Learning Multi-Instance Deep Ranking and Regression Network for Visual House Appraisal

Xiaobai Liu, *Member, IEEE*, Qian Xu, Jingjie Yang, Jacob Thalman, Shuicheng Yan, and Jiebo Luo, *Fellow, IEEE*

Abstract—This paper presents a weakly supervised regression model for the *visual house appraisal* problem, which aims to predict the value of a house from its photos and textual descriptions (e.g., number of bedrooms). The key idea of our approach is a multi-layer neural network, called *multi-instance Deep Ranking and Regression* (MiDRR) net, which jointly solves two coupled tasks: ranking and regression, in the multiple instance setting. The network is trained using weakly supervised data, which do not require intensive human annotations. We also design a set of human heuristics to promote deep features through imposing constraints over the solution space, e.g., a house with three bedrooms often has a higher value than that with only two bedrooms. While these constraints are specific to the studied problem, the developed formula can be easily generalized to the other regression applications. For test and evaluation purposes, we collect a comprehensive house image benchmark that includes 900,000 photos from 30,000 houses recently traded in the USA, and apply the proposed MiDRR net to predict house values. Extensive evaluations with comparisons demonstrate that additional usage of imagery data as well as human heuristics can significantly boost system performance and that the proposed MiDRR net clearly outperforms the alternative methods.

Index Terms—Deep learning, ranking, regression, social network, multi-instance learning, house photos

1 Introduction

Background In modern housing market, house appraisal is a necessary step for all involved parties, including buyers, sellers, lenders, underwriters, and Realtors. To trade a house, it is a common practice to utilize a large number of house photos to attract potential buyers and persuade them to make offers. To automatically estimate the value of a house, we need to answer a key question: *what does an experienced human appraiser look for to value a home?* Basically, we have found that the following aspects are most significant when assessing house values: textual or visual features of houses; comparisons to similar properties, and assessment of the surrounding area.

To study the impacts of the above factors over house pricing, we collect a database from public on-line real estate markets, e.g., Multiple Listing Service (MLS), that document hundreds of thousands of houses and their trading histories. For each house, we download both textual features (e.g., house size, number of rooms, land types, etc.) and house photos, and take the mostly recent deal price as the true house value. With this dataset, our objective is to study an effective automatic approach to accurately predict the value of a house from its textual and visual features, i.e., visual house appraisal.

Visual house appraisal is a challenging multi-modal problem because it is actually solving the regression of high-dimensional data, including both textual features and house photos. Moreover, there has been studies showing that it is difficult, even for human beings, to perform quantitative perceptions [1]. The study is also related to visual persuasion [18], a standing fundamental problem in Artificial Intelligence (AI) field.

1.1 Overview of the Proposed Method

In this work we cast the visual house appraisal task as a weakly supervised learning problem. For each house, we use both textual features (e.g., number of bedrooms, school rating) and house photos as inputs. A house photo often covers a portion of the house, e.g., kitchen room, living room, backyard, etc., and is expected to have different values. For example, a well-modeled kitchen room is often more expensive than an original kitchen room, or a well-maintained backyard with colorful flower fields would have a higher value than a less-maintained backyard. However, it is in general difficult to assess the separate values of these individual house parts. In contrast, it is relatively easy to access the sold price for the whole house, e.g., from the housing website. Therefore, the task of learning a visual house appraisal model falls in the multi-instance setting [35], for which only bag-level labels are available. In this work, a bag represents a house and an instance represents a house

- *X. Liu, J. Yang, and J. Thalman are with the Department of Computer Science, San Diego State University (SDSU), San Diego, CA 92182. E-mail: xiaobai.liu@sdsu.edu, jj-iceflower@hotmail.com, pthalman@gmail.com.*
- *Q. Xu is with XreLab Inc., San Diego, CA 92182. E-mail: langhter0124@gmail.com.*
- *S. Yan is with the Qihoo/360 Company, Beijing 100015, China, and the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 119077. E-mail: eleyans@nus.edu.sg.*
- *J. Luo is with the Department of Computer Science, University of Rochester, Rochester, NY 14627. E-mail: jluo@cs.rochester.edu.*

Manuscript received 1 Feb. 2017; revised 20 Nov. 2017; accepted 24 Nov. 2017. Date of publication 10 Jan. 2018; date of current version 5 July 2018. (Corresponding author: Xiaobai Liu.)
Recommended for acceptance by L. B. Holder.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2018.2791611

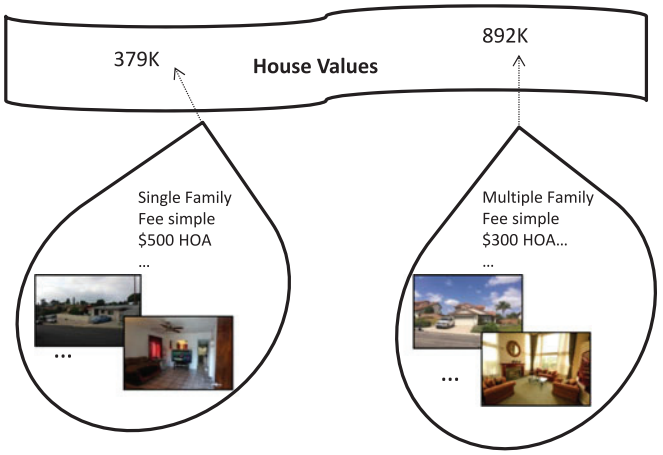


Fig. 1. Visual house appraisal. The objective of this work is to learn a prediction function to estimate the value of a house from its textual descriptions and house photos. *Top row*: The target solution space that explicitly embeds house values. *Bottom row*: Each house is considered as a bag of instances, where every instance indicates a house photo plus textual features.

photo. We augment each instance with the textual features of the house to get a cross-modal representation for individual part of the house. Fig. 1 demonstrates the proposed visual house appraisal method. The objective of our method is to learn to map a house instance into a target solution space.

Our approach is motivated with a fact: given the textual features and photos of two houses, ranking them according to their values is relatively easier than directly predicting their respective absolute house prices [1], [28], [31]. Taking the house photos in Fig. 1 for an example, one can easily identify that the house on the left hand side has lower value than that on the right hand side, whereas it is challenging to estimate their respective values. In fact, in house appraisal we human beings often employ common sense knowledge to help perceive which one is more expensive. For example, a house of multiple family is more expensive than a house of simple family in the same neighborhood; a house with swimming pool is more expensive than a house without pool; just to mention a few. The relative orders of houses, once estimated, can be used to regularize the estimation of the values of individual houses; and vice versa. Therefore, it is mutually beneficial to solve these two tasks in a tightly coupled fashion.

We propose to learn a multi-purpose representation for simultaneous regression and ranking purposes. To leverage the recent advantages in deep feature learning [15], [19], we parameterize the representation using a multi-layer neural network. Our network takes as inputs a house instance, i.e., a house photo plus textual house features, and comprises of three subnetworks: i) subnetwork A, a convolution neural network that processes imagery data; ii) subnetwork B, a feed forward network that processes textual data; and iii) subnetwork C, a feed forward neural network that is fed with the fusion of the output activations of the subnetworks A and B. A fusion layer is used to accumulate the activations of subnetworks A and B and is connected to a ranking loss. The whole network architecture is trained with an objective function for regression purpose. Both loss functions are formulated in the multiple instance setting. The proposed network, called *Multi-instance Deep Ranking and Regression network* (MiDRR), can be trained with the standard back propagation (BP) algorithm in an end-to-end fashion [19]. We develop an alternating algorithm to optimize the two objectives iteratively.

For test and evaluation purposes, we create a comprehensive image dataset for studying the visual house appraisal problem, which is the first one is it catalog. The dataset includes about 900,000 house photos of 30,000 houses in California, U.S.A., which were traded in an online website. For each house, we use its sold price as the ground-truth price label. We evaluate the proposed method on the collected dataset and compare it to the alternative algorithms. The average appraisal error of our approach is less than 1 percent the house value (e.g., $5,000 for a house of $500,000), which is encouraging considering that the house prices in our dataset vary between $102,000 and $2,100,000. We also demonstrate that using house photos, in addition to textual features, can significantly improve the accuracy of house appraisal.

1.2 Contributions and Paper Organization

The two major *contributions* of this paper include (i) a comprehensive image benchmark for studying the multi-modality house appraisal problem, which will be released for public use in order to foster research in this novel direction; and (II) an effective multi-instance deep ranking and regression method for visual house appraisal that outperforms the other appraisal methods with good margins.

Organization. In the rest of this paper, we review the related literature in Section 2, and introduce the proposed multi-instance deep regression and ranking method in Section 3. Then, we discuss how to apply MiDRR for visual house appraisal task in Section 4. Last, we report the experimental results in Section 5.

2 Background and Related Works

This work is closely related to three research streams in the fields of data engineering and machine learning.

Multi-dimensional Regression is one of the foundational tasks in statistics and machine learning. In the past literature, it was addressed with a wide variety of methods, including support vector machine [3], [25], generalized linear model [23], boosting [33], random forest [7], and neural network [30]. In addition to these efforts, researchers also proposed to jointly optimize ranking loss and regression loss and achieved impressive results on image retrieval [37] or click predictions [28]. However, these algorithms are restricted to their shallow representations which are not robust against noises or outliers, especially while dealing with large-scale high-dimensional data. In this work, we develop a multi-instance deep ranking and regression method to fill in the gap, and demonstrate its potentials as an effective solution to visual house appraisal.

Multiple Instance Learning (MIL) was proposed to deal with weakly supervised data [10], where a bag includes multiple instances and only bag-level labels are available. MIL problems have been addressed with a number of machine learning algorithms, including SVM [2], neural network [9], [27], [40], and boosting [33]. These multi-instance algorithms have wide applications in computer vision, e.g., face detection [33], categorization [32], segmentation [36] image retrieval [35], and medical image

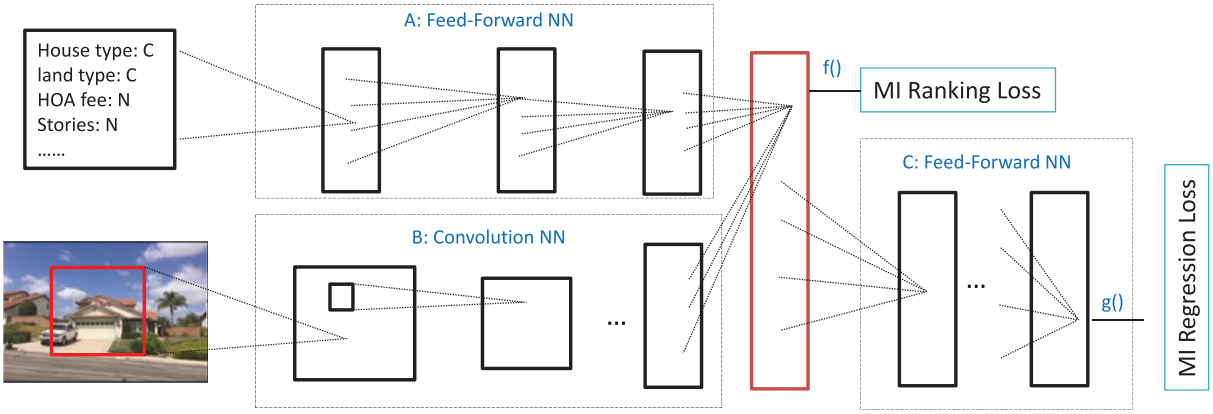


Fig. 2. Multi-instance Deep Ranking and Regression (MiDRR) network. It is comprised of three sub-networks: i) A feed-forward network that takes as inputs textual features; ii) a Convolution Neural Network (CNN) that takes as inputs images; and iii) a feed-forward network for regression purpose. These three networks are trained with loss functions for ranking and regression purposes in the multi-instance setting.

analysis, e.g., Mammogram classification [41]. In this work, we propose to train a deep regression and ranking network in the setting of multi-instance, which is a novel learning schema in its catalog.

Deep Neural Network (DNN) has shown great potentials on many aspects of data-intensive engineering systems, e.g., image classification [6], [14], [39], landmark recognition [38], and image retrieval [8], etc., mostly driven by the availability of large-scale labeled data, e.g., ImageNet [20]. The proposed MiDRR net is motivated with two recent DNN techniques. The first one is to formulate DNN in the multi-instance setting in order to utilize weakly annotated data [11], [36]. The other one is to learn a deep convolution network with ranking loss [16] for image annotations problems [13]. Moreover, Belagiannis et al. [4] proposed a robust optimization method for learning deep regression models to address outliers or difficult samples. We extend these efforts to jointly optimize ranking and regression losses in the multi-instance setting. The proposed MiDRR net is also motivated by the previous efforts in mixing multiple sub-networks for boosting system performance, e.g., deep classification and reconstruction net [12], and network in network [21]. Similarly, we also employ multiple networks to learn deep features.

3 Multi-Instance Deep Ranking and Regression Net

Notations. Consider a set of training samples $(\mathbf{x}_i, y_i)$ where $\mathbf{x}_i$ denote a bag of instances, and y_i the related continuous labels. Our goal is to learn a prediction function $g_w()$ that can map an instance to the target label space and a rank function $f_w()$ that can sort two instances according to their continuous labels. For both cases, w indicates the parameters to be learned from training data. As aforementioned, the goal of this work is aimed at learning deep representations of $g_w()$ and $f_w()$ from weakly supervised data in the multi-instance setting.

3.1 Deep Ranking and Regression Net

We introduce a multi-layer neural network, called multi-instance Deep Ranking and Regression net, to parameterize the prediction function $g_w()$ as well as the ranking function $f_w()$. Fig. 2 illustrates the architecture of MiDRR net. The input is an instance with both textual and imagery data of a house and the outputs are the continuous labels and ranks. MiDRR comprises of three subnetworks. A feed-forward neural network (FNN) and a convolution neural network (CNN) are used to process the textual inputs and imagery inputs, respectively. The last hidden layers of these two networks are fully connected to a fusion layer, as is shown in red, which functions in two aspects. On the one hand, the output activations of the fusion layer are fully connected to a single output neuron, denoted by $f_w()$, whose output is used to rank two instances according to their continuous labels. We denote the parameters for the MiDRR net by w, with abuse of notation. On the other hand, the activations of the fusion layer are further fed to a FNN whose output neuron, denoted as $g_w()$, returns the continuous label. In the following two sections, we will derive the mathematical definitions of the two loss functions.

3.2 Multi-Instance Ranking Loss

The function $f_w()$ is used to map an instance to a scalar of rank. To learn the optimal $f_w()$, we collect a set of ordered pairs of bags, denoted by (i, j), where the bag i has a higher rank than the other bag j, and use these ordered bag-pair to learn the network parameters w. Let x_{ik} denote the kth instance of the ith bag. For any pair (i, j), the function $f_w()$ should follow the multiple instance constraint [34]: the maximal rank of the instances of the bag i should be higher than the rank of any instance in the bag j. Formally, we have a set of constraints

$$\max_k f_w(x_{ik}) \; > \; f_w(x_{jl}), \forall k, l, \forall (i, j), \tag{1}$$

which can be used to confine the feasible solution of w for the ranking purpose.

Accordingly, the ranking loss can be formulated as

$$\min_{w, \xi_{ijl}} \sum_{i,j,l} \xi_{ijl} + \frac{\lambda}{2} ||w||^2, \tag{2}$$

$$s.t., \quad \max_k f_w(x_{ik}) - f_w(x_{jl}) \; > \; 1 - \xi_{ijl}, \tag{3}$$

$$\xi_{ijl} \geq 0, \forall k, l, \forall (i, j), \tag{4}$$

where λ is a constant parameter. Following [24], we can replace the maximal operator on the left hand side of Eq. (3)

TABLE 1
Exemplar Textual Houses Features

Name	Type	Example values
house type	**C**ategory	attached, detached
land type	**C**ategory	single family, multiple family, townhouse, condo
HOA fee	**N**umber	100, 300, 588,...,
county	**C**ategory	San Diego, Poway, Le Mesa
built year	**N**umber	1987, 1988, 1989,...,
number of garage parking spaces	**N**umber	0, 1, 2, 3, ...,
number of bedroom	**N**umber	1, 2, 3, ...,
lot size	**N**umber	3999
has fencing	**B**oolean	yes, no
description	**T**ext	move-in ready ... located... bamboo floors...updated recently...

There are four feature types: boolean, number, category, and text, where text means the feature value consists of a few sentences.

by rewriting x_{ik} as the convex combination of the instances in the bag i.

3.3 Multi-Instance Regression Loss

The regression function $g_w()$ is used to predict the continuous label of the input instance which is passed through the three subnetworks. Herein, $g_w()$ follows the conventional multiple instance constraints [27], [33]: for all the instances in a bag, the maximal prediction of $g_w()$ should be equal to the ground-truth value which is only available for every bag. We have the following equation:

$$y_i = \max_k g_w(x_{ik}), \tag{5}$$

or, equivalently

$$y_i = g_w(x_{i,d(i)}), \quad s.t., \quad g_w(x_{i,d(i)}) \geq g_w(x_{ik}), \forall k, \tag{6}$$

where $d(i)$ indexes the instances of the bag i that has the maximal prediction $g_w()$. The regression loss function for the MiDRR net is defined as

$$\min \frac{1}{2} \sum_i \|y_i - g_w(x_{i,d(i)})\|^2 \tag{7}$$

$$s.t., \quad g_w(x_{i,d(i)}) \geq g_w(x_{ik}). \tag{8}$$

To minimize $Q()$, we can use the alternative optimization strategy. With $d(i)$ fixed, $Q()$ is a quadratic convex function with respect to the weight parameters w; with w at the present step, we can evaluate $g_w()$ to retrieve $d(i)$. For an unseen bag, its prediction can be made by passing all its instances through the MiDRR net and finding the maximal prediction of $g_w()$.

3.4 Learning by Back-Propagation

To solve the optimal parameters w, we convert Eqs. (2) and (7) into their unconstrained forms and use the back-propagation algorithm [19] to update the network parameters w iteratively. At each iteration, we calculate the derivatives of the loss function regarding w, and update the corresponding parameters with back-propagated gradients. The stochastic gradient descent method with a momentum term is used. The momentum weight is set to be 0.9, and mini-batch size to be 32. We set the global learning rate to be 0.004 at the beginning, and decay it over epochs.

TABLE 2
Commonsense Knowledge over Five Textual Features

Property	A	B	Order	Valid %
#Bedroom	3	2	A≥B	72
Room Area	1000 f^2	1500 f^2	B≥A	81
Lot Size	10000 f^2	8000 f^2	A≥B	78
Built Year	1988	1965	A≥B	62
#Parking Spaces	1	3	B≥A	59

The price order of two houses A and B are determined based on their property values. Column 5: the percentages of house pairs in training data that comply with the corresponding knowledge. Only house pairs of the same zipcode are counted.

4 MiDRR for Visual House Appraisal

In this section, we discuss the implementation of MiDRR for predicting house prices from both textual and imagery data.

Setting. Suppose that we have a set of houses, and each house is provided with a set of textual features and a collection of house photos. We consider each house as a bag of instances and every instance represents a house photo plus the textual house features, e.g., size, number of rooms, etc. House values are provided at bag-level. In online markets, a house usually has 20-50 photos, and is described with multiple textual features. Table 1 summarizes part of the textual features. A complete list of 48 textual features used in this work can be found in the Appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TKDE.2018.2791611.

Ranking Loss with Semantic-Awareness. The constraints in Eq. (3) are non-sense if the two house photos are not of the same category. It is meaningless to rank, for example, a garden and a kitchen room. Therefore, we categorize all house photos, e.g., gardens, bath-room, living room, etc., and apply the constraints in Eq. (3) only for pairs of photos belonging to the same category. Let c_{ik} denote the category of the image x_{ik}, we define the Semantic-aware objective function for ranking purpose as follows:

$$\min_{w,\epsilon_{ijl}} \quad R(w, \epsilon_{ijl}) = \sum_{i,j,l} \epsilon_{ijl} + \frac{\lambda}{2} ||w||^2 \tag{9}$$

$$s.t., \max_k f_w(x_{ik}) - f_w(x_{jl}) \; > \; 1 - \epsilon_{ijl}, \forall c_{ik} = c_{jl} \tag{10}$$

$$\xi_{ijl} \geq 0, \forall l, \forall (i,j). \tag{11}$$

To rank a pair of testing houses (i, j), we can pass each of their instances through the MiDRR net and test if Eq. (1) holds.

Regression Loss with Heuristics Constraints. We further introduce a set of human heuristics to regularize the learning of the regression loss function Eq. (7). These heuristics are derived from commonsense knowledge that are widely used by human beings while perceiving house prices. For example, a house with three bedrooms is often more expensive than a house with one bedroom; a house of single family has higher value than an apartment. Although individual constraints might be weak, the ensemble of such constraints can provide useful information to aid in the prediction of house prices. Table 2 summarizes five types of common sense knowledge as well as their instances. In the last column, we show the portion of ordered house pairs in the training set that comply with the constraints (in the respective rows).

TABLE 3
Configurations of Subnetworks

Name	Types	Inputs	Outputs	Layers
Subnetwork A	FNN	textual features (183D)	30D	5
Subnetwork B	CNN	images (240×240)	4096	8
Subnetwork C	FNN	fused vectors (4126D)	30	5

FNN: Feed-forward Neural Network; CNN: Convolution Neural Network.

Formally, let C denote the set of house-pairs, $(i, j) \in C$, where the house i has lower value than house j according to a type of commonsense knowledge. We revise the regression loss function in Eq. (7) as

$$\min \quad Q(\{d(i)\}, w) = \frac{1}{2}\sum_i \|y_i - g_w(x_{i,d(i)})\|^2 + \beta \sum_{i,c} \zeta_{ic} \tag{12}$$

$$s.t. \quad g_w(x_{i,d(i)}) \geq g_w(x_{ik})$$

$$g_w(x_{ik}) - g_w(x_{cl}) \; > \; 1 - \zeta_{ic}, \forall (i, c) \in \mathbf{C}. \tag{13}$$

Note that it is possible that two houses in C have conflicting orders according to different commonsense knowledge. The proposed formula simply uses these commonsense as soft constraints to confine the feasible solution space.

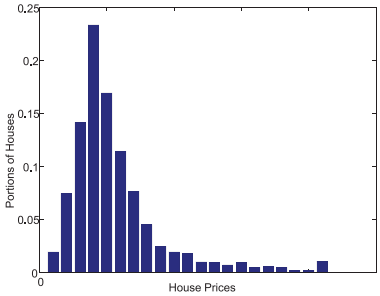
Implementation of Subnetwork A. We implement the feed forward neural network for processing textual features. We extract different types of textual features as follows. (i) For each field of text, we extract a binary bag-of-the-word feature vector, for which each dimension indicates if a word appears in the text. The vector is of 4053-dimension. We apply PCA method to reduce the dimension to be 40. (ii) For the fields of number, we scale all values to be within 0 and 1, by dividing the respective maximal values. (iii) For the boolean and category types, we introduce an augmented feature for each of the possible values. Taking land type for instance, we augment four more features, representing whether a house is a single family, multiple family, townhouse, or apartment, respectively. The Appendix, available in the online supplemental material, summarizes the list of textual features used in this work. We extract a 183 dimensional feature vector for each house, which specifies the input layers of the subnetwork A. The subnetwork A includes 5 hidden layers, with 80, 60, 50, 40 and 30 neuron units, respectively. We use the tangent function for both hidden and output activations.

Implementation of Subnetwork B. We implement the convolution neural network as follows. We resize all house photos so that the longer dimension (vertical or horizontal) of every photo is 500 pixels. For each house photo, we slide a subwindow at a step of 20 pixels, and at each step, we crop subregions of three different resolutions: 240×240, 200×200, and 160×160 pixels. We use the CNN architecture proposed in [19] that is a mixture of convolution layers, pooling layers, fully connected layers and dropout layers. We set the convolution filter size to be squares of 11, 9, 5 respectively. For all the layers but the output layer, we used the rectified linear units (ReLU) as activation functions. Each densely connected layer has output size of 4096. Dropout layers follow each of the densely connected layers with a dropout ratio of 0.6.

Implementation of Subnetwork C. We use a fully connected layer to fuse the output activations of the above two networks, as shown in Fig. 2. The size of the fusion layer is 4126. It is connected with the output neuron used for the ranking loss Eq. (9). We use a FNN to warp the output activations of the fusion layer to a single house value. This network includes 5 hidden layers with 200, 100, 80, 60, and 30 neurons, respectively. Tangent function is used as the activation functions for hidden or output units.

Table 3 summarizes the configurations of each subnetwork used in this work.

Photo Categorization. We divide all house photos into 11 categories: bed room, living room, kitchen room, bath room, backyard, dining room, pool, garage, storage, neighborhood and others. We collect 200 photos for each of the 11 categories, densely crop sub-regions and use the cropped images to train a CNN network with classification loss [19]. To classify a photo, we perform a forward pass over the learned CNN network. The prediction will be used to prune the invalid constraints in Eq. (9). We utilize automatic categorization results to reduce manual efforts, for two reasons. First, the categorization accuracy is relatively high (91 percent on a small validation set). Second, a single constraint obtained from mis-classified photos has little effect over the final results considering that the total number of constraints is relatively huge (hundreds). Moreover, every single constraint is used as a soft constraint in this work and none of them will dominate the training results.



(a)

# zipcode	92
# images	901,347
# houses	30,141
min-price	$102K
max-price	$2,100K
# properties	48

(b)

Fig. 3. Statistics of our dataset. Left column: Histogram of house prices ($100,000 per bin). Right column: statistics of our house dataset.

5 Experiments

In this section, we apply the proposed MiDRR method over the multi-modal house price prediction problem and compare it to the alternative methods.

Dataset. To study the visual house appraisal problem, we collect an image dataset that includes textual features, house photos, and trading history, all of which are available in the online real estate market.[1] Our dataset includes about 900,000 photos of 30,141 houses. For each house, we use the mostly recent traded price as its true house value. Fig. 3 plots in the left column a histogram of house prices where the horizontal direction indicates the house prices (starting form $100,000). Each bin indicates a range of $100,000. In the right column we show several statistics of the dataset. We can observe that about one fourth houses are within the range of $400,000-$500,000. Table 1 illustrates parts of the textual house features. To our best knowledge, this is the first comprehensive image benchmark for studying the multi-modality visual house appraisal problem.

1. http://www.redfin.com

TABLE 4
Results of Regression

Methods	Photos	Ranking Loss $R()$	Regression Loss $Q()$	Semantic-awareness	Commonsense	Error (K\$)	Error (‰)
MiDRR-5A	Y	Y	Y	Y	Y	**4.3** ($\pm$0.34)	**5.1** ($\pm$0.23)
MiDRR-5B	Y	Y	Y	Y		5.2 ($\pm$0.46)	7.2 ($\pm$0.21)
MiDRR-5C	Y	Y	Y		Y	6.3 ($\pm$0.53)	8.7 ($\pm$0.32)
MiDRR-5D	Y	Y	Y			6.8 ($\pm$0.48)	9.2 ($\pm$0.43)
MiDRR-3A		Y	Y		Y	15.7 ($\pm$0.59)	15.2 ($\pm$0.98)
MiDRR-3B		Y	Y			16.8 ($\pm$0.67)	15.3 ($\pm$1.34)
MiDRR-2A			Y		Y	21.5 ($\pm$0.66)	21.1 ($\pm$1.21)
MiDRR-2B			Y			23.6 ($\pm$0.59)	23.4 ($\pm$0.75)
DF-A [7]	Y	-	-	-	-	21.8 ($\pm$0.87)	21.8 ($\pm$2.22)
DF-B [7]	-	-	-	-	-	23.9 ($\pm$0.89)	25.9 ($\pm$2.34)
NN-A [30]	Y	-	-	-	-	24.5 ($\pm$2.34)	20.9 ($\pm$3.47)
NN-B [30]	-	-	-	-	-	23.6 ($\pm$2.14)	21.2 ($\pm$4.15)
Boosting-A [33]	Y	-	-	-	-	23.6 ($\pm$3.34)	23.1 ($\pm$4.71)
Boosting-B [33]	-	-	-	-	-	24.9 ($\pm$4.53)	24.7 ($\pm$5.12)
SVR-A [2]	Y	-	-	-	-	31.6 ($\pm$5.82)	24.5 ($\pm$5.34)
SVR-B [2]	-	-	-	-	-	34.8 ($\pm$5.98)	24.0 ($\pm$4.79)
ResNet-153	Y	-	Y	-	-	9.2 ($\pm$1.12)	10.3 ($\pm$1.23)
RobustDR [4]	Y	-	Y	-	-	14.7 ($\pm$2.47)	18.5 ($\pm$4.01)
VGG-16	Y	-	Y	-	-	11.6 ($\pm$1.05)	12.3 ($\pm$0.98)

The letter'Y' in Columns 2-6 indicates that the respective algorithms use the components of house photos (Pht), ranking loss (R), regression loss (Q), Semantic-awareness (S), or common sense knowledge (C), respectively.

We split the dataset into three subsets, including 15,000, 5,000, and 10,141 houses, used for training, validation and testing, respectively. We manually annotate the training/validation images with 11 categories, and use them to train and evaluate the CNN classification model.

Variants of Our Approach. The proposed approach comprises of three sub-networks and two loss functions, and uses both textual features and house photos for predicting house values. In order to study the individual contributions of these components, we implement and evaluate a few variants of the proposed approach.

- *MiDRR-1A*, a ranking net that processes textual features with the subnetwork A. Since there is only one instance for every house, Eq. (10) degenerates to be a max-margin constraint between house pairs.
- *MiDRR-2A*, a regression net that uses the subnetwork A, subnetwork C and regression loss. The output activations of the subnetwork A are fully connected to the fusion layer. There is only one instance in every bag, and in Eq. (12), d(i) is fixed to be 1. Therefore, the constraints in Eq. (12) are not valid. The common sense constraints in Eq. (13) are still applicable.
- *MiDRR-3A*, a joint ranking and regression net that uses the subnetwork A and subnetwork C. It combines the MiDRR-1A and MiDRR-2A, providing a MiDRR net that only uses textual house features.
- *MiDRR-4A*, a ranking net that uses the subnetwork A and subnetwork B. Both textual and photos are used for learning a ranking function $f_w()$ in the multiple instance setting.
- *MiDRR-5A*, a joint ranking and regression net that uses all the three subnetworks and two objective functions, i.e., Eqs. (9) and (12).

Among the above algorithms, the first three only use textual features and thus only generate a single instance for every bag, for which the multiple instance setting degenerates to the conventional supervised setting. MiDRR-1A and MiDRR-4A are only used for ranking purpose, MiDRR-2A is only used for regression purpose, MiDRR-3A and MiDRR-5A can be used for both ranking and regression purposes.

We further introduce more variants of our method to analyze the effects of two components: photo categorization and common sense knowledge. For each of the algorithms MiDRR-2A, MiDRR-3A and MiDRR-5A that use the regression loss, we implement another variant that does not utilize common sense knowledge, denoted by *MiDRR-2B*, *MiDRR-3B* and *MiDRR-5B*, respectively. In these algorithms, the constraint collection $\mathbf{C}$ is set to be empty. For *MiDRR-4A* and *MiDRR-5A* that use both ranking loss and house photos, we implement two more algorithms without the photo categorization component, denoted by MiDRR-4C and MiDRR-5C, respectively. These algorithms will extract constraints between images that might be of different categories. For *MiDRR-5A*, we further implement a variant without the components of common sense knowledge nor photo categorization, denoted by *MiDRR-5D*. Tables 4 and 5 summarize the usage of individual components, including i) *Pht*., using house photos as inputs; ii) $R()$, using the ranking loss function (9); iii) $Q()$, using the regression loss function (12); iv) *S*, using Semantic awareness or photo categorization, i.e., (10); v) *C*, using common sense knowledge, i.e., (13).

Evaluation Metrics. We apply the above methods over our image dataset for two purposes: regression of house values and ranking of house pairs. For the regression purpose, we calculate the error between the prediction value and true value for every single house. We use two error units: i) the absolute difference between the prediction and true price, and ii) the error percentage, i.e., the absolute error over the

TABLE 5
Results of Ranking

Methods	Photos	*Ranking Loss*$R()$	*Regression Loss*$Q()$	Semantic-awareness	Commonsense	Ranking Accuracy (%)
MiDRR-5A	Y	Y	Y	Y	Y	89.1
MiDRR-5B	Y	Y	Y	Y		86.3
MiDRR-5C	Y	Y	Y		Y	87.2
MiDRR-5D	Y	Y	Y			83.8
MiDRR-4A	Y	Y		Y		77.6
MiDRR-4C	Y	Y				76.5
MiDRR-3A		Y	Y		Y	65.0
MiDRR-3B		Y	Y			63.1
MiDRR-1A		Y				57.1
MiDRR-1C		Y				54.9
DR-A [13]	Y	-	-	-	-	63.3
DR-B [13]	-	-	-	-	-	59.1
MIR-A [5]	Y	-	-	-	-	55.6
MIR-B [5]	-	-	-	-	-	54.7

In Columns 2 through 6, we use 'Y' to indicate whether the respective algorithms use the house photos (Pht), ranking loss (R), regression loss (Q), Semantic-awareness (S), or common sense knowledge (C), respectively.

true house price. We average these two metrics over all the testing samples and calculate their standard deviations as well. For the ranking purpose, we simply count the percentage of wrong predictions for the testing samples.

Baselines. We compare the proposed algorithms to other popular methods for regression or ranking purposes in the multi-instance setting. The baselines for *Regression* include: 1) *SVR*, Multiple-instance Support Vector Regression [2]. 2) Multiple-instance *Boosting* [33]; 3) *DF*, Multiple-instance Decision Forests [7] and 4) *NN*, Multiple-instance Neural Network [30]. The baselines for *Ranking* include: 5) *DR*, Deep Ranking algorithm [13]; and 6) *MIR*, Multiple Instance ranking [5].

We evaluate the above algorithms for two settings: *A*, that uses both textual features and photos in the multi-instance setting; *B*, that uses textual features only and the multiple instance settings degenerate to the supervised setting. Among these algorithms, *NN* and *DR* can directly process house photos. For other algorithms, we pass every photo through the CNN net trained for photo categorization and use the output activations of the last hidden layer as visual feature vector. This feature extraction has been widely used for image related tasks [13], [19]. We use the matlab implementation of the above baselines provided by the authors or other contributors. To obtain the optimal parameters for these baselines, we use a 10-fold cross-validation procedure over the validation subset.

We also implement three network-based regression methods for comparisons, which are trained with full supervisions. To get the instance-level labels, we simply assign the price of each house to its house photos. The first method *RobustDR* is proposed by Belagiannis et al. for learning robust deep regression models [4]. Taking house photos as inputs, the network of *RobustDR* consists of five convolution layers, followed by two fully connected layers. For each photo instance, we concatenate the textual features of the house to the first fully connected layer. In this way, the second fully connected layer serves as a fusion layer for jointly exploring textual and imagery inputs of houses. The second and third methods employ the widely used network structures, VGG network [29] and ResNets [17], respectively, both of which have achieved great performance on multiple image benchmarks. In particular, we use the 16-layer *VGG-16* and the 152-layer *ResNet-152*. Following the same strategy as *RobustDR*, we concatenate the textual houses features to the fully connected layers of these two networks, and optimize the least square regression loss. We follow the original articles to pre-train the above networks on ImageNet [20] and fine-tune the network parameters over the collected image dataset. It is noteworthy that the proposed MiDDR method can use alternative (deeper) network structures (e.g., VGG or ResNet) as the subnetwork B to take advantages of newly developed network structures.

Qualitative Results. Fig. 4 visualizes exemplar results of the MiDRR-5A algorithm. In each column we show two photos of the same house. For every house photo we only highlight the instance (in Red Box) that achieves the highest price prediction $g_w()$. The photos of the top row are of 'backyard', and that of the bottom row are of 'neighborhood'. The true house price is plotted on the top of each box. We can observe that price predictions of house instances are consistent with human perceptions. In particular, the houses in left-hand columns feature cozy furniture, well-maintained plants, nicely upgraded ceiling, or other patterns which lead to higher house appraisal than the houses in other columns. The proposed method is capable of modeling these visual patterns during training and making consistent and accurate price predictions.

Fig. 6 showed three houses for which the proposed method MiDDR can accurately predict their values whereas the baseline methods VGG-16 and MiDDR-3 had relatively large estimation errors. This is because that MiDDR-3 only uses textual features and VGG-16 does not use ranking losses. It is noteworthy that MiDDR-3 and MiDDR-5 preserve the relative orders of the three houses, i.e., $\mathrm{A} < \mathrm{B} < \mathrm{C}$, since they are both trained for joint ranking and regression purposes. In contrast, the VGG-16 method assigns the highest value to house B, and the second

Fig. 4. Exemplar results of visual house appraisal. The true house prices are shown on the top of each box. Every house photo is overlaid with the instance (in red) that achieves the highest price prediction $g_w()$. Top: House photos of backyard. Bottom: House photos of neighborhood.

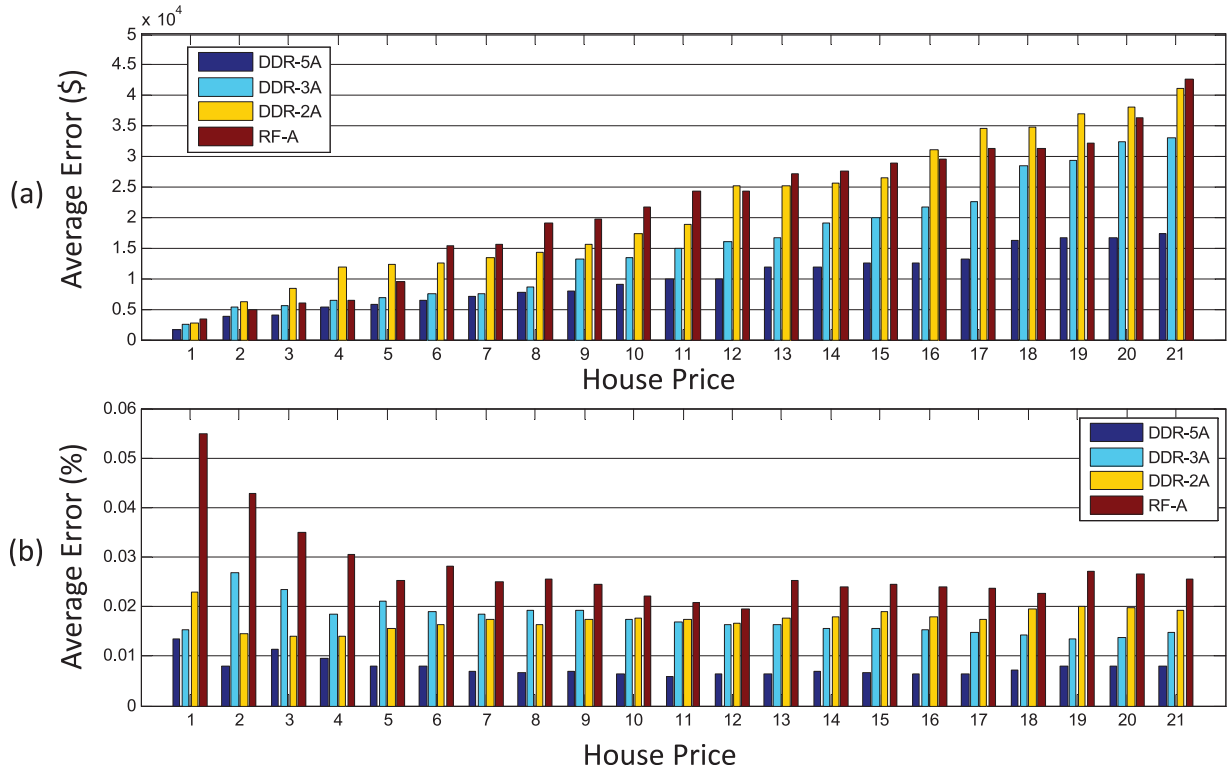


Fig. 5. Results of multi-modal house price prediction at different price ranges. We divide the price range (from 101K to 2200K) into 21 bins (x-direction). Vertical-direction: (a) Average errors in $ or (b) percentages of the house value, respectively.

highest value to house A. These comparisons directly demonstrate the advantages of solving ranking and regression losses in a coupled fashion.

Fig. 7 showed two houses for which the estimations of MiDDR-5 are not better than that of MiDDR-3 and VGG-16. The results of MiDDR-5 are however reasonable because the visual appearances of these two houses are not compliant with their respective true values. For example, the house D has a badly maintained backyard and an original living room, and is expected to have a lower price than the provided true value. Similarly, the house E is expected to have a higher price than its true value. Our method captured these appearance distinctions, and assigned a relatively low value to house D and a relatively high value to house E. These results demonstrate the capabilities of the proposed method and also verify that the use of house photos can improve system accuracies

Quantitative Results. Table 4 reports the average errors and standard deviations of various regression algorithms. Fig. 5 plots the average errors at different price ranges. Note that the absolute errors for low-priced houses are relatively lower than that for the high-priced houses, which are intuitively reasonable. Table 5 reports the error rates of pair-wise house ranking. From the above results, we can obtain the following observations. i)The proposed MiDRR net achieved the minimal errors in terms of both metrics. In particular, the average error is less than $5,000 which is encouraging in practice considering that the house prices vary between $100,000 and $2,000,000. ii) Exploiting house photos with different frameworks, including MiDRR net,



House A House B House C

	True Value (K\$)	VGG-16 (K\$)	MiDDR-3A (K\$)	MiDDR-5A (K\$)
House A	454	531	410	**448**
House B	749	681	671	**735**
House C	849	492	789	**851**

Fig. 6. Exemplar results of various house appraisal methods. Top: Photos of three houses A, B, and C. Bottom (table): True house values (column 2), and the house values estimated by VGG-16, MiDDR-3A and MiDDR-5A, respectively. Note that MiDDR-3A only uses textual house features.



House D



House E

	True Value (K\$)	VGG-16 (K\$)	MiDDR-3A (K\$)	MiDDR-5A (K\$)
House D	415	426	**410**	395
House E	399	413	**405**	431

Fig. 7. Failure examples of the proposed house appraisal methods. Top row: Three house photos of the house D. Middle row: Three house photos of the house E. Bottom (table): True house values (column 2), and the house values estimated by VGG-16, MiDDR-3A, and MiDDR-5A, respectively. The visual appearances of these two houses are not compliant with their respect true-values because, for example, the house A is expected to have a lower price than its true value due to its poor appearance. Similarly, the house B is expected to have a higher price than its true value.

RF, NN or boosting, can reduce the appraisal errors. Taking MiDRR for example, the average error by MiDRR-3A is \$15,700 which can be reduced to be \$4,300 if additionally accessing house photos. iii) Joint regression and ranking was proved to be effective, in particular for the proposed MiDRR net. This can be verified from the comparisons between MiDRR-5A and others, or the comparisons between MiDRR-3A and MiDRR-2A. iv) The proposed components, Semantic awareness (S) and commonsense knowledge (C), can further reduce system errors. Moreover, results showed that the proposed method can achieve much better accuracies than the most recent network-based methods for regression purposes. While our method can benefit from the advanced network structures as well, it is out of the research scope of this work to exhaustively test the combinations of the proposed framework and various network structures. In our experiment setting, MiDDR-5 utilizes house photos and house textual features as inputs for learning the regression and ranking network, whereas MiDDR-3 only utilizes textual house features. The large improvements over regression errors (from $K\$15.7$ to $K\$4.3$) are expected given that tens of thousands of imagery data can provide significant amount of information for learning deep representations. Moreover, the VGG net based method utilizes a fully supervised regression loss defined over house instances (i.e. photos) and bag-level labels, without using instance-level labels, and is expected not to be effective. In fact, the performance of $K\$11.6$ is not bad since this clearly outperformed the traditional multi-instance learning methods (e.g., NN-A, SVR-A). In contrast, MiDDR-5 employs a multi-instance formula and simultaneously trains the regression and ranking losses.

6 Conclusion

In this work, we studied a novel image task, i.e., visual house appraisal, and formulated it as a weakly supervised learning problem. Our efforts are two folds. On the one hand, we collect a comprehensive image benchmark for studying the visual house appraisal problem, which is the first one in its catalog. We implemented multiple regression methods including the proposed regression methods and

exhaustively evaluated them on the proposed dataset. The collected dataset along with the baseline methods will be released to foster research in this novel direction. On the other hand, we developed a multi-instance learning method for visual house appraisal, which can leverage both textual and imagery data to jointly train a deep representation for both ranking and regression purposes. Extensive experiments with comparisons showed that our approach can estimate house values with high-accuracy. Analysis of individual components clearly demonstrated the technical soundness of the proposed solution.

The developed techniques have applications over a wide variety of multi-dimensional regression problems in the multiple instance setting, e.g., predicting crime rate of an area from both textual descriptions and community photos [26], rating of fashion products (purse, clothes) from customer's reviews and photos [22], and diagnosing disease using phenotypic features and imagery scans. For all these applications, there are often more than one photo available for each sample, resulting in the multi-instance setting. Our method can effectively integrate features of multiple modalities and provide a unified framework for utilizing various commonsense knowledge or heuristics. We will study these novel research topics in the future works.

Acknowledgments

Xiaobai Liu is supported by the NSF grant (no.1657600), DARPA SIMPLEX program (no. 58723A) and ONR grant (No. N00014-17-1-2867). Jiebo Luo is supported in part by the New York State through the Goergen Institute for Data Science. Xiaobai Liu and Qian Xu contributed equally to this work.

References

[1] S. Alkire, "Subjective quantitative studies of human agency," *Social Indicators Res.*, vol. 74 no. 1, pp. 217–260, 2005.
[2] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. 15th Int. Conf. Neural Inf. Process. Syst.*, 2002, pp. 577–584.
[3] D. Basak, S. Pal, and D. Patranabis, "Support vector regression," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2007, pp. 203–224.
[4] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab, "Robust optimization for deep regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2830–2838.
[5] C. Bergeron, J. Zaretzki, C. Breneman, and K. Bennett, "Multiple instance ranking," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 48–55.
[6] S.-Y. Cho and Z. Chi, "Genetic evolution processing of data structures for image classification," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 2, pp. 216–231, Feb. 2005.
[7] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Found. Trends Comput. Graph. Vis.*, vol. 7, no. 2, pp. 81–227, 2012.
[8] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surveys*, vol. 40, no. 2, 2008, Art. no. 5.
[9] A. S. d'Avila Garcez and G. Zaverucha, "Multi-instance learning using recurrent neural networks," in *Proc. Int. Joint Conf. Neural Netw.*, 2012, pp. 1–6.
[10] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, no. 1, pp. 31–71, 1997.
[11] P. B. D. Kotzias, M. Denil, and N. de Freitas, "Deep multi-instance transfer learning," in *Proc. Neural Inf. Process. Syst. Deep Learn. Representation Workshop*, 2014, pp. 1–8.
[12] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proc. 32nd Int. Conf. Int. Conf. Mach. Learn.*, 2015, pp. 1180–1189.
[13] Y. Gong, Y. Jia, T. K. Leung, A. Toshev, and S. Ioffe, "Deep convolutional ranking for multilabel image annotation," in *Proc. Int. Conf. Learn. Representations*, 2014, pp. 1–8.
[14] M. Guillaumin, J. Verbeek, and C. Schmid, "Multimodal semi-supervised learning for image classification," in *Proc. 23rd IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 902–909.
[15] P. A. Gutiérrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, and C. Hervas-Martinez , "Ordinal regression methods: Survey and experimental study," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 1, pp. 127–146, Jan. 2016.
[16] T. H. Haveliwala, "Topic-sensitive PageRank: A context-sensitive ranking algorithm for web search," *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 784–796, Jul.–Aug. 2003.
[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[18] J. Joo, W. Li, F. F. Steen, and S.-C. Zhu, "Visual persuasion: Inferring communicative intents of images," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 216–223.
[19] A. Krizhevsky, I. Sutskever, M. M. G. E. Hinton, and Y. LeCun, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[20] 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.
[21] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. Int. Conf. Learn. Representations*, 2015, pp. 1–8.
[22] S. Liu, et al., "Hi, magic closet, tell me what to wear!" in *Proc. 20th ACM Int. Conf. Multimedia*, 2012 pp. 619–628.
[23] O. Mangasarian and E. Wild, "Generalized linear models," *J. Amer. Stat. Assoc.*, vol. 95, no. 452, pp. 1320–1324, 2000.
[24] O. Mangasarian and E. Wild, "Multiple instance classification via successive linear programming," *J. Optimization Theory Appl.*, vol. 137, pp. 555–568, 2008.
[25] J. J. Pan, J. T. Kwok, Q. Yang, and Y. Chen, "Multidimensional vector regression for accurate and low-cost location estimation in pervasive computing," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 9, pp. 1181–1193, Sep. 2006.
[26] L. Porzi, S. R. Bulò, B. Lepri, and E. Ricci, "Predicting and understanding urban perception with convolutional neural networks," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 139–148.
[27] J. Ramon and L. D. Raedt, "Multi instance neural networks," in *Proc. ICML Workshop Attribute-Value Relational Learn.*, 2000, pp. 53–60.
[28] D. Sculley, "Combined regression and ranking," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2010, pp. 979–988.
[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556, 2014.
[30] D. F. Specht, "A general regression neural network," *IEEE Trans. Neural Netw.*, vol. 2, no. 6, pp. 568–576, Nov. 1991.
[31] B.-Y. Sun, J. Li, D. D. Wu, X.-M. Zhang, and W.-B. Li, "Kernel discriminant learning for ordinal regression," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 6, pp. 906–910, Jun. 2010.
[32] S. Vijayanarasimhan and K. Grauman, "Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–8.
[33] P. A. Viola, J. C. Platt, and C. Zhang, "Multiple instance boosting for object detection," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2006, pp. 1417–1424.
[34] J. Wang, et al., "Learning fine-grained image similarity with deep ranking," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 1386–1393.
[35] J. Wu, Y. Wu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3460–3469.
[36] J. Wu, Y. Zhao, J.-Y. Zhu, S. Luo, and Z. Tu, "MILCut: A sweeping line multiple instance learning paradigm for interactive image segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, 2014, 256–263.
[37] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang, "Ranking with local regression and global alignment for cross media retrieval," in *Proc. 17th ACM Int. Conf. Multimedia*, 2009, pp. 175–184.
[38] K.-H. Yap, T. Chen, Z. Li, and K. Wu, "A comparative study of mobile-based landmark recognition techniques," *IEEE Intell. Syst.*, vol. 25, no. 1, pp. 48–57, Jan./Feb. 2010.

[39] X.-T. Yuan, X. Liu, and S. Yan, "Visual classification with multitask joint sparse representation," *IEEE Trans. Image Process.*, vol. 21, no. 10, pp. 4349–4360, Oct. 2012.
[40] Z.-H. Zhou and M.-L. Zhang, "Neural networks for multiinstance learning," in *Proc. Int. Conf. Intell. Inf. Technol.*, 2002, pp. 455–459.
[41] W. Zhu, Q. Lou, Y. S. Vang, and X. Xie, "Deep multi-instance networks with sparse label assignment for whole mammogram classification," arXiv:1612.05968, 2016.

Xiaobai Liu received the PhD degree from the Hua zhong University of Science and Technology, China. He is currently an assistant professor of compuer science with the San Diego State University (SDSU), San Diego. His research interests focus on scene parsing with a variety of topics, e.g., joint inference for recognition and reconstruction, commonsense reasonzing, etc. He has published 30+ peer-reviewed articles in top-tier conferences (e.g., ICCV, CVPR, etc.) and leading journals (e.g., the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *IEEE Transactions on Image Processing*, etc.) He received a number of awards for his academic contribution, including the 2013 Outstanding Thesis Award by CCF(China Computer Federation). He is a member of the IEEE.

Qian Xu received the BS degree from the School of Science, Beihang University, Beijing, China, in 2006, and the master's and the doctoral degrees from the Department of Statistics, San Diego State University, in 2011 and 2017, respectively. She is the co-founder and president of XreLab Inc., San Diego, California. Her research interest falls in the various statistical models and their applications in computer vision.

Jingjie Yang received the master's degree from the Department of Computer Science, San Diego State University (SDSU). Her research interests fall in the areas of social media image analysis and understanding.

Jacob Thalman received the master's degree from the Department of Computer Science, San Diego State University (SDSU). His research interests fall in the area of social media image analysis and understanding. He is now an analytics analyst I at the Pawnee Leasing Corporation.

Shuicheng Yan is chief scientist at the Qihoo/360 company, and also the dean's chair associate professor with the National University of Singapore. His research areas include machine learning, computer vision, and multimedia, and he has authored/co-authored hundreds of technical papers over a wide range of research topics, with Google Scholar citation more than 20,000 times and H-index 66. He is an ISI Highly-cited researcher of 2014, 2015, and 2016, respectively. His team won seven times or received or honorable-mention prizes in PASCAL VOC and ILSVRC competitions, along with more than 10 times best (student) paper prizes. He is also IAPR fellow.

Jibo Luo joined the University of Rochester, in Fall 2011 after more than 15 prolific years at the Kodak Research Laboratories, where he was a senior principal scientist leading research and advanced development. He has been involved in numerous technical conferences, including serving as the program co-chair of ACM Multimedia 2010, IEEE CVPR 2012, and IEEE ICIP 2017. He has served on the editorial boards of the *IEEE Transactions on Pattern Analysis and Machine Intelligence*, the *IEEE Transactions on Multimedia*, the *IEEE Transactions on Circuits and Systems for Video Technology*, the *ACM Transactions on Intelligent Systems and Technology*, the *Pattern Recognition*, the *Machine Vision and Applications*, and the *Journal of Electronic Imaging*. He is a fellow of the SPIE, IEEE, and IAPR.