Chapter 1

Introduction

1.1 Background

Automation has been the main driving force of industrialization and technological innovations. Throughout history, we have delegated mundane or dangerous tasks to machines, which allowed the economies of scale and the foundation of the modern world of unprecedented prosperity and comfort. However, until recently, the main area of automation has been limited to mechanical tasks that do not require high-level cognition. For instance, factories automated an assembly line by decomposing it into countless tasks that only require simple repetitive motion. Tasks that require perception or cognition such as sorting out defects and picking up items in disarray have been the last frontier of the automation. Similarly, in the transportation sector, automating the delivery of goods and taxiing passengers has been one of the greatest challenges.

Many of these frontiers of automation require high-level cognition and predicate on reliable visual perception. For example, in autonomous driving, an autonomous agent needs to identify the road, or areas safe to drive on, to detect pedestrians and other vehicles, to track their location, speed, and to predict their paths. Similarly, in automated warehouses, robots need to identify objects and their shapes to grasp, handle, and package items. In virtual reality and augmented reality applications, we need to track the location of the head-mounted display on a user as well as identifying objects around the user to simulate an environment that users can experience the seamless integration of the illusion and the world.

Specifically, all these tasks require estimating various types of 3D information from sensors. For instance, in autonomous driving and navigation, an agent must localize itself first in the 3D space and in a high-definition 3D map. This process requires identifying geometry and finding correspondences between the map and the sensory inputs. Also, it must recognize all moving objects that might interfere with its path and this recognition includes estimating the velocity, location, and relative distance of others from itself in the 3D space as (Fig. 1.1). In augmented reality or mixed reality

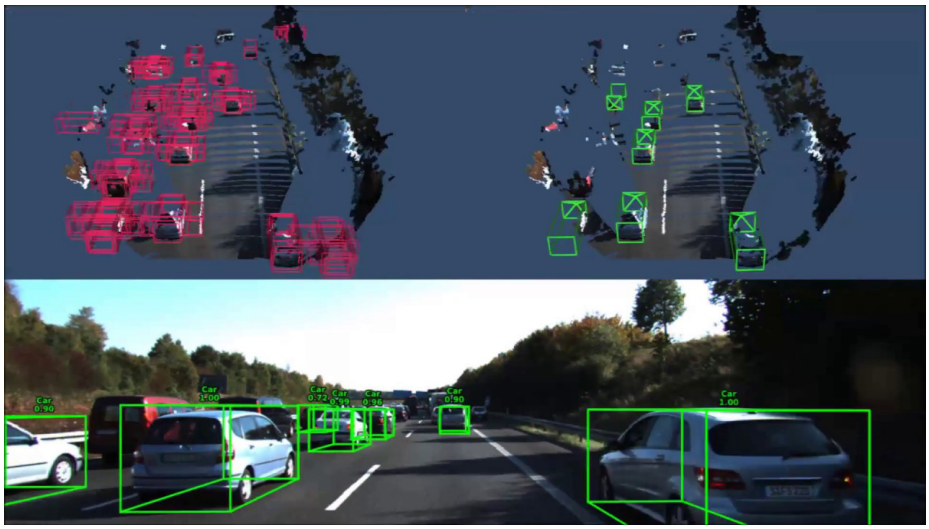

Figure 1.1: Object detection and pose estimation in 3D by Ku *et al.* [5]: (bottom) an autonomous driving setup and 3D bounding boxes overlaid on top of detected cars. (top) The same scene in 3D space from bird's-eye-view (BEV) perspective.

applications, a head-mounted device must localize itself in the 3D environment first to portray relevant information to a user as well as identify objects to accurately simulate the illusion. For example, if a user needs a direction, it needs to localize itself in the world and points the direction to the user. Also, in mixed reality games, a head-mounted device must identify the 3D geometry of the surrounding 3D world to display virtual objects accurately on top of the 3D geometry with correct lighting and shadow (Fig. 1.2).

All these perception problems require an understanding of the underlying 3D world and are known as 3D perception problems. Formally, we define 3D perception as an ability to represent, identify, and interpret visual three-dimensional data to understand the underlying three-dimensional structure in the real world. As it is an important problem domain with various applications, 3D perception includes diverse sub-problems such as generating the 3D structure from single or multi-view 2D images (3D reconstruction), estimating 3D orientation or pose of an object (3D pose estimation), localizing an observer or an agent in 3D space, and semantic understanding of the 3D structure of the objects and scenes.

As there are many important applications in 3D perception, many algorithms have been proposed to tackle these problems; and recently, convolutional neural networks (convnets) have proven to be one of the most effective algorithms for 3D perception. Specifically, 3D convolutional network for reconstruction [2, 4], 3D object detection [11, 9, 6], and semantic segmentation [8, 10] have shown that neural networks beyond image domain can be efficient for 3D perception. Some of these methods can even process new types of 3D data modalities that traditional image-based convolutional neural networks cannot process such as 3D point clouds from Time-of-Flight (ToF) scanners. The data

Figure 1.2: 3D reconstruction of an indoor scene by Niessner *et al.* [7].

from ToF scanners are a list of 3D coordinates where the sensors observed reflection from and are scattered in the 3D space or even the 4D spatio-temporal space. There are even higher-dimensional data readily available to use in 3D perception and some of which arise in registration problems where we want to fit an image or a 3D scan to another image or 3D scan.

In this dissertation, we propose a set of high-dimensional convolutional neural networks for 3D reconstruction to 4D spatio-temporal perception and 6-dimensional convolutional networks for registration. We also verify our approach on high-dimensional synthetic datasets that range from 4-dimension to 32-dimension.

1.2 Thesis Outline

This dissertation is structured into three parts: 3D reconstruction, representation learning, and registration. We summarized the contents of each chapter in Tab. 1.1. Each chapter deals with one of the multiple problems in spatially high-dimensional spaces.

1.2.1 Reconstruction

3D reconstruction is the first step that generates 3D point clouds or meshes from a set of images. In this chapter, we present supervised reconstruction methods using 3D convolutional neural networks that take a set of images as input and generates a 3D occupancy pattern in a grid. To train the network, we use a large-scale 3D shape dataset to generate a set of images rendered from various viewpoints. A random subset of these images from various viewpoints forms a set of inputs and the target labels are discretized 3D shapes in a rectangular grid where 1 indicates the interior of the shape and 0 otherwise. We validate the approach on real image datasets and analyze the method on

Table 1.1: Overview of the dissertation

	Chapters	Contents	Dimension
Part I: Reconstruction	Ch. 2	Fully-supervised reconstruction	3D
	Ch. 3	Weakly-supervised reconstruction	
Part II: Representation learning	Ch. 4	Sparse tensor networks	3D, 4D
	Ch. 5	Spatio-temporal segmentation	
	Ch. 6	Geometric feature learning	
Part III: Registration	Ch. 7	Geometric pattern recognition	4D, 6D
	Ch. 8	Global registration	

various aspects including invariance to some of the failure modes of classical reconstruction methods. In Chapter 3, we relax the requirements of the networks to weaker and cheaper supervision for 3D reconstruction. Note that the supervised reconstruction requires 3D shapes as targets which are created by experts or requires a large number of images to create a 3D model. Instead, the framework makes use of foreground masks in addition to unlabeled real 3D shapes and generates the same 3D occupancy grid as an output. However, as silhouettes can only provide convex hull of the 3D shapes, we create an additional constraint that guides the reconstructions to be valid shapes by defining a constraint that penalizes invalid 3D shapes. As the constraint is defined in a high-dimensional space and requires a semantic understanding of shapes, we learn the constraint with adversarial training. Combined with the learned constraint, we train the reconstruction system with as few as 1 image and the corresponding silhouette and show that the proposed model can sufficiently reconstruct objects without direct 3D supervision.

1.2.2 Representation Learning

3D representation learning is the first step of high-level perception in many applications. However, for large-scale 3D scenes, it is difficult to use traditional dense tensor representations for 3D perception due to the memory and computational complexity. Unlike image-domain representations, 3D scans or point clouds capture the surface of objects or a scene which results in a sparse set of points or meshes of the surface to represent 3D data. Thus, using 3D volumetric tensors to save 3D data is inefficient as it requires more memory and computation to save and process empty spaces between the 3D surfaces of the scene. Instead, sparse tensor representations have been introduced to process and learn sparse data efficiently. Neural networks that can process these sparse tensors are known as the sparse tensor networks and we introduce the basic operations for these networks.

We first apply these networks on a sequence of 3D scans and propose 4-dimensional convolutional neural networks for the spatio-temporal perception that can directly process such 3D-videos using high-dimensional convolutions. The network takes a sequence of For this, we adopt sparse tensors [3, 1] and propose the generalized convolution which encompasses all discrete convolutions. To implement the generalized convolution, we create an open-source auto-differentiation library for sparse tensors that provides extensive functions for high-dimensional convolutional neural networks. We create 4D spatio-temporal convolutional neural networks using the library and validate them on various 3D semantic segmentation benchmarks and proposed 4D datasets for 3D-video perception. To overcome challenges in the high-dimensional 4D space, we propose the hybrid kernel, a special case of the generalized convolution, and the trilateral-stationary conditional random field that enforces spatio-temporal consistency in the 7D space-time-chroma space. Experimentally, we show that convolutional neural networks with only generalized convolutions can outperform 2D or 2D-3D hybrid methods by a large margin. Also, we show that on 3D-videos, 4D spatio-temporal convolutional neural networks are robust to noise, outperform 3D convolutional neural networks, and are faster than the 3D counterpart in some cases.

In the next chapter, we use these sparse tensor networks for geometric representation learning. The geometric features capture local shapes accurately for correspondences and registration. Unlike many state-of-the-art methods that require computing low-level features as input or extracting patch-based features with the limited receptive field, we present fully-convolutional geometric features. Given a 3D scan of an environment, the network processes the input fully convolutionally and extracts low dimensional features densely on all points in the 3D scan. We also present new metric learning losses that dramatically improve performance. Fully-convolutional geometric features are compact, capture broad spatial context, and scale to large scenes. We experimentally validate our approach on both indoor and outdoor datasets. Fully-convolutional geometric features achieve state-of-the-art accuracy without requiring prepossessing, are compact (32 dimensions), and are 600 times faster than the most accurate prior method.

1.2.3 Registration

In the last part of the dissertation, we discuss high-dimensional pattern recognition problems in image and 3D registration. In Chapter 7, we present high-dimensional convolutional networks for high-dimensional pattern recognition problems in 2D and 3D registration. We first propose high-dimensional convolutional networks from 4 to 32 dimensions and analyze the geometric pattern recognition capacity of high-dimensional sparse tensor networks for high-dimensional linear regression problems. We form high-dimensional sparse tensors with discretized coordinates of data points, which we then predict the likelihood that each point belongs to the main pattern or noise. Next, we show that the 3D correspondences form hyper-surface in a 6-dimensional space and use the proposed

high-dimensional convnets for predicting point-wise likelihood whether the point belongs to the 6-dimensional hyper-surface or noise. Finally, we use image correspondences as inputs, which form a 4-dimensional hyper-conic section, and show that the high-dimensional convolutional networks are on par with many state-of-the-art multi-layered perceptrons.

In the next chapter, we extend the proposed high-dimensional convnets for differentiable 3D registration. We propose three core modules for this: a 6-dimensional convolutional neural network for correspondence confidence prediction; a differentiable Weighted Procrustes method for closed-form pose estimation; and a robust gradient-based SE(3) optimizer for pose refinement. Experiments demonstrate that our approach outperforms the state-of-the-art learning-based and classical methods on real-world data while maintaining efficiency.

Bibliography

[1] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *CVPR*, pages 3075–3084, 2019.

[2] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[3] Benjamin Graham. Spatially-sparse convolutional neural networks. *arXiv preprint arXiv:1409.6070*, 2014.

[4] JunYoung Gwak, Christopher B Choy, Manmohan Chandraker, Animesh Garg, and Silvio Savarese. Weakly supervised 3d reconstruction with adversarial constraint. In *3D Vision (3DV), 2017 Fifth International Conference on 3D Vision*, 2017.

[5] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven Waslander. Joint 3d proposal generation and object detection from view aggregation. *IROS*, 2018.

[6] D. Maturana and S. Scherer. VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition. In *IROS*, 2015.

[7] Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Marc Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6):1–11, 2013.

[8] Gernot Riegler, Ali Osman Ulusoy, and Andreas Geiger. Octnet: Learning deep 3d representations at high resolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[9] Shuran Song and Jianxiong Xiao. Deep Sliding Shapes for amodal 3D object detection in RGB-D images. In *CVPR*, 2016.

[10] Lyne P Tchapmi, Christopher B Choy, Iro Armeni, JunYoung Gwak, and Silvio Savarese. Segcloud: Semantic segmentation of 3d point clouds. *International Conference on 3D Vision (3DV)*, 2017.

[11] Dominic Zeng Wang and Ingmar Posner. Voting for voting in online point cloud object detection. In *Robotics: Science and Systems*, volume 1, pages 10–15607, 2015.