ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Autonomous Systems Lab
Prof. Roland Siegwart

Master Thesis

Leveraging Deep Learnt Scene Completion for Fast Autonomous Exploration Planning and Mapping

Autumn Term 2021

Supervised by:
Lukas Schmid
Victor Reijgwart

Author:
Mansoor Nasir Cheema

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

LEVERAGING DEEP LEARNT SCENE COMPLETION FOR FAST AUTONOMOUS EXPLORATION PLANNING AND MAPPING

Authored by (in block letters):

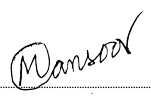
For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
CHEEMA	MANSOOR NASIR

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Munich, Germany 14.12.2021	

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.

Contents

Abstract

Micro Aerial Vehicles (MAVs) have demonstrated an increasing capability to independently explore and navigate unknown environments, enabling widespread applications in search, inspection and monitoring. However, their exploration capability is constrained by their ability to accurately map their surroundings from onboard sensors that are inherently susceptible to noise and occlusions. Current mapping approaches require robots to observe their surroundings from abundant viewpoints to plan effectively. 3D Semantic Scene Completion (SSC) Networks [1] have been shown to complete partial scenes by inferring occluded regions using deep learnt priors. The thesis introduces a novel Scene Completion based semantic mapping and exploration planning pipeline that facilitates efficient exploration without having to rely on redundant observations. Various fusion strategies are investigated for building the scene completion aware map and a probabilistic strategy is proposed to incrementally fuse scene completions into a grid based volumetric map. A conventional informative exploration planner is extended to take advantage of deep-learned priors by the introduction of a scene completion gain that incorporates the completions in the planning pipeline. The complete autonomous exploration pipeline is evaluated in a realistic simulation. The results indicate that the proposed system achieves superior exploration capability while maintaining the safety of conventional approaches.

Symbols

Acronyms and Abbreviations

ETH	Eidgenössische Technische Hochschule
ASL	Autonomous Systems Lab
SSC	Semantic Scene Completion
SC	Scene Completion
ML	Machine Learning
MAV	Micro Aerial vehicles
TSDF	Truncated Signed Distance Fields
ESDF	Euclidean Signed Distance Fields

Chapter 1

Introduction

1.1 Motivation

Micro Aerial Vehicles (MAVs) have experienced widespread adoption in consumer and industrial applications given their agility, maneuverability, and ability to operate in diverse environments. In recent years, MAVs have demonstrated an increasing level of autonomy, from following predefined trajectories to autonomously navigating unknown environments. The autonomous exploration capability is very suitable for information gathering and monitoring tasks, which find applications in industrial inspections [1] [2], crop monitoring[3], search and rescue missions [4], particularly in situations not convenient or safe for humans.

Autonomous exploration in GPS denied environments without a-priori map information is challenging as the robot has to concurrently construct a map while relying on it for planning. The mapping process from ubiquitous visual sensors is sensitive to occlusions, causing missing regions and artifacts in the mapped space. Current methods rely on redundant observations from multiple views to ensure occluded regions are mapped, which affects the planning capability of the robot. In contrast to robots, humans are very efficient in moving around in completely unknown environments and can perceive much more than they can see. Humans rely on their inherent familiarity with the structural patterns in the world and can extrapolate from limited views.

The thesis proposes a novel method to address the limitations of current approaches by introducing a human inspired mapping system that infers occluded regions from partial observations. The method takes advantage of recent advances in Semantic Scene Completion (SSC) [5], which use deep learning to learn structural priors and predict occluded parts. The SSC based mapping and planning pipeline[1] extends current state of the art approaches [6] [7] with the following contributions:

- Scene completed mapping that employs deep learnt priors to extrapolate occluded regions from partial observations
- Introduction of probabilistic fusion strategy to incrementally build scene completed mapping
- Smart exploration planner that leverages scene completion for planning in unknown regions while ensuring safety constraints.

[1] https://github.com/mansoorcheema/ssc_3d_planning

1.2 Problem statement

This thesis addresses the problem of developing a scene completion based mapping and planning approach $\mathcal{P}$ that maximizes exploration of voxelized space $\mathcal{W} \in \mathbb{R}^3$. Let $\omega \subset \mathcal{W}$ be total explorable space, $\mathcal{T} \sim \mathcal{P}$ be the trajectory generated by approach $\mathcal{P}$, the observed space by a trajectory $\mathcal{T}$ as $\mathcal{O}^{\mathcal{T}} \subset \omega$ where $\mathcal{O}^{\mathcal{T}} = \mathcal{O}^{\mathcal{T}}_{free} \cup \mathcal{O}^{\mathcal{T}}_{occupied}$, the objective is to find optimal approach $\mathcal{P}^*$ that generates trajectories maximizing objective from equation. 1.3 while ensuring safety constraints. The safety constraints require identification of unoccupied space with very high precision and detection of all obstacles correctly to ensure collision free trajectories.

$$\omega := \{v | v \in \mathcal{W} \ \wedge \ \texttt{explorable}(v)\} \tag{1.1}$$

$$\texttt{observed}^{\mathcal{T}}(v) = \begin{cases} 1, & \text{if } v \in \mathcal{O}^{\mathcal{T}}_{free} \\ 1, & \text{if } v \in \mathcal{O}^{\mathcal{T}}_{occupied} \\ 0 & \text{otherwise} \end{cases} \tag{1.2}$$

$$\arg\max_{\mathcal{T} \sim \mathcal{P}} \sum_{v \in \omega} \texttt{observed}^{\mathcal{T}}(v) \tag{1.3}$$

$$\text{s.t} \quad \texttt{occupancy}(\mathcal{O}^{\mathcal{T}}(v)) = \texttt{occupancy}(\omega(v)) \tag{1.4}$$

$$\text{s.t} \quad \texttt{precision}(\mathcal{O}^{\mathcal{T}}_{free}) \rightarrow 1 \tag{1.5}$$

$$\text{s.t} \quad \texttt{recall}(\mathcal{O}^{\mathcal{T}}_{occupied}) \rightarrow 1 \tag{1.6}$$

In above equations, $\texttt{occupancy}(\mathcal{O}^{\mathcal{T}}(v))$ denotes the occupancy status of voxel v in discretized observed space from trajectory $\mathcal{T}$. The possible occupancy states are free, occupied and unknown for unobserved space. Similarly, $\texttt{occupancy}(\omega(v))$ denotes the true status of space at discretized voxel location v.

Chapter 2

Related work

2.1 3D Semantic Scene Completion

Deep Learning has enabled remarkable results for visual recognition tasks by learning geometric and semantic representations, implicitly capturing relevant features for a given problem [8]. Such learned representations are particularly effective for extrapolating underlying patterns to complete partial observations and have proven success in solving ill-posed problems like monocular depth estimation [9][10][11] or single view reconstruction [12].

Song et.al [13] proposed SSCNet, an end-to-end 3D CNN to solve a similar ill-posed problem of semantic scene completion from a single depth image. SSCNet demonstrated that jointly learning semantic and geometric completion enhanced the capability of scene completion, which they attribute to close inter winding nature of semantics and geometry. DDRNet [14] followed the semantic scene completion introduced by SSCNet and proposed additionally incorporating RGB images for semantics. Furthermore, DDRNet addressed the computation intensity of SSCNet and introduced efficient 3D convolutions by decomposing a 3D convolution into 3 sub-convolutions across channels, while surpassing SSCNet in both performance and efficiency.

Liu et. al [15] proposed that the position is decisive for the scene completion problem as the boundaries and corners of objects are critical for defining the structure of objects compared to regions without significant structural variation. To address this, they introduced PALNet- a CNN that used a weighted cross-entropy loss function for weighting different regions based on their geometrical significance. Furthermore, PALNet introduced a hybrid architecture for extracting 2D features from the depth map in addition to 3D feature extraction from TSDF encoded depth map. The PALNet improved upon DDRNet without a significant increase in network parameters.

CCPNet[16] addressed low-resolution completions from earlier SSC networks by using separated kernels for 3D convolutions and an efficient feature aggregation technique to significantly reduce the computation complexity, thus eliminating the need for down-sampling input resolution. CCPNet's novel feature aggregation module captured multi-scale feature contexts from different convolutional layers and then combined the features in a cascaded manner with low-level local features. In contrast, earlier approaches [13][14][15] combined features in a serial or parallel mechanism that affected understanding of multi-scale contexts. By incorporating grouped convolutions, Cascaded Feature Aggregation Module with dilated convolutions, CCPNet introduced the fastest and best performing single-stage end-to-end SSC network. Followed by CCPNet, a more complex multistage approach 3D sketch-

aware semantic scene completion [17] offered marginally increased performance with a tremendous increase in complexity.
In spite of significant research on Semantic Scene Completion (SSC) approaches, there remain limited works targeting online real-time scene completion. Chen et. al [18] proposed SCFusion- a lightweight 3D CNN based real-time semantic scene completion and incremental fusion pipeline. Their SSC Network is based on image in-painting approach [19] and completes missing regions in partially scanned volumes using binary masks marking areas for scene completion. The network consists of a GAN-based adversarial network trained to generate realistic completions enforced by a discriminator and a supervised cross-entropy loss based on ForkNet [20].

2.1.1 Incremental Fusion

Though scene completions have been geared for robotic navigation purposes, there remains limited research on incremental scene completed mapping. At the time of writing, SCFusion [18] remains the first and only approach for building an incremental semantic scene completed map. SCFusion [18] uses a probabilistic grid-based voxelized map as volume representation. The measured depth image is projected to 3D and fused probabilistically into the global map. Similarly, the sub-volumes in view frustum are extracted and used as an input to SCFusionNetwork for completion. The scene-completed volumes are then fused into the global map, albeit with a lower probability compared to the scanned measurements.
SCFusion [18] provides a valuable groundwork for incremental scene completed mapping that could potentially be employed for robot navigation tasks, however, SCFusion [18] fuses the scanned pointcloud and scene completions into a single map that poses safety concerns. Additionally, SCFusion only fuses the occupancy which restricts the planning capability of the robot as the robot has to identify free regions where it could potentially move.
To address these limitations, A two-map-based approach is introduced that maintains the incremental scene completed map separately from the scanned map, and has the following advantages.

- Ensures safety of conventional planning approaches by relying on the scanned map for the safety-critical stage of planning
- Uses valuable scene completions for non-safety-critical information planning
- Fuses both free and occupied space using multiple fusion strategies
- Possibility to leverage the combination of measured map and predicted map depending upon planning use-case and risks (controllable safety and performance trade-off)

2.2 3D Volumetric Mapping

3D mapping process involves capturing a geometrical representation of a scene from visual and geometrical sensors, where sensors like depth cameras or laser scanners are used to generate a 3D pointcloud of a surface and subsequently estimate scene geometry. 3D mapping is a well-studied problem and finds its use cases across a wide range of applications from Augmented Reality, 3D Reconstruction, to Robotic perception. In recent years, dense volumetric mapping approaches have experienced rapid adoption for perception and navigation tasks with availability of efficient real-time implementations like [6] for robotic applications. Volumetric mapping

approaches represent the environment using grid-based occupancy structures arranged in voxels where each voxel stores surface information. Such dense voxelized representation provide detailed surface occupancy information in $\mathbb{R}^3$ up to voxel resolution, making it convenient to identify free traversable regions for path planning. Alternate map representations like Meshes and Surfels [21] though accurately capture surface information, don't readily provide dense occupancy information required for planning. Some of the most widely used dense volumetric mapping surface representations [22] are introduced below:

2.2.1 Truncated Signed Distance Fields

Curless et. al [23] introduced a volumetric approach for implicitly modeling geometry as Signed Distance Fields by denoting the surface as zero crossing of the signed distance value. The volume is often represented by discrete voxel structures arranged in a grid storing the distance to the nearest surface. The distance value for each voxel is calculated by averaging fused pointcloud measurements using ray casting technique. TSDF based approaches have experienced widespread adoption since Kinect fusion [24] introduced real-time dense reconstruction from affordable handheld devices. The suitability of TSDF voxels for encoding geometrical information and 3D mesh generation using Marching Cubes [25] has contributed to its use for 3D reconstruction and robot mapping tasks. Kinect Fusion, however, had limited suitability for mapping large environments efficiently. Niessner et al. [26] introduced voxel hashing approach enabling real-time 3D mapping of large-scale environments. Voxel hashing organized voxels in blocks for efficient search irrespective of environment size and efficiently stored voxels only in regions with observations. Modern approaches for real-time volumetric mapping for robotic applications like Voxblox [6] follow similar approach of voxel hashing for robotic applications.

2.2.2 Euclidean Signed Distance Field

Euclidean Signed Distance Field (ESDF) is an approach similar to SDF and TSDF that encodes Euclidean distance to the nearest surface boundary. ESDF addresses the limitations of SDF and TSDF approaches that store projected distance information, which is better suited for rendering than planning. Voxblox [6] proposed a real-time approach for incrementally constructing Euclidean signed distances from TSDF by addressing shortcomings of earlier ESDF approaches where ESDF was generated on a batch by batch basis. ESDF volumetric maps find a very important use case in robot planning advantaging from convenient euclidean distance formulation. ESDF maps have been used for a wide range of planning applications from trajectory optimization based [27] approaches, to widely used sampling [7][28] based approaches.

2.2.3 Probabilistic Occupancy Mapping

Another widely used approach for volumetric mapping involves fusing scanned pointcould probabilistically. This approach has been tested with grid-based volumetric mapping [6] and Tree-based structures [29]. In contrast to binary occupancy mapping methods, probabilistic methods allow continuous values for detailed surface representation.
The drawback of grid-based volumetric approaches is the uniform voxel resolution evident in environments containing objects of various sizes that demand increased resolution in complex areas. Octree [29] based approaches work around this limitation to model arbitrary environments at various resolutions without prior as-

sumptions. Octree offers a very efficient data structure for storing and processing occupancy data.

2.3 Exploration Planning

Exploration planning addresses autonomous exploration of unknown environments to inspect, observe and map a target environment. It differs from the classical path planning approaches with specified start and destination, as the absence of an itinerary and initial map makes it more challenging to map and explore concurrently. Frontier-based classical approach by Yamauchi et al. [30] focuses on identifying regions at the boundary of observed free space and unexplored regions, encouraging robots to uncover further area while extending the observed map. This classical approach has been shown to perform well in exploring diverse environments from large open areas to small indoor cluttered spaces.
Another approach by H. González-Baños et al. [31] follows a Next Best View (NBV) [32] scheme and relies on concurrently exploring sub-regions and patching them together to build a larger map. In contrast to the Next Best View(NBV) approach from [32], the authors [31] propose ranking viewpoint choices based on how good a viewpoint is and how much volume can potentially be explored from that point. The limitation of this approach is that alignment and merging of sub-models is not robust as it struggles with small obstacles at the boundaries of sub-models which are lost while merging explored sub-regions due to inevitable small model alignment errors.
Shen et al. [33] introduced an efficient frontier-based 3D exploration planner designed specifically for MAVs. Shen et al. [33] identify potential frontiers by evolving stochastic differential equations simulating the system of particles and relate higher particle expansion to larger explorable regions. Once the frontiers are identified, the known map is expanded by flying toward those regions. The authors demonstrated that their approach surpasses traditional frontier-based approaches in both 3D and 2D. Their approach proposes mapping only occupied regions while ignoring unoccupied areas which makes it challenging to use with standard planning frameworks that use free-space to plan.

2.3.1 Informative Path Planning

Informative Path Planning helps robots estimate information gained from potential trajectories and prefer trajectories that maximize an objective function. It provides the flexibility to adapt a general planning problem as an exploration problem by formulating objective criteria promoting the desired exploratory behavior.
Informative gains have been explored to adapt widely used sampling-based general path planners. Schmid et al [7] proposed RRT* inspired informative planner by associating information gain to sampled paths and expanding an existing trajectory tree. Similarly, Glocal [28] uses a graph-based local planner consisting of viewpoints affixed by edges with assigned gains and costs. The exploratory behavior is encouraged by counting the unexplored voxels as an information gain, similar to [7] and encouraging the planner to plan in the direction of large unexplored voxels. In the following chapters, this gain-based formulation shall be extended to promote smarter autonomous exploration.

Chapter 3

Method

3.1 Overview

A complete pipeline for deep learnt mapping and exploration planning was developed for this thesis by extending a conventional dense mapping and planning [6] framework to advantage from scene completions. Furthermore, an incremental fusion framework is proposed for maintaining a grid-based volumetric map for fusing scene completions. The figure 3.1 shown an overview of the approach with the highlighted modules in green showing the modules that were extended or introduced for this thesis. A summary of the workflow is briefly introduced below, followed by the description of each module in the subsequent sections.

1. Robot Simulation provides a RGB and depth scan from on-board imaging sensors
2. Mapping module generates a colored pointcloud from RGB-D scan and fuses the pointcloud into a TSDF voxel-based volumetric mapping following Voxblox[6]
3. Scene Completion module predicts the completed geometry from the depth map as a volume of semantic probabilities
4. The semantic scene completed volume is fused probabilistically into a grid-based occupancy mapping
5. The measured TSDF map and predicted occupancy map are used together for providing the planner with occupancy information
6. Planner generates and evaluates trajectories employing the scene completed mapping

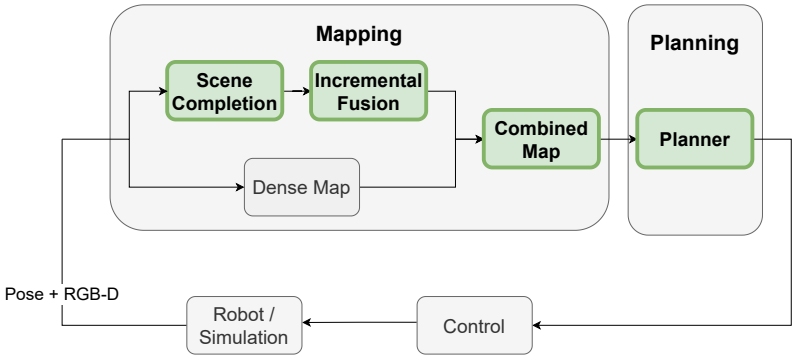


Figure 3.1: Method Overview

3.2 3D Scene Completion

The 3D scene completion module is a Deep Learning based semantic scene completion (SSC) network adapted from an open-source implementation[34] of PALNet [15]. The network completes per-frame depth scans and predicts a voxelized 3D volume with semantic probabilities. The architectural and optimization specifics are introduced briefly in the following sections.

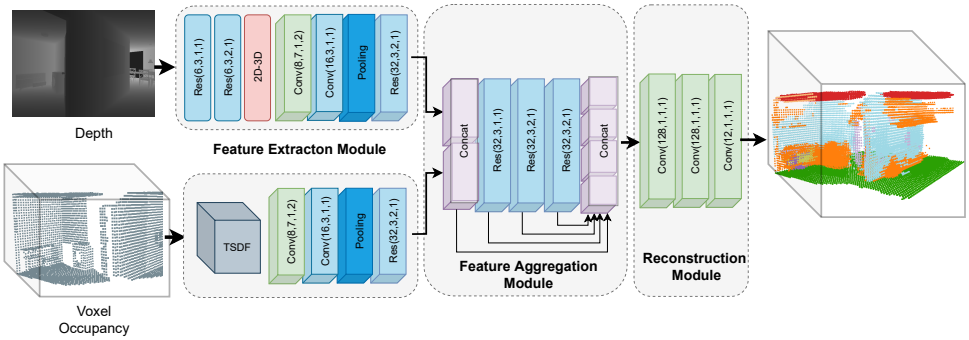


Figure 3.2: PALNet based 3D semantic scene completion network. 2D Layers are represented as planar blocks while 3D blocks are used to indicate a 3D layer with BlockName(filters, kernel size, dilation, stride) indicating the block parameters.

3.2.1 Architecture

The scene completion network is based on a 2D-3D hybrid convolutional neural network (depicted in fig. 3.2) utilizing residual 3D convolutional blocks and follows on earlier SSC works by using dilated convolutions for efficiency and contextual awareness.

Learning robust low-level 3D features is vital for scene completions to ensure geometric integrity. Two parallel streams extract low level features from depth image and TSDF encoded geometry respectively. The 2D feature extraction branch exploits the efficiency of 2D Convolutions to harvest geometrical features from the depth image and projects extracted features to 3D. The 3D feature extraction branch uses 3D convolutions for extracting geometrical features from 3D encoded flipped TSDF volume[13]. Subsequently, both streams are concatenated into a larger

volume and used as input for feature aggregation. The following feature aggregation block uses 3D convolutions with a larger effective kernel and constant stride to concurrently capture features and down-sample the volume without max pooling. For increased Contextual Awareness, multi-stream extracted features are blended and further processed by a series of dilated 3D convolutions with increasing effective receptive field sizes for capturing features at multiple scales, and finally aggregating features serially. Finally, convolution layers with 1x1 kernel size are used to reduce channels and reconstruct the semantically completed volume.

3.2.2 Optimization

Dataset

The SSC network is trained on NYU - RGBD [35], an indoor real-world dataset consisting of 1449 scans along with 3D semantic annotations of office and room environments. The depth map data is further used to generate fixed-size flipped TSDF encoded volume in an offline step using the data processing script from SSCNet[13].

Loss function

The network is trained end-to-end using a single supervised multi-class weighted Cross-Entropy loss in divergence from the proposed Position Aware loss in the original paper[15]. The weighted Cross-Entropy loss equation 3.1 is shown below where N is the total number of samples, C total semantic classes, w_c weight assigned to class C, y_{nc} being the true semantic class and $\hat{y}_{nc}$ being the predicted class.

$$L_{WCE} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{c=1}^{C} w_c y_{nc} \log \hat{y}_{nc} \tag{3.1}$$

Training

The network is trained for 50 epochs with an initial learning rate of 0.01, exponential decay of 0.1 every 10 epochs, and batch size of 1. The inputs to the network are generated by encoding the depth maps as flipped TSDF volume of size $240\times144\times240$ volume flipped. The network predicts a scene completed volumetric probabilities of dimension $60 \times 36 \times 60$ which is denoted by $\hat{y}$ in equation 3.1. Given the large imbalance in free space and objects, the term w_c is set to the inverse frequency of the sample size for each class c.

3.3 Scene Completed Incremental Mapping

The core contribution of this thesis is the development of a **Scene Completed Mapping framework** that builds upon conventional dense volumetric mapping framework Voxblox [6] and additionally introduces functionality for constructing and maintaining a global scene completed predicted map. An overview of the **Scene Completed Mapping framework** is presented in Figure 3.3 along with a brief description below.

1. **Voxblox** framework acts as a primary mapping module responsible for fusing measured depth and RGB scans. It's complemented by Scene Completed volumetric mapping. In the following sections, it shall be denoted as **Measured Map or Observed Map**.

2. **3D SSC network** takes the Depth scan as input like Voxblox, encodes the scan as fixed-size volume, and completes that by predicting per class semantic probabilities.

3. **Scene Completed Mapping** maintains the scene completed global map. It receives scene completions from the 3D SSC network, transforms the completed volume to global frame, and forwards relevant voxels to the Scene Fusion module.

4. **Scene Fusion** calculates fusion updates for potential voxels and integrates fused voxels into the global occupancy map.

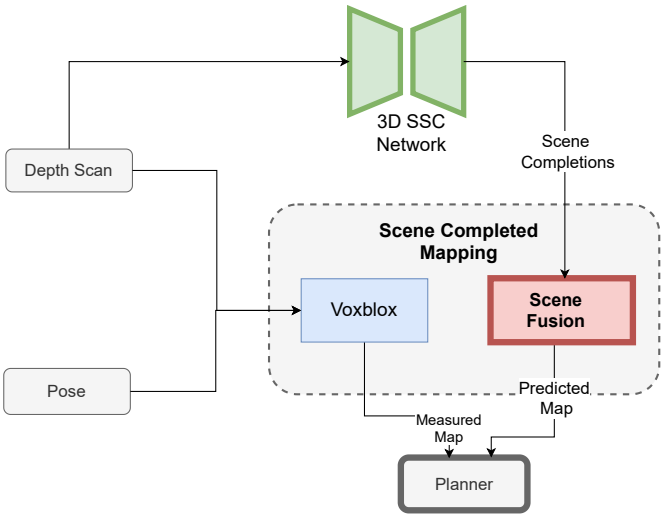


Figure 3.3: Incremental scene completed mapping pipeline. Modules with contributions are highlighted with a bold outline.

The details of the aforementioned components are presented in the following paragraphs.

3.3.1 Semantic Scene Completed Map

A dedicated scene completed map is maintained by incrementally fusing scene completions from partial RGB-D scans. In each frame, a semantic pointcloud is constructed from the RGB-D scan and encoded as the flipped-TSDF voxelized volume of $240 \times 144 \times 240$ voxels. The SSC Network predicts the semantic occupancy probabilities for this volume at a down-scaled resolution of $60 \times 36 \times 60$. The predictions are up-sampled and finally fused back into the global map using the fusion approaches from section 3.3.1. Our proposed approach contrasts current incremental scene completed mapping approaches [18] by performing scene completions on measured depth maps, which eliminates the overhead of extracting to-be-completed sub-volumes. Another factor for the proposed approach is the limited research on SSC Networks addressing scene completions from probabilistic volumes.

Mapping representation

The scene completed global map is represented as a volumetric grid of fixed size voxels arranged in cubic blocks where each voxel contains the probability value (in the log-odds state) indicating its occupancy. Additionally, a label indicating

the semantic category and status from "free", "occupied", or "unknown" is stored. The map follows the Hashing-based representation introduced in Voxel Hashing [26] implemented in the Voxblox framework. The probabilistic approach addresses the shortcomings of discrete categorical occupancy grids as the continuous probability value provides fine surface boundary at the transitions around $\frac{1}{2}$ without the limitation of TSDF, which only maintains voxel data within truncation distance of the surface. The log-odds probabilistic representation simplifies the fusions and retains the advantage of TSDF based maps for fine surface representation. It's additional advantage over TSDF based map is a probabilistic interpretation of the surface.

Incremental Fusion and Integration

The category wise semantic probability predictions from the scene completion network are combined for all object classes to get occupancy probability which is incrementally fused in a probabilistic manner into the global predicted map as shown in fig. 3.4 using a Bayesian approach from [36]. The cumulative occupancy probability of a voxel v at time t, $P_t(v)$ is calculated from predicted probability at time t denoted by $p_t(v)$ along with earlier predictions following eq. 3.2. The initial prior $p_{t_0}(v)$ is assumed to be uniform with value of $\frac{1}{2}$.

$$P_t(v) = \frac{\prod_{i=1}^{t} p_i(v)}{\prod_{i=1}^{t} p_i(v) + \prod_{i=1}^{t}(1 - p_i(v))} \tag{3.2}$$

For incremental fusion, eq. 3.2 is reformulated to update previous cumulative probability$P_{t-1}(v)$ by fusing current predictions $p_t(v)$ to get updated cumulative probability $P_t(v)$ as shown in eq.3.3.

$$P_t(v) = \frac{P_{t-1}(v)p_t(v)}{P_{t-1}(v)p_t(v) + (1 - P_{t-1})(1 - p_t)} \tag{3.3}$$

The equation can be simplified by adapting to log-odds space where the cumulative probability takes the form $\log\left(\frac{P_t(v)}{1-P_t(v)}\right)$ denoted by simplified notation $L_t(v)$ in following equations. Similarly single updates are represented as $\log\left(\frac{p_t(v)}{1-p_t(v)}\right)$ denoted by $l_t(v)$.

$$\underbrace{\log\left(\frac{P_t(v)}{1 - P_t(v)}\right)}_{\text{Updated voxel value}} = \underbrace{\log\left(\frac{P_{t-1}(v)}{1 - P_{t-1}(v)}\right)}_{\text{Previous voxel value}} + \underbrace{\log\left(\frac{p_t(v)}{1 - p_t(v)}\right)}_{\text{new measurement}} \tag{3.4}$$

$$L_t(v) = L_{t-1}(v) + l_t(v) \tag{3.5}$$

3.3.2 Fusion strategies

By adapting log-odds probability representation, incremental updates $l_t(v)$ can be calculated via different weighing strategies. Even though the scene completion network predicts probabilities, to investigate and correct the biases and uncertainty, various strategies are explored which are stated as follows.

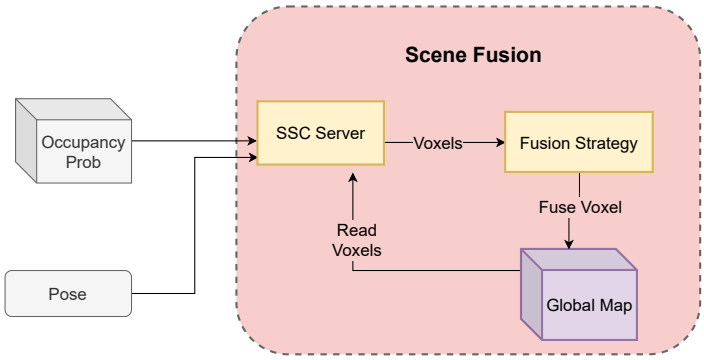


Figure 3.4: Fusion of a scene completed volume into the predicted global map

Log-odds

The probability confidence from the SSC network is directly employed to calculate the updates using the formula in eq 3.6.

$$l_t(v) = \log\left(\frac{p(v)}{1-p(v)}\right) \tag{3.6}$$

Naive

A naive approach without probabilistic fusion where the updates are only fused if the voxel has been not been observed or is empty. It is used as a baseline for analyzing the improvements from incremental fusion.

$$l_t(v) = \begin{cases} \log\left(\frac{p(v)}{1-p(v)}\right) & \text{if } p(v) > \frac{1}{2} \text{ or } v \notin \mathit{Observed} \\ 0, & \text{otherwise} \end{cases} \tag{3.7}$$

SCFusion

A fusion strategy introduced by cheng Wu et al. [18] addresses the fusion of noisy scene completions into a probabilistic occupancy map. The authors propose only fusing occupied space and assign very low constant confidence to predictions from scene completion network with $p_{occ_{const}} = 0.51$ while fusing sensor measured with higher probability.
A predicted voxel v is considered occupied when network predicted probability is greater than $\frac{1}{2}$ and free other wise.

$$l_t(v) = \begin{cases} \log\left(\frac{p_{occ_{const}}}{1-p_{occ_{const}}}\right), & \text{if } p_t(v) > \frac{1}{2} \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

Proposed Strategy

SCFusion proposal to treat predictions as low confidence measurements helps control excessive noise arising from overconfident network predictions but it does not help exploration planning as conservative planner treat unknown space as occupied leading to insignificant gain. To exploit scene completions for planning, fusing free space is crucial for Planner, which is fused in a similar way with constant weighted probability. Another advantage of using weighted probabilities for free and occupied

space is to compensate for networks bias towards predicting free space.

$$l_t(v) = \begin{cases} \log(\frac{p_{occ_{const}}}{1-p_{occ_{const}}}), & \text{if } p_t(v) > \frac{1}{2} \\ \log(\frac{p_{free_{const}}}{1-p_{free_{const}}}), & \text{otherwise} \end{cases} \tag{3.9}$$

Counting

A strategy to fuse free and occupied space predictions from network where predictions are counted and fused only when there is a switch in voxel state from free to occupied and vice versa.

$$l_t(v) = \begin{cases} 0 & \text{if } \texttt{Count}(p_t(v)) = \texttt{Count}(p_{t-1}(v)) \\ \log(\frac{p_{occ_{const}}}{1-p_{occ_{const}}}), & \text{else if } p_t(v) > \frac{1}{2} \\ \log(\frac{p_{free_{const}}}{1-p_{free_{const}}}), & \text{otherwise} \end{cases} \tag{3.10}$$

3.4 Combining measured and predicted maps

3.4.1 Scene Completed Mapping

The Planner requires the mapping information for various sub-modules with varying safety requirements which the proposed two map approach of keeping both a measured dense map and predicted scene completed map fulfills aptly depending upon the level of safety required and the mapping uncertainty. The Measured dense map is denuded from the uncertainty introduced by the scene completion process making it vital when safety is the concern, while the predicted map on the contrary contains scene completions that address occlusion arising with induced uncertainty making it more suitable for safety agnostic situations.

Information Planning

The information planning utilizes the mapping information for information gain to identify promising regions offering maximum exploration without having to execute a trajectory relaxing the need for safety constraints. Information Planning can withstand minor mapping uncertainties while benefiting from increased explored volume.

Traversability Planning

The safety-critical part of planning that verifies potential paths for traversability ensuring collision-free paths. Utilizing an uncertain map for traversability planning shall render planning ineffective regardless of mapped volume if the planner can't explore without collision.

3.4.2 Mapping approaches for Planning

This section illustrates how the planner can be fed optimum perceptual information available from two parallel maps each with different qualities depending on the planning situation. To combine the maps optimally for various planning stages, the approaches stated in table. 3.1 are developed.
Maps can be combined in the following ways for respective planning situations:

- **Measured** - Mapping module utilizes the dense measured map for the respective query.

Table 3.1: Approaches to combine the predicted and measured map for planning

Approach	Traversability Planning	Information Planning
Conventional	Measured	Measured
Conservative	Measured	Hierarchical
Conservative+	Measured	Predicted
Cautiously Optimistic	Hierarchical-II	Hierarchical-II
Optimistic	Hierarchical	Hierarchical
Over Optimistic	Predicted	Predicted

- **Predicted** - The Mapping module utilizes the incrementally fused scene completed map for fulfilling the occupancy requests by the planner.
- **Hierarchical** - First measured map is consulted for querying the occupancy status. In case of no observation, the Predicted map is queried.
- **Hierarchical-II** - Similar to the hierarchical approach, the predicted map is consulted in absence of observations but only if the predictions are high confidence.

Conventional

Conventional Planners rely on the measured map for all stages of path planning making them very safe at the expense of efficiency.

Conservative

A conservative approach advantages from the fact that not all stages of Planning are safety-critical by consulting predicted map along with the measured map in a hierarchical fashion for Information Planning only. For Traversability checks only the measured Map is used which makes it exactly as safe as the Conventional approach while exploiting scene completion.

Conservative+

Conservative+ approach is similar to conservative where completions are only used for Information Planning but in contrast to the conservative approach, it considers predicted map for Information planning assuming the predictions are reliable.

Cautiously Optimistic

The cautiously optimistic approach utilizes the predicted map cautiously in the absence of observations by only considering high confidence predictions in the Hierarchical-II pattern. In contrast to the conservative approach, the specified predicted map is also employed for traversability planning.

Optimistic

An Optimistic approach utilizes measured and predicted Maps hierarchically for both information and traversability planning. This approach provides Planner with maximum visual information regardless of certainty.

Over optimistic

This approach contrasts the conventional approach by only considering the predicted Map for all stages of planning and makes a decisive case to evaluate the capability of the Scene Completed Map.

3.5 Scene Completion Aware Planner

3.5.1 Planner Overview

To make the planner achieve the most from the scene completed mapping, the planner is extended to exploit scene completions smartly at various stages of planning. For this work, a conventional exploration planner from Schmid et al.[7] was enhanced with scene completed mapping along with additional gain formulations to make the planner smarter safely, and effectively.
The overview of the proposed Planner is depicted in figure. 3.5. The components of scene completion aware planner are introduced briefly.

- **Scene Completed Mapping** - A module wrapping a reference to the measured and predicted maps from the mapping pipeline is integrated within the Planner.
- **Trajectory Generator** - Generates new trajectory segments by exploiting the scene completed mapping and forwards the segments to trajectory gain evaluator module
- **Trajectory Gain Evaluation** - Calculates informative gain for a trajectory segment by consulting the Scene Completed Map.

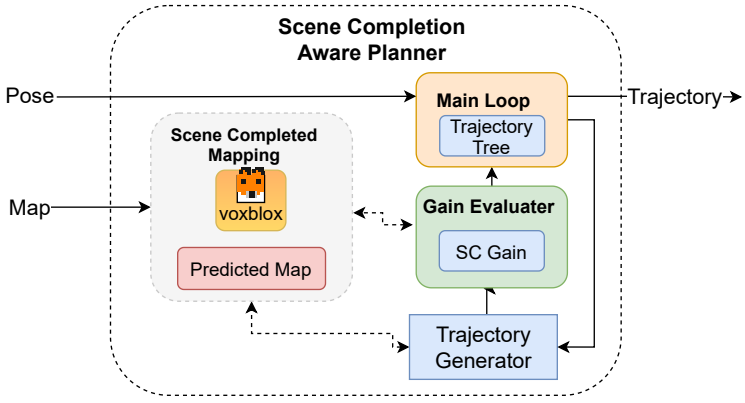


Figure 3.5: Extension of conventional exploration planner from Schmid et al.[7] with scene completed mapping and gain.

3.5.2 Gain Formulations

To maximize the utility of scene completed mapping, the proposed Informed Planner is updated with a novel objective function that promotes using scene completions safely to achieve the following objectives.

1. Explore unknown areas

2. Improve the quality predicted map by refocusing on regions with higher uncertainty
3. Use scene completed map to identify potential free volumes that are not measured

To encourage such behavior by the planner, the following gain formulation is proposed.

$$\mathcal{G}_{unknown}(v) = \begin{cases} 0 & \text{if } \texttt{Observed}(v) \\ 1, & \text{otherwise} \end{cases} \tag{3.11}$$

$$\mathcal{G}_{unmeasured}(v) = \begin{cases} 0 & \text{if } \texttt{Measured}(v) \\ 0 & \text{if } p(v) > \frac{1}{2} \\ 1, & \text{otherwise} \end{cases} \tag{3.12}$$

$$\mathcal{G}_{confidence}(v) = \begin{cases} 0 & \text{if } p(v) > \max \\ 0 & \text{if } p(v) < 1 - \max \\ \log\left(\frac{\max}{1-\max}\right) - \texttt{abs}(\left(\log\left(\frac{p(v)}{1-p(v)}\right)\right), & \text{otherwise} \end{cases} \tag{3.13}$$

$$\mathcal{G}_{total}(v) = \underbrace{\alpha\mathcal{G}_{unknown}(v)}_{\text{conventional gain}} + \underbrace{\beta\mathcal{G}_{unmeasured}(v) + \gamma\mathcal{G}_{confidence}(v)}_{\text{scene completion gain}} \tag{3.14}$$

- **Unknown Voxel Gain** promotes the exploration of regions that are neither measured nor predicted. This gain has the highest weight given exploration is the primary objective.
- **Unmeasured Voxel Gain** promotes the exploration of regions that are predicted as free space (i.e having occupancy probability less than $\frac{1}{2}$) but not measured. Its stems from the intuition that free predicted regions are likely to be measured as such, and by setting the β weight to low, large volumes of free space would out weight paths with a very few unknown voxels resulting in large exploration volumes.
- **Confidence Gain** is a log-based gain that specifically addresses voxels having very low confidence. An occupied voxel has high confidence when occupancy $p(v) \to 1$ and similar free voxels have high confidence when their occupancy probability $p(v) \to 0$. The log-odds-based probabilistic representation is undefined at extreme values of 0 and 1 so cut-off values are enforced near extremes.

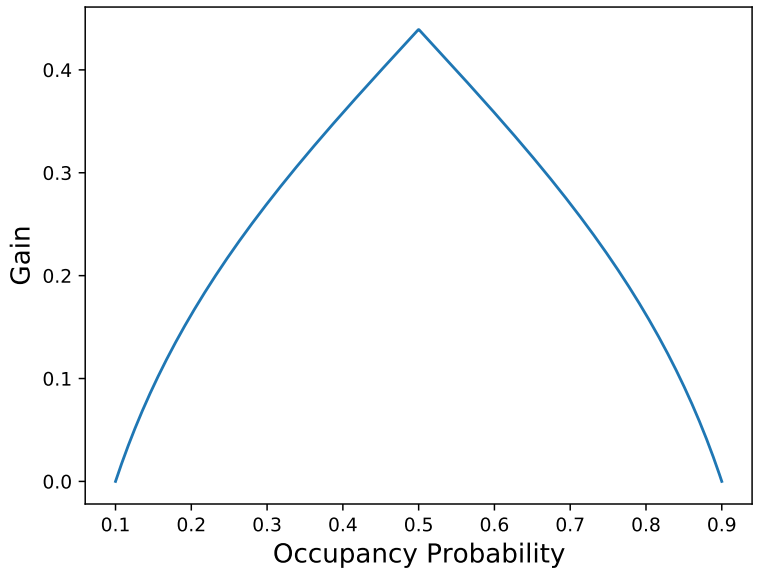


Figure 3.6: A plot of confidence gain $\gamma\mathcal{G}_{confidence}(v)$ with $\gamma = 0.2$ representing the gain relative to voxel occupancy probability. Voxels with $\sim\frac{1}{2}$ occupancy probability have very high uncertainty being on the verge of occupancy status transitions. The confidence gain promotes certainty in uncertain regions by encouraging paths leading to increased observations in the specified regions.

Chapter 4

Experimental Setup

4.1 Simulation Setup

A simulation environment with realistic physics and photo-real visuals assists debugging, accelerates development and enables evaluations safely and efficiently. To develop and evaluate the capabilities of the proposed mapping and planning framework, a MAV equipped with visual sensors was simulated using Airsim simulator [37] configured in Unreal Engine 4 as visualized in fig. 4.1.

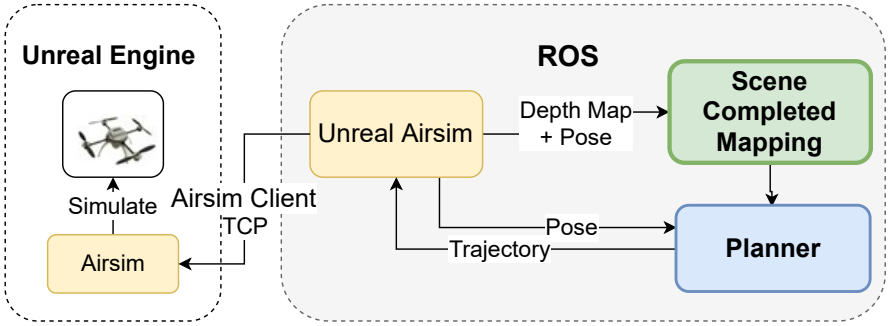


Figure 4.1: An overview of simulation setup with Unreal Engine, Airsim and Unreal_Airsim for interfacing simulation with mapping and Planning modules over ROS

4.1.1 Unreal Engine 4

Unreal is a game engine for editing and creating detailed 3D visual content for immersive virtual worlds making it very suitable to develop and render simulated environments for robot exploration in real-time. For this work, Unreal Engine 4.25.6 was used.

4.1.2 Microsoft Airsim

Airsim [37] is a realistic MAV physics simulator developed for Unreal and Unity Engines complementing detailed visuals with realistic physics. It is developed as an Unreal Engine Plugin that interacts with Unreal API and provides relevant access to end-users via Remote Procedural Calls (RPC) based client-side API named AirLib. Airsim further supports attaching RGB, Depth, Infrared cameras, and Lidar while

providing real-time access to sensor measurements and MAV state. Airsim version 1.2 was used for the simulated setup.

4.1.3 Unreal Airsim

Unreal Airsim is a framework developed by Schmid and Reijgwart [38] that enables communication with the Airsim Server in Unreal Engine 4 via AirSim Client API. It provides ROS interface to the simulated MAV exposing the following non exhaustive capabilities.

- MAV control commands for position and velocity
- Access to MAV state and collision detection
- Retrieving Depth and RGB images from a camera mounted on MAV

4.2 Scene setup

An indoor environment in a home setting is set up in Unreal Engine. The scene contains furniture, small objects, and hallways providing a realistic and moderate exploration testbed. The problem at hand is the autonomous exploration of this scene demonstrated in figures 4.2 4.3 and analyzing the mapping and exploration capability of the proposed system.

Figure 4.2: A Large living room with furniture of various sizes and objects rendered in Unreal Engine

4.2.1 Sensor Setup

A generic AirSim Micro Aerial Vehicle (MAV) equipped with a RGB and Depth camera is employed as the exploration vehicle. The sensor configurations for the simulation are specified in table. 4.1.

Figure 4.3: Top down view of the simulated environment in Unreal Engine used for experiments

Parameters	Value
Image Width	640 pixels
Image height	480 pixels
FOV	90 °
Position	[0.3, 0.0, 0.0]
Roll, pitch, yaw	[0.0, 0.0, 0.0]

Table 4.1: Sensor configuration for depth and RGB cameras in Airsim. The camera pose is specified in MAV frame.

4.3 Experiment Setup

The overview of general parameters for the proposed mapping and planning framework[1] used for all experiments. For convenience, only non-default parameters are specified. Experiment-specific parameters shall be introduced in the respective contexts.

4.3.1 Mapping

As introduced in section 3.3, the proposed mapping system uses Voxblox for TSDF mapping along with a Occupancy grid based scene completed map implemented built on Voxblox. Both maps are defined in Global coordinate frame, have aligned origin and orientation enabling 1:1 voxel correspondence.

4.3.2 Planning

The planner from [7] is used as baseline planner with exploration planning configuration and parameters from table. 4.3.

[1] https://github.com/mansoorcheema/ssc_3d_planning/tree/SSC+VoxbloxOccupancyMap-v-0.2

Parameters	Value
Voxel Size	8 cm
Block Size	16 voxels
Truncation Distance	32 cm

Table 4.2: Mapping Parameters for Voxblox and scene completed mapping

Parameters	Value
Max velocity	3.0 m /s
Max acceleration	3.0 m/s
Collision radius	0.1 m
Min path length	0.2 m
Maximum yaw rate	2.6 rad/s

Table 4.3: General parameters for MAV used for all experiments

4.4 Evaluation Metrics

4.4.1 Scene completion Metrics

Semantic scene completion networks, though differing in input modalities predict fixed size voxelized volumetric scene completions trained using multi-class cross-entropy loss, with some exceptions like Scan Complete [39]. Grid-based gravity aligned scene completions are compared voxel-wise with ground truth labels. The preferred metrics for evaluating semantic scene completed volumes are mean Intersection over Union (m IoU), precision, and recall.

Precision

Multi-class predictions for different object classes are interpreted as a single occupied class and compared voxel by voxel using the standard precision formula 4.1 with TP, FP, FN indicating true positives, false positive and false negatives respectively.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.1}$$

Recall

Recall is calculated similarly voxel by voxel and indicates how much of total volume is observed correctly.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.2}$$

Intersection over Union

IOU can either be calculated for each semantic class as shown in the equation 4.4 or in the binary form, similar to the way precision and recall are calculated in the last section by merging all occupied classes into a single class according to eq. 4.3.

$$\text{IOU} = \frac{TP}{TP + FP + FN} \tag{4.3}$$

$$\text{m-IOU} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FP_c + FN_c} \tag{4.4}$$

4.4.2 Exploration Metrics

An exploration planner's primary objective is to explore maximum volume efficiently and safely. To evaluate such capability, following safety critical questions of main concern

- How much of total volume is explored correctly?
- How precise is the free space explored?
- How much of total occupied volume is identified as occupied?

To evaluate these metrics, the environment in 4.3 is discretized into voxels. The total explorable free and occupied space from the discretized ground truth environment is calculated to enable objective comparison up to voxel resolution. Similarly measured and predicted maps are by construction a grid of voxels aligned in the same world frame making voxel by voxel comparison feasible.
To answer the posed evaluation questions, the following appropriate metrics are evaluated for both measured and predicted maps separately, where observed relates to a voxel status in the respective map.

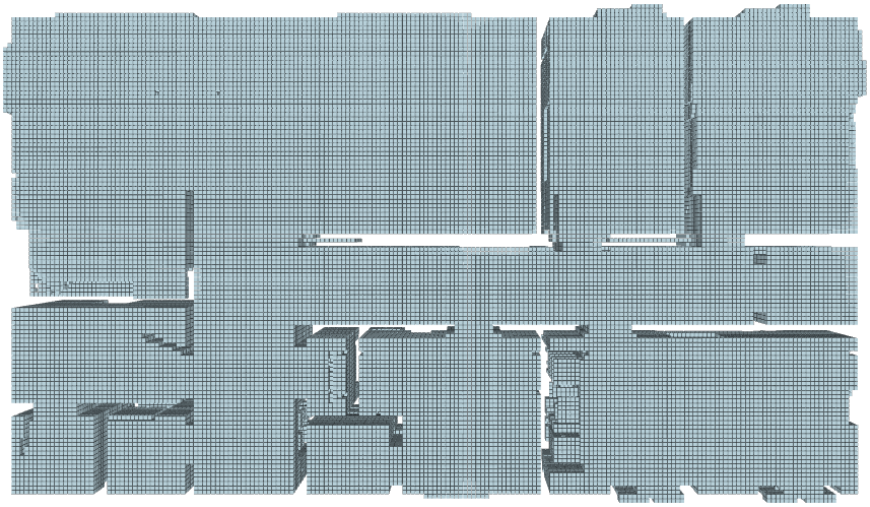


Figure 4.4: The environment in 4.3 is discretized at resolution of 8 cm and explorable free space (cyan voxels) is identified by frontier propagation from robot start position

Precision of observed free space

The precision of explored regions identified as free space is critical for planning, which otherwise lead to collisions when MAV attempts to pass through. The precision is calculated using the formula in eq. 4.5 where TP_{free}, FP_{free} refers to correctly and incorrectly observed free space voxels respectively.

$$\text{Precision} = \frac{TP_{free}}{TP_{free} + FP_{free}} \tag{4.5}$$

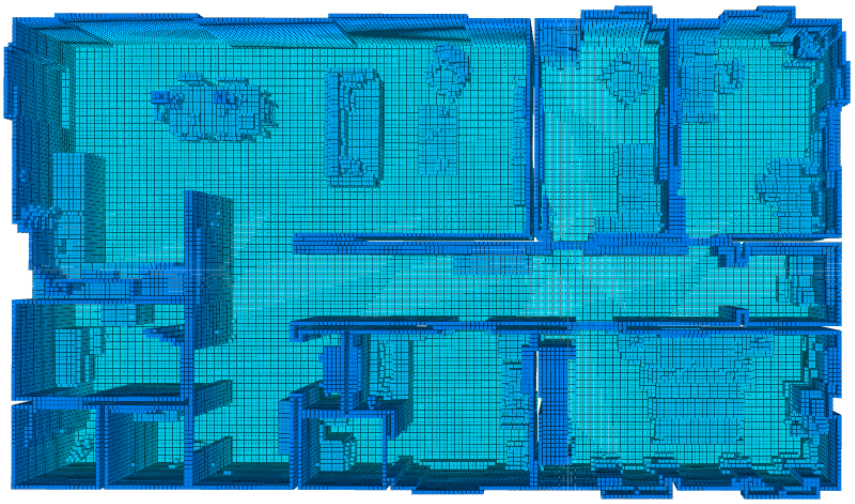


Figure 4.5: The explorable occupied space enclosing the explorable free space from fig. 4.4 color coded over height.

Recall of occupied space in explored regions

Along with the precision of free space, maximum of the occupied space should be correctly identified. In contrast, the recall of free space is less safety-critical as not identifying a free sub-region won't result in a collision. $TP_{occupied}$ refers to correctly observed occupied space while $FN_{occupied}$ refers to occupied space in the discretized environment 4.3 that is observed as free.

$$\text{Recall} = \frac{TP_{occupied}}{TP_{occupied} + FN_{occupied}} \tag{4.6}$$

Coverage

Coverage relates to the amount of explored volume as a fraction of total explorable volume. The eq. 4.7 is used to evaluate coverage metrics where "Total Voxels" refer to total explorable occupied and free voxels visualized in fig. 4.4.

$$\text{Coverage} = \frac{TP_{free} + TP_{occupied}}{\text{Total Voxels}} \tag{4.7}$$

Chapter 5

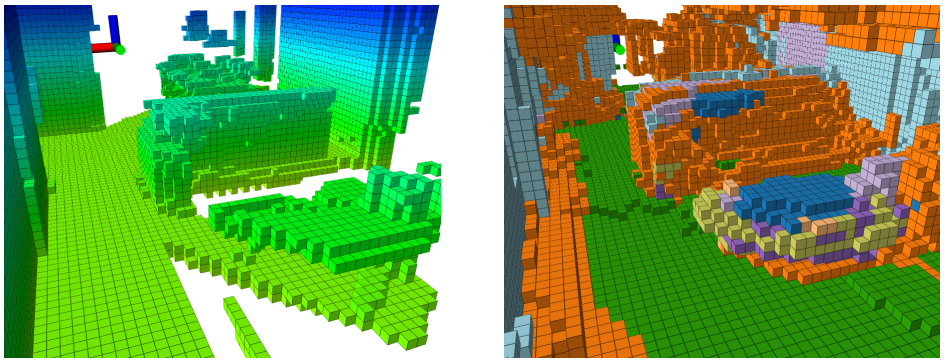
Results

5.1 3D Scene Completion

The proposed SSC network trained on NYU-RGBD [35] dataset is evaluated on NYU-RGBD test set for quantitative performance analysis. Furthermore, the performance of the SSC network is evaluated in a simulated environment significantly different from the training data to evaluate the exploration potential in unknown environments.

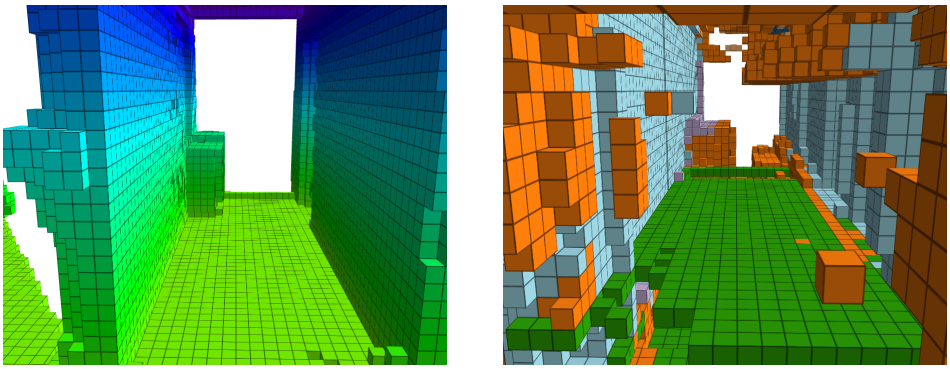
5.1.1 Qualitative Results

The proposed 3D SSC network performs remarkably well on completing structures with a consistent pattern like floor and walls as evident in figure 5.1. Partial objects like sofa and table are inferred acceptably with some noise on edges. In some cases, the SSC network mistakenly predicts objects in free regions as seen in figure 5.2, affirming the hypothesis from this thesis that relying on predictions alone is insufficient for effective planning. It further suggests fusing noisy scene completions into the measured map can compromise safety.



Measured map Scene completed map

Figure 5.1: Scene completed map fills in missing floor, table and sofa. Labels: Green voxels represent floor, oranges represent furniture like sofa, blue table, yellow chair and cyan for walls.



Measured map Scene completed map

Figure 5.2: Scene completions are noisy and often over predicted in corners making reliance on predictions for planning insufficient.

5.1.2 Quantitative Results

Given the compute intensity of SSC networks and limited computation capability available on-board MAVs, specifically single-stage open-source approaches were evaluated. The results summarized in table 5.1 indicate that the scene completion performance of the proposed approach does not match published state-of-the-art open-source approaches, especially in semantic scene completions though the results are closer for class agnostic scene completion. The proposed approach is based on down-scaled PalNet[15] that enables the algorithm to perform significantly faster on a modest consumer-grade mobile CPU (Intel i7-8565U) comparable to an onboard CPU on a MAV.

Approach	IOU	m-IOU	Params	mean. FPS
PALNet	55.6	**34.1**	0.22 M	-
SCFusion	-	-	2.7 M	0.56
SSC Net	55.1	24.7	0.93M	-
DDRNet	**61.0**	30.4	**0.20M**	0.33
Proposed Network	54.5	24.42	0.21M	**1.07**

Table 5.1: Comparisons of semantic scene completions networks performance on NYU-RGBD dataset[33]. FPS are measured on a mobile Intel i7-8665U CPU with builtin Intel Graphics Adapter. The parameters for SCFusion were evaluated from open source implementation which might not reflect the model in the paper.

5.2 Incremental Fusion

5.2.1 Map quality per fusion method

The fusion strategies from section 3.3.1 are evaluated on identical input sequence and the safety critical metrics for map quality are evaluated. First, the quantitative results are presented in fig. 5.3 and 5.4. The qualitative results are presented in fig. 5.5. The result surprisingly indicate that the Naive strategy performs very well on safety-critical metrics for the precision of free space and the recall for occupied is very high. However, a potential explanation is that it naively overwrites all unobserved and free space with predictions leaving only a small portion of voxels

free behind that were not overwritten throughout the experiment suggesting those voxels have repeatedly been predicted as free space, coinciding with a high likelihood of actual free space. Similarly, large occupied predictions drive up the recall. The decrease in IOU of occupied voxels suggests that the excessive occupied fused voxels are often not correct which can be observed in the Qualitative results in fig. 5.5. The proposed approach surpassed all other approaches across metrics specially IOU. The standard log-odds struggles as the Network predicted certainties are biased towards free space, and the flexibility of the proposed strategy of weighting free and occupied predictions help with this issue. Similarly, the absence of SCFusion from the precision graph is due to the fact that the approach only predicts occupied space hence the precision is calculated as NAN. Surprisingly the recall is also absent which is another consequence of the aforementioned reason as all the predictions are exclusively occupied space and the recall metric of occupied space is calculated by finding all the occupied voxels in the observed region and finding intersection with the predictions which is always 1.

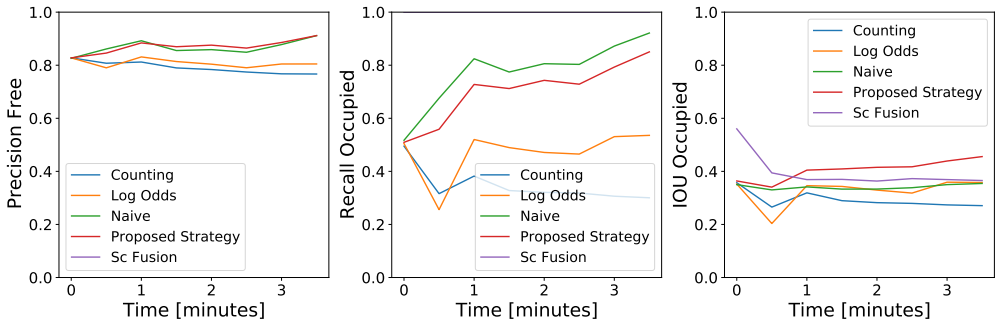


Figure 5.3: Evaluations for safety critical metrics

To evaluate the net improvement from scene completed map outside sensor measurements, voxels corresponding to scanned measurements were discarded from evaluation and the results could be observed in fig. 5.4. The results have decreased as expected from discarding measured voxels, especially the precision seems to decline over time. Its explained by the fact that at the start scene completions are predicted in yet to be measured regions which are overwritten when they are measured later contributing to the decrease. Over time, almost all the map is measured. In contrast to results in fig. 5.3, any value above zero is net gain. Conclusively, the scene completion helps the most when there is not enough measured map which makes it very suitable for exploring unknown environments.

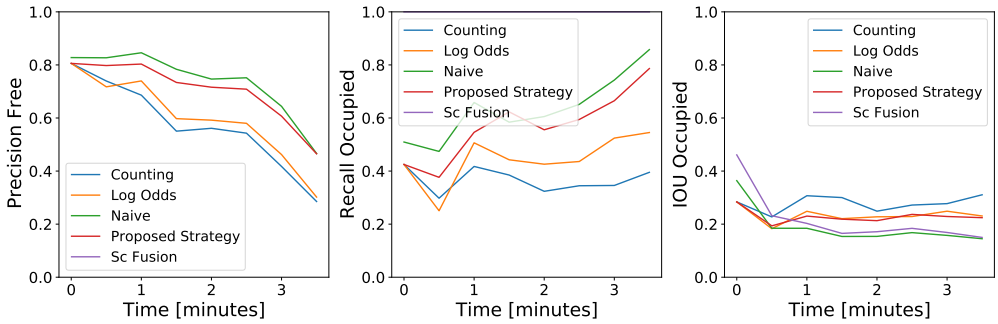
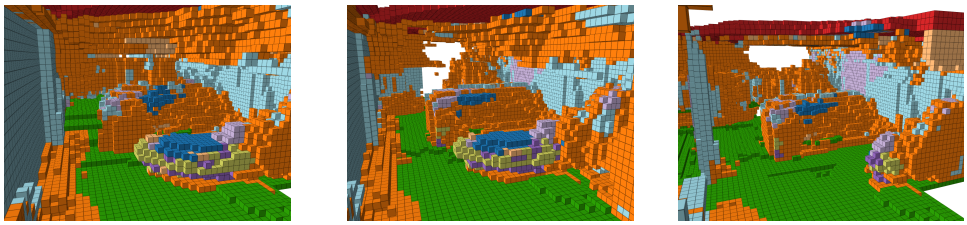


Figure 5.4: Evaluations for safety critical metrics excluding measured map



Naive Proposed strategy Log Odds

Figure 5.5: Qualitative results at identical time for the fusion methods. Notice that Naive approach has excessive cluttering. The log odds approach is very sensitive to free space predictions and the image shows that the table and left wall is discarded during fusions.

5.3 Combining measured and predicted map

To evaluate the approaches introduced in section 3.4 for utilizing the measured and predicted maps for planning, the following objectives were considered :

- The combination approach with best perception quality
- The combination of maps per planning situation that helps planner explore the most efficiently without compromising on safety

Mapping Results

To evaluate the quality of the measured, predicted, and hierarchical maps. 10 experiments were performed with the same baseline planner without taking those scene completions into account for planning and the mean safety-critical metrics with a confidence interval of 95 % were calculated. The results are presented in fig 5.6.

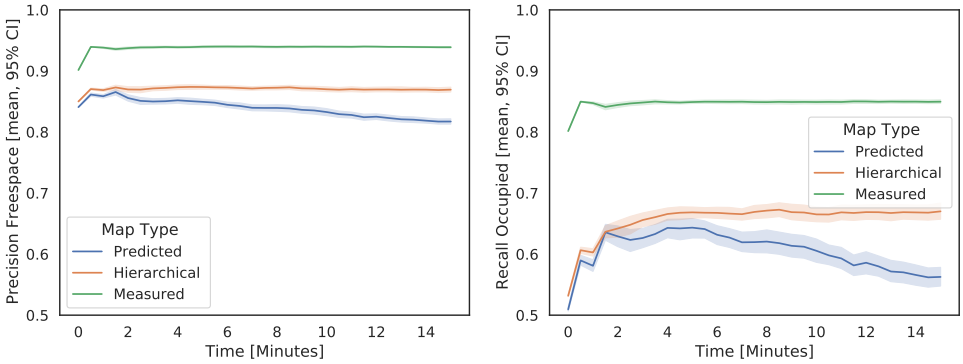


Figure 5.6: Safety critical Metrics for planning tasks demonstrate that neither predicted nor hierarchical map match the measured map quality.

The results indicate that the measured map has significantly higher quality compared to the predicted map. Furthermore, combining it with the predicted map lowered its quality as expected but is there any advantage by combining measured with the hierarchical map? To address this question, coverage metrics were calculated for each map type and results are presented in fig. 5.7 . The coverage

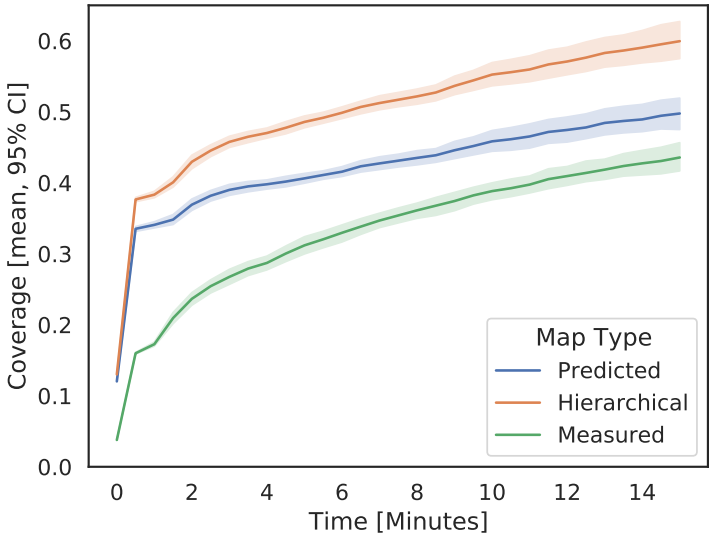


Figure 5.7: Coverage evaluations for predicted map, measured map and hierarchical map.

results for map types indicate that the robot is not able to visualize most of the environment by relying on the measured map alone. The hierarchical inclusions of scene completions enhance coverage in occluded regions (coverage includes correctly explored regions) thus increasing the robot's perceptual awareness beyond what the robot could measure. The hierarchical approach surpasses utilizing either map alone.

Planning Results

The last section showed measured map surpassed predicted map and hierarchical configuration of measured and predicted maps for map quality. However, hierarchical map had higher coverage. To verify how well does the planner performs given various ways of using the predicted and measured maps, approaches from the table 3.1 were evaluated. 10 runs of 15 minutes were performed for each approach. The coverage of the measured map is calculated for each approach and the results are presented in fig. 5.8. This coverage is different compared to the one from the last section that compared the different types of maps that were generated with the same planner without utilizing the scene completions for planning as expected. We shall observe that scene completion helps the planner make better planning decisions leading to larger exploration reflected in the measured map.
Results indicate approaches using scene completion with measured maps for planning have best results than using either map.
Conservative+ approach shows worse results than conventional planner which shows just using the predicted map for information planning reduces result which indicates map quality limitation especially prediction of false-positive occupied predictions. On the other hand Conservative approach performs better than the conventional planner by just using the hierarchical approach for information planning, which suggests supplementing measured map with predicted map helps, and the performance of the cautiously optimistic planner supports this conclusion that has even better results as it even prefers high confidence predictions over the measured map. The

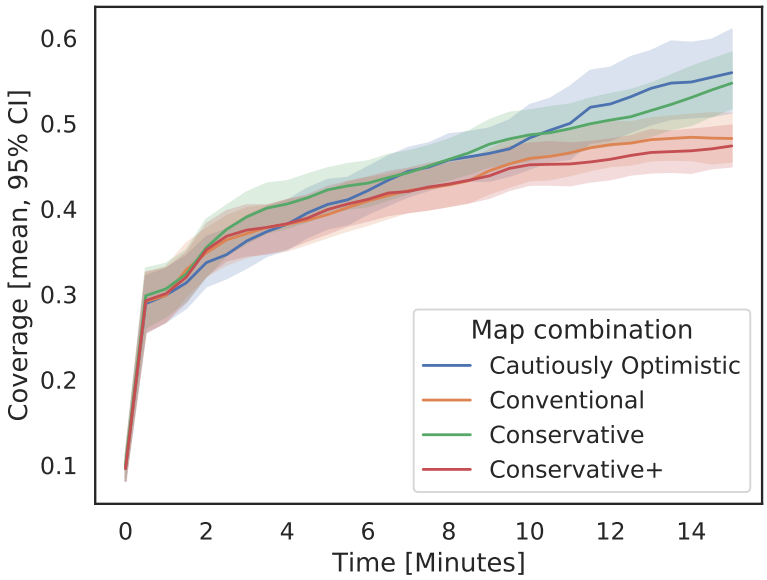


Figure 5.8: Coverage evaluation for different approaches of querying measured and predicted maps introduced in sec. 3.4 for planning.

following conclusions could be drawn from these results.

- Predicted map helps with exploration but only if used with measured map in which case it performs better than either map
- Cautiously optimistic approach shows that low confidence predictions are responsible for the decreased performance of Conservative+ and high confidence predictions can further improve the results.

5.3.1 Safety Analysis

Scene completion helps exploration but the performance of the conservative+ approach suggests that the quality of the predicted map is not precise which the earlier mapping evaluation metrics also supported raising safety concerns.
To evaluate these concerns, the safety of experiments performed in the last section was evaluated. The results are compiled from 10 runs for each approach and are presented in fig. 5.9.

The concerns turn out to be true if the predictions are to be relied upon for safety-critical components of the planning pipeline as shown by optimistic and over-optimistic approaches. As the reader would have noticed, these approaches are not included in the coverage metrics as the planner never made it to the end without colliding. The cautiously Optimistic approach fares significantly better by relying on high confidence predictions even for safety-critical planning components indicating high confidence prediction can be very reliable for superior exploration capability
Similarly, the performance of the Conservative approach shows its as safe as a conventional planner that does not use scene completion and surprisingly also performed as well as the cautiously Optimistic approach. The reason is that the Con-

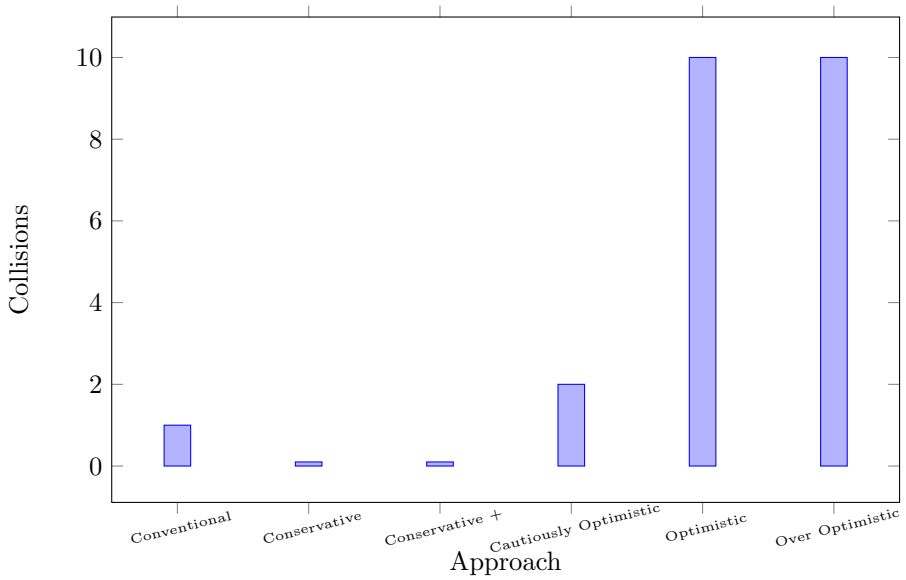


Figure 5.9: Number of collisions for approaches from table. 3.1 across 10 runs

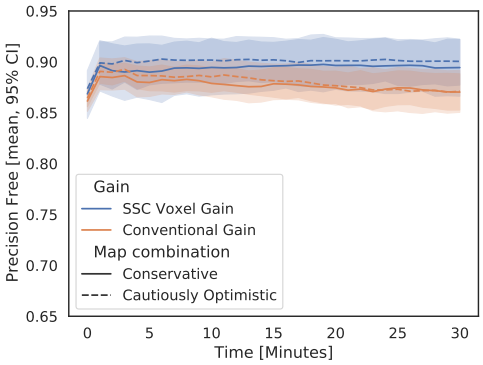
servative approach does not use the measured map for safety-critical traversability planning sufficing only from scene completions for information planning.

5.4 Scene Completion aware Planner

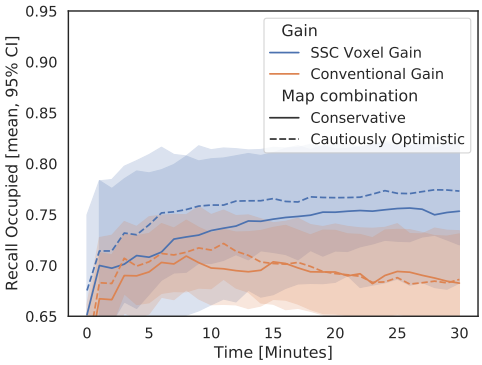
The last experiments showed the improved performance of conservative and cautiously optimistic approaches for mapping over conventional mapping, however, the planner did not incorporate the scene completions for exploration planning. In section 3.5 a gain formulation was introduced to make the planner smarter by incorporating scene completions in the planning process. In this section, the performance of the proposed planner shall be evaluated.
The experiments are performed for the Conservative and Cautiously Optimistic map combining strategies and are grouped into with and without the proposed gain. Given the stochastic nature, each experiment was performed 5 times, and mean values are plotted with confidence intervals. Furthermore, for these experiments, the time of the experiment was increased from 15 to 30 minutes. The results are presented categorized by quality and coverage.

5.4.1 Quality Results

The quality of the predicted map is analyzed to verify the effect of the proposed gain that was designed to improve the predicted map so that it can help the planner explore better trajectories. The results support the hypothesis by indicating the advantage of proposed gain over conventional gain across Conservative and Cautiously Optimistic approaches for safety critical quality metrics. The improvement co relates to the dependence of the approaches on scene completion. Conservative approach only uses predicted map for information planning while Cautiously Optimistic approach uses it for traversability planning as well.



Predicted map's precision for free voxels



Predicted map's recall for occupied voxels

Figure 5.10: A comparison of the quality metrics for the predicted map generated by planners with and without using the proposed gain. All of these methods use scene completions for planning.

5.4.2 Coverage Results

The last section discussed the quality of the predicted map. In this section, the contribution of the predicted map for planning shall be discussed, which is assessed by the coverage of the measured map constructed while planning. As mentioned before, the quality metrics like precision and recall of the measured map are independent of planning strategy. The results in fig. 5.11 follow the earlier pattern

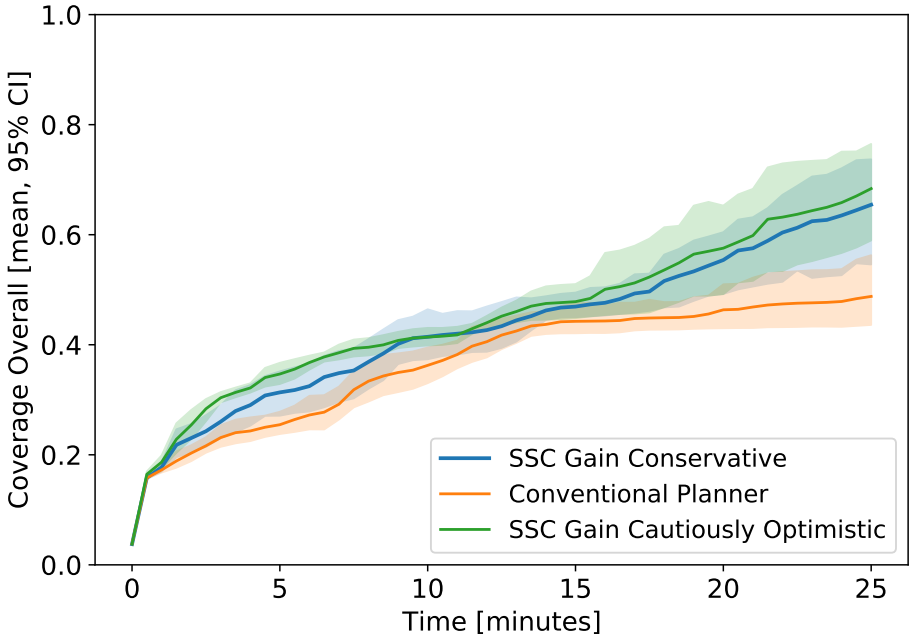


Figure 5.11: Comparison of the conventional planner with scene completion aware planner. The coverage is evaluated for the measured map built by these planners.

of improved coverage for Conservative and Cautiously Optimistic approaches over Conventional approach, which further improve coverage using the SSC Gain. The

safety of the Conservative approach remains exactly the same as the Conventional approach given the predictions are not used for collision checking. The performance of the Cautiously Optimistic approach suggests greater exploration capability can be achieved at the expense of safety. The improvement does not appear to be significant over the conservative approach especially given the additional safety risks, nonetheless, it does provide further direction of exploring scene completion for traversability planning and suggests the results could be boosted with higher quality predicted map.

Qualitative Results- Conventional Planner

The qualitative result of the plots from fig. 5.11 for conventional planner that does not use scene completion is shown in fig. 5.12. From 5 runs, run with median coverage is presented recorded at 25 minutes. The conventional planner uses only measured map for planning with gain function $\mathcal{G}_{unknown}(v)$ from sec. 3.5.2.

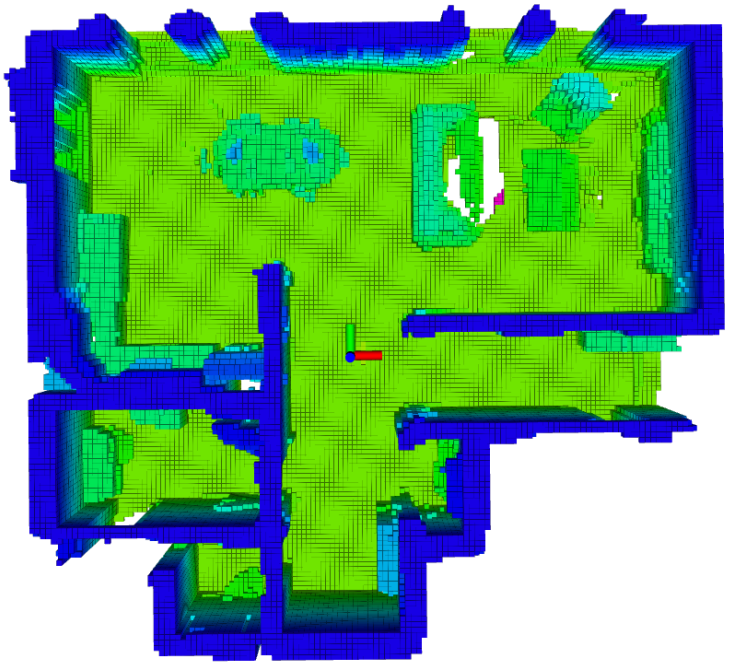


Figure 5.12: Map measured by conventional exploration planner

Qualitative Results- Conservative with Scene Completion Gain

The qualitative result of the conservative planner with scene completion gain is shown in fig. 5.13. From 5 runs, run with median coverage is presented recorded at 25 minutes.

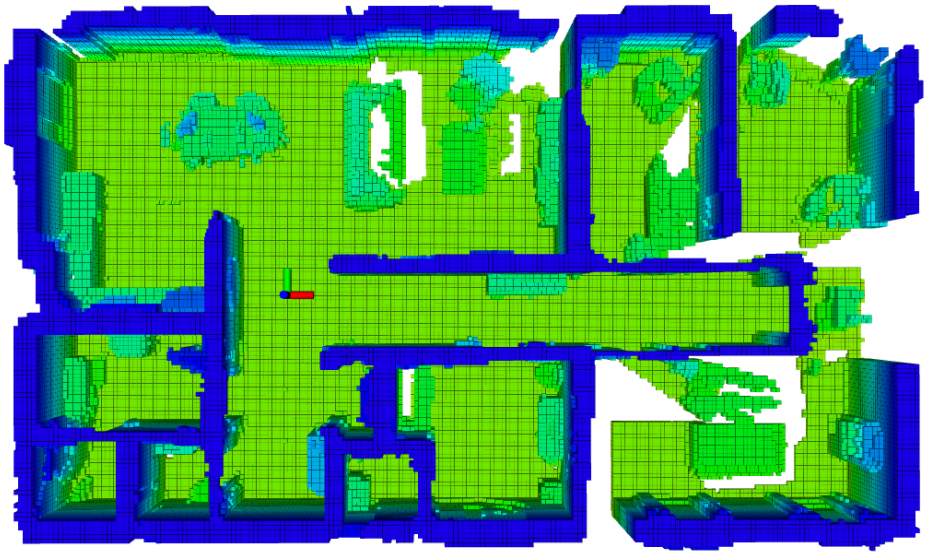


Figure 5.13: Map measured by SSC gain based exploration planner that uses conservative approach for utilizing measured and predicted maps. The conservative approach uses predicted map only for non-safety-critical information planning.

Qualitative Results- Cautiously Optimistic with Scene Completion Gain

The qualitative result of cautiously optimistic planner with scene completion gain from fig. 5.11 is shown in fig. 5.14. From 5 runs, run with median coverage is presented recorded at 25 minutes.

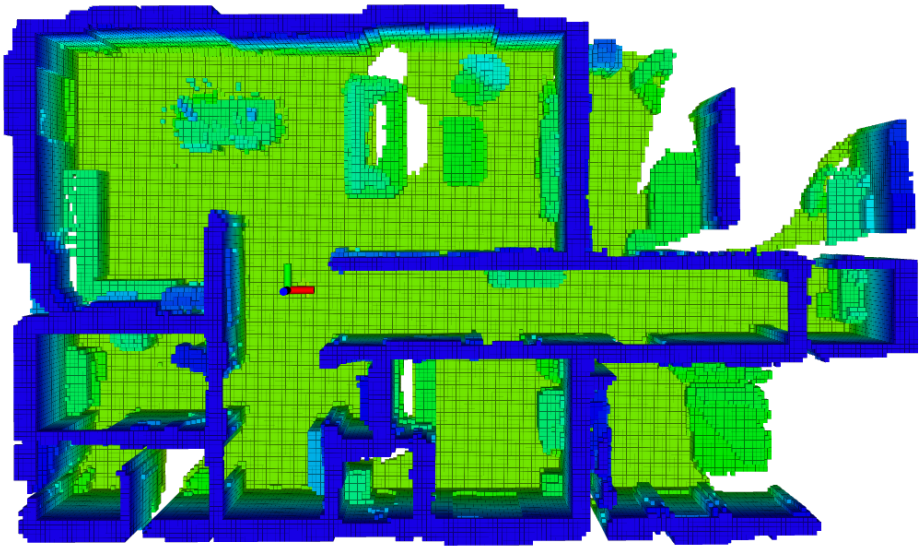


Figure 5.14: Map measured by SSC gain based exploration planner that cautiously uses high confidence voxels in the absence of measured map for all stages of planning.

Chapter 6

Conclusions and Future Work

6.1 Conclusion

In this thesis, semantic scene completion was investigated to address the limitations of conventional mapping arising from occlusions by inferring unknown regions and using the scene completed mapping for autonomous exploration in unknown regions. A deep-learning-based scene completed mapping framework was introduced that provides functionality for incrementally building scene completed maps from partial measurements in real-time. The framework is built upon an open-source mapping stack "Voxblox" that has been used by various state-of-the-art exploration planners making an adaption of this work convenient. Various fusion strategies were investigated to probabilistically fuse the completions into a global map. A log odds based probabilistic fusion strategy was proposed that addressed the limitations of current fusion approaches [18] for fusing noisy scene completions.
Safety concerns arising from using noisy scene completions for planning were addressed via two-tier mapping approach constituting measured and predicted maps in parallel. Probabilistic fusion strategies were explored to incorporate scene completions into the planning pipeline without violating safety concerns. This was achieved by advantaging from informative planner enabling fine-grained map utilization depending upon safety requirements. The measured map was used for safety-critical components of planning, while low confidence predicted map was utilized for non-safety-critical "Information Planning" part of the planning algorithm.
To maximize the utilization of scene completed mapping, a novel information gain formulation was introduced for the baseline informative planner to leverage completions for the planning problem. The proposed gain increased the certainty of the predicted map while identifying potential explorable regions not yet observed by the sensors, thereby improving the exploration efficiently.
The proposed scene completion-aware planning system was evaluated for safety, mapping quality, and exploration capability in a simulation setup with photo-realistic visuals and realistic physics using Airsim Simulator. A MAV with a depth and RGB camera was used as an exploration vehicle. The mapping and planning framework was integrated into a complete exploration planning pipeline. The results indicate the proposed scene completion aware planner was able to explore in challenging situations where the conventional planner would struggle, particularly in regions across obstacles with narrow entries where the proposed gain would identify potential explorable free regions while the conventional planner would prefer other unknown regions that might turn out to be an obstacle and not lead to new

explorable sub-regions.
The final evaluation of the proposed scene completion aware planning framework demonstrated > 34 % exploration improvement over conventional planner while maintaining identical safety across multiple sets of experiments.

6.2 Future Work

The results suggest the capability of the deep learned mapping and planning pipeline could be further improved by learning the gain function or sampling strategy via an end-to-end neural network to exploit scene completion for the planning problem. Furthermore, the improvements from the scene completions aware gain function suggest additional semantic information for trajectory gain evaluations could potentially enhance the planner as scene semantics have been shown to contribute to the scene completion tasks indicating strong co-relation with structural priors.

Bibliography

[1] J. Nikolic, M. Burri, J. Rehder, S. Leutenegger, C. Huerzeler, and R. Y. Siegwart, "A uav system for inspection of industrial facilities," *2013 IEEE Aerospace Conference*, pp. 1–8, 2013.

[2] S. Omari, P. Gohl, M. Burri, M. Achtelik, and R. Y. Siegwart, "Visual industrial inspection using aerial robots," *Proceedings of the 2014 3rd International Conference on Applied Robotics for the Power Industry*, pp. 1–5, 2014.

[3] R. Khanna, M. Möller, J. Pfeifer, F. Liebisch, A. Walter, and R. Y. Siegwart, "Beyond point clouds - 3d mapping and field parameter measurements using uavs," *2015 IEEE 20th Conference on Emerging Technologies & Factory Automation (ETFA)*, pp. 1–4, 2015.

[4] F. Colas, S. Mahesh, F. Pomerleau, M. Liu, and R. Y. Siegwart, "3d path planning and execution for search and rescue ground robots," *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 722–727, 2013.

[5] L. Roldão, R. de Charette, and A. Verroust-Blondet, "3d semantic scene completion: a survey," *ArXiv*, vol. abs/2103.07466, 2021.

[6] H. Oleynikova, Z. Taylor, M. Fehr, R. Y. Siegwart, and J. I. Nieto, "Voxblox: Incremental 3d euclidean signed distance fields for on-board mav planning," *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1366–1373, 2017.

[7] L. Schmid, M. Pantic, R. Khanna, L. Ott, R. Siegwart, and J. Nieto, "An efficient sampling-based method for online informative path planning in unknown environments," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1500–1507, April 2020.

[8] Y. LeCun, K. Kavukcuoglu, and C. Farabet, "Convolutional networks and applications in vision," *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pp. 253–256, 2010.

[9] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *NIPS*, 2014.

[10] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6602–6611, 2017.

[11] Y. Kuznietsov, J. Stückler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2215–2223, 2017.

[12] D. Hoiem, A. A. Efros, and M. Hebert, "Automatic photo pop-up," *ACM Trans. Graph.*, vol. 24, pp. 577–584, 2005.

[13] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. A. Funkhouser, "Semantic scene completion from a single depth image," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 190–198, 2017.

[14] J. Li, Y. Liu, D. Gong, Q. Shi, X. Yuan, C. Zhao, and I. D. Reid, "Rgbd based dimensional decomposition residual network for 3d semantic scene completion," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7685–7694, 2019.

[15] Y. Liu, J. Li, X. Yuan, C. Zhao, R. Siegwart, I. Reid, and C. Cadena, "Depth based semantic scene completion with position importance aware loss," 2020.

[16] P. Zhang, W. Liu, Y. Lei, H. Lu, and X. Yang, "Cascaded context pyramid for full-resolution 3d semantic scene completion," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7800–7809, 2019.

[17] X. Chen, K.-Y. Lin, C. Qian, G. Zeng, and H. Li, "3d sketch-aware semantic scene completion via semi-supervised structure prior," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4192–4201, 2020.

[18] S. cheng Wu, K. Tateno, N. Navab, and F. Tombari, "Scfusion: Real-time incremental scene reconstruction with semantic completion," *2020 International Conference on 3D Vision (3DV)*, pp. 801–810, 2020.

[19] R. Köhler, C. J. Schuler, B. Schölkopf, and S. Harmeling, "Mask-specific inpainting with deep neural networks," in *GCPR*, 2014.

[20] Y. Wang, D. J. Tan, N. Navab, and F. Tombari, "Forknet: Multi-branch volumetric semantic completion from a single depth image," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8607–8616, 2019.

[21] H. Pfister, M. Zwicker, J. van Baar, and M. H. Gross, "Surfels: surface elements as rendering primitives," *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, 2000.

[22] R. Szeliski, "Computer vision - algorithms and applications," in *Texts in Computer Science*, 2011.

[23] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996.

[24] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. W. Fitzgibbon, "Kinectfusion: Real-time dense surface mapping and tracking," *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pp. 127–136, 2011.

[25] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, 1987.

[26] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger, "Real-time 3d reconstruction at scale using voxel hashing," *ACM Transactions on Graphics (TOG)*, vol. 32, pp. 1 – 11, 2013.

[27] H. Oleynikova, M. Burri, Z. Taylor, J. I. Nieto, R. Y. Siegwart, and E. Galceran, "Continuous-time trajectory optimization for online uav replanning," *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5332–5339, 2016.

[28] L. Schmid, V. Reijgwart, L. Ott, J. Nieto, R. Siegwart, and C. Cadena, "A unified approach for autonomous volumetric exploration of large scale environments under severe odometry drift," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4504–4511, July 2021.

[29] A. Hornung, K. M. Wurm, M. Bennewitz, C. Stachniss, and W. Burgard, "Octomap: an efficient probabilistic 3d mapping framework based on octrees," *Autonomous Robots*, vol. 34, pp. 189–206, 2013.

[30] B. Yamauchi, "A frontier-based approach for autonomous exploration," *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97. 'Towards New Computational Principles for Robotics and Automation'*, pp. 146–151, 1997.

[31] H. H. González-Baños and J.-C. Latombe, "Navigation strategies for exploring indoor environments," *The International Journal of Robotics Research*, vol. 21, pp. 829 – 848, 2002.

[32] C. I. Connolly, "The determination of next best views," *Proceedings. 1985 IEEE International Conference on Robotics and Automation*, vol. 2, pp. 432–435, 1985.

[33] S. Shen, N. Michael, and V. R. Kumar, "Autonomous indoor 3d exploration with a micro-aerial vehicle," *2012 IEEE International Conference on Robotics and Automation*, pp. 9–15, 2012.

[34] J. Li, "palnet_opensource," https://github.com/waterljwant/SSC.

[35] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[36] C. T. Loop, Q. Cai, S. Orts, and P. A. Chou, "A closed-form bayesian fusion equation using occupancy probabilities," *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 380–388, 2016.

[37] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics*, 2017.

[38] L. Schmid and V. Reijgwart, "unreal_airsim," https://github.com/ethz-asl/unreal_airsim.

[39] A. Dai, D. Ritchie, M. Bokeloh, S. E. Reed, J. Sturm, and M. Nießner, "Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4578–4587, 2018.