Learning Scene Functionality via Activity Prediction and Hallucination

Ruizhen Hu, Zeyu Huang, Haojing Shen, Manolis Savva, Oliver van Kaick, Angel Chang, Ariel Shamir, *Member, IEEE,* Hao Zhang, *Senior Member, IEEE,* and Hui Huang, *Senior Member, IEEE*

Abstract—We introduce a deep learning approach to acquire and substantiate functional understanding of 3D indoor scenes via human *activity prediction* and *hallucination*. Specifically, we design a deep neural network which takes an *unlabeled* indoor scene without human presence as input and predicts a set of *activity maps*, each of which is a *probability distribution* over the scene indicating the likelihood that a particular human activity is supported at a specific position. Next, we develop an *activity localization* network which takes a scene and a set of predicted activity maps as input and outputs the location, orientation, and pose types of *one or more* humans to substantiate scene functionalities reflected by the activity maps. Finally, we position actual human poses into the scene to "hallucinate" the activities. To train our two deep networks, we crowdsourced a large dataset of activity annotations over SUNCG, collecting descriptions of common human activities that are present in a large collection of 3D indoor scenes. As our data acquisition does not require localization of activity labels onto human-scene interactions, it is scalable and allows us to collect the large volume of data necessary to train the deep neural networks. We show that our method is able to predict different types of activities involving one or more humans for different scene inputs, as well as multiple activities supported by a single scene. We further insert humans with the correct pose types, positions, and orientations at appropriate locations in a scene, to illustrate the corresponding activities and scene functionality. Finally, we will release our large-scale dataset of $11{,}000$+ rooms annotated with $34{,}000$+ activity labels for the benefit of the research community.

Index Terms—3D scene functionality, activity prediction, human hallucination

1 Introduction

In our daily lives, we constantly act on, and react to, the 3D environment that surrounds us to carry out our activities. These activities define how we function as humans. At the same time, the environment and the 3D objects therein have been designed and constructed to *afford* human activities and serve their intended functions. Hence, it is arguable that the ultimate goal of object and scene understanding from an AI perspective is at the *functional* level.

In his seminal book, "The ecological approach to visual perception", James L. Gibson [1] argued that we use layout, 3D, and functions as a way to perceive and understand the world around us [2]. Gibson coined the term "affordance" of the environment as what it offers an agent, what it provides or furnishes. Thus, the term implies the *complementarity* of the agent and the environment. Following in this spirit, we are interested in functional understanding of indoor scenes by learning the "*what*", "*where*", and "*how*" of the *human activities* afforded by a 3D environment. An activity can involve multiple objects and *one or more* humans, where multiple affordances are present. For example, the activity "have conversation" may involve at least two humans sitting on chairs.

In this paper, we introduce a deep learning approach to acquire and substantiate functional understanding of 3D indoor scenes via human *activity prediction* and *hallucination* (Figure 1). Specifically, we develop a deep neural network which takes an *unlabeled* indoor scene without human presence as input and outputs a set of *activity maps*. An activity map is a *probability distribution* over the scene indicating the likelihood that a specific human activity is supported at a given location. Next, we develop an *activity localization* network which takes a scene and a set of predicted activity maps as input and outputs the location, orientation, and pose types of *one or more* humans to substantiate scene functionalities reflected by the activity maps. Finally, we position actual human poses into the scene, based on the activity localization, to "hullucinate" the activities.

Prior approaches to functional scene or object understanding use human pose fitting [3], [4], [5] and what is referred to as the mediated approach [2], where an intermediate semantic or 3D representation is first inferred before functional analysis. In contrast, our activity prediction relies on a deep neural network to *directly* perceive scene functionality from an unlabeled RGBD image representation of the scene, inferring activity maps, as well as object semantics along the way. Recently, Fouhey et al. [2] argued in favor of direct inference over the mediated approach. The former also follows more faithfully Gibson's original idea of affordance analysis [1].

A key challenge to deep learning–based functionality analysis lies in acquiring large volumes of human-scene interaction data with fine-grained annotations. Recording the real interactions of one person and assigning action labels

- *Ruizhen Hu, Zeyu Huang, Haojing Shen, and Hui Huang are with College of Computer Science & Software Engineering, Shenzhen University. Emial: ruizhen.hu@gmail.com.*
- *Manolis Savva, Angel Chang, and Hao Zhang are with Simon Fraser University.*
- *Oliver van Kaick is with Carleton University.*
- *Ariel Shamir is with The Interdisciplinary Center.*

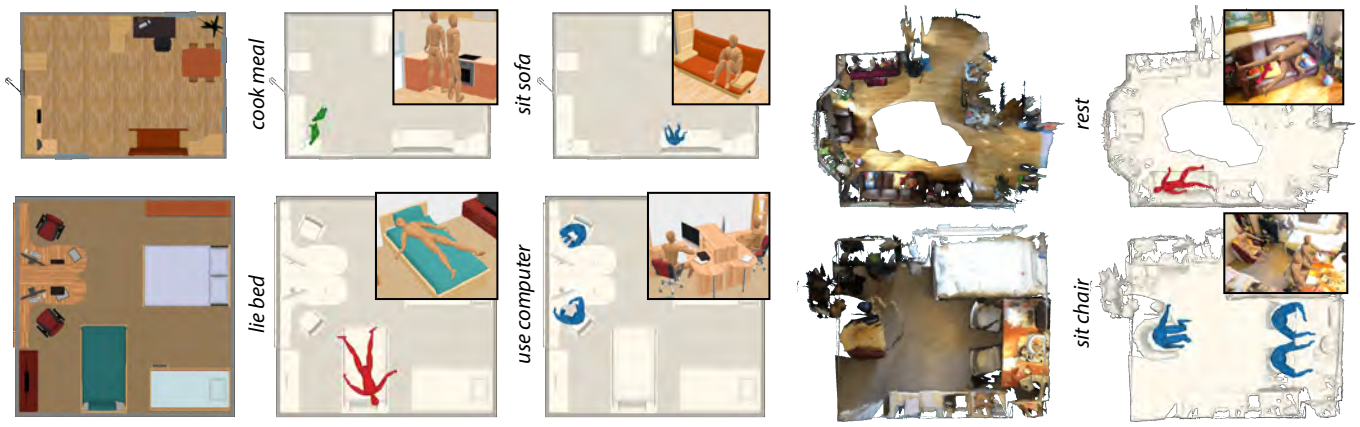


Fig. 1: We develop deep neural networks for functional understanding of 3D scenes, through human *activity prediction* and *hallucination* in both synthetic (left) and reconstructed (right) indoor environments. In each of the four examples, the leftmost image shows an input scene with no humans. Images to the right show predicted activities and their hallucination with associated human poses to illustrate the scene functionalities (top-down view with 3D view inset).

to all interactions is possible, as in SceneGrok [5] for 14 scenes and 45 recordings, but this is hardly scalable. The task becomes intractable when multi-person activities are at play. The vision community has contributed large-scale image [6] and video [7] benchmarks with human-object interaction labels. Such data is valuable for boosting the performance of human-object interaction detection or classification. But they lack 3D geometric information and are subject to the constraints of a first person field of view, limiting their usefulness for reasoning about activities involving multiple people over the space of an entire room.

We train our activity prediction network using *top-down view* RGBD images of 3D indoor scenes from SUNCG [8] in which posed human models reflect common activities. Each scene is annotated with labels summarizing human activities therein, but there are *no* explicit correspondences between the labels and individual human models. Specifically, our labels are obtained from crowdsourced freeform texts describing human activities in each scene. Texts are easier to gather than explicitly tracking and pairing labels to human-scene interactions. However, working with such *unpaired* training data poses a *weakly supervised learning* problem. We address this challenge using an approach relying on two stages. The first stage is responsible for *localizing* an activity label to regions in the scene with human presence, and the second stage uses such localized data to train a network to predict a set of activity maps for scenes without human presence.

We train our localization network with the same top-down view data. Given an input view (without human presence) and an activity map corresponding to a specific activity label, the network returns a set of bounding boxes each reflecting a human position in the scene, along with parameters specifying the orientation and pose type. With this information, actual posed human characters can be added to the scene to visualize the supported functionalities.

Our work contributes the first deep neural network for direct prediction of human activities in indoor scenes. While the top-down view images we work with do not retain full 3D scene geometries, they capture the essence of the "layout, 3D, and functional" characterizations as originally envisioned by Gibson [1]. Moreover, such images are sufficiently expressive in describing a large variety of human indoor activities [9], [10] and they are naturally supported by conventional convolutionary processing using neural networks. By leveraging the large volume of indoor scenes with activity annotations we collected, our method is more scalable and more general than current alternatives, allowing the prediction of a larger class of human activities that can be composed together to generate realistic activities involving multiple people. Finally, we release our large-scale dataset of 11,000+ rooms annotated with 34,000+ activity labels for the benefit of the research community.

2 Related Work

In recent years, there has been a significant amount of work on recognition, prediction, and synthesis of human activities and human-scene interactions, with the ultimate goal of acquiring a functional understanding of the world. For the recognition or detection task, the input scene contains both humans and the 3D objects with which they interact. One group of works focuses on solving the separate problems of pose estimation and action recognition (e.g., [11], [12]), while another group of works jointly approach the problem as the detection of human-object interactions or agent-verb-object pairs in images (e.g., [13], [14], [15], [16], [17]). The prediction task, on the other hand, focuses on inferring plausible human activities or interactions in a scene without human presence. Our work falls into this latter category.

Datasets for interaction recognition. An important front in the recognition of human-object interactions has also been the development of adequate training sets for learning-based approaches. Chao et al. [6] introduce a dataset of human-object interactions encoded as verb-object pairs that describe entire images, showing that the dataset improves a set of baseline approaches. This dataset was extended to enable the use of a convolutional network that detects humans and

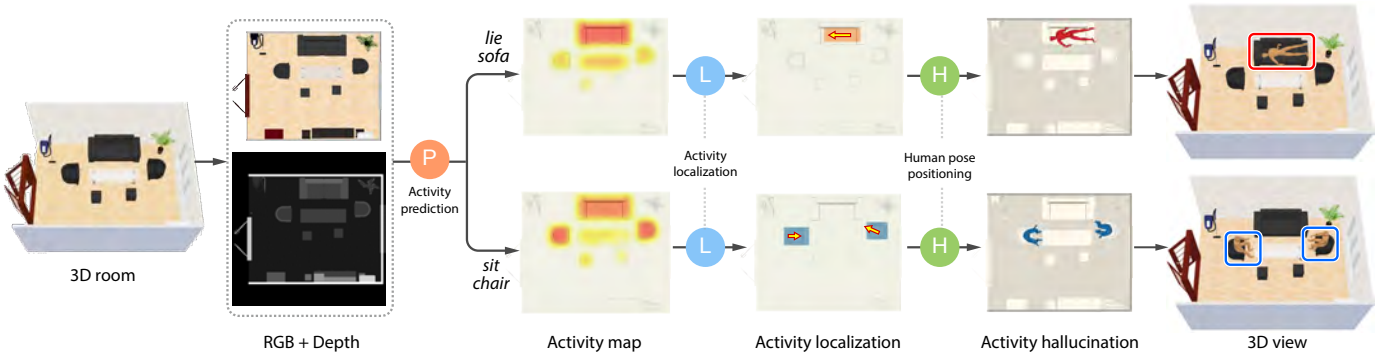


Fig. 2: Overview of our activity prediction and hallucination method: given an unlabeled indoor scene without human presence, represented by a top-down RGBD view, a deep network **P** predicts activity maps that indicate *where* and *what* different activities can be supported in the scene. Next, an activity localization network **L** predicts coarse parameters (i.e., human bounding boxes, types and orientations) which guide the positioning of human poses (labelled by **H**) to illustrate human activities predicted by the action maps.

objects, followed by another network that classifies their interaction [18]. Fouhey et al. [7] annotate lifestyle VLogs with interaction labels to create an unbiased dataset for the understanding of human interactions. As an application, they use the dataset for the task of identifying objects that the hands of the human agents interact with. Qi et al. [19] perform human-object interaction understanding with a graph parsing network.

Affordance prediction and labeling. Roy and Todorovic [20] label images with affordance labels such as "walkable" and "sittable", based on deep networks that directly infer the affordances from input images. Lüddecke and Wörgötter [21] specialize the detection of affordances to the level of object parts. Nagarajan et al. [22] learn to predict human-object interaction affordances in images from weakly supervised first person and third person video data. On the other hand, Zhu et al. [23] employ physics-based simulation to infer forces acting on body parts as people interact with an object, so as to derive human utilities related to affordances. Human activities often involve multiple objects and agents, while affordance is typically tied to a single object. Our work proposes a deep neural network for prediction of activities involving potentially multiple people and objects, and is trained in a weakly supervised fashion from freeform text descriptions of activities.

Another approach for affordance prediction is to *directly hallucinate* human poses that can fit a context of an image or 3D scene. In the seminal work "What makes a chair a chair", Grabner et al. [3] detect affordances by fitting human poses, so as to locate regions in a depth image that support the "sitting" function of chair objects. Jiang et al. [24] hallucinate human poses for an image based on the context of objects detected in the scene. Such information then guides a robot in finding a suitable location for placing a new object into the scene [25]. Chao et al. [26] employ a deep network and adequate training data to address the related problem of human pose forecasting, which is the prediction of human poses for future frames of a given image. In a similar fashion, Kim et al. [4] automatically predict the pose of a human for using a specific 3D object. The method first predicts contact points where the human touches the object and then infers the most adequate pose given an instance of a particular object category. This form of human pose fitting is an indirect approach, which is harder to scale to fine-grained activity prediction than direct inference from scene geometry. Furthermore, for multi-person activities, it is difficult to pre-determine the number of humans and the appropriate regions of interests, making the approach less scalable.

Activity map prediction. A more general form of affordance labeling is the prediction of activity maps for a scene, as in our approach. Rhinehart and Kitani [27] use a first-person view camera to collect activity observations and a matrix factorization approach to predict maps on new scenes. This method establishes a mapping between features from the annotated observations and features from the unseen scenes to infer the activities in the unseen scene. However, their approach relies on accumulating information from egocentric video and only predicts from a set of six manually annotated activity labels that are implicitly defined for a single human.

Most closely related to our work is SceneGrok, by Savva et al. [5], which uses an RGBD camera to observe and track single persons as they interact with a 3D environment. These observations are used to train a classifier that can predict action maps for previously unseen environments. Specifically, their activity prediction is based on human pose fitting, relying on a naive dense random sampling over spatial locations and poses for a single person and evaluating a random forest classifier at each sample. As a result, SceneGrok had runtimes of minutes for a single activity prediction on a single input room scan. Prediction for activities that jointly involve groups of people would be computationally intractable, especially when the total number of people is not known ahead of time.

In contrast, a direct prediction approach using our deep neural network approach runs at real-time rates and predicts regions for a large compositional set of activities that can involve multiple people. A significant increase in the size of the training data played a critical role in making our

approach more scalable and more general. Compared to the 14 scenes with 45 recordings of single-person interactions in SceneGrok, our dataset consists of more than 11,000 scenes with more than 34,000 activity labels. This large-scale dataset allows us to extract implicit correspondences between text describing an activity and the observation of the activity itself with an approach bearing similarities to adaptive attention-based mechanisms for image captioning [28].

Affordance- and functionality-based synthesis. While computer vision research has mainly focused on recognition tasks related to indoor scenes, the computer graphics community has placed more emphasis on automatic or interactive synthesis of 3D scene geometry [9], [10], [29], [30], [31], [32], [33], [34], [35]. There are some prior works related to activity prediction as they use observations or example data to illustrate the usage of scenes and objects, or to modify scenes according to human activities. Fisher et al. [36] generate 3D scenes that allow humans to perform the same activities as in scenes captured with noisy and incomplete 3D scans. Ma et al. [37] learn an action model from photographs, which captures actions involving humans and objects. The action model is then used to alter 3D scenes according to the possible actions carried out by humans. Savva et al. [38] gather observations of people performing actions in 3D environments with skeletal tracking. The observations serve as training data to learn probabilistic graph models called PiGraphs, which link human poses with the geometry and layout of objects near the person. The PiGraphs can then be sampled to provide snapshots depicting interactions that can be carried out in the scenes. Fu et al. [39] define a set of human activity-related object relations and demonstrate that they can be used to improve scene synthesis. Qi et al. [40] use a probabilistic grammar model encoding human contextual relations with Markov Random Fields to synthesize scenes. They predict human affordances as probabilities of position offsets of a person relative to a specific object category. Moreover, Hu et al. [41] hallucinate 3D scenes that illustrate the use of a 3D object.

3 Overview

Figure 2 shows the overview of our method. Given an input scene, our method predicts activity maps for the scene, which reflect the activities that different regions of the scene support. We then perform activity hallucination to localize and place human poses in the scene, illustrating the predicted activities.

We encode the input scene as a top-down view image I with RGB and depth channels. The output of activity prediction consists of activity probability maps P_a for each label L_a, where $P_a(x, y)$ reflects the probability of position $I(x, y)$ supporting activity a. The output of activity localization is a set of bounding boxes B, each specifying the position of a human in the top-view I. Additional parameters specify the 2D orientation and coarse pose type (standing, sitting, or lying) of the human, which are used to place a human figure for activity hallucination.

We first construct an activity dataset by employing workers to gather activity labels for a dataset of scenes. Then, we use the collected labels to prepare training data composed of word maps M_w, activity maps M_a, and ground-truth human bounding boxes H_a for the scenes. Based on this data, we then train deep networks for activity prediction and hallucination. We introduce an activity prediction network that, given a 3D scene without human presence, predicts the activity map corresponding to a specific activity label. We also introduce an activity localization network that predicts the bounding boxes, orientation, and pose type of human models, which guide the placement of 3D human poses into the given scene to illustrate different types of activities. The method used for activity prediction and hallucination are described in Section 4. Dataset construction and training data preparation are described in Sections 5 and 6.

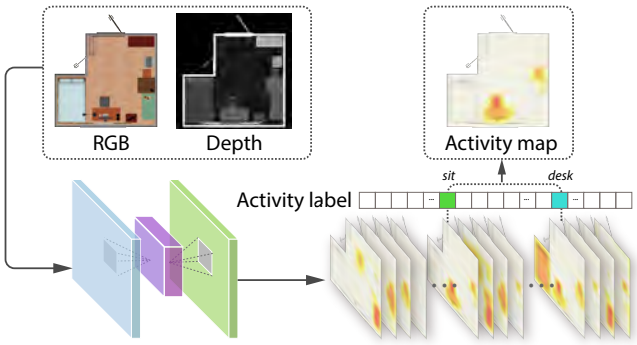


Fig. 3: Our activity prediction network. The network is an encoder-decoder which takes as input the top-down view image of a scene represented with RGB and depth channels and outputs a probability map for each of the $\omega = 34$ words in our dataset. Then, for each activity label, we take one word map (for verbs) or merge two word maps (for verb-noun tuples) to obtain the corresponding activity map.

4 Activity Prediction and Hallucination

Activity prediction. We predict the activity maps P_a with a neural network (see Figure 3). The input is a top-down view I of a scene represented with RGB and depth channels, and the outputs are word maps $\{P_w\}$ that localize the regions in the scene corresponding to the words, e.g., *sit*, *lie*, *sofa*, etc. Each map provides a probability $P_w(x, y)$ reflecting how well a position $I(x, y)$ relates to the word w. Then, to localize an activity label L_a (represented as a verb-noun tuple or only a verb encoded as a multi-hot vector), we take the corresponding word map or merge two word maps to obtain the corresponding activity map P_a, e.g., *sit sofa* or only *sit*.

We predict all the word maps first and then merge them into activity maps as needed, instead of predicting activity maps directly. We do this as some of the activity maps are highly correlated due to sharing a common verb or noun. By predicting all the word maps together and then deriving activity maps through intersection, we are able to both recover the correlation between the words in the input activity labels and the corresponding regions in the maps.

Since most of the activities are strongly correlated with specific object categories, and the object category informa-

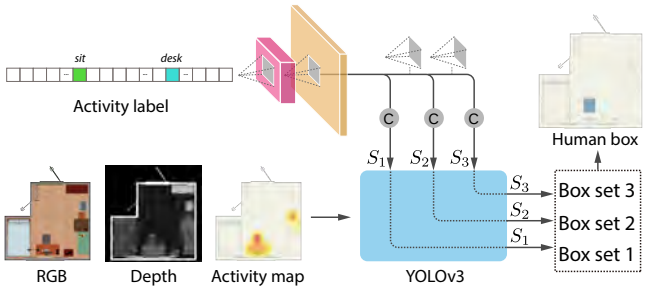


Fig. 4: Our activity localization network. The architecture is based on the YOLOv3 [42] approach and takes as input the RGBD representation of top-down view image of a room, an activity label and the corresponding predicted activity map, and predicts bounding boxes with corresponding human pose types.

tion is available for all the scenes in our dataset, we also pre-train the encoder-decoder network to output the object categories and the segmentation of the input image. Thus, the activity maps and object categories are inferred from the same input layers, although we use different layers at the end of the network according to the type of output. The network structure details can be found in the supplementary material.

Activity localization. Given an input scene encoded as an RGB-D image I and the map P_a of a specific activity label L_a inferred by the prediction network, we use a second deep network to localize the activities by detecting the locations and types of human poses needed to be palced into the scene to illustrate how the selected activity can be carried out.

Specifically, our goal is to infer a set of pose parameters $H = \{h_1, \ldots, h_n\}$ of one or more humans participating in the activity, where a pose parameter h_i encodes the pose type (standing, sitting, and lying), and the 2D position in the top-view of the scene. This task is similar to detecting objects, so we base our network on the YOLOv3 framework [42], which has been shown to provide state-of-the-art object detections [43].

We use the activity label of the input map as a conditional term for the detection, concatenating this label with the output of the convolutional network (see Figure 4). The remaining fully-connected layers learn to detect boxes and types of human poses according to the concatenated features, also outputting confidence values for the pose type and box coordinates. Note that the YOLOv3 network outputs boxes with three different cell resolutions, so that the information coming from the conditional term is passed to all three channels. For our results, we consider only boxes detected with confidence greater than 0.3 and use non-maximum suppression to get the final set of output boxes. See the supplemental material for the network architecture details.

We estimate the orientation of the pose with an additional neural network. The orientation of each pose in a room is defined by a rotation angle specifying its front direction in the reference frame of the room. We represent each bounding box predicted by the hallucination network as a mask B with the same resolution of I and with 1 inside the bounding box and 0 outside. The orientation prediction network then takes as input I with B as an extra channel, and outputs the rotational angle. This network is trained using the pairs of bounding box and rotational angle for all observed poses in the training dataset. We train a network for each human pose type and, during inference, predict the orientation for each predicted bounding box separately.

Human pose positioning. With the predicted bounding box, type and orientation information of each pose, we then position a human pose model of the predicted type by centering it at the bounding box centroid and orienting it such that its front direction matches the predicted orientation, as illustrated in Figure 2. Then, starting with the initial human pose of the given type, we run a ragdoll physics simulation with collision detection to adapt the pose to the surrounding scene. The human model is dropped from 0.5m above the scene and allowed to settle so that its limbs conform to the surrounding furniture. All the poses have limits set for how much the joints can rotate. For sitting and standing poses, joint angles are fixed for the upper body and entire body, respectively. Moreover, since sitting poses frequently occur in front of working surfaces such as desks, we detect the projection points of the feet and the hip on the top surface of the scene, and if the projection point of the feet is higher than that of the hip, we disable collisions for the leg and feet joints and re-drop to avoid unnaturally high leg poses.

5 Data Collection

Activity dataset. Our dataset is based on scenes from the SUNCG dataset [8], which we annotated with activity labels. We selected scenes containing one or more models of humans, which represent approximately 8% of the dataset (32K rooms). We then performed a pre-filtering of this initial dataset to ensure that the data has enough instances of each room type and object.

Data pre-filtering. We first selected rooms that contain 1 to 5 humans, and restricted the floor area to be between 10 to 60 m^2, which covers 77% of the data. We also restricted the room types to *bedroom*, *living room*, *office*, *kitchen*, and *dining room*. These are the five most frequently appearing types, providing a dataset of about 18K rooms. Moreover, we kept only the rooms that contain the 60 most frequently appearing object categories, to prevent certain objects from appearing only a few times in the data. We also removed rooms that contain objects from the "outdoor" and "bathroom" categories.We obtained 17K rooms after the filtering by object types. Finally, to keep the categories of rooms balanced, we subsample the largest category (bedroom) so that it has a similar number of rooms as the other categories, resulting in a final dataset of 12K rooms.

User annotation. To reduce the amount of bias present in the selection of labels, we annotated the pre-filtered dataset with activity labels in two rounds. The first round provided total freedom to the annotators allowing them to enter freeform text descriptions of the activities in the rooms, enabling us to collect an unbiased sample of activity labels. The second

round was more constrained by the set of annotations collected in the first round, to ensure the collection of a larger set of data with consistent labels.

In the first round, we presented a set of rooms one by one to a group of workers. The annotators provided succinct sentences describing each action by a person or group of people in a room. An image of the data collection interface along with the instructions is shown in the supplementary material. The results of annotating 2,000 rooms were then used to create a set of verb-noun labels for subsequent annotation. Specifically, we selected the top 100 most frequently used verbs and nouns in the descriptions. We give more details on the verb-noun parsing in the next paragraph. In the second round, a set of 10,000 rooms were annotated with a similar interface to the first pass, with the addition of an auto-complete interface that suggests verbs and nouns from the top 100 most frequent verbs and nouns. Thus, users could still enter freeform text with autocompletion suggestions based on the first round. The users could use a "Cannot do" button to indicate they were not sure how to label the scene. In total, we obtained annotations for 12K rooms, with about 6 sentences per room on average.

Annotation post-processing. After user annotation, we post-process the annotations to ensure consistency and establish a structured activity label set. We use the Stanford CoreNLP pipeline [44] to convert the freeform text into a set of 44 activity labels composed of verb and noun or verb-only tuples (e.g., "sit", "sit chair", and "sit sofa"). With these labels, we then filter the annotated rooms for annotation consistency. The details of this dataset post-processing procedure are described in the supplemental material.

6 Training Data Preparation

Although our dataset consists of scenes containing human poses assigned with activity labels, we do not have a *direct correspondence* between the humans and labels, which is necessary for training our activity prediction and hallucination networks. Thus, we first transform our dataset into training data with localized activities corresponding to the activity labels using a neural network encoder.

For the preparation of the training data, we use our initial data to train an activity recognition network. The network takes as input a top-down view of a scene and outputs the probable activity labels for the scene. The architecture of the network consists of an encoder and a fully-connected layer that jointly learn a multi-hot classifier of each scene. We encode the activity labels in a multi-hot vector representation that allows for composition of the verbs and nouns in our activity label set. As illustrated in Figure 3, we split the verb-noun tuples to create a separate entry for each verb and noun in the multi-hot vector, also creating entries for the isolated verbs if necessary. In this process, we obtain 16 different verbs and 18 different nouns, creating a multi-hot vector of dimension $\omega = 34$. Note that the original dataset contains 44 activity labels stored as tuples of verbs and nouns or verbs only.

To identify spatial regions corresponding to each word, we use the class peak response method of Zhou et al. [45]

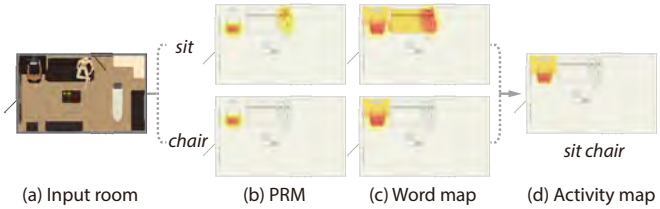


Fig. 5: Creation of an activity map: we combine the PRM for a word with an object mask to obtain the word map, and intersect the two word maps to obtain the activity map for a verb-noun label. The position probabilities of the word and activity maps are shown as heat maps ranging from white (minimum) through yellow and up to red (maximum value).

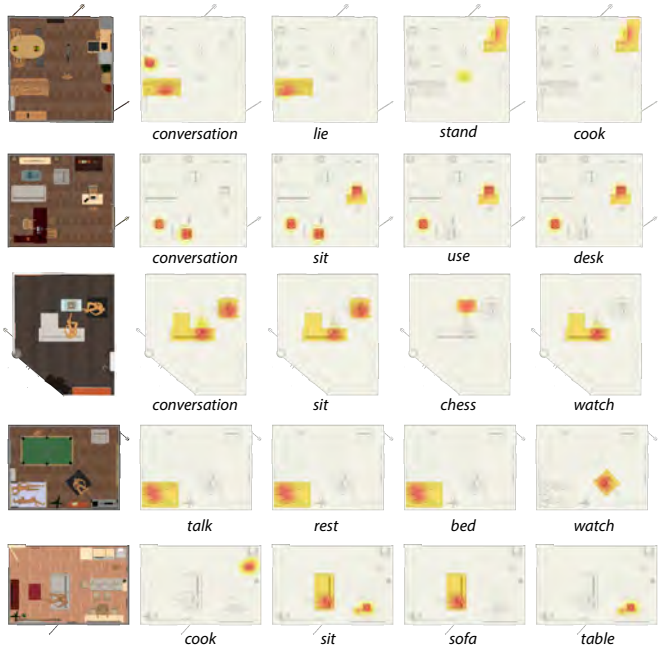


Fig. 6: Examples of training data for activity map prediction. For each input scene on the left, we show different word maps generated from our human activity annotated data. Note the natural emergence of correlations between related verbs and objects, such as *cook* and a kitchen counter, and subtle differences between maps corresponding to different but related words, e.g., *conversation* and *lie* or *conversation* and *sit* in the first two rows.

to derive a Peak Response Map (PRM) for each word. A PRM is a visual map that identifies the regions where the encoder provides peak responses for a specific label. PRMs are computed by applying a peak stimulation process to the learned network, followed by a peak backpropagation that creates the visual map for a given word [45]. We iterate over every word entry in our activity label representation and create their corresponding PRMs.

To prepare the training data for activity prediction, we identify the objects intersected by the detected peak regions, and extend the response maps so that they fully include each intersected object. Specifically, we combine the PRM and mask of intersected objects with equal weights. The object mask is composed of pixels with values of 1 (object

present) or 0, while the PRM is also normalized by mapping the pixel values to the $[0, 1]$ range. The outcome of this step is a set of word maps M_w for each word w, which provide a correspondence between each verb or noun and spatial regions of the scenes; see Figure 5 (b-c) for two examples. These word maps are used to train our activity map prediction network. Figure 6 shows examples of the variety of word maps that we obtain.

For training activity localization, instead of using the word maps, we obtain the combined activity map M_a for each activity label L_a and the corresponding human pose parameters H_a. More specifically, for verb-noun activity labels, we take the intersection of the two word maps corresponding to the verb and noun entry of the label, and assign the maximum entry value of the two maps to the combined map. For verb-only labels, we simply take the corresponding word map. An example of this construction is shown in Figure 5 (c-d). For the corresponding human pose information, we find human poses in the training set scenes that intersect the activity map, and use the bounding boxes of individual human poses with their coarse pose label (standing, sitting, or lying) as the training set H_a. Please refer to the supplemental material for examples of the training dataset and detailed statistics.

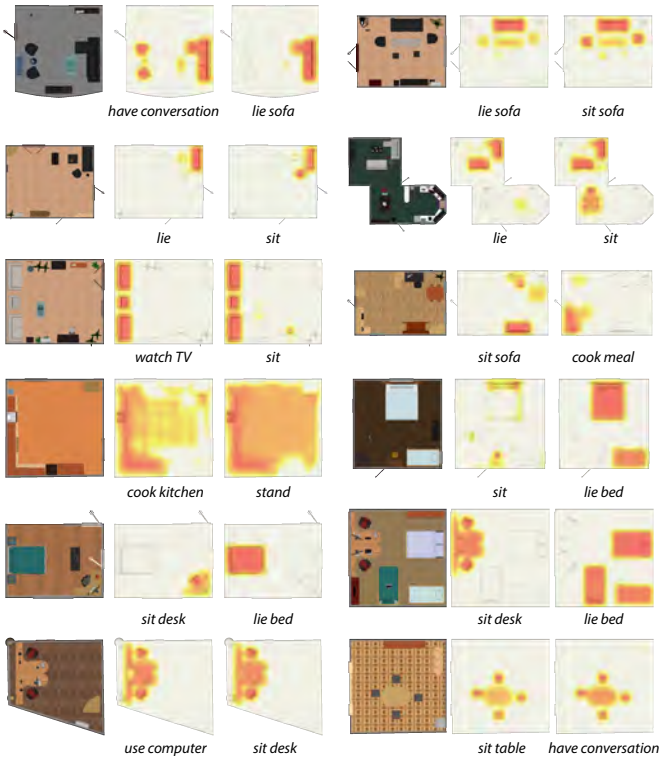


Fig. 7: Examples of activity prediction from the test set. For each room, we show activity maps for two different input activity labels.

7 Results and Evaluation

We show results for activity prediction and hallucination, and evaluate each step with comparisons to other methods.

7.1 Qualitative results

Activity prediction. Figure 7 shows examples of activity map prediction. Our predictions capture subtle differences between activities that can happen in the same space. In the left example of the first row, we predict high probability for the two single chairs when given the label *have conversation*, while excluding them for *lie sofa* since they are too small to support lying down. For the left example in the third row, *watch TV* gives high probability for sittable regions facing the TV while excluding the small *sit* region on the side. In the right example of the third row, we successfully detect activities that usually appear in a different type of room within a multi-functionality room. For example, *sit sofa* usually appears in *living rooms* while *cook meal* usually appears in *kitchens*. In the fifth row, we see predictions of the same activity label detected in rooms with different layouts. In the last row, highly-correlated activity labels such as *use computer* and *sit desk* result in similar maps.

Activity hallucination. Figure 8 shows examples of hallucinated activities demonstrated by human poses placed based on our predicted activity maps. Standing poses are inserted most frequently in *kitchens* close to the cabinet with activity labels such as *cook meal*, *cook kitchen*, and *stand kitchen*, as shown in the first three rows. Lying poses are always related to activities such as *lie bed* and *lie sofa*. The results in the fourth to sixth row show that we reliably detect bed and sofa regions and place the human poses with correct positions and orientations in those regions, accounting for different bed orientations and positions. Moreover, our method can place different numbers of human poses, e.g., two people lying in the same bed. Sitting poses are related to a different set of activities, such as *have conversation*, *watch TV*, and *use computer*. For activities involving multiple persons, our method inserts more than one human pose in the same region, with orientations such that the humans appropriately face each other. For *watch TV*, one or more sitting poses are usually generated and placed facing the TV. For *use computer*, the final human pose numbers and locations depend on the number of desks and chairs supporting the activity.

7.2 Quantitative evaluation and comparisons

Our method involves three steps: i) activity recognition for training data preparation, ii) activity prediction, and iii) activity hallucination. Here, we quantitatively evaluate each step. Our dataset is split into 70%, 15%, 15% portions for training, validation, and testing.

Activity recognition. We first use a multi-label classifier to find the correspondence between activity labels and local regions in the scene, which takes the 3D room with human poses as input and produces a set of activity labels as output. We evaluate different ways of encoding the input room: RGB, depth, and RGBD images. The average classification accuracy and F1 score are (91.6%, 0.26) for RGB, (92.9%, 0.33) for depth, and (93.1%, 0.37) for RGBD. We can see that depth provides more useful information than RGB, and using both RGB and depth provides the best performance.

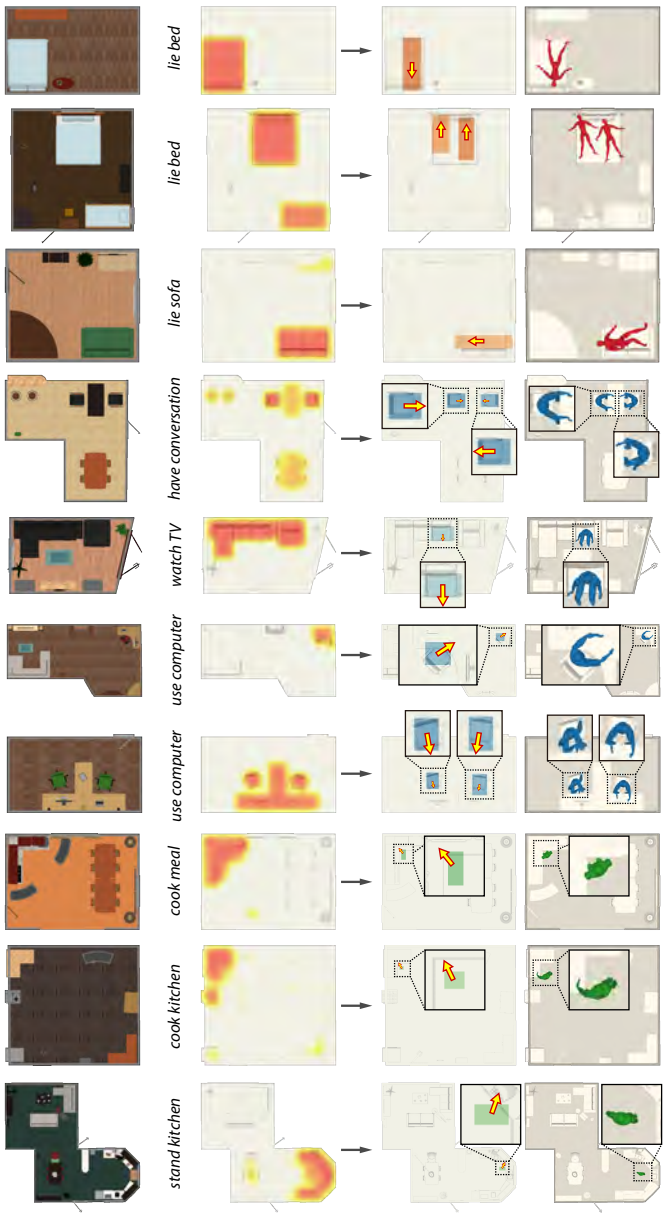


Fig. 8: Examples of human activity hallucination. From left to right, we show the input room, activity map corresponding to the input activity label, the detected human bounding boxes with predicted pose type and orientation, and the final hallucinated human poses in the input room. See the supplementary material for additional 3D views. The predicted poses are positioned and oriented realistically, demonstrating how people use objects during the input activity.

Once the network is trained, we can extract a word map for each scene and each word, and then derive the activity maps corresponding to different activity labels as we described in Section 6. This output forms our training data for the activity map prediction and human hallucination on scenes without humans. An alternative approach is to take rooms without human presence as input and train the multi-label classifier so that the activity map for those scenes can be extracted directly from the network without the need for an

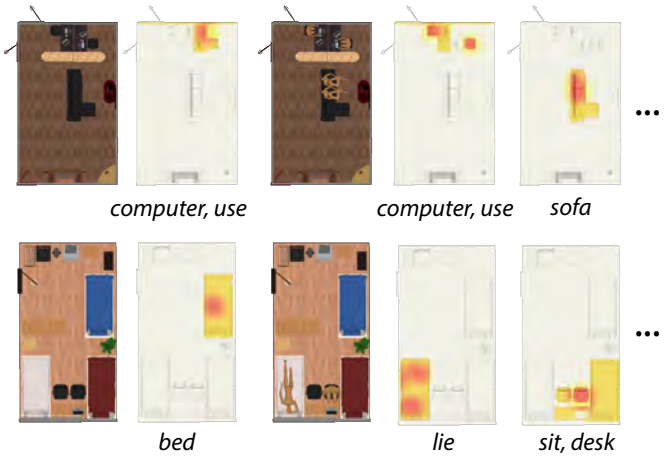


Fig. 9: Comparison of word maps obtained from input rooms with and without observed humans. Left: the only word map obtained for the given scene without human. Right: some examples of word maps obtained for the same scene but with a human present.

activity prediction network. However, this approach has the disadvantage that it does not account for human activities that were not annotated for the room, but that are still possible (e.g., not all chairs will have a person sitting on them in our training data). This partial annotation leads to incorrect penalization of true generalization capability for activity classification and region localization. Figure 9 shows a comparison of extracted word maps from scenes with and without humans. The average classification accuracy and F1 score from input RGBD rooms without humans are 90.4% and 0.23, respectively, which are lower than the 93.1% and 0.37 that we obtain with the use of humans.

Activity prediction. To form the training data for activity prediction, we take the RGBD representation of 3D rooms as input and use the corresponding word maps extracted from the multi-label classifier for activity recognition as the output that the network should predict. Since we use an encoder-decoder network as commonly used for semantic segmentation tasks, we pretrain the network for object category segmentation, obtaining 96.5% accuracy. Then, we replace the last few layers with layers that output word maps, and fine-tune the network on the data that we prepared.

To evaluate the predicted activity maps, we collected "ground truth" activity map annotations from crowdworkers. This is necessary as the original human pose placements that we used for our training data only offer a sparse sample of potential positions where an action might take place, and would penalize generalization to valid regions (e.g. placing a sitting pose in a chair where no pose had been observed). To collect this data, we use a similar interface to the one used for activity label collection. We present a top-down image view of a room, and ask the worker to draw regions where a person would be able to perform a given activity label such as "sit chair". The workers then use an arbitrary number of bounding boxes to delineate the region. If an activity was not possible in the room, the worker could mark it as "Cannot do". In this way, we constructed

more complete activity maps that can be used as ground truth for evaluating our predictions. Some of the example annotations can be found in the last column of Figure 11.

With these annotated activity maps, we evaluate activity map prediction by computing the recall, precision, and F1 scores of our predictions relative to the annotations. The recall is computed pixel-wise between predicted activity map and ground truth map with a threshold value of 0.1. Our definition of precision accounts for the continuous value confidence of the activity maps by taking the intersection between the ground truth and the predicted map, and normalizing by the summed confidence of the predicted map.

We evaluate against four baselines, which include one object-centric baseline and three variations of our method:

- *ObjectRegions*: perform object detection and obtain the activity maps by mapping object labels to activity labels.
- *HumanRegions*: direct prediction of activity maps but using observed human poses and intersected object regions in raw data as training data.
- *ConditionalActMapI*: conditional prediction with training data consisting of only rooms with their supported activity labels and corresponding activity maps.
- *ConditionalActMapII*: conditional prediction with training data not only consisting of rooms with positive activity labels but also same number of rooms with negative activity labels together with their empty activity maps.

The *ObjectRegions* baseline uses object detection to infer the activity maps. After detecting object regions and their labels are mapped to activity labels according to a set of rules, e.g, a region such as *sofa* is mapped to *sit sofa*. Activity labels without concrete objects, e.g., *cook meal*, are derived from related objects, e.g., *kitchenware*. Finally, verb-only activity labels are derived from the union of all detected object labels that involve the verb in the rules. The specific rules used in the mapping are listed in the supplementary material. Comparison against this baseline justifies the use of activity map detection instead of simply performing object detection.

The *HumanRegions* baseline does not use the multi-label classification training data preparation procedure. Instead, it uses the annotated activity labels and observed human poses in the training data directly. However, since the activity labels do not have correspondences to regions in the room, we extract a mask over all human poses in the scene and combine it with the object mask consisting of objects intersected by the human, as we did for word map generation. Then, we use this as the word map for all the verbs and nouns associated with the given room to train the word map prediction network. Comparison against this baseline justifies the use of multi-label classification for training data preparation.

The *ConditionalActMap* baseline is a conditional version of the network which takes as input the top-down view I of a scene as well as an activity label L_a that we would like to localize in the scene, and then outputs an activity map P_a directly that localizes the specified activity in the scene in the form of a probability $P_a(x, y)$ reflecting how well a pixel $I(x, y)$ supports the activity. The activity map for each room with respect to each activity label is obtained as described in Section 6. However, since most of the rooms only support a small set of activities with limited human poses presented, the number of negative examples that we obtain is more than 10 times the number of positive examples. Since the positive examples are more reliable for activity map prediction on scenes without humans, we evaluated two settings for *ConditionalActMap*. The first setting uses only positive examples, while the second uses a sample of negative examples with the same size as the set of positive examples. Comparison against this baseline justifies the choice of predicting all the word maps first and then merging them into activity maps when needed.

TABLE 1
Evaluation of our activity map prediction method against baseline methods. Metrics report recall, precision and F1 score against ground truth activity map. Higher values are better.

Method	Recall	Precision	F1
ObjectRegions	0.37	0.41	0.34
HumanRegions	0.46	0.29	0.30
ConditionalActMapI	0.28	0.13	0.13
ConditionalActMapII	0.29	0.21	0.19
Ours	**0.49**	**0.44**	**0.44**

Table 1 reports the performance of our method against the baselines. We see that our method obtains the best performance. Figure 10 shows visual comparisons of our method to the object-centric baseline. The object-centric baseline fails to produce correct results when: 1) The object segmentation network predicts incorrect object labels; 2) The object segmentation is correct but the detected object regions do not fully cover the activity region; 3) Activities with the same noun but different verbs may correspond to different activity regions. Corresponding examples are shown in the different rows of Figure 10. Case 1: the TV bench is labeled as *desk* and thus added into the activity region for *use desk*. Although this case is rare in the results, as the segmentation network obtains 96.5% segmentation accuracy, cases 2 and 3 are still common even with a perfect segmentation. Case 2: an activity like *look computer* should include the object regions such as *desk* and *chair* around *computer*, but not all such regions in the scene. However, this kind of contextual information cannot be easily added to the object-centric method. Case 3: *sofa* can be associated with both *lie* and *sit* verbs, however, the two activities require different sofa sizes. This implies that the method should consider additional information beyond only the object categories involved in the activity.

Figure 11 compares the different activity prediction methods. We see that the *HumanRegions* baseline cannot leverage the more specific correspondences between activity labels and regions that we obtain from the multi-label classifier. Instead, the baseline considers all observed humans in the training data when learning to predict any activity label

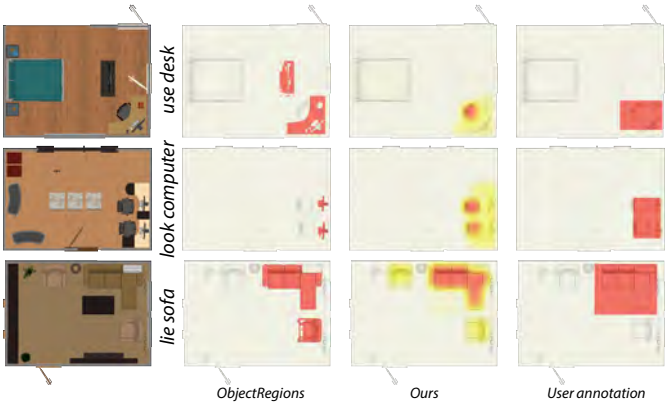


Fig. 10: Comparison of our method to the object-centric baseline.

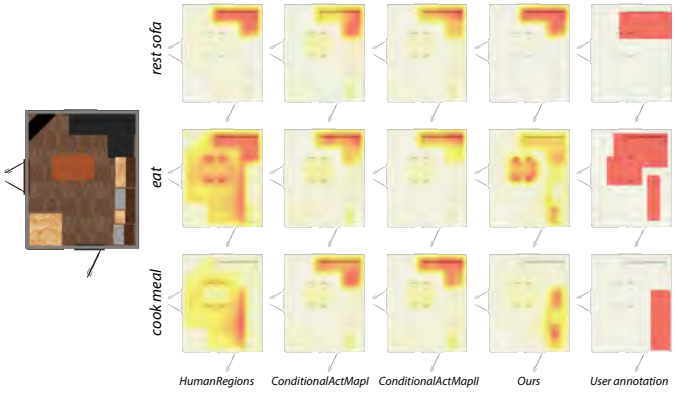


Fig. 11: Comparison of different activity prediction methods.

assigned to a room. The effect is that the network predicts much larger regions in general. We also see that the two *ConditionalActMap* baselines both tend to output similar maps for different activity labels. The main reason could be that it is difficult for the network to pass information from the multi-hot input activity label to later layers that guide the final pixel-level prediction.

Activity hallucination. Given the RGBD representation of the input room and the predicted activity map corresponding to an input activity label, our goal in activity hallucination is to put a human pose in the scene to illustrate the activity. This is achieved in three steps: human pose type and location detection, pose orientation prediction, and 3D human pose insertion. We present a quantitative evaluation for the first two steps.

For human pose type and location detection, the input is a top-down view I of a scene with RGB and depth channels, the activity map M_a, and the activity label L_a, giving us a combined input of $I + M_a + L_a$. The outputs are human bounding boxes associated with each pose type. We formulate this as an object detection problem (where we predict a bounding box for the human pose and a label for its type). Thus, we can evaluate the method using a standard mAP (mean average precision) against the ground truth. We compare this approach against ablations using three alternative input configurations: $I + L_a$, M_a alone, and $M_a + L_a$. Table 2 reports the results. As we can see from the comparison, using $I + L_a$ as input without providing the activity map has the worst performance. Using M_a alone and training the network for all kinds of activity labels together also underperforms. When we add L_a and M_a to the input image I, we observe the best performance.

TABLE 2
Evaluation of our activity hallucination. Numbers report mean average precision (mAP). Higher values are better.

Method	$I + L_a$	M_a	$M_a + L_a$	$I + M_a + L_a$
mAP	87.9%	90.3%	93.1%	**93.5%**

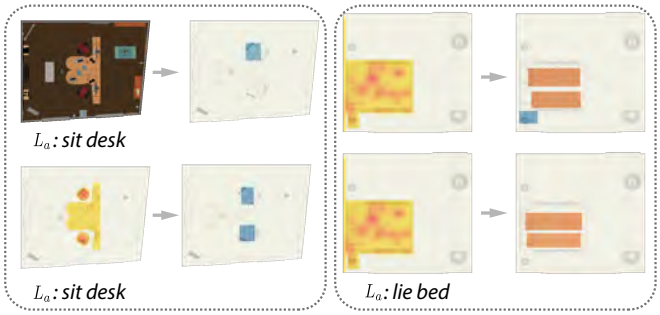


Fig. 12: Comparisons of our human pose detection approach with different input configurations. Left: comparison between $I + L_a$ and $M_a + L_a$. Right: comparison between M_a and $M_a + L_a$.

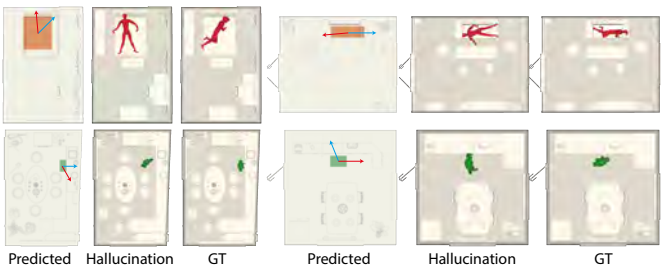


Fig. 13: Examples of activity hallucination results with large orientation error but still valid. We show two examples for lying poses (first row) and two examples for standing poses (second row). For each example, from left to right, we show the predicted human box with ground-truth orientation (blue arrow) and predicted orientation (red arrow), the scene with a hallucinated human in the predicted orientation, and the ground-truth pose orientation.

Figure 12 shows selected comparisons between $I + L_a$ and $M_a + L_a$, and also comparisons between M_a and $M_a + L_a$. On the left, we can see that without the activity map, providing the image input I results in missing some regions that do support the corresponding activity. On the right, we see that when taking the activity map M_a alone without providing the label L_a as input, human boxes are predicted on the highlighted regions without considering whether the pose itself is valid or reflective of the corresponding label L_a.

For the pose orientation prediction, we train one network for each pose type. The average prediction angle error and

standard deviation in degrees are: (25, 37), (58, 53), and (53, 50) for sitting, standing and lying poses, respectively. Despite large angle error values, we observe that in several cases the poses are in fact meaningful compared to the ground truth; see Figure 13 for some examples.

In the supplementary material, we also compare our human pose type and location detection method to the object-centric baseline, and provide more details about the implementation of the baseline.

7.3 Applicability to real scans

To show the generality of our method and its potential applicability to scans of real rooms, we re-train all the networks with input scenes represented using top-down view images but encoding only the depth channel. Then, we test the prediction on scans of scenes from ScanNet [46]. Despite the fact that our networks are trained on synthetic data, we found that they can already be used to predict reasonable activity regions in the scanned scenes, along with the corresponding human pose types and orientations. Figure 14 shows activity prediction and hallucination results on scanned scenes. We see that our method provides reasonable predictions such as *sit chair* and *use computer*, although noisy regions can lead to false positives such as *wash sink*.

To perform a quantitative evaluation of the prediction accuracy on real scans, we also collected ground-truth activity maps for all the scenes of ScanNet with room types that have corresponding training data in SUNCG, which results in more than 700 labeled scenes. We compare our method to the object-centric baseline as in the evaluation for synthetic data. The F1 scores that we obtain are 0.27 and 0.12 for our method and the object-centric baseline, respectively. The cases when the object-centric baseline fails are similar to the ones we listed when comparing on the SUNCG dataset. Figure 15 shows visual comparisons of our method to the object-centric baseline. We see that the object-centric baseline fails to correctly segment the objects in many examples, which directly leads to incorrect activity maps. Although there are cases where the object regions were detected correctly, the detected object regions do not fully cover the regions required by related activities, e.g., *look computer*, *sit table*, and *sit desk*. Moreover, when using the object-centric method, some regions cannot be further filtered based on specific requirements of the activities, e.g., *sit sofa* vs. *lie sofa*.

7.4 Application to activity-based scene retrieval

We use our method to demonstrate a scene retrieval application, where scenes are retrieved given a query set of *multiple activities*. For each activity, we compute a score of how well a scene supports the activity by summing the confidence across the entire activity map predicted for the scene. Then, we rank the scenes based on the average of the scores for all the activities in the given set.

Figure 16 shows examples of scenes retrieved for different sets of activity labels. We see that all the retrieved scenes support multiple uses. The first row shows that rooms with beds and sofas or chairs are needed for combining the activities *rest bed* and *have conversation*. The second row shows that hybrid rooms which are a mix of kitchens, dining rooms, and living rooms, are retrieved for supporting the combination of the activities *cook kitchen*, *sit table*, and *lie sofa*. The last row shows that scenes which are a mix of living rooms and offices are retrieved for supporting the combination of the activities *play chess*, *watch television*, and *use computer*.

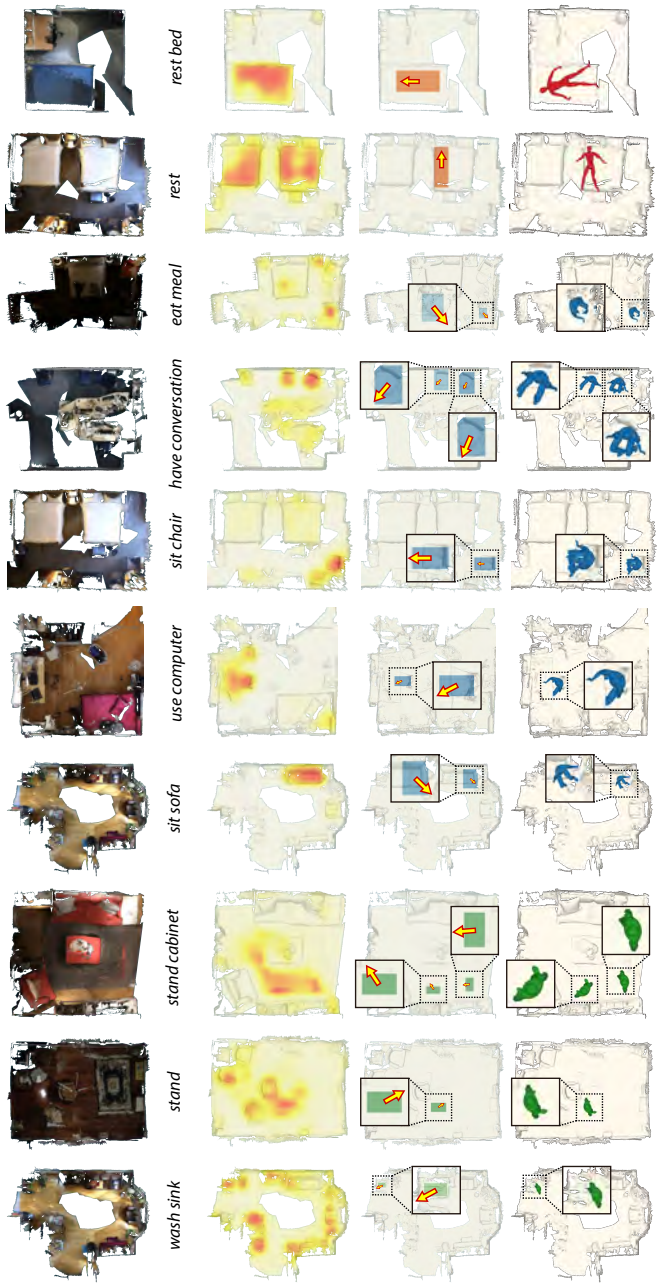


Fig. 14: Activity prediction and hallucination results on scans of real rooms from ScanNet [46]. See the supplementary material for additional 3D views.

8 Conclusion and Future Work

The key to obtaining a functional understanding of indoor scenes is learning the "what", "where", and "how" of human activities supported by an environment. We have

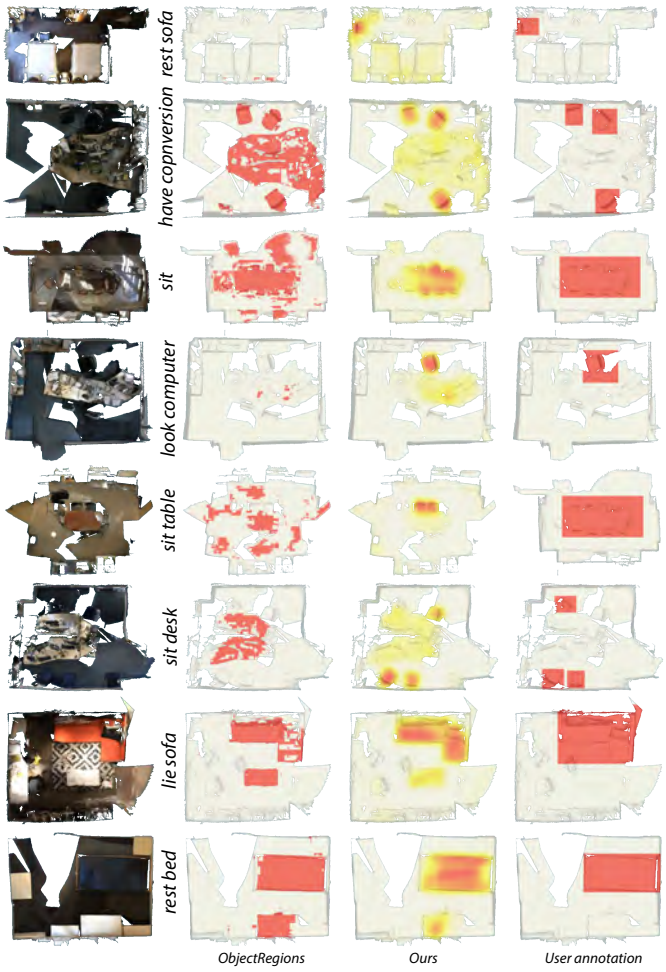


Fig. 15: Comparison of our method to the object-centric baseline on rooms from ScanNet [46]. Our method is more precise and produces results more similar to the ground-truth in many of the cases.

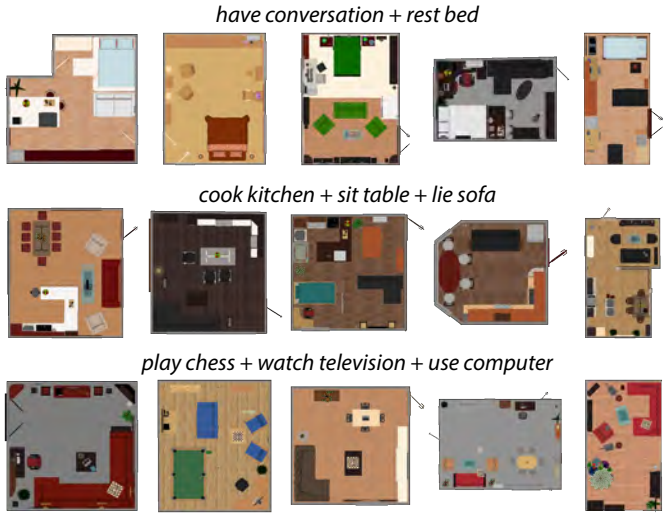


Fig. 16: Example results of activity-based scene retrieval performed with our method, showing how hybrid rooms that support multiple activities are retrieved for queries involving multiple activity labels.

developed a deep learning-based approach for such a functional scene understanding via human activity prediction and hallucination. Our method takes as input an unlabeled top-down view of a 3D scene and outputs a set of activity maps indicating the likelihood of various human activities that are supported over the input scene. Further, our activity localization network takes an input scene with an activity map and predicts where and how human models can be placed into the scene to substantiate functional activities reflected by the action map. To train our deep networks, we crowdsourced a dataset of activity annotations over SUNCG, collecting descriptions of common human activities in a large collection of 3D indoor scenes. As our data acquisition does not require localization of activity labels onto human-scene interactions, it is scalable and allows us to collect the large volume of data necessary for neural network methods.

Our current framework is limited to analyzing top-down views, which are natural for convolutional networks, but they do not retain full 3D scene geometries. In addition, our current scheme for human pose positioning is still rather rudimentary. Fine-grained pose synthesis is a stand-alone and challenging problem in its own right, which likely requires additional data to provide the full 3D configuration of human agents interacting with 3D scenes. To the best of our knowledge, such large-scale datasets are unavailable. Constructing such data from videos is an interesting avenue for future work. Lastly, our activity annotations have focused on human-scene interactions and do not include close human-to-human interactions. Extending our approach by incorporating methods from the human motion modeling and character animation literature is a promising direction to address these technical limitations.

We would also like to extend our functional analysis to first person views. Our method could be combined with a domain adaptation method to target egocentric image or video data with or without observed humans. An exciting direction for future work is to integrate our annotated 3D scene dataset with common activity video data such as the VLOG dataset [7]. One possibility to address these tasks is to train our networks on virtual scans of synthetic scenes, to obtain networks applicable to partial views such as depth images inferred from input frames of videos. Another promising direction is to develop a scene plausibility score based on how well one or more human activities can be afforded by a given indoor scene. Based on such a score we can perform scene synthesis or scene refinement, by re-arranging or inserting object to improve the functional plausibility of a 3D scene.

References

[1] J. J. Gibson, *The ecological approach to visual perception*. Boston: Houghton Mifflin, 1979.

[2] D. F. Fouhey, X. Wang, and A. Gupta, "In defense of the direct perception of affordances," in *European Conf. on Computer Vision*, 2016.

[3] H. Grabner, J. Gall, and L. V. Gool, "What makes a chair a chair?" in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 1529–1536.

[4] V. G. Kim, S. Chaudhuri, L. Guibas, and T. Funkhouser, "Shape2pose: Human-centric shape analysis," *ACM Trans. on Graph (SIGGRAPH)*, vol. 33, no. 4, pp. 120:1–12, 2014.

[5] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, and M. Nießner, "SceneGrok: inferring action maps in 3D environments," *ACM Trans. on Graph (SIGGRAPH Asia)*, vol. 33, no. 6, pp. 212:1–212:10, 2014.

[6] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng, "HICO: a benchmark for recognizing human-object interactions in images," in *Proc. Int. Conf. on Comp. Vis. (ICCV)*, 2015, pp. 1017–1025.

[7] D. F. Fouhey, W. Kuo, A. A. Efros, and J. Malik, "From lifestyle VLOGs to everyday interactions," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] K. Wang, M. Savva, A. X. Chang, and D. Ritchie, "Deep convolutional priors for indoor scene synthesis," *ACM Trans. on Graph (SIGGRAPH)*, vol. 37, no. 4, pp. 70:1–70:14, 2018.

[10] K. Wang, Y. an Lin, B. Weissmann, M. Savva, A. X. Chang, and D. Ritchie, "PlanIT: Planning and instantiating indoor scenes with relation graph and spatial prior networks," *ACM Trans. on Graph (SIGGRAPH)*, vol. 38, no. 4, p. to appear, 2019.

[11] A. Toshev and C. Szegedy, "DeepPose: Human pose estimation via deep neural networks," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1653–1660.

[12] T.-H. Vu, C. Olsson, I. Laptev, A. Oliva, and J. Sivic, "Predicting actions from static scenes," in *European Conf. on Computer Vision*, 2014.

[13] B. Yao and L. Fei-Fei, "Recognizing human-object interactions in still images by modeling the mutual context of objects and human poses," *Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 9, pp. 1691–1703, 2012.

[14] V. Delaitre, D. F. Fouhey, I. Laptev, J. Sivic, A. Gupta, and A. A. Efros, "Scene semantics from long-term observation of people," in *European Conf. on Computer Vision*. Springer, 2012, pp. 284–298.

[15] B. Yao, J. Ma, and L. Fei-Fei, "Discovering object functionality," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2512–2519.

[16] G. Gkioxari, R. Girshick, P. Dollár, and K. He, "Detecting and recognizing human-object interactions," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[17] K. Kato, Y. Li, and A. Gupta, "Compositional learning for human object interaction," in *European Conf. on Computer Vision*, 2018, pp. 234–251.

[18] Y.-W. Chao, Y. Liu, X. Liu, H. Zeng, and J. Deng, "Learning to detect human-object interactions," in *Proc. IEEE Winter Conference on Applications of Computer Vision*, 2018.

[19] S. Qi, W. Wang, B. Jia, J. Shen, and S.-C. Zhu, "Learning human-object interactions by graph parsing neural networks," in *European Conf. on Computer Vision*, 2018.

[20] A. Roy and S. Todorovic, "A multi-scale cnn for affordance segmentation in rgb images," in *European Conf. on Computer Vision*. Springer, 2016, pp. 186–201.

[21] T. Lüddecke and F. Wörgötter, "Learning to segment affordances," in *Int. Conf. on Computer Vision Workshop (ICCVW)*. IEEE, 2017, pp. 769–776.

[22] T. Nagarajan, C. Feichtenhofer, and K. Grauman, "Grounded human-object interaction hotspots from video," in *arXiv preprint arXiv:1812.04558*, 2018.

[23] Y. Zhu, C. Jiang, Y. Zhao, D. Terzopoulos, and S.-C. Zhu, "Inferring forces and learning human utilities from videos," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 3823–3833.

[24] Y. Jiang, H. Koppula, and A. Saxena, "Hallucinated humans as the hidden context for labeling 3D scenes," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2013, pp. 2993–3000.

[25] Y. Jiang and A. Saxena, "Hallucinating humans for learning robotic placement of objects," in *Experimental Robotics*. Springer, 2013, pp. 921–937.

[26] Y.-W. Chao, J. Yang, B. Price, S. Cohen, and J. Deng, "Forecasting human dynamics from static images," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3643–3651.

[27] N. Rhinehart and K. M. Kitani, "Learning action maps of large environments via first-person vision," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 580–588.

[28] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375–383.

[29] L.-F. Yu, S. K. Yeung, C.-K. Tang, D. Terzopoulos, T. F. Chan, and S. Osher, "Make it home: automatic optimization of furniture arrangement," *ACM Trans. on Graph*, vol. 30, no. 4, pp. 86:1–12, 2011.

[30] P. Merrell, E. Schkufza, Z. Li, M. Agrawala, and V. Koltun, "Interactive furniture layout using interior design guidelines," *ACM Trans. on Graph*, vol. 30, no. 4, pp. 87:1–10, 2011.

[31] M. Fisher, D. Ritchie, M. Savva, T. Funkhouser, and P. Hanrahan, "Example-based synthesis of 3D object arrangements," *ACM Trans. on Graph*, vol. 31, no. 6, pp. 135:1–11, 2012.

[32] Z. Sadeghipour, Z. Liao, P. Tan, and H. Zhang, "Learning 3D scene synthesis from annotated RGB-D images," *Computer Graphics Forum (SGP)*, vol. 35, no. 5, pp. 197–206, 2016.

[33] L.-F. Yu, S. K. Yeung, and D. Terzopoulos, "The clutterpalette: An interactive tool for detailing indoor scenes," *IEEE Trans. Visualization & Computer Graphics*, vol. 22, no. 2, pp. 1138–1148, 2016.

[34] M. Li, A. G. Patil, K. Xu, S. Chaudhuri, O. Khan, A. Shamir, C. Tu, B. Chen, D. C. Or, and H. Zhang, "GRAINS: Generative recursive autoencoders for indoor scenes," *ACM Trans. on Graph*, vol. 38, 2019.

[35] R. Ma, A. G. Patil, M. Fisher, M. Li, S. Pirk, B.-S. Hua, S.-K. Yeung, X. Tong, L. J. Guibas, and H. Zhang, "Language-driven synthesis of 3d scenes using scene databases," *ACM Trans. on Graph*, vol. 37, no. 6, 2018.

[36] M. Fisher, M. Savva, Y. Li, P. Hanrahan, and M. Nießner, "Activity-centric scene synthesis for functional 3d scene modeling," *ACM Trans. on Graph (SIGGRAPH Asia)*, vol. 34, no. 6, pp. 179:1–179:13, 2015.

[37] R. Ma, H. Li, C. Zou, Z. Liao, X. Tong, and H. Zhang, "Action-driven 3D indoor scene evolution," *ACM Trans. on Graph (SIGGRAPH Asia)*, vol. 35, no. 6, pp. 173:1–13, 2016.

[38] M. Savva, A. X. Chang, P. Hanrahan, M. Fisher, and M. Nießner, "PiGraphs: Learning Interaction Snapshots from Observations," *ACM Trans. on Graph (SIGGRAPH)*, vol. 35, no. 4, pp. 139:1–12, 2016.

[39] Q. Fu, X. Chen, X. Wang, S. Wen, B. Zhou, and H. Fu, "Adaptive synthesis of indoor scenes via activity-associated object relation graphs," *ACM Trans. on Graph*, vol. 36, no. 6, p. Article 201, 2017.

[40] S. Qi, Y. Zhu, S. Huang, C. Jiang, and S.-C. Zhu, "Human-centric indoor scene synthesis using stochastic grammar," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[41] R. Hu, Z. Yan, J. Zhang, O. van Kaick, A. Shamir, H. Zhang, and H. Huang, "Predictive and generative neural networks for object functionality," *ACM Trans. on Graph (SIGGRAPH)*, vol. 37, no. 4, pp. 151:1–151:13, 2018.

[42] J. Redmon and A. Farhadi, "YOLOv3: an incremental improvement," in *arXiv preprint arXiv:1804.02767*, 2018.

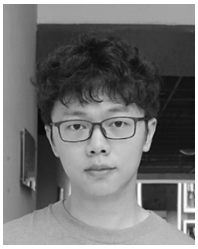
[43] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[44] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky, "The Stanford CoreNLP natural language processing toolkit," in *Association for Computational Linguistics (ACL) System Demonstrations*, 2014, pp. 55–60.

[45] Y. Zhou, Y. Zhu, Q. Ye, Q. Qiu, and J. Jiao, "Weakly supervised instance segmentation using class peak response," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[46] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, 2017, p. 1.

Ruizhen Hu is an Assistant Professor at Shenzhen University, China. She received her Ph.D. from the Department of Mathematics, Zhejiang University. Before that, she spent two years visiting Simon Fraser University, Canada. Her research interests are in shape analysis, geometry processing and fabrication.



Zeyu Huang received the bachelor's degree in software engineering from Shenzhen University in 2019. He is currently working toward the Master degree in Shenzhen University. His research interest include computer graphics.

Haojing Shen received the bachelor's degree in information and computing science from Shenzhen University in 2019. He is currently working toward the Master degree in Shenzhen University. His research interest include machine learning.

Manolis Savva is an Assistant Professor in the School of Computing Science at Simon Fraser University in Vancouver, Canada and a visiting researcher at Facebook AI Research. His research focuses on analysis, organization and generation of 3D content through a human-centric lens of "common sense" semantics. The methods that he works on are stepping stones towards a holistic form of 3D scene understanding revolving around people, with applications in computer graphics, computer vision, and robotics.

Oliver van Kaick is an Assistant Professor at Carleton University, Ottawa, Canada. He received a Ph.D. from the School of Computing Science at Simon Fraser University (SFU). Oliver was then a postdoctoral researcher at SFU and Tel Aviv University. Oliver's research is concen trated in shape analysis and geometric modeling.

Angel Chang is a visiting researcher at Facebook AI Research. She will be starting at Simon Fraser University in Fall 2019 as Assistant Professor. Prior to this, she was a research scientist at Eloquent Labs working on dialogue. She received her Ph.D. in Computer Science from Stanford, where she was part of the Natural Language Processing Group and advised by Chris Manning. She worked on text to 3D scene generation, and the ShapeNet project. In general, She is interested in the semantics of shapes and scenes, the representation and acquisition of common sense knowledge, and reasoning using probabilistic models.

Ariel Shamir is the Dean of Efi Arazi school of Computer Science at the Interdisciplinary Center in Israel. He received a B.Sc. and M.Sc. degrees in math and computer science Cum Laude from the Hebrew University in Jerusalem, and a Ph.D. in computer science in 2000. His research interests include image & video processing, geometric modeling, computer graphics, fabrication, visualization, and machine learning. He is a member of the ACM SIGGRAPH, IEEE Computer, Eurographics, and AsiaGraphics societies.

Hao Zhang received the BMath and MMath degrees from the University of Waterloo, all in computer science, and the PhD degree from the Dynamic Graphics Project (DGP), University of Toronto. He is a full professor in the School of Computing Science, Simon Fraser University (SFU), Canada, where he directs the graphics (GrUVi) lab. His research is in computer graphics with a focus on geometry modeling, shape analysis, 3D content creation, and computational design and fabrication. He is a senior member of the IEEE.

Hui Huang received the PhD degree in computational math from Wuhan University, in 2006, and the another PhD degree in applied math from the University of British Columbia, in 2008. She is a distinguished professor of Shenzhen University, where she directs the Visual Computing Research Center, College of Computer Science and Software Engineering. Her research interests include computer graphics and scientific computing, focusing on point-based modeling, geometric analysis, 3D acquisition, and creation. She is a senior member of the IEEE.