Elsevier Editorial System(tm) for Pattern Recognition Letters

Manuscript Draft

Manuscript Number:

Title: Learning Domain-invariant Feature for Robust Depth-image-based 3D Shape Retrieval

Article Type: SI:DLPR

Keywords: Discriminative neural network; cross-domain; depth images; 3D shape retrieval

Corresponding Author: Professor Yi Fang,

Corresponding Author's Institution: New York University

First Author: Jing Zhu

Order of Authors: Jing Zhu; John-Ross Rizzo; Yi Fang

Pattern Recognition Letters

Authorship Confirmation

Please save a copy of this file, complete and upload as the "Confirmation of Authorship" file.

As corresponding author I, Yi Fang, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Signature  Date March 9th, 2017

List any pre-prints:

Relevant Conference publication(s) (submitted, accepted, or published):

Jing Zhu, Fan Zhu, Edward wong, and Yi Fang, Learning Pairwise Neural Network Encoder for Depth Image-based 3D Model Retrieval, ACM MM 2015

Justification for re-publication:

Cover Letter

Dear Editor,

We are pleased to submit a research article entitled “Learning Domain-invariant Feature for Robust Depth-image-based 3D Model Retrieval” for consideration as a publication in *Pattern Recognition Letters Special Issue on Deep Learning for Pattern Recognition (DLPR)*.

In this article, we addressed the challenging issues in depth-image-based 3D model retrieval problem with a new design of pairwise neural network architecture extended from our recent work “*Learning Pairwise Neural Network Encoder for Depth Image-based 3D Model Retrieval*”, which has been published on ACM Multimedia Conference in 2015. The significant improvements can be concluded as 1) we improved the loss function of our pairwise neural networks by enforcing an additional constraint on intra-class margin to the object function for depth-image-based 3D model retrieval, and as a consequence, a deep domain-invariant representation with maximum inter-class margin and minimum intra-class variance can be learned for the two domains; 2) we conduct experiments on one more 3D dataset ModelNet10 to validate our improved approach; and 3) we also provide more analysis on the experimental performance of the approach, including example results of 3D shape retrieval and visualization of the learned domain-invariant features. The enhancement contributes the substantial difference (over 70%) between the conference paper and the submitting article.

This article has not been published, or under consideration for publication elsewhere. We would like to suggest the following international leaders with expertise in computer vision field to review our paper:

1) Dr. Hui Wu, IBM US, wuhu@us.ibm.com
2) Dr. Xiang Bai, HUST, xbai@hust.edu.cn
3) Dr. Xiaowei Zhou, University of Pennsylvania, xiaowz@seas.upenn.edu
4) Dr. Li Liu, University of East Anglia, liuli1213@gmail.com

We appreciate your time for reviewing our work. We are looking forward to your favorable decision soon.

Sincerely,

Jing Zhu, John-Ross Rizzo and Yi Fang

Research Highlights (Required)

To create your highlights, please type the highlights against each `\item` command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

- A robust domain-invariant feature is learned.
- A model connecting two discriminative neural networks is proposed.
- Loss function with metric learning is defined to handle cross-domain issues.
- Learned features obtain superior performance on depth-image-based 3D shape retrieval.
-

ELSEVIER

Pattern Recognition Letters
journal homepage: www.elsevier.com

Learning Domain-invariant Feature for Robust Depth-image-based 3D Shape Retrieval

Jing Zhu[a,b,d], John-Ross Rizzo[e], Yi Fang[a,c,d,**]

[a]*NYU Multimedia and Visual Computing Lab, USA*
[b]*Department of Computer Science and Engineering, NYU Tandon School of Engineering, New York, USA*
[c]*Department of Electrical and Computer Engineering, NYU Tandon School of Engineering, New York, USA*
[d]*Department of Electrical and Computer Engineering, New York University Abu Dhabi, Abu Dhabi, UAE*
[e]*Depts. of Rehabilitation Medicine and Neurology, NYU Langone Medical Center*

ABSTRACT

In the recent years, 3D shape retrieval has been garnering increased attention in a wide range of fields, including graphics, image processing and computer vision. Meanwhile, with the advances in depth sensing techniques, such as those used by the Kinect and 3D LiDAR device, depth image of 3D objects can be acquired conveniently, leading to rapid increases depth image dataset. In this paper, different from most of the traditional cross-domain 3D shape retrieval approaches, focused on the RGB-D image-based or sketch-based shape retrieval, we aim to retrieve shapes based only on depth image queries. Taking advantage of, but not limited to hand-crafted features, we proposed to learn a robust domain-invariant representation between shapes and depth image domains by constructing a pair of discriminative neural networks on the cross-domain data. Specifically, the networks are connected by a loss function with constraints on both inter-class and intra-class margins, which minimizes the intra–class variance while maximizing the inter-class margin among data from two domains (depth image and 3D shape). Our experiments on the NYU Depth V2 dataset (with Kinect-type noise) and two 3D shape (CAD model) datasets (SHREC 2014 and ModelNet) demonstrate that our proposed technique performs superiorly over existing state-of-the-art approaches on depth-image-based 3D shape retrieval tasks.

1. Introduction

3D shape retrieval has become an important topic in computer vision field with a wide range of applications in engineering, manufacturing, product design, and the medical field. Compared to the within-domain retrieval using 3D shape as queries, cross-domain shape retrieval, such as sketch-based shape retrieval and RGB image-based shape retrieval (Li et al. (2014a); Eitz et al.; Li et al. (2014c)), is a more attractive yet challenging problem. In recent years, due to the emergence of the low-cost depth sensor, e.g. the Kinect and 3D LiDAR systems, RGB-D images of objects can be captured easily. As a consequence, a number of large-scale RGB-D image datasets have become available and precipitated the problem of cross-domain shape retrieval. Although RGB-D images provide large amounts of information about objects and enables promising result towards shape retrieval, processing the complex RGB-D images usually requires higher computational costs on memory and time. A shape retrieval system driven only on depth images may be a more efficient and effective system. As shown in Figure 1, given a depth image query, a depth image-base shape retrieval system can return a set of relevant 3D models from a large database. As an example, product design users that simply capture the depth image of objects could greatly facilitate automated relevant 3D model selection, expediting the step-wise industrial process.

Due to the high diversity between 2D depth image and 3D shape representation format, it is very difficult to build a shape retrieval system by directly matching the depth image queries to corresponding shapes. We can also find the variations from some examples of depth images and their corresponding in Figure 2. To tackle the variation challenge, it is intuitively to convert the cross-domain data into the same domain. For ex-

**Corresponding author
e-mail: yfang@nyu.edu (Yi Fang)

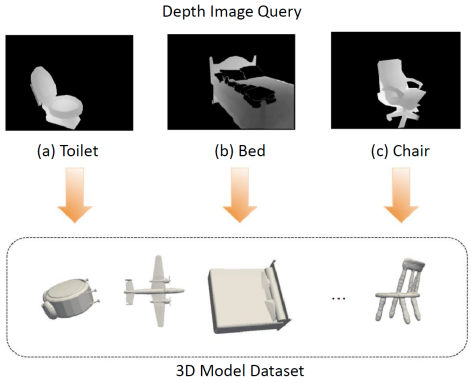

Fig. 1. Illustration of depth image-based 3D model retrieval. Given a depth scan of query sample, a set of relevant 3D models in a large database that are from the same category as query's can be retrieved.

Fig. 2. Examples of depth images (from NYU Depth V2 dataset) and their corresponding 3D shapes (from ModelNet dataset). As we can see from the figure, there is great variation between the representation between the 2D depth images and 3D shapes.

ample, in the early attempt, instead of directly retrieving 3D shapes from object depth images, Wang et al. (2014) transform the depth image-based shape retrieval problem to a reconstructed model-based shape retrieval problem, which takes a noisy 3D model (reconstructed based on depth images from multiple views) as input and output a relevant CAD model from the dataset. However, in most popular depth image dataset, especially for indoor scene, it is not practical for user to capture depth images of single object from many views. As we can see from the Figure 2, objects in most depth images are incomplete and captured from only one single angle, which makes the retrieval task even more challenging.

Recently, inspired by the successful application of autoencoders in the computer vision field, Feng et al. (2016) propose to first render depth images for each 3D model in the dataset and then train an autoencoder for each 3D model based on their rendered images. Finally, given an input depth image, they get a reconstructed depth image from each autoencoder and retrieve the 3D model by applying a potential model on the reconstructed depth images. The performance on a small subset of popular depth image datasets is promising, however, it is not easy to apply the approach on large-scale datasets since it needs to train one autoencoder for each 3D model. Therefore, we consider to construct a more generative deep neural network to learn a cross-domain representation for all 3D models and depth images. Due to the distinctive intrinsic properties between depth image domain and 3D shape domain, it is hard to build a network directly on 3D shapes and depth images, and we are seeking a way enable the connection between two domains.

On the other hand, hand-crafted features have shown their excellent performance on many challenging computer vision problems, for example, 3D SIFT have been proved its effectiveness with leading and robust results on 3D shape retrieval (Darom and Keller (2012)). However, most existing hand-crafted features are designed for single domain, either 2D images or 3D shapes, and due to the discrepancy, it is difficult to find an effective cross-domain feature work on both 2D depth image and 3D shapes. Although it is unpractical to design a cross-domain hand-crafted features manually, we can still utilize the advantages of existing hand-crafted features to first reduce the within-domain variation and then train our network upon the extracted features to handle the cross-domain issues. For better learning, we use two deep neural networks in our proposed model, one for the depth images and the other for 3D shapes. To connect the two networks, we define a loss function with constraints on both inter-class and intra-class margin, which maps the distinctive inputs into the same target space by minimizing the (intraclass) difference between cross-domain data within the same category while maximizing the (interclass) variation among data from different categories. Finally, the outputs of the leaned networks are considered as the domain-invariant representation for given cross-domain data, and relevant 3D shape can be retrieved by directly comparing the output features from the networks.

The experimental results on popular datasets, where depth images are from NYU Depth V2 dataset and 3D models come from SHREC 2014 database and ModelNet dataset, suggest that our proposed model significantly outperform other state-of-the-art approaches. What's more, once the networks are trained, we can perform efficient shape retrieval on given depth image queries since it only requires some matrix computation.

In summary, the main contributions of our work include:

★ To address the challenging depth-image-based 3D shape retrieval problem, we propose to learn a domain-invariant feature representations for cross-domain data.

★ To successfully learn domain-invariant features, we design a model with two neural networks connected and optimized by a loss function that maximizes the inter-class margin while minimizing the intra-class variance between heterogeneous cross-domain (Depth Image and 3D model) data.

★ The proposed method has been successfully validated on large datasets with superior performance over state-of-the-arts methods, including those applied on depth image-based shape retrieval, transfer learning methods used on similar task, and the approach that directly uses the original features of depth images and 3D models as representations for retrieval.

2. Related Work

Although cross-domain shape retrieval has received many attentions for years, most of them are on sketch-based or image-based shape retrieval. Recently, with the increase of the depth image datasets, researchers start to look at the depth image-based shape retrieval problem. In this section, we review three key components in depth image-based shape retrieval, including datasets, features and neural network.

2.1. Dataset

Started from decades ago, extending effort has been paid on building 3D shape datasets. Most of current popular 3D datasets contain thousands even millions of manually design CAD models for different kinds of objects. For example, SHREC 2014 Benchmark (Li et al. (2014c,b)) is one of the most popular 3D dataset in computer vision field, which is usually used for evaluation on sketch-based 3D shape retrieval. The Princeton ModelNet (Wu et al. (2015)) is another well known 3D model dataset providing a collection of clean CAD model for more than 600 categories. Other publicly available datasets, e.g. Princeton Shape Benchmark (Shilane et al. (2004)) and ShapeGoogle dataset (Bronstein et al.), are also widely used for within domain 3D shape retrieval.

With the advanced development of RGB-D cameras, a large number of RGB-D image datasets of different objects have been created. NYU depth V2 dataset (Silberman et al. (2012)) is a recent released RGB-D dataset providing a large collections of both RGB and depth image for diverse indoor objects, such as cup, desk, etc. The availability of such large scale RGB-D image dataset enables many application of using RGB-D image in solving computer vision problems, such as shape reconstruction and object detection (Gupta et al. (2014)), and leads the popularity of depth images in the graphic and computer vision communities.

2.2. Features

Due to the long history of research on 2D images, a lot of hand-crafted features are well defined for different tasks, such as image classification and object recognition, most of which are based on or extended from the classic bag-of-word model, e.g. SIFT features (Lowe (1999)), SPM features (Lazebnik et al. (2006)), ScSPM feature (Yang et al. (2009)), etc. Getting inspiration from those 2D image processing approaches, a number of hand-crafted features have been created to address the 3D shape retrieval challenges, such as calculating the probability distribution on geometric properties of an 3D model (Osada et al. (2002)) and finding the symmetry of shapes (Kazhdan et al. (2002)). Besides the above global descriptors, local characteristics have also been used for more robust shape retrieval. For example, Bronstein et al. (2011) bag the values of multiscale diffusion heat kernel as features to represent 3D models and Darom and Keller (2012) have successfully extend the well-known SIFT features Lowe (1999) on 3D shapes and achieve outstanding performance on shape retrieval task.

In addition to the hand-crafted features, learning-based feature are getting more and more popular to address either image or shape problems. As a special machine learning paradigm, transfer learning are mainly used to tackle with the domain mismatch problem. Most current existing transfer learning methods (Hu and Fang (2014); Li et al. (2014d); Zhu and Shao (2014)) operate at the features learning level, and aim to obtain a unified representation for two or more mismatched domains (e.g., sketch images vs. 3D shapes, images vs. 3D shapes and images vs. texts). Rasiwasia et al. (2010) address the image-to-text and text-to-image retrieval problem by investigating the correlations between two modalities, and measuring the effectiveness of abstraction. In their work, both the canonical correlation analysis (CCA) and the use of abstraction are proved to be effective for retrieval task. To evaluate the contributions of each separate component, three approaches – correlation matching (CM), semantic matching (SM) and semantic correlations matching (SCM) – were proposed for correlation modeling, the abstraction method and the joint working mode of both approaches, respectively.

2.3. Neural Network

Inspired by biological neural networks, artificial neural network is a system containing a number of processing elements, which provides dynamic outputs according to external inputs. The simplest type of neural network is the perceptron, created by Rosenblatt (1958). Later, Werbos (1974) introduce the backpropagation algorithm making neural networks even more popular for machine learning. Nowadays, neural network techniques have achieved great success on various real-world applications, such as biomedicine (Liu and Yang (2015)), energy (Barbounis et al. (2006)), telecommunications (Yuhas and Ansari (2012)), geophysics (Aminzadeh et al. (2013)), etc. In addition to the traditional neural network structure, there has been increasing interest in deep learning neural networks in recent years (LeCun et al. (1998); Krizhevsky et al. (2012); Sermanet et al. (2013)), especially in convolutional neural networks (CNN). Malinowski et al. (2015) proposed an approach for task of answering of questions about images by combining a CNN with a LSTM into an end-to-end architecture that predict answers conditioning, and Pfister et al. (2015) utilized CNN to estimate human pose in videos by combining information across the multiple frames using optical flow.

Feng et al. (2016) attempt to apply neural network on depth image-based shape retrieval. Multiple autoencoders are trained on rendered depth images from corresponding 3D models, one autoencoder for one model. Given a depth image query, the retrieval is performed based on the reconstructed depth images generated from each autoencoder. Generalizing their method on large dataset could be very expensive since every model needs to train its own autoencder for representation. Despite the effort from Feng et al., Zhu et al. (2015) recently propose to build a pair of neural networks for depth image-based 3D shape retrieval. However, random variables are assigned as target vectors to connect the networks in their work, which makes the retrieval performance heavily depend on the initialization of the random values. Their consideration on within-class variation only also limits the performance of shape retrieval. In this paper, we focus on eliminating these shortcomings by connect-

ing the network pair with a loss function that constraints the inter-class difference as well as the intra-class variation.

3. Approach

We propose to learn a domain-invariant representation for depth image-based shape retrieval using two discriminative neural networks, one for each domain. In the section, basic concepts in neural network are first introduced in section 3.1 and followed by a presentation for our network structured in section 3.2.

3.1. Discriminative Neural Network

A classic single-layer neural network consists of an input layer, a hidden layer and a target layer. Except the input layer, both hidden layer and output layer are constructed by a number of basic components – neurons, which computes the output of given input from the previous layer based on a predefined activation function. Due to the different purposes of tasks, the activation function in neurons could be varied, e.g. sigmoid function, tangent function or binary function. Take the neuron at hidden layer as an example, the output of each neuron can be calculated as:

$$h_k = f\left(\boldsymbol{x}_n * \boldsymbol{W} + b\right), \tag{1}$$

where $f(x_i)$ is the activation function, and W denotes the weight for each input and b is the bias term. More generally, the output of neuron at any layer can be represented as:

$$y_k^l = f\left(\boldsymbol{y}_i^{l-1} * \boldsymbol{W}^{\boldsymbol{l}} + b^l\right), \tag{2}$$

where y_k^l is the output value of the k^{th} neuron at layer l. A loss function is usually defined on the output at the last layer in the network, which is compared with a desired target vector. One typical loss function has the following form:

$$\boldsymbol{E} = \arg\min_{\boldsymbol{W},\boldsymbol{b}} \frac{1}{N} \sum_{i=1}^{N} \|\hat{\boldsymbol{y}}_i - \boldsymbol{y}_i\|_2^2 + \lambda \sum_{l=1}^{L} \|\boldsymbol{W}^l\|_F^2, \tag{3}$$

where N is the total number of training samples, $\hat{\boldsymbol{y}}_i$ is the target vector for i^{th} sample, L is the number of layers and λ is the regularization term, which restricts the sparsity of weights. When training sample processing through the network, the network parameters are optimized based on the error computed by the above equation using classic back-propagation algorithm. During training, for any training sample, the output of the network is getting closer to its target vector.

3.2. Cross-domain Matching

In this section, we propose a method to match samples from two domains (depth image domain and 3D shape domain) without any reconstruction on either domain. We first provide an illustration of the cross-domain matching problem. Then, we propose our approach to use two networks handle diverse data from different domains. By projecting the cross-domain data into same feature space, the comparison between depth image and 3D model can be conducted directly.

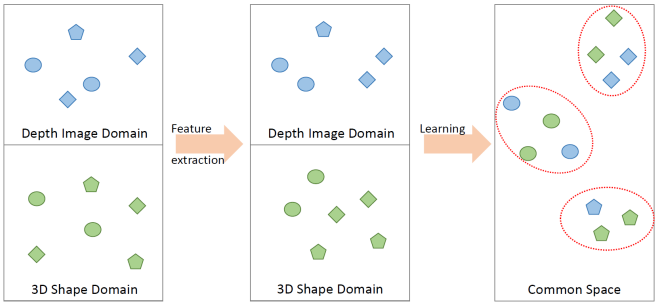

Fig. 3. The illustration of our approach. Shapes in Blue represent the samples in depth image domain while shapes in Green denote samples in 3D shape domain. As shown in left, the samples in the original domain might lack of discrimination. Although samples within same domain could be more classified after feature extraction, samples in two domains are still in different feature space (described in middle). Our learning model maps the samples into the same feature space and enables the comparison.

3.2.1. Illustration

Depth image-based shape retrieval is a typical cross-domain matching problem. Since the intrinsic variance between the two domains, it is difficult for us to do the retrieval with direct matching in their original domains. Therefore, we consider to map the highly discriminative features (from two domains) to a common feature space, where the closer feature points are more likely to share the same class label and the further points are more likely belong to different class. Our idea is clearly illustrated in Figure 3. Shapes in Blue represent the samples in depth image domain while shapes in Green denote samples in 3D shape domain. Samples in different shapes are from different classes while samples with same shape come from the same class. As we can discover from the figure, samples in the original domains might lack of discrimination with some kind of mix. Hand-crafted features can be extracted from original samples, and after that samples within same domain could be more classified. Most of the within domain retrievals can achieve good performance utilizing hand-crafted features, however, samples in two domains are still in different feature space. In this paper, we propose to build a learning model using neural network, which maps the two domain samples into the same feature space that cross-domain samples belonging to same class have similar representation. This also enables the easy comparison for matching on cross-domain data.

3.2.2. Network Architecture

Improving the data smoothness, the basic discriminative neural network is usually used on data from the same domain. In order to reduce the discrepancy between two domains (depth image domain and 3D shape domain), we adopt two neural networks in our model, one for the depth image domain and the other for 3D shape domain. The two networks are "aligned" with one loss function at the target layer. Figure 4 shows the architecture of our network, where same network design is used for both depth image and 3D shape.

Let $\boldsymbol{d}_i$, $\boldsymbol{s}_j$ be the extracted features from depth image and shape, which are also inputs for depth image and 3D shape network, respectively. Then the output of network can be com-

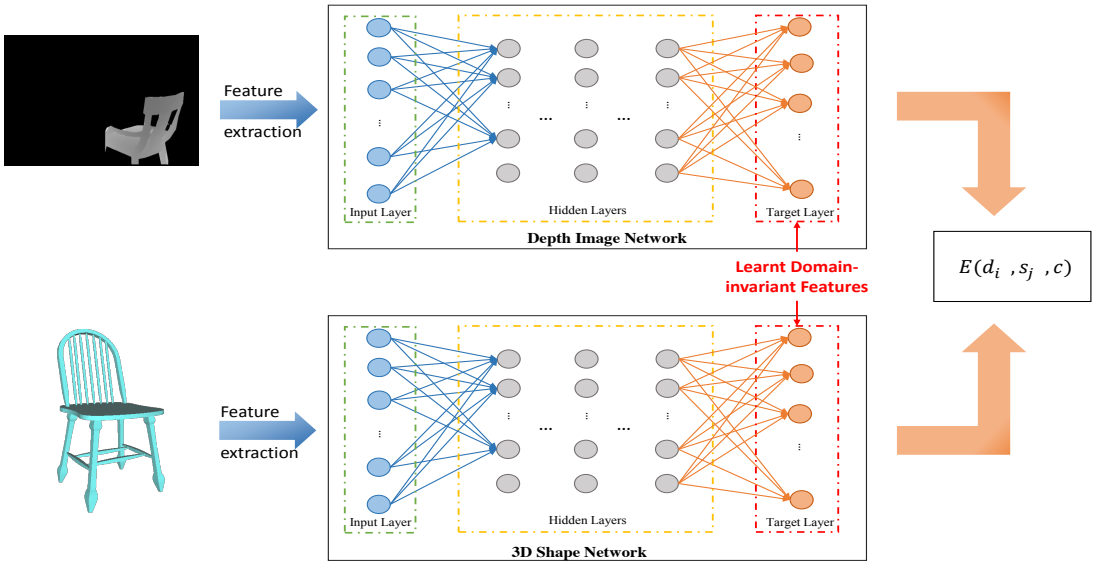

Fig. 4. The structure of our proposed method, where two neural networks with same architecture, one for each domain, are used to handle the cross-domain issues. The domain-invariant feature can be learned by connecting the two networks with a loss function on the outputs.

puted following:

$$\begin{aligned} &\text{for} \quad l = 1, \quad \boldsymbol{D_i^l} = f\left(\boldsymbol{d}_i * \boldsymbol{W_d^l} + b_d^l\right) \\ &\text{for} \quad l = 2, ..., L, \quad \boldsymbol{D_i^l} = f\left(\boldsymbol{D_i^{l-1}} * \boldsymbol{W_d^l} + b_d^l\right), \end{aligned} \tag{4}$$

where $\boldsymbol{D_i^l}$ denotes the output of each layer in the depth image network given input d_i, and $\boldsymbol{W_d^l}$, $\boldsymbol{b_d^l}$ are the depth image network parameters. Sigmoid function $f(z)$ is adopt as the activation function for the neurons in our network:

$$f(z) = \frac{1}{1 + exp(-z)}. \tag{5}$$

Similarly, the output of 3D shape network can be obtained from:

$$\begin{aligned} &\text{for} \quad l = 1, \quad \boldsymbol{S_j^l} = f\left(\boldsymbol{s}_j * \boldsymbol{W_s^l} + b_s^l\right) \\ &\text{for} \quad l = 2, ..., L, \quad \boldsymbol{S_j^l} = f\left(\boldsymbol{S_j^{l-1}} * \boldsymbol{W_s^l} + b_s^l\right), \end{aligned} \tag{6}$$

where $\boldsymbol{S_j^l}$ denotes the output of each layer in the 3D shape network for input s_j, and $\boldsymbol{W_s^l}$, $\boldsymbol{b_s^l}$ are shape network parameters. To connect the two networks, a loss function is designed based on the output at the target layer (last layer of the networks). In the traditional neural network, the loss function is defined as the variance between the output of the network and its target vector, however, it is pretty hard to define a perfect target vector for cross-domain data by human being, but it is possible to learn a feature space suitable for cross-domain data using machine learning approaches. Inspired by metric learning technique, our network takes a pair of samples as input at each time during training, one from each domain, and our loss function is composed by two terms: the inter-class margin and the intra-class margin. If the sample pair comes from the same category, then the variance between outputs of networks is considered as the intra-class margin. Otherwise, the variance can be seen as inter-class margin. The loss function has the following form:

$$\begin{aligned} \boldsymbol{E(d_i, s_j, c_{ij})} = \frac{1}{N_p} \sum_{i=1}^{N_d} \sum_{j=1}^{N_s} (c_{ij} \|\boldsymbol{D_i^L} - \boldsymbol{S_j^L}\|_2^2 - (1 - c_{ij}) \|\boldsymbol{D_i^L} - \boldsymbol{S_j^L}\|_2^2) \\ + \lambda \sum_{l=1}^{L} (\|\boldsymbol{W_d^l}\|_F^2 + \|\boldsymbol{W_s^l}\|_F^2), \end{aligned} \tag{7}$$

where $\boldsymbol{N_p}$ is the number of training pairs, $\boldsymbol{N_d}$ is the number of depth image samples, $\boldsymbol{N_s}$ is the number of shape samples and $\boldsymbol{c_{ij}}$ is the relationship label between the samples $\boldsymbol{d_i}$ and $\boldsymbol{s_j}$. If the two inputs are from the same class, then $\boldsymbol{c_{ij}}$ equals to 1, otherwise, $\boldsymbol{c_{ij}} = 0$. By minimizing the loss generated from Eq. 7, a common feature space can be learnt where data from both domains with minimum intra-class margin and maximum inter-class margin.

The network training process is an optimization problem which aims to get optimum parameters of the network so that the loss is as small as possible. We adopt the classic backpropagation algorithm, which can efficiently compute the partial derivatives and update the parameters with gradients, to obtain the optimum parameter. Once we obtain the optimum $\hat{\boldsymbol{W}}_d$, $\hat{\boldsymbol{b}}_d$, $\hat{\boldsymbol{W}}_s$ and $\hat{\boldsymbol{b}}_s$, given any depth image queries or 3D models, outputs of corresponding networks at the target layers are extracted as the domain-invariant representations. Please note that the two networks can be used to generate domain-invariant features from either depth images or 3D models, independently. 3D models are then retrieved based on computing the Euclidean distance between output features from the depth image query and the 3D models:

$$Dist(\hat{\boldsymbol{D}}_i, \hat{\boldsymbol{S}}_j) = \sqrt{\sum_{k=1}^{m} (\hat{\boldsymbol{d}}_i^k - \hat{\boldsymbol{s}}_s^k)^2}. \tag{8}$$

where $\hat{\boldsymbol{D}}_i, \hat{\boldsymbol{S}}_j$ denote the domain-invariant features for the i^{th} depth image query and the j^{th} 3D model from the database, m

is the dimension of the output features. The difference between the learnt features of the depth image and the 3D model from the same category should be small while the variance is large among those from varied class. We rank the distances computed by Eq. 8 in ascending order for each query and generate a distance matrix. Then, 3D models with smaller distances in the matrix are retrieved as relevant ones for each query.

4. Experiments

To validate the performance of our proposed method, we comprehensively evaluate our algorithm on one large depth image dataset and two 3D model datasets by conducting experiments with various settings, and provide retrieval performance using common evaluation metrics and precision-recall curves. In all experiments, our method outperforms the state of the arts, which demonstrates that our proposed model can successfully learn the domain-invariant features for both domains (depth image and 3D shape).

4.1. Datasets

Depth Image Dataset The queries of our proposed method are labeled depth images from NYU Depth V2 dataset (Silberman et al. (2012)), which is comprised of video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect. The NYU Depth V2 dataset contains $1,449$ densely labeled pairs of aligned RGB and depth images with 894 object categories. Since there may be more than one object on a single image, we extract the corresponding depth image regions with different class labels and collect them as multiple object images. In order to guarantee that both the depth image dataset and the 3D model dataset have the same object categories, we use $2,517$ depth images for 10 categories in the NYU Depth Dataset V2. The $2,517$ depth images are further split into training dataset and test dataset with 1:1 ratio.

3D Model Datasets The 3D models for our experiment are from two recent large datasets. One is the large-scale extended SHREC 2014 sketch-based 3D shape retrieval benchmark (Li et al. (2014c,b)) and the other is the ModelNet dataset (Wu et al. (2015)) from Princeton. The SHREC 2014 benchmark contains $8,987$ 3D shapes from 171 categories. The number of 3D shapes in each category varies from 1 to 632. In order to match the object categories of depth image dataset, the database is constructed by selecting 3D models in corresponding categories, which contains $2,174$ 3D shapes from 7 categories. All of the $2,174$ 3D models will be used for both training and testing. For another test on ModelNet dataset, a subset of $4,315$ clean 3D models from 10 categories (ModelNet10) (corresponding to the sample categories from NYU Depth V2 dataset) are used as the 3D shape dataset. The ratio for samples in training set and test set is 1:1.

4.2. Compared Approaches

To evaluate the performance, we compare our method with other state-of-the-art method (Zhu et al. (2015)) that address the same depth-image-based 3D shape retrieval problem. Besides that, we also select some state-of-the-art transfer learning approaches used for similar cross-modality retrieval and the approach that retrieves models using extracted raw features for comparison.

The Non-transfer Approach (NT) Without any further learning and processing, the non-transfer approach directly retrieve 3D models utilizing the original extracted features as representation for depth image and 3D shapes. The shape retrieval is performed by computing the Euclidean distance between retracted features of queries and features of 3D models.

Transfer Learning Methods Correlation matching (CM) and Semantic Correlations Matching (SCM) are state-of-the-art transfer learning approaches to cross-modality multimedia retrieval (Rasiwasia et al. (2010)). CM learns correlations between two domains with canonical correlation analysis (CCA) (Hotelling (1936)), and the maximal cross-modality correlations are used for retrieval. SCM is an extension of CM with a higher level of abstraction of the domains, where logistic regression is performed on the maximal correlations that are obtained from CM. We directly applied the source code released by the authors on the depth image dataset and the 3D model datasets. We use the source code given by authors.

Pairwise Neural Network (PNN) The state-of-the-art approach for cross-domain 3D shape retrieval with depth images as queries. Zhu et al. (2015) train a pair of neural networks based on the features extracted from samples independently. Identical target vectors are assigned for samples from the same category for both networks. The outputs from the hidden layer of the neural network pair are extracted as features to retrieve relevant 3D models for given depth images. The code is provided by authors.

Proposed Method (Ours) We construct our model using two deep neural networks, one for each domain. Our model is learning a domain-invariant features for both domains by a loss function that minimizes the intra-class distance while maximizing the inter-class variance of cross-domain data. Given a depth image query, shapes are retrieved based on the similarity of features generated by the final outputs of the networks.

4.3. Evaluation Protocol

In our experiments, the performance of all compared approaches for retrieval are evaluated based on the common widely used six evaluation metrics (Shilane et al. (2004)) and precision-recall (PR) curves.

Evaluation Metrics The evaluation metrics include six quantitative statistics (Nearest Neighbor, First Tier, Second Tier, E-Measure, Discounted Cumulated Gain and Average Precision) for evaluating the match retrieval results. Nearest Neighbor (NN) is the average percentage of the closest 3D models that belong to the same category as the depth image queries. Supposed C denotes the total number of 3D models that are in the same category of the query's, First Tier (FT) is the mean percentage of 3D models that are in the same category as the queries' within the top $|C| - 1$ matches, and Second Tire (ST) is the mean percentage of relevant 3D models within the top $2 * |C| - 1$ matches for all queries. E-Measure (E) is the mean

of E_q computed by the precision (P_{32}) and recall (R_{32}) of the first 32 retrieved models for every query as $E_q = \frac{2}{\frac{1}{P_{32}} + \frac{1}{R_{32}}}$. With the assumption that matches appearing closer to the top of the ranked list are more relevant, larger weights are assigned for the matches near the top, and Discounted Cumulated Gain (DCG) is the average weighted sum of correct results of a query in the ranked list. Average Precision (AP) computes the average precision of the retrieval for all queries. For all six statistics, higher values indicate better performance.

Precision-Recall (PR) Curves To visualize the performance of retrieval results, precision-recall curves are used to indicate the relation between precision and recall for all queries. They are generated by calculating the standard 11-point interpolated average precision at different recall levels of $0.0, 0.1, \cdots, 0.9, 1.0$. For a recall level i, interpolated precision is the maximum precision at any recall level that is larger than or equal to i. After obtaining the 11 average precision points, we plot them on a two-dimensional graph with recall on the x-axis and precision on the y-axis.

4.4. Experimental Settings

Before setup our network model, we first extract original features from depth images and 3D models, which are used as inputs for our networks. We follow the Sparse Coding Spatial Pyramid Matching (ScSPM) (Yang et al. (2009)) framework to generate features for depth image network. After obtaining 21504-dimensional ScSPM features, Principal Component Analysis (PCA) (Wold et al. (1987)) is applied to the features, and we reduce the dimension of the depth ScSPM features to 1000. For 3D shapes, we extract Local Depth Scale-Invariant Features Transform (LD-SIFT) (Darom and Keller (2012)) features and then fit the LD-SIFT feature to a Bag-of-Words (BoW) model to get 1000-dimensional histogram features for each 3D model from the dataset.

In the experiments, our network model is constructed by two 3-layer neural networks with 500 hidden layer size and 1000 target layer size, one for each domain. We report results from two experiments in this paper. For the two experiments, depth image all from NYU Depth V2 dataset, but 3D models are varied. In the first experiment, the 3D models are selected from SHREC 2014 benchmark dataset, while in the other test, 3D models are collected from the ModelNet dataset. The network structure remains same for the two tests. We can obtain a 1000-dimensional cross-domain representation for both the depth image queries and the 3D models after applying our model. When training the neural networks, the learning rate β and regularization term λ are set to different values in different experiments.

4.5. Shape Retrieval on SHREC 2014 Dataset

Following the experiment setting given by Zhu et al. (2015), we test our method using 5-class samples (the first 5 categories as displayed in Table 1) and 7-class samples (as displayed in Table 1) from the datasets. Our model is trained with a depth image training set containing 50% of the depth images random selected from each category (670 for 5-category test and $1,014$ for 7-category test). The rest of the depth images are used as for testing. All 3D models are used in both training and testing.

Table 1. Numbers of samples in each category in constructed dataset, where depth images and 3D models are from NYU Depth V2 dataset and SHREC 2014 benchmark, respectively.

Category	bathtub	bed	chair	desk	dresser	night stand	table
Depth images	57	318	654	197	111	148	539
3D models	109	467	712	204	203	218	402

Table 2. Performance metrics comparison of depth-image-based 3D shape retrieval on the NYU Depth V2 dataset and the SHREC 2014 benchmark.

	NN	FT	ST	DCG	E	AP
5 categories						
NT	0.04	0.21	0.39	0.71	0.03	0.21
CM	0.12	0.19	0.39	0.71	0.03	0.20
SCM	0.20	0.18	0.38	0.70	0.02	0.20
PNN	0.52	0.39	0.58	0.78	0.06	0.42
Ours	**0.53**	**0.56**	**0.75**	**0.84**	**0.07**	**0.63**
7 categories						
NT	0.23	0.14	0.30	0.66	0.02	0.15
CM	0.14	0.14	0.27	0.65	0.02	0.15
SCM	0.14	0.14	0.27	0.65	0.03	0.15
PNN	0.37	0.26	0.40	0.71	0.05	0.28
Ours	**0.40**	**0.44**	**0.61**	**0.79**	**0.05**	**0.51**

We compared our approach with recent PNN (Zhu et al. (2015)) and the state-of-the-art transfer learning methods: correlation matching (CM) and semantic correlations matching (SCM) (Rasiwasia et al. (2010)). We also compared our method with the non-transfer approach, which directly utilizes the original depth image extracted ScSPM features and 3D model extracted LD-SIFT features for retrieval. The statistic results are reported in Table 2 with six standard evaluation metrics and in Figure 5 with precision-recall (PR) curves generating from all compared methods. The retrieval performance of our method is obtaining with the learning rate β and regularization term λ set to 0.02 and 0.0005, respectively.

Experimental results suggest that the proposed method achieves outstanding performance when compared with other methods (PNN, CM, SCM and NT) on 7-category and 5-category datasets under the 6 metrics. As we can see in Table 2 and Figure 5, our method consistently leads large margin over other state-of-the-art PNN approach, transfer learning methods and non-transfer approach on the six performance metrics and the PR-curves. This demonstrates the significant improved performance of our model over other methods on cross-domain retrieval and the importance of adding inter-class term on our loss function.

4.5.1. Shape Retrieval on ModelNet 10 Dataset

In this section, we use a subset of ModelNet dataset with 10-category 3D models as our 3D shape database (ModelNet10

Table 3. Numbers of samples in each category of constructed dataset, where depth images are from NYU Depth V2 dataset and 3D models come from ModelNet10 dataset.

Category	bathtub	bed	chair	desk	dresser	monitor	night stand	sofa	table	toilet
Depth images	57	318	654	197	111	118	148	307	539	68
3D models	148	514	869	237	244	485	245	690	445	418

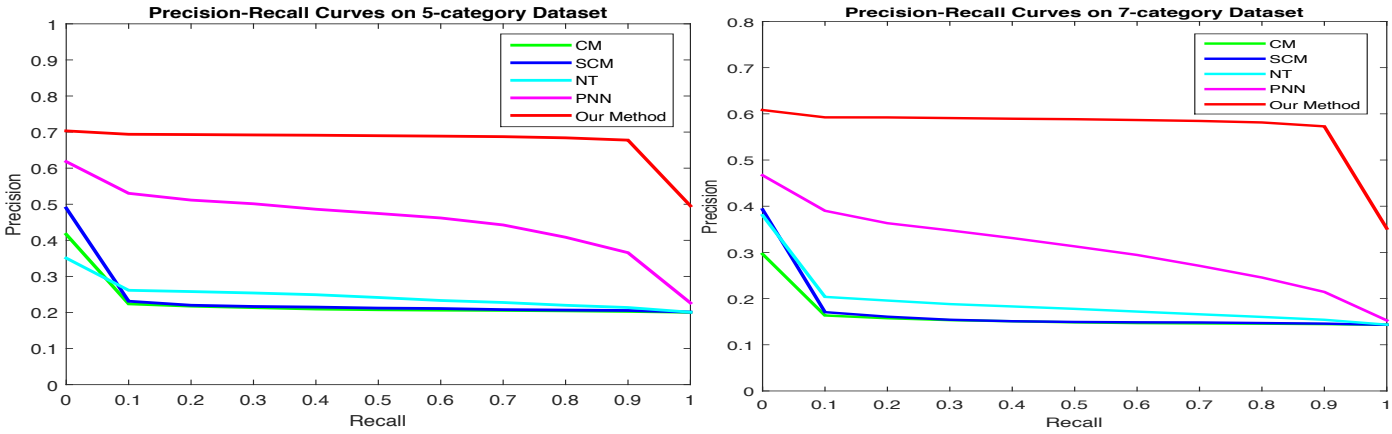

Fig. 5. Precision-Recall plot for performance comparison of state-of-the-art methods on NYU Depth V2 dataset and SHREC 2014 benchmark. The left plot shows the comparisons on 5-category dataset and the right one shows the comparisons on 7-category dataset.

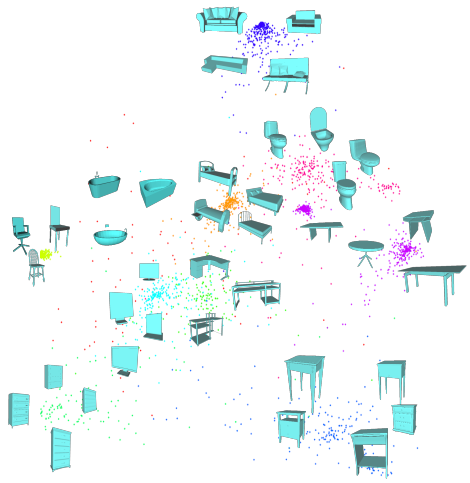

Fig. 6. Visualization of the learned domain-invariant features using our method. Points represent the learned features for both depth image and shape domains. In the figure, we only show some shapes as examples to demonstrate the corresponding class. Cross-domain data from the same category is assigned with same color.

Table 4. Performance metrics comparison of depth-image-based 3D model retrieval on the ModelNet10 dataset and NYU Depth V2 dataset.

	NN	FT	ST	DCG	E	AP
NT	0.14	0.13	0.25	0.68	0.01	0.14
CM	0.10	0.12	0.24	0.68	0.01	0.13
SCM	0.18	0.18	0.31	0.71	0.02	0.20
PNN	0.14	0.13	0.26	0.65	0.03	0.15
Ours	**0.33**	**0.33**	**0.46**	**0.76**	**0.04**	**0.41**

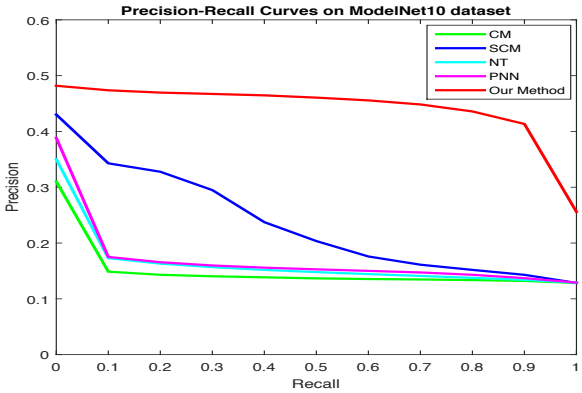

Fig. 7. Precision-Recall plot for performance comparison of state-of-the-art methods on NYU Depth V2 dataset and ModelNet 10 dataset.

dataset). The depth image queries are the same with the experiment on SHREC 2014 dataset. The numbers of samples in 10 categories (*bathtub*, *bed*, *chair*, *desk*, *dresser*, *monitor*, *night stand*, *sofa*, *table*, *toilet*) of the two datasets are given in Table 3. Both the depth images and 3D models are split into training set and testing set with 1:1 ratio. Therefore, there are $1,261$ depth image and $2,160$ 3D models in the training set, while $1,256$ depth image and $2,155$ 3D models in the testing set.

When training our model on ModelNet10 dataset, we set the learning rate β and regularization term λ to 0.001 and 0.0001, respectively. We first provide an example to visualize our learned domain-invariant features in Figure 6 by simply reducing the dimension of the learned features to two using PCA algorithm. Points in the same color represent the cross-domain samples from the same category, and some example shapes are placed next to their corresponding points for better review. As we can see from the visualization figure, most of the cross-domain samples have similar features if they are in the same class. The effective domain-invariant feature learning enables the better performance when applying our proposed method on depth image-based shape retrieval.

We also present the statistic results in Table 4 and compare the precision-recall curve against the state of the art methods in Figure 7. From the Figure 7, we can see that our method significantly outperforms other comparison methods. More importantly, the whole curve decreases much slower than other methods when the recall increases, which suggests that our method is more stable. The performance gain of our method is more than 10% when recall reaches 1. As show in Table 4, our method performs better in every metric against other methods, which further demonstrates our method is superior. On the other hand, we observe that PNN surprisingly performed the same as the without learning NT method. The cause might come from the predefined target vector with random values, which could be similar for different classes. The successful retrieval on both

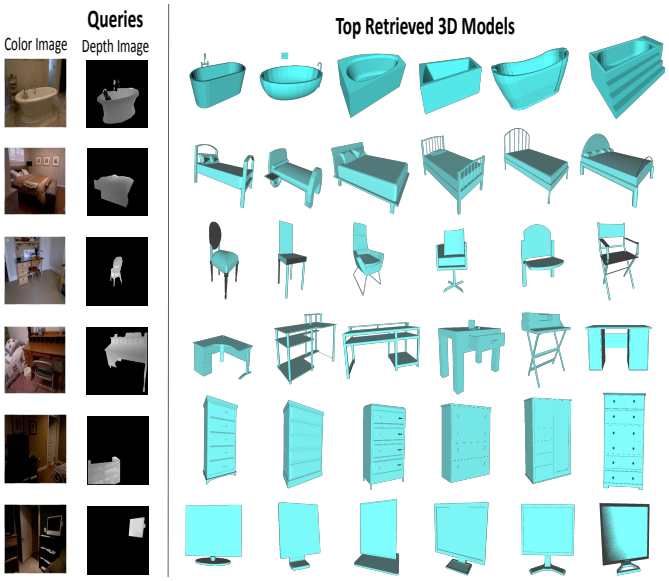

Fig. 8. Examples of successful retrieved results using our proposed method (3D models are from the ModelNet10 dataset). Please note that we did not use any information from color images and only use depth images as queries. From top to bottom the queries are *bathtub, bed, chair, desk, dresser* and *monitor*. Each row shows the top 6 retrieval results for corresponding query.

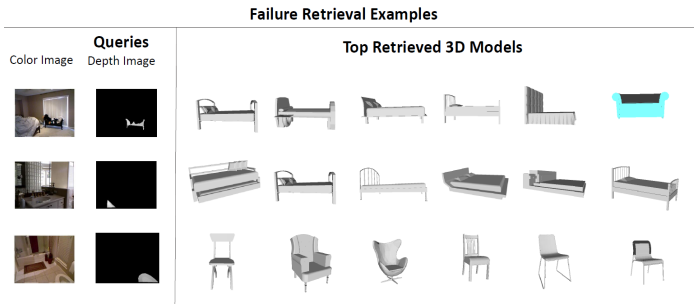

Fig. 9. Examples of failure retrieved results. From top to bottom the queries are *sofa, table* and *toilet*, following with their top 6 retrieval results. Shape in Cyan denote correct retrieval and shape in Gray denote the incorrect retrieval.

SHREC 2014 dataset and ModelNet10 dataset greatly demonstrate the robustness of our proposed method.

Finally, we visualize some example of successful retrieval results for queries from different category in Figure 8. For better review, corresponding color image is provided for each query in the first column, but please note that we did not use any information from color images. The second column shows the depth queries and each row shows the top 6 retrieval results. The retrieval results demonstrate our method is powerful in learning the features for cross-domain data. In addition, Figure 9 presents some failure retrieval results, in which queries from *sofa* (in first row) and *table* (in second row) retrieve *bed* as their top results. A query from *toilet* might be matched to some *chair* models. We conclude the possible reason to 1) the significant incompleteness of the object in some depth images, for example, the second and third query in Figure 9 only provide a very small part of object; 2) since the depth images are captured from the real-world environment, there are some occlusion of objects, e.g. the first query of *sofa* is fully covered by a lot of stuffs, such as toys, cloths and cushions, making it have similar view with *bed*. Although there are some failure cases, the statics evaluation strongly demonstrate the effectiveness of our method with superior performance on depth image-based shape retrieval.

5. Translational Applications

A method that optimizes 3D-shape identification, as rendered by depth images, has great capacity to facilitate computer vision-focused object identification. More specifically, a system that is able to minimize the computational costs of time and memory will enable the re-allocation of processing power to additional undertakings. This becomes pertinent to translational medical applications that leverage deep learning techniques with real-time object detection/categorization needs. In these approaches, identified 3D shapes may aid in the algorithmic strategies deployed to fuse such information with other inputs, towards the goal of minimum delay semantic labeling with object identification. This operation could take place in parallel with local scene understanding, juxtaposing object identity with concurrent spatial understanding in dynamic environments.

In the visually impaired setting, wearable devices that are configured as assistive technology platforms are one such application where these approaches become attractive. Systems that focus on real-time spatial understanding along with on-board navigation instructions will require expedited obstacle identification methodologies that are performed simultaneously with a host of additional tasks and sub-tasks (Patel et al. (2016); Rizzo et al. (2016); Rizzo et al.). While deep learning and computer vision techniques are still in their nascent stages, as applied to the aforementioned use cases, these systems have the potential to drastically improve the mobility profile of those with low vision or blindness and to reverse the untoward co-morbidities that arise as a result of increased immobility, often a byproduct of trips, falls, and injuries (Sengupta et al. (2015); Van Landingham et al. (2012); Harwood (2001); Lord et al. (2010); McLean et al. (2014); Popescu et al. (2011)).

6. Conclusions

In this work, we propose to learn a domain-invariant feature using a deep neural network in an effort to address the challenging problem of depth-image-based 3D shape retrieval. In order to minimize the discrepancy between highly diverged depth images and 3D models, we build a neural network pair for the depth images and the 3D models, while connecting the network pair at their target layers. Instead of enforcing identical fixed target values at the output layers of both networks, we add a constraint on the inter-class margins of the loss function, and enable the neural network pair to adjust the target values towards minimum intra-class variance and maximum inter-class margin during the training process. Our proposed method was successfully validated on the NYU Depth Dataset V2, the extended SHREC 2014 3D shape retrieval benchmark, and Princeton ModelNet dataset. Experimental results showed that our approach outperformed the state-of-the-art PNN method,

other transfer learning methods, and the paradigm that retrieves 3D models by directly using the original extracted features from depth images (ScSPM) and 3D models (3D SIFT). The large improvement margins of our method over existing techniques demonstrates its excellent capacity for cross-domain data representation. Moreover, since our model does not require the correspondence information across different domains, it can be easily generalized to solve real-world problems, including those that focus on translational medical applications.

References

Aminzadeh, F., Sandham, W., Leggett, M., 2013. Geophysical applications of artificial neural networks and fuzzy logic. volume 21. Springer Science & Business Media.

Barbounis, T.G., Theocharis, J.B., Alexiadis, M.C., Dokopoulos, P.S., 2006. Long-term wind speed and power forecasting using local recurrent neural network models. Energy Conversion, IEEE Transactions on 21, 273–284.

Bronstein, A.M., Bronstein, M.M., Bustos, B., Castellani, U., Crisani, M., Falcidieno, B., Guibas, L.J., Kokkinos, I., Murino, V., et al., . Shrec 2010: robust feature detection and description benchmark .

Bronstein, A.M., Bronstein, M.M., Guibas, L.J., Ovsjanikov, M., 2011. Shape google: Geometric words and expressions for invariant shape retrieval. ACM Transactions on Graphics (TOG) 30, 1.

Darom, T., Keller, Y., 2012. Scale-invariant features for 3-d mesh models. Image Processing, IEEE Transactions on 21, 2758–2769.

Eitz, M., Richter, R., Boubekeur, T., Hildebrand, K., Alexa, M., . Sketch-based shape retrieval. .

Feng, J., Wang, Y., Chang, S.F., 2016. 3d shape retrieval using a single depth image from low-cost sensors, in: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. pp. 1–9.

Gupta, S., Girshick, R., Arbeláez, P., Malik, J., 2014. Learning rich features from rgb-d images for object detection and segmentation, in: European Conference on Computer Vision. Springer, pp. 345–360.

Harwood, R.H., 2001. Visual problems and falls. Age and ageing 30, 13–18.

Hotelling, H., 1936. Relations between two sets of variates. Biometrika , 321–377.

Hu, K., Fang, Y., 2014. 3d laplacian pyramid signature, in: Computer Vision-ACCV 2014 Workshops, Springer. pp. 306–321.

Kazhdan, M., Chazelle, B., Dobkin, D., Finkelstein, A., Funkhouser, T., 2002. A reflective symmetry descriptor, in: European Conference on Computer Vision, Springer. pp. 642–656.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, pp. 1097–1105.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), IEEE. pp. 2169–2178.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86, 2278–2324.

Li, B., Lu, Y., Godil, A., Schreck, T., Bustos, B., Ferreira, A., Furuya, T., Fonseca, M.J., Johan, H., Matsuda, T., et al., 2014a. A comparison of methods for sketch-based 3d shape retrieval. Computer Vision and Image Understanding 119, 57–80.

Li, B., Lu, Y., Li, C., Godil, A., Schreck, T., Aono, M., Burtscher, M., Fu, H., Furuya, T., Johan, H., et al., 2014b. Extended large scale sketch-based 3d shape retrieval .

Li, B., Lu, Y., Li, C., Godil, A., Schreck, T., Aono, M., Burtscher, M., Fu, H., Furuya, T., Johan, H., et al., 2014c. Shrec' 14 track: Extended large scale sketch-based 3d shape retrieval, in: Eurographics Workshop on 3D Object Retrieval 2014 (3DOR 2014), pp. 121–130.

Li, W., Duan, L., Xu, D., Tsang, I.W., 2014d. Learning with augmented features for supervised and semi-supervised heterogeneous domain adaptation. Pattern Analysis and Machine Intelligence, IEEE Transactions on 36, 1134–1148.

Liu, F., Yang, L., 2015. A novel cell detection method using deep convolutional neural network and maximum-weight independent set, in: Medical Image Computing and Computer-Assisted InterventionMICCAI 2015. Springer, pp. 349–357.

Lord, S.R., Smith, S.T., Menant, J.C., 2010. Vision and falls in older people: risk factors and intervention strategies. Clinics in geriatric medicine 26, 569–581.

Lowe, D.G., 1999. Object recognition from local scale-invariant features, in: International Conference on Computer Vision, Ieee. pp. 1150–1157.

Malinowski, M., Rohrbach, M., Fritz, M., 2015. Ask your neurons: A neural-based approach to answering questions about images, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1–9.

McLean, G., Guthrie, B., Mercer, S.W., Smith, D.J., et al., 2014. Visual impairment is associated with physical and mental comorbidities in older adults: a cross-sectional study. BMC medicine 12, 181.

Osada, R., Funkhouser, T., Chazelle, B., Dobkin, D., 2002. Shape distributions. ACM Transactions on Graphics (TOG) 21, 807–832.

Patel, N., Pan, Y., Li, Y., Khorrami, F., Rizzo, J., Hudson, T., et al., 2016. Robust object detection and recognition for the visually impaired. 1st Intl Wksh On Deep Learning for Pattern Recognition (DLPR) .

Pfister, T., Charles, J., Zisserman, A., 2015. Flowing convnets for human pose estimation in videos, in: The IEEE International Conference on Computer Vision (ICCV).

Popescu, M.L., Boisjoly, H., Schmaltz, H., Kergoat, M.J., Rousseau, J., Moghadaszadeh, S., Djafari, F., Freeman, E.E., 2011. Age-related eye disease and mobility limitations in older adults. Investigative ophthalmology & visual science 52, 7168–7174.

Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G.R., Levy, R., Vasconcelos, N., 2010. A new approach to cross-modal multimedia retrieval, in: International Conference on Multimedia, ACM. pp. 251–260.

Rizzo, J.R., Hudson, T.E., Shoureshi, R.A., 2016. Smart wearable systems for enhanced monitoring and mobility. International Conferences on Modern Materials and Technologies (CIMTEC); 2016; Perugia, Italy .

Rizzo, J.R., Pan, Y., Hudson, T., Wong, E., Fang, Y., . Sensor fusion for ecologically valid obstacle identification: Building a comprehensive assistive technology platform for the visually impaired. 7th Intl Conf on Modeling, Simulation & Applied Optimization (ICMSAO); 2017; Sharjah, U.A.E .

Rosenblatt, F., 1958. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review 65, 386.

Sengupta, S., Nguyen, A.M., Van Landingham, S.W., Solomon, S.D., Do, D.V., Ferrucci, L., Friedman, D.S., Ramulu, P.Y., 2015. Evaluation of real-world mobility in age-related macular degeneration. BMC ophthalmology 15, 9.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y., 2013. Overfeat: Integrated recognition, localization and detection using convolutional networks. arXiv preprint arXiv:1312.6229 .

Shilane, P., Min, P., Kazhdan, M., Funkhouser, T., 2004. The princeton shape benchmark, in: Shape modeling applications, IEEE. pp. 167–178.

Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from rgbd images, in: European Conference on Computer Vision. Springer, pp. 746–760.

Van Landingham, S.W., Willis, J.R., Vitale, S., Ramulu, P.Y., 2012. Visual field loss and accelerometer-measured physical activity in the united states. Ophthalmology 119, 2486–2492.

Wang, Y., Feng, J., Wu, Z., Wang, J., Chang, S.F., 2014. From low-cost depth sensors to cad: Cross-domain 3d shape retrieval via regression tree fields, in: European Conference on Computer Vision. Springer, pp. 489–504.

Werbos, P., 1974. Beyond regression: New tools for prediction and analysis in the behavioral sciences .

Wold, S., Esbensen, K., Geladi, P., 1987. Principal component analysis. Chemometrics and intelligent laboratory systems 2, 37–52.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., Xiao, J., 2015. 3d shapenets: A deep representation for volumetric shapes, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1912–1920.

Yang, J., Yu, K., Gong, Y., Huang, T., 2009. Linear spatial pyramid matching using sparse coding for image classification, in: Computer Vision and Pattern Recognition, IEEE. pp. 1794–1801.

Yuhas, B., Ansari, N., 2012. Neural networks in telecommunications. Springer Science & Business Media.

Zhu, F., Shao, L., 2014. Weakly-supervised cross-domain dictionary learning for visual recognition. International Journal of Computer Vision 109, 42–59.

Zhu, J., Zhu, F., Wong, E.K., Fang, Y., 2015. Learning pairwise neural network encoder for depth image-based 3d model retrieval, in: Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, ACM. pp. 1227–1230.

Related material (NOT FOR PUBLICATION IN THIS PAPER)

Click here to download Related material (NOT FOR PUBLICATION IN THIS PAPER): p1227-zhu.pdf

LaTeX Source Files

Click here to download LaTeX Source Files: prletter_LDF.zip