Deep Correlated Holistic Metric Learning for Sketch-Based 3D Shape Retrieval

Guoxian Dai, Jin Xie, and Yi Fang

***Abstract*— How to effectively retrieve desired 3D models with simple queries is a long-standing problem in computer vision community. The model-based approach is quite straightforward but nontrivial, since people could not always have the desired 3D query model available by side. Recently, large amounts of wide-screen electronic devices are prevail in our daily lives, which makes the sketch-based 3D shape retrieval a promising candidate due to its simpleness and efficiency. The main challenge of sketch-based approach is the huge modality gap between sketch and 3D shape. In this paper, we proposed a novel deep correlated holistic metric learning (DCHML) method to mitigate the discrepancy between sketch and 3D shape domains. The proposed DCHML trains two distinct deep neural networks (one for each domain) jointly, which learns two deep nonlinear transformations to map features from both domains into a new feature space. The proposed loss, including discriminative loss and correlation loss, aims to increase the discrimination of features within each domain as well as the correlation between different domains. In the new feature space, the discriminative loss minimizes the intra-class distance of the deep transformed features and maximizes the inter-class distance of the deep transformed features to a large margin within each domain, while the correlation loss focused on mitigating the distribution discrepancy across different domains. Different from existing deep metric learning methods only with loss at the output layer, our proposed DCHML is trained with loss at both hidden layer and output layer to further improve the performance by encouraging features in the hidden layer also with desired properties. Our proposed method is evaluated on three benchmarks, including 3D Shape Retrieval Contest 2013, 2014, and 2016 benchmarks, and the experimental results demonstrate the superiority of our proposed method over the state-of-the-art methods.**

***Index Terms*— Sketch-based 3D shape retrieval, deep correlated holistic metric learning, discrepancy across different domains, mitigate.**

I. Introduction

With the advanced development of digitalization techniques, 3D models are widely available in our daily lives across many areas, such as 3D printing, medical imaging and entertainment. The vast amounts of 3D model lead to the pressing demand for effectively searching the desired 3D models. Traditional text-based search could not work well for two main reasons, 1) Only a small number of 3D models are available with text descriptions, which is too limited to retrieve desired 3D models. 2) It is often very hard to describe the very detailed information of complex 3D models simply with texts. Therefore, researchers proposed content-based 3D model retrieval framework, which mainly includes two categories, example-based 3D shape retrieval and sketch-based 3D shape retrieval. Most of the existing works fall into the first group, which is provided with a query 3D model and returns similar models [1]–[7]. Example-based 3D shape retrieval is quite straightforward, however, not convenient, since people usually don't always have the desired 3D model query available before hand. Recently, the sketch-based 3D shape retrieval has received more and more attentions from computer vision and computer graphics community [8]–[11]. Compared to the example-based framework, the sketches are much more convenient and easier to get, even a young kid could draw simple and comprehensive sketches. Apart from simpleness, sketch is also informative since it is very easy for people to understand the class labels for simple query sketches.

Despite all the advantages of sketch-based 3D shape retrieval, actually, it is a quite challenging problem. First, sketch and 3D shape come from two different modalities with huge gap. And features extracted from both modalities follow quite different distributions, which makes it very difficult to directly retrieve 3D shapes from sketch queries. Secondly, sketches are usually very simple with only several lines. The simpleness, on the contrary, also makes the sketch contain very limited information. The 3D shapes look visually similar as the query sketches only from some certain view angles. Generally, it is very hard to find the "best views" to project 3D shapes, which makes both sketches and 3D shapes similar.

The main challenge for sketch-based 3D shape retrieval is the domain discrepancies between these two modalities. In this work, we proposed a novel deep correlated holistic metric learning (DCHML) method to mitigate the discrepancies between sketch and 3D shape domains. Specifically, we first extract low-level features for both sketches and 3D shapes. For 2D sketch, we use pre-trained AlexNet [12] to extract features; for 3D sketch, we extracted histogram of oriented distance (HOD) in [13]; for 3D shape, we extract 3D-SIFT feature [14], which is extended from the well-known 2D SIFT [15]. The extracted 3D-SIFT is further encoded by locality-constrained linear coding (LLC) [16] to get a global shape descriptor. Then we learn two deep neural networks to transform the raw features of both domains into a new

Manuscript received May 29, 2017; revised December 10, 2017 and January 31, 2018; accepted March 3, 2018. Date of publication March 19, 2018; date of current version April 12, 2018. This work was supported by the ADEC Award for Research Excellence 2015, titled "Deep Cross-Domain Model for Conceptual Design." The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jianfei Cai. *(Corresponding author: Yi Fang.)*

The authors are with the NYU Multimedia and Visual Computing Lab and the Department of Electrical and Computer Engineering, New York University Abu Dhabi 129188, United Arab Emirates, and also with the Department of Electrical and Computer Engineering, New York University Tandon School of Engineering, Brooklyn, NY 11201 USA (e-mail: yfang@nyu.edu).

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

Digital Object Identifier 10.1109/TIP.2018.2817042

feature space, mitigating the domain discrepancy as well as maintaining the discrimination. The loss of the proposed network includes two parts, discriminative term which is constructed with pairwise distance within each domain and correlation term which is constructed with pairwise distance across different domains. The former one minimizes the variations of the deep learned features from the same class and maximizes the variations of the deep learned features from different classes within each domain; the latter one aims to alleviate the domain discrepancy, making the distributions of features from both domains as consistent as possible. Apart from adding the proposed loss at the output layer, similar loss is also imposed at the hidden layer to guide features in hidden layer also with desired properties. And it could further increase the robustness of deep learned features at the output layer. We verify our proposed method on three large scale benchmarks, 3D Shape Retrieval Contest (SHREC) 2013, 2014 and 2016 benchmarks, and the experimental results demonstrate the superiority of our proposed method.

This work is an extended version of conference paper [17]. There are three main differences against the conference version [17]: 1) We impose additional loss at the hidden layer to help the convergence of training and improve the retrieval performance. 2) We did more experiments to verify the proposed methods, including sketch-sketch within-domain retrieval and retrieval on SHREC 2016 dataset. 3) We also add one more subsection to discuss the effects of different parameter settings to the final retrieval performance.

The main contribution of our work is that we develop a novel deep correlated holistic metric learning method for sketch-shape cross-domain retrieval, which jointly trains two deep neural networks to learn two deep nonlinear transformations, one for each domain. Most of the existing deep metric learning methods focus on one domain applications. Different from those methods, we propose a correlation loss based deep metric learning to learn discriminative and consistent features across different domains. The proposed method can be viewed as an extension of classic deep metric learning methods. Here, the "holistic" means the proposed loss is not only imposed at the output layer but also at the hidden layer. By forcing features in the hidden layer are discriminative within each domain and consistent across different domains, the features at the output layer could become more discriminative and consistent. And the sketch-shape cross-domain retrieval performance could be further improved.

The rest of the paper is organized as follows. In Section II we introduce the related work; In Section III, we present our proposed deep correlated holistic metric learning for sketch-based 3D shape retrieval; In Section IV, we show the experimental results of our proposed method on three well-known large scale benchmarks; In Section V, we conclude our work.

II. Related Work

Most of the existing works about 3D shape retrieval is the example-based framework [1], [3], [5]–[7], [18]–[32], which could be roughly classified into three categories, projection based methods, diffusion based methods and deep learning based methods. For the projection based methods, 3D shapes are projected into 2D images, so that classic image features are adopted to construct shape descriptor, such as LFD [2] and ED [33]. For the diffusion based methods, 3D shape descriptors are derived based on heat diffusion or probability distribution of quantum particles, such as HKS [23], SIHKS [34] and WKS [35]. All the aforementioned methods are just handcrafted, inspired by the great success of deep learning in 2D images areas, deep learning is also introduced to 3D areas for shape retrieval [6], [7], [30], [31]. Xie *et al.* [7] use discriminative auto-encoder to extract robust shape descriptor in the hidden layers. Bai *et al.* [30] adopts convolutional neural network on the depth projections of 3D model to learn shape descriptor. Bai *et al.* [31] also proposed a two layer encoding framework for 3D shape matching. In addition, Shi *et al.* [36] adopted a cylinder projection and row-wise max-pooling to learn a robust 3D shape representation.

Except for the example-based framework, the sketch-based framework is another promising candidate for retrieving desired 3D shapes. Currently, there are very few works about sketch-based 3D shape retrieval. Zhu *et al.* [37] adopted a cross-domain neural network to mitigate the discrepancy between sketch and 3D shape. Funkhouser *et al.* [38] used spherical harmonics to compare similarities of different models, and designed a search engine supporting 3D models and 2D sketches as queries. Daras and Axenopoulos [39] proposed a unified 3D shape retrieval system supporting multimedia queries by projecting 3D models into a group of 2D images. The similarities among different models are determined by features extracted from 2D images. Bronstein *et al.* [26] applied bag-of-features (BoF) [40], which was widely used in 2D computer vision, for 3D shape retrieval. In addition, Eitz *et al.* [8] further adopted BoF with Gabor local line based feature (GALIF) for sketch-based 3D shape retrieval. Apart from BoF encoding scheme, locality-constrained linear coding (LLC) [16] is another encoding scheme widely applied in image classification, by maintaining the locality property. Biasotti *et al.* [41] applied LLC scheme for textured 3D shape retrieval. Tasse and Dodgson [42] proposed a novel cross-domain retrieval method by embedding different modality samples into the semantic vector representation of shape class. Xie *et al.* [43] proposed to learn a barycenter, which could effectively aggregate features from different 2D projections of 3D shapes. Apart from the aforementioned algorithms, large scale datasets have also been recently proposed to evaluate the performance of different methods, such as SHREC 2013 [11], [44] and SHREC 2014 [45], [46]. Sketches of both datasets come from a latest large sketch collection [47]. The 3D shapes of SHREC 2013 are mainly collected from Princeton Shape Benchmark [3], while the shapes of SHREC 2014 come from various sources, such as [3], [48]–[50]. Different comparison results are reported for both datasets. For SHREC 2013, the best reported result in [44] is from view clustering and shape context matching (SBR-VC). For SHREC 2014, the best reported result in [46] is from overlapped pyramid of HOG and similarity constrained manifold ranking, by Tatsuma *et al.*

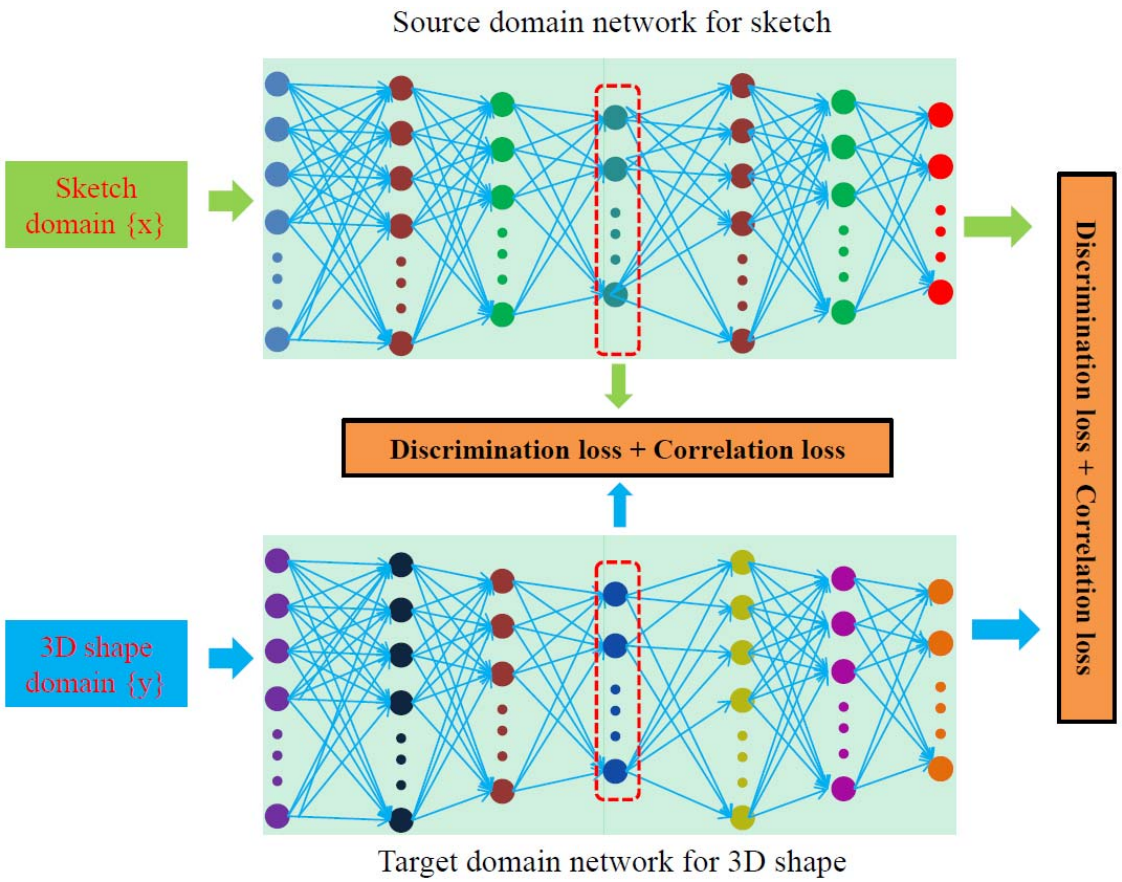

Fig. 1. The detailed framework of our proposed deep correlated holistic metric learning network. The whole network structure mainly includes two components, source domain network and target domain network. The proposed loss function is imposed at both output layer and hidden layer.

Recently, deep metric learning has received more and more attentions from computer vision community. Compared to traditional metric learning with a simple linear transformation [51]–[53], deep metric learning inherits advantages from the existing deep learning techniques [12], [54], [55] and could learn much more complex, powerful nonlinear transformation. Chopra *et al.* [56] adopts Siamese network to learn image similarities for face verifications. Generalizing the ideas in both [56] and large margin distance metric learning [57], Hu *et al.* [58] proposed a discriminative deep metric learning for face verification, with a marginal distance between positive pair and negative pair. Instead of randomly selecting training pairs in [58], Song *et al.* [59] considers all the possible positive pairs and negative pairs in the training set for deep metric learning. Different from deep metric learning with Siamese network in [56] and [58], which adopts pairwise training strategy with two input samples, Hoffer and Ailon [60] adopts triplet network for deep metric learning, which uses three identical networks with three input examples, one base example, its positive example and negative example. All the above deep metric leaning methods assume that training and testing examples follow the same distributions, which is actually too restricted for real applications. Thus, Hu *et al.* [61] proposed a deep transfer learning to deal with this scenario, by imposing maximum mean discrepancy criterion [62], [63] at the hidden layer of network. Lee *et al.* [64] and Xie and Tu [65] also imposed additional losses at the hidden layer of network to boost the performance of image classification and edge detection. Except for the application of deep metric learning in 2D image areas, it is also introduced to 3D shape areas [66]–[68]. Wang *et al.* [68] extended the Siamese network for sketch-based 3D shape retrieval by using two based Siamese networks, one for sketch domain and one for 3D shape domain. Their method is based on a strong assumption that all the 3D models are stored upright, which makes it much easier to choose the project view of 3D model. Such assumption can hardly be guaranteed in real application, and without such assumption, it is actually very hard to choose the "best" projection view. The projection results could change greatly, as the view changes.

It is noted that there are three works, which are most related to our proposed method, including DCML [17], Wang *et al.* [68] and Shape2Vec [42]. DCML [17] is a prepliminary version of the proposed method. Compared to DCML, we imposed additional loss at the hidden layers, which could not only make the training process converge faster, but also improve the performance. Wang *et al.* [68] used similar framework as our proposed method, by extending Siamese network [56], [69] for sketch-based 3D shape retrieval. There are several differences between [68] and our proposed method, 1) [68] needs to project 3D model into two different views with a strong assumption that all models are

stored upright as default. In fact, the projection results could change dramatically as the projection view changes. However, our proposed method doesn't need projection, neither does the upright assumption. 2) We put a marginal distance between examples from different classes, similar as [58], while [68] does not embed marginal distance for metric learning. 3) Wang *et al.* [68] only imposed loss at the output layer, while our proposed method adds loss on both the output and hidden layers to learn more robust features for retrieval. Similar to the proposed method, Shape2Vec [42] also mapped samples into a new feature space. The key difference is that Shape2Vec explicitly used semantic vector representation of shape class as target vector. However, the proposed method learn the target vector from data, which could generalize better from unseen class examples, compared to Shape2Vec.

III. Method

We proposed a novel deep correlated holistic metric learning method for sketch-based 3D shape retrieval. Fig. 1 shows the detailed framework of our proposed method. The proposed networks consist of two components, one for sketch domain, referred as source domain network (SDN), and one for 3D shape domain, referred as target domain network (TDN). The proposed method trains both deep neural networks simultaneously with proposed loss at both output layer and hidden layer. The loss function includes two terms, discrimination term and correlation term, which minimizes intra-class variations and maximizes inter-class variations within each domain, meanwhile guarantees the distribution-consistency across different domains.

The proposed method mainly includes two steps: 1) extracting low-level features for both sketches and 3D shapes. 2) Learning two deep nonlinear transforms to map features of both domains from the original space into a nonlinear feature space, increasing the discrimination of features within each domain as well as mitigating the discrepancy across different domains. The details for each step are introduced as follows.

A. Feature Extraction

Features for both sketches and shapes are extracted separately.

1) Sketch: Our proposed method is verified with both 2D sketch query and 3D sketch query. For 2D sketch, inspired by the outstanding performance of convolutional neural network (CNN) in feature learning [12], [55], [70], we fine-tune AlexNet [12] on sketch dataset. The AlexNet mainly includes 5 convolutional layers and 3 fully connected layers. After fine-tuning, we extract the visual features in "*fc7*" layer. The feature dimension is 4096. For 3D sketch, we extracted histogram of oriented distance (HOD) in [62].

2) 3D Shape: Inspired by Lowe's SIFT [15] in 2D images, [14] extended it into 3D mesh and proposed 3D-SIFT by detecting interest points. We first extract 3D-SIFT for 3D shapes, which are further encoded with the LLC [16] scheme to get a global shape descriptor. Readers could refer to [14] and [16] for more detailed information about 3D-SIFT and LLC.

B. Deep Correlated Holistic Metric Learning

We denote training examples from source domain (sketch domain) and target domain (3D shape domain) as $S = \{x_1, x_2, x_3, \ldots\}$ and $T = \{y_1, y_2, y_3, \ldots\}$, respectively. The transfer functions for SDN and TDN are denoted as $f^s : x \to f^s(x)$ and $f^t : y \to f^t(y)$, respectively. In addition, W_k^s, W_k^t and b_k^s, b_k^t are the weights and bias, connecting layer k and layer $k+1$ of SDN and TDN, respectively; the activations of the i-th example x_i from S and j-th example y_j from T at the k-th layer of SDN and TDN are denoted as $a_k^{i,s}$ and $a_k^{j,t}$, respectively,

$$\begin{aligned} a_{k+1}^{i,s} &= \sigma(W_k^s a_k^{i,s} + b_k^s) = \sigma(r_{k+1}^{i,s}) \\ a_{k+1}^{j,t} &= \sigma(W_k^t a_k^{j,t} + b_k^t) = \sigma(r_{k+1}^{j,t}) \end{aligned} \tag{1}$$

where $\sigma(x)$ is the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$, $r_{k+1}^{i,s} = W_k^s a_k^{i,s} + b_k^s$, $r_{k+1}^{j,t} = W_k^t a_k^{j,t} + b_k^t$. K_s and K_t denotes the total number of layers for SDN and TDN, respectively. Thus, the nonlinear transfer function $f^s(x_i)$ and $f^t(y_j)$ across K_s and K_t layers of SDN and TDN respectively, can be represented as follows,

$$f^s(x_i) = a_{K_s}^{i,s} f^t(y_j) = a_{K_t}^{j,t}. \tag{2}$$

The features sampled from different domains, sketch and 3D shape, suffer the domain discrepancy. Such discrepancy makes it very difficult to directly conduct across-domain retrieval. To effectively perform cross-domain retrieval, the features from both domains should address the following two issues: 1) within each domain, the features should be as discriminative as possible, 2) across different domains, the distributions of features from both domains should be as consistent as possible. To this end, we proposed a novel deep correlated holistic metric learning method to mitigate the discrepancy across different domains as well as increase the discrimination within each domain. The proposed method learns two distinct deep neural networks (different weights and different structures) simultaneously to transform features from both domains into a nonlinear feature space. The proposed loss includes two terms, discriminative loss and correlation loss. For each loss term, it is not only imposed on the output layer but also the hidden layer to guarantee features in both output layer and hidden with desired properties. And it could further improve the performance, compared to only imposing loss at output layer. The proposed loss function L is formulated as follows,

$$\begin{aligned} L &= \alpha L^d + (1-\alpha) L^c + \lambda \Big(\sum_{i=1}^{K_s} \|W_i^s\|_F^2 + \sum_{j=1}^{K_t} \|W_j^t\|_F^2 \Big) \\ &= \alpha \big(\beta L_h^d + (1-\beta) L_o^d \big) + (1-\alpha) \big(\beta L_h^c + (1-\beta) L_o^c \big) \\ &\quad + \lambda \Big(\sum_{i=1}^{K_s} \|W_i^s\|_F^2 + \sum_{j=1}^{K_t} \|W_j^t\|_F^2 \Big) \end{aligned} \tag{3}$$

where L^d denotes the discriminative loss, including losses at both the output layer L_o^d and hidden layer L_h^d. And L^c denotes the correlation loss, including losses at both the output layer L_o^c and hidden layer L_h^c. L_o^d and L_h^d aim to minimize intra-class distance of deep transformed features and maximize inter-class

distance of deep transformed features to a large margin h within each domain for both output layer and hidden layer respectively. And L_o^c and L_h^c optimize the pairwise across-domain distance to mitigate the distribution inconsistency across different domains for both output layer and hidden layer respectively. α is the weight to balance between discrimination term and correlation term. β is the weight to balance between hidden layer loss and output layer loss. λ is the weight for regularization term, avoiding over-fitting.

1) Discrimination Term: The discrimination term aims to minimize intra-class distance and maximize inter-class distance within each domain. To guarantee the deep learned feature as discriminative as possible, we not only impose discriminative loss at the output layer but also the hidden layer of the network, including L_o^d and L_h^d, respectively. For each term, the discriminative loss includes two parts for both source domain and target domain respectively.

$$\begin{aligned} L_o^d &= L_{o,s}^d + L_{o,t}^d \\ L_h^d &= L_{h,s}^d + L_{h,t}^d. \end{aligned} \tag{4}$$

where $L_{o,s}^d$ and $L_{o,t}^d$ denote the discriminative loss at the output layer for both source domain and target domain, respectively. While $L_{h,s}^d$ and $L_{h,t}^d$ denote the discriminative loss at the hidden layer for both source domain and target domain, respectively.

For the positive pair examples, their distances are minimized; for negative pair examples, their distances are maximized to a large margin. Thus, $L_{o,s}^d$ and $L_{o,t}^d$ are formulated as follows,

$$\begin{aligned} L_{o,s}^d = &\sum_{x_i,x_j \in P^s} \|a_{K_s}^{i,s} - a_{K_s}^{j,s}\|_2^2 \\ &+ \sum_{x_i,x_j \in N^s} \max\{0, \beta_o - \|a_{K_s}^{i,s} - a_{K_s}^{j,s}\|_2^2\} \end{aligned} \tag{5}$$

$$\begin{aligned} L_{o,t}^d = &\sum_{y_i,y_j \in P^t} \|a_{K_t}^{i,t} - a_{K_t}^{j,t}\|_2^2 \\ &+ \sum_{y_i,y_j \in N^t} \max\{0, \beta_o - \|a_{K_t}^{i,t} - a_{K_t}^{j,t}\|_2^2\} \end{aligned} \tag{6}$$

where P^s and N^s denote the sets of positive pair and negative pair in source domain S, respectively. While P^t and N^t denote the sets of positive pair and negative pair in target domain T. β_o is the marginal distance for negative pair at the output layer.

Assume the discriminative loss is imposed at the H_s-th and H_t-th hidden layer of both source domain and target domain networks. Thus, $L_{h,s}^d$ and $L_{h,t}^d$ could be formulated similarly.

The overall discriminative loss L^d is imposed at both the output layer and hidden layer of the network, which aims to minimizes intra-class distance and maximizes inter-class distance to a large margin within both source and target domains. Through imposing discriminative loss at the hidden layer, the features in the hidden layers are encouraged to be discriminative, which could help improve the discriminative property of features at the output layer.

2) Correlation Term: Features from both domains follow different distributions, which makes it hard to directly retrieve desired objects across different modalities. Thus, a correlation term is further imposed to maintain the distribution consistency across different domains. The correlation term is the key part to build the connections between sketch domain and 3D shape domains, which is constructed with pairwise distance across different domains. Specifically, the correlation term includes two types of pairwise across-domain distances. Similar as the discriminative loss, the correlation loss is also imposed at both the output layer and hidden layer of the networks, L_o^c and L_h^c,

$$\begin{aligned} L_o^c &= L_{o,1}^c + L_{o,2}^c \\ L_h^c &= L_{h,1}^c + L_{h,2}^c \end{aligned} \tag{7}$$

where $L_{o,1}^c$ and $L_{o,2}^c$ denote the pairwise across-domain distance at the output layer, which could be formulated as follows,

$$\begin{aligned} L_{o,1}^c = &\sum_{x_i,y_j \in P^c} \|a_{K_s}^{i,s} - a_{K_t}^{j,t}\|_2^2 \\ &+ \sum_{x_i,y_j \in N^c} \max\{0, \beta_o - \|a_{K_s}^{i,s} - a_{K_t}^{j,t}\|_2^2\} \\ L_{o,2}^c = &\sum_{c^s,c^t} \sum_{\substack{\forall x_i,x_j \in c^s \\ \forall y_i,y_j \in c^t}} R(x_i,x_j,y_i,y_j) \\ &- \sum_{c^s,d^t}^{c \neq d} \sum_{\substack{\forall x_i,x_j \in c^s \\ \forall y_i,y_j \in d^t}} R(x_i,x_j,y_i,y_j) \end{aligned} \tag{8}$$

where P^c and N^c denote the sets of positive pairs and negative pairs across different domains. L_1^c directly minimizes the distances of positive pair across-domain examples, and maximizes the distances of negative pair across-domain examples to a large margin h, making the distributions of two domains as similar as possible. c^s and c^t denote the set of examples with class label c for source domain and target domain respectively. Except for $L_{o,1}^c$, $L_{o,2}^c$ is further imposed to guarantee the distribution-consistency across different domains. R are formulated as follows,

$$R(x_i,x_j,y_i,y_j) = \left(\sqrt{\|a_{K_s}^{i,s} - a_{K_s}^{j,s}\|_2^2} - \sqrt{\|a_{K_t}^{i,t} - a_{K_t}^{j,t}\|_2^2}\right)^2 \tag{9}$$

In $L_{o,2}^c$, x_i and x_j are from the same class, so do y_i and y_j. $R(x_i,x_j,y_i,y_j)$ is constructed as the difference between within-class distances across two domains. Although in $L_{o,1}^c$ the outputs of two networks with the same label, i.e., $a_{K_s}^{i,s}$ and $a_{K_t}^{j,t}$, are enforced to be near, it cannot guarantee that the local neighborhood structures of $a_{K_s}^{i,s}$ and $a_{K_t}^{j,t}$ are also similar in the high-dimensional feature space. Therefore, we define $R(x_i,x_j,y_i,y_j)$ to describe the similarity between the local neighborhood structures of $a_{K_s}^{i,s}$ and $a_{K_t}^{j,t}$ so that the distributions of the samples with the same label are consistent across two domains. If (x_i,x_j,y_i,y_j) are with the same label, R is minimized to encourage the local neighborhood structures of $a_{K_s}^{i,s}$ and $a_{K_t}^{j,t}$ to be similar, increasing the association between two domains. Otherwise, R is maximized to encourage the local neighborhood structures of $a_{K_s}^{i,s}$ and $a_{K_t}^{j,t}$ to

be dissimilar, decreasing the association between two domains. In addition, the correlation losses at hidden layers, $L^c_{h,1}$ and $L^c_{h,2}$ could be formulated similarly.

The main idea of imposing additional loss at hidden layer is that the performance of features at output layer could be further improved by encouraging features in the hidden layers also with desired property. By adding losses at hidden layers, the training of the network could be guided from early layer, apart from the output layer. Hence, we could provide a much stronger supervision for training. The loss error is not only back-propagated from the output layer, but also from the hidden layer, which could avoid the vanishing gradient problem and help the network converge stably. Each term in the proposed loss function Eq. 3, is differentiable, thus our proposed method could be optimized through back-propagation with stochastic gradient descent. The gradients come from two error paths, one is from the output layer and one is from the H-th hidden layer. For the layers before H-th layer, the gradients are summation from both paths, while for the layers after H-th layer, the gradients are only from the output layer.

IV. Experimental Results

Our proposed method is evaluated on three well-known benchmarks, SHREC 2013 [3], SHREC 2014 [45] and SHREC 2016 [13]. To verify the effectiveness of our proposed method, we first compare the experimental results between only imposing loss at output layer and imposing loss at both output layer and hidden layer. In addition, we not only conduct sketch-shape, cross-domain retrieval, but also sketch-sketch, within-domain retrieval to comprehensively evaluate the performance of our proposed method. Besides, we also compared our proposed method with the state-of-the-art methods using several common metrics, including nearest neighbor (NN), first tier (FT), second tier (ST), discounted cumulative gain (DCG) and mean average precision (mAP). Precision-recall curve is also provided to visualize the performance of our proposed method. Overall, the experimental results demonstrate that our proposed method could outperform the state-of-the-art methods.

A. Implementation Details

In this subsection, we mainly introduce the implementation details for our proposed method. For feature extraction, the 2D sketch feature was extracted from the "*fc7*" layer of AlexNet [12] with the feature size of 4096; the 3D sketch feature was extracted from histogram of oriented distance (HOD) in [62], with the size of $320 = 64 \times 5$; the feature size of 3D-SIFT for 3D shape is set to 128, in addition, the size of the codebook for LLC is set to 4096, which is generated by regular k-means. The network structures for the sketch and 3D shape domains are set to [2000 1000 100] and [2000 1000 500 100], respectively. The network structure doesn't include input layer size, which is determined by the input feature size, for 2D sketch, it is set to 4096; for 3D sketch, it is set to 320; for 3D shape is set to 4096. The loss

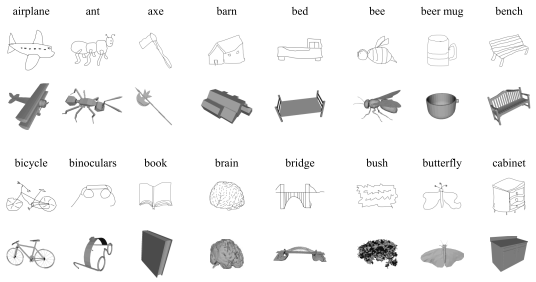

Fig. 2. Example of sketches and shapes from SHREC 2013 dataset.

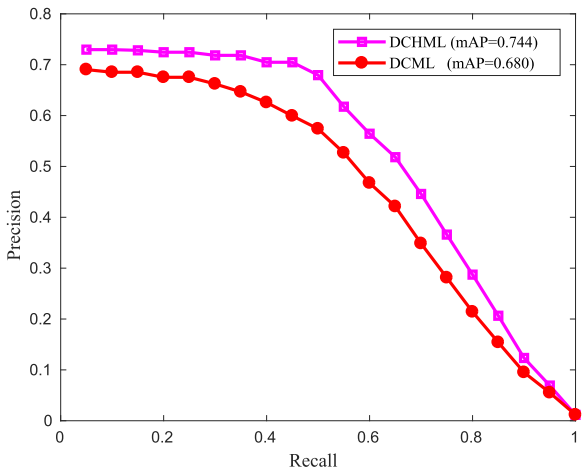

Fig. 3. Precision-recall curves for imposing loss only at output layer and imposing loss at both output layer and hidden layer on SHREC 2013.

is imposed at both output layer and hidden layer, specifically, the 3th hidden layers of SDN and TDN. In addition, the momentum rate is set to 0.1, learning rate is set to 0.015. The marginal distances for the output layer and hidden layer are set to 3 and 50, respectively.

B. Retrieval on SHREC 2013 Dataset

In this section, we evaluate our proposed method on SHREC 2013 benchmark. SHREC 2013 [11], [44] is large scale benchmark to evaluate algorithms for sketch-based 3D shape retrieval. The benchmark is created by collecting common classes from both the Princeton Shape Benchmark [3] and sketch dataset [47]. Fig. 2 shows some examples of sketches and shapes from SHREC 2013 dataset. There are 1258 shapes and 7200 sketches in SHREC 2013, which are grouped into 90 classes in total. The number of shapes in each class is not equal, about 14 in average. While the number of sketch for each class is equal, 80 in total, 50 for training and 30 for testing.

To demonstrate the effectiveness of our proposed method, we first compare the experimental results between imposing loss only at output layer, denoted as DCML, and imposing loss at both output layer and hidden layer, denoted as DCHML. Fig. 3 shows the precision-recall curves of both DCML and DCHML. As we can see from Fig. 3, by imposing additional loss at the hidden layer of the network, the performance could be furhter improved, compared to only imposing loss at the

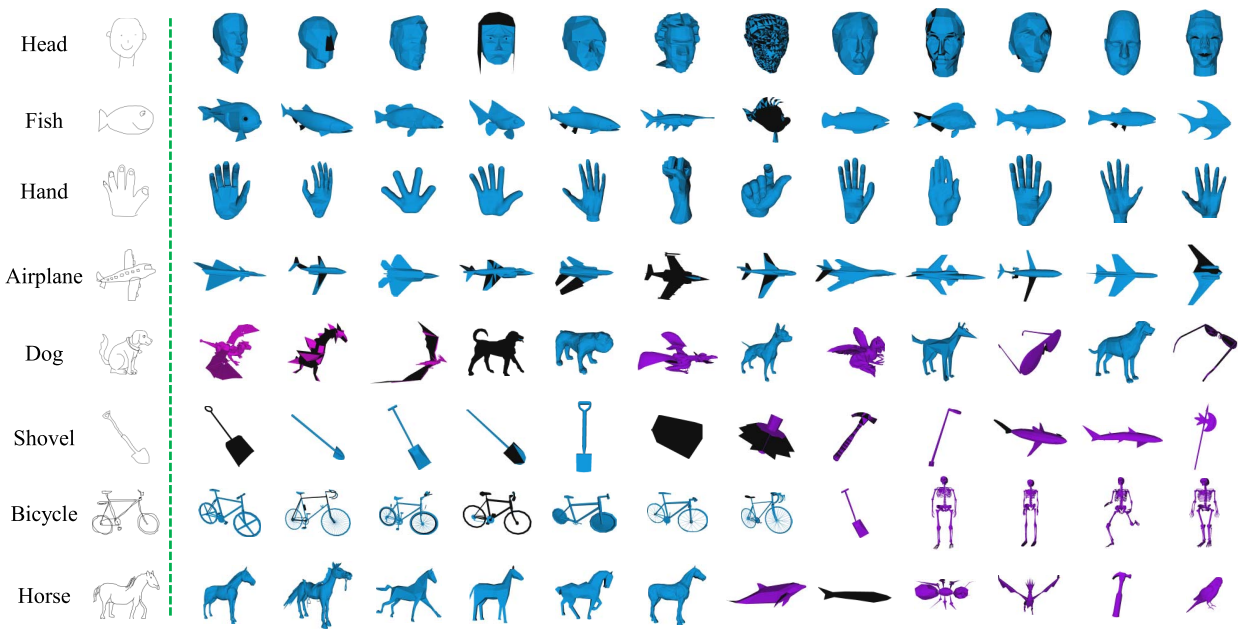

Fig. 4. Illustration of retrieved examples on SHREC 2013 dataset. The query sketch is listed on the left first column, and the top 12 retrieved 3D models are listed on the right side, according to their ranking orders. The correct retrieved examples are marked with blue color, while the incorrect retrieved examples are marked with purple color.

output layer. The mAP is increased from 0.680 to 0.744, with the gain of 0.064.

Fig. 4 shows some retrieved examples on SHREC 2013 dataset. The query sketches are listed on the left first column, namely, head, fish, hand, airplane, dog, shovel, bicycle and horse. The top 12 retrieved models are listed on the right side, based on their ranking order. The correct retrieved models are marked with blue color, while the wrong results are marked with purple color. As we can see from Fig. 4, for the classes of head, fish, hand and airplane, all the retrieved 12 models are correctly relevant; for the classes of shovel, bicycle and horse, the proposed method first retrieved correct examples, and then wrong examples, because there are too few examples in these three classes, less than 12. For the class, dog, the proposed method mistakenly retrieved several wrong examples, due to the geometrical similarity among these 3D shapes.

PCA is adopted to reduce the dimension of the deep learned features from 100 to 2 for visualization, as shown Fig. 5. All the features are grouped in different color by their class labels. As we can see in Fig. 5, features with the same label are grouped together, while features with different labels are separated away. This is just a coarse visualization of our proposed method, which could roughly verify the effectiveness of our proposed method.

Precision-recall curve is a common metric to visually compare the retrieval performances of different algorithms. Fig. 6 shows the precision-recall curves on SHREC 2013 dataset of our proposed method as well as state-of-the-art methods reported in [11]. The magenta curve indicates the performance of our proposed method. As we can see in Fig. 6, our proposed methods could significantly outperform state-of-the-art methods. As the recall value increases, the precision value of our proposed method stays at least double times higher than other methods, in addition, decrease slower when recall is small. Except precision-recall curve, we also compare the performance of our proposed methods with state-of-the-art methods based on other standard metrics. Table. I shows the comparison results on SHREC 2013 dataset. As we can see in Table. I, the DCML without additional loss at hidden layer could already outperform state-of-the-art methods. The proposed DCHML could achieve even better performance, about more than 30% gain in average compared the best reported state-of-the art method [68]. The experimental results demonstrate the effectiveness and superiority of our proposed method.

To comprehensively verify the performance of the deep learned features, we further conduct sketch-sketch (SS)

TABLE I

PERFORMANCE COMPARISON WITH THE STATE-OF-THE-ART METHODS ON SHREC 2013 DATASET

	NN	FT	ST	E	DCG	mAP
[11]	0.164	0.097	0.149	0.085	0.348	0.116
[10]	0.279	0.203	0.296	0.166	0.458	0.250
[71]	0.110	0.069	0.107	0.061	0.307	0.086
[72]	0.017	0.016	0.031	0.018	0.240	0.026
[68]	0.405	0.403	0.548	0.287	0.607	0.469
DCML	0.690	0.642	0.718	0.351	0.773	0.680
DCHML	**0.730**	**0.715**	**0.773**	**0.368**	**0.816**	**0.744**

TABLE II

PERFORMANCE COMPARISON OF SKETCH-SKETCH RETRIEVAL ON SHREC 2013 DATASET

	NN	FT	ST	E	DCG	mAP
[68]	0.431	0.352	0.514	0.298	0.679	0.373
DCHML	0.759	0.755	0.829	0.591	0.873	0.787

Fig. 5. Visualization of the deep learned sketch features and shape features on SHREC 2013 dataset. The features are grouped in different colors by class label.

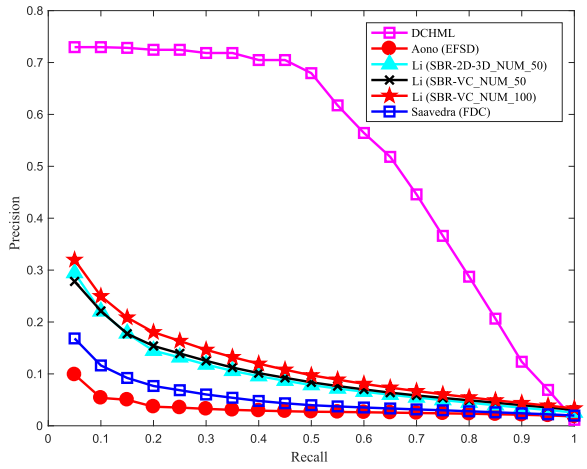

Fig. 6. Performance comparison of precision-recall curve on SHREC 2013 dataset.

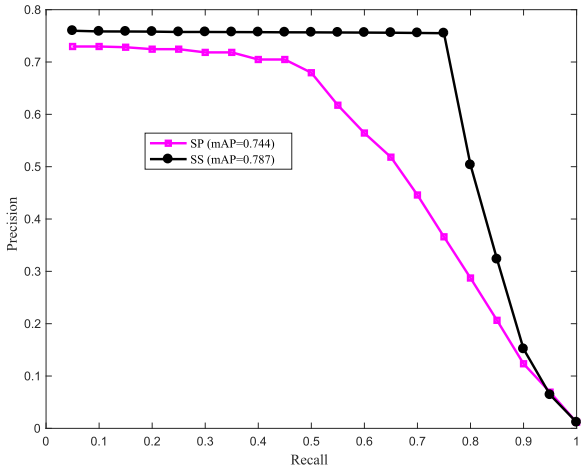

Fig. 7. Precision-recall curves of sketch-shape and sketch-sketch retrieval on SHREC 2013 dataset.

retrieval within sketch domain. Table.II shows the performance comparison between our proposed method with [68] based on common evaluation metrics. Our proposed method could significantly outperform [68] in all criterion. In addition, we also compare the retrieval performance between sketch-shape (SP) and sketch-sketch retrieval. Fig. 7 shows the precision-recall curve for both sketch-shape and sketch-sketch retrievals on SHREC 2013 dataset. Both tasks are using the same query sketches, however, aiming for different target outputs from two modalities. Intuitively, the latter one should have better performance, since the sketch-sketch retrieval is within-domain task, which is easier, compared to sketch-shape across-domain task. The experimental results meet expectations, as shown in Fig. 7. The precision value of sketch-sketch retrieval is steadily higher than sketch-shape retrieval as the recall value increases. Besides, the mAP for sketch-sketch retrieval is 0.787, while the mAP for sketch-shape is only 0.744.

C. Retrieval on SHREC 2014 Dataset

In this subsection, we test our proposed method on SHREC 2014 dataset [45]. SHREC 2014 is a large scale benchmark for sketch-based 3D shape retrieval, which consists of shapes from various datasets, such as SHREC 2012 [48], the Toyohashi Shape Benchmark (TSB) [49], the McGill 3D Shape Benchmark (MSB) [73], *etc.* The dataset has about 13680 sketches and 8987 3D models in total, grouped into 171 classes.

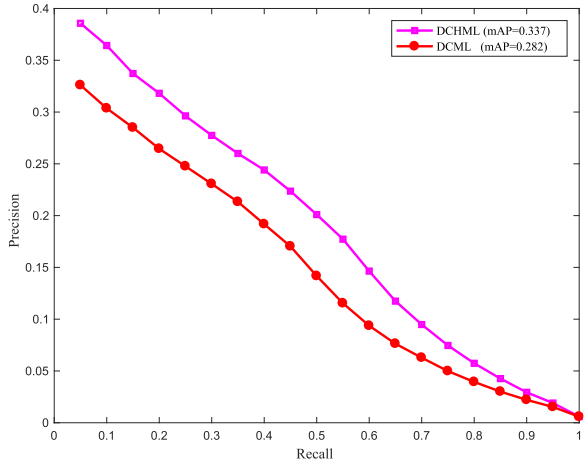

Fig. 8. Performance comparison between imposing loss only output layer and imposing loss at both output layer and hidden layer on SHREC 2014.

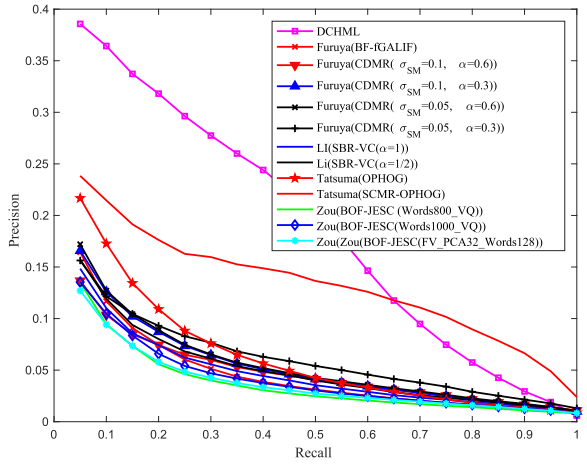

Fig. 9. Performance comparison of precision-recall curve on SHREC 2014 dataset.

SHREC 2014 dataset is quite challenge due to its diversity of categories and large variations within class. The number of shapes in each class varies from less than 10 to more than 300, while the number sketches for each class is equal to 80, 50 for training and 30 for testing.

We first compare the experimental results between the loss only at the output layer and the loss at both hidden layer and output layer to demonstrate the effectiveness of imposing additional loss at hidden layers. Fig. 8 shows the precision-recall curves of both methods. Through imposing additional loss at hidden layer, The performance is significantly improved. The precision value of DCHML is steadily higher than that of DCML. In addition, the mAP is increased from 0.282 to 0.337, with the gain of 0.055.

Fig. 9 shows the precision-recall curves of our proposed method with other state-of-the-art methods on SHREC 2014 dataset. The magenta curve denotes our proposed method. As we can see in Fig. 9, when recall is about less than 0.65, the precision of our proposed method is higher than other methods; while when recall is about larger than about 0.65, the precision of our proposed method drops below other method. In fact, people are generally more interested in the top retrieved objects, instead of latter objects. Based on such assumption, the precision-recall curve indicates the advantages of our proposed method.

TABLE III
PERFORMANCE COMPARISON OF DIFFERENT METHODS ON SHREC'14 DATASET

	NN	FT	ST	E	DCG	mAP
[10]	0.109	0.057	0.089	0.041	0.328	0.054
[11]	0.095	0.050	0.081	0.037	0.319	0.050
[49]	0.160	0.115	0.170	0.079	0.376	0.131
[68]	0.239	0.212	0.316	0.140	0.495	0.228
DCML	0.351	0.276	0.335	0.174	0.500	0.282
DCHML	**0.403**	**0.329**	**0.394**	**0.201**	**0.544**	**0.336**

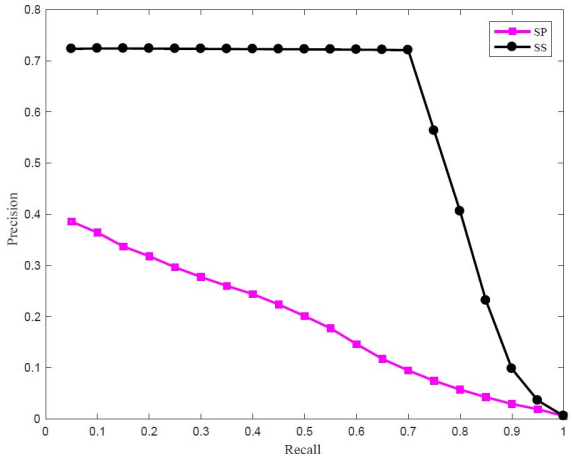

Fig. 10. Precision-recall curve for sketch-shape and sketch-sketch retrieval on SHREC 2014 dataset.

We also compare our proposed method with state-of-the-art methods, based on other common evaluation metrics, such as NN, FT, ST, DCG and mAP. Table. III shows the comparison results of our proposed method and other methods on SHREC 2014 dataset. The DCML with proposed loss only at output layer could already outperform other methods in all metrics. The proposed DCHML could achieve even better performance. The experimental results demonstrate the effectiveness of our proposed method for sketch-shape across-domain retrieval.

In addition, we also conduct sketch-sketch retrieval on SHREC 2014 to further verify the performance of our proposed methods. Fig. 10 shows the precision-recall curve of sketch-sketch (SS) retrieval and sketch-shape (SP) retrieval on SHREC 2014 dataset. The precision-recall curve of SS retrieval is significantly higher than that of SP retrieval, which is reasonable, since sketch-sketch retrieval is a within-domain task, while sketch-shape retrieval is a cross-domain task. And the former one is easier compared to the latter one. Besides, we also compare the sketch-sketch and sketch-shape retrieval performance based on standard criterion, as shown in Table. IV. The performance of SS retrieval could outperform SP retrieval in all criterion.

TABLE IV

PERFORMANCE COMPARISON BETWEEN SKETCH-SHAPE AND SKETCH-SKETCH RETRIEVAL SHREC 2014 DATASET

	NN	FT	ST	E	DCG	mAP
SP	0.403	0.329	0.394	0.201	0.544	0.336
SS	0.725	0.720	0.802	0.564	0.851	0.753

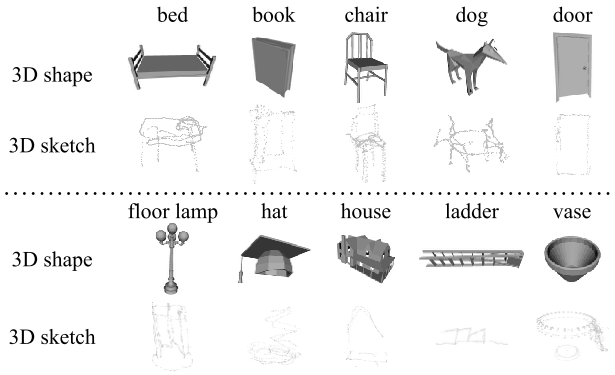

Fig. 11. Examples of 3D sketches and shapes from SHREC 2016.

TABLE V

PERFORMANCE COMPARISON ON SHREC 2016 DATASET

	NN	FT	ST	E	DCG	mAP
CNN-Siamese [13]	0.000	0.031	0.108	0.048	0.293	0.072
DCHML	**0.117**	**0.106**	**0.148**	**0.086**	**0.327**	**0.147**

D. Retrieval on SHREC 2016 Dataset

Different from SHREC 2013 and 2014 with 2D sketches, SHREC 2016 is a new benchmark, which evaluates the performance of different algorithms by using 3D sketch queries to retrieve 3D shapes [13]. The 3D sketches of SHREC 2016 come from [74] and [75]. There are 300 3D sketches in total, which are divided into 30 groups, For each group, 7 models are used for training, and 3 for testing. The 3D models of SHREC 2016 come from SHREC13STB [44]. There are 1258 models total, classified into 90 groups. The number of shapes for each group is not equal, around 13 models in average. Among the 30 classes of sketches, only 21 classes have relevant 3D models. Fig. 11 shows some examples of 3D sketch and shapes from SHREC 2016. As we can see in Fig. 11, the 3D sketches are just sparse point cloud. For 3D sketch, we extract raw feature, histogram of oriented distances (HOD) [13]. Specifically, we choose 64 bins for 5 different angles. Table. V shows the performance comparison between our proposed method and state-of-the-art methods. As we can see from Table. V, our proposed DCHML could outperform CNN-Siamese [13].

E. Parameter Discussion

In this section, we mainly discuss the effects of different parameter settings to the experimental results, including α and β. Specifically, We conduct experiments on SHREC 2013 to study the effects.

1) Parameter α: The proposed loss function mainly includes two terms, discrimination term and correlation term.

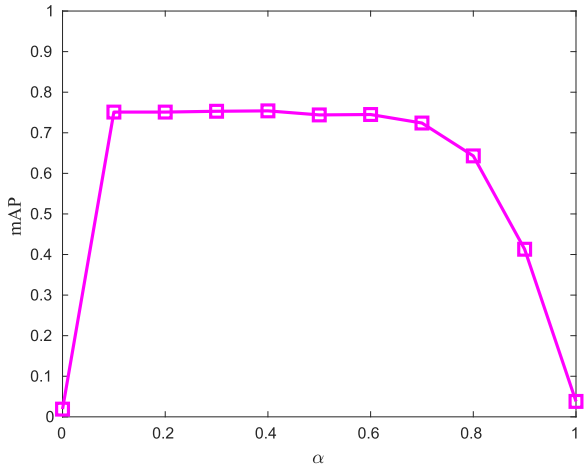

Fig. 12. The mAP of the proposed method vs α.

And both losses are imposed at both the output layer and hidden layer. The discrimination loss is built with pairwise distance within each domain. And it is mainly used to make features discriminative within each domain, which could be trained independently for both domains. However, the correlation loss is built with pairwise distance across different domains, which is the key part to make the distribution consistent across different domains. α is the weight to balance between two terms. α denotes the weight of discrimination loss, while $1-\alpha$ denotes the weight of correlation loss. Fig. 12 shows the mAP vs. α. There are three main conclusions from Fig. 12, 1) both terms are very essential in the proposed methods, since the performance is very bad for $\alpha = 0$ with only the correlation term and $\alpha = 1$ with only the discrimination term. 2) The proposed method is very robust to α, when α changes from 0.1 to 0.7. 3) The correlation term plays a relatively more important role, compared to discrimination term. For discrimination term, the weight of 0.1, is enough to maintain the performance, however for for correlation term, the weight is at least 0.3 to maintain the performance. The experimental results suggests that we need to assign larger weight to correlation term, compared to discrimination term.

2) Parameter β: Most of existing deep metric learning methods only impose contrastive loss at the output layer of the deep neural network. We impose the additional losses at the hidden layer of the neural network, which could make training converge faster and further increase the performance. We did additional experiments to verify the effects of assigning different weights to hidden layer loss and output layer loss. β denotes the weight for hidden layer loss, while $1-\beta$ denotes the weight for output layer loss. We change the balance weight β from 0 to 1. The performance is shown in Fig. 13. When β changes from 0 to 0.1, the mAP is creased from 0.67 to 0.74, which demonstrates that the additional hidden layer loss could help improve the retrieval performance. When β changes from 0.1 to 0.8, the mAP remains stable, which indicates the robustness of our proposed method to β. Finally, when β changes from 0.8 to 1, the performance begins to drop, particularly when $\beta \geq 0.9$, it decreases drastically. The experimental results indicate that the weight for the output

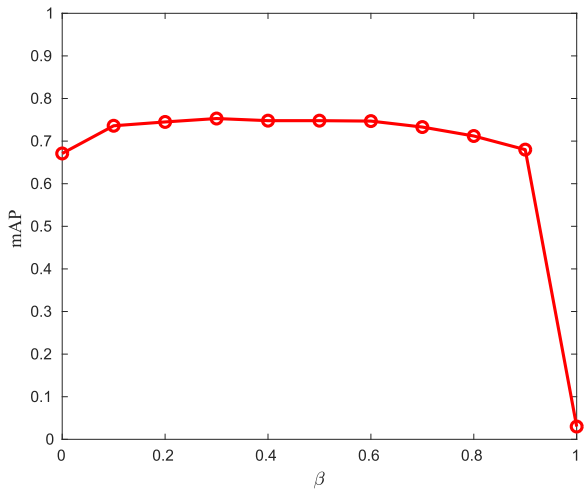

Fig. 13. The mAP of the proposed method vs. β.

layer should be larger, since it plays to key role for retrieval. When the weight for the output layer loss is too small, the networks almost learn nothing for the last few layers, thus the performance is very bad for using the output features.

V. Conclusions

In this work, we developed a novel deep correlated holistic metric learning for sketch-based 3D shape retrieval. Specifically, we first extract raw features for both sketches and 3D shapes separately. Then features from both domains are mapped into a new feature space through our proposed deep correlated holistic metric learning network. The overall loss function of the proposed networks mainly includes two terms, discrimination term and correlation term. The former one aims to minimize intra-class distance and maximize inter-class distance to a large margin within each domain. While, the latter one aims to maintain the distribution-consistency of features across different domains. In addition, the proposed loss is not only imposed at the output layer but also the hidden layer. And the performance of features at the output layer could be further improved by encouraging features in the hidden layer also with desired properties. Our proposed method is evaluated on SHREC 2013, 2014 and 2016, and the experimental results demonstrate superiority over the state-of-the-art methods.

Currently, the proposed method only focuses on two modalities, sketch and 3D shape. One possible future direction would be extending it to multi-modalities, such as sketch, 3D shape and depth image. The extension could be done by simply constructing pairwise distances within each domain, pairwise distance across all the different domains. Since generative adversarial networks (GANs) can maintain distributions between two domains consistent, another interesting direction would be combining GANs with the proposed method.

References

[1] J. W. Tangelder and R. C. Veltkamp, "A survey of content based 3D shape retrieval methods," *Multimedia Tools Appl.*, vol. 39, no. 3, pp. 441–471, Sep. 2008.

[2] D.-Y. Chen, X.-P. Tian, Y.-T. Shen, and M. Ouhyoung, "On visual similarity based 3D model retrieval," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 223–232, Sep. 2003.

[3] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, "The princeton shape benchmark," in *Proc. Shape Model. Appl.*, Jun. 2004, pp. 167–178.

[4] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3D shape descriptors," in *Proc. Symp. Geometry Process.*, vol. 6. 2003, pp. 156–164.

[5] N. Iyer, S. Jayanti, K. Lou, Y. Kalyanaraman, and K. Ramani, "Three-dimensional shape searching: State-of-the-art review and future trends," *Comput.-Aided Des.*, vol. 37, no. 5, pp. 509–530, 2005.

[6] Y. Fang *et al.*, "3D deep shape descriptor," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 2319–2328.

[7] J. Xie, Y. Fang, F. Zhu, and E. Wong, "Deepshape: Deep learned shape descriptor for 3D shape matching and retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1275–1283.

[8] M. Eitz, R. Richter, T. Boubekeur, K. Hildebrand, and M. Alexa, "Sketch-based shape retrieval," *ACM Trans. Graph.*, vol. 31, no. 4, 2012, Art. no. 31.

[9] B. Gong, J. Liu, X. Wang, and X. Tang, "Learning semantic signatures for 3D object retrieval," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 369–377, Feb. 2013.

[10] T. Furuya and R. Ohbuchi, "Ranking on cross-domain manifold for sketch-based 3D model retrieval," in *Proc. Int. Conf. Cyberworlds (CW)*, Oct. 2013, pp. 274–281.

[11] B. Li *et al.*, "A comparison of methods for sketch-based 3D shape retrieval," *Comput. Vis. Image Understand.*, vol. 119, pp. 57–80, Feb. 2014.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[13] B. Li *et al.*, "SHREC'16 track: 3D sketch-based 3D shape retrieval," in *Proc. Eurograph. Workshop 3D Object Retr.*, 2016, pp. 1–8.

[14] T. Darom and Y. Keller, "Scale-invariant features for 3-D mesh models," *IEEE Trans. Image Process.*, vol. 21, no. 5, pp. 2758–2769, May 2012.

[15] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[16] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 3360–3367.

[17] G. Dai, J. Xie, F. Zhu, and Y. Fang, "Deep correlated metric learning for sketch-based 3D shape retrieval," in *Proc. AAAI*, 2017, pp. 4002–4008.

[18] A. S. Mian, M. Bennamoun, R. Owens, D. Mathers, and G. L. Hingston, "3D face recognition by matching shape descriptors," in *Proc. IVCNZ*, vol. 4. 2004, pp. 23–28.

[19] A. S. Mian, M. Bennamoun, and R. A. Owens, "Matching tensors for pose invariant automatic 3D face recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)-Workshops*, Sep. 2005, p. 120.

[20] A. Mian, M. Bennamoun, and R. Owens, "2D and 3D multimodal hybrid face recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 344–355.

[21] M. Elad, A. Tal, and S. Ar, "Content based retrieval of VRML objects—An iterative and interactive approach," in *Multimedia*. Vienna, Austria: Springer, 2002, pp. 107–118.

[22] D. V. Vranic, D. Saupe, and J. Richter, "Tools for 3D-object retrieval: Karhunen-Loeve transform and spherical harmonics," in *Proc. IEEE 4th Workshop Multimedia Signal Process.*, Oct. 2001, pp. 293–298.

[23] J. Sun, M. Ovsjanikov, and L. Guibas, "A concise and provably informative multi-scale signature based on heat diffusion," *Comput. Graph. Forum*, vol. 28, no. 5, pp. 1383–1392, 2009.

[24] A. M. Bronstein, M. M. Bronstein, R. Kimmel, M. Mahmoudi, and G. Sapiro, "A Gromov–Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching," *Int. J. Comput. Vis.*, vol. 89, nos. 2–3, pp. 266–286, 2010.

[25] A. Belyaev and M. Garland, "Laplace-Beltrami eigenfunctions for deformation invariant shape representation," in *Proc. Eurograph. Symp. Geometry Process.*, 2002, pp. 1–9.

[26] A. M. Bronstein, M. M. Bronstein, L. J. Guibas, and M. Ovsjanikov, "Shape Google: Geometric words and expressions for invariant shape retrieval," *ACM Trans. Graph.*, vol. 30, no. 1, 2011, Art. no. 1.

[27] G. Lavoué, "Combination of bag-of-words descriptors for robust partial shape retrieval," *Vis. Comput.*, vol. 28, no. 9, pp. 931–942, 2012.

[28] B. Leng, S. Guo, X. Zhang, and Z. Xiong, "3D object retrieval with stacked local convolutional autoencoder," *Signal Process.*, vol. 112, pp. 119–128, Jul. 2015.

[29] R. Litman, A. Bronstein, M. Bronstein, and U. Castellani, "Supervised learning of bag-of-features shape descriptors using sparse coding," *Comput. Graph. Forum*, vol. 33, no. 5, pp. 127–136, 2014.
[30] S. Bai, X. Bai, W. Liu, and F. Roli, "Neural shape codes for 3D model retrieval," *Pattern Recognit. Lett.*, vol. 65, pp. 15–21, Nov. 2015.
[31] X. Bai, S. Bai, Z. Zhu, and L. J. Latecki, "3D shape matching via two layer coding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp. 2361–2373, Dec. 2015.
[32] S. Bai, X. Bai, Z. Zhou, Z. Zhang, and L. J. Latecki, "GIFT: A real-time and scalable 3D shape search engine," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5023–5032.
[33] J.-L. Shih, C.-H. Lee, and J. T. Wang, "A new 3D model retrieval approach based on the elevation descriptor," *Pattern Recognit.*, vol. 40, no. 1, pp. 283–295, 2007.
[34] M. M. Bronstein and I. Kokkinos, "Scale-invariant heat kernel signatures for non-rigid shape recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 1704–1711.
[35] M. Aubry, U. Schlickewei, and D. Cremers, "The wave kernel signature: A quantum mechanical approach to shape analysis," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 1626–1633.
[36] B. Shi, S. Bai, Z. Zhou, and X. Bai, "DeepPano: Deep panoramic representation for 3-D shape recognition," *IEEE Signal Process. Lett.*, vol. 22, no. 12, pp. 2339–2343, Dec. 2015.
[37] F. Zhu, J. Xie, and Y. Fang, "Learning cross-domain neural networks for sketch-based 3D shape retrieval," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 3683–3689.
[38] T. Funkhouser *et al.*, "A search engine for 3D models," *ACM Trans. Graph.*, vol. 22, no. 1, pp. 83–105, 2003.
[39] P. Daras and A. Axenopoulos, "A 3D shape retrieval framework supporting multimodal queries," *Int. J. Comput. Vis.*, vol. 89, nos. 2–3, pp. 229–247, 2010.
[40] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Comput. Vis.*, Oct. 2003, pp. 1470–1477.
[41] S. Biasotti *et al.*, "Retrieval and classification methods for textured 3D models: A comparative study," *Vis. Comput.*, vol. 32, no. 2, pp. 217–241, 2015.
[42] F. P. Tasse and N. Dodgson, "Shape2Vec: Semantic-based descriptors for 3D shapes, sketches and images," *ACM Trans. Graph.*, vol. 35, no. 6, 2016, Art. no. 208.
[43] J. Xie, G. Dai, F. Zhu, and Y. Fang, "Learning barycentric representations of 3D shapes for sketch-based 3D shape retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3615–3623.
[44] B. Li *et al.*, "SHREC'13 track: Large scale sketch-based 3D shape retrieval," in *Proc. 6th Eurograph. Workshop 3D Object Retr.*, 2013, pp. 89–96.
[45] B. Li *et al.*, "SHREC'14 track: Extended large scale sketch-based 3D shape retrieval," in *Proc. Eurograph. Workshop 3D Object Retr.*, 2014, pp. 121–130.
[46] B. Li *et al.*, "A comparison of 3D shape retrieval methods based on a large-scale benchmark supporting multimodal queries," *Comput. Vis. Image Understand.*, vol. 131, pp. 1–27, Feb. 2015.
[47] M. Eitz, J. Hays, and M. Alexa, "How do humans sketch objects?" *ACM Trans. Graph.*, vol. 31, no. 4, pp. 44:1–44:10, 2012.
[48] B. Li *et al.*, "SHREC'12 track: Generic 3D shape retrieval," in *Proc. 3DOR*, 2012, pp. 119–126.
[49] A. Tatsuma, H. Koyanagi, and M. Aono, "A large-scale shape benchmark for 3D object retrieval: Toyohashi shape benchmark," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2012, pp. 1–10.
[50] D. V. Vranić and D. Saupe, "3D model retrieval," in *Proc. SCCG*, 2000, p. 89.
[51] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, "Information-theoretic metric learning," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 209–216.
[52] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? Metric learning approaches for face identification," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 498–505.
[53] M. Köstinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 2288–2295.
[54] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
[55] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[56] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 1. San Diego, CA, USA, Jun. 2005, pp. 539–546.
[57] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *J. Mach. Learn. Res.*, vol. 10, pp. 207–244, Feb. 2009.
[58] J. Hu, J. Lu, and Y.-P. Tan, "Discriminative deep metric learning for face verification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1875–1882.
[59] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese. (2015). "Deep metric learning via lifted structured feature embedding," [Online]. Available: https://arxiv.org/abs/1511.06452
[60] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *Proc. Int. Workshop Similarity-Based Pattern Recognit.*, 2015, pp. 84–92.
[61] J. Hu, J. Lu, and Y.-P. Tan, "Deep transfer metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 325–333.
[62] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola, "A kernel method for the two-sample-problem," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 513–520.
[63] S. J. Pan, J. T. Kwok, and Q. Yang, "Transfer learning via dimensionality reduction," in *Proc. AAAI*, vol. 8. 2008, pp. 677–682.
[64] C.-Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. Artif. Intell. Stat.*, 2015, pp. 562–570.
[65] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1395–1403.
[66] I. Lim, A. Gehre, and L. Kobbelt, "Identifying style of 3D shapes using deep metric learning," *Comput. Graph. Forum*, vol. 35, no. 5, pp. 207–215, Aug. 2016.
[67] H. Zhu, J.-B. Weibel, and S. Lu, "Discriminative multi-modal feature fusion for RGBD indoor scene recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2969–2976.
[68] F. Wang, L. Kang, and Y. Li, "Sketch-based 3D shape retrieval using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1875–1883.
[69] J. Bromley *et al.*, "Signature verification using a 'Siamese' time delay neural network," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 7, no. 4, pp. 669–688, 1993.
[70] K. He, X. Zhang, S. Ren, and J. Sun. (2015). "Deep residual learning for image recognition." [Online]. Available: https://arxiv.org/abs/1512.03385
[71] J. M. Saavedra, B. Bustos, T. Schreck, S. M. Yoon, and M. Scherer, "Sketch-based 3D model retrieval using keyshapes for global and local representation," in *Proc. 3DOR*, 2012, pp. 47–50.
[72] P. Sousa and M. J. Fonseca, "Sketch-based retrieval of drawings using spatial proximity," *J. Vis. Lang. Comput.*, vol. 21, no. 2, pp. 69–80, 2010.
[73] K. Siddiqi, J. Zhang, D. Macrini, A. Shokoufandeh, S. Bouix, and S. Dickinson, "Retrieving articulated 3-D models using medial surfaces," *Mach. Vis. Appl.*, vol. 19, no. 4, pp. 261–275, 2008.
[74] B. Li *et al.*, "3D sketch-based 3D model retrieval," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, 2015, pp. 555–558.
[75] B. Li *et al.*, "KinectSBR: A Kinect-assisted 3D sketch-based 3D model retrieval system," in *Proc. 5th ACM Int. Conf. Multimedia Retr.*, 2015, pp. 655–656.

Guoxian Dai received the master's degree from Fudan University, China. He is currently pursuing the Ph.D. degree with the Department of Computer Science and Engineering, New York University Tandon School of Engineering. His current research interests focus on 3D shape analysis, such as 3D shape retrieval and cross-domain 3D model retrieval.

Jin Xie received the Ph.D. degree from the Department of Computing, The Hong Kong Polytechnic University. He is currently a Research Scientist with New York University Abu Dhabi and New York University Tandon School of Engineering. His research interests include computer vision and machine learning. He is currently focusing on 3D computer vision with convex optimization and deep learning methods.

Yi Fang received the Ph.D. degree from Purdue University with a focus on computer graphics and vision. He was a Research Intern with Siemens, Princeton, NJ, USA, with one year industry experience, a Senior Research Scientist with Riverain Technologies, Dayton, OH, USA, and a Senior Staff Scientist with the Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, with six months academic experience, and he joined New York University Abu Dhabi as an Assistant Professor of electrical and computer engineering. He is currently involved in the development of state-of-the-art techniques in large-scale visual computing, deep visual learning, deep cross-domain and cross-modality model, and their applications in engineering, social science, medicine, and biology.