Mining on Heterogeneous Manifolds for Zero-Shot Cross-Modal Image Retrieval

Fan Yang,[1,2] **Zheng Wang,**[2,*] **Jing Xiao,**[2] **Shin'chi Satoh**[1,2]

[1]The University of Tokyo, Japan
[2]National Institute of Informatics, Japan
{yang, wangz, jing_xiao, satoh}@nii.ac.jp

Abstract

Most recent approaches for the zero-shot cross-modal image retrieval map images from different modalities into a uniform feature space to exploit their relevance by using a pre-trained model. Based on the observation that manifolds of zero-shot images are usually deformed and incomplete, we argue that the manifolds of unseen classes are inevitably distorted during the training of a two-stream model that simply maps images from different modalities into a uniform space. This issue directly leads to poor cross-modal retrieval performance. We propose a bi-directional random walk scheme to mining more reliable relationships between images by traversing heterogeneous manifolds in the feature space of each modality. Our proposed method benefits from intra-modal distributions to alleviate the interference caused by noisy similarities in the cross-modal feature space. As a result, we achieved great improvement in the performance of the thermal *v.s.* visible image retrieval task. The code of this paper: https://github.com/fyang93/cross-modal-retrieval

1 Introduction

The ongoing prosperity of deep learning lowers the barrier to produce reliable image representations. Given robust image representations, previous studies have made tremendous progress on the task of image retrieval. Although most image retrieval tasks are enclosed in a single modality, *e.g.* all images are from the RGB domain, a rising trend is to break through this limitation so that images can be queried across different modalities. However, a huge performance gap can still be observed between intra-modal and cross-modal image retrieval tasks (Ye et al. 2018b; Wang et al. 2019; 2018). We specifically focus on the zero-shot cross-modal image retrieval where the testing images are sampled from classes that never appeared in the training set.

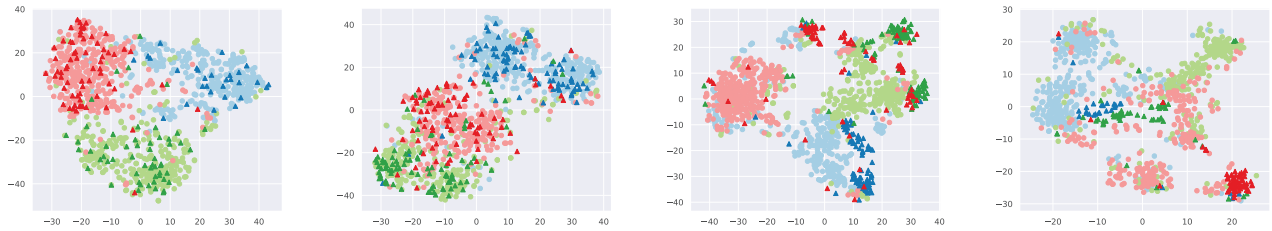
For a single-modal task, it is general to fine-tune a model pre-trained on ImageNet (Deng et al. 2009) and extract image features from its intermediate layers. While for a cross-modal task, the images come from two subsets corresponding to two different modalities. In this regard, the cross-modal model usually has two streams where each of them handles images input from each modality. Those two streams join together at the tail of the model, mapping image features from two modalities into a shared feature space, with a classifier attached afterward (Ye et al. 2018a; 2018b). Although such a model memorized how to map training images of the same class into a concentrated cluster whichever modalities they are from, we argue that aligning classes in the training set across different modalities inevitably *distorts the remaining manifolds consists of samples of unseen classes*.

Fig. 1 visualizes this phenomenon. Fig. 1a shows the feature distribution of digits 7 to 9 in the MNIST dataset where features are extracted by a single-modal model which is trained by using other digits in the MNIST dataset only. While Fig. 1b shows the feature distribution for the SVHN dataset. Although imperfect, it can be observed that most image features are concentrated around their class centers. The cross-modal model, instead, is trained with images from both modalities (*i.e.* modalities of MNIST and SVHN) and results in a cluttered distribution in its uniform feature space (Figs. 1c and 1d). Particularly, it can be observed that the manifold of digit 9 in Fig. 1d is isolated into several blobs. Needless to say, this issue directly results in poor cross-modal retrieval performance.

Some may ascribe such a phenomenon to the over-fitting, but we have another plausible interpretation. Suppose that we have images of lemon, lime, and mango. In the color domain, lemon is more similar to mango than lime in terms of color. While in the monochrome domain, lemon is closer to lime than mango in terms of shape. These domain-specific relationships are helpful in forming the feature distribution in every single domain. However, when we train a cross-modal model by using all images of lemon and lime, the conflicts among those relationships (lemon and lime are close in shape but far in color) ask both domains to adapt its feature space for each other. As a result, the feature space of each domain is distorted and the feature manifold of the unseen class mango is very likely to be torn into parts.

In this work, we propose to leverage the domain-specific information by mining on manifolds in different single-modal feature spaces. Our idea is straightforward. Now that

*corresponding author



(a) MNIST → MNIST (b) SVHN → SVHN (c) SVHN → MNIST (d) MNIST → SVHN

Figure 1: Visualization of feature distributions [t-SNE (Maaten and Hinton 2008)] for zero-shot digits 7 (●▲), 8(●▲) and 9(●▲) in single-modal spaces (a)(b) and cross-modal spaces (c)(d), while other digits (0∼6) are used for training. Samples from the query set and the gallery set are represented by triangles and circles respectively. The subcaptions are formatted by {dataset whose testing set (7∼9) is used as query} → {dataset whose training set (7∼9) is used as gallery}.

the manifolds of zero-shot images in the cross-modal feature space are distorted and possibly incomplete, while the single-modal spaces have complete yet well-shaped manifolds. We take the nearest neighbors of a query in the cross-modal feature space as the initial state, then apply random walk on the corresponding single-modal feature space. In other words, we use the cross-modal feature space as a bridge to connect heterogeneous manifolds of different modalities. Our main contributions are 1) we show the proofs that mapping images from different modalities into a uniform feature space distorts manifolds of unseen classes; 2) we propose a novel random walk scheme which benefits from the reliable feature distributions created by single-modal models; 3) we validate our method on visible *v.s.* thermal datasets and achieves significant performance improvement.

The rest of the paper is structured as follows. After introducing the related work, we formulate our proposed bi-directional random walk scheme in Section 3. Section 4 describes our experimental settings and Section 5 demonstrates the results, showing that our proposed method significantly improves retrieval performance before concluding.

2 Related Works

Existing approaches for cross-modal image retrieval mainly focused on the representation learning to coalesce feature spaces in different modalities, such as visible and infrared modalities. (Wu et al. 2017) introduced a deep zero-padding framework where images from different modalities share a common convolutional network during the feature generation. However, instead of training a common network, it is natural to benefit from the intrinsic context inside each modality by using a two-stream model. The literature is rich in the applications of the two-stream model with various training strategies. Some early researches trained the model to generate hashing code as features (Cao et al. 2016; 2017). Different loss functions designed for metric learning, *e.g.* contrastive loss, ranking loss, and sphere loss were employed in recent works (Ye et al. 2018a; 2018b; Hao et al. 2019). (Dai et al. 2018) adopted a generative adversarial training strategy which was first used in the text *v.s.* image retrieval (Wang et al. 2017). The key idea is to update image features to fool the discriminator so that it cannot distinguish the source modality of the feature, while the discriminator is trained to categorize features into their source modalities at the same time. Another recent trend is to apply data augmentation by converting images to other modalities with generative adversarial networks (Wang et al. 2019).

Besides the representation learning performed by the above methods, it is equivalently important to choose a proper search algorithm. Based on the k-NN search, average query expansion is a widely used technique to boost performance (Chum et al. 2007). Other than that, a better image retrieval performance can also be achieved through mining on manifolds. (Zhou et al. 2004) first proposed to conduct a random walk on an undirected graph where each image is connected to its locally nearest neighbors. Starting from an initial state, such a random walk process spreads out the weight on the query node iteratively until it reaches a stable state. The weights on other nodes are then regarded as the similarity scores to the query. Since the graph represents local relationships between images, it allows traversing over manifolds regardless of their shapes and ranges. (Donoser and Bischof 2013) used the term diffusion to describe this technique. Recent works focused on improving diffusion's speed and performance. (Bai et al. 2017a; 2017b) proposed the regularized ensemble diffusion to suppress the noisy similarities between false-positive image pairs in the graph. (Iscen et al. 2018) extended diffusion to regional features and achieved better retrieval performance than using global features only. (Yang et al. 2019) decoupled diffusion into online and offline processes, making it efficient during the search time. In addition, they achieved better performance by conducting late truncation in large scale datasets.

3 Proposed Method

We propose a bi-directional random walk scheme for cross-modal image retrieval. Such a random walk approach was also called diffusion in single-modal image retrieval (Donoser and Bischof 2013; Iscen et al. 2017; Yang et al. 2019). We avoid using the term diffusion since

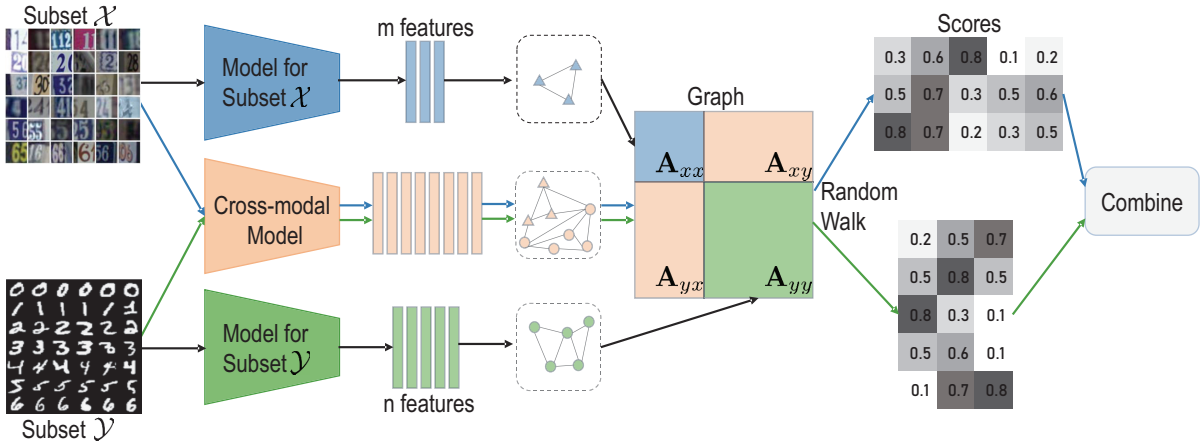


Figure 2: Overview of our proposed bi-directional random walk scheme. The black arrows show the pipeline to obtain the graph represented by an adjacency matrix, while the blue and green paths describe the random walk processes in both directions.

our problem setting is not just diffusing similarity scores in the gallery for the given query but in both directions.

3.1 Problem Definition

For cross-modal image retrieval, we have two subsets containing images of two different modalities. We name the two subsets $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ and $\mathcal{Y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$. For each subset, we train a single-modal model to obtain feature representations. We denote $f_x : \mathcal{X} \mapsto \mathcal{R}^d$ and $f_y : \mathcal{Y} \mapsto \mathcal{R}^d$ as feature extraction operations for each subset using the corresponding models. We also train a cross-modal model, serving as a bridge to relate two subsets, to project all images to feature vectors in a uniform space where the inputs from two modalities are mapped into. This operation is denoted by $f_u : \{\mathcal{X}, \mathcal{Y}\} \mapsto \mathcal{R}^d$.

Now that the cross-modal image retrieval problem is depicted as ranking the images in $\mathcal{Y}$ by their similarities to the given query image $\mathbf{x}_i \in \mathcal{X}$, or ranking $\mathcal{X}$ by the given query $\mathbf{y}_j \in \mathcal{Y}$. A naive solution is to apply k-NN search in $\{f_u(\mathbf{x}_1), \ldots, f_u(\mathbf{x}_n)\}$ ($\{f_u(\mathbf{y}_1), \ldots, f_u(\mathbf{y}_n)\}$) with the query $f_u(\mathbf{y}_i)$ ($f_u(\mathbf{x}_i)$). However, as we mentioned in Section 1, the manifolds of unseen classes in the uniform feature space are very likely to be deformed during the training of a two-stream model. In severe cases such as searching for digit 7 in SVHN by the given queries in MNIST (see Fig. 1d, blue), k-NN search can find either a small portion of the positive samples or negative ones. To alleviate this issue, we propose to construct graphs for heterogeneous manifolds in the single-modal spaces and apply a random walk across them.

3.2 Graph Construction

We first consider a large graph representing affinities between all images, then show how to split it up. The entire dataset $\{\mathcal{X}, \mathcal{Y}\}$ contains $m + n$ images, so we allocate a $\mathcal{R}^{(m+n)\times(m+n)}$ adjacency matrix to record their similarity scores to each other:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{xx} & \mathbf{A}_{xy} \\ \mathbf{A}_{yx} & \mathbf{A}_{yy} \end{bmatrix}, \tag{1}$$

where $\mathbf{A}_{xx} \in \mathcal{R}^{m\times m}$, $\mathbf{A}_{xy} \in \mathcal{R}^{m\times n}$, $\mathbf{A}_{yx} \in \mathcal{R}^{n\times m}$, and $\mathbf{A}_{yy} \in \mathcal{R}^{n\times n}$.

To benefit from the domain-specific information, we compute $\mathbf{A}_{xx}$ and $\mathbf{A}_{yy}$ by using their features from two isolated single-modal models. This process is shown in Fig. 2 by black arrows. Following (Iscen et al. 2017), we set element a_{ij} $(i \neq j)$ in $\mathbf{A}_{xx}$ to $\langle f_x(\mathbf{x}_i), f_x(\mathbf{x}_j)\rangle$ only if $f_x(\mathbf{x}_i)$ and $f_x(\mathbf{x}_j)$ are in each other's k nearest neighbors, and 0 otherwise. We omit the description to obtain $\mathbf{A}_{yy}$ since it is alike.

Notice that $\mathbf{A}_{xy}$ and $\mathbf{A}_{yx}$ are used to save the similarity scores between images from different modalities. Therefore, it requires a measurement in a shared feature space. As a result, element a_{ij} $(i \neq j)$ in $\mathbf{A}_{xy}$ is $\langle f_u(\mathbf{x}_i), f_u(\mathbf{y}_j)\rangle$ only if $f_u(\mathbf{x}_i)$ and $f_u(\mathbf{y}_j)$ are in each other's k nearest neighbors, and 0 otherwise. $\mathbf{A}_{yx}$ is the transposition of $\mathbf{A}_{xy}$.

Since we only concern about the affinities between each image and its locally nearest neighbors, the adjacency matrix is sparse and costs a limited amount of memory.

3.3 Random Walk

Now we have constructed the graph represented by an adjacency matrix. We still need to convert it to a transition matrix for the next random walk. This step is called normalization. Since the adjacency matrix consists of blocks containing similarity scores measured in different spaces, here we break down the adjacency matrix and apply normalization individually to each block. Specifically, $\mathbf{A}_{xx}$ and $\mathbf{A}_{yy}$ are symmetrically normalized by

$$\mathbf{S}_{xx} = \mathbf{D}_{xx}^{-1/2}\mathbf{A}_{xx}\mathbf{D}_{xx}^{-1/2}, \ \mathbf{S}_{yy} = \mathbf{D}_{yy}^{-1/2}\mathbf{A}_{yy}\mathbf{D}_{yy}^{-1/2}, \tag{2}$$

where elements on the diagonal of $\mathbf{D}_{xx}$ and $\mathbf{D}_{yy}$ are the row-wise sums of $\mathbf{A}_{xx}$ and $\mathbf{A}_{yy}$ respectively, while the off-diagonal elements are all zeros. On the other hand, $\mathbf{A}_{xy}$ and $\mathbf{A}_{yx}$ are ℓ_1 normalized to $\mathbf{S}_{xy}$ and $\mathbf{S}_{yx}$ by row. Similar to

the structure of the adjacency matrix, the transition matrix is defined as

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_{xx} & \mathbf{S}_{xy} \\ \mathbf{S}_{yx} & \mathbf{S}_{yy} \end{bmatrix}. \tag{3}$$

Once the transition matrix is prepared, we can start the random walk from any wanted initial states. The state vector at t-th step of random walk is denoted by $\mathbf{f}^t = [\mathbf{f}_x^{t\top}, \mathbf{f}_y^{t\top}]^\top$, where $\mathbf{f}_x^t \in \mathcal{R}^m$ and $\mathbf{f}_y^t \in \mathcal{R}^n$. On each random walk stage, we concern about the retrieval result of one query image. Therefore, we set the initial state $\mathbf{f}^0$ as a one-hot vector, where the position of 1 implies the index of the query. Another parameter $\alpha \in (0, 1)$ is introduced to ensure that the random walk can finally converge to a stable state. Then each iteration step of random walk is defined as

$$\mathbf{f}^{t+1} = \alpha \mathbf{S}\mathbf{f}^t + (1-\alpha)\mathbf{f}^0. \tag{4}$$

After assigning Eq. (3) into Eq. (4), we obtain

$$\begin{aligned} \mathbf{f}_x^{t+1} &= \alpha \mathbf{S}_{xx}\mathbf{f}_x^t + \alpha \mathbf{S}_{xy}\mathbf{f}_y^t + (1-\alpha)\mathbf{f}_x^0, \\ \mathbf{f}_y^{t+1} &= \alpha \mathbf{S}_{yx}\mathbf{f}_x^t + \alpha \mathbf{S}_{yy}\mathbf{f}_y^t + (1-\alpha)\mathbf{f}_y^0. \end{aligned} \tag{5}$$

When the query image lies in the subset $\mathcal{X}$ ($\mathcal{Y}$), we only concern the similarity scores between the query and each image in subset $\mathcal{Y}$ ($\mathcal{X}$). It means that we only need to compute either $\mathbf{f}_x^\infty$ or $\mathbf{f}_y^\infty$ rather than both of them. For instance, during the computation of $\mathbf{f}_y^\infty$, $\mathbf{S}_{xx}$ and $\mathbf{S}_{xy}$ are regarded as zero matrices so that $\mathbf{f}_x^t$ is always fixed. Similarly, $\mathbf{f}_y^t$ is fixed during the computation of $\mathbf{f}_x^\infty$. As a result, we obtain the closed-form solution for $\mathbf{f}_x^\infty$ and $\mathbf{f}_y^\infty$:

$$\begin{aligned} \mathbf{f}_x^\infty &= \alpha(1-\alpha)(\mathbf{I} - \alpha \mathbf{S}_{xx})^{-1}\mathbf{S}_{xy}\mathbf{f}_y^0, \\ \mathbf{f}_y^\infty &= \alpha(1-\alpha)(\mathbf{I} - \alpha \mathbf{S}_{yy})^{-1}\mathbf{S}_{yx}\mathbf{f}_x^0. \end{aligned} \tag{6}$$

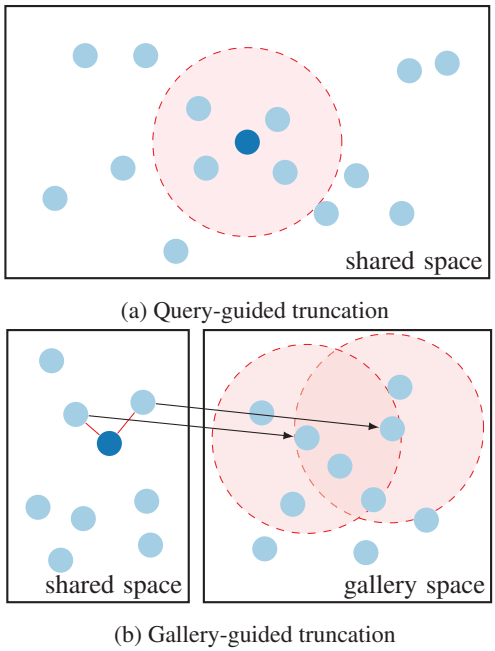
For brevity, the constant scale $\alpha(1-\alpha)$ is ignored and $\mathbf{I} - \alpha\mathbf{S}_{(\cdot)}$ is denoted by $\mathcal{L}_{(\cdot)}$ by convention. Eventually, Eq. (6) is simplified to

$$\mathbf{f}_x^\infty \propto \mathcal{L}_{xx}^{-1}\mathbf{S}_{xy}\mathbf{f}_y^0, \ \mathbf{f}_y^\infty \propto \mathcal{L}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{f}_x^0, \tag{7}$$

where $\mathbf{f}_x^0$ and $\mathbf{f}_y^0$ are one-hot vectors representing the index of query image in subset $\mathcal{X}$ and $\mathcal{Y}$.

3.4 Graph Truncation

Truncation is a necessary technique in applying random walk to large datasets. The key idea is to search in a proper range of the query's neighborhood rather than the entire gallery. Recently, (Yang et al. 2019) showed that truncation can not only enhance efficiency but also improves retrieval performance by ruling out potential false-positive samples. There are two main types of truncation (Iscen et al. 2018; Yang et al. 2019). Fig. 3 demonstrates the difference between them. The query-guided truncation conducts a k-NN search and uses found samples as the truncated gallery. This requires the query and the gallery to be in the same space. While the gallery-guided truncation decides the search range by combining several neighborhood ranges around the query's nearest neighbors. In other words, a few nearest neighbors of the query can help to apply truncation in another feature space (see Fig. 3b). In our scenario,



(a) Query-guided truncation

(b) Gallery-guided truncation

Figure 3: Comparison between the query-guided truncation and the gallery-guided truncation, where the query is in dark blue while the gallery samples are in light blue. The pink area represents the final range to search in.

it is preferable to apply truncation in the single-modal space of the gallery side instead of the noisy cross-modal space. Therefore, our solution is to look for a few nearest neighbors of the query in the cross-modal space and then apply the gallery-guided truncation in the single-modal space.

From the perspective of math, the truncation can be either achieved by slicing the matrices to only reserve concerned rows and columns or simply replacing the similarities out of interest to zeros. For simplicity, we do not revise the notations of matrices to refer to the truncated ones. Unless otherwise stated, Laplacian matrices ($\mathcal{L}_{xx}$ and $\mathcal{L}_{yy}$) are truncated by default throughout this paper.

3.5 Bi-directional Retrieval

As afore-mentioned, we aim at the bi-directional retrieval instead of the unidirectional retrieval from the query side to the gallery side. That means we view both subsets $\mathcal{X}$ and $\mathcal{Y}$ equivalently, and each of them can be either the query or the gallery. In this regard, we exploit heterogeneous manifolds of both modalities by conducting random walk on the constructed graphs.

Eq. (7) shows the convergence of the random walk process. Scores in the final state of the random walk are important cues for ranking images by their similarities to the query on manifolds. For instance, starting from the i-th query image in the subset $\mathcal{Y}$, the final state vector $\mathbf{f}_x^\infty$ is the i-th column vector of the matrix $\mathcal{L}_{xx}^{-1}\mathbf{S}_{xy}$ whose elements are also viewed as scores for the subsequent ranking. In other words, $\mathcal{L}_{xx}^{-1}\mathbf{S}_{xy} \in \mathcal{R}^{m\times n}$ and $\mathcal{L}_{yy}^{-1}\mathbf{S}_{yx} \in \mathcal{R}^{n\times m}$ consists of the similarity scores of each query image in $\mathcal{Y}$ and $\mathcal{X}$ respec-

tively. We define the result of our bi-directional retrieval as

$$\mathbf{B} = \lambda \mathcal{L}_{xx}^{-1} \mathbf{S}_{xy} + (1-\lambda)(\mathcal{L}_{yy}^{-1} \mathbf{S}_{yx})^{\top}, \tag{8}$$

where λ is weight for balancing the scores obtained from both directions. Note that we do not choose to normalize similarity scores. In the later experiment (see Fig. 5), we show that normalizing similarity scores for each query has no positive effect on the performance.

4 Datasets and Setup

4.1 Datasets

We first introduce two well-known digit classification datasets briefly, then describe two person re-identification datasets with both visible and thermal (infrared) images.

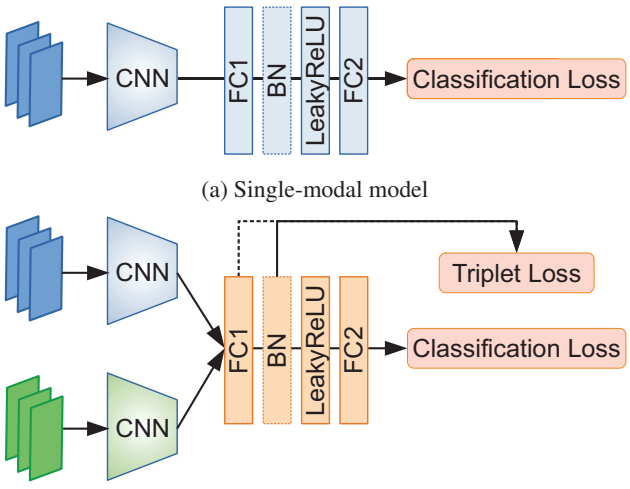
MNIST and SVHN are both digit datasets. The MNIST dataset (LeCun et al. 1998) consists of handwritten digit images from 0 to 9 of size 28×28. Among them, 60,000 binary images are in the training set and 10,000 in the testing set. The SVHN (Street View House Numbers) dataset (Netzer et al. 2011), as implied by its name, consists of RGB images of digits. It contains 73,257 training and 26,032 testing images of size 32×32.

RegDB contains images of 412 persons. 10 visible images and 10 thermal images were captured for each person by a dual-camera system (Nguyen et al. 2017). We take the evaluation protocol in (Ye et al. 2018b) where the entire dataset was divided into a training set and a testing set. On the testing stage, query images and gallery images are from different modalities for evaluating the cross-modal image retrieval. Since the dataset was randomly split, it is recommended to run the whole procedure for 10 trials to obtain statistically stable results.

SYSU-MM01 consists of visible and thermal images of 491 identities captured by 6 cameras including 4 RGB cameras and 2 infrared cameras (Wu et al. 2017). Those RGB cameras and IR cameras work in light and dark scenarios respectively. The training set contains 22,258 visible images and 11,909 thermal images of 395 persons. The testing set contains 3,803 thermal query images where 96 persons appeared, and 301 visible images randomly sampled for each person as the gallery set. Following the evaluation protocol in (Ye et al. 2018b; Wang et al. 2019), we adopt the most challenging single-shot all-search mode evaluation protocol. However, it should be clarified that we adopt the entire gallery set during the graph construction, and randomly select 301 entries for each person in the result of our bi-directional retrieval for the subsequent evaluation.

4.2 Model Fine-tuning

Our main concern is to exploit the manifolds in feature distributions rather than features themselves. So we merely adopt two baseline models (see Fig. 4) for feature extraction. Fig. 4a shows the model for images in a single modality, while Fig. 4b shows the cross-modal model taking images from two different modalities as input. ResNet18 serves as the backbone CNN for MNIST and SVHN datasets, while



(a) Single-modal model

(b) Cross-modal model

Figure 4: Overview of our adopted models. The batch normalization is only used for RegDB and SYSU-MM01 datasets. For digit images, we use features extracted from the FC1 layer, while for person images, we use features extracted from the batch normalization layer.

ResNet50 is used for RegDB and SYSU-MM01 datasets (He et al. 2016). Both of them are pre-trained on the ImageNet dataset. We take the backbone's convolutional layers followed by an adaptive average pooling layer as the feature provider. The rest part of the models is shared. Following (Ye et al. 2018a), image features flow into a fully-connected layer (FC1), followed by a batch normalization block and a leaky ReLU. Here, the dimension of the output of FC1 is set to 512 for all datasets. Notice that we only use batch normalization in training models for RegDB and SYSU-MM01 datasets. At the end of the model, an FC layer (FC2) is attached for classification.

We then fine-tune the models for each dataset. For each class, we fine-tune a single-modal model by using cross entropy as the criterion for classification. These fine-tuned single-modal models function as f_x and f_y defined in Section 3.1 where we take image representations from the FC1 layer for digit images and the batch normalization block for person images. The training of the cross-modal model is a bit more difficult. As shown in Fig. 4b, average pooled features from two streams are mapped into a uniform space by a shared FC1 layer. Besides the classification loss, *i.e.* cross entropy, a batch hard triplet loss (Hermans, Beyer, and Leibe 2017) is introduced to push features of the same class to gather closer. The overall loss function for the cross-modal model is

$$L = L_{class} + \beta L_{tri}, \tag{9}$$

where β is a weight on the triplet loss. In our experiments, we consistently set $\beta = 0.2$ and the margin of triplet loss to 0.6 through validation.

During the training, 7 digits (0 to 6) are used for the MNIST and SVHN datasets only. The remaining three digits are then regarded as zero-shot classes, which is coherent

Query	Gallery	Rank-1	Rank-5	Rank-10	mAP
MNIST	MNIST	96.4%	98.5%	99.0%	69.1%
SVHN	SVHN	89.0%	97.0%	98.5%	65.4%
SVHN	MNIST	59.0%	80.0%	84.9%	48.3%
MNIST	SVHN	45.2%	83.3%	95.6%	46.7%

Table 1: Intra-modal and cross-modal image retrieval performance by using k-NN search on digit datasets.

with Fig. 1.

5 Experiments

We show the effectiveness of the proposed method on the above datasets by conducting cross-modal image retrieval in this section. We particularly explain the way we perform retrieval on the MNIST and SVHN datasets. Those two datasets are made for classification rather than retrieval, therefore we have to manually set up two subsets for each of them, *i.e.* the query set and the gallery set. As described in Fig. 1, we use the training set of digits 7 to 9 as the gallery, while samples in the testing set are used as queries. In the MNIST dataset, the numbers of images in the query and gallery set are 3,011 and 18,065 respectively. While the SVHN dataset has 15,299 images in the gallery and 5,274 query images.

In addition to the proposed bi-directional random walk on heterogeneous manifolds, we also conduct experiments on homogeneous manifolds for comparison. It only differs from the heterogeneous way in constructing graphs, namely, it computes $\mathbf{A}_{xx}$ and $\mathbf{A}_{yy}$ by using cross-modal features. Here, the homogeneous manifolds refer to manifolds in the cross-modal uniform feature space. Although they consist of features from different modalities, we regarded them as homogeneous manifolds since all the images are intentionally mapped into a uniform feature space.

5.1 Digit Image Retrieval

We first perform the naive k-NN search for all combinations of two digit datasets. Table 1 shows the results. The gap between the performance of intra-modal and cross-modal is obvious. As we discussed, the distortion on manifolds of unseen classes caused by aligning classes in the training set from different modalities is part of the reason.

Table 2 shows the retrieval performance by using the proposed bi-directional random walk scheme. The results are obtained by mining on homogeneous manifolds in the cross-modal feature only or by mining on heterogeneous manifolds in the single-modal feature spaces bridged by the cross-modal feature space. It is obvious that mining on heterogeneous manifolds outperforms mining on homogeneous manifolds in terms of rank-1 and mAP. Some may notice that the values of rank-5 to rank-20 are sometimes even lower when mining on heterogeneous manifolds. The explanation of this issue lies in the pipeline of our approach. The proposed method tries to look for true-positive samples among the single-modal feature space guided by a few locally nearest neighbors of the query found in the cross-modal feature space. This means that our method can obtain reliable retrieval results if only some true-positive samples exist in the initial search results in the cross-modal feature space, but still fails when none of the true-positives are found in the very first step. After all, our work focuses on filtering noisy similarities by mining on heterogeneous manifolds but not correcting the cross-modal feature distribution. Mining on heterogeneous manifolds fails when the initial k-NN search results contain none of the true-positives, while mining on homogeneous manifolds performs the random walk inside the cross-modal feature space and may find the first true-positive sample on the manifold which is not included in the nearest neighbors measured by the Euclidean distance. It is possible that we use the results of mining on homogeneous manifolds as the initial state for the random walk process of mining on heterogeneous manifolds, but it will double the computational cost which is not worthy. The improvements on mAP are more convincing. Mining on heterogeneous manifolds achieved 9.1% (48.3% to 57.4%) and 5.4% (52.0% to 57.4%) improvement compared with the baseline and mining on homogeneous manifolds when searching SVHN's query images in MNIST's gallery. Correspondingly, it outperforms the baseline and mining on homogeneous manifolds by 13.6% (46.7% to 60.3%) and 8.6% (51.7% to 60.3%) when searching MNIST's query images in SVHN's gallery.

	Manifold	Rank-1	Rank-5	Rank-10	mAP
S→M	Homo	61.0%	**76.2%**	**81.7%**	52.0%
S→M	Hetero	**61.7%**	74.0%	79.4%	**58.4%**
M→S	Homo	46.7%	**84.0%**	**96.4%**	51.7%
M→S	Hetero	**47.1%**	83.8%	92.2%	**60.3%**

Table 2: Cross-modal digit image retrieval performance by mining on homogeneous and heterogeneous manifolds using the proposed bi-directional random walk scheme. The MNIST and SVHN datasets are denoted by their initials and the first column implies the search direction formatted by {dataset of query} → {dataset of gallery}.

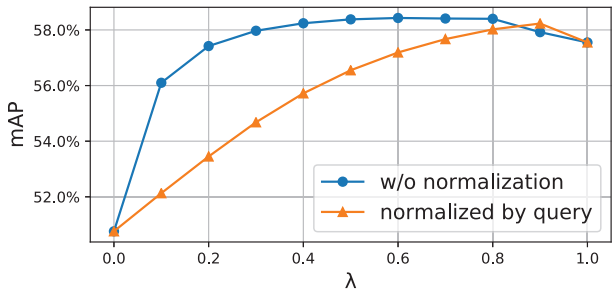


Figure 5: Performance *v.s.* the hyperparameter λ. The blue curve is plotted by ensembling similarity scores obtained from both search directions without normalization. While the orange one is plotted by ensembling similarity scores normalized by each query.

(a) Thermal image feature space (b) Visible image feature space (c) Cross-modal uniform feature space

Figure 6: Visualization of feature distributions in the thermal, visible, and cross-modal spaces. Images of 10 identities in SYSU-MM01's testing set are sampled and used to provide features. Features from the thermal and visible modalities are separately denoted by circles and triangles.

Dataset	Method	Rank-1	Rank-5	Rank-10	Rank-20	mAP
RegDB	One-stream (Wu et al. 2017)	13.1%	–	33.0%	42.5%	14.0%
	Two-stream (Wu et al. 2017)	12.4%	–	30.4%	41.0%	13.4%
	Zero-Padding (Wu et al. 2017)	17.8%	–	34.2%	44.3%	18.9%
	TONE (Ye et al. 2018a)	16.9%	–	34.0%	44.1%	14.9%
	HCML (Ye et al. 2018a)	24.4%	–	47.5%	56.8%	20.8%
	Baseline	24.5%	36.8%	44.4%	55.1%	23.9%
	Proposed (homogeneous)	30.1%	**45.5%**	**53.1%**	**66.8%**	29.3%
	Proposed (heterogeneous)	**31.1%**	42.3%	47.0%	58.6%	**32.1%**
SYSU-MM01	One-stream (Wu et al. 2017)	12.0%	–	49.7%	66.7%	13.7%
	Two-stream (Wu et al. 2017)	11.7%	–	48.0%	65.5%	12.9%
	Zero-Padding (Wu et al. 2017)	14.8%	–	54.1%	71.3%	16.0%
	TONE (Ye et al. 2018a)	12.5%	–	50.7%	68.6%	14.4%
	HCML (Ye et al. 2018a)	14.3%	–	53.2%	69.2%	16.2%
	BDTR (Ye et al. 2018b)	17.0%	–	55.4%	72.0%	19.7%
	cmGAN (Dai et al. 2018)	26.9%	–	67.5%	80.6%	27.8%
	Baseline	25.3%	51.5%	65.7%	80.2%	26.7%
	Proposed (homogeneous)	33.9%	60.3%	**75.6%**	83.8%	34.9%
	Proposed (heterogeneous)	**35.9%**	**61.5%**	73.0%	**86.1%**	**38.0%**

Table 3: Performance comparison with our baseline and previous works.

In addition, Fig. 5 shows the effect of the hyperparameter λ defined in Eq. (8) by measuring the mAP when searching digits from SVHN in MNIST. Usually, the two single-modal feature spaces are complementary and combining them by certain weights performs better than each side only (in the case $\lambda = 0$ or 1), which can also be interpreted as an ensemble trick. Since the graphs are normalized before the random walk, the similarity scores obtained from both directions are at the same numerical scale. It is observed that normalizing the scores again does not help improve the performance.

5.2 Person Image Retrieval

To provide more evidence of the effectiveness of our proposed method, we also conduct experiments on person re-identification datasets. Since persons in the testing set of two selected datasets never appear in their training set, it is fair to regard the person re-identification on these two datasets as zero-shot image retrieval tasks.

Fig. 6 visualizes the feature distributions for the SYSU-MM01 dataset. In the single-modal spaces of thermal and visible images, features of each identity are relatively grouped in clusters. However, in the cross-modal feature space, some samples are off the manifolds of their corresponding classes. We show this issue by highlighting two typical cases in Fig. 6c for identity #49 (in green) and identity #85 (in red). This, again, supports our assumption that mapping images from different modalities into a uniform space distorts the manifolds of zero-shot data.

Table 3 shows the retrieval performance of our proposed method and other previous works. We did not put much effort into training the baseline but it is still comparable to some methods we list. Mining on heterogeneous mani-

folds outperforms our baseline and mining on homogeneous manifolds in the cross-modal feature space only, which is in coherence with the results in Table 2. As discussed in the above section, the proposed method (heterogeneous) fails when the initial k-NN search results contain none of the true-positives, while mining on homogeneous manifolds performs the random walk inside the cross-modal feature space and may find the first true-positive sample on the manifold which is not included in the nearest neighbors measured by the Euclidean distance. This issue sometimes leads to worse rank-5 to rank-20 in the RegDB dataset. Unlike the rank values on the cumulative match characteristic (CMC) curve (DeCann and Ross 2013), the mAP protocol cares about both the precision and the recall. Our proposed method improved the mAP by 8.2% (23.9% to 32.1%) and 11.3% (26.7% to 38.0%) in the RegDB and SYSU-MM01 datasets respectively, which are no doubt significant improvements.

6 Conclusion

In this paper, we propose a novel bi-directional random walk scheme for zero-shot cross-modal image retrieval to exploit the heterogeneous manifolds in different single-modal spaces. To the best of our knowledge, this is the very first time that the random walk on manifolds is used in the cross-modal retrieval task. We show by experiments that our approach consistently achieves better retrieval performance on digit datasets as well as person re-identification datasets.

Acknowledgement

This study is supported by CREST (JPMJCR 1686).

References

Bai, S.; Bai, X.; Tian, Q.; and Latecki, L. J. 2017a. Regularized diffusion process for visual retrieval. In *AAAI*, 3967–3973.

Bai, S.; Zhou, Z.; Wang, J.; Bai, X.; Jan Latecki, L.; and Tian, Q. 2017b. Ensemble diffusion for retrieval. In *ICCV*, 774–783.

Cao, Y.; Long, M.; Wang, J.; Yang, Q.; and Yu, P. S. 2016. Deep visual-semantic hashing for cross-modal retrieval. In *ACM SIGKDD*, 1445–1454.

Cao, Y.; Long, M.; Wang, J.; and Liu, S. 2017. Collective deep quantization for efficient cross-modal retrieval. In *AAAI*, 3974–3980.

Chum, O.; Philbin, J.; Sivic, J.; Isard, M.; and Zisserman, A. 2007. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, 1–8.

Dai, P.; Ji, R.; Wang, H.; Wu, Q.; and Huang, Y. 2018. Cross-modality person re-identification with generative adversarial training. In *IJCAI*, 677–683.

DeCann, B., and Ross, A. 2013. Relating roc and cmc curves via the biometric menagerie. In *BTAS*, 1–8. IEEE.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*, 248–255.

Donoser, M., and Bischof, H. 2013. Diffusion processes for retrieval revisited. In *CVPR*, 1320–1327.

Hao, Y.; Wang, N.; Li, J.; and Gao, X. 2019. Hsme: Hypersphere manifold embedding for visible thermal person re-identification. In *AAAI*, 8385–8392.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 770–778.

Hermans, A.; Beyer, L.; and Leibe, B. 2017. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*.

Iscen, A.; Tolias, G.; Avrithis, Y.; Furon, T.; and Chum, O. 2017. Efficient diffusion on region manifolds: Recovering small objects with compact cnn representations. In *CVPR*, 2077–2086.

Iscen, A.; Tolias, G.; Avrithis, Y.; and Chum, O. 2018. Mining on manifolds: Metric learning without labels. In *CVPR*, 7642–7651.

LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P.; et al. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86(11):2278–2324.

Maaten, L. v. d., and Hinton, G. 2008. Visualizing data using t-sne. *Journal of machine learning research* 9(Nov):2579–2605.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning.

Nguyen, D.; Hong, H.; Kim, K.; and Park, K. 2017. Person recognition system based on a combination of body images from visible light and thermal cameras. *Sensors* 17(3):605.

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

Wang, Z.; Ye, M.; Yang, F.; Bai, X.; and Satoh, S. 2018. Cascaded sr-gan for scale-adaptive low resolution person re-identification. In *IJCAI*, 3891–3897.

Wang, Z.; Wang, Z.; Zheng, Y.; Chuang, Y.-Y.; and Satoh, S. 2019. Learning to reduce dual-level discrepancy for infrared-visible person re-identification. In *CVPR*, 618–626.

Wu, A.; Zheng, W.-S.; Yu, H.-X.; Gong, S.; and Lai, J. 2017. Rgb-infrared cross-modality person re-identification. In *ICCV*, 5380–5389.

Yang, F.; Hinami, R.; Matsui, Y.; Ly, S.; and Satoh, S. 2019. Efficient image retrieval via decoupling diffusion into online and offline processing. In *AAAI*, volume 33, 9087–9094.

Ye, M.; Lan, X.; Li, J.; and Yuen, P. C. 2018a. Hierarchical discriminative learning for visible thermal person re-identification. In *AAAI*.

Ye, M.; Wang, Z.; Lan, X.; and Yuen, P. C. 2018b. Visible thermal person re-identification via dual-constrained top-ranking. In *IJCAI*, 1092–1099.

Zhou, D.; Weston, J.; Gretton, A.; Bousquet, O.; and Schölkopf, B. 2004. Ranking on data manifolds. In *NeurIPS*, 169–176.