rather than directly comparing the two modalities. The image discriminator $D^v(\cdot)$ takes the semantic features produced by the generator $G^v(\cdot)$ of ImgNet and $G^l(\cdot)$ of LabNet as inputs, while the text discriminator $D^t(\cdot)$ takes the semantic features generated by $G^t(\cdot)$ of TxtNet and $G^l(\cdot)$ of LabNet as inputs. The outputs of both discriminators are binary decision scores, either "0" or "1," denoting the semantic features of each instance belong to whether the image modality or the text modality.

Let $\{p_i^v\}_{i=1}^{n_s}$ and $\{p_i^l\}_{i=1}^{n_s}$, where $p_i^l \in \{0,1\}$ represents the modality labels of the semantic features $\mathbf{f}_i^v$ and $\mathbf{f}_i^l$ of ImgNet and LabNet for discriminator $D^v(\cdot)$, respectively; $\{q_i^t\}_{i=1}^{n_s}$, $q_i^t \in \{0,1\}$ and $\{q_i^l\}_{i=1}^{n_s}$, $q_i^l \in \{0,1\}$ are the modality labels assigned to the semantic features $\mathbf{f}_i^v$ and $\mathbf{f}_i^t$ of TxtNet and LabNet for discriminator $D^t(\cdot)$. The two discriminators act as two adversaries to train binary classifier given the semantic features and modality labels. As such, the objective function of adversarial learning in our TANSS can be derived as follows:

$$\begin{aligned}\mathcal{L}^a = &\sum_{i=1}^{n_s}\left(\left\|D^v\left(\mathbf{f}_i^v\right) - p_i^v\right\|_2^2 + \left\|D^v\left(\mathbf{f}_i^l\right) - p_i^l\right\|_2^2\right)\\ &+\sum_{i=1}^{n_s}\left(\left\|D^t\left(\mathbf{f}_i^t\right) - q_i^t\right\|_2^2 + \left\|D^t\left(\mathbf{f}_i^l\right) - q_i^l\right\|_2^2\right).\end{aligned} \tag{11}$$

F. Optimization

As the aforementioned loss functions in each subnetwork of TANSS are calculated in different positions; here, we first denote the parameters in LabNet as θ^l, and the parameters in ImgNet and TxtNet as θ^v and θ^t, respectively. Since, in the three subnetworks, the coefficients $\{\alpha^l, \alpha^{v,t}\}$, $\{\beta^l, \beta^{v,t}\}$, and

Algorithm 1 Iterative Learning Procedure of Our TANSS

Input: Source set instances $\{(\mathbf{v}_i, \mathbf{t}_i, y_i)\}_{i=1}^{n_s}$, the label embeddings $\{\mathbf{e}_i\}_{i=1}^{c_s}$, $\{\mathbf{e}_i\}_{i=1}^{c_t}$ of seen and unseen labels.

Output: Model parameters θ^l, θ^v, θ^t, θ^a.

1: Initialize hyper-parameters α, β, γ, learning rate μ.
2: **repeat**
3: Update θ^l by $\theta^l \leftarrow \theta^l - \mu \nabla_{\theta^l} \frac{1}{c_s}(\mathcal{L}^l + \mathcal{L}^{v,t} - \mathcal{L}^a)$.
4: Update $\theta^{v,t}$ by $\theta^* \leftarrow \theta^* - \mu \nabla_{\theta^*} \frac{1}{n_s}(\mathcal{L}^l + \mathcal{L}^{v,t} - \mathcal{L}^a)$, $* = v, t$.
5: Update θ^a by $\theta^a \leftarrow \theta^a - \mu \nabla_{\theta^a} \frac{1}{n_s}(\mathcal{L}^l + \mathcal{L}^{v,t} - \mathcal{L}^a)$.
6: **until** Objective function of Equ. (12) converges or reaches maximum iterations.
7: Obtain generators $G^v(\cdot)$, $G^t(\cdot)$, $G^l(\cdot)$ that generate common features for images, texts and label embeddings, respectively.

$\{\gamma^l, \gamma^{v,t}\}$ correspondingly play the same roles on the generators, the regressors, and the classification parts, we simplify the notations as α, β, and γ, respectively. Besides, we also denote the parameters of the two discriminators in adversarial learning part as θ^a. With the above definitions, the semantic feature generators and the modality discriminators in TANSS can beat each other with a minimax game. Finally, the overall objective function of TANSS that consists of the losses in the three subnetworks along with the adversarial learning part can be formulated as

$$\min_{\theta^l, \theta^v, \theta^t} \max_{\theta^a} \mathcal{L}^l + \mathcal{L}^{v,t} - \mathcal{L}^a. \tag{12}$$

Our goal is to seek for the optimal values of $\hat{\theta}^l$, $\hat{\theta}^v$, $\hat{\theta}^t$, and $\hat{\theta}^a$ for optimizing (12) via an adversarial training manner

$$\left(\hat{\theta}^l, \hat{\theta}^v, \hat{\theta}^t\right) = \arg\min \mathcal{L}^l + \mathcal{L}^{v,t} - \mathcal{L}^a \tag{13}$$

$$\hat{\theta}^a = \arg\max \mathcal{L}^l + \mathcal{L}^{v,t} - \mathcal{L}^a. \tag{14}$$

Here, we optimize the objective function in (13) and (14) through iterative optimization similar as [14], [27], and [45]. Specifically, we first optimize the $\mathcal{L}^l$ for θ^l by exploring label embeddings. Then, we optimize $\mathcal{L}^{v,t}$ for $\theta^{v,t}$ by fixing θ^l. During this process, self-supervised learning manner is adopted for learning the semantic features of two modalities. Finally, we optimize $\mathcal{L}^a$ over θ^a by fixing θ^l and $\theta^{v,t}$. The standard optimization algorithm of stochastic gradient descent (SGD) can be naturally used to learn the model parameters. The entire learning procedure in our TANSS is demonstrated in Algorithm 1.

G. Retrieval on Unseen Cross-Modal Data

Once we have obtained the model parameters of the generators $G^v(\cdot)$ and $G^t(\cdot)$ in ImgNet and TxtNet, in the testing stage, we can utilize them to generate semantic features in the learned common subspace for unseen data for zero-shot cross-modal retrieval. Specifically, for the image samples $\{\mathbf{v}_j\}_{j=1}^{n_t}$ and text samples $\{\mathbf{t}_j\}_{j=1}^{n_t}$ in the target set $\mathcal{O}_t$, their semantic features can be generated as $\{\mathbf{f}_j^v\}_{j=1}^{n_t}$ with $\mathbf{f}_j^v = G^v(\mathbf{v}_j)$, and $\{\mathbf{f}_j^t\}_{j=1}^{n_t}$ with $\mathbf{f}_j^t = G^t(\mathbf{t}_j)$, respectively. Finally, the cross-modal retrieval can be accomplished by measuring the cosine similarity score of the pairwise cross-modal samples in $\{\mathbf{f}_j^v\}_{j=1}^{n_t}$ and $\{\mathbf{f}_j^t\}_{j=1}^{n_t}$.

Except for the zero-shot cross-modal retrieval, our TANSS can also classify the unseen labels for image and text samples in $\mathcal{O}_t$, or find the related images and texts at the same time given a specific label. Similar as the evaluation in existing ZSL work, the semantic features of unseen labels can be obtained by the generator $G^l(\cdot)$ of LabNet. Given the label embeddings $\{\mathbf{e}_c^l\}_{c=1}^{c_t}$ of unseen labels, the semantic features can be computed as $\{\mathbf{f}_c^l\}_{c=1}^{c_t}$ with $\mathbf{f}_c^l = G^l(\mathbf{e}_c^l)$. Then, the classification or retrieval of image (text) samples can be cast to a nearest neighbor search problem by computing the distances between semantic features of image (text) samples and unseen labels. In the experiment, we show the capability of our TANSS on the zero-shot cross-modal retrieval task, as well the label classification and label retrieval tasks.

IV. Experiments

A. Experimental Setup

Datasets and Features: Three benchmark datasets, that is, Wikipedia [16], Pascal Sentence [56], and NUS-WIDE [57] are adopted in our experiments, which have been widely used in the recent studies [12], [14], [17], [18], [50]. Table I provides the basic information of the three datasets and, here, we briefly describe the general information of them are as follows.

1) Wikipedia has 2866 instances of image–text pairs that are crawled from the Wikipedia website, and each instance has one label from ten categories.
2) Pascal Sentence consists of 1000 images with each image belonging to one of the predefined 20 categories. Each image is also described with five textual sentences.
3) NUS-WIDE contains around 270 000 images with associated tags, and the images belong to at least one of the 81 categories.

Due to the coarse annotations in the original dataset, we adopt two different subsets that contain images with more accurate annotations.

1) Similar as the settings in [18] and [31], we first choose the images along with their tags that exclusively come from the largest ten categories, ensuring that each image–text pair only has one class label. Finally, there are 71 602 pairs remaining in the subset *NUS-WIDE-S1*.
2) According to the settings in several cross-modal hashing methods [39], [40], [44], we select the images that have one or more tags belonging to the largest ten categories.

Finally, there are 186 577 image–text pairs with multilabels in the subset *NUS-WIDE-S2*.

Recently, the convolutional neural network (CNN) has shown its advance on extracting more effective image representation and has been applied in various computer vision tasks. For this reason, in our experiments, for each image on all datasets, we make use of the popular CNN model, that is, VGGNet [28] to extract the 4096-D CNN features from its fc7 layer's activations. Similar as the previous work [12], we use DV [30] model pretrained on Google News to extract 300-D features for representing texts on the Wikipedia and Pascal Sentence datasets. For the tags in the NUS-WIDE dataset, we

TABLE I
GENERAL STATISTICS OF ALL DATASETS AND THE DETAILED SETTINGS ON STANDARD AND ZERO-SHOT RETRIEVAL TASKS. HERE "I," "T," "L," "S," AND "U" DENOTE IMAGES, TEXTS, LABELS, SEEN, AND UNSEEN, RESPECTIVELY. "UNC" IS SHORT FOR "UNCERTAIN" AND "-" MEANS NO EVALUATION

Datasets	Basic information					Standard retrieval			Zero-shot retrieval		
	Pairs	Labels	Feature (I)	Feature (T)	Feature (L)	Training	Test	Label	Training	Test	Labels (S/U)
Wikipedia	2,866	10	VGG	DV	WV	2,173	693	10	UNC	UNC	5/5
Pascal Sentence	1,000	20	VGG	DV	WV	800	200	20	UNC	UNC	10/10
NUS-WIDE-S1	71,602	10	VGG	WV	WV	42,941	28,661	10	UNC	UNC	5/5
NUS-WIDE-S2	186,577	10	VGG	WV	WV	184,577	2,000	10	-	-	-

utilize the WV [29] model also pretrained on Google News to extract 300-D feature vector for each tag and then average overall all tags for the text representation. Similarly, the label embeddings for labels on all datasets are also 300-D vectors extracted by the WV model.

Retrieval Tasks: We conduct two different cross-modal retrieval tasks: 1) the standard retrieval and 2) the zero-shot retrieval on all datasets similar as in [17] and [27]. Each task has two scenarios: 1) Img2Txt and 2) Txt2Img that take one modality data as query, that is, images (text), to retrieve related items in the other modality, that is, texts (images). The detailed settings on the two tasks are also list in Table I, and the instructions are depicted as follows.

1) *Standard Retrieval:* We adopt the default data (training/test) split originally provided by each dataset for the standard retrieval task. That is, for each dataset, the image–text pairs in the training data are considered as the source set and those in the test data are treated as the target set. It is worth mentioning that both the two sets include all the classes in the dataset.
2) *Zero-Shot Retrieval:* Each dataset is also partitioned into a source set for training and a target set for testing. Similar as the data split protocol used in [27], to make disjoint classes, we choose the image–text pairs from one half classes as the source set and the pairs from the other half classes as the target set for each dataset. We randomly shuffle the classes in the source and target set, to eliminate the affect of different class combinations in the two sets. Therefore, the number of training/test instances are uncertain since it depends on the shuffle process (as "UNC" shown in Table I). Note that the NUS-WIDE-S2 dataset cannot form disjoint classes between the source and target set as each image–text pair in it has multiple labels, therefore, it is not used for zero-shot retrieval task.

Details of Network: In the self-supervised semantic learning subnetwork, we build the generator and regressor of LabNet with fully connected layers with dimensions [4096, 1024, K], where K denotes the nodes of semantic feature layer (in common subspace). For the modality-specific feature learning subnetworks, we build the generators of ImgNet and TxtNet with three fully connected layers of [4096, 4096, K] to generate semantic features with each layer following a tanh layer except the last one. The regressors of ImgNet and TxtNet have the reverse structure as the corresponding generators. In addition, we build the two discriminators for adversarial learning using three fully connected layers with dimensions [4096, 2048, 1], which project the generated semantic features into a single value to distinguish real or fake.

While training our TANSS, we update the parameters of the discriminators less often than the parameters in the subnetworks to make the semantic feature learning more stable. This strategy is commonly adopted in recent methods [14], [58]–[60]. We take the Adam optimizer to jointly optimize the parameters in the three subnetworks, where the mini-bath size is set as 64 and the initial learning rate as $\mu = 0.0001$. We tune the hyperparameters α, β, and γ, and the dimension of common semantic space as K on all datasets, where the sensitivity of these hyperparameters is presented in the latter ablation study (in Fig. 9).

Compared Methods and Evaluation Metric: In our experiment, we compare the proposed TANSS approach to ten state-of-the-art methods that are originally proposed for standard retrieval task or zero-shot retrieval task. Among the methods for standard retrieval task, CCA [16], CCA-3V [5], and JFSSL [7] are the shallow methods, while Corr-AE [11], DCCA [13], Deep-SM [61], CMDH [12], and ACMR [14] are the DNN-based methods. Besides, MASLN [27] and DANZCR [18] are the two latest approaches designed for zero-shot retrieval task. All the experiments are conducted under the same configurations for fair comparison. In particular, for each instance (either an image or a text) in the target set, we generate its common representation by the learned linear projection functions or deep models by all the methods. Then, the standard and zero-shot retrieval can be accomplished by computing the cosine similarities of pairwise instances. We conduct the evaluation for the compared methods with source codes or our implementations 20 times on all datasets, and average the results of different runs to obtain the final performance.

The mean average precision (MAP) score that has been commonly adopted in [6], [7], and [14]–[16] is adopted as the primary evaluation metric. Specifically, we first compute the average precision of all queries in the target set, and then take their mean value as the final MAP score. Note that when calculating the MAP score on the NUS-WIDE-S2 dataset that has multilabel image–text pairs, we follow the previous methods [39], [44], [45] and define the retrieved instances that have at least one label with the query as the groundtruth relevant instances. Besides, we also consider another metric called cumulative matching characteristics (CMCs) curve [17], which elaborately measures the rank of the first match in the top-r returned candidate for each input query. The two metrics complement each other and provide different aspects for retrieval performance assessment.

TABLE II
STANDARD RETRIEVAL COMPARISON IN TERMS OF MAP WITH STANDARD DEVIATION ON THREE SINGLE-LABEL DATASETS

Methods	Wikipedia			Pascal Sentences			NUS-WIDE-S1		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA [16]	0.264	0.272	0.268	0.234	0.205	0.220	0.421	0.433	0.427
CCA-3V [5]	0.423	0.386	0.405	0.362	0.315	0.339	0.446	0.495	0.471
JFSSL [7]	0.514	0.537	0.526	0.489	0.516	0.503	0.491	0.527	0.505
DCCA [13]	0.566 ± 0.018	0.552 ± 0.021	0.559	0.578 ± 0.007	0.602 ± 0.016	0.590	0.533 ± 0.023	0.547 ± 0.015	0.540
Corr-AE [11]	0.543 ± 0.006	0.558 ± 0.006	0.551	0.562 ± 0.004	0.589 ± 0.005	0.576	0.504 ± 0.010	0.551 ± 0.003	0.528
Deep-SM [61]	0.644 ± 0.023	0.669 ± 0.021	0.657	0.645 ± 0.028	0.667 ± 0.020	0.656	0.650 ± 0.011	0.589 ± 0.014	0.597
CMDH [12]	0.646 ± 0.033	0.677 ± 0.016	0.662	0.673 ± 0.028	0.692 ± 0.021	0.683	0.623 ± 0.024	0.604 ± 0.016	0.614
ACMR [14]	0.632 ± 0.014	0.683 ± 0.006	0.658	0.655 ± 0.022	0.679 ± 0.018	0.667	0.626 ± 0.026	0.578 ± 0.015	0.602
MASLN [27]	0.651 ± 0.023	0.706 ± 0.027	0.679	0.668 ± 0.011	0.707 ± 0.019	0.688	0.634 ± 0.014	0.629 ± 0.029	0.632
DANZCR [18]	0.663 ± 0.014	0.712 ± 0.031	0.688	0.685 ± 0.019	0.701 ± 0.026	0.693	0.647 ± 0.033	0.626 ± 0.019	0.637
TANSS (Ours)	**0.679 ± 0.015**	**0.733 ± 0.019**	**0.706**	**0.702 ± 0.024**	**0.724 ± 0.026**	**0.713**	**0.665 ± 0.018**	**0.642 ± 0.023**	**0.654**

B. Results on Standard Retrieval

We first assess our TANSS approach and the compared baselines on the standard retrieval task. The MAP scores of TANSS as well as the compared baselines on the three single-class Wikipedia, Pascal Sentences, and NUS-WIDE-S1 datasets are reported in Table II. Based on the comparison results, we can make several essential observations. First, the baselines CCA, CCA-3V, and JFSSL are inferior to the other DNN-based approaches, as the DNN-based approach is advantageous to obtain more effective common semantic space due to the powerful network architecture in them. Second, among the DNN-based approaches, the CMDH is the best baseline that is designed for standard retrieval task, as it takes a two-stage learning procedures to fully explore the intramodal properties which boost the semantic feature learning in the common subspace. Nevertheless, the methods MASLN, DANZCR, and our TANSS that are developed for zero-shot retrieval task remarkably outperforms CMDH on all dataset. It indicates that the label embeddings used in these zero-shot retrieval methods can not only boost the knowledge transfer across seen and unseen labels but also bridge the modality gap and enhance the learned common semantic features. Third, our TANSS achieves the highest average MAP scores for both Img2Txt and Txt2Img retrieval tasks compared with the latest methods MASLN, DANZCR, and AgNet. Specifically, compared with the best average MAP scores of DANZCR, our TANSS gains remarkable accuracy improvement from 0.688 to 0.706, 0.693 to 0.713, and 0.637 to 0.654 on average for the three datasets, respectively. We can conclude that our TANSS is advantageous to learn more compact common space than DANZCR and MASLN on standard retrieval task, since the label embeddings are less effective to be directly treated as common space in DANZCR or conditional information for autoencoder reconstruction in MASLN.

We then compare the standard retrieval performance of the above methods on the multilabel NUS-WIDE-S2 dataset. Note that our TANSS and the compared methods MASLN and DANZCR are originally proposed to use the label embeddings of one single class for semantic features generation. In the multilabel situation, we take the averaged label embeddings of the multiple classes for each image–text pair, to keep coherent with the single-label case. The comparison results of all the methods are demonstrated in Table III. First, it can be observed that the values of the averaged MAP scores of all methods are generally larger than those in Table II, the reason is that in the multiclass situation, it is much easier to retrieve the matched instances that share at least one class label with the query instance. However, in the single-label case, exactly matching the instances of one single class with the query is more challenging. Second, among the compared methods, MASLN performs inferior to Deep-SM, CMDH, and ACMR. It indicates that it may be not effective on the multilabel case, since using the averaged word embeddings may lose important information that captures the semantic associations of multiple labels. Third, compared with DANZCR, the self-supervision scheme used in our TANSS is more effective to incorporate the semantic associations of multiple classes during the semantic feature learning procedure, and helps TANSS to achieve the best retrieval performance.

TABLE III
STANDARD RETRIEVAL COMPARISON IN TERMS OF MAP WITH STANDARD DEVIATION ON THE MULTILABEL NUS-WIDE-S2 DATASETS

Methods	NUS-WIDE-S2		
	Img2Txt	Txt2Img	Avg.
CCA [16]	0.546	0.525	0.536
CCA-3V [5]	0.557	0.565	0.561
JFSSL [7]	0.663	0.591	0.627
DCCA [13]	0.673 ± 0.014	0.618 ± 0.009	0.646
Corr-AE [11]	0.692 ± 0.006	0.623 ± 0.011	0.658
Deep-SM [61]	0.735 ± 0.024	0.702 ± 0.018	0.719
CMDH [12]	0.742 ± 0.033	0.721 ± 0.028	0.732
ACMR [14]	0.746 ± 0.027	0.707 ± 0.012	0.727
MASLN [27]	0.735 ± 0.024	0.714 ± 0.016	0.725
DANZCR [18]	0.781 ± 0.035	0.734 ± 0.032	0.758
TANSS (Ours)	**0.796 ± 0.018**	**0.762 ± 0.025**	**0.779**

C. Results on Zero-Shot Retrieval

We then report the zero-shot retrieval results of the proposed TANSS and its counterparts on three datasets in Table IV. Compared with the results in Table II, here, in Table IV, the MAP scores of all methods deteriorate drastically, indicating that the zero-shot retrieval task is more challenging than the standard retrieval task. Besides, we can also see that for the compared methods, there is no obvious gap between the DNN-based methods and the shallow ones. For example,

TABLE IV
ZERO-SHOT RETRIEVAL COMPARISON IN TERMS OF MAP SCORE WITH STANDARD DEVIATION ON THREE SINGLE-LABEL DATASETS

Methods	Wikipedia			Pascal Sentences			NUS-WIDE-S1		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
CCA [16]	0.195	0.182	0.189	0.187	0.179	0.183	0.305	0.311	0.308
CCA-3V [5]	0.223	0.203	0.213	0.217	0.196	0.207	0.334	0.326	0.330
JFSSL [7]	0.242	0.217	0.230	0.246	0.241	0.244	0.352	0.358	0.355
Corr-AE [11]	0.218 ± 0.016	0.204 ± 0.014	0.211	0.225 ± 0.004	0.203 ± 0.010	0.214	0.341 ± 0.011	0.338 ± 0.005	0.340
DCCA [13]	0.235 ± 0.006	0.21 ± 0.009	0.223	0.238 ± 0.004	0.239 ± 0.009	0.239	0.346 ± 0.006	0.352 ± 0.002	0.349
Deep-SM [61]	0.272 ± 0.026	0.246 ± 0.010	0.259	0.253 ± 0.018	0.244 ± 0.010	0.249	0.356 ± 0.014	0.366 ± 0.022	0.361
CMDH [12]	0.282 ± 0.036	0.252 ± 0.021	0.267	0.266 ± 0.022	0.241 ± 0.011	0.254	0.367 ± 0.015	0.369 ± 0.013	0.368
ACMR [14]	0.296 ± 0.010	0.258 ± 0.014	0.277	0.284 ± 0.014	0.273 ± 0.008	0.279	0.376 ± 0.012	0.388 ± 0.013	0.382
MASLN [27]	0.313 ± 0.023	0.269 ± 0.026	0.291	0.297 ± 0.021	0.281 ± 0.017	0.289	0.389 ± 0.026	0.394 ± 0.019	0.392
DANZCR [18]	0.306 ± 0.031	0.283 ± 0.020	0.295	0.306 ± 0.025	0.297 ± 0.010	0.302	0.384 ± 0.013	0.407 ± 0.021	0.396
TANSS (Ours)	**0.322 ± 0.015**	**0.303 ± 0.009**	**0.313**	**0.318 ± 0.021**	**0.327 ± 0.016**	**0.323**	**0.404 ± 0.021**	**0.421 ± 0.018**	**0.413**

the shallow method JFSSL performs even better than Corr-AE and DCCA of the DNN-based methods in some cases, as the semantic inconsistence across seen and unseen labels may limit the adaption ability of DNN-based methods. In addition, among the DNN-based methods, ACMR performs better than CMDH on this task tough it is worse than CMDH on the standard retrieval task, showing that its adversarial learning scheme is capable to reduce the discrepancy of different modalities and is beneficial for the zero-shot retrieval task. Nevertheless, our proposed TANSS still gains significant improvements compared with the best counterpart DANZCR from 0.295 to 0.313, 0.302 to 0.323, and 0.396 to 0.413 on three datasets, respectively. The advantage of TANSS is that it utilizes the label embeddings of both seen and unseen labels as external knowledge, which alleviates the inconsistent semantics between source and target data. In addition, the self-supervised learning and adversarial learning schemes adopted in TANSS again ensure the knowledge transfer and generalization ability across the source set and the target set data in the learned subspace when producing the common representations for zero-shot retrieval.

According to the comparisons with all the counterparts on both standard and zero-shot retrieval tasks in Tables II and IV, our TANSS method achieves the best performance that can be attributed to the following aspects.

1) The three subnetworks mutually boost each other to learn more effective common semantic space and the adversarial learning scheme further alleviates the modality gap and ensures the semantic consistency across different modalities.
2) The adopted self-supervision scheme fully exploits word vectors of both seen and unseen labels to supervise the semantic feature learning process, and the cycle-consistence constraints enhance the knowledge transfer to the unseen labels and benefit zero-shot retrieval performance.

Furthermore, Fig. 4 shows the CMC curves of the proposed TANSS method and several counterparts on the three single-label datasets on the zero-shot retrieval task, where all curves are with the top-rank number r in [1, 15]. We can see that with different level of r, our TANSS method consistently obtains the highest matching rating compared with the counterparts, which again reflects its superior performance on the zero-shot cross-modal retrieval.

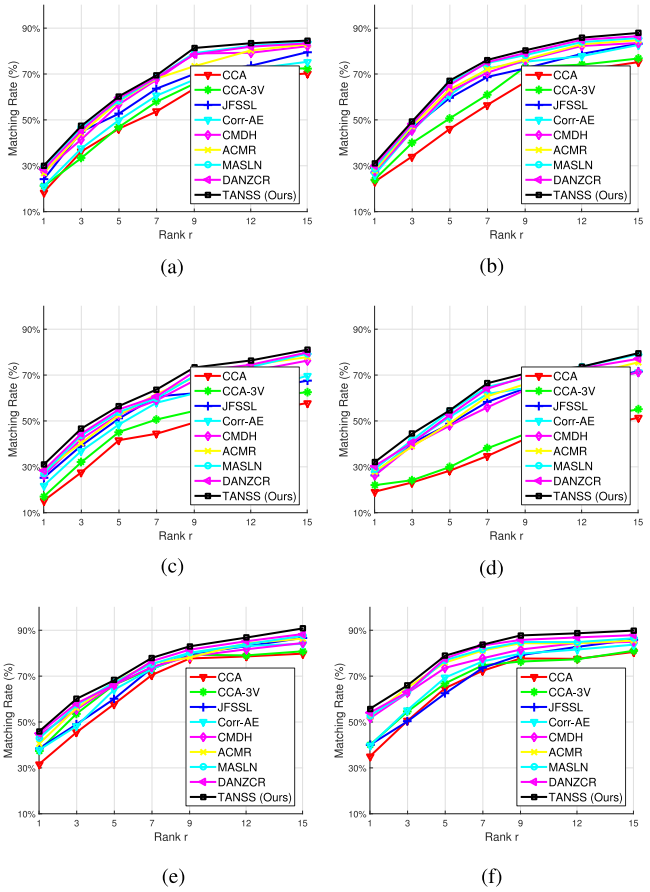


Fig. 4. CMC curves of the proposed TANSS method as well as its counterparts on all datasets for both Img2Txt and Txt2Img tasks under zero-shot retrieval scenario. (a) Img2Txt on Wikipedia. (b) Txt2Img on Wikipedia. (c) Img2Txt on Pascal Sentence. (d) Txt2Img on Pascal Sentence. (e) Img2Txt on NUS-WIDE-S1. (f) Txt2Img on NUS-WIDE-S1.

D. Ablation Study on TANSS

1) Results of Label Classification and Retrieval on Unseen Data: Similar as the existing ZSL approaches that utilize the label embeddings as intermediate features for classification

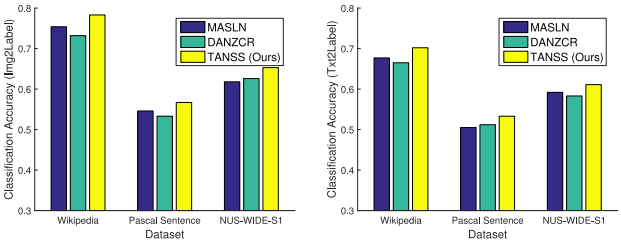


Fig. 5. Comparisons of our TANSS with two latest zero-shot cross-modal retrieval methods on (a) Img2Label and (b) Txt2Label tasks.

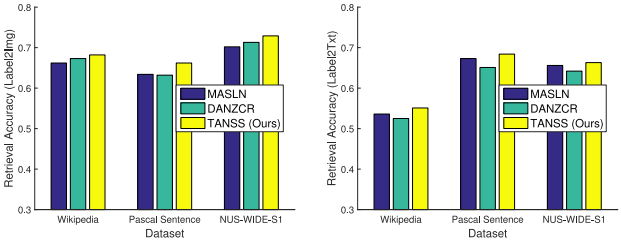


Fig. 6. Comparisons of our TANSS with two latest zero-shot cross-modal retrieval methods on (a) Label2Img and (b) Label2Txt tasks.

and retrieval on unseen data; here, we can also investigate the capability of TANSS for the two tasks on the target set data. As aforementioned in Section III-G, here we define two cases *Img2Label* and *Txt2Label* for the classification task, that is, classifying the unseen labels for both the image and text samples in the target set. For the retrieval task, we also define two cases *Label2Img* and *Label2Txt* that search the related images or texts in the target set using unseen labels. It is feasible to accomplish the two tasks since the semantic features generated from the images, texts, and label embeddings can be directly compared in the learned common semantic space in TANSS. We evaluate accuracy by comparing the predicted/retrieved results with the groundtruth labels for the two tasks.

Figs. 5 and 6 illustrate the classification and retrieval accuracy of TANSS compared with other two latest zero-shot retrieval approaches MASLN and DANZCR on all datasets. We can clearly seen that TANSS outperforms the two counterparts for both tasks on all datasets. Therefore, it indicates that the common semantic space learned by TANSS is more effective for feature representations of unseen images and texts, as well as label embeddings. Actually, label classification and retrieval tasks on unseen cross-modal data are important for more efficient data management, from this aspect, the proposed TANSS is promising to meet the requirements of the two tasks.

2) Visualization of the Learned Semantic Space: In this experiment, we further explore the effectiveness of the semantic features obtained by our TANSS on the target set data. We take the NUS-WIDE-S1 dataset as testbed, and randomly choose 500 images and 500 text samples in both the source and the target set. Then, the learned common semantic space from TANSS to generate the semantic features for the total 1000 samples. To make clear comparison, we also use the

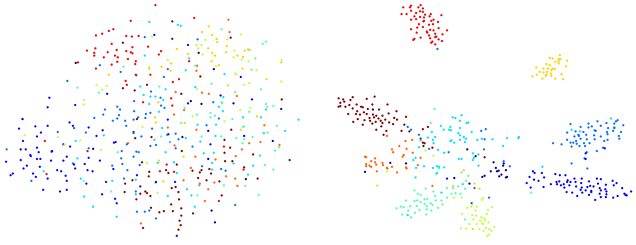


Fig. 7. t-SNE visualization on the NUS-WIDE-S1 dataset for the chosen target set samples. (a) Raw features. (b) Semantic features obtained by our TANSS. Clusters with different colors belong to different classes.

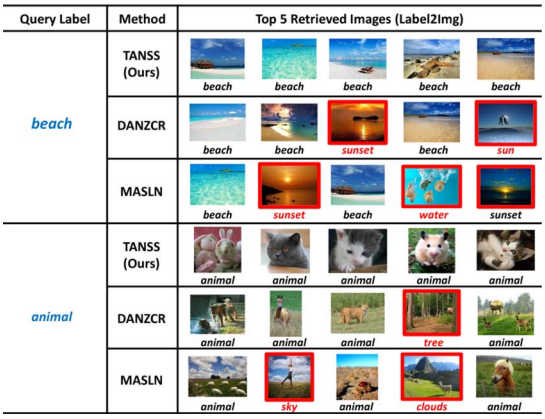


Fig. 8. Typical Label2Img retrieval results on the NUS-WIDE-S1 dataset obtained by our TANSS and the counterparts DANZCR and MASLN. Note that the incorrect retrieval results are marked in red rectangles for each query label.

principal component analysis (PCA) to perform dimension reduction on the original representations of images and texts to obtain the same dimension (100-D) in our experiments. Then, the distribution of these selected samples are visualized via the t-SNE [62] tool, where Fig. 7 shows the two distribution scenarios of their raw features and the learned semantic features. We can see that in Fig. 7(a), the samples of two modalities scatter and hardly be separated, while in Fig. 7(b), they form discriminative clusters and the samples belong to the same class are in the same cluster. Therefore, it indicates that under the supervision of class embedding, our TANSS approach is effective to model the correlations of different modalities. Meanwhile, the knowledge gap between the seen classes and the unseen classes can be bridged in the aligned clusters and benefits the zero-shot cross-modal retrieval.

Furthermore, we evaluate the effectiveness of the learned semantic space by conducting the aforementioned *Label2Img* retrieval task. Since our TANSS approach is advanced to synthesize semantic features for a given class label, the Label2Img retrieval can be accomplished by computing the similarities between the synthesized semantic features of the class label and the images in the learned semantic space. Fig. 8 provides two Label2Img retrieval exemplars by our TANSS and the counterparts DANZCR and MASLN on the NUS-WIDE-S1 dataset for each query (unseen) class label. We can observe

TABLE V
COMPARISON OF TANSS AND ITS DEGRADED BASELINES ON ZERO-SHOT RETRIEVAL TASK FOR ALL DATASETS

Methods	Wikipedia			Pascal Sentences			NUS-WIDE-S1		
	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.	Img2Txt	Txt2Img	Avg.
TANSS	0.322	0.303	0.313	0.318	0.327	0.323	0.404	0.421	0.413
TANSS (without $\mathcal{L}^l$)	0.295	0.276	0.286	0.291	0.286	0.289	0.373	0.389	0.381
TANSS (without $\mathcal{L}^a$)	0.302	0.291	0.297	0.301	0.306	0.304	0.385	0.397	0.391
TANSS (without $\mathcal{L}^a$ and $\mathcal{L}^l$)	0.274	0.262	0.268	0.268	0.275	0.271	0.363	0.367	0.365
TANSS (with L_1 norm)	0.257	0.245	0.251	0.261	0.266	0.264	0.341	0.355	0.348

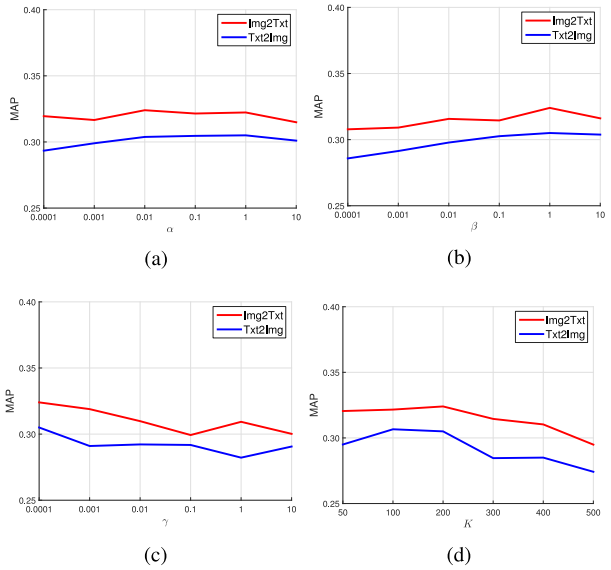


Fig. 9. Sensitivity analysis of the hyperparameters on the Wikipedia dataset. (a) α. (b) β. (c) γ. (d) K.

that for the two query unseen class labels "beach" and "animal," the matched top-5 images by our TANSS approach are all semantically related to the labels. However, for the DANZCR and MASLN methods, there are some incorrect matches with different groundtruth class labels in the results. For example, images of "water," "sunset," and "sun" are retrieved for query label *beach*, and images of "tree," "sky," and "clouds" are retrieved for query label *animal*. It again indicates that the synthesized semantic features of the unseen class labels by our TANSS are more coherent with the semantic features of the images with the same labels and, thus, benefit the Label2Img retrieval performance.

3) Analysis on Parameter Sensitivity: In our previous experiments, the values of the hyperparameters α, β, and γ in the objective function (12) are empirically set, where the three parameters balance the importance of each loss term in the three subnetworks of TANSS. In this experiment, we explicitly investigate their effects on TANSS. Specifically, the Wikipedia dataset is chosen as the testbed, and we vary the values of α, β, and γ in the range of $\{0, 10^{-4}, 10^{-3}, \ldots, 1, 10\}$ and assess their impact for zero-shot retrieval result on the target set. Specifically, the evaluation is performed by changing one parameter while fixing the others. Fig. 9 shows the sensitivity analysis of the three parameters. We can observe that TANSS can perform stable with a wide range of the parameters α and β, that is, α in $[10^{-2}, 1]$ and β in $[10^{-1}, 10]$, respectively. It indicates that the impact of pairwise similarity loss of both semantic features and transformed label embeddings are robust in each subnetwork. However, the optimal values of the parameter γ needs to be tuned, that is, values around 1, which shows that the classification loss contributes equally to other loss terms.

We further investigate how the nodes K of semantic feature layer (in common subspace) affect the performance of TANSS. Fig. 9(d) shows the zero-shot retrieval results of TANSS with different value of K in range of [50, 500]. It can be observed that when K is considerable small, that is, [50, 200], TANSS obtains better performance compared with larger K in [300, 500]. With $K = 200$, TANSS achieves the best performance. Therefore, it shows that the dimension of the common semantic space is also an important factor that has impact on TANSS, and in practice, smaller dimensions generally formulate more compact semantic features that improve the retrieval accuracy.

4) Further Discussions on TANSS: As the proposed TANSS consists of three different subnetworks, to explore the importance of each subnetwork on TANSS, in this experiment, we design several baselines by excluding specific component in TANSS. Specifically, according to the final objective function of TANSS in (12), there are three baselines: 1) "TANSS *without* $\mathcal{L}^l$" that ignores the LabNet subnetwork of self-supervised learning; 2) "TANSS *without* $\mathcal{L}^a$" that disables the adversarial learning; and 3) "TANSS *without* $\mathcal{L}^l$ and $\mathcal{L}^a$" that eliminates the two factors above. Besides, we also consider another baseline, that is, "TANSS *with* L_1-norm" that uses the L_1-norm in (4).

Table V shows the performance of TANSS and its degraded baselines on the zero-shot retrieval task for all datasets. Apparently, it can be observed that the four baselines have much worse performance than the original TANSS on all datasets, showing that both self-supervised learning and adversarial learning are crucial for TANSS on zero-shot retrieval. Specifically, the self-supervised learning has considerably more impact than adversarial learning on TANSS, as the results of TANSS *without* $\mathcal{L}^a$ is better than TANSS *without* $\mathcal{L}^l$ on all datasets. In addition, without self-supervision and adversarial learning, the baseline TANSS *without* $\mathcal{L}^l$ and $\mathcal{L}^a$ obtains the worse performances, as it is similar as the previous methods Deep-SM and CMDH that take a basic two-branch network for modeling the correlations between image and text modalities. Moreover, we can also observe that the result of the baseline "TANSS *with* L_1-norm" is inferior to the original TANSS and the other three baselines. The primary difference in them is the usage of the L_1-norm versus the L_2-norm. Generally, the L_1-norm is advanced to capture

the property of sparsity in feature vector; however, the semantic features in our TANSS are expected to be dense vectors that approximate the real-valued label embeddings. Therefore, using the L_1-norm leads to lossy information when establishing the cycle-consistency in (4). It indicates that in zero-shot retrieval scenario, the L_2-norm is superior and more robust to the L_1-norm on modeling the semantic consistency of the semantic features and the label embeddings for both seen and unseen labels.

V. Conclusion

In this paper, an effective method called TANSS has been proposed for *zero-shot* cross-modal retrieval. TANSS consists of three subnetworks that forms an end-to-end network structure. Two semantic feature learning subnetworks ImgNet and TxtNet capture the intrinsic data structures of different modalities and preserve the modality relationships in the common semantic space; a self-supervised semantic subnetwork LabNet leverages the word vectors of both seen and unseen labels as side information to supervise the semantic feature learning process and enhance to transfer knowledge to unseen labels. Two discriminators are additionally equipped upon the three subnetworks via an adversarial training style, to preserve the consistency of the semantic features between different modalities, which well strengthens their semantic correlation. The extensive experiments on both *standard* and *zero-shot* cross-modal retrieval tasks have clearly validated the advantage of our TANSS approach comparing with our previous conference version and other state-of-the-art approaches.

We plan to conduct our future work in the following directions: 1) more advanced semantic feature learning algorithms can be utilized for zero-shot cross-modal retrieval; 2) additional unlabeled multimodal data can be exploited to boost the current self-supervision scheme; and 3) other external knowledge, for example, label hierarchy and knowledge graph, will be incorporated to fully build the association between seen and unseen labels for knowledge transfer.

References

[1] L. Chen, D. Xu, I. W. Tsang, and X. Li, "Spectral embedded hashing for scalable image retrieval," *IEEE Trans. Cybern.*, vol. 44, no. 7, pp. 1180–1190, Jul. 2014.

[2] X. Zhou *et al.*, "Graph convolutional network hashing," *IEEE Trans. Cybern.*, to be published.

[3] M. Zhang, Y. Yang, H. Zhang, Y. Ji, H. T. Shen, and T.-S. Chua, "More is better: Precise and detailed image captioning using online positive recall and missing concepts mining," *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 32–44, Jan. 2019.

[4] D. R. 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, Dec. 2004.

[5] Y. Gong, Q. Ke, M. Isard, and S. Lazebnik, "A multi-view embedding space for modeling Internet images, tags, and their semantics," *Int. J. Comput. Vis.*, vol. 106, no. 2, pp. 210–233, 2014.

[6] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2088–2095.

[7] 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.

[8] 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, p. 22, 2019.

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

[10] 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.

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

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

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

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

[15] 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.

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

[17] R. Liu, Y. Zhao, L. Zheng, S. Wei, and Y. Yang, "A new evaluation protocol and benchmarking results for extendable cross-media retrieval," *CoRR*, vol. abs/1703.03567, 2017.

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

[19] 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.

[20] 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.

[21] 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.

[22] X. Xu, X. Zhou, F. Shen, L. Gao, H. T. Shen, and X. Li, "Fusion by synthesizing: A multi-view deep neural network for zero-shot recognition," *Signal Process.*, vol. 164, no. 1, pp. 354–367, 2019.

[23] X. Xu, T. M. Hospedales, and S. Gong, "Transductive zero-shot action recognition by word-vector embedding," *Int. J. Comput. Vis.*, vol. 123, no. 3, pp. 309–333, 2017.

[24] J. Qin *et al.*, "Zero-shot action recognition with error-correcting output codes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1042–1051.

[25] C. Doersch and A. Zisserman, "Multi-task self-supervised visual learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2070–2079.

[26] A. Kolesnikov, X. Zhai, and L. Beyer, "Revisiting self-supervised visual representation learning," *CoRR*, vol. abs/1901.09005, 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 Retriev.*, 2018, pp. 46–54.

[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 470–477.

[29] T. Mikolov, K. Chen, and G. Corrado, "Jeffrey dean: Efficient estimation of word representations in vector space," in *Proc. Int. Conf. Learn. Represent. Workshop*, 2013, pp. 2312–2319.

[30] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1188–1196.

[31] Y. Zhuang, Y. Wang, F. Wu, Y. Zhang, and W. Lu, "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in *Proc. AAAI Conf. Artif. Intell.*, 2013, pp. 1070–1076.

[32] W. Wang, R. Arora, K. Livescu, and J. Bilmes, "On deep multi-view representation learning," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 1083–1092.

[33] X. Xu, L. He, H. Lu, L. Gao, and Y. Ji, "Deep adversarial metric learning for cross-modal retrieval," *World Wide Web*, vol. 22, no. 2, pp. 657–672, 2019.

[34] J. Song, Y. Guo, L. Gao, X. Li, A. Hanjalic, and H. T. Shen, "From deterministic to generative: Multimodal stochastic RNNs for video captioning," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.

[35] N. Srivastava and R. Salakhutdinov, "Learning representations for multimodal data with deep belief nets," in *Proc. Int. Conf. Mach. Learn. Workshop*, 2012, pp. 1–8.

[36] J. Song, Y. Yang, Y. Yang, Z. Huang, and H.-T. Shen, "Inter-media hashing for large-scale retrieval from heterogeneous data sources," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2013, pp. 785–796.

[37] 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.

[38] 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.

[39] D. Wang, Q. Wang, and X. Gao, "Robust and flexible discrete hashing for cross-modal similarity search," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2703–2715, Oct. 2018.

[40] D. Wang, X.-B. Gao, X. Wang, and L. He, "Label consistent matrix factorization hashing for large-scale cross-modal similarity search," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published.

[41] Y. Zhuang, Z. Yu, W. Wang, F. Wu, S. Tang, and J. Shao, "Cross-media hashing with neural networks," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 901–904.

[42] Q.-Y. Jiang and W.-J. Li, "Deep cross-modal hashing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3270–3278.

[43] J. Zhang, Y. Peng, and M. Yuan, "SCH-GAN: Semi-supervised cross-modal hashing by generative adversarial network," *IEEE Trans. Cybern.*, to be published.

[44] C. Deng, Z. Chen, X. Liu, X. Gao, and D. Tao, "Triplet-based deep hashing network for cross-modal retrieval," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3893–3903, Aug. 2018.

[45] C. Li, C. Deng, N. Li, W. Liu, X. Gao, and D. Tao, "Self-supervised adversarial hashing networks for cross-modal retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1342–1350.

[46] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.

[47] J. Li, K. Lu, Z. Huang, L. Zhu, and H. T. Shen, "Transfer independently together: A generalized framework for domain adaptation," *IEEE Trans. Cybern.*, vol. 49, no. 6, pp. 2144–2155, Jun. 2019.

[48] J.-Y. 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. 1166–1174.

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

[50] X. Huang, Y. Peng, and M. Yuan, "MHTN: Modal-adversarial hybrid transfer network for cross-modal retrieval," *IEEE Trans. Cybern.*, to be published.

[51] Z. Ji, Y. Sun, Y. Yu, Y. Pang, and J. Han, "Attribute-guided network for cross-modal zero-shot hashing," *CoRR*, vol. abs/1802.01943, 2018.

[52] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[53] H. Zhang, Y. Sun, L. Liu, X. Wang, L. Li, and W. Liu, "ClothingOut: A category-supervised GAN model for clothing segmentation and retrieval," *Neural Comput. Appl.*, pp. 1–12, Aug. 2018. [Online]. Available: https://citation-needed.springer.com/v2/references/10.1007/s00521-018-3691-y?format=bibtex&flavour=citation

[54] Y. Ji, H. Zhang, and Q. M. J. Wu, "Saliency detection via conditional adversarial image-to-image network," *Neurocomputing*, vol. 316, pp. 357–368, Nov. 2018.

[55] 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.

[56] C. Rashtchian, M. Young, P. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's Mechanical Turk," in *Proc. NAACL HLT Workshop Creating Speech Lang. Data Amazon's Mech. Turk*, 2010, pp. 674–686.

[57] 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 Retriev.*, 2009, p. 48.

[58] L. He, X. Xu, H. Lu, Y. Yang, F. Shen, and H. T. Shen, "Unsupervised cross-modal retrieval through adversarial learning," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2017, pp. 1153–1158.

[59] M. Rosca, B. Lakshminarayanan, D. Warde-Farley, and S. Mohamed, "Variational approaches for auto-encoding generative adversarial networks," *CoRR*, vol. abs/1706.04987, 2017.

[60] Y. Li, N. Xiao, and W. Ouyang, "Improved generative adversarial networks with reconstruction loss," *Neurocomputing*, vol. 323, pp. 363–372, Jan. 2019.

[61] 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.

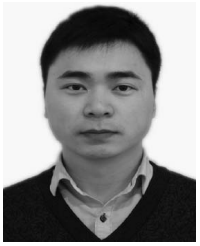
[62] L. V. D. Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

Xing Xu received the B.E. and M.E. degrees from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2012, respectively, and the Ph.D. degree from Kyushu University, Fukuoka, Japan, in 2015.

He is currently with the School of Computer Science and Engineering, University of Electronic of Science and Technology of China, Chengdu, China. His current research interests include multimedia information retrieval, pattern recognition, and computer vision.

Huimin Lu received the M.S. degrees in electrical engineering from the Kyushu Institute of Technology, Kitakyushu, Japan, and Yangzhou University, Yangzhou, China, in 2011, and the Ph.D. degree in electrical engineering from the Kyushu Institute of Technology in 2014.

From 2013 to 2016, he was a JSPS Research Fellow. He is currently an Associate Professor with the Kyushu Institute of Technology, a Visiting Professor with Shanghai Jiao Tong University, Shanghai, China, and an Excellent Young Researcher of the Ministry of Education, Culture, Sports, Science and Technology, Tokyo, Japan. His current research interests include computer vision, robotics, artificial intelligence, and ocean observing.



Jingkuan Song received the Ph.D. degree in information technology from the University of Queensland, Brisbane, QLD, Australia, in 2014.

He is a Professor with the University of Electronic Science and Technology of China, Chengdu, China. His current research interests include large-scale multimedia retrieval, image/video understanding using hashing, graph learning, and deep learning techniques.

Mr. Song was a recipient of the Best Paper Award in ICPR (Best Paper Honorable Mention Award) in 2017.

Yang Yang (M'16) received the bachelor's degree in computer science from Jilin University, Changchun, China, in 2006, the master's degree in computer science from Peking University, Beijing, China, in 2009, and the Ph.D. degree in computer science from the University of Queensland, Brisbane, QLD, Australia, in 2012.

He is currently with the University of Electronic Science and Technology of China, Chengdu, China. His current research interests include multimedia content analysis, computer vision, and social media analysis.

Heng Tao Shen (SM'10) received the B.Sc. (First Class Hons.) and Ph.D. degrees in computer science from the Department of Computer Science, National University of Singapore, Singapore, in 2000 and 2004, respectively.

He is currently a Professor of National "Thousand Talents Plan," the Dean of the School of Computer Science and Engineering, and the Director of the Center for Future Media with the University of Electronic Science and Technology of China, Chengdu, China. His current research interests include multimedia search, computer vision, artificial intelligence, and big data management.

Dr. Shen was a recipient of the best paper awards from international conferences, including the Best Paper Award from ACM Multimedia in 2017 and the Best Paper Award-Honorable Mention from ACM SIGIR in 2017. He has served as the PC Co-Chair for ACM Multimedia 2015. He is currently an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.

Xuelong Li (M'02–SM'07–F'12) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China.

He is currently a Full Professor with the School of Computer Science and the Center for Optical Imagery Analysis and Learning, Northwestern Polytechnical University, Xi'an, China.