MIME: Human-Aware 3D Scene Generation

Hongwei Yi[1] Chun-Hao P. Huang[2*] Shashank Tripathi[1] Lea Hering[1] Justus Thies[1] Michael J. Black[1]
[1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]Adobe Inc.
{firstname.lastname}@tuebingen.mpg.de chunhaoh@adobe.com

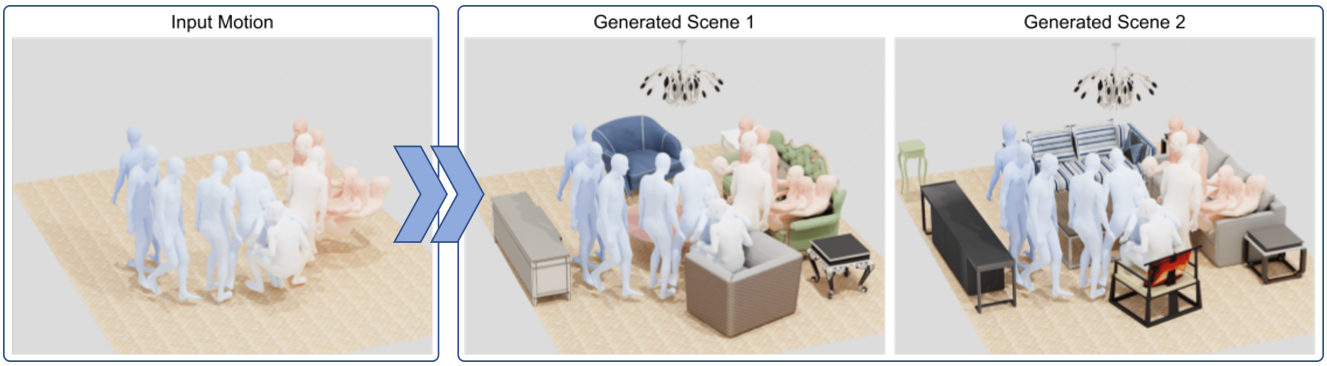

Figure 1. **Estimating 3D scenes from human movement.** Given 3D human motion (left), e.g. from motion capture or body-worn sensors, we reconstruct plausible 3D scenes in which the motion could have taken place. Our generative model is able to produce multiple realistic scenes (right) that take into account the locations and poses of the person, with appropriate human-scene contact.

Abstract

Generating realistic 3D worlds occupied by moving humans has many applications in games, architecture, and synthetic data creation. But generating such scenes is expensive and labor intensive. Recent work generates human poses and motions given a 3D scene. Here, we take the opposite approach and generate 3D indoor scenes given 3D human motion. Such motions can come from archival motion capture or from IMU sensors worn on the body, effectively turning human movement into a "scanner" of the 3D world. Intuitively, human movement indicates the free-space in a room and human contact indicates surfaces or objects that support activities such as sitting, lying or touching. We propose MIME (Mining Interaction and Movement to infer 3D Environments), which is a generative model of indoor scenes that produces furniture layouts that are consistent with the human movement. MIME uses an auto-regressive transformer architecture that takes the already generated objects in the scene as well as the human motion as input, and outputs the next plausible object. To train MIME, we build a dataset by populating the 3D FRONT *scene dataset with 3D humans. Our experiments show that MIME produces more diverse and plausible 3D scenes than a recent generative scene method that does not know about human movement. Code and data are available for research at* https://mime.is.tue.mpg.de.

[*]This work was performed when C.P. H. was at the MPI-IS.

1. Introduction

Humans constantly interact with their environment. They walk through a room, touch objects, rest on a chair, or sleep in a bed. All these interactions contain information about the scene layout and object placement. In fact, a mime is a performer who uses our understanding of such interactions to convey a rich, imaginary, 3D world using only their body motion. Can we train a computer to take human motion and, similarly, conjure the 3D scene in which it belongs? Such a method would have many applications in synthetic data generation, architecture, games, and virtual reality. For example, there exist large datasets of 3D human motion like AMASS [38] and such data rarely contains information about the 3D scene in which it was captured. Could we take AMASS and generate plausible 3D scenes for all the motions? If so, we could use AMASS to generate training data containing realistic human-scene interaction.

To answer such questions, we train a new method called MIME (Mining Interaction and Movement to infer 3D Environments) that generates plausible indoor 3D scenes based on 3D human motion. Why is this possible? The key intuitions are that (1) A human's motion through free space indicates the lack of objects, effectively *carving out* regions of the scene that are free of furniture. And (2), when they are in contact with the scene, this constrains both the type and placement of 3D objects; e.g., a sitting human must be sitting on something, such as a chair, a sofa, a bed, etc.

To make these intuitions concrete, we develop MIME,

which is a transformer-based auto-regressive 3D scene generation method that, given an empty floor plan and a human motion sequence, predicts the furniture that is in contact with the human. It also predicts plausible objects that have no contact with the human but that fit with the other objects and respect the free-space constraints induced by the human motion. To condition the 3D scene generation with human motion, we estimate possible contact poses using POSA [23] and divide the motion in contact and non-contact snippets (Fig. 2). The non-contact poses define free-space in the room, which we encode as 2D floor maps, by projecting the foot vertices onto the ground plane. The contact poses and corresponding 3D human body models are represented by 3D bounding boxes of the contact vertices predicted by POSA. We use this information as input to the transformer and auto-regressively predict the objects that fulfill the contact and free-space constraints; see Fig. 1.

To train MIME, we built a new dataset called *3D-FRONT HUMAN* that extends the large-scale synthetic scene dataset 3D-FRONT [18]. Specifically, we automatically populate the 3D scenes with humans; i.e., non-contact humans (a sequence of walking motion and standing humans) as well as contact humans (sitting, touching, and lying humans). To this end, we leverage motion sequences from AMASS [38], as well as static contact poses from RenderPeople [47] scans.

At inference time, MIME generates a plausible 3D scene layout for the input motion, represented as 3D bounding boxes. Based on this layout, we select 3D models from the 3D-FUTURE dataset [19] and refine their 3D placement based on geometric constraints between the human poses and the scene.

In comparison to pure 3D scene generation approaches like ATISS [46], our method generates a 3D scene that supports human contact and motion while putting plausible objects in free space. In contrast to Pose2Room [43] which is a recent pose-conditioned generative model, our method enables the generation of objects that are not in contact with the human, thus, predicting the entire scene instead of isolated objects. We demonstrate that our method can directly be applied to real captured motion sequences such as PROX-D [22] *without finetuning*.

In summary, we make the following contributions:

- a novel motion-conditioned generative model for 3D room scenes that auto-regressively generates objects that are in contact with the human and do not occupy free-space defined by the motion.

- a new 3D scene dataset with interacting humans and free space humans which is constructed by populating 3D FRONT with static contact/standing poses from RenderPeople and motion data of AMASS.

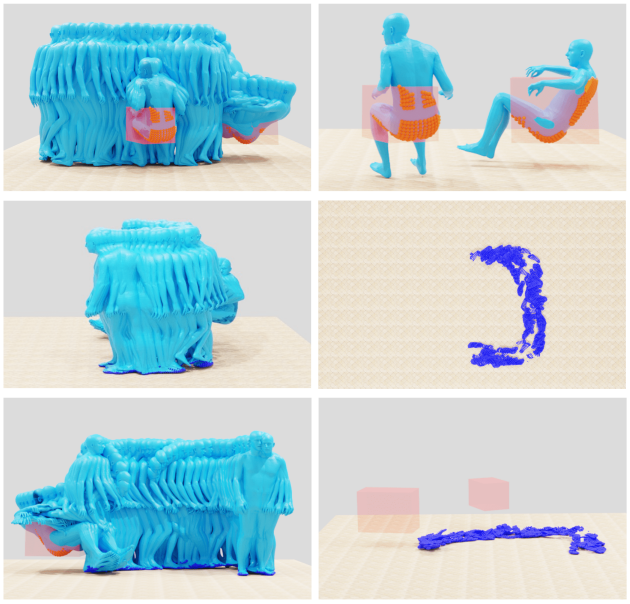

Figure 2. We divide input humans into two parts: contact humans and free-space humans. We extract the 3D bounding boxes for each contact human, and use non-maximum suppression on the 3D IoU to aggregate multiple humans in the same 3D space into a single contact 3D bounding box (orange boxes). We project the foot vertices of free-space humans on the floor plane, to get the 2D free-space mask (dark blue).

2. Related Work

Generative Scene Synthesis (No People). Most prior work on indoor scene synthesis, ignores the human and is based on (1) procedural modeling with grammars[11, 32, 41, 45, 49, 50, 60]; (2) graph neural networks [13, 33, 35, 37, 50, 65, 80–82, 82]; (3) auto-regressive neural networks [53, 66]; or (4) transformers [44, 46, 67]. Some methods leverage lexical text [6] or a sentence [7] as input to guide the 3D scene synthesis. Fisher et al. [16] take 3D scans as input and synthesize the corresponding 3D object arrangements. This is extended to also include functionality aspects in the reconstruction [17]. Recently, ATISS [46] performs scene synthesis using a transformer-based architecture. ATISS takes a floorplan as input and auto-regressively generates a 3D scene that is represented as an unordered set of objects.

All methods mentioned above do not take human motion into consideration to guide the 3D scene synthesis. In contrast, we generate 3D scenes that are compatible with the humans defined by a given input motion. Specifically, the objects in the generated scene should support the human motion (e.g., a chair or couch for sitting) and should not intersect with the path of a walking human, To exploit these insights, we build upon the auto-regressive scene synthesis architecture of ATISS [46] and incorporate contact and free-space information into the pipeline.

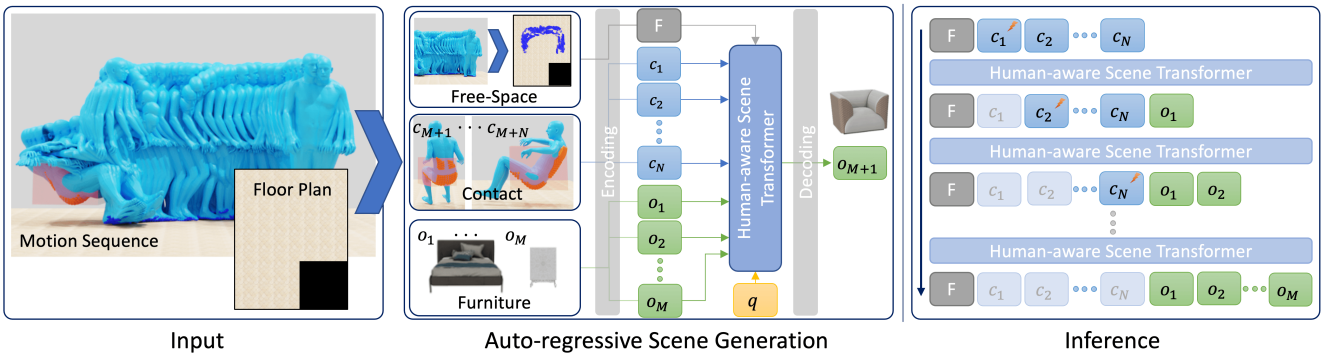

Figure 3. Method overview. During training, our method generates the $(M+1)$-th object through a transformer encoder and a decoding module conditioned on the free space concatenated with the floor plan, contact humans $c_{j=1}^{N}$, other existing objects $o_{j=1}^{M}$ and a learnable query q. We minimize the negative log-likelihood between the distribution of the generated object $M+1$ and the ground truth. During inference, we start from the floor plan, the free space and input contact humans $c_{i=1}^{N}$ and assign the contact label of the first human as 1 by default, to auto-regressively generate objects. At each step, we remove the contact humans that are already supported by the previously generated object and generate next objects until the *end symbol* is generated.

Human-aware Scene Reconstruction. Qi et al. [51] propose a method that synthesizes a 3D scene based on a human's affordance map with a spatial And-Or graph. PiGraphs [54] learns a probability distribution over human pose and object geometry from interactions. However, it does not model the free space carved out by movement. Similarly, recent methods explore how to estimate a 3D scene from human behaviors and interactions. Mura et al. [42] predict a "3D floor plan" from a 2D human walking trajectory without modelling objects or handling contact information. Ye et al. [73] design a contact predictor to estimate temporally coherent contact vertices on an input human motion sequence, and manually select plausible objects to interact with humans and other objects in free space. Nie et al. [43] propose Pose2Room, which predicts 3D objects inside a room from 3D human pose trajectories in a probabilistic way by learning a 3D object arrangement distribution. While it predicts contacted objects, it does not generate objects in free space. In addition, it cannot take floor plans as input. We find these crucial in our experiments since object arrangements are highly related to the floor plan; e.g. furniture like a bed is designed to go against a wall.

Human-Scene Interaction Datasets. Many datasets exist for understanding humans [15, 29, 34, 36, 55, 58, 59, 71, 75, 79] or scenes [14, 72, 76, 77] in separation, but relatively few address humans and scenes [27, 70, 74] together. Human bodies are commonly captured using optical markers [8, 30, 56], IMU sensors [28, 64], and multiple RGB cameras [31, 39, 78]. See [61] for a comprehensive review. These datasets contain only humans, forgoing the 3D environments which the subjects interact with, e.g., floor plane, walls, furniture. In contrast, real 3D scene datasets such as Matterport3D [5], ScanNet [9] and Replica [57] are captured primarily through time-of-flight sensors, where humans are excluded since only static content is reconstructed. Consequently, despite having a large variety of scenes, they are not suitable for modeling human-scene interaction.

To train MIME, we need diverse scene arrangements given a set of sparse or continuously-moving bodies. While recent real datasets [2, 10, 20, 22, 26, 40, 62, 68] capture both humans and environments, they fail to provide sufficient variety because the *a priori* scanned scenes are static and only the subject moves. This limits the variety of scenes that can be practically captured. Hassan et al. [21] use MoCap to capture a person interacting with objects like chairs, sofas and tables. They augment the dataset by changing the size and shape of the objects and update the human pose using inverse kinematics. However, the data does not capture full scenes which we aim to generate. Composited or synthetic datasets such as [1, 3, 47] are widely used for human mesh recovery, but the human-scene interactions are very limited. Pose2Room [43] and GTA-IM [4] are the closest to our needs. However, they represent humans with 3D skeletons, which cannot represent contact between the body surface and the scene. In addition, the variety of scene arrangements is limited. To address these limitations, we introduce a new dataset called *3D-FRONT HUMAN*, which is generated by populating 3D scenes from 3D FRONT [18] with humans that move and interact with the scene.

3. Method

Given input motion of a human and an empty or partially occupied room of a specific kind (e.g., bedroom, living room, etc.) with its floor plan, we learn a generative model that populates the room with objects that support the human interactions. To this end, we propose a human-aware auto-

regressive model that represents scenes as *one* unordered set of objects. We divide the objects into contact objects and non-contact objects based on the human-object interaction. Contact objects are those that humans interact with, while non-contact objects can be placed anywhere in the free space of a room. These non-contact objects enrich the content and potential functionality of a room.

Figure 3 overviews the method. In the following, we describe our human-aware scene synthesis model, MIME, which consists of two components: (1) a generative scene synthesis method based on 3D bounding boxes with object labels, and (2) a 3D refinement method that takes 3D human-scene interactions into account to optimize the placement of the generated objects. In Sec. 4, we detail the dataset generation process; this dataset is used to train our model.

3.1. Generative Human-aware Scene Synthesis

Given humans $\mathcal{H}$ and a floor plan $\mathcal{F}$, our goal is to generate a "habitat" $\mathcal{X} = \{\mathcal{H}, \mathcal{F}, \mathcal{S}\}$ where the 3D scene $\mathcal{S}$ can support all human interactions and motions. In contrast to the pure 3D scene generation methods [44, 46], we focus on leveraging information from human motion to guide the 3D scene generation. We extract two types of information from the input motion and the corresponding human bodies: (i) contact humans $\mathcal{C}$ and (ii) free-space humans. We use POSA [23], to automatically label the vertices of the posed human meshes which are potentially in contact with an object. Free-space humans are those that are only in contact with the ground plane, $\mathcal{F}$. These define a binary mask that we call the free-space mask $\mathcal{E}$ (for "Empty"), which is constructed by taking the union of all projected foot contact points on $\mathcal{F}$. This free-space mask $\mathcal{E}$ defines the region of a room that is free from objects since a human can stand and walk there. See the dark blue "footprints" in Fig. 2.

Given all contact humans, we compute the bounding boxes around their contact vertices; see the orange dots and boxes in Fig. 2. We keep only the non-overlapping boxes using non-maximum suppression and denote these as c_i. The collection of contact boxes is referred to as $\mathcal{C} = \{c_i\}_{i=1}^{N}$. Instead of storing all contact vertices of all bodies, our features are compact and encode complementary information. The contact humans, represented by $\mathcal{C}$, indicate where to locate an object.

We represent a 3D scene $\mathcal{S}$ as an unordered set of objects, consisting of two kinds of objects based on human-object interaction. Objects in contact with the input human are referred to as contact objects $\mathcal{O} = \{o\}_{i=1}^{N}$, while non-contact objects $\mathcal{Q} = \{q\}_{i=1}^{M}$ are without any human interaction. Formally, a 3D scene is the union of contact and non-contact objects: $\mathcal{S} = \mathcal{O} \cup \mathcal{Q}$.

The free-space mask $\mathcal{E}$, the floor plan $\mathcal{F}$, the contact humans $\mathcal{C}$ as well as the already existing objects $\mathcal{S}$ are input to an auto-regressive transformer model. Each input is encoded with a respective encoder, detailed below. The log-likelihood of the generation of scene $\mathcal{S}$ including contact objects for contact humans and non-contact objects in free space is

$$\log p(\mathcal{S}) = \log p(\mathcal{O}|\mathcal{F}, \mathcal{E}, \mathcal{C}) + \log p(\mathcal{Q}|\mathcal{F}, \mathcal{E}, \mathcal{C}). \quad (1)$$

To calculate the likelihood of all generated contact objects $\mathcal{Q}$, we accumulate the likelihood of every contact object:

$$p(\mathcal{O}|\mathcal{F}, \mathcal{E}, \mathcal{C}) = \sum_{\hat{\mathcal{O}} \in \pi(\mathcal{O})} \prod_{j \in \hat{\mathcal{O}}} p\left(o_j \mid o_{<j}, \mathcal{F}, \mathcal{E}, c_{\geq j}\right),$$

where $p\left(o_j \mid o_{<j}, \mathcal{F}, \mathcal{E}, c_{\geq j}\right)$ is the probability of generating the j_{th} object conditioned on the input floor plan, free-space humans, the rest of contact humans and the previously generated objects, and π is the random permutation function for those generated contact objects in the scene. The likelihood of all non-contact objects $\mathcal{Q}$ is computed by replacing the input contact humans with the corresponding generated contact objects. During training, we remove all contact humans inside the room, thus, all contact objects $\mathcal{O}$ can be treated as non-contact objects $\mathcal{Q}'$:

$$\begin{aligned} p(\mathcal{Q}|\mathcal{F}, \mathcal{E}, \mathcal{C}) &= p(\mathcal{Q}|\mathcal{F}, \mathcal{E}, \mathcal{O}) \\ &= p(\mathcal{Q}|\mathcal{F}, \mathcal{E}, \mathcal{Q}') \\ &= \sum_{\hat{\mathcal{Q}} \in \pi(\mathcal{Q}+\mathcal{Q}')} \prod_{j \in \hat{\mathcal{Q}}} p\left(q_j \mid q_{<j}, \mathcal{F}, \mathcal{E}\right). \end{aligned}$$

We follow [46] to use Monte Carlo sampling to approximate all different object permutations during training; this makes our model invariant to the order of the generated objects.

Free-Space Encoder. The 2D free-space mask $\mathcal{E}$ is encoded together with the 2D floor plan $\mathcal{F}$ using a ResNet-18 [24]. The encoded feature provides the information to the transformer encoder about where an object can be placed.

Contact Encoder. We represent the contact humans as 3D bounding boxes, which consist of the contact label I, the contact (sitting, touching, lying) or object class category k, the translation t, the rotation r, and the size s. At each autoregressive step, we generate an object in the scene. When generating an object for a contact human, we set the contact label I of one contact human to 1 while the others are labeled 0. This label highlights the contribution of the specific contact human to the next generated contacted object. Note that we remove contact humans from the input set if they are already in contact with an existing object in the scene. Otherwise, we encode the j_{th} input contact human by applying the contact encoder E_θ:

$$E_\theta : (I_j, k_j, t_j, r_j, s_j) \rightarrow (I_j, \lambda(k_j), p(t_j), p(r_j), p(s_j)),$$

where $\lambda(\cdot)$ is a learnable embedding for the contact class category k, and $p(\cdot)$ [63] is the positional encoding for the translation t, rotation r and size s.

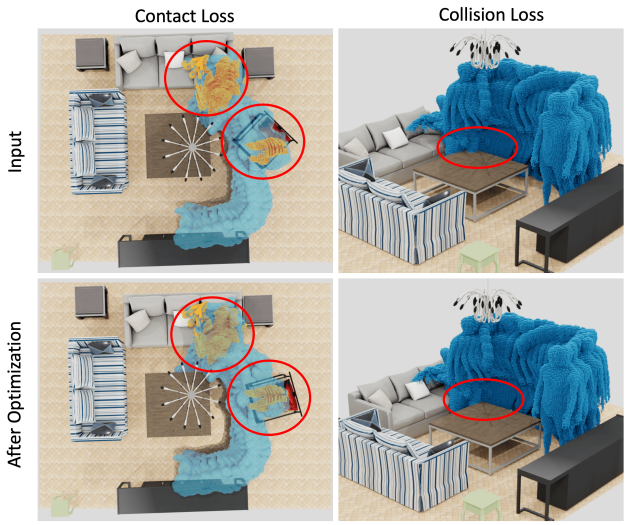

Figure 4. Scene refinement with the collision and contact loss from MOVER [74]. In this example, the contact loss leads to a refinement of the sofa and chair, while the collision loss resolves the intersections of the humans with the table.

Furniture Encoder. The furniture encoder computes the embedding of existing objects in the room:

$$E_\theta : (I_j = 0, k_j, t_j, r_j, s_j) \rightarrow (0, \lambda(k_j), p(t_j), p(r_j), p(s_j)).$$

Note that the furniture encoder is sharing the same network weights as the contact encoder. The contact labels of the objects are all zero, where $j \in [1, M]$.

Scene Synthesis Transformer. We pass the free-space feature F, context embedding $T_{i=1}^{M+N}$, and a learnable query vector $q \in \mathbb{R}^{64}$ into a transformer encoder τ_θ [12, 63] without any positional encoding [63], similar to ATISS[46], to predict the feature $\hat{q}$ that is used to generate the next object:

$$\tau_\theta(F, T_{i=1}^{M+N}, q) \rightarrow \hat{q}.$$

To decode the attributes $(\hat{k}, \hat{t}, \hat{r}, \hat{s})$ of the generated object o_{M+1} from $\hat{q}$, we follow the same design as ATISS [46]. Specifically, we employ an MLP for each attribute in a consecutive fashion. Given $\hat{q}$, we first predict the class category label $\hat{k}$, then we predict the $\hat{t}$, $\hat{r}$ and $\hat{s}$ in this specific order, where the previous attribute will be concatenated with the input $\hat{q}$ for the next attribution prediction.

3.2. Training and Inference.

We train our model on the training set of *3D-FRONT HUMAN*, by maximizing the log-likelihood of each generated scene $\mathcal{S}$ in Eq. (1). During training, we select a human-populated scene in *3D-FRONT HUMAN* and add a random permutation $\pi(\cdot)$ on all N contact and M non-contact objects. We randomly select the $m_{th}+1$ as the generated object, where $m \in [0, N+M]$. Note that, $m = 0$ represents an empty scene, while $m = N + M$ indicates the generated scene is already full and the class label k of the predicted object is an extra *end symbol*. Our model predicts the attribute distribution of the generated object, conditioned on the floor plan $\mathcal{F}$, free space $\mathcal{E}$, previous m objects and contact humans $\mathcal{C}$; see Fig. 3. To enable our model to generate both contact and non-contact objects, we apply data augmentation by adding or dropping input contact humans.

During inference, we start with an empty floor plan F with input humans including free-space humans $\mathcal{E}$, and contact humans $\mathcal{C}$. We auto-regressively predict a new object including its attributes. By default, we set the contact label of the first contact human to 1, and the rest to 0. After each generation step, we remove contact humans that are already in contact, by computing the 2D IoU of the human bounding box and the generated object by projecting them on the ground plane. Specifically, if the IoU is larger than 0.5, we remove the contact human from the input. Once the *end symbol* is generated, the scene synthesis is finished.

3.3. 3D Scene Refinement

The generated scene from our model is represented with 3D bounding boxes. Based on the bounding box size and class category label, we retrieve the closest mesh model from 3D FUTURE [19]. To improve the human-scene interaction between the generated scenes and input humans, we apply the collision loss and the contact loss from MOVER [74] to refine the object position, as can be seen in Fig. 4. We calculate a unified SDF volume and accumulate all contact vertices for all humans in the 3D space, and jointly optimize the object alignment to improve human-object contact and resolve 3D interpenetrations between humans and the scene. The MOVER contact loss weight and the collision loss weight are $1e5$ and $1e3$ respectively.

4. Dataset Generation of 3D-FRONT HUMAN

To enable 3D scene generation from humans, we need a dataset that consists of large numbers of rooms with a wide variety of human interactions. Since no such dataset exists, we generate a new synthetic dataset by automatically populating the 3D rooms in 3D FRONT [18] with interactive humans. We name the resulting dataset *3D-FRONT HUMAN*. To populate the rooms of 3D FRONT with people, we insert humans with contact and humans that stand or walk in free space, as shown in Fig. 5. We represent people with the SMPL-X model [48] and add contact humans from RenderPeople [47] by randomly assigning plausible interactions to different contactable objects in the room. Specifically, we allow for three types of contact interactions: touching, sitting, and lying. In Fig. 5 (bottom), we put a lying down person on a bed, and multiple humans interact with a nightstand or wardrobe. In the free space, we put a random number

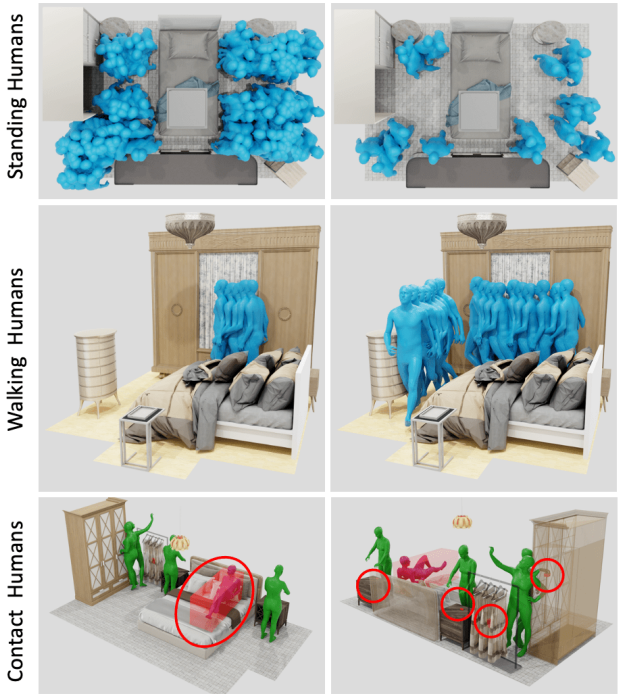

Figure 5. Illustration of populated 3D scenes in *3D-FRONT HUMAN*. Given a room, a random number of static "standing" people and "walking" motion sequences with random start positions and directions are automatically added in the free space. "Contact humans" are added for objects that support interactions like "sitting", "lying", or "touching".

of static standing people and add multiple walking motion clips from AMASS [38] with random start positions and directions to the scene, and remove humans that intersect with objects.

5. Experiments

We qualitatively and quantitatively evaluate our method and compare with two baselines. Specifically, we compare to the 3D scene generation method ATISS [46] and the human-aware scene reconstruction method Pose2Room [43].

Evaluation Datasets. Our human-populated dataset *3D-FRONT HUMAN* contains four room types: 1) 5689 bedrooms, 2) 2987 living rooms, 3) 2549 dining rooms and 4) 679 libraries. We use 21 object categories for the bedrooms, 24 for the living and dining rooms, and 25 for the libraries. We independently train our model four times on the four room types. Following our baseline ATISS [46], for each room type, we split the data 80%, 10%, 10% into training, validation and test sets. We train and validate MIME on the training and validation sets respectively, and evaluate it on the test set. Since ATISS [46] does not provide a pretrained model, we retrain it with the official code[1] following the same training strategy on the original 3D FRONT dataset.

To evaluate the effectiveness and generalization of our method, we compare MIME with Pose2Room [43] on PROX-D [22], a real-world dataset with human motions captured with an RGB-D camera. Note that Pose2Room needs a sequence of human motions that are in contact with objects which our *3D-FRONT HUMAN* does not provide. Thus, fine-tuning of Pose2Room on our dataset is not possible.

Evaluation Metrics. We compare MIME with the baselines on the *test* split of the *3D-FRONT HUMAN* dataset by measuring: (i) the plausibility of human-scene interaction and (ii) the realism of the generated scenes. We propose an *interpenetration metric* (↓) to evaluate the collision between the generated objects and the free-space, by computing the ratio of the violated and non-violated free-space using the 2D projection of the generated objects:

$$L_{\text{inter}} = \left(\sum_{j=1}^{M} \sum_{p \in O_j} \mathcal{E}(p) \right) / \sum_{p \in \mathcal{E}} \mathcal{E}(p),$$

where p denotes each pixel on the floor plan image. We calculate the *2D IoU* and *3D IoU* between generated objects and input contact bounding boxes to measure the human-object interaction. To evaluate the realism and diversity of generated scenes, we follow [46, 80] and calculate the FID [25] score (at 256^2 resolution) between a bird's-eye view orthographic projections of the generated and real scenes from the *test* set, as well as the category KL divergence. We compute the FID score 10 times and report the mean and variance of it.

5.1. Human-aware Scene Synthesis.

In Fig. 6, we visualize the ability of our method to generate plausible 3D scenes from input motion and floor plans for different types of rooms; we also show our baseline methods for comparison. See Sup. Mat. for more examples. Note that the original ATISS [46] model generates a 3D scene only based on the floor plan, without taking the humans into account. Thus, generated scenes from ATISS violate free space constraints and are not consistent with the human contact. In an additional experiment, we extend ATISS to take information about the human motion as input. Specifically, we adapt the 2D input floor plan to also contain the free space information of the walking and standing humans. However, ATISS still generates objects in free space, while generating implausible object configurations such as the white closet inside the bed (Fig. 6, top). In contrast, MIME generates plausible 3D scenes that have fewer interpenetrations with the free space and support interacting humans; e.g. a bed beneath a lying person and a chair under a sitting person.

[1] https://github.com/nv-tlabs/ATISS/commit/6b46c11.

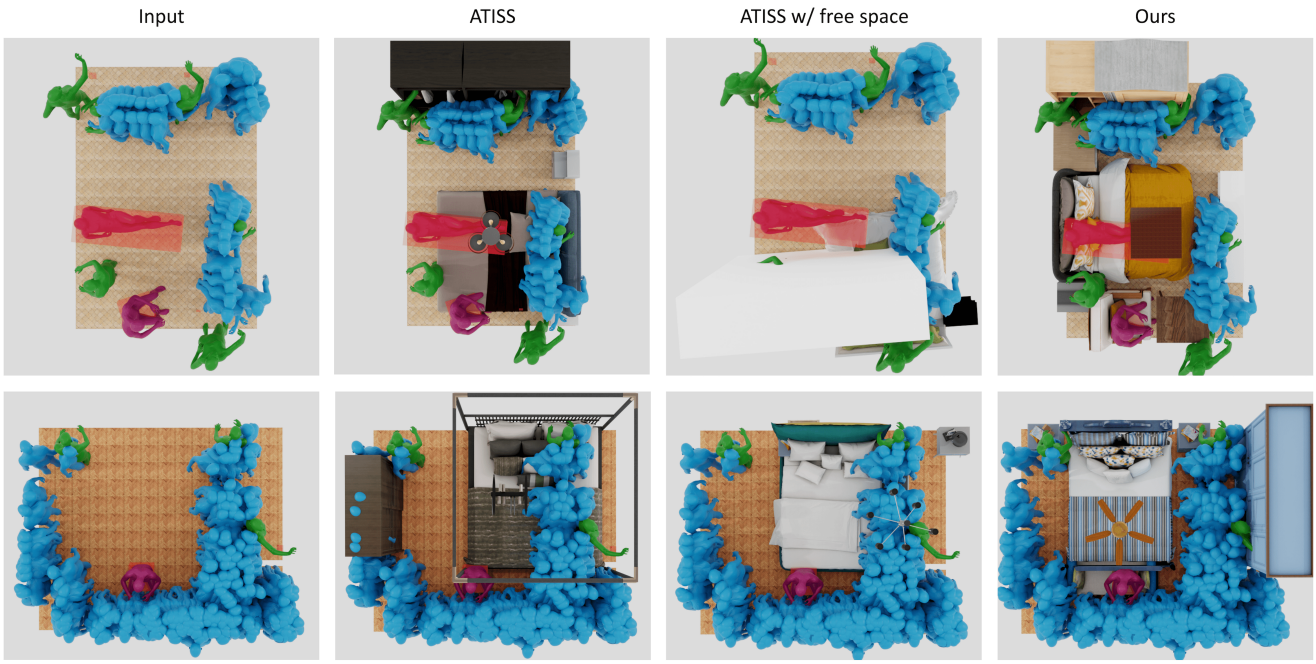

Figure 6. Qualitative comparison on the test split in *3D FRONT HUMAN*. Given free space and contact humans as input, MIME generates more plausible scenes in which the contact humans interact with the contact objects and the free space humans have fewer collisions with all the generated objects, in comparison to the baselines. We also show the original ATISS w/ or w/o the free space mask as input. All results are w/o refinement. Top and bottom rows represent two different example inputs.

	Interpenetration(↓)		2D IoU(↑)		3D IoU(↑)		FID Score (↓)		Category KL Div. (↓)	
	ATISS [46]	Ours	ATISS [46]	Ours	ATISS [46]	Ours	ATISS [46]	Ours	ATISS [46]	Ours
Bedroom	0.348	**0.129**	0.472	**0.939**	0.376	**0.756**	**70.21**±1.80	74.18±2.19	**0.028**	0.044
Living	0.129	**0.050**	0.480	**0.971**	0.360	**0.920**	**130.61**±1.27	150.03± 1.00	**0.004**	0.053
Dining	0.121	**0.047**	0.163	**0.959**	0.122	**0.769**	**45.99** ± 0.90	76.75 ± 1.45	**0.004**	0.037
Library	0.139	**0.106**	0.351	**0.725**	0.390	**0.570**	**93.16** ± 2.59	118.34±2.94	**0.066**	0.093

Table 1. Quantitative comparison on the *test* split of the *3D-FRONT HUMAN* dataset. The interpenetration metric, 2D IoU and 3D IoU are used to evaluate human-scene interaction in generated scenes. The FID score (reported at 256^2) and category KL divergence are used to evaluate the realism and diversity of generated scenes w.r.t. the ground truth scenes.

The observations in the qualitative comparison are also confirmed by a quantitative evaluation in Tab. 1. MIME achieves significant improvements on human-scene interaction evaluation metrics compared with ATISS. Note, since our scene generation is constrained by the input human motion, the diversity scores (FID, KL divergence), ATISS has lower (better) scores because it is not human-aware. Note that this is not a failure/limitation of MIME, as diversity is reduced by the human motion constraints.

To evaluate the generalization of our method, we test it on the PROX-D [22] dataset with the 3D bounding box annotation from [74]. We use it *without finetuning*, and use the motions to generate scenes. We compare our method with Pose2Room [43], which predicts 3D objects from a motion sequence of 3D skeletons. Note that Pose2Room can only predict contact objects - it does not predict an entire scene which is the goal of our method. Figure 7 presents a qualitative comparison of the methods, while quantitative metrics are reported in Tab. 2. We compute the mean average precision with 3D IoU 0.5 (mAP@0.5) to evaluate the 3D object detection accuracy for those contact objects only. Note that both methods are probabilistic generative models. Therefore, we use the same 5 input motions and sample 10 scenes for each motion sequence, and report the mean value of the 3D IoU following Pose2Room. Our method achieves better 3D object detection accuracy compared to Pose2Room *without pretraining*.

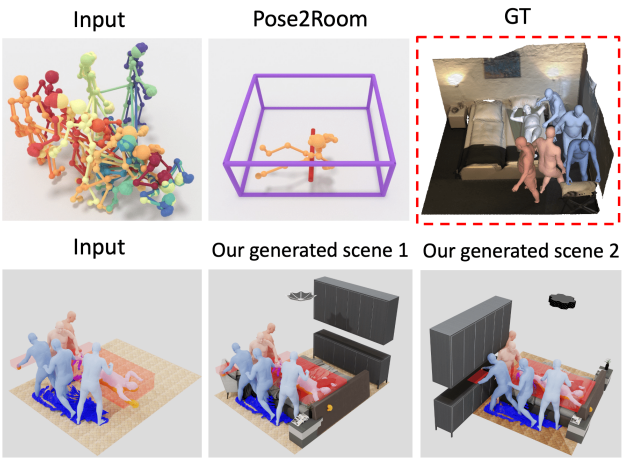

Figure 7. Evaluation on PROX-D [22, 74]. Compared with Pose2Room [43], MIME (w/o finetuning and w/o refinement) can not only generate more accurate contact objects, but it also generates objects appropriately in free space. GT = ground truth.

Method	3D IoU
P2R-Net [43] w/o pretrain	5.36
Ours (MIME) w/o pretrain	**8.47**

Table 2. Comparisons on 3D object detection accuracy (mAP@0.5) using the PROX-D *qualitative* dataset [22].

5.2. Ablation Study

In Fig. 8, we evaluate the influence of the density of free-space humans, and the number of contact humans that we provide as input to MIME. We observe that MIME generates contact objects according to the number of contact humans and, as the density of free-space humans increases, MIME generates fewer objects in the scenes. We also experiment with varying sizes of input floor plans. Larger floor plans result in more objects generated, given the same input motion; see Sup. Mat.

6. Limitation and Discussion

Given a sequence of human motions, MIME generates diverse and plausible scenes with which the humans interact. We assume that the generated scenes are static. In future work moving objects have to be explored, as humans move objects, open doors, or grasp objects like a cup, handle, etc.

MIME, like ATISS, needs a pre-defined room floor plan layout as input. The resolution of the 2D floor plan is coarse (64×64 represents 6.2×6.2 meter square); i.e., 1 pixel is around 10 centimeters wide, which is extracted as a 512 dimension feature by ResNet-18. Introducing a finer floor plan representation, such as dividing one floor plan into multiple patches (cf. ViT [52]), sampling points around the boundary [69], or simply enlarging the feature dimension

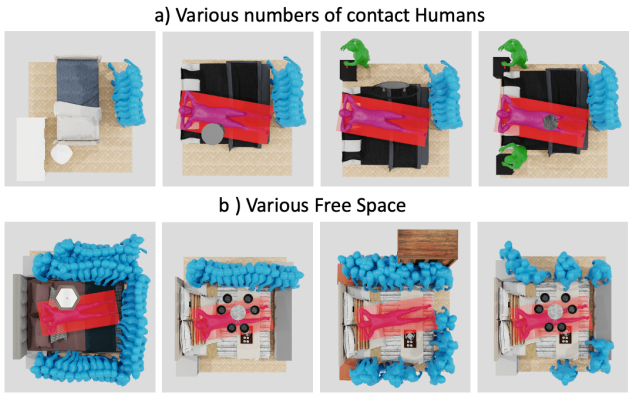

Figure 8. Ablation study on different numbers of contact humans and different density of free space humans. In a), with more contact humans as input, the generated scenes contain more occupied objects. In b), more "free space humans" in a room leads to fewer generated objects in a scene.

could improve the generated object placement, resulting in fewer collisions between the humans and the free space. Another interesting direction is to estimate a floor plan and 3D object layout jointly from input humans.

During inference, MIME uses a hand-crafted 2D IoU metric between the generated objects and the input contact humans to remove contacted humans. In future work, a network could learn this information. Our model directly estimates 3D bounding boxes as a 3D scene representation, followed by a scene refinement that places the mesh models into the scene. Learning to directly estimate the mesh models from the interacting humans is another promising direction.

7. Conclusion

We have introduced MIME, which generates furniture layouts that are consistent with input human movement and contacts. To train MIME, we built a new dataset called *3D-FRONT HUMAN*, by populating humans into the large-scale synthetic scene dataset [18]. We have demonstrated that our method can generate multiple realistic scenes, where the input motion can take place. We believe that MIME is a building block for generating synthetic training data at scale in which humans interact with objects in a scene.

Acknowledgments. We thank Despoina Paschalidou, Wamiq Para for useful feedback about reimplementing ATISS, and Yuliang Xiu, Weiyang Liu, Yandong Wen, Yao Feng for the insightful discussions, Radek Daněček, Peter Kulits for proofreading, and Benjamin Pellkofer for IT support. This work was supported in part by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B.

Disclosure. https://files.is.tue.mpg.de/black/CoI_CVPR_2023.txt

References

[1] Eduard Gabriel Bazavan, Andrei Zanfir, Mihai Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. HSPACE: Synthetic parametric humans animated in complex environments. *arXiv*, 2021. 3

[2] Bharat Lal Bhatnagar, Xianghui Xie, Ilya Petrov, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. BEHAVE: Dataset and method for tracking human object interactions. In *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022. 3

[3] Zhongang Cai, Mingyuan Zhang, Jiawei Ren, Chen Wei, Daxuan Ren, Zhengyu Lin, Haiyu Zhao, Lei Yang, and Ziwei Liu. Playing for 3d human recovery. *arXiv*, 2021. 3

[4] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qizhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[5] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from rgb-d data in indoor environments. In *International Conference on 3D Vision (3DV)*, 2017. 3

[6] Angel Chang, Will Monroe, Manolis Savva, Christopher Potts, and Christopher D Manning. Text to 3d scene generation with rich lexical grounding. *arXiv*, 2015. 2

[7] Angel X Chang, Mihail Eric, Manolis Savva, and Christopher D Manning. Sceneseer: 3d scene design with natural language. *arXiv*, 2017. 2

[8] CMU Graphics Lab. CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu/, 2000. 3

[9] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[10] Yudi Dai, YiTai Lin, XiPing Lin, Chenglu Wen, Lan Xu, Hongwei Yi, Siqi Shen, Yuexin Ma, and Cheng Wang. Sloper4d: A scene-aware dataset for global 4d human pose estimation in urban environments. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[11] Jeevan Devaranjan, Amlan Kar, and Sanja Fidler. Meta-sim2: Learning to generate synthetic datasets. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 2018. 5

[13] Xinhan Di, Pengqian Yu, Hong Zhu, Lei Cai, Qiuyan Sheng, Changyu Sun, and Lingqiang Ran. Structural plan of indoor scenes with personalized preferences. In *European Conference on Computer Vision (ECCV)*, pages 455–468. Springer, 2020. 2

[14] Mingyu Ding, Yuqi Huo, Hongwei Yi, Zhe Wang, Jianping Shi, Zhiwu Lu, and Ping Luo. Learning depth-guided convolutions for monocular 3d object detection. In *Computer Vision and Pattern Recognition (CVPR)*, pages 1000–1001, 2020. 3

[15] Qi Fang, Kang Chen, Yinghui Fan, Qing Shuai, Jiefeng Li, and Weidong Zhang. Learning analytical posterior probability for human mesh recovery. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[16] Matthew Fisher, Daniel Ritchie, Manolis Savva, Thomas Funkhouser, and Pat Hanrahan. Example-based synthesis of 3d object arrangements. *Transactions on Graphics (TOG)*, 31(6):1–11, 2012. 2

[17] Matthew Fisher, Manolis Savva, Yangyan Li, Pat Hanrahan, and Matthias Nießner. Activity-centric scene synthesis for functional 3d scene modeling. *Transactions on Graphics (TOG)*, 34(6):1–13, 2015. 2

[18] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *International Conference on Computer Vision (ICCV)*, pages 10933–10942, 2021. 2, 3, 5, 8

[19] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision (IJCV)*, pages 1–25, 2021. 2, 5

[20] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (HPS): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4318–4329, 2021. 3

[21] Mohamed Hassan, Duygu Ceylan, Ruben Villegas, Jun Saito, Jimei Yang, Yi Zhou, and Michael Black. Stochastic scene-aware motion prediction. In *International Conference on Computer Vision (ICCV)*, Oct. 2021. 3

[22] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision (ICCV)*, pages 2282–2292, 2019. 2, 3, 6, 7, 8

[23] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J. Black. Populating 3D scenes by learning human-scene interaction. In *Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2, 4

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Conference on Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. 6

[26] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[27] Siyuan Huang, Zan Wang, Puhao Li, Baoxiong Jia, Tengyu Liu, Yixin Zhu, Wei Liang, and Song-Chun Zhu. Diffusion-based generation, optimization, and planning in 3d scenes. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[28] Yinghao Huang, Manuel Kaufmann, Emre Aksan, Michael J Black, Otmar Hilliges, and Gerard Pons-Moll. Deep inertial poser: Learning to reconstruct human pose from sparse inertial measurements in real time. *Transactions on Graphics (TOG)*, 37(6):1–15, 2018. 3

[29] Yangyi Huang, Hongwei Yi, Weiyang Liu, Haofan Wang, Boxi Wu, Wenxiao Wang, Binbin Lin, Debing Zhang, and Deng Cai. One-shot implicit animatable avatars with model-based priors. *arXiv*, 2022. 3

[30] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, jul 2014. 3

[31] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2017. 3

[32] Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[33] Mohammad Keshavarzi, Aakash Parikh, Xiyu Zhai, Melody Mao, Luisa Caldas, and Allen Y Yang. SceneGen: Generative contextual scene augmentation using scene graph priors. *arXiv*, 2020. 2

[34] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. NIKI: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

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

[36] Tingting Liao, Xiaomei Zhang, Yuliang Xiu, Hongwei Yi, Xudong Liu, Guo-Jun Qi, Yong Zhang, Xuan Wang, Xiangyu Zhu, and Zhen Lei. High-Fidelity Clothed Avatar Reconstruction from a Single Image. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[37] Andrew Luo, Zhoutong Zhang, Jiajun Wu, and Joshua B Tenenbaum. End-to-end optimization of scene layout. In *Computer Vision and Pattern Recognition (CVPR)*, pages 3754–3763, 2020. 2

[38] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision (ICCV)*, pages 5442–5451, Oct. 2019. 1, 2, 6

[39] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular rgb. In *International Conference on 3D Vision (3DV)*. IEEE, sep 2018. 3

[40] Aron Monszpart, Paul Guerrero, Duygu Ceylan, Ersin Yumer, and Niloy J. Mitra. iMapper: interaction-guided scene mapping from monocular videos. *Transactions on Graphics (TOG)*, 38(4):92:1–92:15, 2019. 3

[41] Pascal Müller, Peter Wonka, Simon Haegler, Andreas Ulmer, and Luc Van Gool. Procedural modeling of buildings. In *Proceedings of the international conference on Computer graphics and interactive techniques (SIGGRAPH)*, pages 614–623, 2006. 2

[42] Claudio Mura, Renato Pajarola, Konrad Schindler, and Niloy Mitra. Walk2map: Extracting floor plans from indoor walk trajectories. In *Computer Graphics Forum (CGF)*, volume 40, pages 375–388, 2021. 3

[43] Yinyu Nie, Angela Dai, Xiaoguang Han, and Matthias Nießner. Pose2room: understanding 3d scenes from human activities. In *European Conference on Computer Vision*, pages 425–443. Springer, 2022. 2, 3, 6, 7, 8

[44] Wamiq Reyaz Para, Paul Guerrero, Niloy Mitra, and Peter Wonka. COFS: Controllable furniture layout synthesis. In *SIGGRAPH Conference Papers*, 2023. 2, 4

[45] Yoav IH Parish and Pascal Müller. Procedural modeling of cities. In *Proceedings of the international conference on Computer graphics and interactive techniques (SIGGRAPH)*, pages 301–308, 2001. 2

[46] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. ATISS: Autoregressive transformers for indoor scene synthesis. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 2, 4, 5, 6, 7

[47] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2, 3, 5

[48] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[49] Aayush Prakash, Shaad Boochoon, Mark Brophy, David Acuna, Eric Cameracci, Gavriel State, Omer Shapira, and Stan Birchfield. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *International Conference on Robotics and Automation. (ICRA)*, pages 7249–7255. IEEE, 2019. 2

[50] Pulak Purkait, Christopher Zach, and Ian Reid. Sg-vae: Scene grammar variational autoencoder to generate new indoor scenes. In *European Conference on Computer Vision (ECCV)*, pages 155–171. Springer, 2020. 2

[51] Siyuan Qi, Yixin Zhu, Siyuan Huang, Chenfanfu Jiang, and Song-Chun Zhu. Human-centric indoor scene synthesis using stochastic grammar. In *Computer Vision and Pattern Recognition (CVPR)*, pages 5899–5908, 2018. 3

[52] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *International Conference on Computer Vision (ICCV)*, pages 12179–12188, 2021. 8

[53] Daniel Ritchie, Kai Wang, and Yu-an Lin. Fast and flexible indoor scene synthesis via deep convolutional generative models. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6182–6190, 2019. 2

[54] Manolis Savva, Angel X. Chang, Pat Hanrahan, Matthew Fisher, and Matthias Nießner. PiGraphs: Learning Interaction Snapshots from Observations. *ACM Transactions on Graphics (TOG)*, 35(4), 2016. 3

[55] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[56] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International Journal of Computer Vision (IJCV)*, 87(1):4–27, 2010. 3

[57] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv*, 2019. 3

[58] Jingxiang Sun, Xuan Wang, Lizhen Wang, Xiaoyu Li, Yong Zhang, Hongwen Zhang, and Yebin Liu. Next3d: Generative neural texture rasterization for 3d-aware head avatars. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[59] Yu Sun, Qian Bao, Wu Liu, Tao Mei, and Michael J Black. TRACE: 5D Temporal Regression of Avatars with Dynamic Cameras in 3D Environments. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[60] Jerry O Talton, Yu Lou, Steve Lesser, Jared Duke, Radomír Měch, and Vladlen Koltun. Metropolis procedural modeling. *Transactions on Graphics (TOG)*, 30(2):1–14, 2011. 2

[61] Yating Tian, Hongwen Zhang, Yebin Liu, and Limin Wang. Recovering 3d human mesh from monocular images: A survey. *arXiv*, 2022. 3

[62] Shashank Tripathi, Lea Müller, Chun-Hao P. Huang, Taheri Omid, Michael J. Black, and Dimitrios Tzionas. 3D human pose estimation via intuitive physics. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Conference on Neural Information Processing Systems (NeurIPS)*, 2017. 4, 5

[64] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, pages 614–631, 2018. 3

[65] Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X Chang, and Daniel Ritchie. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *Transactions on Graphics (TOG)*, 38(4):1–15, 2019. 2

[66] Kai Wang, Manolis Savva, Angel X Chang, and Daniel Ritchie. Deep convolutional priors for indoor scene synthesis. *Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2

[67] Xinpeng Wang, Chandan Yeshwanth, and Matthias Nießner. Sceneformer: Indoor scene generation with transformers. In *International Conference on 3D Vision (3DV)*, 2021. 2

[68] Zhe Wang, Liyan Chen, Shaurya Rathore, Daeyun Shin, and Charless Fowlkes. Geometric pose affordance: 3D human pose with scene constraints. *arXiv*, 2019. 3

[69] Qiuhong Anna Wei, Sijie Ding, Jeong Joon Park, Rahul Sajnani, Adrien Poulenard, Srinath Sridhar, and Leonidas Guibas. Lego-net: Learning regular rearrangements of objects in rooms. *arXiv*, 2023. 8

[70] Xianghui Xie, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Visibility aware human-object interaction tracking from single rgb camera. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[71] Yuliang Xiu, Jinlong Yang, Xu Cao, Dimitrios Tzionas, and Michael J. Black. ECON: Explicit Clothed humans Optimized via Normal Integration. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[72] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European Conference on Computer Vision (ECCV)*, pages 674–689. Springer, 2020. 3

[73] Sifan Ye, Yixing Wang, Jiaman Li, Dennis Park, C Karen Liu, Huazhe Xu, and Jiajun Wu. Scene synthesis from human motion. In *SIGGRAPH Asia Conference Papers*, pages 1–9, 2022. 3

[74] Hongwei Yi, Chun-Hao P. Huang, Dimitrios Tzionas, Muhammed Kocabas, Mohamed Hassan, Siyu Tang, Justus Thies, and Michael J. Black. Human-aware object placement for visual environment reconstruction. In *Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 5, 7, 8

[75] Hongwei Yi, Hualin Liang, Yifei Liu, Qiong Cao, Yandong Wen, Timo Bolkart, Dacheng Tao, and Michael J. Black. Generating holistic 3D human motion from speech. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[76] Hongwei Yi, Shaoshuai Shi, Mingyu Ding, Jiankai Sun, Kui Xu, Hui Zhou, Zhe Wang, Sheng Li, and Guoping Wang. Segvoxelnet: Exploring semantic context and depth-aware features for 3d vehicle detection from point cloud. In *International Conference on Robotics and Automation. (ICRA)*, pages 2274–2280. IEEE, 2020. 3

[77] Hongwei Yi, Zizhuang Wei, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Pyramid multi-view stereo net with self-adaptive view aggregation. In *European Conference on Computer Vision (ECCV)*, pages 766–782. Springer, 2020. 3

[78] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. HUMBI: A large multiview dataset of human body expressions. In *Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[79] Hongwen Zhang, Siyou Lin, Ruizhi Shao, Yuxiang Zhang, Zerong Zheng, Han Huang, Yandong Guo, and Yebin Liu. Closet: Modeling clothed humans on continuous surface with explicit template decomposition. In *Computer Vision and Pattern Recognition (CVPR)*, June 2023. 3

[80] Song-Hai Zhang, Shao-Kui Zhang, Wei-Yu Xie, Cheng-Yang Luo, and Hong-Bo Fu. Fast 3d indoor scene synthesis with discrete and exact layout pattern extraction. *arXiv*, 2020. 2, 6

[81] Zaiwei Zhang, Zhenpei Yang, Chongyang Ma, Linjie Luo, Alexander Huth, Etienne Vouga, and Qixing Huang. Deep generative modeling for scene synthesis via hybrid represen-

tations. *Transactions on Graphics (TOG)*, 39(2):1–21, 2020.

[82] Yang Zhou, Zachary While, and Evangelos Kalogerakis. Scenegraphnet: Neural message passing for 3d indoor scene augmentation. In *International Conference on Computer Vision (ICCV)*, pages 7384–7392, 2019. 2