Scene Synthesis from Human Motion

Sifan Ye*
Stanford University
United States of America
sifan.ye@cs.stanford.edu

Yixing Wang*
Stanford University
United States of America
yiw998@stanford.edu

Jiaman Li
Stanford University
United States of America
jiamanli@stanford.edu

Dennis Park
Toyota Research Institute
United States of America
dennis.park@tri.global

C. Karen Liu
Stanford University
United States of America
karenliu@cs.stanford.edu

Huazhe Xu†
Stanford University
United States of America
huazhexu@stanford.edu

Jiajun Wu†
Stanford University
United States of America
jiajunwu@cs.stanford.edu

(a) Human Motion Input (b) Synthesized Scene with Semantic Labels (c) Synthesized Scene with Textures

Figure 1: From a human motion sequence, SUMMON synthesizes physically plausible and semantically reasonable objects.

ABSTRACT

Large-scale capture of human motion with diverse, complex scenes, while immensely useful, is often considered prohibitively costly. Meanwhile, human motion alone contains rich information about the scene they reside in and interact with. For example, a sitting human suggests the existence of a chair, and their leg position further implies the chair's pose. In this paper, we propose to synthesize diverse, semantically reasonable, and physically plausible scenes based on human motion. Our framework, **S**cene Synthesis from H**UM**an **M**oti**ON** (SUMMON), includes two steps. It first uses ContactFormer, our newly introduced contact predictor, to obtain temporally consistent contact labels from human motion. Based on these predictions, SUMMON then chooses interacting objects and optimizes physical plausibility losses; it further populates the scene with objects that do not interact with humans. Experimental results demonstrate that SUMMON synthesizes feasible, plausible, and diverse scenes and has the potential to generate extensive human-scene interaction data for the community.

* and † indicate equal contribution. https://sites.google.com/stanford.edu/summon

SA '22 Conference Papers, December 6–9, 2022, Daegu, Republic of Korea

ACM ISBN 978-1-4503-9470-3/22/12.
https://doi.org/10.1145/3550469.3555426

CCS CONCEPTS

• **Computing methodologies** → **Motion processing**; **Shape inference**; **Scene understanding**.

KEYWORDS

Scene synthesis, motion analysis, activity understanding

ACM Reference Format:
Sifan Ye, Yixing Wang, Jiaman Li, Dennis Park, C. Karen Liu, Huazhe Xu, and Jiajun Wu. 2022. Scene Synthesis from Human Motion. In *SIGGRAPH Asia 2022 Conference Papers (SA '22 Conference Papers), December 6–9, 2022, Daegu, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3550469.3555426

1 INTRODUCTION

Capturing, modeling, and synthesizing realistic human motion in 3D scenes is crucial in a spectrum of applications such as virtual reality, game character animation, and human-robot interaction. To facilitate research in this area, a plethora of datasets [Hassan et al. 2019; Mahmood et al. 2019; Savva et al. 2016] have been curated to capture human motion. For example, Bhatnagar et al. [2022] collected trajectories of humans manipulating objects. The PROX-E dataset [Zhang et al. 2020a] contains human contact with a scene mesh. However, building high-quality, large-scale datasets annotated with both diverse human motions and rich 3D scenes remains challenging. This is mainly because current data capture pipelines depend on costly devices, such as MoCap systems, structure cameras, and 3D scanners, and therefore can only be conducted in laboratory settings, which entails limited physical space and scene

diversity. Inspired by recent advances in modeling 3D human poses and their contact with environments, we aim to address these challenges by exploring a new possibility: *can we learn to synthesize the scenes only from human motion?* If successful, our system will also have many potential applications beyond data collection, such as providing suggestions during the creation of virtual environments based on artists' motions in VR.

Recent works have proposed to estimate room layouts based on human trajectories and learned room priors [Nie et al. 2022]. However, only semantics, not affordances, was considered in the reconstructed layouts. Yi et al. [2022] proposed to reconstruct scene objects from visual inputs and then use Human-Scene Interactions (HSIs) to further improve the feasibility. While such a method produces physically plausible reconstructions, it requires additional visual inputs so that the reconstructed scenes are restricted.

We propose **S**cene Synthesis from H**UM**an **M**oti**ON** (SUMMON), a method that predicts feasible object placements in a scene based solely on 3D human pose trajectories, as shown in Figure 1. SUMMON consists of two modules: a human-scene contact prediction module and a scene synthesis module. The human-scene contact prediction module, named ContactFormer, leverages existing HSI data to learn a mapping from human body vertices to the semantic label of the objects that are in contact. ContactFormer advances previous methods [Hassan et al. 2021b] by incorporating temporal cues to enhance the consistency in label prediction in time. Given the estimated semantic contact points, the scene synthesis module first searches for objects that fit the contact points in terms of semantics and physical affordances to the agent; it then populates the scene with other objects that have no contact with humans, based on human motion and objects inferred from previous steps.

We conduct our experiments using the PROXD [Hassan et al. 2019] and the GIMO [Zheng et al. 2022] datasets. In terms of contact estimation, ContactFormer outperforms previous single-frame contact prediction methods [Hassan et al. 2021b]. In terms of scene synthesis, our proposed system shows more realistic, physically plausible, and diverse scenes than baselines, using various metrics and human evaluation.

Our contributions are threefold. First, we propose SUMMON, a system that synthesizes semantically reasonable, physically plausible, and diverse scenes based only on human motion trajectories. Second, as a part of SUMMON, we propose a contact prediction module ContactFormer that outperforms existing methods by modeling the temporal consistency in semantic labels. Third, we demonstrate that the scenes synthesized by SUMMON consistently outperform existing methods both qualitatively and quantitatively.

2 RELATED WORKS

Scene affordance learning. Learning affordance from human-scene interaction has caught much attention recently [Chen et al. 2019; Chuang et al. 2018; Delaitre et al. 2012; Fouhey et al. 2012; Gupta et al. 2011; Wang et al. 2019a; Zhu et al. 2014]. In the literature, researchers study how to put human skeletons in a scene. For example, Wang et al. [2017] proposed to learn the affordance from sitcom videos for positioning skeletons in a static image. Li et al. [2019a] introduced a generative model of 3D poses to predict plausible human poses in a scene. Along with developing better human body representations, there have been methods that try to put a 3D human body into the scene [Zhang et al. 2020a]. More recently, POSA [Hassan et al. 2021b] learns a model that augments a SMPL-X human body model vertices with contact probability and semantic labels to place human poses in a 3D scene mesh. Blinn et al. [2021] proposed a fitting and comfort-based loss to train an affordance-aware model to generate chairs that fit a human body pose. Several works also try to collect or generate data that involve human-scene interactions. For example, VirtualHome [Puig et al. 2018] provides a simulated 3D environment where humanoid agents can interact with 3D objects. BEHAVE [Bhatnagar et al. 2022] provides a dataset of real full-body human parameterized using SMPL interacting and manipulating objects in 3D with contact points. Our work takes an additional step from the affordance learning works: we first learn to understand the affordance, then leverage them to synthesize scenes that can be used for other related tasks.

Human motion synthesis. Motion synthesis is a long-standing problem in computer graphics and vision [Brand and Hertzmann 2000; Holden et al. 2016; Kovar and Gleicher 2003; Park et al. 2002; Spallone 2015]. Xu et al. [2020] proposed a hierarchical way to generate long-horizon motion by using a memory bank to retrieve short-horizon reference clips. Harvey et al. [2020] proposed to predict motion robustly with additional embeddings. Recently, many works also take the environment into consideration [Hassan et al. 2021a; Rempe et al. 2021; Wang et al. 2021a]. For example, Wang et al. [2021a] combined long-term human motion synthesis conditioned on a scene mesh with affordance optimization to generate realistic human trajectories. SAMP [Hassan et al. 2021a] learns generalized interaction for object classes across different instances of that class. Our work is trying to solve the inverse problem that generates plausible scenes given human motion trajectories.

Scene synthesis. Our work is also closely related to synthesizing plausible 3D scenes and room layout [Li et al. 2019b; Luo et al. 2020; Purkait et al. 2020; Ritchie et al. 2019; Wang et al. 2019b, 2021b; Zhang et al. 2020b; Zhou et al. 2019]. For example, ATISS [Paschalidou et al. 2021] learns an autoregressive generative model for furniture placement. It can be used for generating plausible novel room layouts, completing a scene given existing objects, and suggesting possible placements given spatial constraints. Another work, Pose2Room [Nie et al. 2022], predicts bounding boxes of objects conditioned on 3D human pose trajectory. MOVER [Yi et al. 2022] reconstructs 3D objects constrained by 3D human body predictions from monocular RGB videos. Unlike these prior methods, our model generates not only layouts but also affordable objects with only human trajectories.

3 METHOD

We aim to predict a set of furniture objects and a physically plausible 3D configuration of them only from human motion sequences. We first introduce the human body and contact representation in Sec. 3.1. SUMMON generates a temporally consistent contact semantic estimation for each vertex of the human body to retrieve suitable objects (Sec. 3.2). Then we optimize object placement based on the contact locations and physical plausibility (Sec. 3.3). An illustration of our method is shown in Figure 2.

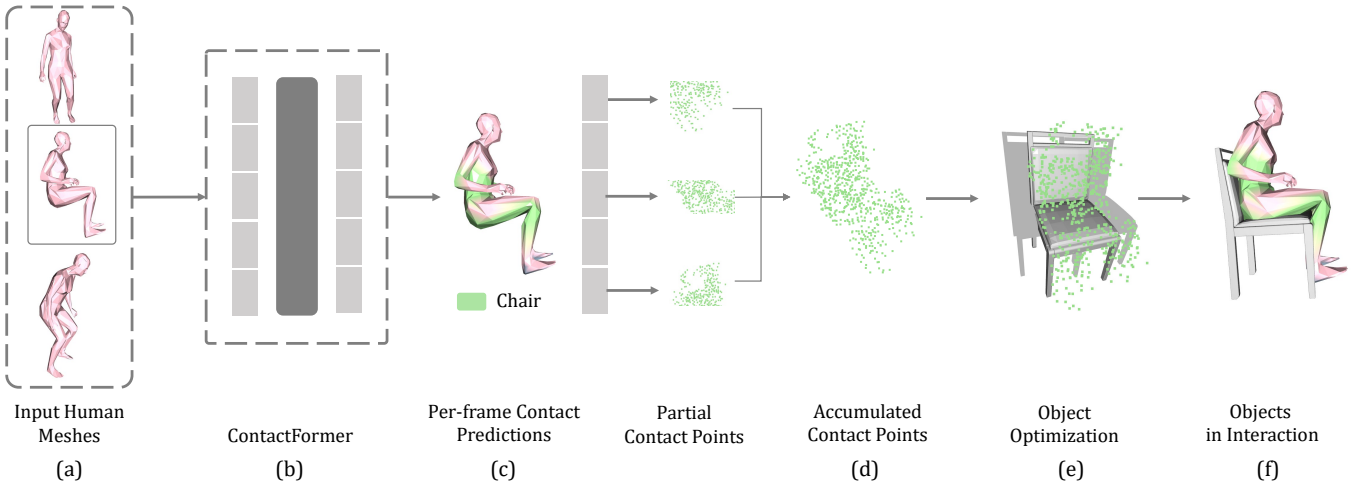

Figure 2: The overview of SUMMON: (a) an input sequence of human body meshes interacting with a scene, (b) the ContactFormer that predicts per-frame contact labels, (c) per-frame contact predictions, (d) estimated contact points, (e) synthesized objects, and (f) objects in interaction.

3.1 Human Body and Contact Representation

We use a modified version of SMPL-X [Pavlakos et al. 2019] as the representation of human body poses. Specifically, we parameterize the human body with $M(\theta, \beta) : \mathbb{R}^{|\theta|\times|\beta|} \rightarrow \mathbb{R}^{3N}$, where θ denotes pose parameters, β denotes coefficients in a learned shape space, and N is the number of vertices in a SMPL-X body mesh. For computation efficiency, we downsample the vertices from 10,475 to 655 points, following the prior work by Hassan et al. [2021b].

We represent contact information by per-vertex features. For each vertex $v_b \in V_b$, where V_b is all vertices of a human body, we use a one-hot vector f to represent the contact semantic label for that vertex. Each vector f has a length of $|f| = C + 1$, where C is the number of object classes. We introduce an extra "void" class to represent vertices without contact. We use F to denote the contact semantic labels for all vertices in a body pose.

3.2 Human-Scene Contact Prediction

Our dataset consists of a sequence of paired vertices and contact semantic labels $\{(V_b^1, F^1), (V_b^2, F^2), ..., (V_b^n, F^n)\}$, where V_b^i represents the human body vertices (Figure 2(a)), F^i represents the contact semantic labels for frame i, and n is the varied sequence length. We first train a conditional Variational Autoencoder (cVAE) to learn a probabilistic model of contact semantic labels conditioned on vertex positions. Then we deploy transformer layers on top of the cVAE to improve temporal consistency. We refer to this framework as ContactFormer. An illustration of the overall network architecture is shown in Figure 3.

Contact semantics prediction. We first train a model to predict contact semantic labels for each individual pose. Given a pair of body vertices and contact semantic labels (V_b, F), we first fuse these two components: $I_e = \text{Concat}(V_b, F)$. We feed I_e into a graph neural network (GNN) encoder G_{Enc} to get a latent Gaussian space with the mean H_μ and the standard deviation H_σ. Then we sample a latent vector z from the latent Gaussian space and concatenate it with each vertex position: $I_d = \text{Concat}(V_b, z)$. We feed I_d into a

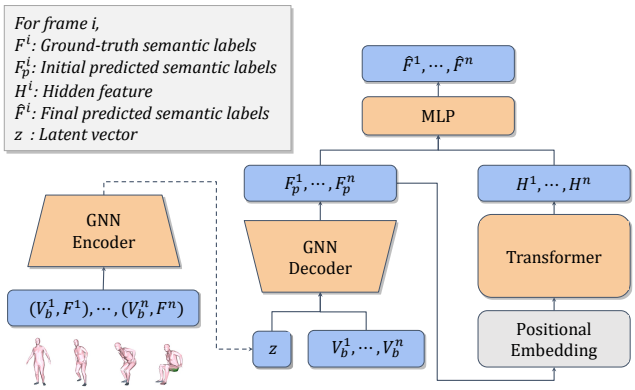

Figure 3: The architecture of ContactFormer. We first use a GNN-based variational autoencoder to encode the contact points. Then a transformer is applied to improve the temporal information fusion. We also add a sinusoidal positional encoding to the output of the GNN decoder.

GNN decoder G_{Dec} to predict the reconstructed contact semantic labels F_p. Note that both GNNs in the encoder and the decoder share the same structure as in Hassan et al. [2021b]. Each vertex feature h_x^k for vertex x at layer k is updated by

$$h_x^k = \text{Linear}(\text{Concat}(\{h_{x'}^{k-1} : x' \in N(x)\})), \quad (1)$$

where $N(x)$ is defined as the m-nearest neighbor vertices of x in a spiral-ordered sequence, as proposed by Gong et al. [2019].

ContactFormer: We train a transformer to extract temporal information from a pose sequence to enhance prediction consistency, as shown in Figure 3. Specifically, given a sequence of pose and contact semantic labels $\{(V_b^1, F^1), ..., (V_b^n, F^n)\}$ from frame 1 to n, we first use the previous model to reconstruct contact semantic labels F_p^i independently for each frame i. We then embed each F_p^i into a hidden feature space, augmenting it with a sinusoidal positional

embedding before feeding it to the transformer module. The output of the transformer module is a sequence of n vectors $\{H_1, ..., H_n\}$. For each frame i, we concatenate H_i with the initial prediction F_p^i and use a multi-layer perceptron (MLP) to get a final prediction $\hat{F}^i$. The final prediction is shown in Figure 2(c).

Training: We optimize the model's parameters by the following loss function:

$$\mathcal{L} = \mathcal{L}_{rec} + \alpha \cdot \mathcal{L}_{KL}, \tag{2}$$

where $\mathcal{L}_{rec}$ is the sum of the categorical cross entropy (CCE) loss between the ground truth semantic label F^i and the model prediction $\hat{F}^i$ for any frame i:

$$\mathcal{L}_{rec} = \sum_i \text{CCE}(F^i, \hat{F}^i), \tag{3}$$

and $\mathcal{L}_{KL}$ is the Kullback-Leibler divergence loss between the latent Gaussian space and the normal distribution $\mathcal{N}$:

$$\mathcal{L}_{KL} = KL(Q(z|F, V_b)||\mathcal{N}). \tag{4}$$

Here we use Q to represent the encoder network in our cVAE combined with the sampling process with the reparameterization trick. Inspired by Higgins et al. [2016], we also multiply $\mathcal{L}_{KL}$ with a weight α to control the balance between the reconstruction accuracy and diversity.

3.3 Scene Synthesis

Contact Object Recovery. Given the accumulated contact points from each frame predicted by ContactFormer (Figure 2(d)), we further reduce spatial prediction noise by performing a local object class majority voting as shown in Figure 4. Then, the vertices of each predicted object class are clustered into possible contact instances V_c, using the shortest length of all object edges in that class as ϵ for clustering. In practice, we downsample the contact vertices to keep later computations tractable.

We then optimize the poses of the object point cloud V_o by minimizing the following losses:

$$\mathcal{L}(V_c, V_o) = \mathcal{L}_{contact} + \mathcal{L}_{pen}. \tag{5}$$

The contact loss $\mathcal{L}_{contact}$ is defined as

$$\mathcal{L}_{contact} = \lambda_{contact} \frac{1}{|V_c|} \sum_{v_c \in V_c} \min_{v_o \in V_o} ||v_c - v_o||_2^2, \tag{6}$$

where $\lambda_{contact}$ is a tunable hyperparameter. This loss encourages the object to be in contact with the predicted human vertices. The penetration loss $\mathcal{L}_{pen}$ is defined as:

$$\mathcal{L}_{pen} = \lambda_{pen} \sum_{d_c^i < t} {d_c^i}^2, \tag{7}$$

where d_c^i are signed distances between the object and the human body sequence, t is the penetration distance threshold. This loss prevents the object from penetrating the human body sequence.

Intuitively, these losses encourage objects to be in contact with human meshes, but not penetrate them. An illustration of the optimized object placement is shown in Figure 2(e). To improve computation efficiency, we choose to compute human SDF from merged human meshes of the motion sequence. To have a consistent scale of loss across different objects, we choose the number of sampled points according to the size of the object.

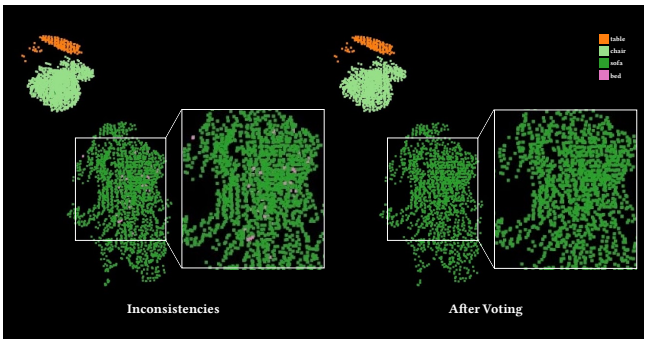

Figure 4: Illustration of the local majority voting. From the zoomed-in box, there are multiple inconsistent points in the original contact points. The pink points represent the semantic label bed, and the green points represent the label sofa. We alleviate this issue by adding majority voting.

Constrained Scene Completion. To obtain a complete scene, we also predict non-contact objects as a scene completion task constrained by 3D human trajectories and existing in-contact objects. The floor is divided into a grid, and each cell is labeled as occupied if feet vertices or object vertices are in close proximity. Considering the furniture categories in a room as a sequence, we train an autoregressive transformer model on the 3D-FRONT dataset [Fu et al. 2021a]. The model takes as input the categories of existing objects and returns a probability distribution of the next potential object category. We sample a category from the distribution and randomly place an object of that category onto an unoccupied floor grid. To prevent the sampled object from penetrating the human body sequence, we further optimize the object's translation and rotation using our $\mathcal{L}_{pen}$ (see Equation 7).

4 EXPERIMENT SETUP

In this section, we introduce the datasets and implementation details for the whole SUMMON framework.

4.1 Datasets

We use the PROXD [Hassan et al. 2019] dataset for training our ContactFormer. PROXD uses RGB-D cameras to capture 20 human subjects interacting with 12 scenes. We represent human poses using the SMPL-X format to reconstruct human body meshes. The pose sequences in PROXD are estimated using SMPLify [Bogo et al. 2016] and contain many jitters. We apply LEMO [Zhang et al. 2021], a learned temporal motion smoothness prior, to produce smooth human motion as training data. Our ground truth per vertex contact semantic labels are generated using scene SDF with contact semantic labels from PROX-E [Zhang et al. 2020a], which extends PROXD by manually annotating the scene meshes with predefined object categories. We define human-scene contact as the signed distance between a human vertex and the scene to be less than 0.05.

We select objects from 3D-FUTURE [Fu et al. 2021b] to be placed into the scenes. 3D-FUTURE is a dataset of categorized 3D models of furniture with their original sizes. We use a selected subset of 3D-FUTURE to reduce candidate search time. To simplify contact estimation and limit predicted object classes to the available ones

in our object dataset, we reduce the contact object categories in the PROX-E dataset from 42 to 8.

We use the GIMO dataset [Zheng et al. 2022] as another test dataset for evaluating the generalization ability of the proposed method on out-of-distribution data.

4.2 Implementation

ContactFormer. For the encoder and decoder GNNs, we choose the number of hidden layers to be 3. The dimension for each hidden vertex feature in the GNNs is 64. In the GNN encoder, we downsample the body vertices after each hidden layer by a factor of 4. We deploy a similar architecture for the transformer layers as used in the previous work [Vaswani et al. 2017]. We provide training details and hyperparameter choices in the supplementary materials.

To compare different architectures' capacities for extracting temporal information, we also implement models that use MLP and LSTM [Greff et al. 2016] modules as the final block on top of the GNN decoder. For the model that uses the MLP module, we deploy a max pooling layer to the output of the GNN decoder along the dimension of vertices. Then we feed it to an MLP block to get the embedding for the whole sequence. The sequence embedding is then fused with the output of the GNN decoder to get the final prediction via a linear projection. For the model that uses the LSTM module, we linearly project the outputs from the GNN decoder into a higher dimensional embedding space and feed them to a bidirectional LSTM layer to extract features for each frame. Frame features are then concatenated with the output from the GNN decoder to obtain final semantic labels.

Contact Object Recovery. To reduce noise in contact semantic estimation, we use majority semantic voting in point cloud clusters with $\epsilon = 0.1$ and $minPts = 10$. In point cloud clustering for object instance fitting, we used different values for ϵ for different classes due to their different sizes.

To place objects into the scene at an appropriate height, we first cluster all the human body vertices that are in contact with the floor. We then take the minimum medians of all clusters as the estimated floor height. Next, we translate the object to place its lowest vertex on the floor.

To avoid local minima, we perform a grid search for translation along the floor plane and rotation around the up axis to warm-start the initial transformation. We then optimize for the same transformation parameters on top of the results from the grid search. In both cases, We use different $\lambda_{contact}$, λ_{pen} and t to accommodate for different properties of object classes. We keep the transformation that achieves the lowest loss as the optimization result.

To achieve scene diversity, we consider inter-class and intra-class diversity. Inter-class diversity is when a human motion is likely to interact with different classes of objects. For example, sitting down can be performed on a chair, a bed, or a sofa. To achieve this, we first sample per-vertex contact semantics based on the contact probability distribution predicted by ContactFormer. During local clustering of contact object recovery, we consider class labels in local clusters as a probability distribution and sample the cluster contact class. Intra-class diversity is when a human motion is likely to interact with different instances of the same object class. To achieve this, we perform grid search and optimization on all the instances from the object class.

Table 1: Results of contact prediction. We use the reconstruction accuracy and the consistency score as metrics. Our ContactFormer clearly outperforms the baselines.

Models	Reconstruction Acc. ↑	Consistency Score ↑
MLP Predictor	0.9082	0.8922
LSTM Predictor	0.9087	0.9209
POSA	0.9106	0.8816
ContactFormer (ours)	**0.9120**	**0.9518**

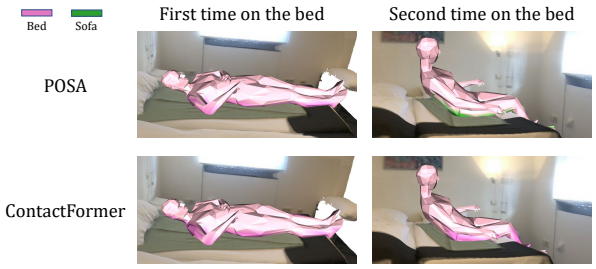

Figure 5: Visualizations of the contact prediction results of POSA and our ContactFormer. Left: Contact predictions from POSA and ContactFormer when the person lies on the bed. Right: Contact predictions from POSA and ContactFormer when the person lies on the bed again after walking around. ContactFormer has better consistency when the person lies in bed for the second time.

5 EVALUATIONS

In this section, we introduce evaluation metrics, baselines, and results on contact prediction and scene synthesis. We encourage the readers to watch the video in the supplementary materials.

5.1 Contact Semantic Prediction

Baselines. We compare with three baselines, including POSA [Hassan et al. 2021b], an architectural variant that uses a multi-layer perceptron (MLP) based predictor, and a temporal information fusion variant that uses a bidirectional LSTM [Greff et al. 2016].

Metrics. We use two metrics for evaluating the contact semantic prediction: reconstruction accuracy and consistency score. The reconstruction accuracy is computed as the average correctness of the predicted label compared with the ground-truth label for each vertex. The consistency score is designed following this intuition: if we accumulate predicted contact points from each frame, close contact points should have consistent contact semantic labels. Hence, this loss is computed as follows: Given a pose sequence and the accumulated contact points, for each point, we compare its predicted contact label with the contact labels of its neighboring points to see if the prediction agrees with the majority of the neighboring contact labels (i.e., a high consistency score).

Results. Table 1 shows the reconstruction accuracy and the consistency score of all methods on the validation set of PROXD. We find that ContactFormer achieves competitive performance in terms

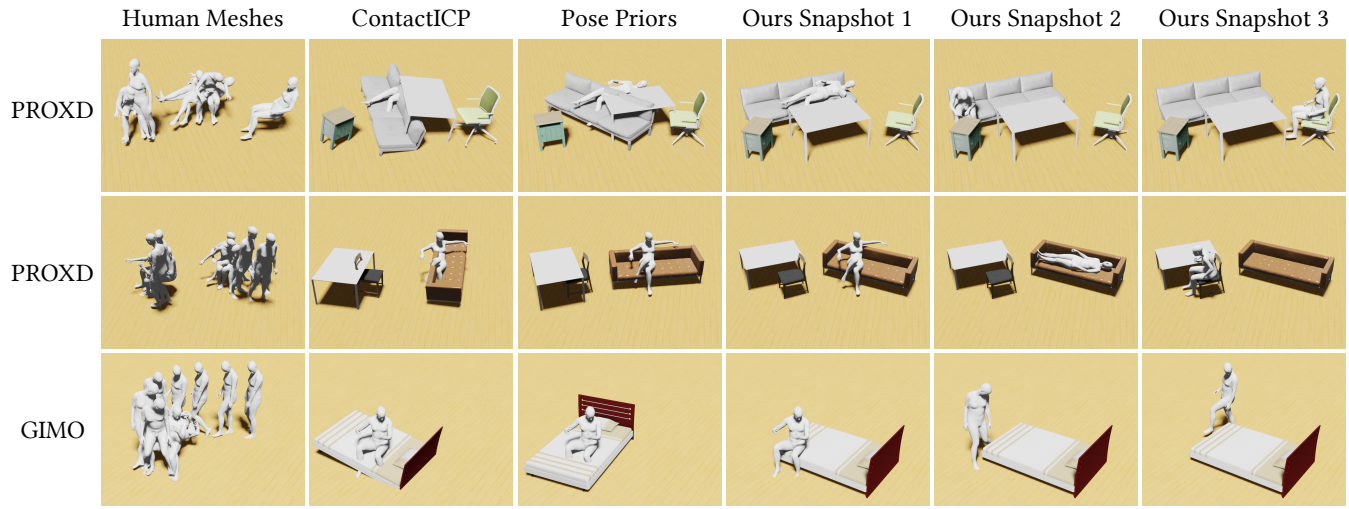

Figure 6: Visualizations of the generated objects based on human trajectories. The human trajectories are from the PROXD dataset and the unseen GIMO dataset. The first column shows the trajectory. The second column shows the results from the ContactICP baseline. The third column shows the results from the Pose Prior baseline. The fourth to sixth columns are snapshots of results generated by SUMMON.

of reconstruction accuracies and significantly outperforms all the baselines in consistency scores. This demonstrates the superiority of the transformer-based architecture in predicting temporally consistent yet accurate contact labels.

Figure 5 visualizes the output contact labels from ContactFormer and POSA. We notice that ContactFormer predicts consistent labels the second time the human tries to lie on the bed, while POSA, due to its lack of temporal information, predicts a different label.

User study. We conduct a user study to evaluate the quality of the contact semantic label predictions, where we compare ContactFormer with POSA. For each pose sequence in the validation dataset, we render a video showing the human motion, predicted contact semantic labels, and the ground truth scene. We show the predicted contact semantic labels by rendering small areas around body vertices in different colors depending on their labels. Each video is rendered from a camera angle that can clearly capture human motion and semantic labels. For each pose sequence, we ask the human subjects the following question: "Which video seems to have a more reasonable contact label prediction?". Among 22 users, 78.12% of the users choose ContactFormer over POSA, believing ContactFormer provides more reasonable and convincing results. This result echoes the quantitative results in Table 1.

5.2 Contact Object Recovery

Baselines. Since our problem is novel and there are no baselines, we devise two reasonable baselines ourselves: contact-informed point cloud registration (ContactICP) based on point-to-point ICP [Besl and McKay 1992] and object alignment with pose priors (Pose Priors) based on the orientation of the hip. We provide the details of those methods in the supplementary materials.

Metrics. We use the *non-collision score* proposed by Zhang et al. [2020a]. This score estimates the collision ratio between human body mesh and scene objects. Since all the methods, including SUMMON, first align the object to the centroid of contact points, contact constraints are naturally satisfied.

Table 2: Non-collision scores for contact object recovery on the smoothed PROXD and the unseen GIMO dataset. For each sequence, the score is computed to be the mean of all possible generated scenes. Higher scores are better.

Method	PROXD	GIMO
ContactICP	0.654	0.820
Pose Priors	0.703	0.798
SUMMON w/o optimization	0.815	0.937
SUMMON (ours)	**0.851**	**0.951**

Results. For each sequence, we compute the mean of the *non-collision scores* for all the objects in the scene. In Table 2, we compare the mean non-collision scores on the smoothed PROXD dataset [Zhang et al. 2021], which was used during training, and the unseen GIMO dataset [Zheng et al. 2022], which also provides SMPL-X parameters for humans interacting with scenes.

We visualize comparisons between our method and the baselines in Figure 6. We find that SUMMON can synthesize objects that are physically plausible and semantically reasonable. ContactICP usually suffers from large penetrations because the contact points might be sparse for registration. While Pose Priors can have seemingly correct object locations and orientations, it often fails to consider physical constraints.

Figure 7 demonstrates various possible scenes generated from the same human motion trajectory by SUMMON. We find that SUMMON can generalize intra-class (e.g., chairs with different appearances) and inter-class (e.g., sofa to a bed). We provide additional examples in the supplementary materials.

Human user study. We follow the same procedure as in Section 5.1. Instead of contact prediction, we present the users with

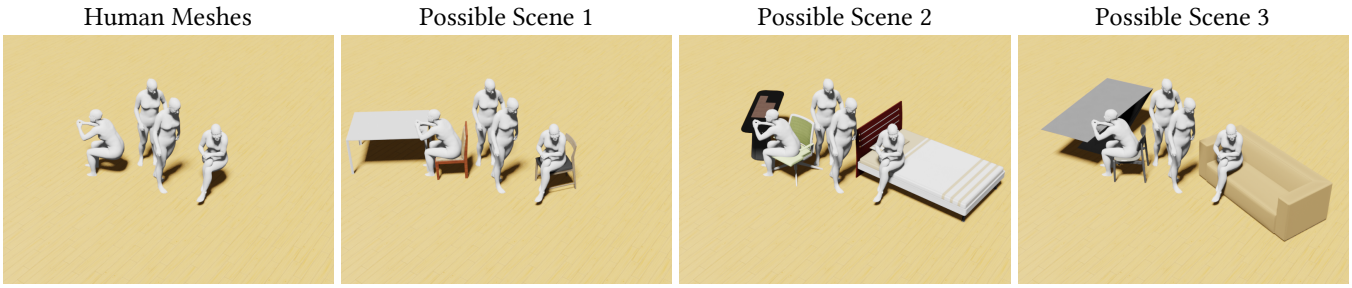

Figure 7: Visualizations of possible alternative object placements generated by SUMMON based on the same human trajectories. In this example, an in-contact object can be a chair, a sofa, or a bed, as long as it does not violate physical constraints. SUMMON can also generate different instances (e.g., chairs) within the same category.

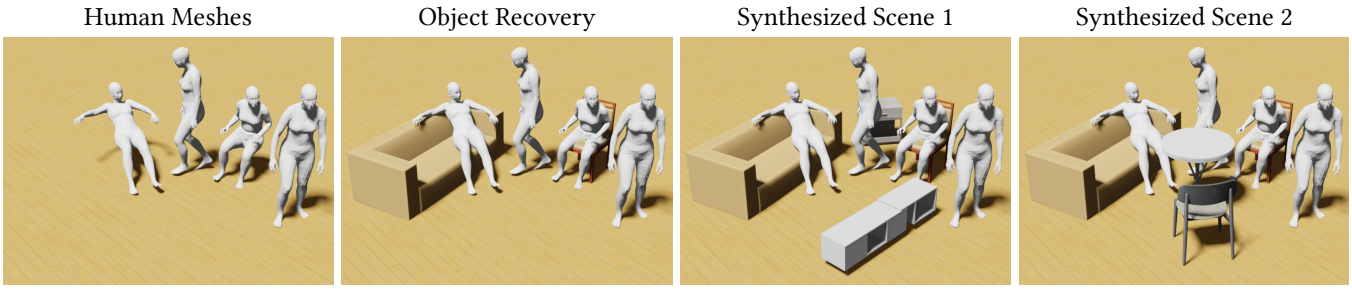

Figure 8: Visualizations of scene completion. Based on all the in-contact objects and human motion trajectories, SUMMON now generates the objects that are not in contact with human meshes. While there is no contact, it makes the scene more complete and introduces the potential for future synthesized human motion sequences to interact with additional objects.

the animated human motion sequences and the predicted objects in the scene, and ask them to choose the most plausible and realistic placement. From the statistics, we find that 74.5% of the users select SUMMON over ContactICP and Pose Priors. We find that Pose Priors has a 23.5% user selection rate, showing that it can produce reasonable results in some cases.

Ablations. We also perform ablation on the optimization objectives. Table 3 shows that both the penetration loss and the contact loss are important for SUMMON. Intuitively, the penetration loss helps the object to avoid a collision, while the contact loss helps to keep the object close to humans. We use both the *non-collision score* and the *contact score*. The *contact score* is computed as the fraction of objects in the scenes that are in contact with the human trajectory Zhang et al. [2020a].

5.3 Scene completion

To generate a full-fledged scene, we train another object generation model following Paschalidou et al. [2021] as in Section 3.3. The model outputs a family of possible objects that does not contact or penetrate human meshes. Using this model, we generate a fuller scene with both in-contact and no-contact objects. Visualized results are in Figure 8. The completed scenes have additional objects, such as a TV stand or a coffee table. While there is no contact between these objects and the human meshes, they make the scene semantically more realistic.

Table 3: Ablation study on the losses. The penetration loss and the contact loss are ablated. We use the non-collision score and the contact score as metrics.

Method	non-collision score ↑	contact score ↑
SUMMON	0.894	1
w/o penetration loss	0.656	1
w/o contact loss	0.995	0.194

6 CONCLUSION

We propose **S**cene Synthesis from H**UM**an **M**oti**ON** (SUMMON), a framework that generates multi-object scenes from a sequence of human interaction. SUMMON leverages human contact estimations and scene priors to produce scenes that realistically support the interaction and the semantic context. The flexibility of SUMMON also enables the synthesis of diverse scenes from a single motion sequence. We hope this can also shed light on generating inexpensive diverse human-scene interaction datasets. In the future, we are interested in exploring the following directions. Since PROXD does not consider soft-body interactions, a potential direction would be considering soft-body deformation of objects such as beds and sofas. Our method considers synthesized scenes to be stationary, hence future works can include movement and rearrangement of furniture during human-scene interaction. As PROXD categorizes all the smaller interaction objects such as books, cups, or TV remotes into a single category, one potential extension to our method would be to include interactions with more specific small objects.

ACKNOWLEDGMENTS

This work is in part supported by the Stanford Human-Centered AI Institute (HAI), the Toyota Research Institute (TRI), Innopeak, Meta, Bosch, and Samsung.

REFERENCES

P.J. Besl and Neil D. McKay. 1992. A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14, 2 (1992), 239–256.

Bharat Lal Bhatnagar, Xianghui Xie, Ilya A. Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. 2022. BEHAVE: Dataset and Method for Tracking Human Object Interactions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 15935–15946.

Bryce Blinn, Alexander Ding, Daniel Ritchie, R Kenny Jones, Srinath Sridhar, and Manolis Savva. 2021. Learning Body-Aware 3D Shape Generative Models. *arXiv preprint arXiv:2112.07022* (2021).

Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J. Black. 2016. Keep it SMPL: Automatic Estimation of 3D Human Pose and Shape from a Single Image. In *European Conference on Computer Vision (ECCV)*. 561–578.

Matthew Brand and Aaron Hertzmann. 2000. Style machines. In *Conference on Computer Graphics and Interactive Techniques*. 183–192.

Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. 2019. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *International Conference on Computer Vision (ICCV)*. 8648–8657.

Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. 2018. Learning to act properly: Predicting and explaining affordances from images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 975–983.

Vincent Delaitre, David F Fouhey, Ivan Laptev, Josef Sivic, Abhinav Gupta, and Alexei A Efros. 2012. Scene semantics from long-term observation of people. In *European Conference on Computer Vision (ECCV)*. 284–298.

David F Fouhey, Vincent Delaitre, Abhinav Gupta, Alexei A Efros, Ivan Laptev, and Josef Sivic. 2012. People watching: Human actions as a cue for single view geometry. In *European Conference on Computer Vision (ECCV)*. 732–745.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 2021a. 3d-front: 3d furnished rooms with layouts and semantics. In *International Conference on Computer Vision (ICCV)*. 10933–10942.

Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 2021b. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision (IJCV)* 129, 12 (2021), 3313–3337.

Shunwang Gong, Lei Chen, Michael Bronstein, and Stefanos Zafeiriou. 2019. Spiralnet++: A fast and highly efficient mesh convolution operator. In *International Conference on Computer Vision Workshops (ICCVW)*. 4141–4148.

Klaus Greff, Rupesh K Srivastava, Jan Koutník, Bas R Steunebrink, and Jürgen Schmidhuber. 2016. LSTM: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems* 28, 10 (2016), 2222–2232.

Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. 2011. From 3d scene geometry to human workspace. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 1961–1968.

Félix G Harvey, Mike Yurick, Derek Nowrouzezahrai, and Christopher Pal. 2020. Robust motion in-betweening. *ACM Transactions on Graphics (TOG)* 39, 4 (2020), 60–1.

Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. 2021a. Stochastic Scene-Aware Motion Prediction. In *International Conference on Computer Vision (ICCV)*. 11354–11364.

Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. 2019. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV)*. 2282–2292.

Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. 2021b. Populating 3D Scenes by Learning Human-Scene Interaction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 14708–14718.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2016. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representation (ICLR)*.

Daniel Holden, Jun Saito, and Taku Komura. 2016. A deep learning framework for character motion synthesis and editing. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–11.

Lucas Kovar and Michael Gleicher. 2003. Flexible automatic motion blending with registration curves. In *Symposium on Computer Animation (SCA)*. 214–224.

Manyi Li, Akshay Gadi Patil, Kai Xu, Siddhartha Chaudhuri, Owais Khan, Ariel Shamir, Changhe Tu, Baoquan Chen, Daniel Cohen-Or, and Hao Zhang. 2019b. Grains: Generative recursive autoencoders for indoor scenes. *ACM Transactions on Graphics (TOG)* 38, 2 (2019), 1–16.

Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. 2019a. Putting Humans in a Scene: Learning Affordance in 3D Indoor Environments. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 12360–12368.

Andrew Luo, Zhoutong Zhang, Jiajun Wu, and Joshua B Tenenbaum. 2020. End-to-end optimization of scene layout. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3754–3763.

Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. 2019. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*. 5442–5451.

Yinyu Nie, Angela Dai, Xiaoguang Han, and Matthias Nießner. 2022. Pose2Room: Understanding 3D Scenes from Human Activities. In *European Conference on Computer Vision (ECCV)*.

Sang Il Park, Hyun Joon Shin, and Sung Yong Shin. 2002. On-line locomotion generation based on motion blending. In *Symposium on Computer Animation (SCA)*. 105–111.

Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. 2021. ATISS: Autoregressive Transformers for Indoor Scene Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*. 12013–12026.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985.

Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. 2018. Virtualhome: Simulating household activities via programs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 8494–8502.

Pulak Purkait, Christopher Zach, and Ian Reid. 2020. Sg-vae: Scene grammar variational autoencoder to generate new indoor scenes. In *European Conference on Computer Vision (ECCV)*. 155–171.

Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. 2021. Humor: 3d human motion model for robust pose estimation. In *International Conference on Computer Vision (ICCV)*. 11488–11499.

Daniel Ritchie, Kai Wang, and Yu-an Lin. 2019. Fast and flexible indoor scene synthesis via deep convolutional generative models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 6182–6190.

Manolis Savva, Angel X Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. 2016. Pigraphs: learning interaction snapshots from observations. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 1–12.

Roberta Spallone. 2015. Digital reconstruction of demolished architectural masterpieces, 3D modeling, and animation: the case study of Turin Horse Racing by Mollino. *Handbook of research on emerging digital tools for architectural surveying, modeling, and representation* (2015), 476–509.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*. 5998–6008.

Jiashun Wang, Huazhe Xu, Jingwei Xu, Sifei Liu, and Xiaolong Wang. 2021a. Synthesizing long-term 3d human motion and interaction in 3d scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 9401–9411.

Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X Chang, and Daniel Ritchie. 2019b. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–15.

Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. 2017. Binge watching: Scaling affordance learning from sitcoms. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2596–2605.

Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. 2021b. Sceneformer: Indoor scene generation with transformers. In *International Conference on 3D Vision (3DV)*. 106–115.

Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless Fowlkes. 2019a. Geometric pose affordance: 3d human pose with scene constraints. *arXiv preprint arXiv:1905.07718* (2019).

Jingwei Xu, Huazhe Xu, Bingbing Ni, Xiaokang Yang, Xiaolong Wang, and Trevor Darrell. 2020. Hierarchical Style-based Networks for Motion Synthesis. In *European Conference on Computer Vision (ECCV)*. 178–194.

Hongwei Yi, Chun-Hao P Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J Black. 2022. Human-aware object placement for visual environment reconstruction. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 3959–3970.

Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. 2021. Learning motion priors for 4d human body capture in 3d scenes. In *International Conference on Computer Vision (ICCV)*. 11343–11353.

Song-Hai Zhang, Shao-Kui Zhang, Wei-Yu Xie, Cheng-Yang Luo, and Hong-Bo Fu. 2020b. Fast 3d indoor scene synthesis with discrete and exact layout pattern extraction. *arXiv preprint arXiv:2002.00328* (2020).

Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. 2020a. Generating 3d people in scenes without people. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 6194–6204.

Yang Zheng, Yanchao Yang, Kaichun Mo, Jiaman Li, Tao Yu, Yebin Liu, Karen Liu, and Leonidas Guibas. 2022. GIMO: Gaze-Informed Human Motion Prediction in Context. In *European Conference on Computer Vision (ECCV)*.

Yang Zhou, Zachary While, and Evangelos Kalogerakis. 2019. Scenegraphnet: Neural message passing for 3d indoor scene augmentation. In *International Conference on Computer Vision (ICCV)*. 7384–7392.

Yuke Zhu, Alireza Fathi, and Li Fei-Fei. 2014. Reasoning about object affordances in a knowledge base representation. In *European Conference on Computer Vision (ECCV)*. 408–424.