

Fan Zhu

Personal Information

Status: Post-doctor

Program: Computer Science and Engineering

School: Tandon School of Engineering, New York University

Website: [https://www.linkedin.com/in/fan-zhu-66b56b73/
?originalSubdomain=ae](https://www.linkedin.com/in/fan-zhu-66b56b73/?originalSubdomain=ae)

Period: From 2014-11 to 2017-03

Biography

I was a post-doctor at New York University and advised by Professor Yi Fang. During my post doctoral period, I was a research assistant in NYU Multimedia and Visual Computing (MMVC) Lab. I am broadly interested in 3D Computer Vision and Deep Learning. Now I am a director of project coordination at Inception Institute of Artificial Intelligence, Abu Dhabi.

Research Project: Learning to synthesize 3d indoor scenes from monocular images

Description

Depth images have always been playing critical roles for indoor scene understanding problems, and are particularly important for tasks in which 3D inferences are involved. In this project, we consider the scenarios where depth images are not available in the testing data, and propose to learn a convolutional long short-term memory (Conv LSTM) network and a regression convolutional neural network (regression ConvNet) using only monocular RGB images. The proposed networks benefit from 2D segmentations, object-level spatial context, object-scene dependencies and objects' geometric information, where optimization is governed by the semantic label loss, which measures the label consistencies of both objects and scenes, and the 3D geometrical loss, which measures the correctness of objects' 6D of estimation. Both quantitative and qualitative experimental results are provided on the NYU-v2 dataset, and we demonstrate that the proposed Conv LSTM can achieve state-of-the-art performance without requiring the depth information.

Method

We propose to learn a convolutional long short-term memory (Conv LSTM) network from purely monocular RGB scene images for estimating 2D object locations and dimensions, and we further build a regression ConvNet for estimating objects' degrees of freedom (Dof) information. Conv LSTM and regression ConvNet are trained separately following an end-to-end learning strategy, where the former maps pixels of object segments within scene images to both semantic object labels and scene labels, while capturing the inter-object and object-scene relations using the recurrent unit, and the latter maps pixels of object segments to parametrized object poses, position and dimension information so as to provide 3D inferences in continuous forms. By enforcing some reasonable geometric constrained factors on objects' Dof, estimations obtained from Conv LSTM and regression ConvNet can jointly provide 3D inferences to an object given only

monocular RGB queries.

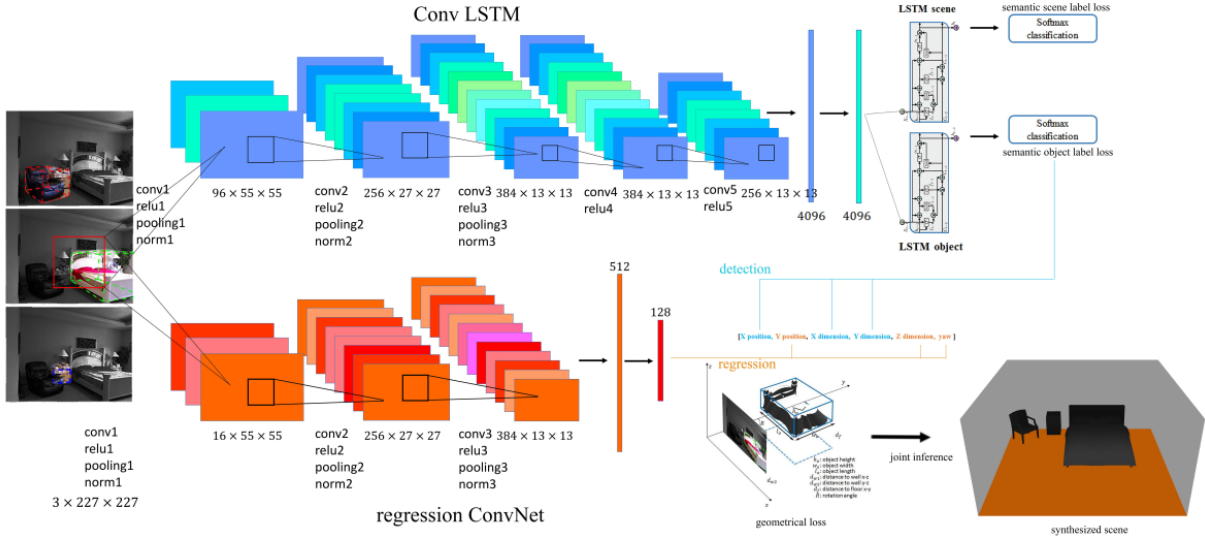


Figure 1: Overview of our approach: Annotations on ground-truth object regions and scene images are used for training the Conv LSTM network and regression ConvNet. The top learning stream demonstrates the Conv LSTM framework, which seamlessly connects CNN with two LSTM modules, where we compute the Softmax output of each LSTM hidden layer values to obtain the semantic scene label loss and the semantic object label loss respectively. The regression ConvNet (shown in the bottom stream) is trained using the geometric loss, which measures the correctness between ground-truth Dof values and regression values obtained by the network. Conv LSTM and regression ConvNet jointly provide inferences for objects’ poses, positions and dimensions in the 3D space, and can eventually generate a 3D scene that agrees with the floor plan of the query monocular image.

Results

Experiments on NYU-v2 dataset demonstrate the effectiveness of introducing the LSTM recurrent unit into a pure ConvNet framework by showing consistent improvements over directly fine-tuned CNN. Also, we demonstrate that training the regression ConvNet from scratch can achieve significantly less error rate than the fine-tuned CNN approach. In addition to achieving state-of-the-art performance on object/scene classification, object detection and object Dof estimation tasks while without requiring any depth information in the test stage, we further demonstrate qualitative results of synthesized 3D scenes that agree with the inferred room floor plan based on monocular indoor scene image.

Methods/Configuration	scene classification
Lin <i>et al.</i> [28]	58.72%
CNN [23]	35.87%
CNN fine-tuned	56.25%
Conv LSTM (salient segments)	57.87%
Conv LSTM (ground-truth segments)	59.63%

Figure 2: Scene classification results on NYU depth V2 dataset.

Methods/Configuration	object classification
Lin <i>et al.</i> [28]	60.49%
CNN [23]	40.22%
CNN fine-tuned	59.63%
Conv LSTM (ground-truth segments)	62.53%

Figure 3: Object classification results on NYU depth V2 dataset.




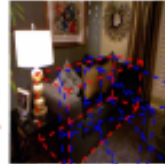
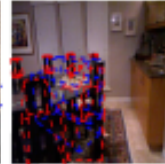
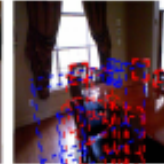











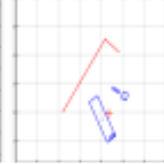
Detection						
Ground-truth						
Estimated						
Rotation error	3.6°	92.3°	88.8°	101.6°	62.5°	31.3°

Figure 4: Qualitative illustrations of estimation results. The top row: ground-truth 3D bounding boxes (red) and estimated 3D bounding boxes for object categories that are directly placed on the floor; the second row: ground-truth top view room layout; the third row: estimated top view room layout; the bottom row: mean objects' rotation error.

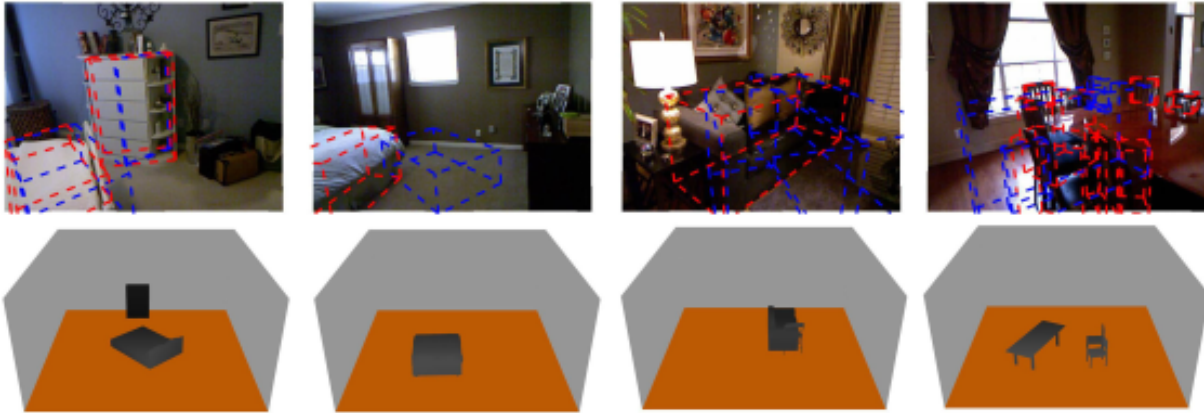


Figure 5: Qualitative illustrations of synthesized 3D scenes.

Research Project: Learning cross-domain neural networks for sketch-based 3d shape retrieval

Description

Sketch-based 3D shape retrieval, which returns a set of relevant 3D shapes based on users' input sketch queries, has been receiving increasing attentions in both graphics community and vision community. In this project, we address the sketch-based 3D shape retrieval problem with a novel Cross-Domain Neural Networks (CDNN) approach, which is further extended to Pyramid Cross-Domain Neural Networks (PCDNN) by cooperating with a hierarchical structure. By constructing cross-domain neural networks at multiple pyramid levels, a many-to-one relationship is established between a 3D shape feature and sketch features extracted from different scales. We evaluate the effectiveness of both CDNN and PCDNN approach on the extended large-scale SHREC 2014 benchmark. Experimental results suggest that both CDNN and PCDNN can outperform state-of-the-art performance, where PCDNN can further improve CDNN when employing a hierarchical structure.

Method

In order to alleviate the domain discrepancy between sketches and 3D shapes, we construct pyramid cross-domain neural networks (PCDNN), which map the mismatched sketch and 3D shape low-level representations to a unified feature space at multiple pyramid levels. The pyramid structure is defined by a fixed hierarchy of rectangular windows, and computes local histogram features within each subdivided image regions. Within each pyramid level, a cross-domain neural network (CDNN) pair with identical target layers for objects of the same category is learned. By extending a CDNN to the pyramid structure, multi-resolution sketch histograms are mapped to corresponding 3D shapes. When sketch queries pass through the learned neural networks, multi-resolution histogram features are computed and fed into corresponding neural networks, followed by which hidden layers are extracted from networks of all levels and concatenated as final representations.

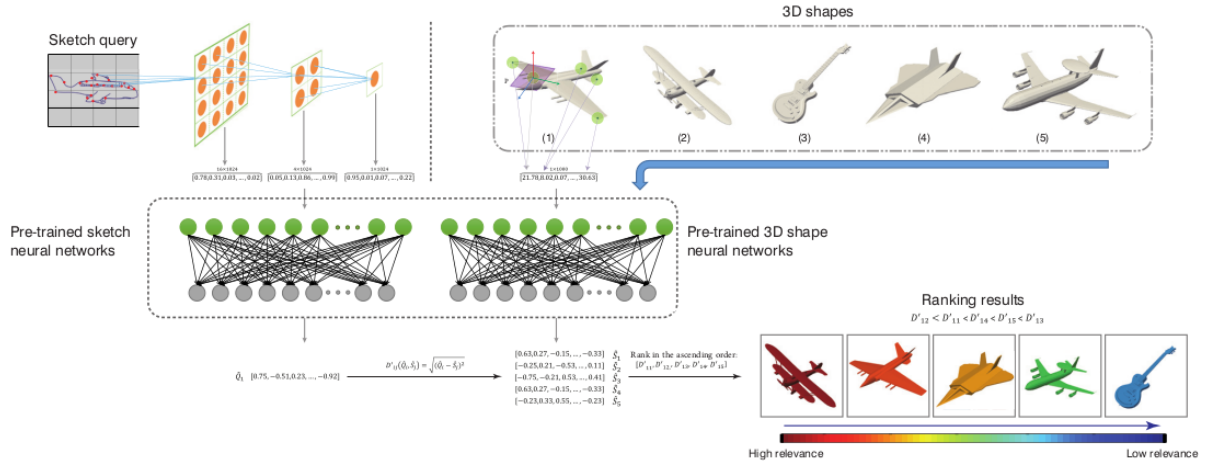


Figure 6: Illustration of the pipeline of our proposed sketch-based 3D shape retrieval framework.

Results

The proposed methods are evaluated on the large-scale extended SHREC 2014 sketch-based 3D shape retrieval benchmark. Sufficient experimental results suggest both methods can achieve the state-of-the-art performance, and PCDNN can further improve CDNN when cooperating with the pyramid learning structure.

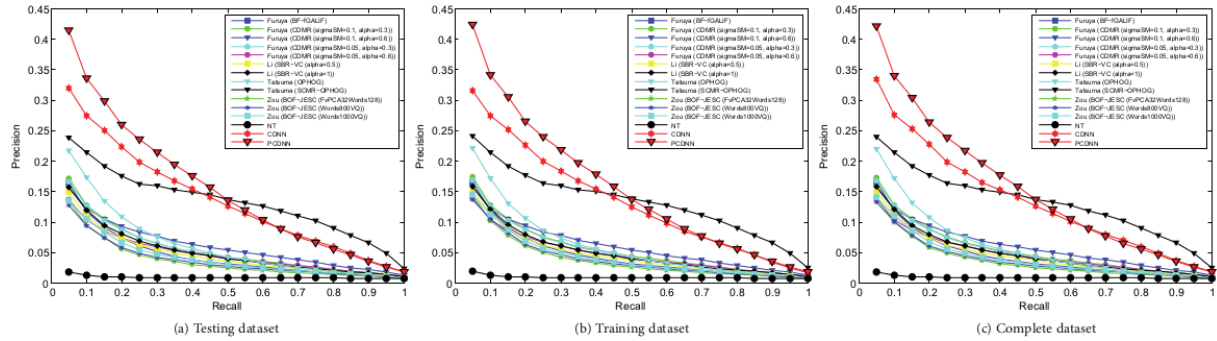


Figure 7: Precision-Recall plot performance comparisons on the extended large-scale SHREC’14 benchmark.

Sketch query	Retrieved 3D shapes				
Elephant	$S_1 = 7.28$	$S_2 = 7.27$	$S_3 = 6.99$	$S_4 = 6.88$	$S_5 = 6.81$
Airplane	$S_1 = 5.63$	$S_2 = 5.51$	$S_3 = 5.48$	$S_4 = 5.33$	$S_5 = 5.31$
Comb	$S_1 = 5.950$	$S_2 = 5.92$	$S_3 = 5.83$	$S_4 = 5.58$	$S_5 = 5.50$
Potted plant	$S_1 = 4.93$	$S_2 = 4.50$	$S_3 = 4.48$	$S_4 = 4.43$	$S_5 = 4.36$
Apple	$S_1 = 5.99$	$S_2 = 5.49$	$S_3 = 5.36$	$S_4 = 6.48$	$S_5 = 6.20$

Figure 8: Illustration of the top-5 3D shape retrieval results for some sketch queries in different categories. The scores on the top of retrieved 3D shapes denote confidence levels of retrieval according to query sketches. Each red rectangular denotes a false positive retrieval.

Research Project: Heat diffusion long-short term memory learning for 3d shape analysis

Description

The majority of prior heat kernel-based strategies of building 3D shape representations fail to investigate the temporal dynamics of heat flows on 3D shape surfaces over time. In this work, we address the temporal dynamics of heat flows on 3D shapes using the long-short term memory (LSTM). We guide 3D shape descriptors toward discriminative representations by feeding heat distributions throughout time as inputs to units of heat diffusion LSTM (HD-LSTM) blocks with a supervised learning structure. We further extend HD-LSTM to a cross-domain structure (CDHD-LSTM) for learning domain-invariant representations of multi-view data. We evaluate the effectiveness of both HD-LSTM and CDHD-LSTM on 3D shape retrieval and sketch-based 3D shape retrieval tasks respectively. Experimental results on McGill dataset and SHREC 2014 dataset suggest that both methods can achieve state-of-the-art performance.

Method

We start by computing the heat kernel features (i.e., HKS) from 3D shapes, and learn the heat diffusion kernel distributions overall all sampling time-steps through HD-LSTM, where the heat diffusion kernel distribution is the histogram of heat diffusion values given a fixed time-step. We then guide the input features towards discriminative 3D shape representations through a supervised LSTM learning structure, where the category information of training samples are supplied to the output end of LSTM in the form of discriminative vectors. When the heat flows sequentially pass through the HD-LSTM, its “forget gate layer” can selectively throw away the previous heat flow from the cell state, and determine how much we decide to update the current state value using the past data. We then extend HD-LSTM to a cross-domain learning structure CDHD-LSTM, which minimizes the cross-domain discrepancy by connecting HD-LSTM to a 3-layer neural network and guiding same-category cross-domain data toward identical targets.

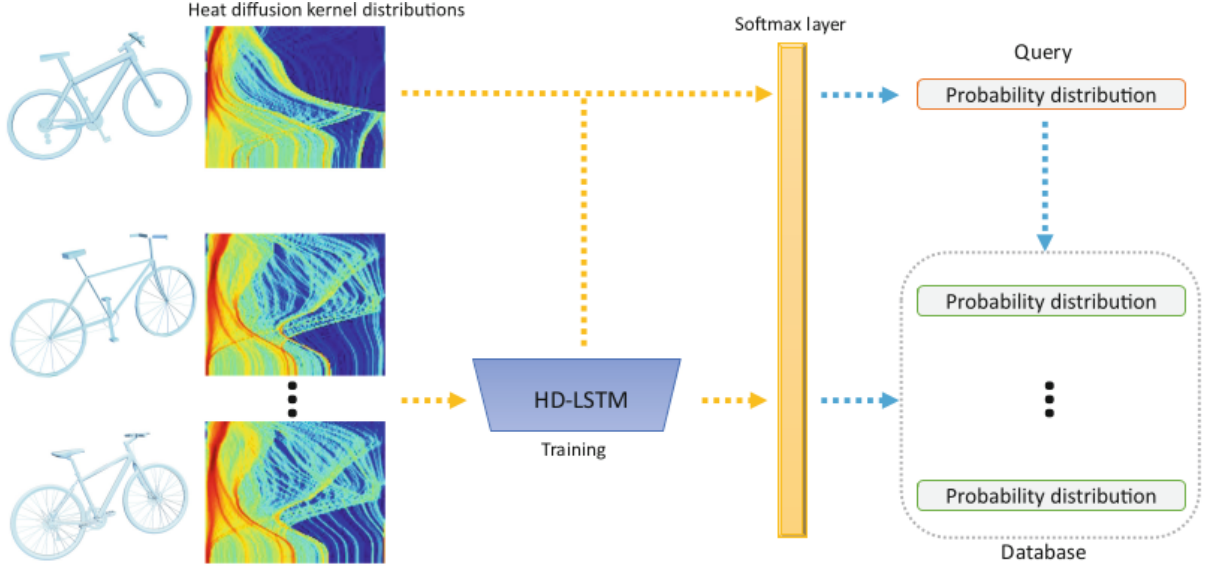


Figure 9: We discover the temporal dynamics of heat diffusions and correspondingly propose HD-LSTM to learn discriminative 3D shape representations based on heat diffusions.

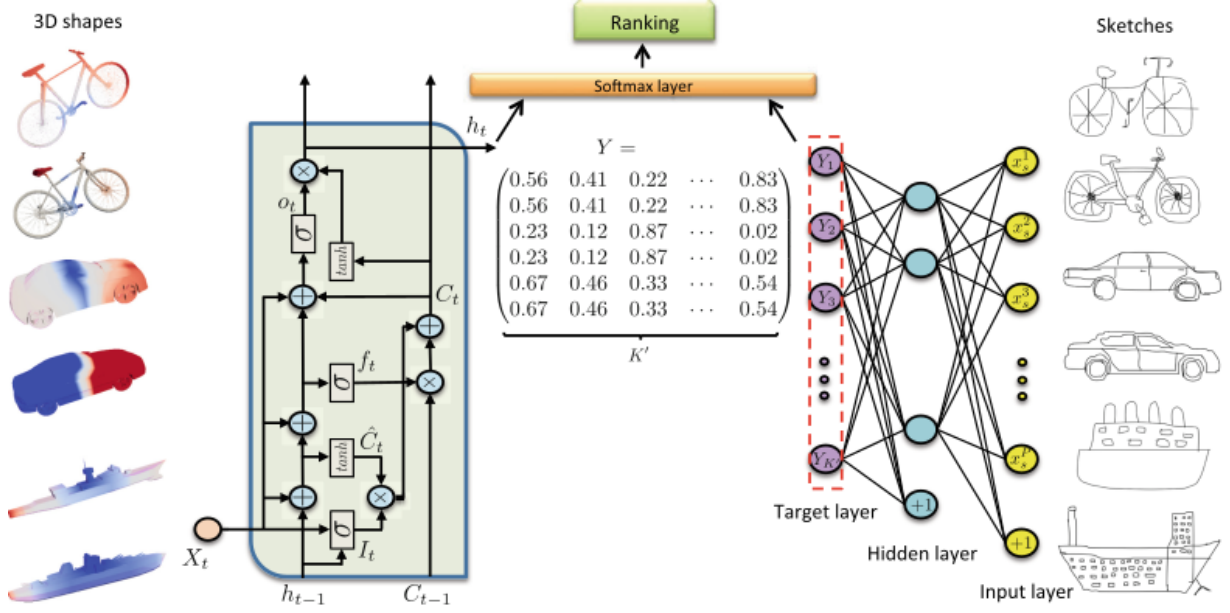


Figure 10: Learning domain-invariant representations for sketch-based 3D shape retrieval using the CDHD-LSTM architecture. CDHD-LSTM is constructed by connecting a 3-layer neural network to HD-LSTM at the output ends, where the connection is established by sharing identical discriminative random vectors for sketches and 3D shapes that come from the same category.

Results

In order to demonstrate the effectiveness of the proposed HD-LSTM method, we conduct experiments on 3D shape retrieval tasks using the McGill dataset. The 5 commonly used evaluation metrics, nearest neighbor (NN), first tier (1-Tier), second tier (2-Tier), discounted cumulated gain (DCG) and average precision (AP) are used for evaluating the performance of the proposed methods and comparison methods. We evaluate the performance of CDHD-LSTM on the sketch-based 3D shape retrieval task using the extended large scale SHREC 2014 dataset. Experimental results on the MacGill shape dataset and the extended SHREC 2014 dataset suggest both HD-LSTM and CDHD-LSTM can achieve state-of-the-art performance.

Methods	NN	1-Tier	2-Tier	DCG	AP
Hybrid BoW [23]	0.95	0.63	0.79	0.88	—
Covariance method [35]	0.97	0.73	0.81	0.93	—
Graph-based method [1]	0.97	0.74	0.91	0.93	—
DeepShape [39]	0.98	0.78	0.83	—	—
BoW	0.80	0.40	0.54	0.70	0.46
HD-LSTM (without softmax)	0.97	0.88	0.83	0.88	0.90
HD-LSTM (with softmax)	0.98	0.92	0.95	0.95	0.94

Figure 11: Performance comparison between the proposed HD-LSTM method and the state-of-the-art methods on the McGill dataset.

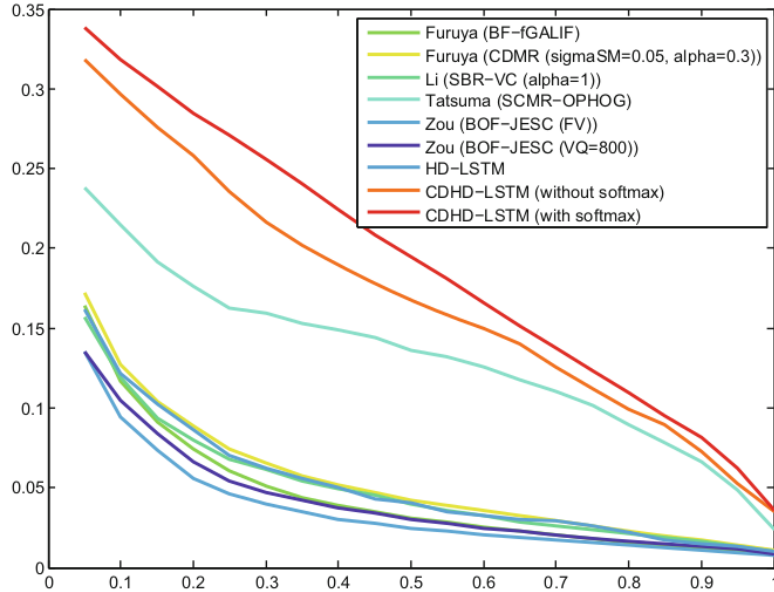


Figure 12: Precision-Recall plot of performance comparisons on the extended large-scale SHREC 14 sketch-based 3D shape retrieval dataset.

Research Project: Boosted cross-domain dictionary learning for visual categorization

Description

A boosted cross-domain categorization framework that utilizes labeled data from other visual domains as the auxiliary knowledge for enhancing the original learning system is presented. The source domain data under a different data distribution are adapted to the target domain through both feature representation level and classification level adaptation. The proposed framework is working in conjunction with a learned domainadaptive dictionary pair, so that both the source domain data representations and their distribution are optimized in order to match the target domain. Using a set of Web images and selected categories from the HMDB51 dataset as the source domain data, the proposed framework is evaluated with both image classification and human action recognition tasks on the Caltech-101 and the UCF YouTube datasets, respectively, achieving promising results.

Method

We introduce a boosted cross-domain categorization (BCDC) framework that utilizes labeled data from other domains as the source data to span the intra-class diversity of the original learning system. In addition to the manually annotated information in the target domain, partially labeled data from another visual domain are provided as the source domain. A boosted classification framework is introduced to work in conjunction with a cross-domain dictionary learning method. Through iteratively updating both the source domain data representations and their distribution, the source domain training instances can be optimized, and thus can help improve the visual categorization tasks in the target domain.

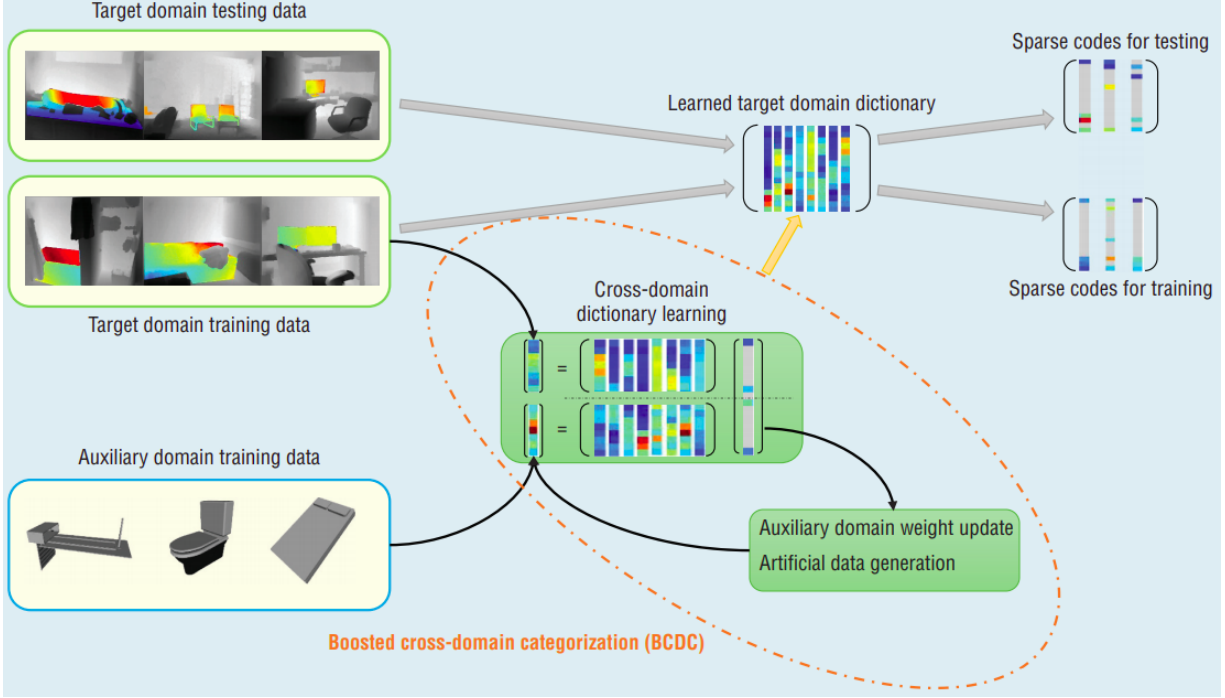


Figure 13: Knowledge transfer between 3D shapes to depth images.

Results

The experiments are conducted on 4 different data sources, where the UCF YouTube dataset and the Caltech-101 dataset are treated as the target domains, and the HMDB51 dataset and some Web images indexed by Google are treated as the source domains. Promising results are achieved on both image classification and action recognition, where knowledge from either the Web or a related dataset is transferred to standard benchmark datasets.

No. images	Algorithm					
	ScSPM (%)	K-SVD (%)	LC-KSVD (%)	TrAdaBoost (%)	WSCDDL (%)	BCDC (%)
30	79.11	79.98	81.32	84.37	86.52	87.34
25	75.05	75.06	79.68	81.46	84.31	85.90
20	65.44	67.40	73.04	79.72	80.02	82.32
15	49.66	54.12	69.23	75.53	77.59	78.69
10	30.65	46.28	64.89	72.87	74.98	76.04

Figure 14: Performance comparison between the proposed BCDC and state-of-the-art methods on the Caltech-101 dataset with source domain data.

No. images	Algorithm				
	ScSPM (%)	K-SVD (%)	LC-KSVD (%)	AdaBoost (%)	BCDC (%)
30	85.36	84.69	85.60	79.46	87.34
25	83.23	82.16	83.47	74.83	85.90
20	80.11	80.07	80.59	74.22	82.32
15	76.66	74.82	76.96	71.91	78.69
10	72.87	72.55	72.37	68.35	76.04

Figure 15: Performance comparison between the BCDC and state-of-the-art methods on the Caltech-101 dataset when the source domain data are only used by the BCDC.

No. actors	Algorithm					
	LLC (%)	KSVD (%)	LC-KSVD (%)	TrAdaBoost (%)	WSCDDL (%)	BCDC (%)
16	79.78	75.43	82.87	82.40	83.26	84.64
9	68.38	64.54	67.14	69.20	72.01	73.05
5	63.35	59.35	63.68	65.46	67.37	68.89

Figure 16: Performance comparison between the BCDC and state-of-the-art methods on the UCF YouTube dataset with source domain data.

No. actors	Algorithm				
	LLC (%)	KSVD (%)	LC-KSVD (%)	AdaBoost (%)	BCDC (%)
16	82.77	74.57	83.15	79.40	84.64
9	68.38	62.63	69.82	69.61	73.05
5	64.84	59.37	65.17	65.52	68.89

Figure 17: Performance comparison between the BCDC and state-of-the-art methods on the UCF YouTube dataset when the source domain data are only used by the BCDC.