A Multi-Scale CNN for Affordance Segmentation in RGB Images

Anirban Roy and Sinisa Todorovic

School of Electrical Engineering and Computer Science
Oregon State University
royani@oregonstate.edu, sinisa@eecs.oregonstate.edu

Abstract. Given a single RGB image our goal is to label every pixel with an affordance type. By affordance, we mean an object's capability to readily support a certain human action, without requiring precursor actions. We focus on segmenting the following five affordance types in indoor scenes: 'walkable', 'sittable', 'lyable', 'reachable', and 'movable'. Our approach uses a deep architecture, consisting of a number of multi-scale convolutional neural networks, for extracting mid-level visual cues and combining them toward affordance segmentation. The mid-level cues include depth map, surface normals, and segmentation of four types of surfaces – namely, floor, structure, furniture and props. For evaluation, we augmented the NYUv2 dataset with new ground-truth annotations of the five affordance types. We are not aware of prior work which starts from pixels, infers mid-level cues, and combines them in a feed-forward fashion for predicting dense affordance maps of a single RGB image.

Keywords: Object affordance, Mid-level cues, Deep learning

1 Introduction

This paper addresses the problem of affordance segmentation in an image, where the goal is to label every pixel with an affordance type. By affordance, we mean an object's capability to support a certain human action [1, 2]. For example, when a surface in the scene affords the opportunity for a person to walk, sit or lie down on it, we say that the surface is characterized by affordance types 'walkable', 'sittable', or 'lyable'. Also, an object may be 'reachable' when someone standing on the floor can readily grasp the object. A surface or an object may be characterized by a number of affordance types. Importantly, affordance of an object exhibits only the possibility of some action, subject to the object's relationships with the environment, and thus is not an inherent (permanent) object's attribute. Thus, sometimes chairs are not 'sittable' and floors are not 'walkable' if other objects in the environment prevent performing the corresponding actions.

Affordance segmentation is an important, long-standing problem with a range of applications, including robot navigation, path planning, and autonomous driving [3–14]. Reasoning about affordances has been shown to facilitate object and action recognition [4, 10, 13]. Existing work typically leverages mid-level visual

cues [3] for reasoning about spatial (and temporal) relationships among objects in the scene, which is then used for detection (and in some cases segmentation) of affordances in the image (or video). For example, Hoiem et. al. [15,16] show that inferring mid-level cues – including: depth map, semantic cues, and occlusion maps – facilitates reasoning about the 3D geometry of a scene, which in turn helps affordance segmentation. This and other related work typically use a holistic framework aimed at "closing the loop" that iteratively improves affordance segmentation and estimation of mid-level cues, e.g., via energy minimization.

Motivated by prior work, our approach to affordance segmentation is grounded on estimation of mid-level cues, including depth map, surface normals and coarse-level semantic segmentation (e.g., general categories of surfaces such as walls, floors, furniture, props), as illustrated in Fig. 1. Our key difference from prior work is that, instead of "closing the loop", we use a *feed-forward* multi-scale convolutional neural network (CNN) in order to predict and integrate the mid-level cues for labeling pixels with affordance types. CNNs have been successfully used for low-level segmentation tasks [17–23]. Multi-scale CNNs have been demonstrated as suitable for computing hierarchical features, and successful in a range of pixel-level prediction tasks [17–19, 23, 24].

Given an RGB image, we independently infer its depth map, surface normals, and coarse-level semantic segmentation using the multi-scale CNN of Eigen et. al. [24]. The three multi-scale CNNs produce corresponding mid-level cues at the output, which are then jointly feed as inputs to another multi-scale CNN for predicting N affordance maps for each of N affordance types. Our estimate of depth map, surface normals, and semantic segmentation can be explicitly analyzed for reasoning about important geometric properties of the scene – such as, e.g., identifying major surfaces, surface orientations, spatial extents of objects, object heights above the ground, etc. We treat the three mid-level cues as *latent* scene representations which are fused by the CNN for affordance segmentation. Therefore, in this paper, we do not evaluate inference of the mid-level cues.

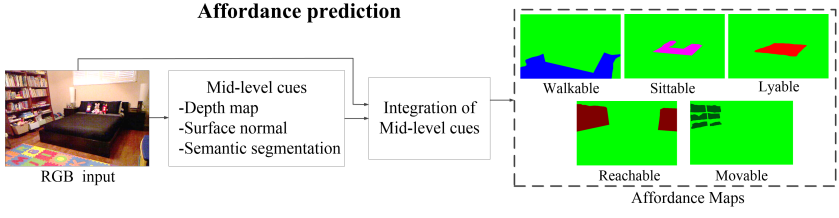

Fig. 1. An overview of our approach: Given an RGB image, we use a multi-scale convolutional neural network (CNN) to compute mid-level cues – including: depth map, surface normals and segmentation of general surface categories (e.g., walls, floors, furniture, props). The CNN also fuses these mid-level cues in a feed-forward manner for predicting five affordance maps for each of the five affordance types considered: 'walkable', 'sittable', 'lyable', 'reachable', and 'movable'.

In this paper, we focus on indoor scenes and typical affordances characterizing objects and surfaces in such scenes. Indoor scenes represent a challenging domain, because of relatively large variations in spatial layouts of objects affecting the feasibility of human-object interactions, and thus affordances. We consider the following five affordance types typical of indoor scenes:

1. **Walkable:** is any horizontal surface at a similar height as the ground that has free space vertically above (i.e., not occupied by any objects), since such a surface would afford a person to comfortably walk on it (even if soft);
2. **Sittable:** is any horizontal surface below a certain height from the ground (estimated relative to the human height) that has free space around and vertically above, as it would afford a person to comfortably sit on it;
3. **Lyable:** is any 'sittable' surface that is also sufficiently long and wide for a person to lie on it;
4. **Reachable:** can be any part of the scene that is within a reachable height for a person standing on the ground, and has free space around so that a person can stand next to it and readily grasp it;
5. **Movable:** is any 'reachable' small object (e.g., book) that can be easily moved by hand, and has free space around so as to afford the moving action.

In our specification, we consider that any 'walkable' surface is also 'standable'; therefore, 'standable' is not included in the above list. Also, we consider that the sitting action can be performed without a back support, which might be different from previous definitions in the literature. Note that almost everything under a certain height can be reachable if a person is allowed to bend, crawl, climb or perform other complex actions. In this paper, we only consider reachability by hand while a person is standing on the floor. Regarding 'movable', our definition may be too restrictive for a case when a relatively large object can be moved (e.g., chair); but, in such cases, the moving action cannot be easily performed.

It is also worth noting that we focus on "immediate" affordances, i.e., an object's capability to *immediately* support a human action, which can be readily executed without any precursor actions. For example, a chair is not immediately 'sittable' if it has to be moved before sitting on it. Therefore, we *cannot* resort to a deterministic mapping between object classes and their usual affordance types (chairs are in general sittable), since affordance types of particular object instances depend on the spatial context.

An obstacle that we have encountered in our work is the lack of datasets with ground truth pixel-wise annotations of affordances. Our literature review finds that most prior work focuses on affordance prediction at the image level, where the goal is to assign an affordance label to the entire image [4, 5, 9, 11, 25–27]. A few exceptions [28, 29] seek to discover similar affordance types as ours in RGB images. They estimate ground truth by hallucinating human skeletons in various postures amidst the inferred 3D layout of the scene.

As human skeletons may provide a limited model for reasoning about certain human-object interactions in the scene, and may not be informative for some of our affordance types (e.g., 'movable'), we have developed a new semi-automated method for generating pixel-wise ground truth annotations of affordances. This

is used to extend the NYU v2 dataset [30] with ground-truth dense affordance annotations, and our quantitative evaluation.

Contributions:

- We extend the NYUv2 dataset [30] with pixel-wise affordance ground truth.
- A new multi-scale deep architecture for extracting and fusing mid-level cues toward predicting dense affordance maps from an RGB image. Note that, unlike previous approaches [9, 31–34], we do not rely on any additional cues based on human-object interaction (e.g., action, pose).

In the following, Sec. 2 reviews prior work, Sec. 3 explains our method for generating affordance ground truth, Sec. 4 specifies our deep architecture for affordance segmentation, Sec. 5 describes how to train our deep architecture, and Sec. 6 presents our experimental results.

2 Prior Work

Predicting affordances has a long history in computer vision [1,2]. Early work has typically considered a rule-based inference for affordance segmentation [35–37]. However, their hand-designed rules are too brittle for real-world indoor scenes abounding with clutter and occlusions.

Some recent approaches reason about affordance via interpreting human actions and human-objects interactions [9,31–34]. For example, recognizing human actions can provide informative cues for predicting affordance [31,32]. Other approaches leverage a fine-grained human pose estimation [33]. These visual cues are also used for predicting affordance of novel objects [9]. One of our key differences from these approaches is that they are aimed at predicting affordance of foreground objects, whereas we aim for a dense pixel-wise labeling.

A related line of work predicts affordance by hypothesizing possible human-object interactions in the scene [28, 38, 39]. For example, [28, 39] use human-skeleton models in various postures. Our approach does not use human skeletons.

Another group of approaches [25–27, 40, 41] focus on affordances of small objects, such as spoon, knife, cup, etc., which are operated by hands. Thus, they address different affordance types from ours, including graspable, cuttable, liftable, fillable, scoopable, etc. In contrast, we consider affordance for human actions that involve the complete human body.

RGB and RGBD videos provide additional temporal cues for interpreting human-object interactions, and thus allow for robust affordance prediction [4, 7, 42, 43]. Also, detecting objects and reconstructing a detailed 3D scene geometry can lead to robust affordance segmentation [25, 28, 42, 44].

We are not aware of prior work which infers and combines mid-level cues in a feed-forward fashion using a deep architecture for predicting dense affordance maps of a single RGB image.

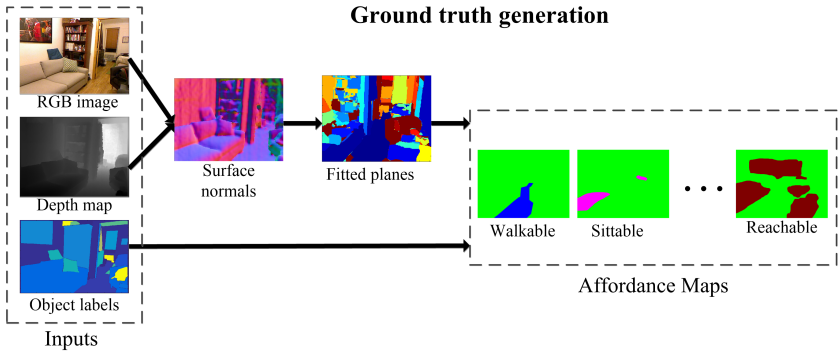

Fig. 2. For generating ground truth, we assume access to RGBD images. First, we compute surface normals from the RGB and depth information. Then, we use the RANSAC algorithm to fit 3D scene surfaces to a piece-wise planar approximation of the scene. The identified surface planes and their plane normals are combined with ground-truth object labels to decide affordance types present at every pixel.

3 Generation of Affordance Ground Truth

This section explains our semi-automated method for generating dense ground truth affordance maps in the NYUv2 dataset [30]. Importantly, for estimating such ground truth, we assume access to RGBD images and their pixel-wise annotations of object class labels. This is in contrast to our setting, where we have no access to depth information and object class labels, i.e., our approach takes only RGB images as input.

The NYUv2 dataset consists of 1449 indoor images with pixel-wise depth maps and object class labels for each image. There are 40 indoor object classes [45], including floor, wall, chair, sofa, table, bed, desk, books, bottle etc. Most of the scenes exhibit complex layouts of objects, clutter, and prominent occlusion. This makes affordance segmentation challenging.

Object class labels vs. affordance labels: Assigning affordance labels to pixels cannot be done using a direct mapping from available object class labels. This is because of two reasons. Different object parts may not support the same affordance (e.g., back-rest of a chair may not be sittable). Also, affordance of a particular object instance depends on the spatial context (e.g., a chair is placed under a table is not immediately sittable by our definition).

It follows that, in addition to object class labels, we also need to consider the spatial layout of objects in the scene for generating a reliable ground truth affordance maps. Thus, we develop an approach to systematically extract some essential geometrical cues from the scene, as explained below.

Understanding 3D scene geometry: We first align the RGB color and depth data, such that the floor represents the X-Y plane and Z axis represents height. From the RGB and depth map, we compute surface normals at every pixel. Then, we use the RANSAC algorithm to fit 3D scene surfaces to a piece-

wise planar approximation of the scene. This allows us to identify vertical and horizontal surface planes relative to the ground plane, as in [30]. For robustness, we allow some margin, such that we also account for near-horizontal and near-vertical surfaces (± 10 degrees of the surface normal). Finally, for each horizontal and vertical surface plane, we compute its height and maximum height from the ground plane, respectively. Also, for every surface plane, we estimate its size, and if there is a free space around and vertically above. Surrounding clearance is considered at a distance of 1 foot from the surface plane, where distances in the 3D scene are estimated using the camera parameters and the depth data.

Combining scene geometry and ground-truth object labels: Given the aforementioned estimates of horizontal and vertical surface planes in the scene, we identify their ground-truth object class labels. This has two purposes: (a) to constrain the set of candidate affordance types that could be associated with each plane, and (b) to enforce smoothness in our generation of affordance ground truth. To this end, for each affordance type, we specify a list of object classes appearing in the NYUv2 dataset that could be characterized by that type. For example, objects that could be 'sittable' are {chair, bed, sofa, desk, table, ... }; objects that could be 'walkable' are {floor, floor-mat}; objects that could be 'lyable' are {bed, sofa, table, ... }. The detailed list of NYUv2 objects and affordances they could support is provided in the supplemental material.

After determining the object class labels of the surface planes, the above-mentioned manually specified affordance-object pairs are used for hypothesizing candidate affordances of each plane. The candidates are further constrained per affordance definitions, stated in Sec. 1 and specified in Table 1, taking into account the plane's size, height, and surrounding and vertical clearances. Thus, when the plane's size, height or clearance does not satisfy the definition of a particular candidate affordance, this candidate is removed from the solution. For example, a horizontal plane, estimated at 3 feet from the ground and with vertical clearance, whose majority ground-truth class is 'bed', could be 'sittable' and 'lyable'. But if the plane's size has been estimated as too small to comfortably accommodate a full human body, the plane is labeled only as 'sittable'.

Affordance type	Definition				
	surface type	height(h)	size(s)	clearance above	clearance side
Walkable	horizontal	$h \leq 1/0.3$	$s \geq 2.5/0.23$	Yes	No
Sittable	horizontal	$1.5/0.45 \leq h \leq 3.5/1$	$s \geq 1/0.1$	Yes	Yes
Lyable	horizontal	$1.5/0.45 \leq h \leq 3.5/1$	$s \geq 10/0.9$	Yes	Yes
Reachable	horizontal/vertical	$1.5/0.45 \leq h \leq 7/2.1$	N/A	No	Yes
Movable	horizontal/vertical	$1.5/0.45 \leq h \leq 7/2.1$	$s \leq 2.5/0.23$	Either of two	

Table 1. Definitions of affordance types for surfaces identified in the scene. The heights are given in feet/meters, and sizes are given in feet^2/meters^2. We consider the maximum convex area of a surface to estimate its size. For all measurements, we allow $\pm 10\%$ tolerance to ensure robustness.

Note that our approach to generating ground truth differs from that presented in [28, 29]. Their approach hallucinates a human skeleton model in the

scene to determine the ground-truth affordance labels. Specifically, they convolve a human skeleton corresponding to a particular human action with the 3D voxelized representation of the scene. Such an approach would not generate ground truth which respects our affordance definitions. For example, a skeleton representing a standing person can fit on top of a desk or a table, and as a result these surfaces would be labeled as 'walkable' or 'standable' [28]. However, our definition of 'walkable' is based on the expectation that walking on horizontal surfaces with non-zero heights from the ground cannot be readily performed (one needs to climb first). Also, a skeleton representing a sitting person can easily fit on a chair even if there are small objects on the chair preventing a comfortable sitting action. Unlike [28, 29], we explicitly consider all requirements of the affordance definitions in order to generate affordance ground truth.

Manual Correction: The aforementioned automated generation of ground truth is prone to error. This is due to: (a) the stochastic nature of the RANSAC algorithm, (b) challenging elongated and thin surfaces that we fail to separate from the background, and (c) prominent occlusions and clutter that make our estimation of surface normals unreliable. Therefore, we have resorted to visual inspection of our results for corrections. We have used the Amazon mechanical turk to acquire multiple human corrections, and then applied majority voting to determine the final ground truth. Each user has been allowed to either add new regions to an affordance type, or remove wrongly labeled regions. The human corrections have been in relatively small disagreement, considering that our five affordance types are relatively complex cognitive concepts. Hence, the majority vote has helped resolve most disagreements.

Dataset statistics: About 72% of the automatically generated affordance labelings have been corrected by human experts. Each manual correction takes about 30-40 seconds per affordance class. We compute the intersection over union (IoU) measure between the automatically generated and manually corrected ground truth to compute their similarity. The IoU value is 67%, which indicates that the manual correction is necessary. Affordance types 'walkable', 'sittable', 'lyable', 'reachable', and 'movable' appear in 83%, 60%, 22%, 100%, and 93% of the NYUv2 images, respectively. A similar pixel-level statistic estimates that 12%, 5%, 11%, 65% and 11% of pixels are occupied by the corresponding affordance, if that affordance is present in the image. 18% of pixels have multiple distinct affordance labels. We notice that pixels occupied by a particular object instance are not all labeled with the same affordance, as desired. Thus, for example, only 79% of pixels occupied by floors are labeled as 'walkable'.

Some example ground truths are shown in Fig. 3. Additional examples are presented in the supplemental material.

4 Affordance Segmentation with a Multi-Scale CNN

We use four multi-scale CNNs for affordance segmentation, as illustrated in Fig 4. Each of these CNNs has the same architecture (e.g., number of convolutional layers, number and size of convolutional kernels) as the deep network presented

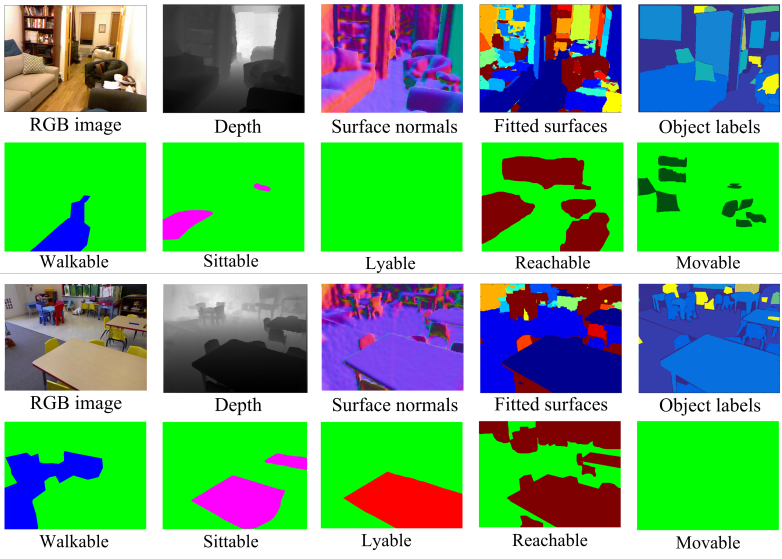

Fig. 3. Examples of our ground truth affordance maps: Top row represents the mid-level cues, i.e., depth map, surface normals and semantic segmentation. The bottom row represents the ground truth affordance maps.

in [24]. The three CNNs are aimed at extracting hierarchical features from the RGB image, via the coarse- and fine-scale networks, for estimating depth map, surface normals, and semantic segmentation, respectively. The coarse-scale CNN is designed to generate feature maps representing global visual cues in the image (e.g., large surfaces, context). The fine-scale CNN is designed to capture detailed visual cues, such as such as small objects, edges and object boundaries. As in [24], the outputs of the coarse-scale network is considered as inputs to the fine-scale network. The fourth CNN also consists of the coarse- and fine-scale networks, which serve for a multi-scale integration of the estimated mid-level cues and pixels of the input image for predicting the five affordance maps.

For semantic segmentation, we consider four high-level categories, including: 'floor', 'structure' (e.g., walls), 'furniture', and 'props' (small objects), defined in [30]. Note that this is in contrast to our method for generating ground truth, where we use as input ground-truth annotations of all fine-grained object classes (40 classes [45]). The four high-level categories allow for robust deep learning, since they provide significantly more training examples than there are instances for each fine-grained object class. Alternatively, we could have tried to conduct semantic segmentation of fine-grained object classes, and used the resulting segmentation for predicting affordance maps. However, such semantic scene labeling would limit the generalizability of our approach to scenes with novel objects.

Coarse-scale network: It takes pixels of the entire image as input, and generates feature maps as output. In this network, we use larger convolutional kernels with higher stride length than those used in the fine-scale network. As

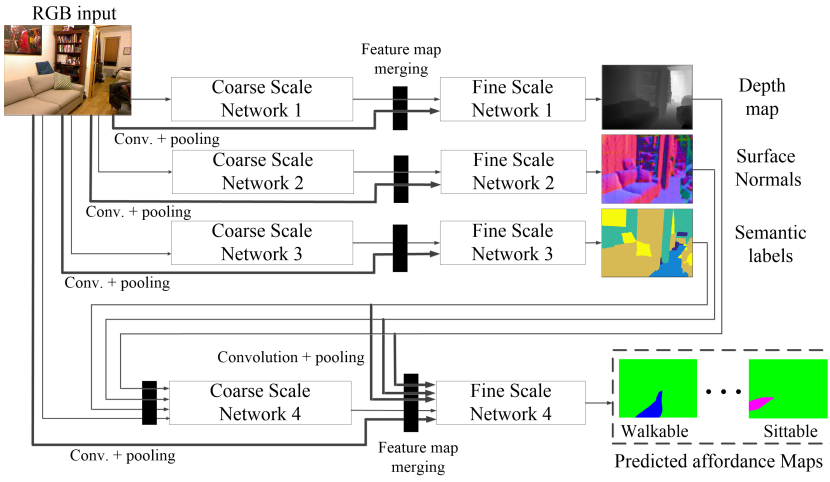

Fig. 4. Our multi-scale CNN architecture for predicting the affordance labels in the indoor scenes. The coarse scale network captures the global features such as context and the fine scale CNN captures finer details such as object boundaries. We combine the mid-level cues with the low-level RGB image to predict the affordance labels. Thin lines represent direct input and bold lines represent a convolution+pooling step, performed on the input before merging with the feature maps [24].

the deep architecture of [46], our coarse-scale network replaces the top fully connected layers by 1×1 convolution layers. The output is then upsampled in order to generate the final feature map. After every convolution+pooling step, the size of the output is reduced, such that the size of final output is 1/16 of the input. This final output is then upsampled to size 1/4 of the input.

Fine-scale network: The final output feature maps of the coarse-scale network and pixels of the image are jointly input to the fine-scale network for making the final prediction of the corresponding mid-level cue. In order to match the size of the RGB image with the size of the feature maps produced by the coarse-scale network, we perform a single step of convolution+pooling of the RGB image, as in [24]. For preserving fine details at the output, the convolution+pooling steps in this network do not reduce the output size. The final output of this network is upsampled to the size of the input RGB image, resulting either one of the mid-level cues or the affordance maps.

Details about both the coarse-scale and fine-scale network such as number of layers, kernel sizes, etc. are provided in the supplemental material. Note that, instead of using the three distinct scales as in [24], we consider only first two scales in our approach for efficiency.

5 Training

Our training of the deep architecture presented in Sec. 4 consists of four tasks aimed at training the four multi-scale CNNs. In each training task, the four coarse-scale networks are initialized with the VGG Net [47]. The fine-scale networks are initialized randomly. After initialization, the coarse-scale and fine-scale networks of each multi-scale CNN are trained jointly so as to minimize a suitable loss function, using the standard sub-gradient method with momentum. The momentum value is set to 0.9. Learning iterations are 2M, 1.5M, 1.5M and 2M for depth prediction, surface normal estimation, semantic segmentation and affordance segmentation respectively. Training takes 6-8 hours/per task and inference takes ≈0.15 sec/per image, on a Nvidia Tesla K80 GPU. In the following, we specify the loss functions used for training.

Multi-scale CNN-1 is trained for depth map prediction. We use a scale-invariant and structure-aware loss function for depth prediction as in [48]. Let d denote a difference between a predicted depth and ground truth in the log scale. Then, this loss function is defined as

$$L_{depth} = \frac{1}{I}\sum_i d_i^2 - \frac{1}{2I^2}\left(\sum_i d_i\right)^2 + \frac{1}{I}\sum_i [(\bigtriangledown_x d_i)^2 + (\bigtriangledown_y d_i)^2] \qquad (1)$$

where i is the index of pixels, I is the total number of pixels, and $\bigtriangledown_x d_i$ and $\bigtriangledown_y d_i$ denote the gradients of the depth difference d along the x and y axes.

Multi-scale CNN-2 is trained for predicting surface normals. The loss function for normals prediction is specified as $L_{norm} = -\frac{1}{I}\sum_i \boldsymbol{n}_i \cdot \hat{\boldsymbol{n}}_i$, where $\boldsymbol{n}_i$ and $\hat{\boldsymbol{n}}_i$ denote the ground truth and predicted surface normals at pixel i, and the symbol '$\cdot$' denotes the scalar product.

Multi-scale CNN-3 and multi-scale CNN-4 are trained for four-class semantic segmentation and predicting five binary affordance maps, respectively. For training both networks, we use the standard cross-entropy loss given the ground-truths of the four semantic categories, and our ground-truth affordance maps.

Data Augmentation: The NYUv2 dataset provides only 795 training images. The size of this training set is not sufficient to robustly train the multi-scale deep architecture. Therefore, we augment the training data by applying random translations, mirror flips, small rotations and contrast modifications. We also apply the same transformations to the corresponding ground truth maps. This results in a three times larger training set. Such data manipulation methods for increasing the training set are common [46, 48, 48].

6 Results

We first explain our experimental setup and then report our results.

Dataset. For evaluation, we use the NYUv2 dataset [30] which consists of 1449 RGBD images of indoor scenes with densely labeled object classes. Though the depth information is available for each image, we predict affordance only

using RGB input. Following the standard split, we use 795 images for training and 654 images for testing. We augment the dataset with the additional five, ground-truth, binary, dense affordance maps of: 'sittable', 'walkable', 'lyable', 'reachable' and 'movable'.

Benchmark datasets for evaluating scene geometry and scene layout, such as the UIUC indoor dataset [49] and geometric context [15], are not suitable for our evaluation, because they do not provide dense ground-truth annotations of object classes and surface normals. This, in turn, prevents us to generate ground truth affordance maps for these datasets. The RGBD video datasets [4, 43] are also not suitable for our evaluation, since our goal is to segment affordance in a single RGB image. Moreover, we focus on five affordance types that are different from those annotated in the RGBD videos of [4, 43] – specifically, our affordances are defined in relation to the entire human body, whereas the RGBD videos show affordances of small objects manipulated by hands. Also, note that a direct comparison with the related approaches that densely predict similar (but not the same) affordance types in indoor scenes [28, 29] would not be possible. Their affordance labels are heuristically estimated, and their ground truth is not yet publicly available. Although we are not in a position to conduct direct comparison with the state of the art, in the following, we present multiple baselines and compare with them.

Evaluation Metric. For quantitative evaluation, we compare the binary ground-truth affordance map with our predicted binary map. Specifically, we compute the ratio of intersection over union (IOU) between pixel areas with value 1, i.e., where affordance is present in the binary map. This is a common metric used in semantic segmentation [24, 46]. Note that this metric is stricter than the pixel wise precision measure or classification accuracy as used in [7, 16, 29].

Baselines. The following baselines are used for our ablation studies, and thus systematically evaluate each component of our approach.

Without predicting depth map (w/o Depth): In this baseline, we do not estimate depth maps, and do not use them for affordance segmentation. As shown in Tab. 2, ignoring depth cues significantly affects performance. This indicates that depth prediction is crucial for affordance segmentation as it helps reason about the 3D geometry of a scene.

Without predicting surface normals (w/o Surf Norm): In this baseline, we ignore surface normals while predicting affordance. Surface normals help estimate a surface's orientation (i.e., horizontal or vertical), and in turn inform affordance segmentation (e.g., a 'walkable' surface must be horizontal). As shown in Tab. 2, ignoring surface normals in this baseline leads to poor performance.

Without predicting semantic labels (w/o Sem): In this baseline, we ignore semantic labels for affordance segmentation. Tab. 2 shows that this baseline gives relatively poor performance, as semantic cues could help constrain ambiguities in affordance reasoning (e.g., floor is likely to be 'walkable').

Without predicting mid-level cues (w/o Mid-level): In this baseline, we ignore all three mid-level cues, i.e., affordance is predicted directly from pixels of the RGB image. Tab. 2 shows that the performance of this baseline is poor.

This suggests that affordance maps cannot be reliably estimated directly from pixels, and that inference of our mid-level cues is critical.

With ground truth cues (w GT): In this baseline, we directly use the ground truth depth maps, surface normals [30] and the semantic labels instead of predicting them from the image. This baseline amounts to an oracle prediction with correct mid-level cues for predicting affordance labels. Results of this baseline are shown in Tab. 2.

	walkable	sittable	lyable	reachable	movable	avg.
w/o Depth	63.23	31.44	36.10	57.24	43.84	46.37
w/o Surf Norm	64.36	32.32	37.77	57.70	44.64	47.36
w/o Sem	62.24	32.42	37.84	58.28	41.70	46.50
w/o Mid-level	58.45	24.63	31.20	50.54	34.20	39.80
w GT	70.43	37.61	43.33	63.41	51.37	53.23
Our approach	66.74	34.44	40.18	60.01	46.42	49.56

Table 2. Every baseline lacks one or more components of our approach and compered in terms of pixel wise IOU accuracy measure on the NYUv2 affordance dataset.

Evaluation of the Network Architecture. In this section, we empirically demonstrate the importance of multi-scale CNN architecture for affordance segmentation. Tab. 3 presents our results when using only coarse- or fine-scale network at a time, which amounts to considering features from a single scale – namely, either global visual cues or fine visual details. Tab. 3 shows that we get better performance when using only a coarse-scale network.

	walkable	sittable	lyable	reachable	movable	avg.
Coarse scale only	62.41	29.01	35.43	55.54	40.25	44.53
Fine scale only	64.67	31.58	37.74	57.86	43.67	47.10
Our approach (both)	66.74	34.44	40.18	60.01	46.42	49.56

Table 3. Comparisons of the approaches with varying the network architechture in terms of pixel wise IOU accuracy measure on the NYUv2 affordance dataset.

Qualitative Results. Fig. 5 illustrates some of our results. As can be seen, some affordance classes may not be present in an image, and a pixel might be assigned multiple affordance labels. Pixels which are not assigned any affordance labels are considered as background.

Failure case. Fig. 6 shows a failure case, where some parts of the floor – under the table – are predicted as 'walkable'. Here, we fail to identify that the table is vertically above the floor, preventing the walking action. In this case, the presence of object clutter and partial occlusion cause our incorrect estimation of the 3D geometry, and consequently the wrong affordance map estimation.

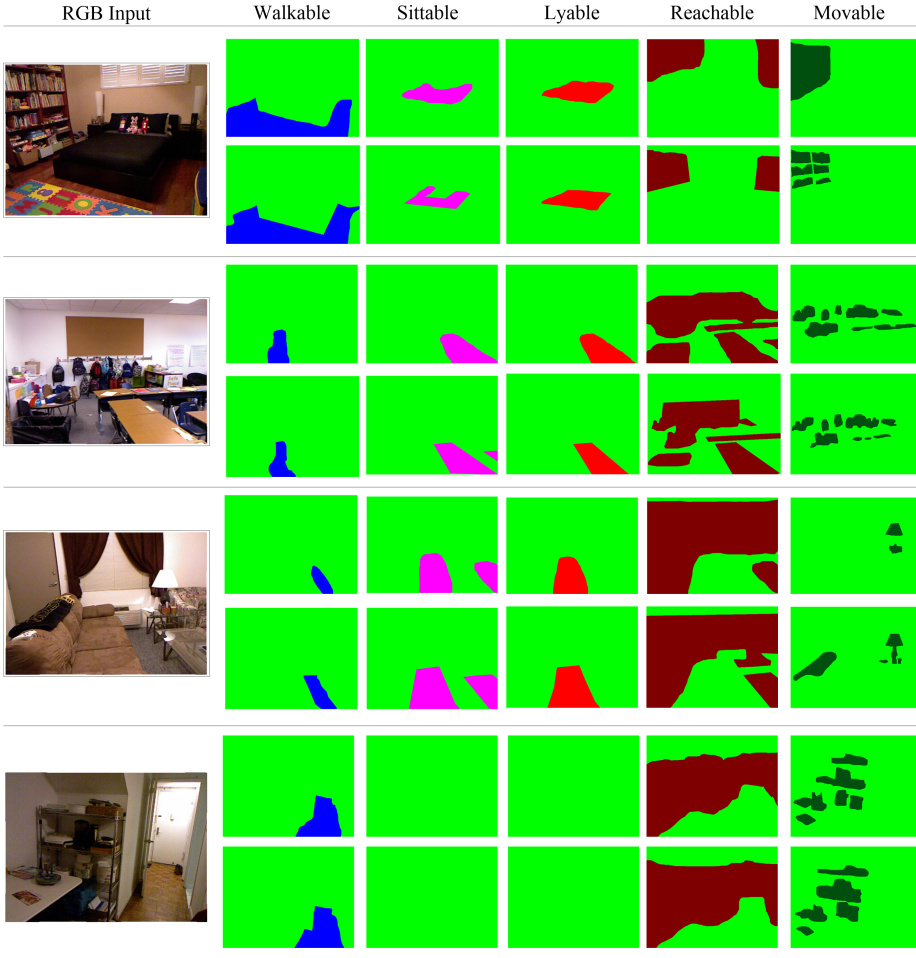

Fig. 5. Qualitative results of affordance segmentation for each type of affrodacne class. For each RGB image, the top row represents the predicted affordance maps and the bottom row represents the ground truth maps.

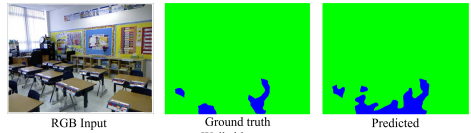

Fig. 6. A failure case where we fail to identify that the table is vertically above the floor, preventing the walking action.

7 Conclusion

We have developed and evaluated a multiscale deep architecture for affordance segmentation in a single RGB image. Three multi-scale CNNs are applied independently to the image for extracting three mid-level cues – namely, depth map, surface normals and semantic segmentation of coarse-level surfaces in the scene. An additional multi-scale CNN is used to fuse these mid-level cues for pixel-wise affordance prediction. For evaluation, we have developed a semi-automated method for generating dense ground-truth affordance maps in images, using RGB and depth information along with ground-truth semantic segmentations as input. This method has been used to augment the NYUv2 dataset of indoor scenes with dense annotations of five affordance types: walkable, sittable, lyable, reachable and movable. Our experiments on the NYUv2 dataset demonstrate that each of the mid-level cues is crucial for the final affordance segmentation, as ignoring any of them significantly downgrades performance. Also, our multi-scale CNN architecture gives a significantly better performance than extracting visual cues at either a coarse or fine scale.

Acknowledgment

This work was supported in part by grant NSF RI 1302700.

References

1. Gibson, J.J.: The theory of affordances. In: Perceiving, Acting, and Knowing: Toward and Ecological Psychology. Erlbaum (1977) 62–82
2. Gibson, J.J.: The ecological approach to visual perception: classic edition. Psychology Press (2014)
3. Barrow, H., Tenenbaum, J.: Recovering intrinsic scene characteristics. Computer vision systems (1978) 3–26
4. Koppula, H.S., Saxena, A.: Anticipating human activities using object affordances for reactive robotic response. Pattern Analysis and Machine Intelligence **38**(1) (2016) 14–29
5. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. International Journal of Robotics Research **32**(8) (2013) 951–970
6. Gupta, A., Kembhavi, A., Davis, L.S.: Observing human-object interactions: Using spatial and functional compatibility for recognition. Pattern Analysis and Machine Intelligence **31**(10) (2009) 1775–1789
7. Fouhey, D.F., Delaitre, V., Gupta, A., Efros, A.A., Laptev, I., Sivic, J.: People watching: Human actions as a cue for single view geometry. International Journal of Computer Vision **110**(3) (2014) 259–274
8. Delaitre, V., Fouhey, D.F., Laptev, I., Sivic, J., Gupta, A., Efros, A.A.: Scene semantics from long-term observation of people. In: ECCV. (2012)
9. Zhu, Y., Fathi, A., Fei-Fei, L.: Reasoning about object affordances in a knowledge base representation. In: ECCV. (2014)

10. Kjellström, H., Romero, J., Kragić, D.: Visual object-action recognition: Inferring object affordances from human demonstration. Computer Vision and Image Understanding **115**(1) (2011) 81–90
11. Yao, B., Ma, J., Fei-Fei, L.: Discovering object functionality. In: ICCV. (2013)
12. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR. (2009)
13. Hoiem, D., Efros, A.A., Hebert, M.: Geometric context from a single image. In: ICCV. (2005)
14. Chen, C., Seff, A., Kornhauser, A., Xiao, J.: Deepdriving: Learning affordance for direct perception in autonomous driving. In: ICCV. (2015)
15. Hoiem, D., Efros, A.A., Hebert, M.: Recovering surface layout from an image. International Journal of Computer Vision **75**(1) (2007) 151–172
16. Hoiem, D., Efros, A.A., Hebert, M.: Closing the loop in scene interpretation. In: CVPR. (2008)
17. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. Pattern Analysis and Machine Intelligence **35**(8) (2013) 1915–1929
18. Pinheiro, P.H., Collobert, R.: Recurrent convolutional neural networks for scene parsing. In: ICML. (2014)
19. Socher, R., Lin, C.C., Manning, C., Ng, A.Y.: Parsing natural scenes and natural language with recursive neural networks. In: ICML. (2011)
20. Couprie, C., Farabet, C., Najman, L., LeCun, Y.: Indoor semantic segmentation using depth information. In: ICLR. (2013)
21. Ning, F., Delhomme, D., LeCun, Y., Piano, F., Bottou, L., Barbano, P.E.: Toward automatic phenotyping of developing embryos from videos. Image Processing (2005)
22. Hariharan, B., Arbeláez, P., Girshick, R., Malik, J.: Simultaneous detection and segmentation. In: ECCV. (2014)
23. Ganin, Y., Lempitsky, V.: N^ 4-fields: Neural network nearest neighbor fields for image transforms. In: ACCV. (2014)
24. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: ICCV. (2015)
25. Zhu, Y., Zhao, Y., Chun Zhu, S.: Understanding tools: Task-oriented object modeling, learning and recognition. In: CVPR. (2015)
26. Myers, A., Kanazawa, A., Fermuller, C., Aloimonos, Y.: Affordance of object parts from geometric features. In: Workshop on Vision meets Cognition, CVPR. (2014)
27. Hermans, T., Rehg, J.M., Bobick, A.: Affordance prediction via learned object attributes. In: ICRA: Workshop on Semantic Perception, Mapping, and Exploration. (2011)
28. Gupta, A., Satkin, S., Efros, A.A., Hebert, M.: From 3d scene geometry to human workspace. In: CVPR. (2011)
29. Fouhey, D.F., Wang, X., Gupta, A.: In defense of the direct perception of affordances. arXiv preprint arXiv:1505.01085 (2015)
30. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV. (2012)
31. Gupta, A., Davis, L.S.: Objects in action: An approach for combining action understanding and object perception. In: CVPR. (2007)
32. Kjellström, H., Romero, J., Martínez, D., Kragić, D.: Simultaneous visual recognition of manipulation actions and manipulated objects. In: ECCV. (2008)
33. Yao, B., Fei-Fei, L.: Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR. (2010)

34. Castellini, C., Tommasi, T., Noceti, N., Odone, F., Caputo, B.: Using object affordances to improve object recognition. Autonomous Mental Development **3**(3) (2011) 207–215
35. Winston, P.H., Binford, T.O., Katz, B., Lowry, M.: Learning physical descriptions from functional definitions, examples, and precedents. Department of Computer Science, Stanford University (1983)
36. Stark, L., Bowyer, K.: Achieving generalized object recognition through reasoning about association of function to structure. Pattern Analysis and Machine Intelligence **13**(10) (1991) 1097–1104
37. Rivlin, E., Dickinson, S.J., Rosenfeld, A.: Recognition by functional parts. Computer Vision and Image Understanding **62**(2) (1995) 164–176
38. Grabner, H., Gall, J., Van Gool, L.: What makes a chair a chair? In: CVPR. (2011)
39. Jiang, Y., Koppula, H., Saxena, A.: Hallucinated humans as the hidden context for labeling 3d scenes. In: CVPR. (2013)
40. Saxena, A., Driemeyer, J., Ng, A.Y.: Robotic grasping of novel objects using vision. International Journal of Robotics Research **27**(2) (2008) 157–173
41. Yu, L.F., Duncan, N., Yeung, S.K.: Fill and transfer: A simple physics-based approach for containability reasoning. In: ICCV. (2015)
42. Xie, D., Todorovic, S., Zhu, S.C.: Inferring dark matter and dark energy from videos. In: ICCV. (2013)
43. Koppula, H.S., Saxena, A.: Physically grounded spatio-temporal object affordances. In: ECCV. (2014)
44. Zhao, Y., Zhu, S.C.: Scene parsing by integrating function, geometry and appearance models. In: CVPR. (2013)
45. Gupta, S., Arbelaez, P., Malik, J.: Perceptual organization and recognition of indoor scenes from rgb-d images. In: CVPR. (2013)
46. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR. (2015)
47. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
48. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NIPS. (2014)
49. Hedau, V., Hoiem, D., Forsyth, D.: Recovering the spatial layout of cluttered rooms. In: ICCV. (2009)