Realistic Indoor Scene Reconstruction from a Single Image

Stanford CS231A Project Final Report

Yunong Liu*
Department of Computer Science
Stanford University
yunongl@stanford.edu

Zhen Wu*
Department of Computer Science
Stanford University
zhenwu@stanford.edu

Harrison Guan*
Department of Mechanical Engineering
Stanford University
hrguan@stanford.edu

Abstract

We propose a novel approach for reconstructing 3D indoor scenes from a single image. Our method addresses the challenges of physical plausibility and limited input views by leveraging a database of 3D CAD models and introducing a contact-aware refinement step. We first apply Set-of-Mark (SoM) prompting to partition the input image into semantically meaningful regions, which are then used to guide a large language model (LLM) in generating a structured scene description. By parsing this description, we extract spatial relationships between objects and use this information to retrieve and align corresponding 3D CAD models. Finally, our contact-aware refinement step adjusts the positions and orientations of the aligned models based on their spatial relationships, ensuring physically plausible object-to-object and object-to-layout interactions. We evaluate our approach on the challenging ScanNet dataset and demonstrate significant improvements over state-of-the-art methods in terms of reconstruction accuracy and perceptual realism. Our results showcase the effectiveness of combining LLMs with contact-aware refinement for single-image 3D indoor scene reconstruction. The code can be found in this link.

1. Introduction

Reconstructing 3D indoor environments from a single image is a challenging problem due to complex layouts, object clutter, and occlusions (13). Existing methods often rely on multi-view inputs or produce reconstructions that deviate from the actual scene geometry (11; 19). Single-image 3D reconstruction has numerous applications, including virtual reality, augmented reality, robotics, and scene understanding. However, the limited information available in a single image makes it difficult to recover accurate 3D geometry and spatial relationships between objects. To address these challenges, we propose a novel approach for single-image indoor scene reconstruction using a database of 3D CAD models. Our method leverages the rich geometric information available in CAD models to generate physically plausible and visually coherent 3D scenes. By combining CAD model retrieval, object alignment, and contact-aware refinement, our approach achieves significant improvements over state-of-the-art methods in terms of reconstruction accuracy and perceptual realism. Our method consists of three main stages. First, we employ a state-of-the-art CAD model retrieval technique (7) to identify the most similar CAD models for each object in the input image. Second, we align the retrieved CAD models to the image using a robust pose estimation algorithm (20). Finally, we introduce a novel contact-aware refinement step that adjusts the positions and orientations of the aligned models based on their spatial relationships, ensuring physically plausible interactions between objects and the scene layout. The main contributions of our work are:

- A framework for single-image indoor scene reconstruction that integrates CAD model retrieval, object alignment, and contact-aware refinement.
- A novel contact-aware post-processing method that improves the physical plausibility of the reconstructed scene by considering object-to-object and object-to-

layout interactions.

- Extensive experiments on the challenging ScanNet dataset (6), demonstrating improved reconstruction accuracy and perceptual realism compared to state-of-the-art methods.

The paper is organized as follows. Section 2 reviews related work, Section 3 describes our methodology, Section 4 presents experimental results, and Section 5 discusses limitations and future work.

2. Related Work

2.1. Single-Image 3D Scene Reconstruction

Single-image indoor scene reconstruction has been an active research area in recent years. Early methods relied on hand-crafted features and geometric constraints to reconstruct 3D scenes from a single image (10; 22). These methods often made strong assumptions about the scene geometry and struggled to handle complex scenes with occlusions and clutter.

With the advent of deep learning, several approaches have been proposed to leverage the power of neural networks for single-image 3D scene reconstruction. One line of research focuses on end-to-end learning of 3D scene representations from a single image. (19) introduced an unsupervised approach using differentiable rendering and learned shape priors, while (11) proposed a coarse-to-fine method utilizing depth estimation, surface normal prediction, and semantic segmentation. These methods have shown promising results in reconstructing 3D scenes from a single image but often struggle with complex scenes and heavily occluded objects.

Another line of research investigates the use of 3D CAD models for single-image scene reconstruction. (7) introduced ROCA, a robust CAD model retrieval and alignment method that leverages dense correspondences and a joint embedding space. (15) proposed Mask2CAD for retrieving and aligning CAD models to objects in an RGB image. These methods have demonstrated impressive performance in reconstructing 3D scenes using CAD models but do not explicitly consider the physical plausibility of the reconstructed scene.

2.2. 3D Object Pose Estimation

Estimating the 3D poses of objects from a single image is a crucial component of indoor scene reconstruction. Traditional methods relied on feature matching and geometric constraints to estimate object poses (3; 18). However, these methods often struggled with occlusions, symmetries, and intra-class variations.

Recent works have explored the use of deep learning for 3D object pose estimation. (24) proposed PoseCNN, a convolutional neural network (CNN) architecture for estimating the 6D poses of known objects from a single image. (16) introduced DeepIM, an iterative refinement framework for 6D pose estimation using deep neural networks. These methods have shown significant improvements over traditional approaches but require a large amount of labeled training data.

Few-shot learning and neural rendering have also been explored for 3D object pose estimation. (23) proposed Neural View Synthesis and Matching (NVSM) for semi-supervised few-shot learning of 3D object pose, demonstrating impressive data efficiency and robustness to occlusion. However, this method focuses on individual objects and does not consider the overall scene context.

2.3. CAD Model Alignment and Pose Refinement

Aligning 3D CAD models to objects in an image is an essential step in single-image indoor scene reconstruction. Early methods relied on hand-crafted features and iterative closest point (ICP) algorithms to align CAD models to 3D point clouds (4; 21). However, these methods often required good initial alignments and struggled with partial observations and occlusions.

Recent works have leveraged deep learning to improve the robustness and accuracy of CAD model alignment. (20) introduced CORENets for predicting object correspondences and 6D poses, while (5) proposed a category-level 6D pose estimation method handling intra-class shape variations. These methods have shown promising results in aligning CAD models to objects in an image but do not explicitly model the physical interactions between objects in the scene.

2.4. Physical Reasoning and Contact-Aware Reconstruction

Incorporating physical reasoning and contact-aware constraints into 3D scene reconstruction is crucial for generating realistic and physically plausible results. Early methods relied on hand-crafted rules and optimization techniques to ensure the stability and plausibility of reconstructed scenes (8; 12). However, these methods often made strong assumptions about the scene geometry and struggled to handle complex object interactions.

Recent works have explored the use of deep learning and physics simulation for contact-aware 3D scene reconstruction. (25) proposed a structured neural network architecture for learning object-centric representations and physical interactions from 3D scene graphs. These methods have shown promising results in generating realistic and physically plausible 3D scenes but require extensive annotations and simulation data.

Our proposed method builds upon the strengths of ROCA (7) by incorporating a contact-aware post-processing step to ensure the physical plausibility of the reconstructed scene. We leverage the power of large language models (LLMs) to generate structured scene descriptions and extract spatial relationships between objects. By combining CAD model retrieval, object alignment, and contact-aware refinement, our approach aims to generate realistic and physically plausible 3D indoor scenes from a single image.

3. Technical Approach

In our proposed method, we first leverage the pre-trained ROCA model (7) for pose estimation and 3D CAD model retrieval. Next, we utilize the GPT-4V (1) language model to generate structured scene descriptions. Finally, we apply a contact-aware post-processing step to ensure physically plausible scene reconstruction. Additionally, we extend our method to generate 3D scenes directly from textual descriptions, enhancing its versatility and applicability.

3.1. Pose Estimation: ROCA

We utilize ROCA (7) to retrieve and align 3D CAD models from a shape database to a single input image. The ROCA model has been trained on the ScanNet dataset (6), which contains 1,513 scans of 707 unique indoor environments, along with CAD model alignments. The method begins by detecting and segmenting objects in an RGB image using a Mask-RCNN (9) backbone and estimating depths with a multi-scale FPN (17). It then aligns these objects by calculating their 9-DoF alignment using estimated depths, 2D features, and instance masks. ROCA employs a weighted Procrustes optimization to refine object rotation and translation, ensuring robust alignment. The method extends to retrieving geometrically similar CAD models for each detected object, leveraging geometry-aware joint embeddings and a nearest-neighbor lookup from a CAD database for real-time inference.

The model is trained in an end-to-end manner, with the total loss defined as follows:

$$L_{\text{align}} = w_{\text{rot}} L_{\text{rot}} + L_{\text{trans}} + L_{\text{scale}} + w_{\text{noc}} L_{\text{noc}} + L_{\text{trans\_initial}} \quad (1)$$

Here, w_{rot} and w_{noc} are weights for the rotational and NOC-correspondence components of the loss, respectively. L_{rot}, L_{trans}, and L_{scale} denote losses for rotation, translation, and scaling errors, aiming to align the 3D model to observed 2D features accurately. $L_{\text{trans\_initial}}$ accounts for errors in the initial translation estimation.

3.2. Scene Description Generation

To generate structured scene descriptions from the input image, we employ a large language model (LLM), specifically GPT-4V (1). GPT-4V is a powerful language model that has been trained on a vast amount of text data, enabling it to understand and generate human-like descriptions. In our approach, we leverage GPT-4V's ability to interpret visual information and generate structured scene descriptions in a predefined domain-specific language (DSL) format.

The DSL format is designed to capture essential details about the objects present in the scene, including their attributes, bounding box coordinates, and spatial relationships with other objects. The DSL consists of a set of predefined syntax rules and keywords that allow for the efficient representation of scene information. For example, each object in the scene is represented by a unique identifier, followed by its class label, bounding box coordinates, and any relevant attributes (e.g., color, size, or material). Spatial relationships between objects are captured using predefined keywords such as ”on,” ”next to,” or ”behind,” along with the corresponding object identifiers.

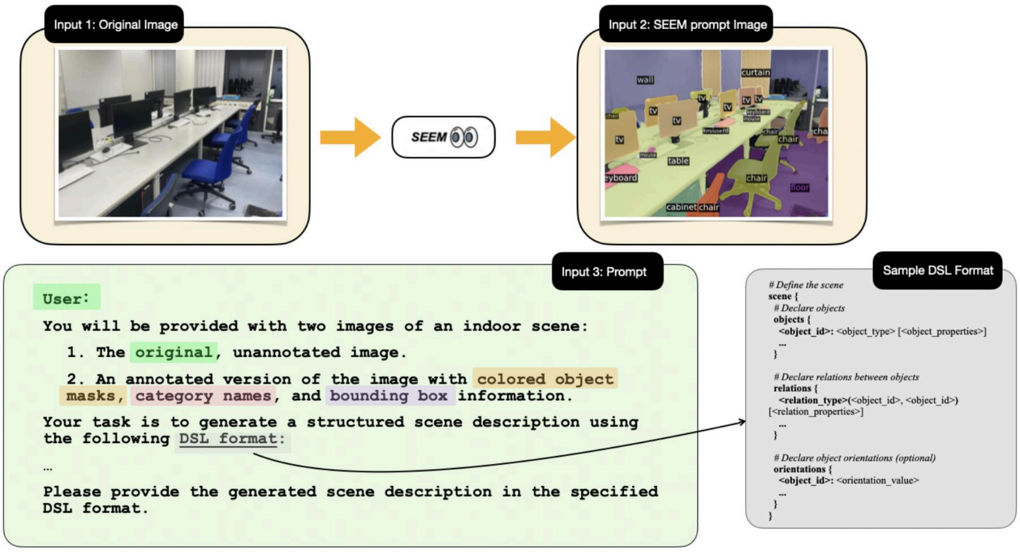
To generate the structured scene description, we first apply Set-of-Mark (SoM) prompting to the input image. SoM prompting utilizes state-of-the-art segmentation models, such as SEEM (26) and SAM (14), to partition the image into semantically meaningful regions. These regions are then overlaid with visual marks, including colored masks and bounding boxes, to provide additional context for the LLM. The SoM-prompted image serves as input to GPT-4V, which analyzes the image and generates a structured scene description in the DSL format.

Using a structured representation in the form of the DSL offers several benefits. First, it allows for efficient parsing and extraction of relevant information, such as object attributes and spatial relationships, which are crucial for the subsequent steps of our approach. Second, the structured format enables easy integration with other components of our pipeline, such as the scene description parsing and spatial relationship extraction modules. Third, the DSL provides a standardized and interpretable representation of the scene, facilitating debugging, analysis, and evaluation of the generated descriptions.

Figure 1 illustrates an example of a generated scene description in the specified DSL format. As shown in the figure, the description captures objects, their attributes, bounding boxes, and spatial relationships, providing a comprehensive representation of the indoor scene.

3.3. Scene Description Parsing and Spatial Relationship Extraction

Once the structured scene description is generated in the DSL format, the next step is to parse it and extract the relevant information, particularly the spatial relationships between objects. While the scene description may include various spatial relationships such as ”on,” ”next to,” and ”behind,” our method primarily focuses on the ”on” relationship for the subsequent steps of estimating the 3D poses of objects and refining their positions and orientations. We



```
User:
You will be provided with two images of an indoor scene:
  1. The original, unannotated image.
  2. An annotated version of the image with colored object
  masks, category names, and bounding box information.
Your task is to generate a structured scene description using
the following DSL format:
...
Please provide the generated scene description in the specified
DSL format.
```

```
# Define the scene
scene {
 # Declare objects
 objects {
  <object_id>: <object_type> [<object_properties>]
  ...
 }

 # Declare relations between objects
 relations {
  <relation_type>(<object_id>, <object_id>)
[<relation_properties>]
  ...
 }

 # Declare object orientations (optional)
 orientations {
  <object_id>: <orientation_value>
  ...
 }
}
```

Figure 1. Example of a generated scene description in the specified DSL format. The description captures objects, their attributes, bounding boxes, and spatial relationships.

utilize LLMs to extract these relationships.

The extracted ”on” relationships are then represented using a directed graph structure, where each node corresponds to an object instance, and each directed edge represents an ”on” relationship between two objects. The direction of the edge indicates which object is on top of the other. This graph-based representation allows for efficient reasoning about the vertical stacking of objects in the scene, which is crucial for the subsequent steps of our approach.

3.4. Contact-Aware Post-Processing

The final stage of our approach focuses on reconstructing the 3D indoor scene by placing the objects according to their estimated poses and the extracted spatial relationships. Our goal is to generate a physically plausible and visually coherent representation of the scene that closely matches the input image. To achieve this, we introduce a contact-aware post-processing step that refines the object positions and orientations based on their spatial relationships and contact constraints.

The contact-aware post-processing step takes as input the estimated object poses from the ROCA model and the extracted ”on” relationship graph from the scene description parsing stage. It then applies a series of refinement operations to ensure that the objects are placed in a physically plausible manner, taking into account their contact points and the stability of the overall scene.

The method begins by finding the global vertical direction. Since all the CAD models are y-up, we determine the vertical direction of all objects from their estimated poses. Specifically, for an object i with an estimated rotation R_i, its vertical direction is $R_i y$, where $y = [0, 1, 0]$. We take the average of the vertical directions of all objects as the global vertical direction:

$$y_{\text{global}} = \text{avg}(y_i). \tag{2}$$

Simultaneously, we take the average of the lowest points of the meshes of all objects on the ground as the ground height.

Next, we start from the ground and traverse the relationship graph in topological order. If there is an edge from object a to object b, it indicates that object b should be placed above object a, which means the height of object b should equal to the height of object a plus the height of the center-of-mass of object b. We calculate the height of object b as follows:

$$\text{height}_\text{b} = \text{height}_\text{a} + \text{height}_\text{b, com}. \tag{3}$$

We repeat this process for all objects, ensuring each object's height is adjusted accordingly.

This structured approach ensures that objects are placed in a physically plausible manner, maintaining stability and coherence in the reconstructed 3D scene.

3.5. An Extension: Generation from Text

Going beyond generating 3D indoor scenes from a single image, we also explore generation from texts. This is practical because, in many scenarios, it is more convenient for users to express their layout design thoughts through text rather than providing an actual image. We explore two approaches in this regard. First, we utilize a large text-to-image model to generate an image from the text, and then use our proposed image-to-scene pipeline to generate the 3D scene. Additionally, as suggested by (2), we ask LLM to directly generate the DSL description of the scene. We then utilize a gradient-based optimization scheme to compute the 3D layout based on this description. This approach bypasses the need for intermediate image generation and directly interprets textual descriptions into spatial arrangements.

Text-to-Image-to-Scene The first approach begins with converting textual descriptions into images. Here, the generated images serve as the basis for our image-to-scene conversion pipeline.

Text-to-DSL-to-Scene In this innovative approach, we bypass image generation and directly transform text descriptions into DSL formats that articulate the spatial and relational attributes of the intended 3D scene. This direct text-to-DSL method relies heavily on the capabilities of advanced Language Models (LLMs) to interpret and translate natural language into structured data that accurately represents a scene's layout. We then follow the gradient-based optimization as proposed in (2) to get the 3D layout.

4. Results

4.1. Dataset

We will use the same dataset as ROCA for consistency and to take advantage of the pre-trained model's learned features and alignments. By utilizing the pre-trained ROCA model, we can focus in developing our contact-aware post-processing step and fine-tuning the network parameters to improve the physical plausibility and accuracy of the reconstructed scene.

4.2. DSL Generation

To generate a structured scene description in the DSL format, we used a standardized prompt for LLMs to generate spatial relationships:

You will be provided with two images:

1. An original indoor scene image.
2. The same image with colored object masks and their corresponding category names and 3D bounding box values.

Your task is to analyze the images and generate a scene description using the following DSL format:

```
scene {
  # Declare objects
  objects: [
    {
      id: <object_id>
      category: <category_name>
      mask_color: (<r>, <g>, <b>)
      bounding_box: {
        position: (<x>, <y>, <z>)
        size: (<width>, <height>, <depth>)
        orientation: (<rx>, <ry>, <rz>)
      }
      [attributes: {
        <attribute_name>: <attribute_value>
        ...
      }]
    },
    ...
  ]

  # Declare spatial relationships
  spatial_relationships: [
    {
      type: <relationship_type>
      subject_id: <object_id>
      object_id: <object_id>
      [properties: {
        <property_name>: <property_value>
        ...
      }]
    },
    ...
  ]
}
```

Guidelines:

- Identify the objects in the scene based on the colored masks and their category names.
- Extract the 3D bounding box information for each object.
- Determine the spatial relationships between the objects based on their positions and orientations.
- Use the provided DSL format to generate a structured scene description.
- If any additional object attributes can be inferred from the images, include them in the object definitions.
- Ensure that the generated DSL is valid and well-formatted.

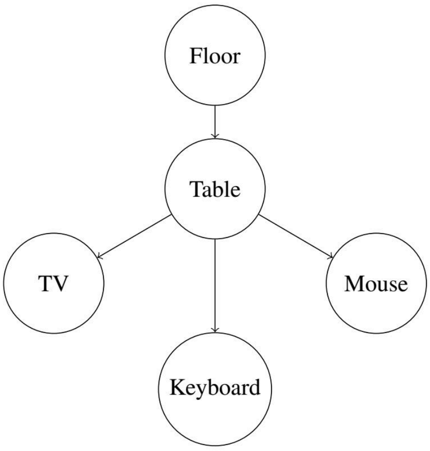
Please provide the scene description in the specified DSL format.

For example, we input the original image and the SEEM prompt Image in Figure 1 and got the following result:

```
   scene {
  # Declare objects
  ...

  # Declare spatial relationships
  spatial_relationships: [
    {
     type: "on"
     subject_id: 2
     object_id: 1
    },
    {
     type: "on"
     subject_id: 3
     object_id: 2
    },
    {
     type: "on"
     subject_id: 4
     object_id: 2
    },
    {
     type: "on"
     subject_id: 5
     object_id: 2
    },
    {
     type: "next_to"
     subject_id: 6
     object_id: 2
    },
    {
     type: "behind"
     subject_id: 7
     object_id: 2
    },
    {
     type: "next_to"
     subject_id: 8
     object_id: 2
    },
    {
     type: "next_to"
     subject_id: 9
     object_id: 2
    }
  ]
}
```

Then the on-relationship was acquired as follows:



4.3. Comparison

To validate the effectiveness of our method, we selected ROCA as the baseline and compared with it. We first evaluated the physical realism of the 3D scenes generated by our method by converting these scenes into URDF format and importing them into a simulation environment. Under the influence of gravity, the scene will eventually reach a steady state as in Figure 2. We assessed the realism of our generated 3D scenes by measuring the offset between the objects in the generated scenes and their positions in the final steady state; smaller offsets indicate greater physical realism. The results, shown in the table below, demonstrate that our method significantly improves the realism of the generated 3D scenes.

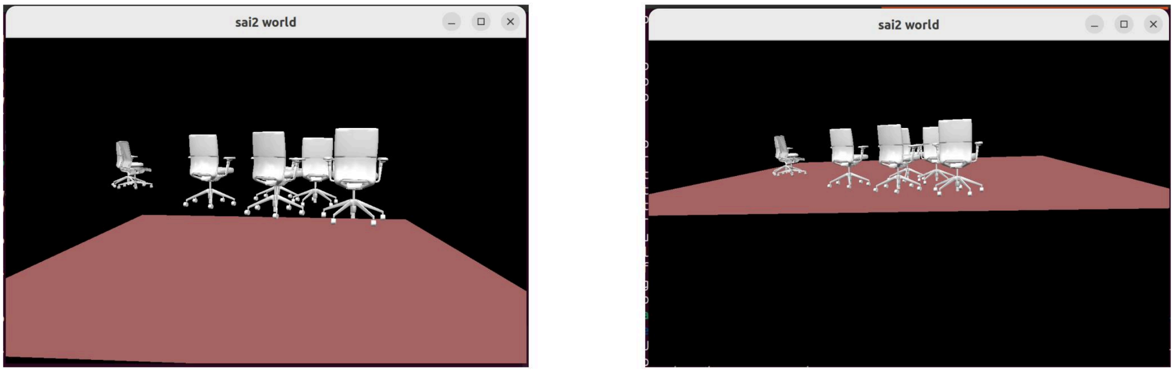
Method	Average Offset
ROCA	0.11 m
Our Method	0 m

Table 1. Comparison of Average Offset

Qualitative comparison results are displayed in Figure 3. It is evident that the baseline method exhibits numerous noticeable issues, such as floating objects or penetration, whereas our method consistently generates correct contacts between objects. We also show the results of overlaying the 3D model onto the input image in Figure 4. The high degree of overlap demonstrates the accuracy of our reconstruction.

5. Conclusion

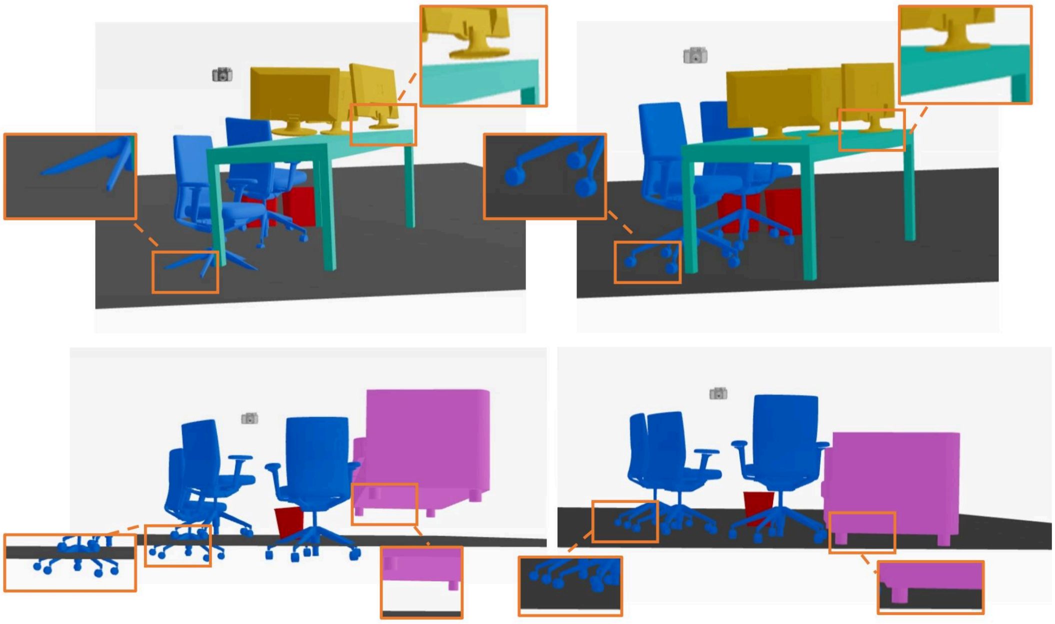
In this paper, we proposed a novel approach for reconstructing realistic 3D indoor scenes from a single image. Our method combines the strengths of large language models (LLMs) and contact-aware refinement to address the



(a) Objects floating in the air.

(b) Objects at steady state.

Figure 2. Objects in simulation.



w/o contact-aware post-processing

w/ contact-aware post-processing

Figure 3. Comparison of our method and the baseline. The baseline results show noticeable issues such as floating objects or penetration, while our method generates correct contact.

challenges of physical plausibility and limited input views. By leveraging Set-of-Mark (SoM) prompting, scene description generation, and spatial relationship extraction, our approach enables the retrieval and alignment of 3D CAD models to the input image. One of the key aspects of our method is the use of a comprehensive domain-specific language (DSL) for representing the structured scene description. Although our current implementation primarily focuses on the ”on” relationship for contact-aware refinement, the DSL is designed to capture a wide range of spatial relationships between objects. This design choice allows for future extensibility and the potential incorporation of additional relationships to further improve the reconstruction quality.

We evaluated our approach on the challenging ScanNet dataset and compared it with the state-of-the-art ROCA

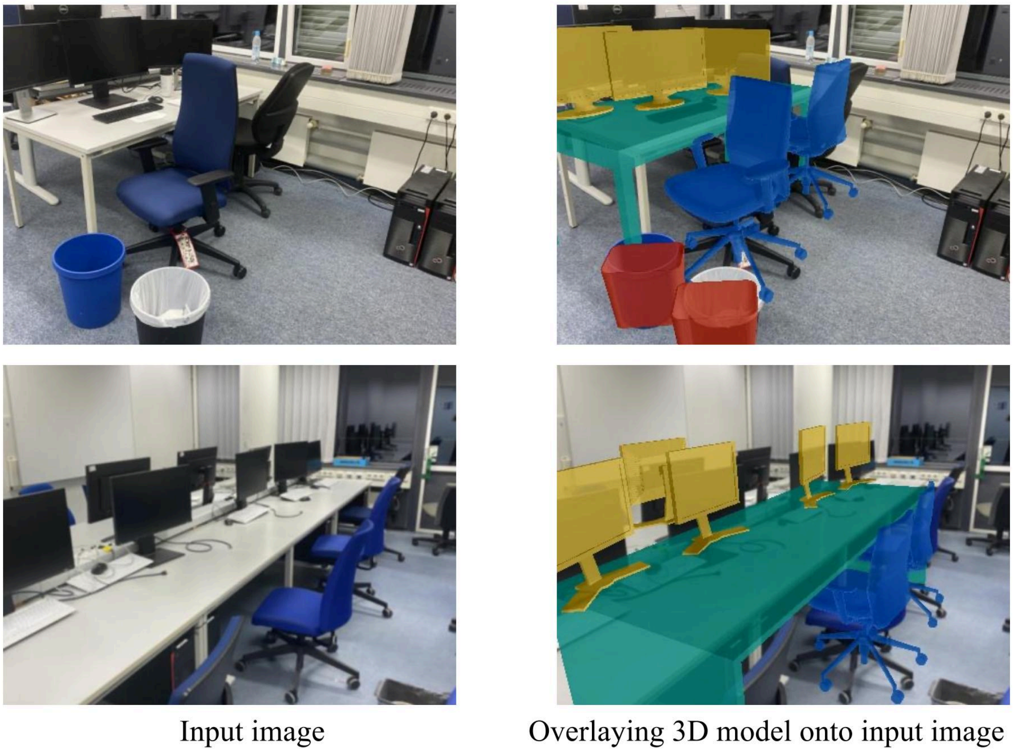


Figure 4. Overlay of the reconstructed 3D model onto the input image.

baseline. The qualitative results showcase the effectiveness of our contact-aware refinement step in generating physically plausible object interactions and resolving issues such as floating objects or interpenetrations. Our method demonstrates improved physical plausibility in the reconstructed scenes, as evidenced by the correct contact between objects in the generated 3D scenes.

To further validate the physical realism of our reconstructions, we converted the generated 3D scenes into URDF format and imported them into a simulation environment. By measuring the offset between the object positions in the generated scenes and their positions in the final steady state under the influence of gravity, we showed that our method produces more physically realistic scenes compared to the baseline. Although our current evaluation focuses on qualitative comparisons and physical plausibility, future work could involve more comprehensive quantitative evaluations to assess the reconstruction accuracy and perceptual realism of our approach.

However, our method has certain limitations that open up avenues for future research. One potential direction is to explore the use of other spatial relationships captured by the DSL, such as ”next to,” ”in front of,” or ”behind,” to handle more complex scene layouts and object interactions. Additionally, incorporating these relationships could help resolve scale ambiguities that may arise when relying solely on the ”on” relationship. Furthermore, extending our approach to handle multiple input views or leveraging additional cues such as depth information could enhance the robustness and accuracy of the reconstructed scenes. Integrating more advanced physics simulation techniques into the contact-aware refinement step could also improve the physical plausibility of the generated scenes. In conclusion, our work demonstrates the potential of combining LLMs with contact-aware refinement for single-image 3D indoor scene reconstruction. By leveraging the power of language models and spatial reasoning, we can generate realistic and physically plausible 3D scenes from limited input views. We believe that our approach lays the foundation for future research in this direction and has the potential to benefit a wide range of applications, from augmented reality and robotics to architectural design and visualization.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 3

[2] Rio Aguina-Kang, Maxim Gumin, Do Heon Han, Stewart Morris, Seung Jean Yoo, Aditya Ganeshan, R Kenny Jones, Qiuhong Anna Wei, Kailiang Fu, and Daniel Ritchie. Open-universe indoor scene generation using llm program synthesis and uncurated object databases. *arXiv preprint arXiv:2403.09675*, 2024. 5

[3] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer, 2006. 2

[4] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. Spie, 1992. 2

[5] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11973–11982, 2020. 2

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2, 3

[7] Can Gümeli, Angela Dai, and Matthias Nießner. Roca: Robust cad model retrieval and alignment from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4022–4031, 2022. 1, 2, 3

[8] Abhinav Gupta, Alexei A Efros, and Martial Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part IV 11*, pages 482–496. Springer, 2010. 2

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[10] Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. In *ACM SIGGRAPH 2005 Papers*, pages 577–584. 2005. 2

[11] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3d scene parsing and reconstruction from a single rgb image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 187–203, 2018. 1, 2

[12] Zhaoyin Jia, Andrew C Gallagher, Ashutosh Saxena, and Tsuhan Chen. 3d reasoning from blocks to stability. volume 37, pages 905–918. IEEE, 2014. 2

[13] Zhizhong Kang, Juntao Yang, Zhou Yang, and Sai Cheng. A review of techniques for 3d reconstruction of indoor environments. *ISPRS International Journal of Geo-Information*, 9(5):330, 2020. 1

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3

[15] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Mask2cad: 3d shape prediction by learning to segment and retrieve. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 260–277. Springer, 2020. 2

[16] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018. 2

[17] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[18] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2

[19] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 1, 2

[20] Giorgia Pitteri, Slobodan Ilic, and Vincent Lepetit. Cornet: generic 3d corners for 6d pose estimation of new objects without retraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 2

[21] Szymon Rusinkiewicz and Marc Levoy. Efficient variants of the icp algorithm. In *Proceedings third international conference on 3-D digital imaging and modeling*, pages 145–152. IEEE, 2001. 2

[22] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. volume 31, pages 824–840. IEEE, 2008. 2

[23] Angtian Wang, Shenxiao Mei, Alan L Yuille, and Adam Kortylewski. Neural view synthesis and matching for semi-supervised few-shot learning of 3d pose. *Advances in Neural Information Processing Systems*, 34:7207–7219, 2021. 2

[24] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 2

[25] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 519–535. Springer, 2020. 2

[26] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *Advances in Neural Information Processing Systems*, 36, 2024. 3