CHAPTER 1

Introduction

1.1 Motivation

With the rapid advancements in science and technology, machines have seamlessly integrated into our daily lives. Now we find ourselves living alongside machines capable of driving cars, organizing our homes, and even assisting in medical surgeries. Central to these advances is the machine's ability to perceive and understand the surrounding environment.

For machines to effectively perceive the three-dimensional world, they need to model the surroundings from sensory data. In particular, accurately representing and reconstructing detailed geometry to their real-life counterparts is vatal for applications in AR/VR, autonomous driving, robotics, etc. Yet, creating detailed geometry from scratch is a labor-intensive task, demanding specialized expertise. Despite the emergence of advanced software and user-friendly modeling tools, challenges like scalability and speed prohibit their large-scale deployment. How to accurately construct geometric details for large scenes at speed is a primary focus of this thesis.

Once the 3D environment is constructed accurately, it is equally important to understand the semantics, affordances, functions, and physical properties of the reconstructed subjects. This kind of holistic understanding is pivotal for machines to really interact intelligently with humans in daily scenarios. However, traditional methods are often tailored for specific tasks, such as 3D semantic segmentation for a limited set of classes, leaving other tasks unaddressed. Achieving a broad understanding of 3D scenes is another objective of this thesis.

Scene representation, i.e. translating observations of an environment, either visual, haptic, auditory, or otherwise, into a concise model of the environment [156, 193], is naturally crucial for machines aiming to tackle complex tasks like accurately reconstructing a realistic scene and having a comprehensive understanding of our world. Recent advances in deep learning, particularly the emergence of Convolutional Neural Networks (CNNs), offer a promising way of deriving robust and powerful scene representations, termed here as *neural scene representations*.

CNNs have revolutionized many computer vision tasks, notably in areas like image classification and depth estimation, showcasing the potential of deep learning in processing visual information. However, much of their prowess is centered on processing 2D information. Transitioning these 2D-focused technologies to 3D environments poses distinct challenges. To effectively model and understand the complex world, it is essential for machines to learn 3D scene representations, enabling a deeper spatial understanding akin to how humans perceive the world.

The goal of this thesis is to pioneer the development of neural scene representations, specifically tailored to accurately reconstruct and comprehensively understand the 3D world. Our roadmap is marked with clear milestones that are all tied together. First, we want to develop a scalable scene representation capable of faithfully reconstructing detailed 3D geometry, spanning from objects to large-scale scenes. Next, with the integration of a novel differentiable point-to-mesh layer, we can represent detailed shapes using just lightweight point clouds, and speed up the 3D reconstruction process. Third, we also investigate a hierarchical neural scene representation that empowers dense RGB-D SLAM applications, specifically for large indoor scenarios. Once obtaining the 3D reconstruction of a scene, the final piece of the thesis is to produce 3D neural scene representations for a plethora of 3D scene understanding tasks, leveraging only a 2D pre-trained model, thus bypassing the need for any costly 3D labeled data.

Overall, this thesis investigates various neural scene representations to produce detailed 3D scene reconstruction efficiently, and subsequently pushes the boundary of 3D scene understanding to another level. In the next section, we will delve into the actual problems and challenges.

1.2 Research Questions and Challenges

In this thesis, we are interested in developing neural scene representations for two different but closely related topics: 3D reconstruction and 3D scene

understanding. We present the following research questions that we try to address in this thesis:

Research Question 1: ***What shape representation is scalable and suitable for detailed 3D reconstruction?***

Shape representations are pivotal for learning-based 3D reconstruction. Explicit shape representations, such as voxels, point clouds, or meshes, have been traditionally favored due to their simplicity. However, each has its limitations: voxels are limited in terms of resolution due to large memory requirements, point clouds discard topological relationships, and predicting mesh-based representations directly via neural networks is challenging. The recent neural implicit representations define shapes implicitly as the level set of a continuous function, parameterized with neural networks [28, 143, 165]. They can model dense surfaces in arbitrary topologies, but often fall short when reconstructing comparably simple objects. Our aim is to advance the neural implicit representations, enabling them to encode complex geometries across diverse topologies and scale to large scenes.

Research Question 2: ***Can we find a representation that is interpretable, lightweight, and facilitates rapid inference?***

As mentioned before, neural implicit representations gained popularity due to their expressiveness and flexibility. However, their reliance on heavy neural networks for encoding surface details often results in slow surface extraction as they require numerous network evaluations in 3D space. This significantly limits its feasibility for applications demanding fast inference. On the other hand, explicit representations like point clouds, require only a few parameters to represent the geometry, and it is very fast to predict. Therefore, our target is to benefit the best from both worlds, leading to a lightweight representation that ensures high-quality reconstruction at low inference times.

Research Question 3: ***How can neural implicit representations be employed for dense SLAM in large scenes?***

While our first two research questions explore optimal shape representations for 3D reconstruction from input point clouds, A more realistic scenario for 3D reconstruction is to densely model a scene solely from an unposed RGB(-D) sequence. This falls into the category of dense visual SLAM. Traditional dense visual SLAM systems are often unable to estimate plausible geometry for unobserved regions. Although recent SLAM approaches using neural implicit representations attain a certain level of predictive power, they

are typically confined to smaller scenes due to their reliance on suboptimal neural scene representations. We want to circumvent this limitation by introducing a novel hybrid representation, enabling the neural-implicit-based SLAM system for large-scale scenes.

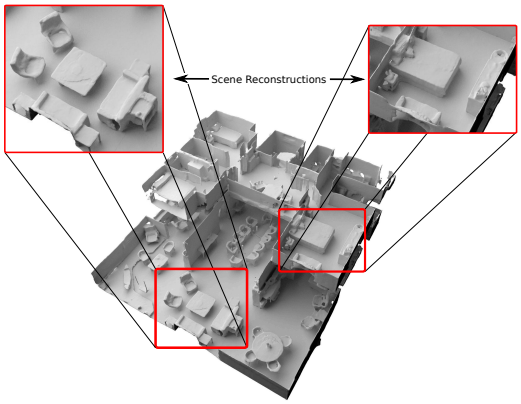
Research Question 4: ***How to generate a unified neural representation for a broad set of 3D scene understanding tasks without any 3D supervision?***

Upon addressing the first three research questions, we can assume having obtained the 3D geometry of a scene. One natural downstream application is the understanding of this reconstructed scene. Previous learning-based methods usually handle one single 3D scene understanding task at a time, in a fully-supervised learning manner. Our aspiration instead is to develop a zero-shot method, producing a neural scene representation capable of inferring 3D semantics, affordances, physical properties, and beyond.

1.3 Contributions

This thesis addresses the research questions outlined earlier and contributes to the instigation of learning neural scene representations for 3D reconstruction as well as 3D scene understanding. Specific contributions are detailed as follows.

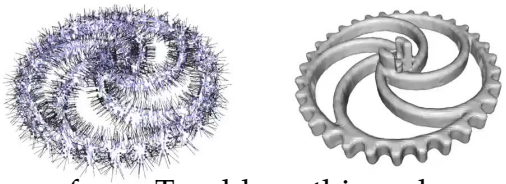
1.3.1 3D Reconstruction with Scalable Neural Representations



Neural implicit representations have emerged as a popular choice for learning-based 3D reconstruction since it they can capture 3D shapes in a continuous, resolution-independent, and topologically flexible manner. However, most implicit-based approaches struggle with complex geometries and larger scenes. This limitation often stems from their simple fully-connected network architecture which does not allow for integrating local information in the observations or incorporating inductive biases such as translational equivariance. To address this, we propose *Convolutional Occupancy Networks*, a novel flexible implicit representation for detailed reconstruction of objects and 3D scenes. Our model incorporates

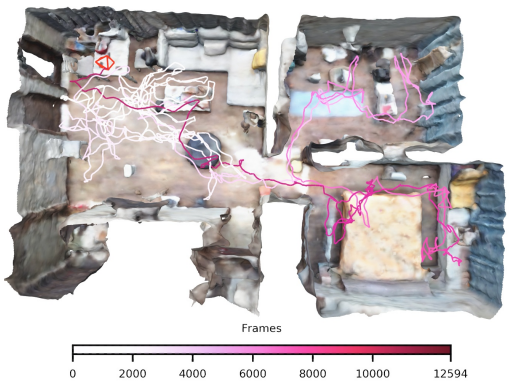
inductive biases by combining convolutional encoders with implicit occupancy decoders, enabling structured reasoning in 3D space. Our evaluations show that our method enables the fine-grained implicit 3D reconstruction of single objects, scales to large indoor scenes, and generalizes well from synthetic to real data.

1.3.2 3D Reconstruction with a Differentiable Poisson Solver



While our scalable neural implicit representations show promising results in detailed reconstruction, the inference process remains time-consuming due to the numerous network evaluations for extracting surfaces. To address this problem, we revisit the classic yet ubiquitous point cloud representation and introduce a differentiable point-to-mesh layer using a differentiable formulation of Poisson Surface Reconstruction (PSR), which allows for a GPU-accelerated fast solution of the indicator function given an oriented point cloud. The differentiable PSR layer bridges the explicit 3D point representation with the 3D mesh via the implicit indicator field, enabling end-to-end optimization. This duality between points and meshes hence allows us to represent shapes as oriented point clouds, which are explicit, lightweight, and expressive. Our *Shape-As-Points* (SAP) model is interpretable, lightweight, and accelerates inference time by one order of magnitude compared to neural implicit representations, but could still produce topology-agnostic, high-fidelity watertight surfaces.

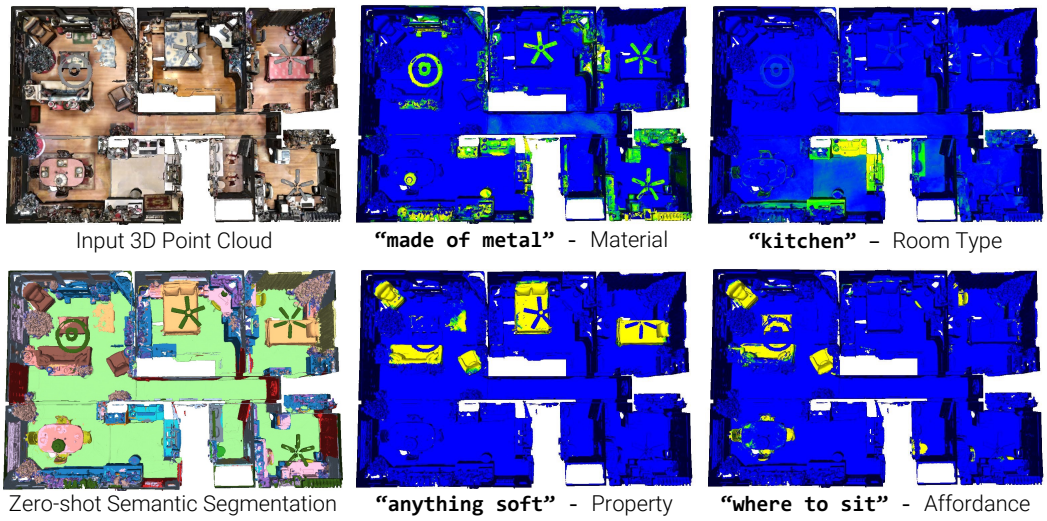
1.3.3 SLAM with Scalable Neural Representations



While our earlier contributions focused on 3D reconstruction from point clouds, a more practical setting for 3D reconstruction is to reconstruct 3D dense scene geometry given only some unposed RGB(-D) sequences with a hand-held camera, i.e. dense visual SLAM. To this end, we present *NICE-SLAM*, a dense SLAM system that employs a hierarchical neural implicit representation. Optimizing this representation with pre-trained geometric priors enables detailed reconstruction on large

indoor scenes, outperforming recent neural implicit SLAM systems in scalability, efficiency, and robustness.

1.3.4 3D Scene Understanding with Large Vision Language Models



Once we obtain the realistic 3D reconstruction of a scene, the aim for the last part of this thesis is high-level perception tasks, such as 3D scene understanding. Traditional 3D scene understanding approaches have largely depended on supervision from benchmark datasets tailored for specific tasks, such as 3D semantic segmentation, often confined to a closed set of classes. Such specialized models, while adept in their designated task, are impractical for many real-world applications as the models lack the flexibility to continuously adapt to new concepts/classes in the scene.

Addressing this challenge, recent advancements, including our work *OpenScene* discussed in Chapter 6 emphasize open-vocabulary 3D scene understanding. This approach allows segmentation and understanding of arbitrary concepts, independent of any fixed closed set of classes. Specifically, given an arbitrary query like a text description or an image of an object, the goal is to segment those parts in the 3D scene that are described by the query. For example, within a reconstructed house as shown above, we are interested in understanding which surfaces are part of "`a bed`" (semantics), "`made of metal`" (materials), "`kitchen`" (room types), "`where to sit`", and which surfaces are "`soft`" (physical property). Such capabilities not only offer a richer understanding but are also pivotal for applications such as facilitating robot navigation in unfamiliar settings or enhancing AR/VR experiences in complicated indoor scenarios, especially when specific annotated labels are sparse.