# Learning Cross-Modal Common Representations by Private–Shared Subspaces Separation

Xing Xu, *Member, IEEE*, Kaiyi Lin, Lianli Gao, *Member, IEEE*, Huimin Lu, *Senior Member, IEEE*, Heng Tao Shen, *Senior Member, IEEE*, and Xuelong Li, *Fellow, IEEE*

***Abstract*—Due to the inconsistent distributions and representations of different modalities (e.g., images and texts), it is very challenging to correlate such heterogeneous data. A standard solution is to construct one common subspace, where the common representations of different modalities are generated to bridge the heterogeneity gap. Existing methods based on common representation learning mostly adopt a less effective two-stage paradigm: first, generating separate representations for each modality by exploiting the modality-specific properties as the complementary information, and then capturing the cross-modal correlation in the separate representations for common representation learning. Moreover, these methods usually neglect that there may exist interference in the modality-specific properties, that is, the unrelated objects and background regions in images or the noisy words and incorrect sentences in the text. In this article, we hypothesize that explicitly modeling the interference within each modality can improve the quality of common representation learning. To this end, we propose a novel model private–shared subspaces separation (P3S) to explicitly learn different representations that are partitioned into two kinds of subspaces: 1) the common representations that capture the cross-modal correlation in a shared subspace and 2) the private representations that model the interference within each modality in two private subspaces. By employing the orthogonality constraints between the shared subspace and the private subspaces during the one-stage joint learning procedure, our model is able to learn more effective common representations for different modalities in the shared subspace by fully excluding the interference within each modality. Extensive experiments conducted on cross-modal retrieval verify the advantages of our P3S method compared with 15 state-of-the-art methods on four widely used cross-modal datasets.**

***Index Terms*—Common representation learning, cross-modal retrieval, subspace learning.**

Manuscript received February 10, 2020; revised April 11, 2020; accepted June 25, 2020. Date of publication August 11, 2020; date of current version May 19, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61976049, Grant 61872064, and Grant 61632007; in part by the Fundamental Research Funds for the Central Universities under Grant ZYGX2019Z015; and in part by the Sichuan Science and Technology Program, China, under Grant 2019ZDZX0008, Grant 2019YFG0003, Grant 2020YFS0057, and Grant 2020YJ0038. This article was recommended by Associate Editor Y.-M. Cheung. *(Corresponding author: Heng Tao Shen.)*

Xing Xu, Lianli Gao, and Heng Tao Shen are with the Center for Future Multimedia 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@gmail.com; lianli.gao@uestc.edu.cn; shenhengtao@hotmail.com).

Kaiyi Lin is with the Center for Future Multimedia and School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China, and also with the School of Software and Microelectronics, Peking University, Beijing 100871, China.

Huimin Lu is with the Department of Mechanical and Control Engineering, Kyushu Institute of Technology, Kitakyushu 8048550, Japan (e-mail: dr.huimin.lu@ieee.org).

Xuelong 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 figures in this article are available at https://doi.org/10.1109/TCYB.2020.3009004.

Digital Object Identifier 10.1109/TCYB.2020.3009004

## I. Introduction

THESE DAYS, the presence of massive multimodal data, such as images, videos, texts, and audios, not only significantly enriches people's daily life but also brings great challenges to multimedia data management and retrieval. Traditional unimodal retrieval applications, such as document retrieval [1], [2] and image/video search [3]–[6], limitedly concentrate on finding the retrieval results for the query data within the same modality. It cannot bridge the connections among different modalities and provide comprehensive and diverse information consisting of multiple modalities for the query. To erase the discrepancy between different modalities, *cross-modal retrieval* that takes a query from one modality to search for related results from another modality has become a research hotspot. Compared with the unimodal retrieval applications, the cross-media retrieval is advanced to provide more flexible and informative results for a query of any modality. However, the so-called "heterogeneity gap," that is, the data of different modalities have diverse representations in different feature spaces, which makes it challenging and difficult to directly measure the cross-modal similarity.

In real scenarios, the data of different modalities are intrinsically correlated from content and semantically consistent on a certain level. For example, the images and textual paragraphs coexisting in a news article on the web may contain relevant semantic topics. Thus, it is crucial to capture the cross-modal correlation and construct appropriate metrics for measuring the semantical relevance of the heterogeneous data. To eliminate the heterogeneity gap, it is natural and intuitive to project the heterogeneous data of different modalities into an intermediate common subspace, where the joint distribution of the heterogeneous data can be modeled by learning their common representations. In the common subspace, data of relevant semantics would have "similar" common representations and be close to each other. Under this situation, the similarities of pairwise instances of different modalities can be measured using conventional distance metrics, for example,

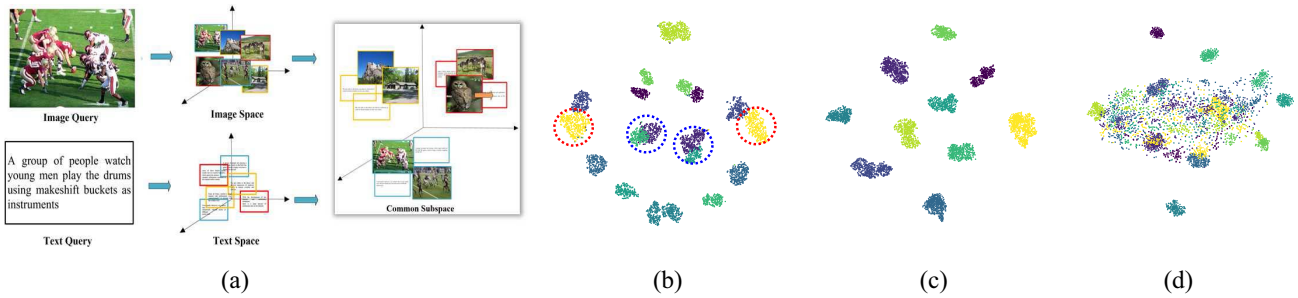



Fig. 1. Illustration of the mainstream framework for common representation learning on cross-modal retrieval paradigm in (a); and the visualizations of the learned common subspace by ACMR [7] in (b), and the learned shared subspace and private subspaces by our P3S method in (c) and (d), respectively.

the Euclidean distance and cosine similarity. Following this idea, during the last decade, a large number of methods [7]–[12] have been proposed to learn common representations for the data of different modalities, which has become the mainstream solution for cross-modal retrieval.

As the flowchart illustrated in Fig. 1(a), the standard solution for common representation learning mostly adopts a two-stage learning procedure: first, generating the separate representations for each modality data to model the intramodality correlation and second, learning the common representations of different modalities by exploring the cross-modal correlation from the separate representations. However, the cross-modal correlation may not be effectively captured since the two stages are performed successively rather than simultaneously. In addition, the first stage may not incorporate important intermodality information that can benefit the second stage as it only pays attention to modeling the intramodality correlation. Taking the recently proposed two-stage method adversarial cross-modal retrieval (ACMR) [7] for example, Fig. 1(b) plots the t-SNE [13] visualization of the common representations of all the training image–text instances of the Wikipedia [8] dataset in the learned common subspace by ACMR. Note that the points marked as circle and cross represent the image and text instances, respectively. It can be observed that the intramodality correlation is well captured as the instances (images or texts) within each modality form several intensive clusters corresponding to their semantic labels. However, some image clusters and text clusters of the same classes (marked in dotted red circles) are far away in the subspace, while some other clusters of different classes (marked in dotted blue circles) overlap in the subspace. This result indicates that the cross-modal correlation is not fully captured by the ACMR and, thus, it fails to correctly measure their semantic relevance.

To address this issue, some studies [14], [15] further put forward to exploit the modality-specific properties as the complementary to get more meaningful separate representation for capturing the rich cross-modal correlation during learning the common representation. Nevertheless, they still follow the two-stage learning paradigm. Furthermore, some recent work [11], [16] proposes a single-stage framework with the advanced scheme, such as adversarial learning and attention mechanism, to preserve the *modality-specific properties* during the cross-modal correlation learning (CCL) procedure. Though these methods have verified the benefit of the modality-specific properties of each modality on common representation learning, they may neglect that the underlying *interference* within each modality that may lead to the negative effect. For example, images usually contain objects or background regions that are unrelated to the textual descriptions, while textual descriptions may also be noisy even incorrect to describe the images. Therefore, it is necessary to first eliminate the interference in the modality-specific properties to provide valuable complementary for common representation learning.

To this end, we propose a novel single-stage method called private–shared subspaces separation (P3S) to learn cross-modal common representations. To explicitly capture the interference in the modality-specific properties, we assume the interference within each modality spans in a feature space (called *private space*) that is mutually exclusive from the common subspace (called *shared space*). In our P3S method, we allocate two private spaces, one for each modality, where the private representations are learned to capture the interference in the modality-specific properties. Meanwhile, a shared space is assigned to be shared between different modalities, where the cross-modal common representations are learned to model the cross-modal correlation. Through finding such shared subspace which is *orthogonal* to the private subspaces in a joint learning procedure, our P3S approach is able to separate the interference of each modality and learn more compact common representations of different modalities in the shared subspace.

Similar to the settings in Fig. 1(b)–(d), respectively, shows the t-SNE visualizations of the training instances of the Wikipedia dataset in the learned shared space and the private spaces obtained by our P3S method. We can observe that in Fig. 1(c), the image and text instances of the same semantic label form much closer clusters (in the same colors) compared with the distributions in Fig. 1(b). In contrast, the scattered distribution of the private representations of the image and text instances indicates that the interference within each modality is effectively captured and excluded from the learned common representations. Therefore, the learned common representations by the proposed P3S are more compact and effective for cross-modal retrieval.

As shown in Fig. 2, our P3S is designed as an end-to-end network structure consisting of three subnetworks: 1) shared subspace learning subnetwork (SNet) is devised to effectively reduce the heterogeneity gap and model the joint distribution of the heterogeneous data in the shared space. Specifically, the SNet contains two shared encoders (one for each modality),

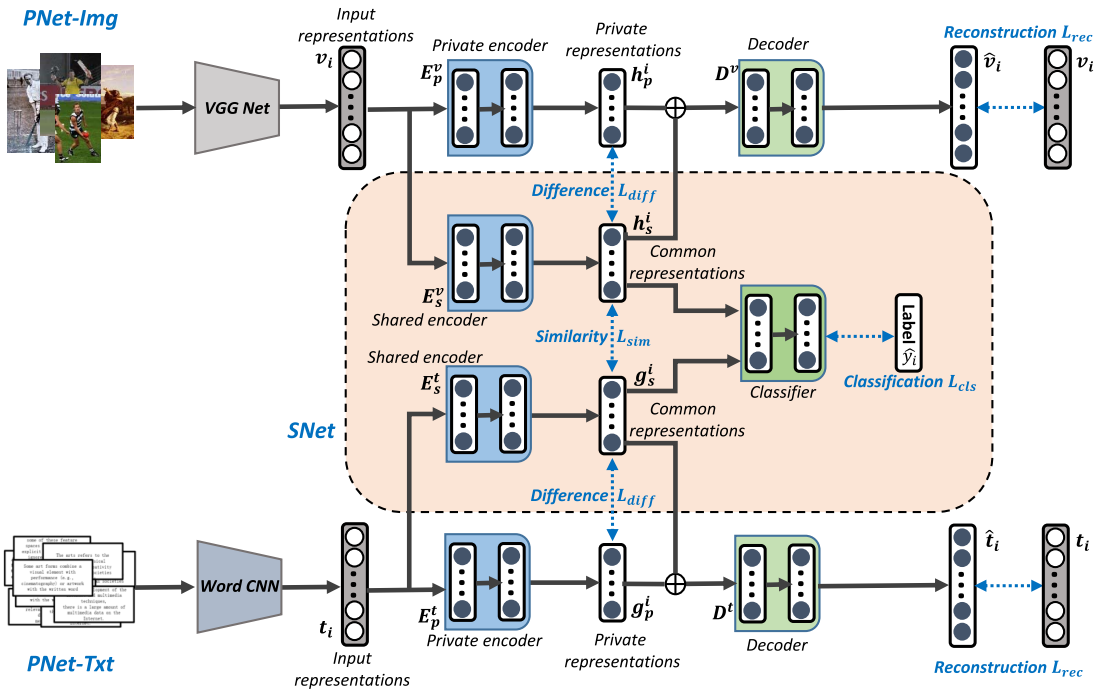



Fig. 2. Framework of the proposed P3S method, including three subnetworks: two PNets (PNet-Img and PNet-Txt) that, respectively, learn the private subspace for each modality, and an SNet that learns the shared subspace for different modalities. The SNet contains a shared encoder for generating common representations and a classifier for preserving their discriminability based on the semantic label information. Each PNet consists of a private encoder for private representations generation and a decoder for the robust reconstruction of the input representations. The pretrained VGG [20] and WordCNN [21] models are adopted to extract the input representations for images and texts, respectively. Note that the input representations of images and texts are denoted by the layer with white nodes, while the hidden layers are marked with gray nodes.

which can learn the common representations by capturing intermodality correlation and intramodality correlation simultaneously. Besides, the SNet also leverages the semantic label information to make the common representations semantically discriminative and 2) two private subspace learning subnetworks (PNet-Img and PNet-Txt) are proposed to form two parallel autoencoders that explicitly capture the interference within each modality. Specifically, each PNet has a private encoder and a decoder, where the former learns the private representations encoding the interference in the private spaces, while the latter models the intramodality reconstruction information to preserve semantic consistency with each modality. Different from the usage of the decoder architecture in existing methods [16]–[19], the decoder in each PNet integrates both the private representations from the private encoder and the common representations from the shared encoder (in the SNet) for more robust reconstruction and semantical consistency.

Our main contributions in this article are three-fold.

1) A novel common representation learning architecture called P3S is proposed for cross-modal retrieval. The proposed P3S method consists of three associate subnetworks: an SNet that learns cross-modal common representations in a shared subspace, and two PNets that obtain private representations of each modality in two private subspaces.
2) We derive four different schemes to effectively model cross-modal correlation in the SNet method considering the pairwise correlation as well as the overall statistical information of the heterogeneous data.
3) We employ the subspace orthogonality constraints in the joint learning procedure of the three subnetworks to explicitly separate the private spaces from the shared space, which ensure that the learned common representations in the shared subspace are more effective for retrieval. We compare our P3S approach with 15 state-of-the-art methods on the widely used multimodal datasets and the experimental results demonstrate the effectiveness of our P3S approach for cross-model retrieval.

## II. Related Work

In this section, we discuss the common representation learning in existing cross-modal retrieval methods. Regarding the diverse learning schemes, three typical categories can be grouped: 1) the traditional methods; 2) the deep neural-network (DNN)-based methods; and 3) the domain adaptation (DA) methods.

*1) Traditional Methods:* Linear projection functions are usually assumed in the traditional methods to map the heterogeneous data into a latent subspace. The pioneering work in these methods is canonical correlation analysis (CCA) [8], which targets maximizing the correlation of the projected features of different modality data in a lower dimensional common subspace. The extensions of CCA, such as kernel-based CCA (KCCA) [22]; multiview CCA [9], [23]; and multilabel CCA [24], have been extended to handle various conditions. In addition to the CCA and its extensions, some

other methods that utilize advanced schemes, such as factor analysis [25], dictionary learning [26], and half-quadratic optimization [27], to learn more effective common representations. A major drawback of the above methods is that they mostly require collecting paired instances of different modalities, which are usually labor costly. To tackle this problem, several methods [14], [28], [29] leverage unlabeled images or documents that can be easily crawled on the web to construct partially labeled graph, which further boost the common representation learning performance.

*2) DNN-Based Methods:* With the powerful nonlinear feature extraction capability of the DNNs, recent studies mainly leverage various DNN structures to capture the complex cross-modal correlation during learning common representations. Ngiam *et al.* [30] proposed a multimodal autoencoder structure to learn the joint representation of multimodal data based on the deep restricted Boltzmann machine. Andrew *et al.* [31] developd a deep version of CCA (DCCA) with stacked nonlinear transformation, which captures more effective correlation for multimodal data. Feng *et al.* proposed a correspondence autoencoder (Corr-AE) [17], which constructs two-way subnetworks based on the structure of autoencoder, where the hidden codes in the middle layer are used as common representations to model the association of different modalities. Peng *et al.* [14] built a cross-modal hierarchical network called CMDH which fully explores the inter- and intra-modality correlation via a two-stage learning strategy. They further propose the CCL [15] method, which utilizes a multitask learning framework to reach a tradeoff between two constrains, that is, jointly preserving the intramodality semantic discrimination and the intermodality pairwise similarity. Recently, the generative adversarial network (GAN) [32] has show advanced capability on modeling the data distribution and learning discriminative representation for unimodal data. A vanilla GAN model contains two modules that play the adversarial game: 1) a generator that produce fake data samples to mimic the distribution of real data samples and 2) a discriminator that distinguishes a sample real or coming from the generator. The two modules in the GAN are trained with competition in an adversarial way, to ensure that they can learn more compact representation of real data samples. Many works have developed GAN-based models for various computer vision problems dealing with unimodal data, such as image synthesis [33], image super-resolution [34], and video prediction [35]. Several studies [7], [12], [16], [36] extend the unimodal GAN to model the joint distribution of multimodal data. These methods employ a feature generator for each modality and a modality discriminator to learn modality-invariant representations under the adversarial learning manner.

As discussed in Section I, the aforementioned DNN-based methods mostly construct one latent subspace to learn the common representations for different modality data. In addition, they usually extract the modality-specific properties as complementary to improve the common representation learning under a two-stage learning paradigm. However, the modality-specific properties may contain interference that deteriorates the effectiveness of the learned common representations. Differently, in our P3S method, we introduce two private subspaces that capture the interference in the modality-specific properties. Notably, we ensure that the common representations are learned in the shared subspace that is orthogonal to the two private subspaces. The three subspaces are jointly learned in one stage to fully model the cross-modal correlation with the complement of the refined modality-specific properties, leading to more compact common representations for retrieval.

It is worth mentioning that in this work we mainly focus on learning *real-valued* common representations. Actually, a group of other studies [29], [37]–[41] concentrate on the efficiency of cross-modal retrieval and target to learn *binary* common representations. Readers can refer to the related survey paper [42] to obtain a more comprehensive review of these cross-modal hashing methods. Nevertheless, the cross-modal hashing methods usually design the learning objective functions with binary constraints, which are intrinsically different from learning real-valued common representations in our work.

*3) Domain Adaptation Methods:* There exists the domain shift problem when applying the trained models from the source domain to the target domain with different data distributions. As the data distributions of different domains can be aligned in high-level correlation space to reduce the discrepancy, many DA methods, including both shallow methods [43]–[45] and DNN-based methods [46]–[49] are proposed to cope with the domain shift problem by learning an adaptive common subspace that can generate domain-invariant representations for the data in both the source and target domains.

Compared to the common representation learning methods on multimodal data discussed above, most works on DA focus on unimodal scenario, that is, the source- and target-domain data are in the same modality (e.g., images). Several research efforts have been devoted to develop DA methods [50]–[53] operating on the source data that spans multiple modalities. Nevertheless, they are limitedly designed for the classification task, which learn the modality-invariant representation by finding the latent association between the multimodal source-domain data and the unimodal target data. Differently, we focus on the common representation learning on the cross-modal retrieval problem, where both the source- and target-domain data are with multiple modalities. We extend the domain-invariant representation learning schemes in the DA methods [44], [47] to the cross-modal retrieval task, and develop effective schemes to capture the cross-modal correlation for common representation learning. These extensions successfully show advantages compared with several conventional schemes proposed for cross-modal retrieval task.

## III. Proposed P3S Method

### A. Problem Formulation

In this work, we consider the bimodal data of images and texts as a general case study. Specifically, suppose we have a training data corpus consisting of $N$ pairwise image–text instances, that is, $\mathcal{O}_{tr} = \{o_p\}_{p=1}^{N}$ and $o_p = (\mathbf{v}_p, \mathbf{t}_p)$, where $\mathbf{v}_p \in \mathbb{R}^{d_v}$ and $\mathbf{t}_p \in \mathbb{R}^{d_t}$, respectively, denote the original feature

vectors of the image and the text for the $p$th instance $o_p$. Note that the dimensionalities $d_v$ and $d_t$ of the visual and textual features are usually different. Furthermore, $y_p$ is the semantic label assigned to $o_p$, and there are $C$ different semantic labels in $\mathcal{O}_{tr}$. Besides, we also have a testing set $\mathcal{O}_{te}$ that have $N'$ instances for each modality, that is, $\{\mathbf{v}'_q\}_{q=1}^{N'}$ and $\{\mathbf{t}'_q\}_{q=1}^{N'}$.

Our goal is to learn an effective common subspace from the training instances $\mathcal{O}_{tr}$ where the common representation of each image instance and each text instance can be extracted. In the testing stage, cross-modal retrieval on the test instances $\mathcal{O}_{te}$ can be directly performed based on their generated common representations. For a query image $\mathbf{v}'_q$ in $\mathcal{O}_{te}$, retrieving the relevant text $\mathbf{t}'_q$ can be performed by measuring their similarity based on their extracted common representations.

### B. Architecture of P3S

Fig. 2 illustrates the overall framework of the proposed P3S method, which consists of three key modules: a shared subspace learning network called SNet, and two private subspace learning networks called PNet-Img and PNet-Txt for image and text modalities, respectively. For each modality, two subspaces, that is, a private subspace that captures the interference and a shared subspace that models the association across two modalities are learned through an autoencoder, respectively. With a modified encoder–decoder structure of the autoencoder, each modality is equipped with two encoders $E_p^*$ and $E_s^*$ and one decoder $D^*$, $* = v, t$ that are all built with several fully connected layers. Specifically, given the input representations of each modality data, the private encoders $E_p^*$ in two PNets encode the interference within each modality and generate the private representations in the private subspaces. Meanwhile, the shared encoders $E_s^*$ in SNet capture the cross-modal correlation and produce the common representations of two modalities in the shared subspace. Additional orthogonality constraints are employed in order to make the common representations in the shared subspace different from the private representations in the private subspaces. Furthermore, both common and private representations are simultaneously fed into the decoder $D^*$, $* = v, t$ in the two PNets for robust reconstruction and semantic consistency of the original features of each modality data. As a result, the obligations of the private and the shared subspaces in each modality is cleared separated, and the common representations in the shared subspace explicitly capture the correlations of two modalities without the interference within each modality. In the following sections, we will elaborately describe the details of each subnetwork.

### C. Learning Shared Subspace in SNet

The fundamental premise for cross-model retrieval is to make each pair of image and text closer together in the common shared subspace to represent closely relevant semantic. We design the SNet to learn the shared subspace from two aspects: 1) the similarity relationship between the input representations of each pair of image and text instances and their corresponding common representations in the shared subspace is well preserved and 2) the common representations of the instances in each modality are discriminative according to their semantic labels.

*1) Measuring the Cross-Modal Similarity:* For an image–text pair $o_i = \{\mathbf{v}_i, \mathbf{t}_i\}$, their common representations $\mathbf{h}_s^i \in \mathbb{R}^m$ and $\mathbf{g}_s^i \in \mathbb{R}^m$ in the shared subspace are generated by the respective shared encoders $E_s^v$ and $E_s^t$ as

$$\mathbf{h}_s^i = E_s^v(\mathbf{v}_i); \quad \mathbf{g}_s^i = E_s^t(\mathbf{t}_i). \tag{1}$$

In the SNet, both $E_s^v$ and $E_s^t$ consist of several fully connected layers with semantic constraints. It is notable that the semantic constraints are usually added to encourage the common representations $\mathbf{h}_s^i$ and $\mathbf{g}_s^i$ in the same class to be as similar as possible irrespective of the modality, which is crucial for the common representation learning.

Actually, the SNet is flexible to incorporate various schemes that address the semantic constraints. Here, we consider two typical schemes that have been widely used for cross-modal retrieval. Moreover, inspired by the related studies in DA, we also develop two other schemes that are based on distribution adaptation. We explicitly discuss the four schemes as follows.

Cross-modal correlation similarity (CMCS) is a commonly used criterion [17], [18], [51] that takes the Euclidean distance to measure the discrepancy of the pairwise image and text instances by their common representations. Intuitively, the common representations of the pairwise instances with the same label in different modalities should be similar to each other. Thus, for all the image–text pairs in the training set, the similarity loss in SNet based on CMCS is formulated as

$$L_{\text{sim}}^{\text{CMCS}} = \frac{1}{N}\sum_{i=1}^{N}\left\|\mathbf{h}_s^i - \mathbf{g}_s^i\right\|^2 \tag{2}$$

where $\mathbf{h}_s^i$ and $\mathbf{g}_s^j$ represent the $i$th image–text pair. By minimizing this loss, the common representations of pairwise instances can be aligned and their discrepancy can be reduced.

Cross-modal adversary (CMA) is recently proposed in [7], [16], and [51] which applies the adversarial learning in GANs for cross-modal retrieval. Specifically, the CMA designs an adversarial training style to learn a generator that generates common representations for image and text instances, and a discriminator that cannot reliably predict the modality of the common representations. Here in the SNet, we consider the shared encoders $E_s^v$ and $E_s^t$ as the generator, and we further configure the discriminator with a special unit of gradient reversal layer (GRL) [48]. During the forward propagation, the GRL unit in the SNet performs identity transform, while reverses the gradient direction by a negative coefficient during the backpropagation for adversarial training.

In the training stage, each instance (image $\mathbf{v}_i$ or text $\mathbf{t}_i$) is assigned with a ground-truth modality label $z_i^* \in \{0, 1\}$, $* = v, t$. The discriminator maps the common representations ($\mathbf{h}_s^i$ or $\mathbf{g}_s^i$) to a prediction score (e.g., sigmoid activation output) of the label $\hat{z}_i^* \in \{0, 1\}$. Essentially, we *maximize* the cross-entropy loss for the modality prediction task as

$$L_{\text{sim}}^{\text{CMA}} = -\frac{1}{N}\sum_{*=v,t}\sum_{i=1}^{N} f_{\text{sigmoid}}\left(z_i^*, \hat{z}_i^*\right) \tag{3}$$

where $f_{\text{sigmoid}}(x, \hat{x}) = -(x \log \hat{x} + (1-x)\log(1-\hat{x}))$ denotes the sigmoid cross-entropy loss function. Maximizing the loss in (3) will reduce the heterogeneity gap of the common representations across different modalities.

Cross-modal maximum mean discrepancy (CMMD) extends the feature adaptation method [54] on data in the same media type to multimodal data. For two inconsistent distributions $P(X)$ and $Q(Y)$ with their embeddings $\mu_X$ and $\mu_Y$, respectively, the original maximum mean discrepancy (MMD) measures the square distance between the embeddings of the two distributions in the reproducing kernel Hilbert space (RKHS). If we consider the common representations of the two modalities span two distributions, the target of the proposed CMMD scheme is to maximize the mean discrepancy of the common representations of pairwise instances of different modalities in the shared subspace.

Let $\mathbf{h}_s^i$, $\mathbf{g}_s^i$, $\mathbf{h}_s^j$, and $\mathbf{g}_s^j$, respectively, denote the common representations of the $i$th and $j$th image and text instances, then the cross-modal similarity-based CMMD for all instances can be derived as a kernel-based distance function between the common representations of pairwise instances

$$L_{\text{sim}}^{\text{CMMD}} = \frac{1}{N^2}\sum_{i,j=1}^{N}\kappa\left(\mathbf{h}_s^i, \mathbf{h}_s^j\right) - \frac{2}{N^2}\sum_{i,j=1}^{N}\kappa\left(\mathbf{h}_s^i, \mathbf{g}_s^j\right) + \frac{1}{N^2}\sum_{i,j=1}^{N}\kappa\left(\mathbf{g}_s^i, \mathbf{g}_s^j\right) \tag{4}$$

where $\kappa(x_i, x_j) = \sum_n \eta_n \exp\{-[1/(2\sigma_n)]\|x_i - x_j\|^2\}$ is the linear combination of multiple radial basis function (RBF) kernels. The weight for the $n$th RBF kernel is denoted as $\eta_n$, and $\sigma_n$ is the standard deviation. It is notable that the RBF kernel is a typical positive-semidefinite kernel that used in the MMD term for kernel embeddings [47], [51], [52]. As a single RBF kernel is a basic Gaussian function, it is well known that the Taylor expansion of the Gaussian function can exhibit a wide and complex distribution of real data. Therefore, the different components of the multiple RBF kernel are more advanced to ensure that the MMD loss value more accurately reveals the discrepancy of the two distributions. By minimizing (4), the shared structures of instances in different modalities can be effectively modeled in their common representations.

Cross-modal correlation alignment (CMCA) is developed to measure the distance between the covariance of the common representations of the instances in different modalities. Unlike the above three schemes that model the correlation of pairwise instances, here it explores the overall statistical property of all instances. Let $\mathbf{H}_s \in \mathbb{R}^{N\times m}$ and $\mathbf{G}_s \in \mathbb{R}^{N\times m}$ be matrices whose rows are the common representations $\{\mathbf{h}_s^i\}_{i=1}^N$ and $\{\mathbf{g}_s^i\}_{i=1}^N$ for image and text instances, where the two matrices have zero mean with $L_2$ normalization. The covariances of $\mathbf{H}_s$ and $\mathbf{G}_s$ are defined as

$$C_s^v = \frac{1}{N-1}\left(\mathbf{H}_s^\top \mathbf{H}_s - \frac{1}{N}\left(\mathbf{1}^\top \mathbf{H}_s\right)^\top \left(\mathbf{1}^\top \mathbf{H}_s\right)\right) \tag{5}$$

$$C_s^t = \frac{1}{N-1}\left(\mathbf{G}_s^\top \mathbf{G}_s - \frac{1}{N}\left(\mathbf{1}^\top \mathbf{G}_s\right)^\top \left(\mathbf{1}^\top \mathbf{G}_s\right)\right) \tag{6}$$

where $\mathbf{1}$ is an $m$-dimensional vector with all elements equals 1. Then, the cross-modal similarity based on CMCA is to align the different distributions by minimizing

$$\mathcal{L}_{\text{sim}}^{\text{CMCA}} = \frac{1}{4m^2}\left\|C_s^v - C_s^t\right\|_F^2 \tag{7}$$

where $\|\cdot\|_F^2$ is the squared Frobenius norm for matrices. Therefore, the second-order statistics of the distributions of instances in different modalities can be modeled based on their common representations.

*2) Preserving the Discriminability in the Shared Space:* The above branch in the SNet addresses the pairwise correlations between different modalities. However, they cannot explicitly capture the class information of the available semantic labels to make the common representations semantically discriminative. To this end, we add a classifier after the shared space to predict the semantic labels of the image or text instances projected in the subspace. Specifically, the classifier contains fully connected layers with the softmax loss function. Then, preserving the discriminability of the common representations can be considered as a classification task of predicting their semantic labels, and the classification loss is defined as

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N}\left[f_{\text{softmax}}\left(\mathbf{h}_s^i, y_i\right) + f_{\text{softmax}}\left(\mathbf{g}_s^i, y_i\right)\right]. \tag{8}$$

Here, $y_i$ is the semantic label for the $i$th image and text instances, and $f_{\text{softmax}}$ is the softmax loss function

$$f_{\text{softmax}}(x, y) = \sum_{c=1}^{C} 1\{y = c\}\log\left[\hat{p}(x, c)\right] \tag{9}$$

where $y$ denotes the label of the instance $x$, and $C$ is the total number of classes. The function $\hat{p}(x, c)$ denotes the softmax probability distribution over $C$ classes of $x$ as $\hat{p}(x, c) = \exp(\psi_c(x))/\sum_{l=1}^{C}\exp(\psi_l(x))$, where $\psi_l$ has the network parameters for the $l$th classifier. We can maximize the classification accuracy jointly for the instances in two modalities by optimizing (8). As a result, it would ensure to obtain semantically discriminative and consistent common representations for all the instances in two modalities.

### D. Learning Private Subspaces in PNet-Img and PNet-Txt

As discussed in Section III-B, in the proposed P3S, both the PNet-Img and PNet-Txt adopt the network structure of autoencoder, which is basically formed by an encoder and a decoder. In addition to the two shared encoders $E_s^v$ and $E_s^t$ in the SNet that explore the shared structures underlying the heterogeneous data, two private encoders $E_p^v$ and $E_p^t$ are separately equipped in the two PNets to capture the interference of the individual modality. For an image–text pair $o_i = \{\mathbf{v}_i, \mathbf{t}_i\}$, their private representations $\mathbf{h}_p^i \in \mathbb{R}^m$ and $\mathbf{g}_p^i \in \mathbb{R}^m$ in the private subspaces are generated by the respective private encoder as

$$\mathbf{h}_p^i = E_p^v(\mathbf{v}_i); \quad \mathbf{g}_p^i = E_p^t(\mathbf{t}_i). \tag{10}$$

The decoders $D^v$ and $D^t$ in the two PNets ensure the common representations in the shared space and the private representations in the private space are combined to

reconstruct the input representations as

$$\hat{\mathbf{v}}_i = D^v\left(\mathbf{h}_p^i + \mathbf{h}_s^i\right); \quad \hat{\mathbf{t}}_i = D^t\left(\mathbf{g}_p^i + \mathbf{g}_s^i\right) \tag{11}$$

where $\hat{\mathbf{v}}_i$ and $\hat{\mathbf{t}}_i$ are the reconstruction of the inputs $\mathbf{v}_i$ and $\mathbf{t}_i$, and $+$ denotes the addition operation for two vectors.[1] Both $D^v$ and $D^t$ consist of several fully connected layers.

Comparing with the existing method [16], [17] that only takes the common representations for reconstruction, here the combination of both the common representations and the private representations enhances the robustness of the reconstruction results. Specifically, we calculate the reconstruction loss in the two PNets as

$$\mathcal{L}_{\text{rec}} = \sum_{i=1}^{N}\left[f_{\text{recon}}\left(\mathbf{v}_i, \hat{\mathbf{v}}_i\right) + f_{\text{recon}}\left(\mathbf{t}_i, \hat{\mathbf{t}}_i\right)\right]. \tag{12}$$

Here, $f_{\text{recon}}$ is a scale-invariant mean-square error function defined as

$$f_{\text{recon}}\left(\mathbf{x}, \hat{\mathbf{x}}\right) = \frac{1}{k}\left\|\mathbf{x} - \hat{\mathbf{x}}\right\|_2^2 - \frac{1}{k^2}\left(\left[\mathbf{x} - \hat{\mathbf{x}}\right] \cdot \mathbf{1}_k\right)^2 \tag{13}$$

where $k$ is the dimension of the input vector $x$, $\mathbf{1}_k$ denotes a vector of ones with length $k$, and $\|\cdot\|_2^2$ is the squared $L_2$ norm. Unlike the traditional mean-squared error loss used in [16], [17], [19], and [30] that relies on a scaling term to penalize the vectors of correct predictions, the function $f_{\text{recon}}$ here is advantaged to compensate for the difference between the elements in two vectors, which allows to reconstruct the overall structure of the data. As a result, the semantic consistency within each modality can be well preserved upon the high-level semantics.

### *E. Subspaces Separation*

In each modality, to encourage the two encoders (shared and private) to encode different aspects of the input representations of the instances, we explicitly separate the shared space and the private space during the learning process. Specifically, we employ the orthogonality constraints between the common and private representations of the instances in each modality to address the difference between the shared and private subspaces. Similar as the definitions of the aforementioned $\mathbf{H}_s$ and $\mathbf{G}_s$, we denote $\mathbf{H}_p = \{\mathbf{h}_p^i\}_{i=1}^N \in \mathbb{R}^{N\times m}$ and $\mathbf{G}_p = \{\mathbf{g}_p^i\}_{i=1}^N \in \mathbb{R}^{N\times m}$ as the matrices whose rows are the private representations for the instances of image and text modality, respectively. Then, the difference loss encourages orthogonality between the representations in the private and shared spaces of the two modalities, which is formulated as

$$\mathcal{L}_{\text{dif}} = \left\|\mathbf{H}_s{}^\top \mathbf{H}_p\right\|_F^2 + \left\|\mathbf{G}_s{}^\top \mathbf{G}_p\right\|_F^2. \tag{14}$$

The constraints in (14) avoid trivial solutions for the model that produce useless representations without separation.

[1]We also evaluate the concatenation operation for the vectors, that is, $[\mathbf{h}_p^i, \mathbf{h}_s^i]$ and $[\mathbf{g}_p^i, \mathbf{g}_s^i]$. However, it has no significant difference with the simple addition.

### *F. Optimization*

To clearly derive the optimization procedure in the proposed P3S method, we first simplify the notations of the parameters in each subnetwork. Specifically, in the SNet, we denote the network parameters of the two shared encoders as $\theta_s$ and the network parameters of the classifier as $\theta_c$. When using the CMA scheme for the $\mathcal{L}_{\text{sim}}$, there are additional parameters $\theta_a$ in the SNet because of the introduced modality classifier in (3). In the two PNets, the parameters of the two private encoders are $\theta_p$ and the two decoders are $\theta_d$. Then, we can correspond the parameters to the specific loss function and formally derive the overall loss function of P3S as

$$\begin{aligned}\mathcal{L}_{\text{all}}\left(\theta_s, \theta_c, \theta_p, \theta_d, \theta_a\right) = \mathcal{L}_{\text{cls}}(\theta_s, \theta_c) + \alpha\mathcal{L}_{\text{rec}}\left(\theta_s, \theta_p, \theta_d\right) \\ + \beta\mathcal{L}_{\text{dif}}\left(\theta_s, \theta_p\right) + \gamma\mathcal{L}_{\text{sim}}(\theta_s, \theta_a)\end{aligned} \tag{15}$$

where $\alpha$, $\beta$, and $\gamma$ are positive weighted coefficiencies that balance each loss term. Our goal is to find the saddle points of (15) for the minimization problem as

$$\left(\hat{\theta}_s, \hat{\theta}_c, \hat{\theta}_p, \hat{\theta}_d\right) = \arg\min_{\theta_s,\theta_c,\theta_p,\theta_d} \mathcal{L}_{\text{all}}\left(\theta_s, \theta_c, \theta_p, \theta_d\right). \tag{16}$$

Note that $\gamma$ in (15) becomes a negative value when using the CMA scheme for the $\mathcal{L}_{\text{sim}}$, as the loss term in (3) needs to be maximized. Then in this case, the solution for (15) turns to a standard minimax optimization task under the adversarial training style as in [7], [16], and [48], that is, finding the saddle points of $\hat{\theta}_s$, $\hat{\theta}_c$, $\hat{\theta}_p$, and $\hat{\theta}_d$ by minimizing (16), while searching for the saddle point of $\theta_a$ by maximizing

$$\hat{\theta}_a = \arg\max_{\theta_a} \mathcal{L}_{\text{sim}}\left(\hat{\theta}_s, \theta_a\right). \tag{17}$$

In practice, we compute the gradient $\nabla_{\theta_*}$ of each parameter and adopt the sophisticated algorithm of stochastic gradient descent (SGD) to update each parameter alternatively. Specifically, given a predefined learning rate $\mu$, the updating rules for the parameters $\theta_p$, $\theta_d$, $\theta_c$, $\theta_s$, and $\theta_a$ are derived as follows:

$$\theta_p \leftarrow \theta_p - \mu\left(\nabla_{\theta_p}\mathcal{L}_{\text{rec}} + \beta\nabla_{\theta_p}\mathcal{L}_{\text{diff}}\right) \tag{18}$$

$$\theta_d \leftarrow \theta_d - \mu\left(\alpha\nabla_{\theta_d}\mathcal{L}_{\text{rec}}\right) \tag{19}$$

$$\theta_c \leftarrow \theta_c - \mu\left(\nabla_{\theta_c}\mathcal{L}_{\text{cls}}\right) \tag{20}$$

$$\theta_s \leftarrow \theta_s - \mu\left(\nabla_{\theta_s}\mathcal{L}_{\text{cls}} + \alpha\nabla_{\theta_s}\mathcal{L}_{\text{rec}} + \beta\nabla_{\theta_s}\mathcal{L}_{\text{diff}} + \gamma\nabla_{\theta_s}\mathcal{L}_{\text{sim}}\right) \tag{21}$$

$$\theta_a \leftarrow \theta_a - \mu\left(\gamma\nabla_{\theta_a}\mathcal{L}_{\text{sim}}\right). \tag{22}$$

The training procedure of the proposed P3S method is summarized in Algorithm 1. For the testing state, after the shared encoders $E_s^v$ and $E_s^t$ in the proposed P3S method are learned, they are used to produce the corresponding common representations $\{\mathbf{h}_s'^q\}_{q=1}^{N'}$ and $\{\mathbf{g}_s'^q\}_{q=1}^{N'}$ for the images $\{\mathbf{v}_q'\}_{q=1}^{N'}$ and texts $\{\mathbf{t}_q'\}_{q=1}^{N'}$ in the testing set $\mathcal{O}_{te}$. Finally, cross-modal retrieval can be performed by calculating the cosine similarities of the common representations $\{\mathbf{h}_s'^q\}_{q=1}^{N'}$ and $\{\mathbf{g}_s'^q\}_{q=1}^{N'}$ of the pairwise image–text instances.

**Algorithm 1** Training Procedure of the Proposed P3S

**Input:** Training instances $\{(\mathbf{v}_i, \mathbf{t}_i, y_i)\}_{i=1}^{N}$, batch size $B$, the number of training iterations $T$, hyper-parameters $\alpha$, $\beta$, $\gamma$ and learning rate $\mu$.

**Output:** Model parameters $\hat{\theta}_s$, $\hat{\theta}_c$, $\hat{\theta}_p$, $\hat{\theta}_d$, $\hat{\theta}_a$.

**repeat**
  Sample image–text pair $\{\mathbf{v}_b, \mathbf{t}_b\}_{b=1}^{B}$ with batch.
  Update $\theta_p$ according to Eq. (18).
  Update $\theta_d$ according to Eq. (19).
  Update $\theta_c$ according to Eq. (20).
  Update $\theta_s$ according to Eq. (21).
  Update $\theta_a$ according to Eq. (22).
**until** Objective function of Eq. (15) converges or reaches the maximum iterations.
The shared encoders $E_s^v$ and $E_s^t$ that generate common representations for image and text instances, respectively.

## IV. Experiments

In this section, to verify the effectiveness of our P3S approach, we take 15 state-of-the-art methods as counterparts and conduct experiments for cross-modal retrieval on four widely used cross-modal datasets. We will first introduce the experimental configurations, and then depict the extensive experiments, including baseline comparison, convergence, and parameter sensitiveness analysis, to comprehensively verify the contribution of each component in the proposed P3S approach.

### A. Experimental Configuration

*1) Datasets and Features:* In our experiments, the selected datasets include three widely used benchmark datasets, that is, Wikipedia, Pascal Sentences, and NUS-WIDE-10k, as well as a large-scale PKU-XMediaNet dataset. Here, we briefly introduce the statistical information of each dataset as follows.

1) *Wikipedia* [8] is the most widely used dataset for cross-modal retrieval, which contains 2866 image–text pairs collected from the Wikipedia website. Each pair only belongs to one of ten high-level semantic categories. We adopt the data partition protocol of [14], [16], and [17] to divide the dataset into a training set of 2173 pairs, a validation set of 231 pairs, and a test set of 462 pairs.
2) *Pascal Sentences* [55] contains 1000 image–text pairs with 20 categories which are generated from PASCAL 2008 development kit. Each image is annotated by five related sentences that form one text document. Following [14], [16], and [17], we select 800 pairs for training, while 100 pairs are for validation and the remaining 100 pairs are for testing.
3) *NUS-WIDE-10k* [56] is a subset cropped from the original NUS-WIDE [57] dataset containing around 270 000 image–text pairs from 81 classes. This subset has over 70 000 pairs, which is collected by choosing the images along with their tags that exclusively come from the largest ten categories.
4) *PKU-XMediaNet* is a newly constructed large-scale multimodal dataset in [58], which contains data of five modalities, that is, text, image, audio, video, and 3-D model. Similar as [11] and [16], we select the media data of images and texts in this dataset to conduct our experiments, where textual descriptions are extracted from Wikipedia articles. Finally, we have 40 000 image–text pairs from 200 categories in total.

For the feature representations, we follow the latest studies [11], [16], [59] to extract convolutional neural-network (CNN) features for the instances in both image and text modalities. In particular, each image is represented by a 4096-D feature vector encoded by the fc7 layer of the 19-layer VGGNet [20], and the feature vector for each text is extracted from the WordCNN [21] with 300 dimensions.

*2) Implementation Details:* For the settings of the network architecture, we build the (shared and private) encoders in SNet and two PNets with three fully connected layers ($d_* \rightarrow 4096 \rightarrow 3072 \rightarrow m$, $* = v, t$) with each layer following the tanh activation function. Here, $m$ is the dimension of the common/private representations in the shared/private subspaces, respectively. The decoders in two PNets have the symmetrical structure as the encoders, with only the dimensions of the three layers being reversed, that is, $m \rightarrow 3072 \rightarrow 4096 \rightarrow d_*$, $* = v, t$. For the classifier in the SNet, we build three fully connected layers ($m \rightarrow 64 \rightarrow 64 \rightarrow C$) and add the softmax activation to the last layer. The Adam optimizer [60] is utilized for training all subnetworks jointly, the initial learning rate $\mu$ is set to 0.001 with the weight decay rate as 0.9, and the mini-batch size is set to 64. We employ the Tensorflow [61] toolkit to implement our P3S method and all the experiments are conducted on a desktop machine with one GeForce GTX 1080 Ti GPU.

As the proposed CMMD scheme in the proposed P3S method utilizes the multiple RBF kernels, we empirically set the parameters weight $\eta_n$ and the standard deviation $\sigma_n$ according to the official implementation of the MMD loss in Tensorflow toolkit.[2] Specifically, follow the default setting in the toolkit, we use 19 RBF kernels in total, where their standard deviation parameters in $[10^{-6}, 10^{-5}, \ldots, 10^{5}, 10^{6}]$, and the weight of each kernel is set to $(1/19)$. The hyperparameters $\alpha$, $\beta$, and $\gamma$ and the dimension $m$ of the shared and private spaces are tuned on all datasets for both the baseline experiments (Section IV-B) and the overall experiments (Section IV-C). A sensitivity analysis of the hyperparameters is provided in Fig. 6.

*3) Evaluation Metrics and Compared Methods:* The proposed P3S approach is compared with 15 state-of-the-art methods, including five traditional methods, namely, CCA [8], CFA [25], KCCA [22], JRL [28], and LGCFL [62]; and ten DNN-based methods, namely, Corr-AE [17], DCCA [31], CMDH [14], DeepSM [63], ACMR [7], CCL [15], CBT [59], MCSM [11], CM-GAN [16], and TANSS [12]. For a fair comparison, all the compared methods take the same features for both image and text modalities as our P3S approach. Specifically, the learned shallow and deep models of all the methods are used to convert the images $\{\mathbf{v}'_q\}_{q=1}^{N'}$ and texts

[2]The toolkit can be found at https://github.com/tensorflow/models/tree/master/research/domain_adaptation.

TABLE I
COMPARISON OF THE RETRIEVAL RESULTS USING DIFFERENT SCHEMES FOR $\mathcal{L}_{\text{sim}}$ IN OUR P3S METHOD ON ALL DATASETS

| Different $\mathcal{L}_{sim}$ | Wikipedia | | | Pascal Sentences | | | NUS-WIDE-10k | | | PKU-XMediaNet | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Img2Txt | Txt2Img | Avg. | Img2Txt | Txt2Img | Avg. | Img2Txt | Txt2Img | Avg. | Img2Txt | Txt2Img | Avg. |
| $\mathcal{L}_{sim}^{CMCS}$ | 0.508 | 0.452 | 0.480 | 0.597 | 0.593 | 0.595 | 0.535 | 0.561 | 0.548 | 0.643 | 0.644 | 0.644 |
| $\mathcal{L}_{sim}^{CMA}$ | 0.499 | 0.447 | 0.473 | 0.610 | 0.596 | 0.603 | 0.531 | 0.562 | 0.546 | 0.634 | 0.635 | 0.634 |
| $\mathcal{L}_{sim}^{CMMD}$ | 0.512 | 0.452 | 0.482 | 0.614 | 0.590 | 0.602 | 0.534 | 0.561 | 0.548 | 0.642 | 0.629 | 0.631 |
| $\mathcal{L}_{sim}^{CMCA}$ | **0.520** | 0.469 | **0.495** | **0.622** | **0.601** | **0.612** | **0.545** | **0.574** | **0.560** | 0.647 | 0.648 | 0.648 |
| $\mathcal{L}_{sim}^{CMCS} + \mathcal{L}_{sim}^{CMCA}$ | 0.519 | 0.465 | 0.492 | 0.608 | 0.593 | 0.601 | 0.542 | 0.561 | 0.552 | 0.642 | 0.645 | 0.644 |
| $\mathcal{L}_{sim}^{CMA} + \mathcal{L}_{sim}^{CMCA}$ | 0.514 | 0.461 | 0.488 | 0.615 | 0.589 | 0.602 | 0.539 | 0.564 | 0.552 | 0.641 | 0.645 | 0.643 |
| $\mathcal{L}_{sim}^{CMMD} + \mathcal{L}_{sim}^{CMCA}$ | 0.512 | **0.476** | 0.494 | 0.615 | 0.590 | 0.603 | 0.542 | 0.559 | 0.551 | **0.658** | **0.657** | **0.658** |

TABLE II
COMPARISON OF THE RETRIEVAL RESULTS OF OUR P3S METHOD AND ITS FIVE BASELINES ON ALL DATASETS

| Baselines | Wikipedia | | | Pascal Sentences | | | NUS-WIDE-10k | | | XMediaNet | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Img2Txt | Txt2Img | Avg. | Img2Txt | Txt2Img | Avg. | Img2Txt | Txt2Img | Avg. | Img2Txt | Txt2Img | Avg. |
| P3S-B1 ($\mathcal{L}_{cls}$) | 0.423 | 0.345 | 0.384 | 0.509 | 0.493 | 0.501 | 0.435 | 0.443 | 0.439 | 0.217 | 0.219 | 0.218 |
| P3S-B2 ($\mathcal{L}_{sim}$, $\mathcal{L}_{dif}$) | 0.501 | 0.449 | 0.475 | 0.600 | 0.581 | 0.591 | 0.504 | 0.532 | 0.518 | 0.557 | 0.554 | 0.556 |
| P3S-B3 ($\mathcal{L}_{sim}$) | 0.500 | 0.455 | 0.478 | 0.594 | 0.578 | 0.586 | 0.504 | 0.539 | 0.521 | 0.615 | 0.614 | 0.614 |
| P3S-B4 ($\mathcal{L}_{dif}$) | 0.502 | 0.464 | 0.483 | 0.577 | 0.571 | 0.574 | 0.519 | 0.543 | 0.531 | 0.612 | 0.611 | 0.612 |
| P3S-B5 ($\mathcal{L}_{rec}$) | 0.517 | 0.467 | 0.492 | 0.589 | 0.574 | 0.582 | 0.539 | 0.567 | 0.553 | 0.592 | 0.592 | 0.592 |
| P3S ($\mathcal{L}_{all}$) | **0.520** | **0.469** | **0.495** | **0.622** | **0.601** | **0.612** | **0.545** | **0.574** | **0.560** | **0.647** | **0.648** | **0.648** |

$\{\mathbf{t}'_q\}_{q=1}^{N'}$ in the testing set $\mathcal{O}_{te}$ into their corresponding common representations $\{\mathbf{h}_s^{\prime q}\}_{q=1}^{N'}$ and $\{\mathbf{g}_s^{\prime q}\}_{q=1}^{N'}$. The cross-modal retrieval is performed by ranking the cosine similarities of the common representations of the pairwise image–text instances.

In order to objectively verify the retrieval performance, two typical retrieval tasks are performed on all datasets: 1) *Img2Txt* that takes a query of an image as a query to retrieve relevant texts from test set and 2) *Txt2Img* that uses a query of text to retrieve relevant image candidates from the test set. The widely used mean average precision (MAP) score is adopted as the evaluation metric on the two retrieval tasks. In particular, MAP is computed as the mean value of the average precision scores of all the queries. Larger MAP score indicates better cross-modal retrieval performance. It should be noted that the original experiments in Corr-AE [17] and ACMR [7] report the MAP scores top 50 returned results, while the rest returned results are not considered. In our experiments, for these two compared methods, we fairly report their results of all the returned results as the other counterparts, Moreover, we also plot the precision–recall (PR) curves that elaborately assess the retrieval results under different conditions.

### *B. Baseline Experiments on P3S*

As our P3S approach consists of three subnetworks with various components, to explicitly investigate the effectiveness of the components and subnetworks, in the following experiment, we first dig inside the P3S and conduct two kinds of baseline experiments as follows. According to the objective function of P3S in (15), to choose the optimal hyperparameters of the loss terms in each baseline, we follow the parameter tuning procedure described in Section IV-D2, to ensure that the hyperparameters in each baseline are effectively tuned.

*1) Effect of Different $\mathcal{L}_{sim}$:* In Section III, we have derived four schemes of $\mathcal{L}_{\text{sim}}$ to measure the cross-modal similarity in the SNet. In this baseline experiment, we compare the effect of the different $\mathcal{L}_{\text{sim}}$ used in our P3S approach.

Table I shows the retrieval performance of our P3S with different schemes for $\mathcal{L}_{\text{sim}}$ on all datasets, where the best results are highlighted in bold font. We first discuss the effect of the four individual schemes $\mathcal{L}_{\text{sim}}^{\text{CMCS}}$, $\mathcal{L}_{\text{sim}}^{\text{CMA}}$, $\mathcal{L}_{\text{sim}}^{\text{CMMD}}$, and $\mathcal{L}_{\text{sim}}^{\text{CMCA}}$. We can observe that using $\mathcal{L}_{\text{sim}}^{\text{CMA}}$ obtains much inferior performance on all datasets comparing with other three baselines, showing that the adversarial learning may not be appropriate to be applied on our P3S. The reason is that the adversarial learning scheme is to find a common subspace that mixes the modality-invariant features across different modalities, and the specific properties in each modality. It may be a conflict with the private–shared subspace separation scheme used in our P3S. As our separation scheme aims to extract the intramodal distinctions in the private subspace of each modality, while preserving the shared structures of different modalities in the shared subspace. We also observe that using $\mathcal{L}_{\text{sim}}^{\text{CMCA}}$ obtains the best performance for P3S on all datasets, indicating that in the shared subspace, aligning the common representations of different modalities are more advantaged than simply measuring their Euclidean distances in $\mathcal{L}_{\text{sim}}^{\text{CMCS}}$ or maximizing the mean discrepancy of the inconsistent representation in $\mathcal{L}_{\text{sim}}^{\text{CMMD}}$.

As the proposed $\mathcal{L}_{\text{sim}}^{\text{CMCA}}$ scheme shows more effective performance among the four schemes, we further investigate the effect of combining it with the other three schemes in the proposed P3S method. The last three rows in Table II show the results of the integrations of $\mathcal{L}_{\text{sim}}^{\text{CMCA}}$ and one of the other three schemes, respectively. We can see that the results of the three combinative baselines are diverse on different datasets. Generally, the two combinations of $\mathcal{L}_{\text{sim}}^{\text{CMCS}} + \mathcal{L}_{\text{sim}}^{\text{CMCA}}$ and $\mathcal{L}_{\text{sim}}^{\text{CMA}} + \mathcal{L}_{\text{sim}}^{\text{CMCA}}$ obtain inferior performance compared to the individual $\mathcal{L}_{\text{sim}}^{\text{CMCA}}$ on all datasets. As these two schemes

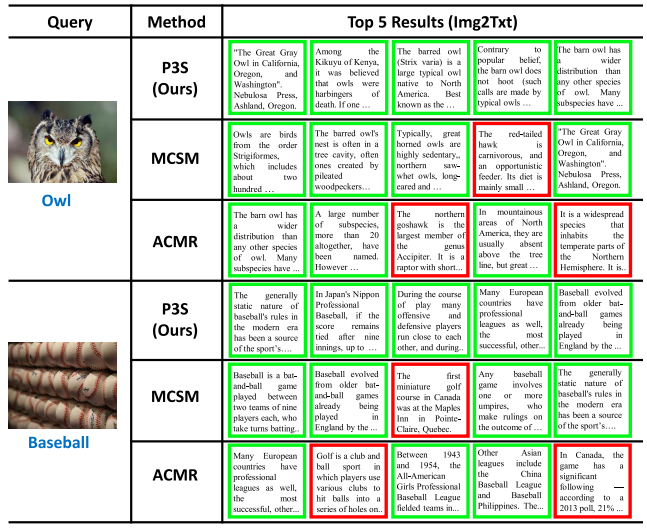



Fig. 3. Examples of the (a) Img2Txt and (b) Txt2Img retrieval results on the PKU-XMediaNet dataset by our P3S approach and two compared methods MSCM [11] and ACMR [7]. The ground-truth semantic label for each query is attached for more informative comparison. Besides, the true matches and the incorrect retrieval results are, respectively, marked by green and red rectangles.

are designed to measure the cross-modal similarity on the pairwise instance level, which may not be compatible to the proposed CMCA scheme for the overall data distribution of all instances. Nevertheless, the performance of $\mathcal{L}_{\text{sim}}^{\text{CMMD}} + \mathcal{L}_{\text{sim}}^{\text{CMCA}}$ seems to be better than the $\mathcal{L}_{\text{sim}}^{\text{CMCA}}$ on a specific task (Img2Txt or Txt2Img) on Wikipedia and PKU-XMediaNet, while being worse on the other two datasets. Though both the two schemes measure the cross-modal similarity on the overall distribution of all instances, they may have an intrinsic difference on estimating the distributions of different modality data. Therefore, we can conclude that due to the potential incompatibility of the four schemes, the combinations of these schemes may not obtain stable and consistent results on all datasets. Based on the above observations, in all the latter experiments, we still use an individual scheme of the more advanced $\mathcal{L}_{\text{sim}}^{\text{CMCA}}$ for the cross-modal similarity learning in the SNet of our P3S method.

*2) Effect of Each Loss Term:* According to the final objective function $\mathcal{L}_{\text{all}}$ in (15), the proposed P3S contains four loss terms, which correspond to different components in the three subnetworks. To investigate the impact of each loss term, we design five variants as the baselines of the P3S by excluding one or two loss terms in (15) during the training procedure. Table II shows the retrieval performance of the original P3S as well as its five baselines on all datasets. Here, $\not{\mathcal{L}}_{*}$ indicates the specific loss term that is excluded from the $\mathcal{L}_{\text{all}}$ when training the baselines. As aforementioned, we use the prototype of $\mathcal{L}_{\text{sim}}^{\text{CMCA}}$ for the loss term $\mathcal{L}_{\text{sim}}$.

The following observations can be obtained from the comparison results in Table II.

1) The baseline P3S-B1 obtains the worst performance among all the baselines, as it is trained in an unsupervised manner without the supervision of the semantic label information. In other words, it fails to learn an effective shared subspace in the SNet that generates less discriminative common representations for different modality data, hence deteriorates the retrieval performance.
2) When excluding the $\mathcal{L}_{\text{sim}}$ and $\mathcal{L}_{\text{dif}}$ in the SNet, the baseline P3S-B2 obtains second-worst performance which shows that capturing the cross-modal similarity in the shared subspace, as well as separating the difference between the shared subspace and the private spaces are two important factors for common representation learning.
3) The retrieval performance of the PS3-B3 is on par with the result of the P3S-B4, showing that explicitly separating the shared subspace and the private subspaces are significant to learn more effective shared subspace. In addition, it facilitates to produce more compact common representations that fully incorporate the shared structures in different modalities.
4) The performance of the P3S-B5 varies depending on the specific dataset. Nevertheless, the reconstruction procedure is a useful complementary that preserves the semantic consistency of the input and the reconstructed representation, and enhances the robustness of the learned shared subspace.
5) The P3S method with all loss terms achieves the best performance comparing with all baselines, as it well balances the importance of different loss terms and learns more effective common representations for retrieval.

### C. Comparisons With the State-of-the-Art Methods

In Table III, we further present the overall comparison on the two retrieval tasks of our P3S method and the 15 state-of-the-art approaches, including both the traditional methods (on the top panel) and the DNN-based methods (on the middle panel). First, when considering the first three datasets of small size, the DNN-based methods are advantageous in capturing the nonlinear correlation across different modality data, hence generally outperforming traditional methods. Among the

TABLE III
MAP SCORES OF CROSS-MODAL RETRIEVAL FOR OUR P3S APPROACH AND OTHER COMPARED METHODS ON ALL DATASETS

| Methods | Wikipedia | | | Pascal Sentences | | | NUS-WIDE-10k | | | PKU-XMediaNet | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Img2Txt | Txt2Img | Avg. | Img2Txt | Txt2Img | Avg. | Img2Txt | Txt2Img | Avg. | Img2Txt | Txt2Img | Avg. |
| CCA [8] (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 [25] (2003) | 0.319 | 0.316 | 0.318 | 0.476 | 0.47 | 0.473 | 0.406 | 0.435 | 0.421 | 0.252 | 0.4 | 0.326 |
| KCCA [22] (2004) | 0.438 | 0.389 | 0.414 | 0.488 | 0.446 | 0.467 | 0.351 | 0.356 | 0.354 | 0.252 | 0.27 | 0.261 |
| JRL [28] (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 [62] (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 [31] (2013) | 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 [17] (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 [14] (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 |
| DeepSM [63] (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 [7] (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 [15] (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 [59] (2018) | 0.516 | 0.464 | 0.490 | 0.602 | 0.583 | 0.593 | 0.522 | 0.550 | 0.536 | 0.577 | 0.575 | 0.576 |
| MCSM [11] (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 |
| TANSS [12] (2019) | 0.518 | 0.459 | 0.488 | 0.608 | 0.594 | 0.601 | 0.539 | 0.551 | 0.545 | 0.582 | 0.574 | 0.578 |
| CM-GAN [16] (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 |
| **P3S (Ours)** | 0.520 | **0.469** | **0.495** | **0.622** | 0.601 | **0.612** | **0.545** | **0.574** | **0.560** | **0.647** | **0.648** | **0.648** |

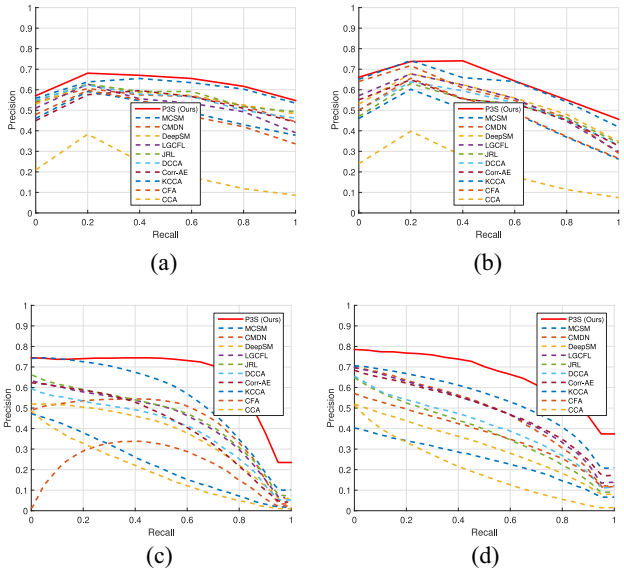



Fig. 4. PR curves of our P3S method and the counterparts for both Img2Txt and Txt2Img tasks on the Pascal Sentences and PKU-XMediaNet datasets. (a) Img2Txt on Pascal Sentences. (b) Txt2Img on Pascal Sentences. (c) Img2Txt on PKU-XMediaNet. (d) Txt2Img on PKU-XMediaNet.

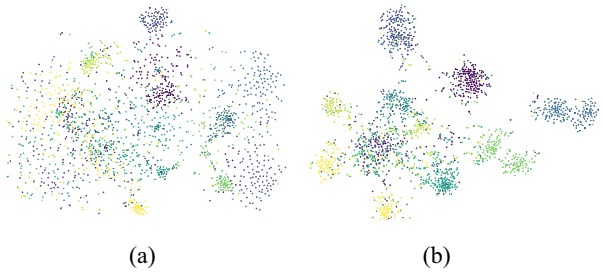



Fig. 5. t-SNE visualizations of the private and the shared subspaces learned in our P3S method on the NUS-WIDE-10k dataset. (a) Learned private subspaces. (b) Learned shared subspace.

DNN-based methods, our P3S obtains the highest averaged MAP scores on all datasets. Specifically, though the compared method CM-GAN obtains a slightly better Img2Txt score on the Wikipedia dataset and Txt2Img score on the Pascal Sentences dataset than our P3S method, it still performs inferior to P3S on the other results. The CM-GAN method employs two coupled GANs as the generative model to estimate the joint distribution of different modalities in a common space. On the contrary, our P3S method introduces two different spaces to explicitly capture the correlated information across different modalities, as well as the intrinsic properties within each modality. Therefore, the proposed P3S method shows the importance of learning more compact common representations, which is another promising aspect for cross-modal retrieval.

Second, when turning to the large-scale PKU-XMediaNet dataset, several DNN-based methods perform inferior to the traditional methods. In particular, the average MAP scores of the DNN-based methods DCCA and DeepSM are clearly worse than the traditional methods JRL and LGCFL. The reason is that the latter two methods utilize advanced constraints, such as sparse regularization and group structures. It is notable that the performance of the compared method CM-GAN remarkably drops on the PKU-XMediaNet dataset comparing to its performance on the three small datasets, indicating that it may not generalize well on real scenario due to the complicate GAN structures in it. Nevertheless, our P3S approach consistently outperforms all the compared methods, showing its effectiveness and generalization capability.

Comparing with all the counterparts, the best overall performance achieved by our P3S method achieves can be attributed to the following aspects.

1) Explicitly separating the shared subspace and the private subspaces in different modalities is advantaged to learn more compact shared subspace that generate more effective common representations for different modality data.
2) The CMCA scheme utilized in the SNet is advanced to explore the shared structures in heterogeneous data distributions, and effectively incorporate the correlations in the learned shared subspace.
3) The P3S method effectively combines the four different loss terms that fully address important factors for cross-modal retrieval, including the supervision of semantic

label information, the modeling of cross-modal correlation, the preservation of semantic consistency, and the P3S. These components jointly ensure the P3S method to learn more effective and discriminative common representations for cross-modal retrieval.

Typical cross-modal retrieval results on the large-scale PKU-XMediaNet dataset of our P3S method and two latest DNN-based compared methods ACMR and MCSM are shown in Fig. 3. Furthermore, Fig. 4 shows the PR curves of two retrieval tasks (Img2Txt and Txt2Img) generated by our P3S methods and several counterparts on Pascal Sentences and PKU-XMediaNet datasets. We can see that the downtrend of all the curves in Fig. 4 are coherent to the results reported in previous studies [11], [51]. The results in Fig. 4 consistently show the superior overall retrieval performance of our P3S as it generally obtains the highest precision scores with various recall levels in the curves.

### D. Further Analysis on P3S

*1) Visualizations of the Shared and Private Subspaces:* In Section I, we have shown the typical t-SNE visualizations of the shared and private subspaces on the Wikipedia dataset by our P3S method. In this experiment, more visualization results on other datasets are provided to demonstrate the effectiveness of the learned subspaces by our P3S method. Specifically, we take the NUS-WIDE-10k dataset as testbed, then use the t-SNE again to visualize the distribution of the 2000 testing image and text instances in both the shared and the private subspaces, using the learned common and private representations, respectively. The visualizations of the two subspaces on the NUS-WIDE-10k dataset are shown in Fig. 5, where in each subfigure, the points marked as circle and cross, respectively, denote the image and text instances. We can observe that the scattered distribution of the private subspace in Fig. 5(a) indicates that the interference within each modality is fully captured in the private representations. Since the interference is effectively excluded, the learned common representations of both image and text instances form several discriminative and compact clusters in the shared subspace, in addition with the supervision of the semantic label information. As a result, the discriminative and compact common representations are beneficial to boost the retrieval performance.

*2) Analysis on Parameter Sensitivity:* In this experiment, we first take both the Wikipedia dataset and PKU-XMediaNet dataset as two testbeds, and conduct experiment to explore the effect of key parameters of our P3S method. Specifically, we consider the hyperparameters $\alpha$, $\beta$, and $\gamma$ in (15) and the dimension $m$ of the shared/private subspaces. In this experiment, we set the numerical range of each hyperparameter in the range [0.001, 1000], and at each time, we fix the value of one parameter and search the best values of the others, which can dynamically reflect their pairwise relations. Besides, for the parameter $m$, we set its range in [50, 800] and increase it doubly in each step. The sensitivity analysis of the four parameters of our P3S is shown in Fig. 6. It can be seen that the performance of our P3S method varies with a different value of the three hyperparameters. In particular, the optimal

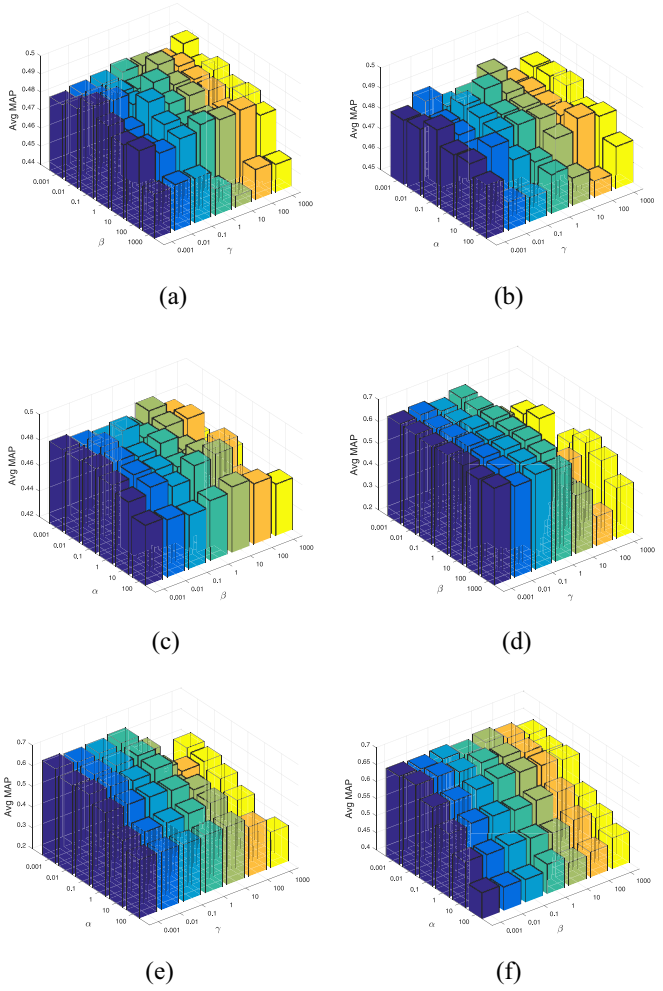
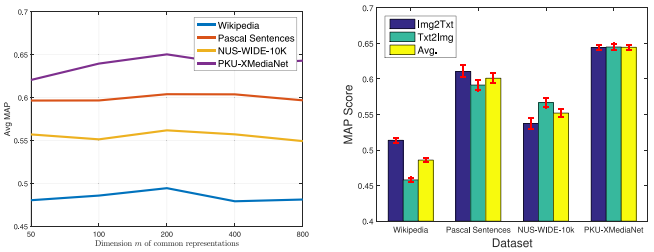



Fig. 6. Sensitivity analysis of the hyperparameters of our P3S method on the Wikipedia dataset and PKU-XMediaNet dataset. (a) Results by fixing $\alpha$ on Wikipedia. (b) Results by fixing $\beta$ on Wikipedia. (c) Results by fixing $\gamma$ on Wikipedia. (d) Results by fixing $\alpha$ on PKU-XMediaNet. (e) Results by fixing $\beta$ on PKU-XMediaNet. (f) Results by fixing $\gamma$ on PKU-XMediaNet.

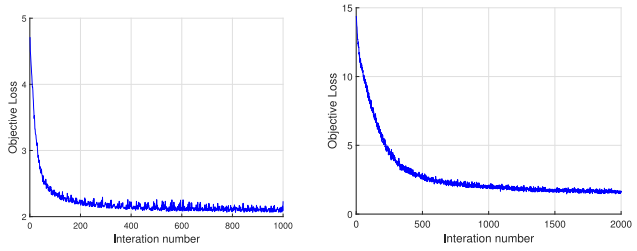
values for $\alpha$, $\beta$, and $\gamma$ are in the range of [0.01, 1] on the Wikipedia dataset and in the range of [0.001, 0.01] on the PKU-XMediaNet dataset. Thus, it indicates that the loss terms behind the three hyperparameters have different contributions for the P3S method on different datasets. On the Wikipedia dataset, when the values for the three hyperparameters are too large or too small (e.g., in [10, 1000] or [0.001, 0.01]), the performance drops severely as the impact of these loss terms becomes trivial. We can find a similar observation on the PKU-XMediaNet dataset. In practice, we can efficiently search for the optimal hyperparameters on the validation set for different datasets.

Regarding the parameter $m$ that directly controls the dimension of both the shared and private subspaces, from Fig. 7(a), we can observe that the P3S achieves the best performance with different optimal values of $m$ on different datasets. Specifically, on the Pascal Sentences dataset, the proposed P3S method obtains the best averaged MAP score when $m$ is round 300, and on the other three datasets, the optimal $m$ is around 200. When the dimension is smaller (e.g., less than 100) or bigger (e.g., more than 400), the performance of




(a) (b)

Fig. 7. (a) Effect of subspace dimension $m$ and the (b) MAP scores with error bars (standard deviation) of the proposed P3S method on all datasets.




(a) (b)

Fig. 8. Convergence experiments of our P3S method on (a) Wikipedia and (b) PKU-XMediaNet datasets.

P3S becomes inferior. Thus, it shows that the dimension of the learned shared subspace is also an important factor as it affects the training efficiency and determines the quality of the learned common representations.

Furthermore, Fig. 7(b) shows the MAP scores with error bars of our P3S method on four datasets, from which we can observe that the results of P3S are insensitive to the network initialization. Therefore, it again demonstrates the robustness of our P3S method.

*3) Analysis on Convergence:* We also present convergence experiments for our P3S approach on the Wikipedia and PKU-XMediaNet datasets to assess its training efficiency. Fig. 8 shows that the curves of the loss value in (15) have clear downtrend on the two datasets. It can be observed that our P3S method can reach the convergence status within 400 iterations on Wikipedia and 1000 iterations on PKU-XMediaNet, respectively, showing its high efficiency for training. In practice, it only requires around 5 and 30 min to obtain the optimal model parameters on the two datasets, respectively. In contrast, several counterparts, such as CCL and MSCM need much longer time (around 3–5 times) to obtain their best performance under the same experimental configurations.

## V. Conclusion

In this article, we have proposed a novel network architecture called P3S to learn compact and discriminative common representations of different modality data for cross-modal retrieval. The proposed P3S consists of three paralleled subnetworks, that is, an SNet and two PNets (PNet-Img and PNet-Txt). The SNet introduces a shared subspace to learn common representations of different modalities by capturing the cross-modal correlation of the heterogeneous data. Moreover, the two PNets assume the private subspace within each modality, in which the private representations are learned to model the interference in the modality-specific properties. Additional soft subspace orthogonality constraints ensure the learned representations in the shared and private subspaces to be dissimilar. Extensive experiments with comprehensive analysis validate the effectiveness of the learned common representations of our P3S method by explicitly separating the shared space and private spaces. For future work, we will explore more advanced schemes to learn the cross-modal correlation and explore our P3S method for other modality data besides images and texts.

## References

[1] H. Turtle and W. B. Croft, "Inference networks for document retrieval," *SIGIR*, vol. 51, no. 2, pp. 124–147, 2017.
[2] F. Heimerl, S. Koch, H. Bosch, and T. Ertl, "Visual classifier training for text document retrieval," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 12, pp. 2839–2848, Dec. 2012.
[3] M. Wang, W. Li, D. Liu, B. Ni, J. Shen, and S. Yan, "Facilitating image search with a scalable and compact semantic mapping," *IEEE Trans. Cybern.*, vol. 45, no. 8, pp. 1561–1574, Aug. 2015.
[4] W. Hu, N. Xie, L. Li, X. Zeng, and S. J. Maybank, "A survey on visual content-based video indexing and retrieval," *IEEE Trans. Cybern.*, vol. 41, no. 6, pp. 797–819, Nov. 2011.
[5] M. Hu, Y. Yang, F. Shen, N. Xie, and H. T. Shen, "Hashing with angular reconstructive embeddings," *IEEE Trans. Image Process.*, vol. 27, no. 2, pp. 545–555, Feb. 2018.
[6] F. Shen, Y. Xu, L. Liu, Y. Yang, Z. Huang, and H. T. Shen, "Unsupervised deep hashing with similarity-adaptive and discrete optimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 3034–3044, Dec. 2018.
[7] B. Wang, Y. Yang, X. Xu, A. Hanjalic, and H. T. Shen, "Adversarial cross-modal retrieval," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2017, pp. 154–162.
[8] N. Rasiwasia *et al.*, "A new approach to cross-modal multimedia retrieval," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2010, pp. 251–260.
[9] 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.
[10] 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.
[11] 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, Aug. 2018.
[12] 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.
[13] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[14] Y. Peng, X. Huang, and J. Qi, "Cross-media shared representation by hierarchical learning with multiple deep networks." in *Proc. IJCAI*, 2016, pp. 3846–3853.
[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] 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. 1–24, 2019.
[17] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 7–16.
[18] 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 ICMR*, 2018, pp. 46–54.
[19] W. Wang, B. C. Ooi, X. Yang, D. Zhang, and Y. Zhuang, "Effective multi-modal retrieval based on stacked auto-encoders," *Proc. VLDB Endow.*, vol. 7, no. 8, pp. 649–660, 2014.

[20] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: arxiv.abs/1409.1556.
[21] Y. Kim, "Convolutional neural networks for sentence classification," in *Proc. EMNLP*, 2014, pp. 1746–1751.
[22] 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, 2004.
[23] M. Xu, Z. Zhu, X. Zhang, Y. Zhao, and X. Li, "Canonical correlation analysis with L2,1-norm for multiview data representation," *IEEE Trans. Cybern.*, early access, Apr. 4, 2019, doi: 10.1109/TCYB.2019.2904753.
[24] V. Ranjan, N. Rasiwasia, and C. Jawahar, "Multi-label cross-modal retrieval," in *Proc. ICCV*, 2015, pp. 4094–4102.
[25] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2003, pp. 604–611.
[26] 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*, 2013, pp. 263–269.
[27] K. Wang, R. He, W. Wang, L. Wang, and T. Tan, "Learning coupled feature spaces for cross-modal matching," in *Proc. ICCV*, 2013, pp. 2088–2095.
[28] 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.
[29] 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.
[30] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML*, 2011, pp. 689–696.
[31] G. Andrew, R. Arora, J. Bilmes, and K. Livescu, "Deep canonical correlation analysis," in *Proc. ICML*, 2013, pp. 1247–1255.
[32] I. J. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2014, pp. 2672–2680.
[33] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier GANs," in *Proc. ICML*, 2017, pp. 2642–2651.
[34] C. Ledig *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. CVPR*, 2017, pp. 105–114.
[35] C. Finn, I. J. Goodfellow, and S. Levine, "Unsupervised learning for physical interaction through video prediction," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 64–72.
[36] G. Song, D. Wang, and X. Tan, "Deep memory network for cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1261–1275, May 2019.
[37] 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*, 2013, pp. 785–796.
[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] 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.
[40] X. Liu, Z. Hu, H. Ling, and Y.-M. Cheung, "MTFH: A matrix TRI-factorization hashing framework for efficient cross-modal retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Sep. 10, 2019, doi: 10.1109/TPAMI.2019.2940446.
[41] H. T. Shen *et al.*, "Exploiting subspace relation in semantic labels for cross-modal hashing," *IEEE Trans. Knowl. Data Eng.*, early access, Jan. 29, 2020, doi: 10.1109/TKDE.2020.2970050.
[42] J. Wang, T. Zhang, J. Song, N. Sebe, and H. T. Shen, "A survey on learning to hash," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 769–790, Apr. 2018.
[43] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised visual domain adaptation using subspace alignment," in *Proc. ICCV*, 2013, pp. 2960–2967.
[44] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proc. AAAI*, 2016, pp. 2058–2065.
[45] S. Li, S. Song, and G. Huang, "Prediction reweighting for domain adaptation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 7, pp. 1682–1695, Jul. 2017.
[46] M. Long, Y. Cao, J. Wang, and M. I. Jordan, "Learning transferable features with deep adaptation networks," in *Proc. ICML*, 2015, pp. 97–105.
[47] K. Bousmalis, G. Trigeorgis, N. Silberman, D. Krishnan, and D. Erhan, "Domain separation networks," in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2016, pp. 343–351.
[48] Y. Ganin *et al.*, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2096–2030, 2016.
[49] B. Sun and K. Saenko, "Deep CORAL: Correlation alignment for deep domain adaptation," in *Proc. ECCV*, 2016, pp. 443–450.
[50] L. Duan, D. Xu, and I. W.-H. Tsang, "Domain adaptation from multiple sources: A domain-dependent regularization approach," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 3, pp. 504–518, Mar. 2012.
[51] 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.
[52] X. Huang and Y. Peng, "Deep cross-media knowledge transfer," in *Proc. CVPR*, 2018, pp. 8837–8846.
[53] M. Mancini, L. Porzi, S. R. Bulò, B. Caputo, and E. Ricci, "Boosting domain adaptation by discovering latent domains," in *Proc. CVPR*, 2018, pp. 3771–3780.
[54] 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, no. 25, pp. 723–773, 2012.
[55] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's Mechanical Turk," in *Proc. NAACL HLT Workshop*, 2010, pp. 139–147.
[56] X. Zhai, Y. Peng, and J. Xiao, "Heterogeneous metric learning with joint graph regularization for cross-media retrieval," in *Proc. AAAI*, 2013, pp. 1198–1204.
[57] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-wide: A real-world web image database from National University of Singapore," in *Proc. ACM ICIVR*, 2009, pp. 48–55.
[58] Y. Peng, X. Zhai, Y. Zhao, and X. Huang, "Semi-supervised cross-media feature learning with unified patch graph regularization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 3, pp. 583–596, Mar. 2016.
[59] J. Qi and Y. Peng, "Cross-modal bidirectional translation via reinforcement learning," in *Proc. IJCAI*, 2018, pp. 2630–2636.
[60] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014. [Online]. Available: arxiv.abs/1412.6980.
[61] S. S. Girija, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," 2016. [Online]. Available: arxiv.org/abs/1603.04467.
[62] 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.
[63] 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.

**Xing Xu** (Member, IEEE) 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.

**Kaiyi Lin** received the B.S. degree in software engineering from the University of Electronic Science and Technology of China, Chengdu, China, in 2018. He is currently pursuing the master's degree with the School of Software and Microelectronics, Peking University, Beijing, China.

His main research interests are multimedia content analysis and social media analysis.

**Lianli Gao** (Member, IEEE) received the B.S. degree in computer science from the University of Electronic Science and Technology of China, Chengdu, China, in 2009, and the Ph.D degree in information technology from the University of Queensland, Brisbane, QLD, Australia, in 2015.

She is currently a Professor with the School of Computer Science and Engineering, University of Electronic Science and Technology of China. Her research interests include semantic web, machine learning, deep learning, computer vision, and the related practical applications.

**Huimin Lu** (Senior Member, IEEE) 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 with the Ministry of Education, Culture, Sports, Science and Technology, Tokyo, Japan. His current research interests include computer vision, robotics, artificial intelligence, and ocean observing.

**Heng Tao Shen** (Senior Member, IEEE) 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 and the Dean of the School of Computer Science and Engineering, an Executive Dean of AI Research Institute, and the Director of the Centre for Future Media, University of Electronic Science and Technology of China, Chengdu, China. He has published 280+ peer-reviewed papers, including 80+ IEEE/ACM Transactions. His research interests mainly include multimedia search, computer vision, artificial intelligence, and big data management.

Prof. Shen received the seven Best Paper Awards from international conferences, including the Best Paper Award from ACM Multimedia 2017 and the Best Paper Award—Honorable Mention from ACM SIGIR 2017. He is/was an Associate Editor of the *ACM Transactions of Data Science*, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, and the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. He is an OSA Fellow and an ACM Distinguished Member.

**Xuelong Li** (Fellow, IEEE) received the Ph.D. degree from the University of Science and Technology of China, Hefei, China, in 2002.

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.