A Qualitative Analysis Strategy Towards AI-Enabled 3D City Reconstruction

Andreas Christodoulides
967321

Submitted to Swansea University in partial fulfilment
of the requirements for the Degree of Master of Science

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

30th September 2023

Declaration

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed Andreas Christodoulides ..... (candidate)

Date 01/10/2023 .....

Statement 1

This work is the result of my own independent study/investigations, except where otherwise stated. Other sources are clearly acknowledged by giving explicit references. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure of this work and the degree examination as a whole.

Signed Andreas Christodoulides ..... (candidate)

Date 01/10/2023 .....

Statement 2

I hereby give my consent for my work, if accepted, to be archived and available for reference use, and for the title and summary to be made available to outside organisations.

Signed Andreas Christodoulides ..... (candidate)

Date 01/10/2023 .....

I would like to dedicate this work to my parents, Maria and Marios.

Abstract

3D urban scene reconstruction is a difficult problem in computer graphics, mainly due to unavoidable noise in real scene datasets and scalability. Whilst the state-of-the-art is able to produce reconstructions of various qualities, very few reconstruction approaches incorporate noise and clutter treatment strategies. Funded by Ordnance Survey, in this dissertation, we identify key areas of potential innovation in order to achieve large-scale and denoised 3D models of cities. To determine these areas of innovation, qualitative data from stakeholder engagement activities have been used to guide a literature survey on recent techniques in 3D reconstruction.

The KJ method, also known as Affinity Diagrams, has been employed to organise and visualise the qualitative data collected. The steps of the method were closely followed where appropriate, and the process of arriving at the finalised diagram is shown. The literature surveyed is then mapped on the Affinity Diagram to identify the strengths and gaps of the literature with respect to our stakeholder's requirements and workflow practices.

Following this approach, the main goal of our stakeholder has been defined as a multi-class reconstruction of urban scenes. Along with the 3D reconstructions themselves, two more areas for future investigation are identified, which combined promise to achieve our stakeholder's goal. The first area identified is semantic capabilities, which are capabilities referring to the ability to understand different objects and shapes within the scene. The second area identified is human-centric quality assurance, where the hypothesis is that the incorporation of people within the pipeline to be developed will greatly aid in achieving scalable multi-class reconstructions of that scale.

Acknowledgements

I would like to thank my parents, my siblings and my friends in both Limassol and Swansea for their support and my supervisory team for their continuous guidance.

Contents

Chapter 1

Introduction

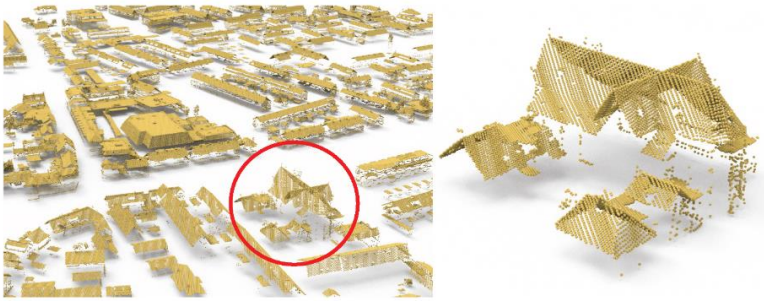
In 1971, Ordnance Survey was established as the national mapping agency of Great Britain [1]. Through the use of the Ramsden theodolite, they created accurate triangulation networks, which allowed for the creation of the first maps of that accuracy. Whilst they initially paved the way for modern map making, since then, the geospatial Industry has evolved, and it is now a need more than ever to incorporate a third dimension in their maps.

Today, data-gathering techniques have greatly evolved. By utilising satellites and aerial vehicles, ultra-accurate images of any landscape are available. These images can be directly converted into 3D through photogrammetry techniques, which convert images into 3D points. These points are called point clouds and are one of the dominant data types for 3D representations. Other point cloud generation techniques include using aerial and land-based LiDAR sensors, which, in essence, send electromagnetic waves in a direction and measure the time it takes for the wave to be reflected.

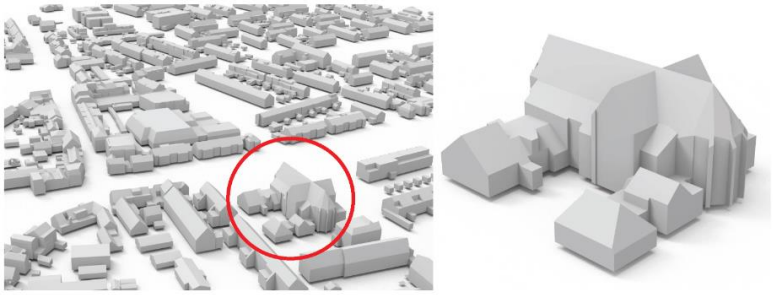
Using point clouds to generate more intuitive 3D meshes has always seen limitations in the unavoidable noise that comes with them. City-scale 3D reconstructions can be a foundation for many Industries, smart cities, digital twins, autonomous vehicles and assessment and management of infrastructure.

Reconstruction of 3D buildings through sensor points is a difficult problem in computer graphics [2]. The optimisation-based Polyfit method has presented a fast technique which provides 3D reconstructions whilst preserving structural details. Improving on their approach, the more recent City3D [3] can achieve building reconstructions on a very large scale, using aerial LiDAR data as input. However, as seen in Figure 1.1 the output shapes do not comprise building and architectural details, geographical details or environment.

On the other hand, deep learning approaches have shown to be very promising in covering these limitations as of late. For example, as seen in Figure 1.2, neural representations, such as NKSR [4], can provide much more detailed 3D reconstructions and do not limit themselves to buildings. However, it is demonstrated through this dissertation that these approaches mostly lack any capabilities of augmenting and interacting with the reconstructed scene. Techniques incorporating such capabilities are shown to have other limitations, such as being able only to infer a single object.



(**a**) Input airborne LiDAR point cloud.



(**b**) Our reconstruction result

Figure 1.1: City3D Reconstruction, taken from [3]

Funded by Ordnance Survey, the aim of this dissertation is to identify key areas of potential innovation, which, as an end result, can provide 3D reconstructions of urban landscapes. As the essence of our centre, we employ human-centred approaches for the identification of said areas of potential innovation. The main findings of the human-centred approaches are used to guide the analysis of a technical survey in the state-of-the-art of 3D reconstruction.

The KJ method [5], also known as affinity diagramming, was employed to analyse qualitative data collected through interactions with our stakeholder. The themes uncovered from the Affinity Diagram relate to our stakeholder's technical requirements, the data modalities available, industrial standards and human-in-the-loop. These themes are used

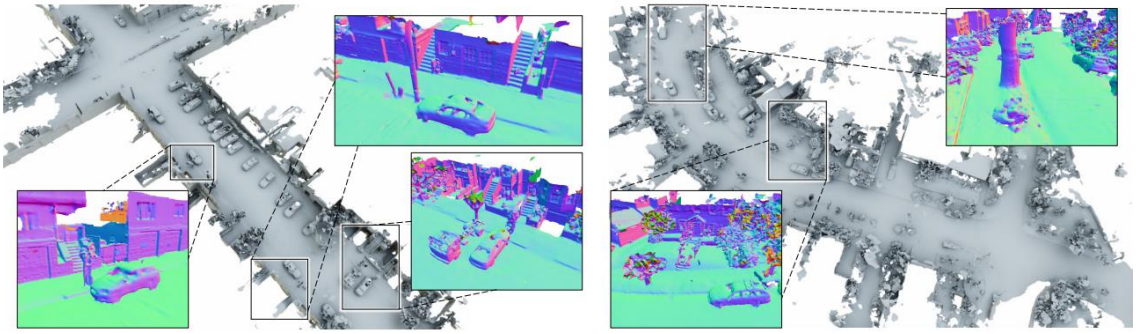


Figure 1.2: NKSR Reconstruction, taken from [4]

to guide our survey and assess the latest approaches in 3D reconstruction. Combining the insights for the Affinity Diagram and our literature survey, we identify key areas of potential innovation.

Following this introduction, Chapter 2 provides the necessary background, Chapter 3 provides the technical survey on 3D reconstruction. In Chapter 4, the methods undertaken to produce this dissertation are shown. In Chapter 5, our main findings are demonstrated. In Chapter 6, we discuss our main findings and in Chapter 7, we explain our limitations and future research that will be built from this project. Conclusions are shown in Chapter 8.

Chapter 2

Background

In this Chapter, some context is provided regarding the topics of this dissertation. In Section 2.1, some knowledge of 3D Vision is provided. In Section 2.1.2, some fundamentals on Human-centred methods and Affinity Diagrams are shown. In Section 2.3, fundamental neural approaches are presented.

2.1 3D City Vision

In this section, some context on 3D Vision techniques is provided. 3D Vision is categorised into 3D reconstruction, classification and segmentation. 3D reconstruction is concerned with capturing the shape from a sensor modality and returning it as a 3D model. Whereas classification and segmentation are semantic tasks which are concerned with understanding individual shapes and objects in the sensor data. In Section 2.1.1, some information on 3D reconstruction is shown and in Section 2.1.2, some preliminaries on semantic tasks.

2.1.1 3D Reconstruction

Reconstruction techniques in 3D spaces can be categorised as explicit or implicit [6]. Explicit surfaces are defined by quantifiable geometrical parameters and can be parametric or triangulated. Parametric surfaces refer to surfaces where primitives are deformed to fit the shape points, whereas triangulated surfaces are formed by connecting input points into triangles. Implicit surfaces solely rely on functions whose isosurface approximates the input data and require post-processing techniques for visualisation, such as marching cubes.

3D reconstruction has been a monumental topic of interest in the area of computer graphics and can trace its steps back to over fifty years ago [7]. William E. Lorensen revolutionised the field with the marching cubes, a method which allows for wrapping explicit surfaces over implicit functions [8]. This technique overcame the main limitation of 3D surfacing at the time, prohibiting computational complexity. Owing to its efficiency and wide applicability, it is widely used today. Since then, techniques of 3D computer vision have been a very popular field of research owing to its many applications, such as simulation and autonomous path navigation.

Only a small portion of the literature actively contributes to achieving 3D maps. As already shown, [3] and [4] utilise aerial and land-based LiDAR data, respectively, to achieve their reconstructions. Other approaches [9–11] identify building geometries, some able to even achieve large-scale reconstructions as seen in Figure 2.1. Others are able to provide both small-scale detailed and large-scale but very rough reconstructions [12], as seen in Figure 2.2. As will be demonstrated in Chapter 3, the majority of the state-of-the-art incorporates deep learning approaches to achieve their reconstructions and does not focus on urban landscapes-related reconstructions.

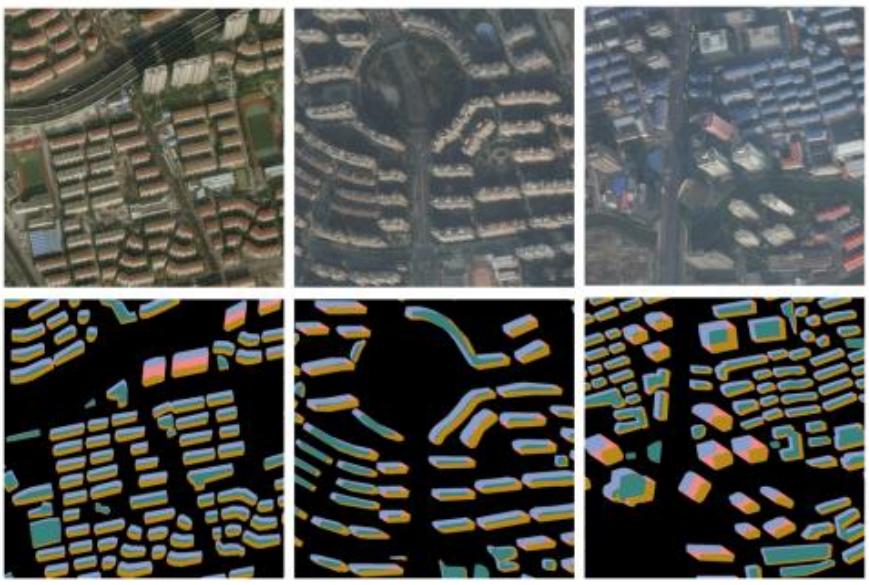


Figure 2.1: MTBR-Net large-scale 3D reconstruction, taken from [11]

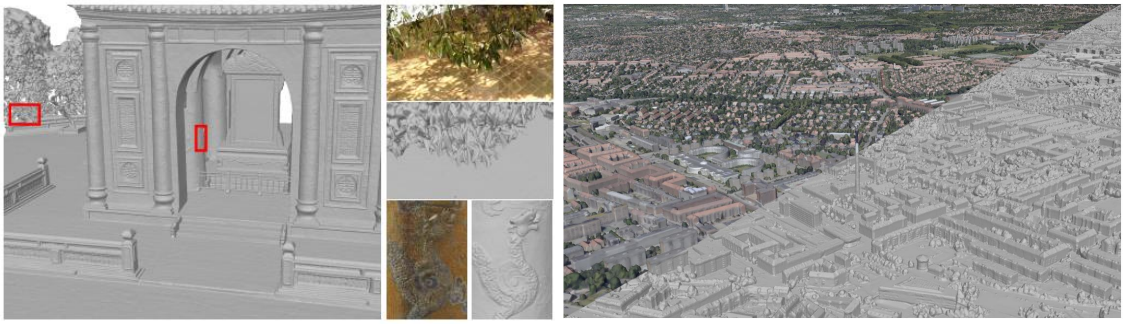


Figure 2.2: Out-of-core reconstruction method, small-scale scene (left), large-scale scene (right), taken from [12]

3D reconstruction is a leading topic in terms of digital products, with the largest technology companies in the world leading it. Where Google Earth Engine utilises AI to analyse enormous datasets from satellite imagery, enabling planet-scale analysis [13]. Microsoft Building Footprints utilises AI to identify map features at scale and has released millions of building footprints as part of their humanitarian efforts [14]. QGIS, an open-source GIS which provides viewing, editing, printing and analysis of geospatial data [15]. ArcGIS has similar functionalities [16]. However, it allows a great deal of data interactivity, which unlocks new horizons for analysis. FugroViewer allows for the visualisation and interpretation of geospatial data in addition to providing terrain models, as seen in Figure . However, FugroViewer outputs subpar 3D reconstructions due to reliance on Triangulation Irregular Networks (TIN). TerraXplorer Pro, developed by Skyline software, in addition to capabilities for quality reconstructions, utilises AI to perform semantic tasks, such as segmentation.

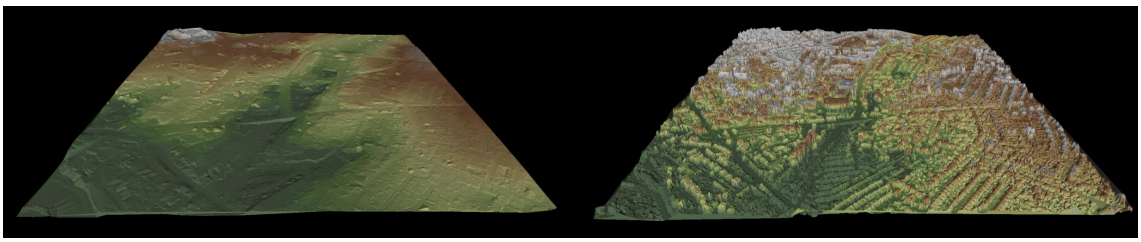


Figure 2.3: FugroViewer Reconstruction of terrain (Bear Earth) model (left) and TIN (right), provided by Ordnance Survey.

2.1.2 Semantic Understanding

Semantic tasks can be categorised as either classification or segmentation [17,18]. Classification is the process of identifying what the data represents and assigning an appropriate label. Segmentation, on the other hand, takes this concept a step further, breaking down the input data into segments and assigning labels to individual components. For point cloud data, classification assigns class labels to specific points based on different global criteria, whereas segmentation predicts point labels based on point-level characteristics [17,18].

In Geographic Information Systems (GIS), semantic information can greatly enrich the data by providing context to the captured geographical coordinates [19]. Such data enrichment can include building installations, door entries and exits of buildings, land use and land cover, to name a few [19]. However, GIS datasets oftentimes include high amounts of noise and clutter, where semantic information can be used to treat appropriately. Said noise and clutter need to be identified and treated carefully, as some industries use information that is meaningless for other industries. To paraphrase from the original quote, one industry's noise is another industry's signal.

It is seen that generalised semantic capabilities, whilst a very active area of research, are vastly different from 3D reconstruction. Approaches focusing on reconstruction are seen to incorporate geometric parameters of buildings to identify them from their edges [9,10], whereas others utilise building-specific semantics [11]. It is shown through this dissertation that, in general, semantic capabilities in 3D reconstruction approaches are limited.

2.2 Human-centred Software Design

2.2.1 Human-Centred Software Design in Industrial Contexts

Over the years, the significance of incorporating User Centred Design (UCD) approaches in industrial contexts has been repeatedly demonstrated. Early work on Human-computer Interaction (HCI) incorporation in Industry has identified that larger organisations mostly employed such approaches at very low scales [20]. Furthermore, it was identified that the interaction of usability studies was avoided due to the perception of being too time-consuming, expensive relative to reward, limited in diversity and applicability, and very complex [20,21]. According to HCI professionals, resource constraints and resistance to user-centred approaches were the major prohibiting factors in employing UCD [20].

Central to enabling these techniques is the role of qualitative data collection. Qualitative data are non-numeric data which are obtained through open-ended and conversational communications aiming to provide an understanding of the underlying reasons, opinions and motivations [22]. Common qualitative data-gathering techniques employed by HCI practitioners include interviews [23], focus groups [23,24], surveys [20] and prototyping [22]. Analysis of qualitative data involves grouping the interviewed population into themes and interpreting data based on the themes developed, [20,22,24]. Another key aspect of thematic analysis is specific participant responses, which can benefit the researcher when attempting to contextualise the themes generated and draw more meaningful insights.

In industrial contexts, geographical data introduce additional complexities. Many industrial applications involve working with geographical information, and the interaction and visualisation systems surrounding these types of data can vary greatly depending on data sources, system purposes and interfaces [25]. Dealing with complex data is a challenge in general for such applications. Different individual users with diverse tasks will often use the system in unique and unpredicted ways, which can vary greatly depending on data sources, system purposes and interfaces [26]. Previous research has demonstrated that individual user's experience with geographic information data is a significant factor in determining potential usability issues with the software being developed. This is exactly what we aim to do through this dissertation, which will show how experts from the field guide our Affinity Diagram and, thus, the survey analysis.

2.2.2 Affinity Diagrams

The KJ method, more commonly known today as Affinity diagramming, is an organisational method developed by anthropologist Jiro Kawakita [5]. It can be used as a quality control method that is focused on allowing creativity when analysing unstructured qualitative data [27]. This is achieved by undertaking a bottom-up approach in data grouping, which eliminates preconceived notions and biases of the researchers [5]. Affinity diagramming has been previously used in a diverse set of applications, including user experience design [28], interactive prototype evaluation [29], and logistics optimisation [30], to name a few. In addition, an interest in digitalising affinity diagramming is evident in recent literature [27, 28], allowing for optimisation of the affinity diagramming process and greatly aiding the researcher when dealing with large sets of quantitative data.

These are the four key steps to be followed when creating an affinity diagram, according to [5]:

1. Label making
2. Label grouping
3. Chart making
4. Written or verbal explanation

The first step undertaken in producing an Affinity Diagram is label making. In this step, qualitative information is transcribed into labels, each representing a single thought or statement. During the second step, said labels are shuffled and grouped based on label statement affinity, i.e., natural relationships between labels. This is an iterative process, where the groups formed over several iterations indicate broader categorisation. The different iterations in this step include carefully reading the labels and performing label and grouping adjustments to minimise group number. This is a crucial step in Affinity Diagramming, as the non-linear method of label grouping is what constitutes the nonlinear nature of this diagram. The third step in creating an Affinity Diagram is chart-making, where the groups formed are placed on a unifying chart. This chart attempts to join groups into larger thematic concepts where relevant and show causal and effect relationships between groups or individual labels. This is a decisive step for revealing inter-relationships and patterns between groups and optimising spatial arrangements to correctly identify the thematic concepts. Finally, a written or verbal explanation of the chart must be supplied to provide the reader with an understanding of the data and enforce the core ideas and patterns evident from the diagram.

2.3 Fundamental Neural Architectures and Representations

This section briefly describes the preliminary information required to follow the rest of the dissertation. The literature surveyed is diverse in that techniques can comprise many modules, each addressing a different function of the overall architecture. This section describes the preliminaries required to follow the rest of the dissertation. These preliminaries include basic and advanced model architectures and neural representations.

2.3.1 Multilayered-Perceptrons (MLP)

MLPs are a simple architecture of neural networks comprised of simple interconnected neurons [31]. They receive and output data in the form of vectors. Data are propagated forward as weights in their hidden layers, which can be one or many. The inputs in each hidden layer are the outputs of the previous one, and new weight sums are computed in the corresponding neurons of the layer. Through an activation function, nonlinearity is achieved.

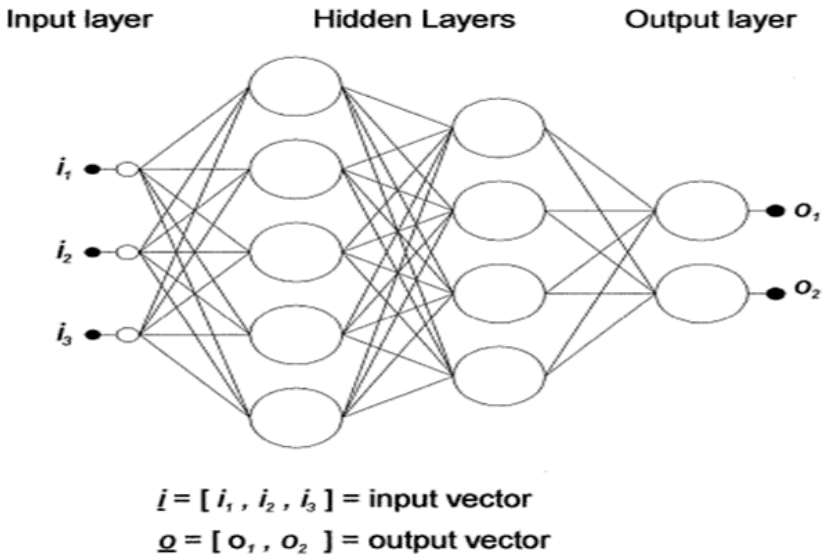


Figure 2.4: A multilayer perceptron with two hidden layers, taken from [31]

2.3.2 Convolutional Neural Networks (CNN)

L. Atlas et al. [32] carried out the first implementation of CNNs to learn dynamic patterns. CNNs differentiate themselves from traditional neural networks in that they employ convolutional layers, which in turn use a filter that computes different feature maps. Many different architectures arise from CNNs. A very popular approach is the U-Net architecture [33]. First applied to biomedical image segmentation, its name is derived from its architecture, which resembles a U-shape, where a contractive fully convolutional network is followed by an expanding one, as seen in Figure 2.5. The contractive and expanding networks are called the downsampling and upsampling stages, respectively.

High-resolution features from the contracting stage are used within the upsampling stage to assemble a more precise output.

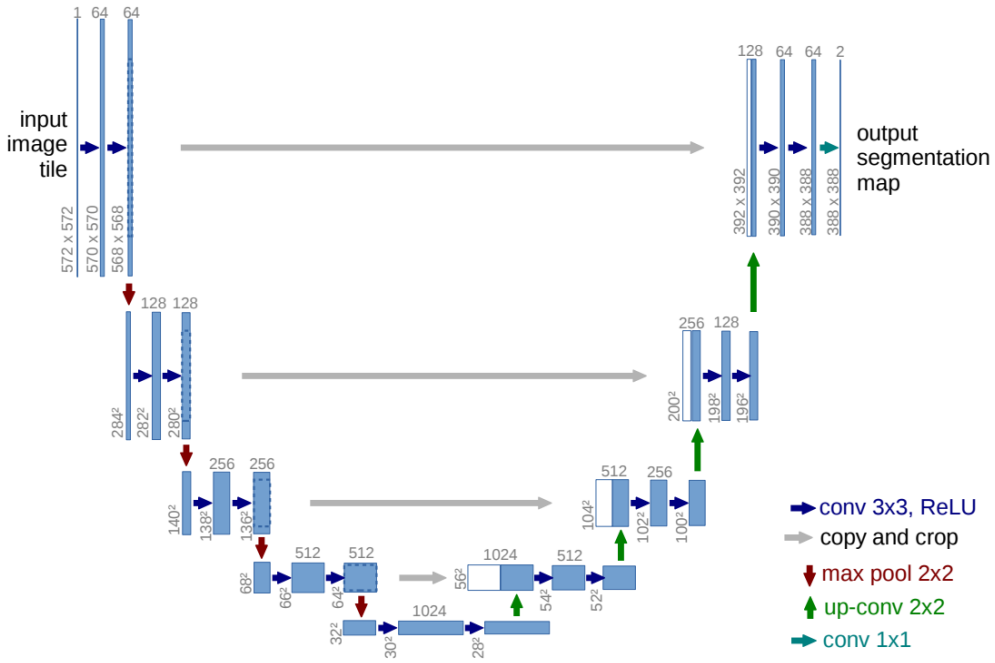


Figure 2.5: U-Net architecture: Each individual blue box corresponds to a multi-channel feature map. The channel number is denoted on top of the box. The white boxes represent copied feature maps. The arrows denote different operations. Taken from [33]

2.3.3 Vision Transformers (ViT)

Vaswani et al., proposed the Transformer models, completely relying upon attention mechanisms to map relationships between inputs and outputs in natural language processing applications [34]. They utilise a fully connected network seen in Figure 2.6, where both encoder and decoder utilise self-attention mechanisms. Research from Google extended this technique into vision applications with the Vision Transformer (ViT) [35]. They overcome the reliance of convolutional networks for image processing by directly applying the transformer architecture on sequences of image patches.

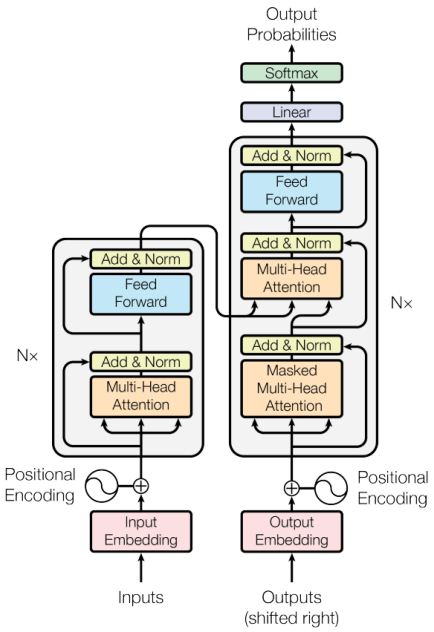


Figure 2.6: Transformer architecture taken from [34]

2.3.4 Variational Autoencoders (VAE) and Vector Quantised - Variational Autoencoders (VQ-VAE)

Kingma and Welling introduced the Variational Autoencoder (VAE) [36], which utilises neural networks. As an initial stage, it maps the input data into a latent space, the encoder stage. It is followed by a decoder, which maps the data from the latent space to the desired output. Van den Oord et al., by learning the latent space with a technique inspired by vector quantisation to develop the Vector Quantised-Variational Auto Encoder (VQ-VAE) [37]. Their VAE architecture includes two shallow CNNs as encoder-decoder.

2.3.5 Diffusion-Denoise Models (DDM)

DDMs, inspired by non-equilibrium thermodynamics and stochastic differential equations, are models that incrementally add noise to a data sample, the diffusion process [38, 39]. They become generative models by undertaking the reverse diffusion process, which inverts diffusion [38].

2.3.6 Neural Radiance Field (NeRF)

The NeRF framework, introduced by Mildenhall et al., utilises a deep, fully connected MLP without any convolutional layers to render the scene as a volume density with directional emitted radiance at any point in space [40]. The volume rendering is represented by a 5D vector, with a Cartesian coordinate in the 3D space and radiance emitted in each direction. An overview of the NeRF method is seen in Figure 2.7

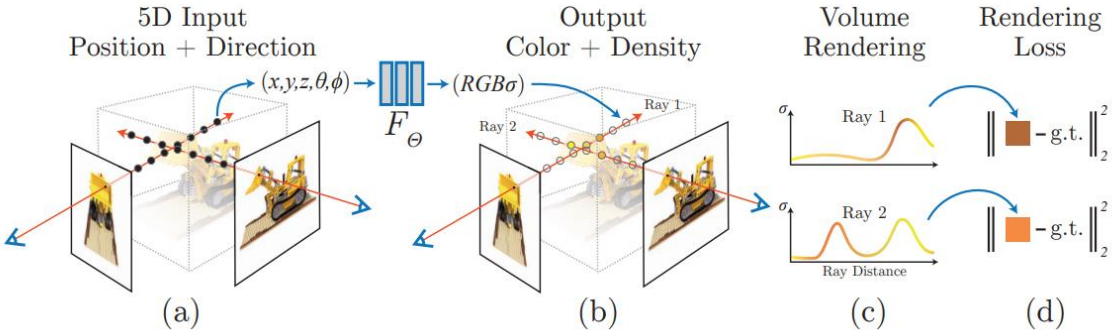


Figure 2.7: Overview of the neural radiance field scene representation and differentiable rendering procedure: Images are synthesised by sampling 5D coordinates along camera rays (a), feeding those locations into an MLP to produce colour and volume density (b), and using volume rendering techniques to composite these values into an image (c). This rendering function is differentiable, so the scene representation can be optimised by minimising the residual between the synthesised and ground truth observed images (d), taken from [40]

2.3.7 Neural Implicit Surfaces: NeuS

The NeuS framework, presented by Wang et al. [41], introduces a neural surface reconstruction method that integrates volume rendering techniques for multi-view reconstruction. The framework is explicitly inspired by NeRF, which represents a scene as a volume density with emitted radiance. Unlike NeRF, NeuS aims to learn a neural implicit surface and is optimized for high-fidelity reconstructions.

2.3.8 Neural Kernel Fields (NKF)

The authors of NKF [42] utilise a kernel with learnable parameters which solve simple positive linear systems and predict an implicit function. This approach directly receives points with surface normal sampled from the set of input points. They utilise the Neural

Spline kernel function to define the data-dependent kernel. Coefficients learned by the kernel function are used to predict new functions for new points.

2.3.9 Backbones

Backbone networks are deep, pre-trained models used for feature extraction, usually at the beginning of a pipeline [43]. Whilst they can be of many architectures, convolutional backbones are the most popular. The Residual neural network (ResNet) is a fully convolutional neural network created for image recognition [44]. The convolutional layers were proposed to alleviate the vanishing gradients during the backpropagation algorithm, which was very common with deep neural networks. On the other hand, PointNet is heavily inspired by CNNs in that it uses weight sharing and maximum pooling whilst not incorporating any convolutional layers [18]. Instead, it uses an MLP shared for all points after an initial spatial transformation network, which attempts to canonicalise the data before processing. PointNet is used for classification and point-level segmentation, where its architecture is seen in Figure 2.8.

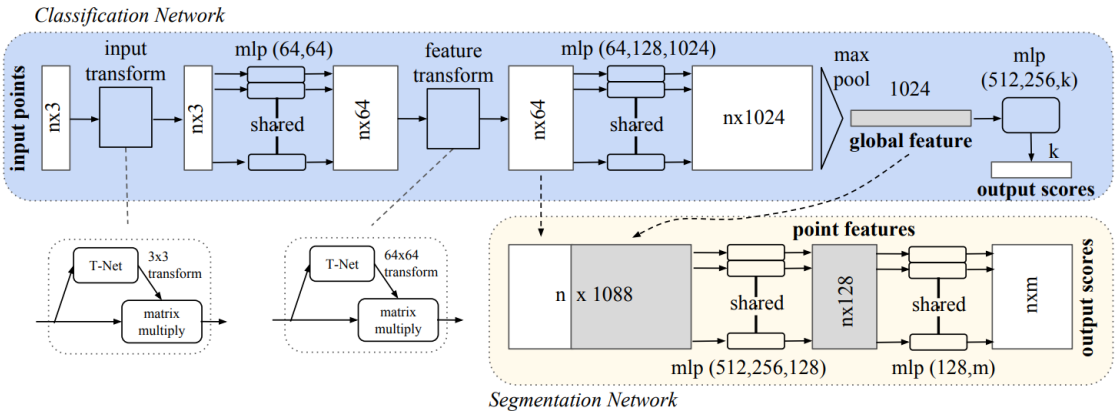


Figure 2.8: PointNet architecture, taken from [18]

Chapter 3

Literature Survey

We present a survey with key elements of analysis being 3D reconstruction techniques along with modelling approaches and architectures. Such reconstruction techniques include both explicit and implicit representations. It is shown that individual modules in the approaches discussed contribute to the resulting reconstructions in different ways, whether used for learning or modelling. Moreover, data input modalities, semantic scene understanding capabilities and human incorporation of the literature are analysed.

The focus of the survey is extended beyond urban-specific reconstructions such that a much better understanding of the state-of-the-art is gained. The analysis themes are guided by our Affinity Diagram, shown in Chapter 5. Following the literature analysis with respect to our themes, a critical analysis is presented. Section 3.1 examines the modeling approaches undertaken by 3D reconstruction techniques. Section 3.2 relate to the data input modalities. Section 3.3 demonstrates the reconstruction modalities of the approaches, i.e., the data outputs. Section 3.4 shows approaches incorporating scene and semantic capabilities. Section 3.5 techniques with human-in-the-loop elements and finally Section 3.6 summarises the insights from this survey.

3.1 Modelling Approaches

Modelling approaches for 3D reconstruction are often seen to be comprised of multiple models and architectures. In Section 2.3, we have shown the fundamental models, representations and architectures that mostly comprise the literature analysed. In Table 3.1, a summary of the modelling approaches used by the literature is visible. Section 3.1.1

focuses on approaches that employ MLP modules. Section 3.1.2, discusses approaches that incorporate convolutional modules. Section 3.1.3 covers approaches that utilise transformer modules. Section 3.1.4 examines VAE and VQ-VAE architectures. Section 3.1.5 explores techniques that employ diffusion models, and finally, Section 3.1.6 delves into optimisation based techniques.

3.1.1 Multi-Layered Perceptrons (MLPs)

MLPs are seen to be the primary technique for producing neural representations. Directly building upon NeRF, [57] utilises an MLP decoder for density field prediction. Authors of [67] chose to save the density and colour modalities of NeRF into explicit voxels. They can directly render volumes, and with a shallow MLP, they can predict colour emission. The authors of RealFusion [58] utilise Instant-NGP to increase their NeRF rendering speed.

On the other hand, [51, 59, 60, 62], Improve upon NeuS, leveraging the cheaper SDF representations [41]. Neuralangelo authors [51] focus on allowing surface normals computation everywhere, improving detail refinement. PermutoSDF [59] utilises MLPs for SDF and colour computations and follows NeuS for approximating their SDFs. NeuDA [60] improves the approximated SDFs by initialising NeuS with a voxel grid comprising 3D positions called "anchors", which are hierarchically concatenated. ShadowNeus utilises the same strategy as NeuS and, thus, an MLP for producing their representations. However, they utilise their proposed "shadow rays", which are more physically correct, based on the assumption that a scene does not emit any light.

MLPs are seen as a default solution for data processing. They are used as decoders by [47, 48, 50, 52]. For [50] and [48], they are used for decoding the attention features. ShapeClipper [47], on the other hand, utilises two MLPs to directly predict SDFs and RGB values. In their pipeline, authors of ALTO [52] utilise PointNet to extract features from point clouds and MLPs to decode attention features. ManhattanSDF [64] through a single MLP can predict SDFs, colour fields and semantic logits, improving reconstruction accuracy by enforcing the Manhattan-world assumption [75]. Authors of [39] utilise MLPs for their noisy point cloud processing.

Some approaches are seen to be entirely comprised of MLPs, such as [65, 66, 72]. TARS [65] utilises DeformNet as an MLP backbone to compute spatial mappings and point features. The mappings and features are then fed to another MLP to compute neural representations of the reconstructed shape. The authors of [66] incorporate an MLP query

Table 3.1: Modelling Approaches used by the literature

Author Papers	Models and Architectures						
	DDM	VAE	VQ-VAE	Transformer	MLP	CNN	Optimisation
LION [38]	×	×	-	-	-	×	-
AutoSDF [45]	-	-	×	×	-	×	-
SDFusion [46]	×	-	×	-	-	-	-
ShapeClipper [47]	-	-	-	×	×	×	-
SRDF [48]	-	-	-	-	×	-	-
OG-INR [49]	-	-	-	-	×	-	-
VolRecon [50]	-	-	-	×	×	-	-
Neuralangelo [51]	-	-	-	-	×	-	-
ALTO [52]	-	-	-	-	×	-	-
BUOL [53]	-	-	-	-	-	×	-
AutoRecon [54]	-	-	-	×	-	-	-
Part Retrieval and Assembly [55]	-	×	-	-	-	-	-
SparseFusion [56]	×	-	-	×	×	×	-
Behind the Scenes [57]	-	-	-	-	×	×	-
PC^2 [39]	×	-	-	-	×	×	-
RealFusion [58]	×	-	-	-	×	-	-
PermutoSDF [59]	-	-	-	-	×	-	-
NeuDA [60]	-	-	-	-	×	-	-
SECAD-Net [61]	-	-	-	-	-	×	-
ShadowNeuS [62]	-	-	-	-	×	-	-
NKSR [4]	-	-	-	-	-	×	-
City3D [3]	-	-	-	-	-	-	×
POCO [63]	-	-	-	-	-	×	-
HEAT [9]	-	-	-	×	-	×	-
ManhattanSDF [64]	-	-	-	-	×	×	-
TARS [65]	-	-	-	-	×	-	-
Predictable Context Prior [66]	-	-	-	-	×	-	-
Direct Voxel Grid Optimisation [67]	-	-	-	-	×	-	-
Out-of-Core [12]	-	-	-	-	-	-	×
Vis2Mesh [68]	-	-	-	-	-	×	-
PSR [69]	-	-	-	-	-	×	-
GaussianFussion [70]	-	-	-	-	-	-	×
Search and Evaluate [10]	-	-	-	-	-	×	-
RetrievalFuse [71]	-	-	-	-	-	×	-
Learning SDF for Multiview Surface Reconstruction [72]	-	-	-	-	×	-	-
VolT [73]	-	-	-	×	-	-	-
3DIAS [74]	-	-	-	-	×	×	-
MTBR-Net [11]	-	-	-	-	-	×	-

network that globalises local features learned from PointNet and an SDF network that produces an implicit scene representation. Authors of [72] provide a self-supervised approach where two different MLPs are utilised, one for SDF computing and a second one with learnable parameters for computing the light field. Said learnable parameters are utilised to overcome the non-differentiable nature of the light rays and allow for the intersection point computation of the rays and the field.

3.1.2 Convolutional Neural Networks (CNNs)

Convolutional encoders are the most popular applications of CNNs, followed by backbones. ResNet is used as a convolutional encoder for [45,47,53,56,57,69]. For AutoSDF [45], this is one of two domain-specific encoders, which are followed by encoders for up-convolutions to 3D. ShapeClipper [47] and SparseFusion [56] find the only convolutional element of their approach in their encoder. In [69], a siamese ResNet encoder is used for plane detection and camera pose estimation modules. In addition, PlanRCNN is used for plan detection, producing features from the ResNet backbone [76]. Authors of BUOL [53] utilise a ResNet encoder and a Deeplab backbone as a three-branch decoder. The decoder is able to predict semantic maps, multi-plane occupancies and a depth map. In addition to their point cloud processing, the authors of PC^2 [39] also process voxels using a Un-Net. A U-Net feature extractor is also used by RetrievalFuse [71]; however, it is used for point cloud processing instead. Authors of NKSR [4] utilise a U-Net encoder and decoder to obtain sparse voxel grids from extracted point features and to reverse the encoding process.

Authors of [9,11,63] rely on other convolutional backbones. Authors of POCO [63] utilise the FKAConv backbone to perform point convolutions and compute vectors in the latent space, where they identify that the convolutional backbone is more efficient than PointNet.Authors of HEAT [9] utilise a ResNet backbone of feature extraction in their architecture, whereas the authors of MTBR-Net [11] find a convolutional backbone in the high-resolution network HR-Net which unlocks their diverse semantic capabilities. Authors of VolT [73] use a pre-trained CNN to generate initial 2D-view embeddings.

Other approaches are fully convolutional [10, 61, 68]. SECAD-Net [61] utilises a standard 3D CNN encoder to extract features from input engineering sketches. A single fully connected layer decodes extrusion parameters of a 2D distance field with heights. Vis2Mesh [68] exploits depth completion for visibility prediction. They utilise three modules with convolutional encoders and decoders. A renderer predicts depth maps,

which are fed into the first module, CoarseVisNet, which predicts pixel visibility. Their DepthCombNet module then receives the sparse depth map and provides a dense and complete depth map. Their final module, FineVisNet, takes the depth maps from the renderer and DepthCombNet as input to predict fine pixel visibility. Finally, in Search and Evaluate [10], they utilise an architecture comprised by two U-Nets. One U-Net is deep and is used for extracting global features of buildings. A shallow U-Net is incorporated to provide pixel-wise classification scores based on predicted building edge accuracy.

3.1.3 Transformers

Transformer modules utilised in the pipelines analysed see varying levels of involvement. The approaches [52] and [71] use attention mechanisms but not transformer modules. For ALTO [52], the authors utilise an attention module to decode the latent space. The authors of [71] approximate the reconstruction task as a composition of cropped chunks and utilise a patch-attention module to only make use of useful features from their generated chunks.

The authors of [9,47,56] use transformers for different tasks in their pipelines. ShapeClipper [47] finds its transformer element in its CLIP backbone, which generates encodings for their image inputs. SparseFusion [56] utilises a feature transformer that is used to predict the colours of novel views. HEAT [9] utilises transformers to incorporate learned features onto edge nodes.

Techniques that utilise more transformer involvement are found in [50,54,73]. Authors of AutoRecon [54] utilise a self-supervised ViT to combine features from images onto point clouds. Through a 3D transformer, they can segment the point cloud into foreground and background regions. In VolRecon [50], they use a view transformer and a ray transformer to learn projection features and ray information. They utilise the ray information to compute their SRDFs for all points along the rays computed. Authors of VolT [73] propose a 3D Vision Transformer framework to make the most out of multi-view images. A 2D-view transformer encoder is used to receive the 2D-view embeddings. With a 3D-volume transformer decoder, they correlate different spatial locations in the global domain to explore relationships between the spatial and view domains.

3.1.4 Variational Autoencoders (VAE) and Vector Quantised - Variational Autoencoders (VQ-VAE)

A limited amount of authors are seen to build their approaches as VAE-based generative networks [38, 55]. The authors of SparseFusion [56] set their overall approach as a VAE to generate plausible, accurate and realistic renderings. Xu et al. [55], on the other hand, set their framework as a VAE such that they make use of the VAE's continuous latent space to allow to transform their binatorial problem of search and retrieval through a database into a continuous optimisation problem. Cheng et al., with AutoSDF [45] and SDFusion [46], make use of a VQ-VAE framework such that they take advantage of the discretised space and alleviate the computational complexity of their DDM. Their approach allows for decoding high-quality outputs whilst allowing for much easier computations.

3.1.5 Diffusion-Denoise Models (DDM)

DDMs are seen to be utilised in a broader fashion. The authors of LION [38], owing to the VAE framework, can train two DDM models on hierarchical latent spaces which combine a global latent representation with a point-structured latent space. Compared to DDMs that operate directly on point clouds, they report better performance. Contrastingly, SDFusion [46] employs DDMs in a VQ-VAE framework where the discredited space allows for the reduction of the high-resolution 3D shapes and thus, the DDM can be trained on latent representations.

Z. Zhou et al., with SparseFusion [56], extend the DDM concept, introducing the Diffusion Distillation procedure. A diffusion model that works on computationally cheap features extracted through a transformer to recover a latent representation of the ground truth image and guide the reconstruction.

L. Melas-Kyriazi et al., with PC^2, incorporate a gradual diffusion process where, at each step, the image features gained through the encoder are projected into the partially denoised point cloud and augmenting each point according to the feature sets. In their follow-up approach [58], they utilise the open-source Stable Diffusion [77] as their DDM of choice to generate diffusion priors, which make up for missing information when reconstructing through Instant-NGP [78].

3.1.6 Optimisation

Whilst not neural approaches, state-of-the-art comprises some non-data-driven approaches [3,12,70]. Huang et al., [3] with City3D, provide an optimisation framework for generating large building reconstructions. They combine footprint and point cloud data to extract segments through the point cloud, and the building is extracted using its footprint. They then extract polylines from a TIN height map and optimise extracted planes and polylines to generate 3D models as meshes. The authors of GaussianFussion build upon the well-known TSDF fusion reconstruction method [79]. They focus on fusing parameters gained by different views, undertaking an algorithmic approach that exploits geodesic curves between and Gaussian measurements. Through a simplex network, they optimise the geodesic curves. N. Poliarnyi proposes the utilisation of the total generalised variation minimisation (TGV) algorithm to generate large-scale scenes whilst delivering a GPU-friendly approach. Using LiDAR data or depth maps, they construct octrees, which are optimised through the TGV by building hierarchical treetops.

3.2 Data Input Modalities

It is shown that our stakeholder is primarily interested in utilising point clouds gained from aerial images and converted through photogrammetry and vehicle LiDAR data. In this section, the input modalities of the different techniques are analysed. Table 3.2 shows a summary of the input modalities used by the literature. Section 3.2.1 shows literature that utilises point cloud inputs. Section 3.2.2 shows techniques that can utilise a single image to infer their reconstructions. In Section 3.2.3, approaches which utilise multi-view images are demonstrated. Finally, in Section 3.2.4, the remaining modalities are seen.

3.2.1 Point Cloud Inputs

Approaches [3,4,12,38,45,46,49], directly receive and operate on point cloud data.AutoSDF [45] splits the point cloud inputs into patches and encodes the patches independently. NKSR [4], utilising oriented point cloud inputs, predicts a voxel hierarchy which enables their technique. Optimisation-based [3] and [12] directly receive and operate on aerial LiDAR data. X. Xu et al. designed their framework to receive a target shape as a point cloud.

The remaining techniques incorporating point clouds utilise backbones to extract latent features from the input point clouds [52, 63, 66], [71]. Where approaches [52] and [66]

Table 3.2: Input modalities used by the literature

Author Papers	Input Modalities						
	Point Cloud	Mesh	Single Image	Language	Multiple Images	Video	RGB-D
LION [38]	×	-	-	-	-	-	-
AutoSDF [45]	×	-	-	-	-	-	-
SDFusion [46]	×	-	×	-	-	-	-
ShapeClipper [47]	-	-	×	×	-	-	-
SRDF [48]	-	-	-	-	×	-	-
OG-INR [49]	×	-	-	-	-	-	-
VolRecon [50]	-	-	-	-	×	-	-
Neuralangelo [51]	-	-	-	-	×	-	-
ALTO [52]	×	-	-	-	-	-	-
BUOL [53]	-	-	×	-	-	-	-
AutoRecon [54]	-	-	-	-	×	×	-
Part Retrieval and Assembly [55]	×	×	-	-	-	-	-
SparseFusion [56]	-	-	-	-	×	-	-
Behind the Scenes [57]	-	-	×	-	-	-	-
PC^2 [39]	-	-	×	-	-	-	-
RealFusion [58]	-	-	×	-	-	-	-
PermutoSDF [59]	-	-	-	-	×	-	-
NeuDA [60]	-	-	-	-	×	-	-
SECAD-Net [61]	-	-	×	-	-	-	-
ShadowNeuS [62]	-	-	-	-	×	-	-
NKSR [4]	×	-	-	-	-	-	-
City3D [3]	×	-	-	-	-	-	-
POCO [63]	×	-	-	-	-	-	-
HEAT [9]	-	-	×	-	-	-	-
ManhattanSDF [64] [59]	-	-	-	-	×	-	-
TARS [65]	-	-	-	-	×	-	-
Predictable Context Prior [66]	×	-	-	-	-	-	-
Direct Voxel Grid Optimisation [67]	-	-	-	-	×	-	-
Out-of-Core [12]	×	-	×	-	-	-	×
Vis2Mesh [68]	×	-	-	-	-	-	-
PSR [69]	-	-	-	-	×	-	-
GaussianFussion [70]	-	-	-	-	-	-	×
Search and Evaluate [10]	-	-	×	-	-	-	-
RetrievalFuse [71]	×	-	-	-	-	-	-
Learning SDF for Multiview Surface Reconstruction [72]	-	-	-	-	×	-	-
VolT [73]	-	-	-	-	×	-	-
3DIAS [74]	-	-	×	-	-	-	-
MTBR-Net [11]	-	-	×	-	-	-	-

extract point features through PointNet. POCO [63] incorporate their convolutional backbone in order to perform point cloud convolutions, and RetrievalFuse [71] utilises a U-Net for feature extraction.

3.2.2 Single Image Inputs

Owing to their ResNet backbone [9,47,53,57,65,74], they can receive a single image to carry out their respective processes. HEAT [9] utilises this backbone to classify edges within images. TARS [65], through its encoder, learns image features and enables their neural representation approach. MTBR-Net [11], on the other hand, utilises its HR-Net backbone so that it can perform its semantic tasks. Approaches [12] and [10] can also receive satellite or aerial images, where [12] requires an additional dimension in depth.

The rest of the techniques utilising single images incorporate DDMs in their pipelines [39,46,58]. PC^2 [39] takes a single image and projects the image features on their noisy point cloud as DDM conditioning. The second technique from these authors [58] utilises a single image for a neural representation approach and makes up for the lack of multi-view availability through their DDM component. Finally, SDFusion [46] carries a similar approach where they learn latent features through DDMs to synthesise 3D shapes.

3.2.3 Multi-view Image Inputs

Approaches that require multi-view images can be conveniently classified into two categories, neural representations and transformers. Neural representations [48,51,59, 60,62,67,72], are first discussed. As seen, authors of [67] are directly inspired by NeRF and follow their implementation, where this is true for NeuS with Neuralangelo [51], PermutoSDF [59] NeuDA [60] and ShadowNeuS [62]. Authors of [48] are utilising the multi-view images for their photo-consistency network.

The transformer-based approaches [50,54,56,73] are also seen to utilise mutli-view images. VolRecon [50] through its view transformer learns different view projection features. SpraseFusion [56] transformer predicts colours in novel views. AutoRecon [54] combines the features learned from the different views and decomposes them into a point cloud representation. In addition, they can also utilise video inputs. VolT [73] approach is based on refining the multi-view representations and then lifting them to 3D.

3.2.4 Other Input Modalities

In this subsection, techniques incorporating unique modalities are discussed [12, 47, 55, 69, 70]. ShapeClipper [47] leveraging on its CLIP embeddings is the only approach seen that can directly use language to infer 3D reconstructions. X. Xu et al. [55] utilise a library of mesh-based shapes in order to learn the primitive shapes to carry out their reconstructions. Optimisation techniques [12] and [70] can use RGB-D data. Finally, L. Jin et al. [69] utilise a pair of images with their camera module.

3.3 Reconstruction modalities

Whilst our stakeholder is primarily interested in explicit representations, throughout this section, it is shown that off-the-shelf modules can achieve such reconstructions. A summary of reconstruction modalities employed by the literature is shown in Table 3.3. Mesh-based reconstructions are shown in Section 3.3.1. Section 3.3.2 shows point-based reconstructions, including point clouds and voxels. For convenience, implicit reconstruction modalities are grouped together and shown in Section 3.3.3. In Section 3.3.4, approaches that provide image reconstructions are demonstrated.

3.3.1 Mesh Reconstructions

The most commonly used reconstruction modality is meshes. Among the approaches presented, it is particularly common to compute an implicit representation and then extract meshes through marching cubes, such as [4, 12, 47, 49, 50, 52, 54, 56, 58, 63–66]. Out of the ordinary are POCO [63], which directly renders their occupancy scores, and NKSR [4], which utilises dual marching cubes directly onto their kernel functions [80]. Other approaches [48] utilise the TSDF fusion method [79] to extract their meshes.

Depending on the representation, it is seen that there are other approaches for quick mesh generation. X Xu et al. [55] utilise the hole-filling method to cover their generated primitives with watertight meshes [81]. Authors of Vis2Mesh [68] utilise a graph-cut based mesh generation [82] to reconstruct the mesh elements. In 3DIAS [74], they utilise the mesh-fusion [83] technique to generate meshes from their 3D CAD models. In GaussianFussion [70], the authors apply the screen poison surface mesh generation [84] to generate meshes from their dense point clouds. As iterated and differing from the

Table 3.3: Reconstruction modalities used by the literature, where abbreviations used are as follows: Signed Distance Function (SDF), Signed Ray Distance Function (SRDF), Point Clouds (PC), Density Field (DF), Radiance Field (RF).

Author Papers	Reconstruction Modalities								
	Mesh	Voxel	SDF	SRDF	PC	DF	RF	CAD	Image
LION [38]	×	-	-	-	-	-	-	-	-
AutoSDF [45]	-	-	×	-	-	-	-	-	-
SDFusion [46]	-	-	×	-	-	-	-	-	-
ShapeClipper [47]	×	-	×	-	-	-	-	-	-
SRDF [48]	×	-	-	×	-	-	-	-	-
OG-INR [49]	×	-	×	-	-	-	-	-	-
VolRecon [50]	×	-	-	×	×	-	-	-	-
Neuralangelo [51]	-	-	×	-	-	-	-	-	-
ALTO [52]	×	-	×	-	-	-	-	-	-
BUOL [53]	-	×	-	-	-	-	-	-	-
AutoRecon [54]	×	-	×	-	-	-	-	-	-
Part Retrieval and Assembly [55]	×	-	-	-	-	-	-	-	-
SparseFusion [56]	×	-	×	-	-	-	-	-	-
Behind the Scenes [57]	-	-	-	-	-	×	-	-	-
PC^2 [39]	-	-	-	-	×	-	-	-	-
RealFusion [58]	×	-	-	-	-	-	×	-	-
PermutoSDF [59]	-	-	×	-	-	-	-	-	-
NeuDA [60]	-	-	×	-	-	-	-	-	-
SECAD-Net [61]	-	-	-	-	-	-	-	×	-
ShadowNeuS [62]	-	-	×	-	-	-	-	-	-
NKSR [4]	×	-	-	-	-	-	-	-	-
City3D [3]	×	-	-	-	-	-	-	-	-
POCO [63]	×	-	-	-	-	-	-	-	-
HEAT [9]	-	-	-	-	-	-	-	-	×
ManhattanSDF [64]	×	-	×	-	-	-	-	-	-
TARS [65]	×	-	×	-	-	-	-	-	-
Predictable Context Prior [66]	×	-	×	-	-	-	-	-	-
Direct Voxel Grid Optimisation [67]	-	×	-	-	-	-	×	-	-
Out-of-Core [12]	×	-	×	-	-	-	-	-	-
Vis2Mesh [68]	×	-	-	-	-	-	-	-	-
PSR [69]	-	-	-	-	-	-	-	-	×
GaussianFussion [70]	×	-	-	-	×	-	-	-	-
Search and Evaluate [10]	-	-	-	-	-	-	-	-	×
RetrievalFuse [71]	×	-	×	-	-	-	-	-	-
Learning SDF for Multiview Surface Reconstruction [72]	×	-	×	-	-	-	-	-	-
VolT [73]	-	×	-	-	-	-	-	-	-
3DIAS [74]	×	-	-	-	-	-	-	×	-
MTBR-Net [11]	-	-	-	-	-	-	-	×	×

rest, authors of City3D [3] optimise extracted planes and polylines extracted from their TIN heightmap to generate their meshes.

3.3.2 Point-based Reconstructions

Point-based reconstructions can be categorised in voxels and point clouds. Starting with voxelised outputs, both [53] and [73] directly output a voxelised output. In BUOL [53], the 2D to 3D lifting technique is designed to output voxels, whereas in VolT [73], the 3D-volume transformer decoder generates a probabilistic voxel output. C. Sun et al. [67] utilise a voxel grid representation in which modalities of interest, such as density and colour, are stored in the grid cells. In regard to the point cloud-based approaches, [39] and [70] are designed to directly output point clouds. The authors of VolRecon [50] generate point clouds from their computed SRDFs using [85].

3.3.3 Implicit Reconstructions

The main implicit reconstruction technique used is SDFs. Many approaches directly output SDFs through directly using MLPs with SDF loss functions [45–47, 49, 54, 56, 64–66]. Approaches [51, 59, 60, 62] follow NeUS implementation for recovering their SDFs. Authors of [52] and [12] follow different approaches for outputting their SDFs, where [52] authors achieve that by utilising their attention module, and the authors of [12] have designed the approach to operate directly on and output SDFs.

The remaining approaches output other types of implicit representations. Authors of [48] and [50] utilise SRDFs as implicit representations, using the more efficient simulated rays. The NeRF-inspired [58] and [67] output a radiance field as a representation, where [67] makes smart use of the conveniently stored density and colour information in their grid cells. Finally, [57] outputs a density field prediction.

3.3.4 Image Reconstructions

Whilst not 3D, some approaches are seen to output 2D imagery comprising useful information that can be utilised for 3D reconstructions such as [9–11, 69]. HEAT [9] and Search and Evaluate [10] output 2D images with building footprints drawn on them. L. Jin et al. [69] output pseudo-3D renderings of merged planes in their approach. Finally, the authors of MTBR-Net [11] output 2D views of 3D geometries within the original images.

3.4 Scene and Understanding Capabilities

Following our Affinity Diagram, our stakeholder is interested in large-scale reconstructions and semantic understanding. In this section, the literature with such capabilities is discussed in an attempt to identify how these capabilities are unlocked. In Table 3.4 a summary of techniques with semantic or scene reconstruction capabilities is shown.

The literature comprising semantic or scene capabilities is seen in Table 3.4. Approaches [45–47] attempt to capture multi-modal semantic relationships by using language embeddings. AutoSDF [45] combines ResNet with BERT to learn naive conditionals and capture semantic relationships across image and language modalities. The same authors with SDFusion [46] combine CLIP and BERT to achieve the same. The Authors of ShapeClipper [47] have taken advantage of observation of similar 3D shapes having similar CLIP embeddings where they improve global shape understanding by grouping similar images together during training.

Other techniques find their semantic capabilities in their backbones [11,53,64]. Authors of BUOL [53] combine a ResNet50 encoder with three decoders to allow 2D-rich prior learning. Their occupancy-aware lifting block lifts the 2D priors into segmented 3D features. In ManhattanSDF [64], they enhance the scene representation by incorporating semantic logits through DeepLabv3+ [86]. These logits are transformed into probabilities indicating surface types such as floors and walls. The authors of MTBR-Net [11] have found their backbone in HR-Net, which allows them to perform their semantic tasks. In addition, they utilise their roof/facade semantics to segment the footprints of buildings.

Approaches [4,9,10,52,74] all have in common that whilst they perform classification or segmentation tasks, they fail to utilise any semantic capabilities. In ALTO [52], the authors estimate that the rich features extracted from PointNet can also be utilised for semantic tasks but have no proof yet. In NKSR [4], they only classify voxels depending on their contribution to the scene. Approaches [9] and [10] utilise edge/corner detection where [9] does so geometrically, and [10] relies on a footprint ground truth to learn generic geometries of buildings.

In Table 3.4 it can also be seen that only three approaches combine scene reconstructions with semantic capabilities [11,53,64]. Approaches [53] and [64] scene understanding capabilities have been tested and proven in indoor scenes only. Contrastingly, [11] addresses large-scale outdoor reconstructions.

Table 3.4: Semantic and Scene Capabilities as seen by the literature

Author Papers	Semantic Capabilities	Scene Capabilities
MTBR-Net [11]	Yes	Yes
BUOL [53]	Yes	Yes
ManhattanSDF [64]	Yes	Yes
AutoSDF [45]	Yes	No
SDFusion [46]	Yes	No
ShapeClipper [47]	Yes	No
AutoRecon [54]	Yes	No
Neuralangelo [51]	No	Yes
ALTO [52]	No	Yes
Behind the Scenes [57]	No	Yes
NeuDA [60]	No	Yes
ShadowNeuS [62]	No	Yes
NKSR [4]	No	Yes
City3D [3]	No	Yes
POCO [63]	No	Yes
HEAT [9]	No	Yes
Predictable Context [66]	No	Yes
Direct Voxel Grid Optimisation [67]	No	Yes
Vis2Mesh [68]	No	Yes
PSR [69]	No	Yes
Search and Evaluate [10]	No	Yes
RetrievalFuse [71]	No	Yes
Learning SDF for Multiview Surface Reconstruction [72]	No	Yes

In addition, variations in the scale of reconstructions are evident. Approaches [45–47,54] whilst having semantic capabilities in that they can understand object classes, they are limited to single object reconstructions. On the other hand, the rest of the approaches do not possess any semantic capabilities but are able to perform scene reconstructions. More specifically, [52, 60, 62, 63, 66, 67, 69, 72] are only addressing small scale scenes, whereas [3, 4, 9, 10, 68] are able to tackle large scale scenes.

3.5 Human-in-the-loop

Guided by our Affinity Diagram, we are looking for methods in which the state-of-the-art literature incorporates people within their approaches. It is evident that the literature

attempting to incorporate people within their pipelines is very limited. Approaches [38, 45–47] use language text as their method of interactivity. All aforementioned approaches incorporate language-only guided 3D generations, where [38] and [47] are able to texturise their shapes via Text2Mesh [87]. In addition, [38] is able to provide interpolations between shapes, whereas SDFusion [46] is also able to perform text-guided shape completion.

Two other methods of interactivity are seen. Neuralangelo [51], utilising commercial software, can load the source point cloud and select the region of interest to be reconstructed. Authors of [49] have taken a more algorithmic approach where a user can directly change octree labels to guide or influence the reconstruction.

3.6 Critical Analysis

Through the analysis provided, it is seen that the most popular input modalities are multi-view images with thirteen entries. As discussed, most methods operating on multi-view images are neural representations and transformer-based techniques. It has been shown that neural representations deal with multi-view images by learning a continuous field, whereas transformers can learn multi-view features from these images.

Single images and point cloud inputs closely follow with twelve entries each. Approaches [9,11,69] can receive a single large-scale image and output building and footprint information. Regarding point clouds, many approaches can directly receive and operate on this modality. It was shown that single images are commonly used with techniques that rely on their convolutional backbones to process them.

Meshes are the most commonly used reconstruction modality, followed by SDFs. Most authors utilise off-the-shelf approaches to convert their modality of preference to meshes. Only two approaches can directly output meshes. However, even these approaches utilise off-the-shelf techniques to compute these meshes. Most approaches using implicit representations as reconstruction modalities use marching cubes or TSDF-fusion to extract meshes. For the techniques that utilise implicit representations but do not provide meshes, it can be understood that whilst they can, the authors chose not to. Finally, it is noteworthy that most approaches that incorporate marching cubes or other mesh extraction techniques utilise explicit metrics such as Chamfer Distance or Intersection over Union to assess their reconstruction qualities.

A pattern is identified regarding semantic and scene reconstruction capabilities, where very few approaches can incorporate both. It can be quickly inferred that most neural

representations that can output scene reconstructions do not incorporate semantic capabilities. As discussed, most techniques that can incorporate semantic capabilities address single-object reconstructions. Only three techniques incorporate semantic capabilities with scene reconstruction, and only one is able to process large-scale outdoor scenes [11]. It should be noted, however, that their reconstructions are simple, lack details and only can identify buildings.

Finally, the clear lack of Human-in-the-loop incorporation is evident. The approaches that were demonstrated to incorporate language-based interaction cater to the Arts Industry, and the shapes they produce do not reflect a physical shape captured from sensors. One approach has shown region selection, which can be useful for addressing scalability issues. Finally, one approach gives the ability to the user to influence the reconstruction outcome by directly intervening in their octree-building process.

Chapter 4

Materials and Method

Throughout this Chapter, the activities undertaken towards the development of this dissertation are documented. These activities include stakeholder engagement activities in the form of meetings and contextual inquiry, the creation of an Affinity Diagram based on data gathered throughout the stakeholder engagement activities and a non-comprehensive survey in 3D object reconstruction. In the remainder of this Chapter, Section 4.1 describes the flow of events, and Section 4.2 describes the stakeholder engagement activities that took place. In Section 4.3 the Affinity Diagram methodology is described. Finally, in Section 4.4, we provide reasoning for our literature collection criteria and for the themes guiding our survey analysis.

4.1 Overall Flow

The general flow of the project is described in Figure 4.1. As soon as the project began, the literature search and analysis for the survey presented was initiated. Afterwards, three stakeholder engagement activities were conducted between the academic and industrial teams. Notes taken throughout these engagement activities were used to create an initial set of labels that enabled the creation of an Affinity Diagram. The Affinity Diagram has allowed for identifying key stakeholder requirements along with other areas of high importance, termed Themes. The state-of-the-art literature collected was subsequently re-analysed to align with the insights gained from the Affinity Diagram.

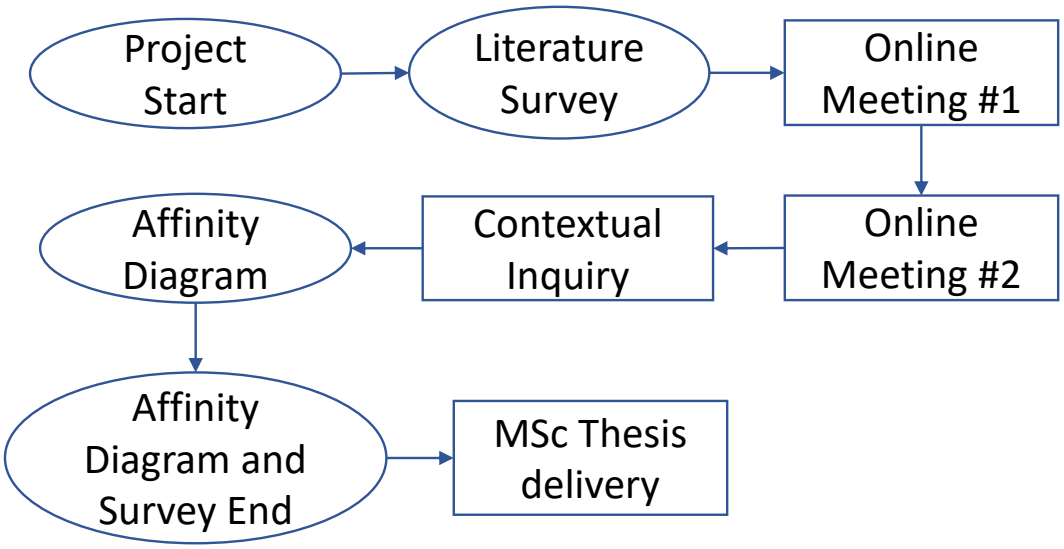


Figure 4.1: Project Overall Flow

4.2 Stakeholder Engagement

The stakeholder engagement activities have taken place in two phases. The first phase includes two online meetings between the industrial and academic teams. The second phase was in the form of contextual inquiry, where the author of this dissertation travelled to our stakeholder's headquarters to first-hand experience current practices and delve deeper into the initial requirements given. In Table 4.1, simple demographic information of the stakeholder participants is shown.

Table 4.1: Stakeholder Participants Demographic Information

Participant	Role	Years of Experience
ST1	Chief Geospatial Scientist	8
ST2	Senior Innovation and Research Scientist	N/A ~10+
ST3	Research Software Engineer Senior Geospatial Scientist	7 3

As iterated, in the first phase, our stakeholder engagement activities took place in the form of online meetings. In both meetings, a short presentation was conducted to set

the scene for our stakeholders and provide useful information on preliminary findings. This has allowed to maximise the efficiency of these meetings and ensure that we receive appropriate guidance and feedback. Extreme care was taken regarding the delivery of said presentations to not bias the participants. For the second meeting, key changes were made to the presentation demonstrated to accommodate for new information extracted from the previous meeting. The slides of both presentations can be seen in Appendix A.

For the second phase, the author of this dissertation has travelled to our stakeholder's headquarters. This contextual inquiry trip has served the purpose of directly observing the organisation's current practices regarding 3D products. In addition, it has aided in gaining further context regarding how this product will be utilised, by whom and the potential capabilities desired. Said capabilities ensure that the pipeline delivered by upcoming research complies with our stakeholder's vision and maximises our contributions towards the said vision.

4.3 Affinity Diagram

The KJ method is closely followed where applicable to create an Affinity Diagram. Raw notes, taken throughout the stakeholder engagement activities, have been first transcribed into labelled statements. Each transcribed label represents a simplified version covering the originating note's essence. Only notes relevant to the scope of this research have been transcribed, and repeating labels have been eliminated.

After several iterations of label grouping and adjusting, individual groups start to form. Preliminary group titles are given to the groups formed when no apparent changes can further be made to help contextualise the patterns starting to appear. Following the nonlinearity of the process, the groups are continuously adjusted and decimated when new grouping patterns are evident. Due to the iterative procedure of the process, the whole grouping process cannot be described in detail, and many iterations are lost. Core steps of the iterations towards the creation of the Affinity Diagram are described in Chapter 5, along with grouping pattern analysis.

4.4 Literature Survey

Based on information collected before the initiation of this dissertation, the literature collection criterion was to incorporate capabilities of 3D object or scene reconstructions.

This criterion ensures that a broad scene of the latest advancements in 3D reconstruction is captured without restricting the search to building or urban-specific reconstructions. Such a restriction would prohibit the search and analysis from having multi-disciplinary elements and hence prohibit a wider understanding of 3D product usages and limitations of the techniques analysed towards our objectives.

Following the stakeholder engagement activities described and the creation of our Affinity Diagram, a re-analysis of the literature ensued based on the Themes discovered. This approach has allowed for a more focused interpretation of the literature assessed and greatly enhanced the understanding gained from the state-of-the-art with respect to our stakeholder's needs.

Chapter 5

Results

In this Chapter, the findings derived from our Affinity Diagram are discussed. In Section 5.1, the construction and analysis of the Affinity Diagram are shown. In Section 5.2, the literature assessed is mapped onto the Affinity Diagram where appropriate, aiming to identify gaps with respect to our stakeholder's needs.

5.1 Affinity Diagram

5.1.1 Construction of Affinity Diagram

Whilst the identified themes, subthemes and categories, and their relationships with individual labels, are analysed in detail in this section, it is imperative to demonstrate the defining steps undertaken to produce the Affinity Diagram. All steps to create the Affinity Diagram were digitalised except the first step - label making, where the notes representing relevant statements were extracted offline.

The first grouping iteration, i.e., the first iteration of the second step within the KJ method, is shown in Figure 5.1. It is evident that some of the core themes that will lead our survey are already formed. Group (a) indicates our stakeholders need to incorporate semantic capabilities within the pipeline to be developed. Labels in group (b) are related to how our stakeholders intend to incorporate human-centredness in this pipeline. Group (c) refers to the need to identify how the Industry of 3D products operates and delivers said products on a large scale. The fourth group identified, group (d), relates to guidance and instructions on how the 3D products should be reconstructed. The final group, group (e), relates to data availability and related challenges.

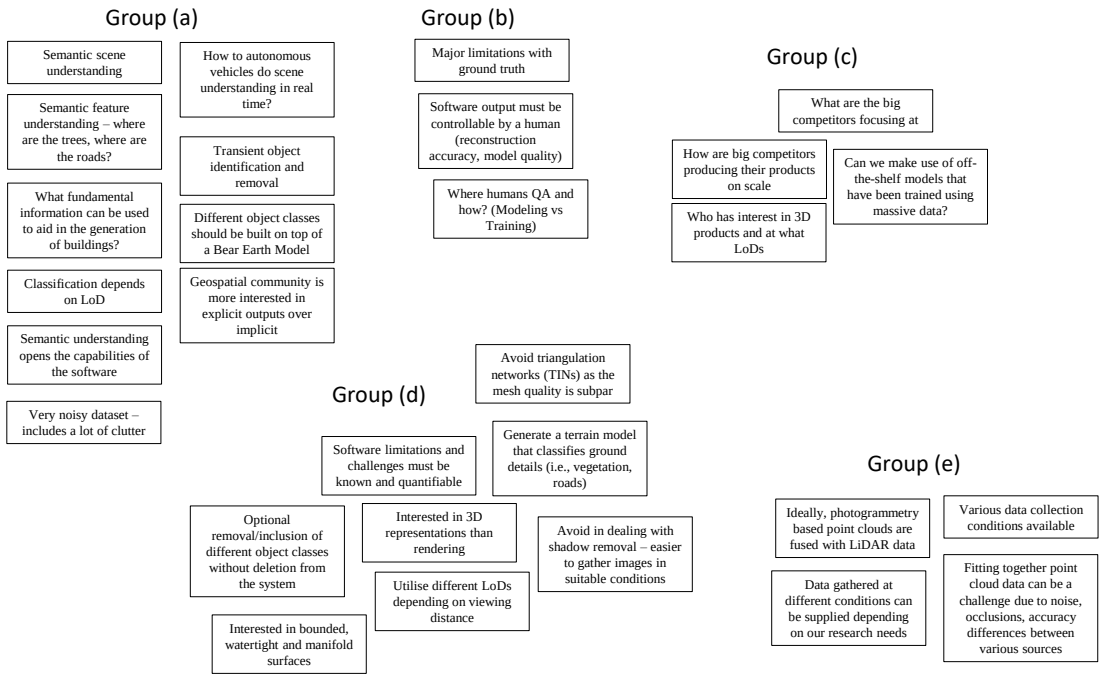


Figure 5.1: First iteration of label grouping process towards the creation of the Affinity Diagram

Following the nonlinearity of the process, labels with extreme similarities are further eliminated, label positioning is adjusted after careful consideration, and an attempt is made to decimate the groups further to uncover relationships hidden in our dataset. In Figure 5.2, the final iteration of the group-making process is seen, where preliminary group titles are given to aid the researcher in contextualising the information when moving to the chart-making step. Evidently, the preliminary groups identified earlier in the process are further solidified. Other than label adjustments, it is seen that group (d) from Figure 5.1 has been reduced into two groups. One group relates to our stakeholder's requirements regarding how the pipeline will be developed and implemented in their current practices. The second group arising from this reduction explicitly relates to the guidance and requirements of the 3D reconstructions themselves.

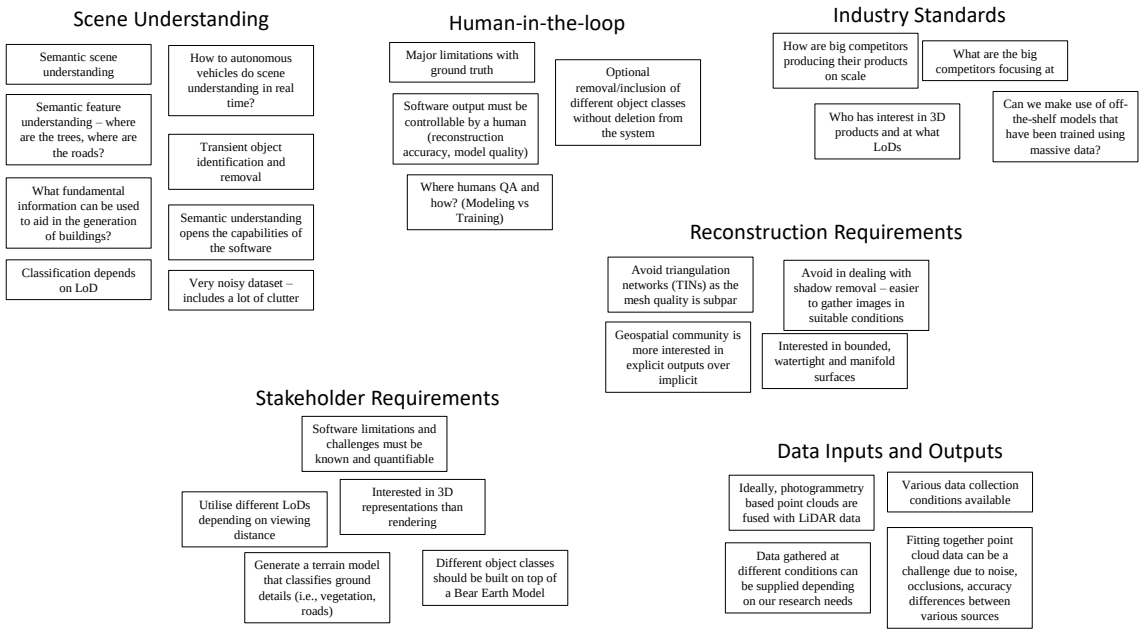


Figure 5.2: Final iteration of label grouping process towards the creation of the Affinity Diagram

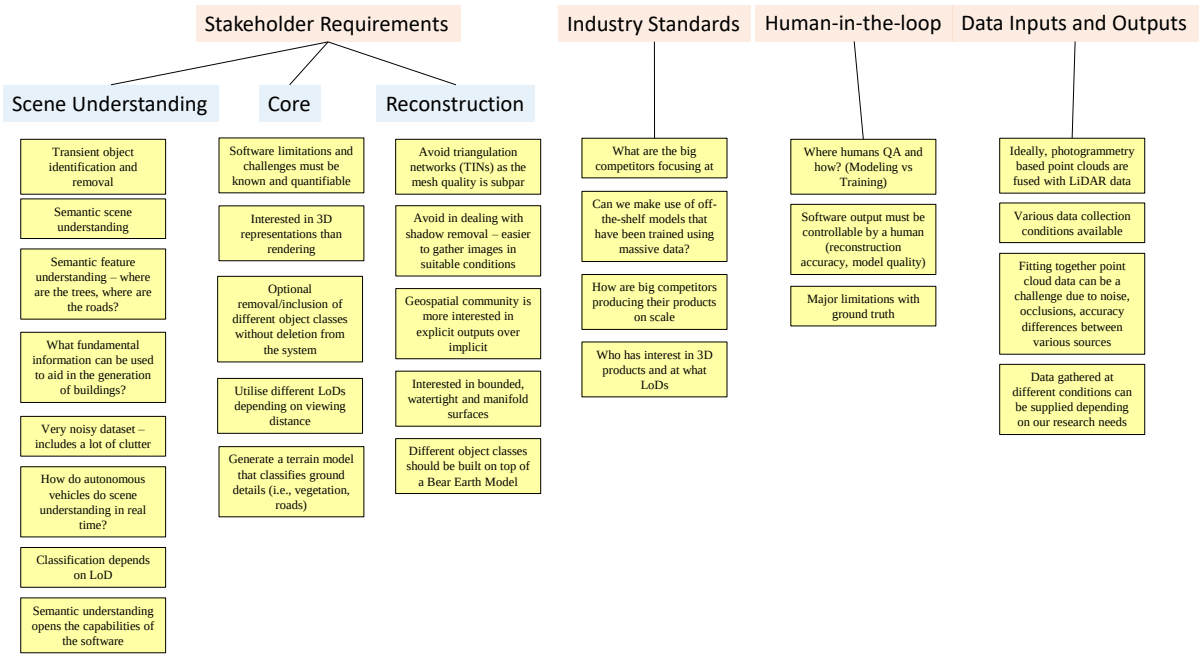


Figure 5.3: First iteration of the chart-making process towards the creation of the Affinity Diagram

Figure 5.3 shows a step in the chart-making process. In this preliminary stage, the main themes and sub-themes that are apparent in the finalised Affinity Diagram are already visible. Without accounting for individual label adjustments, the previously identified Stakeholder Requirements group has taken a larger role in the Affinity Diagram and has been identified as a theme comprised of three requirement-related sub-themes. Namely, the sub-themes are Scene Understanding, Core, previously termed Stakeholder Requirements, and Reconstruction. The remaining themes identified in the chart-making process are the Industry Standards, Human-in-the-loop and Data Inputs and Outputs themes.

The fourth and final step of the KJ method is to provide a written explanation of the resulting Affinity Diagram. In Figure 5.4, the finalised Affinity Diagram is seen, where the themes previously identified remain unchanged. The themes and sub-themes were further reduced wherever possible to make the so-called categories, and the cause and effects are also shown. The hierarchy of the Affinity Diagram is dominated by the themes, followed by sub-themes. Categories are third in the hierarchy and can be incorporated directly into themes or sub-themes. Finally, the remaining labels can be placed under any hierarchical rank.

The first theme explored is the Stakeholders Requirements theme, which is the only theme comprising sub-themes. The sub-themes are the Scene Understanding, Core and Reconstruction requirements. The Scene-Understanding sub-theme comprises the Semantic Meaning and the Applications category. As indicated by the arrow, the Applications category is enabled by the Semantic Meaning category, which comprises labels relevant to the semantic scene and feature understanding requirements. This category also includes labels on how autonomous vehicles can perform scene understanding in real time and what fundamental information can be used when reconstructing 3D models of buildings. The Applications category refers to applications the stakeholder intends to utilise through the semantic semantic capabilities. Such applications include dynamic classifications of objects based on levels of detail (LoDs), dealing with cluttered and noisy datasets, and transient object identification and removal.

The second sub-theme under the Stakeholders Requirements theme is termed Core requirements. This sub-theme includes three autonomous labels and the Functionalities category. Said functionalities include incorporating varying levels of detail depending on viewing distance, optional removal and re-addition of object classes without deletion from the overall system and having a detailed terrain model. The first autonomous label

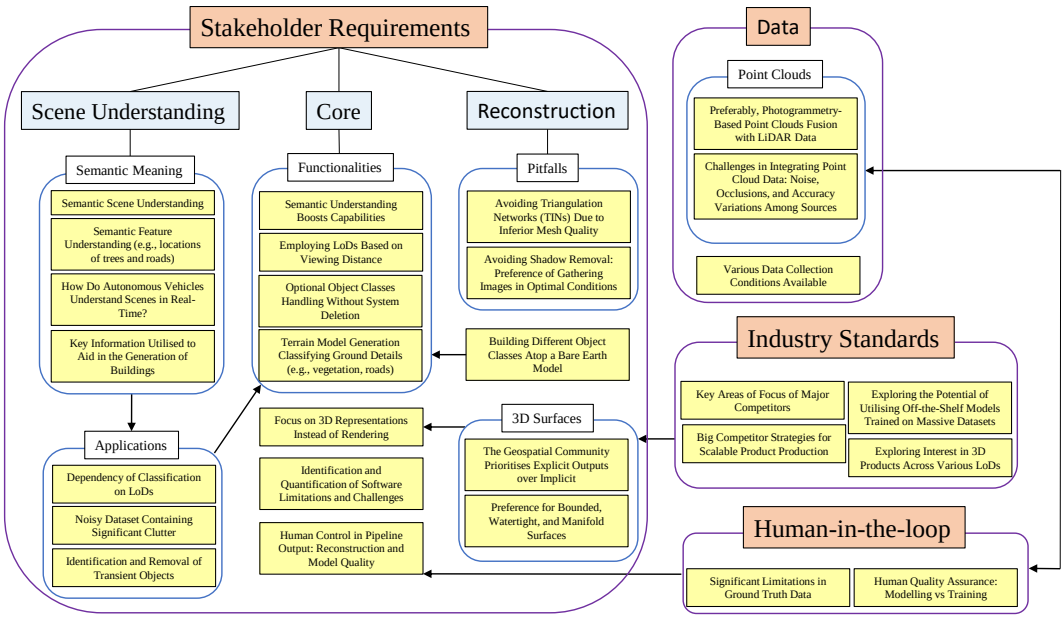


Figure 5.4: Finalised Affinity Diagram

conveys the stakeholder's interest in explicit representations over renderings. The second autonomous core requirement is that the pipeline to be developed must have known and measured limitations. The final autonomous label conveys the need to employ people within the pipeline for quality assurance.

The third and final sub-theme related to our stakeholder's requirements is termed Reconstruction. The lone label in this sub-theme conveys the requirement of building different object classes, such as buildings or vegetation, on top of a Bear Earth Model, which is essentially a terrain model. The Pitfalls category comprises areas the stakeholder wants to avoid delving into. The first identified pitfall is reconstructions based on Triangular Irregular Networks (TINs) due to their subpar quality. The second pitfall the stakeholder identified is that dealing with shadow detection and removal is very laborious and requires a lot of effort. The second category in this sub-theme is termed 3D Surfaces. This is related to the Geospatial community's preference for explicit surfaces over implicit and that the output must be watertight, manifold and bounded.

The second theme analysed is the Data theme, which is concerned with data modalities the stakeholder can provide to enable future research. A category termed Point Cloud has been identified, comprising two labels. The first label relates to our stakeholder's preference for fusing photogrammetry-based point clouds with data derived from ground-based LiDAR devices. The second label in this category relates to problems that arise when dealing with these types of data. Such challenges include variations in noise, occlusions and accuracies between the different point cloud data sources. The autonomous label in this theme suggests a flexible approach from our stakeholders regarding providing data in other formats, such as images and different environmental conditions.

The final two themes are Industry Standards and Human-in-the-loop. There are no sub-themes or categories within these themes. The labels the Industry Standards theme is comprised of essentially convey the need to research big competitor organisations that have products in GIS 3D applications. The stakeholder has expressed the need to identify the areas said competitors are focusing on. In addition, the strategies they undertake to deliver scalable 3D products must be identified. The final autonomous label located in this theme is the requirement to investigate if and how said competitors utilise 3D products across different levels of detail. The Human-in-the-loop theme refers to the unavailability of ground truth. Hence, there is a need to identify where people can engage within the pipeline developed to accommodate for the lack of ground truth.

5.1.2 Affinity Diagram Analysis

Under the Scene Understanding theme, the Semantic Meaning sub-theme clearly indicates the need of the stakeholder to leverage semantic relationships to maximise the capabilities of the reconstruction software. In addition, there is a need for detailed classifications to identify city furniture, city objects and other classes and differentiate them from other classes perceived as noise or clutter, such as transient objects. Furthermore, as the stakeholder is interested in a scalable and fast pipeline, there is a need to research how autonomous vehicles execute scene understanding in real time. The final statement in this sub-theme expresses the need to identify useful information that can be used to aid in the generation of 3D buildings. Essentially, this statement is related to the types of semantic or geometric information that can be leveraged to aid the generation of 3D buildings. The Applications category in this sub-theme conveys how this semantic information can be

employed within future research. These use cases include different class reconstructions depending on the level of detail and clutter/transient object identification and removal.

The second sub-theme analysed is the Reconstruction sub-theme. As already mentioned, the Pitfalls category includes areas to be avoided as instructed by the stakeholder. TINs are avoided as they only provide a single height parameter per square distance, which can greatly limit the quality of the reconstructions. The second category of this sub-theme, the 3D Surfaces, states requirements regarding the stakeholder's preference of employing explicit modalities for the reconstructions, where the main modality of interest is meshes, which must be watertight, manifold and fully bounded. It is shown that the 3D Surfaces category is directly related to the Core sub-theme in that explicit outputs essentially translate into avoidance of rendering. In addition, the surfaces generated must be able to adjust their intersection vertices to be bounded on top of a Bear Earth Model. The autonomous label in this sub-theme is also related to the Core functionalities in that the surfaces built must be controllable by class selection.

Almost all of the themes and sub-themes identified point towards the Core requirements sub-theme. The Functionalities category conveys additional functionalities to reconstruction that the stakeholder wishes to incorporate within the future pipeline. A relation is noticed between the Functionalities and Applications categories. It is established that the user must be able to interact with different levels of detail depending on the distance viewed, where different classes are shown depending on said viewing distance. In addition, the fact that the dataset is noisy and contains significant amounts of clutter might indicate that this noise must be classified and distinguished from clutter. This conclusion is also supported by the lone label under the Reconstruction sub-theme, where some classes, like building installations, whilst considered clutter, might prove to be useful information for many industries. Such a strategy would need to incorporate a Bear Earth Model as a base surface, where different verified classes must be able to be optionally built on top of it. Combined with the insights gained from the Scene Understanding sub-theme, it is evident how the semantic capabilities enable multi-class reconstructions. Finally, the autonomous label in the core sub-theme regarding the stakeholder's preference for 3D representations is seen to convey the same message with 3D Surfaces.

The Data theme, as described, is largely related to the point cloud inputs from different sources and the challenges accompanying this type of data. The relationship between this theme and the Human-in-the-loop theme has been identified as the unavailability

of ground truth. Photogrammetry-based point clouds contain artefacts due to shadows, lighting conditions or lack of redundancy between source images. In comparison, the LiDAR data contain different inaccuracies relating to GIS loss and moving objects. Hence, an expert must be incorporated to ensure that the model quality remains high, possibly through interactive guidance. The autonomous label relates to the collection of new data under different conditions, which further establishes that shadow removal would indeed be a pitfall.

Finally, the Industry Standards theme is associated with the identification of both state-of-the-art and commercial 3D Geospatial products, where the specifics have been discussed in Section 5.1.1. Not shown on the Affinity Diagram due to spatial restrictions is that this theme can have a large effect on the 3D surface reconstruction strategy to be undertaken by future work.

5.2 Survey Mapping on Affinity Diagram

We combine the findings from the Affinity Diagram and the survey. To do so, where applicable, the relevant literature mentioned in the survey presented is mapped onto the Affinity Diagram, as seen in Figure Fig. 5.5. Whilst through the survey, the trends and gaps of the state-of-the-art in regards to our themes have been found and analysed, this mapping allows for a much more holistic analysis.

The Scene Understanding sub-theme under Stakeholder Requirements is first considered. In the Semantic Meaning category, only three techniques have been identified as being capable of scene reconstruction with semantic understanding, of which two are capable of recognising multi-class features. As discussed, four authors were identified to utilise building information for their reconstructions. However, as shown, the resulting qualities are inferior to the reconstructions we are aiming for. No techniques showcased semantic capabilities in conjunction with 3D reconstructions in real-time.

In the Applications category, two major gaps are identified. No techniques focus on transient object detection and removal. Whilst some techniques have the capability of handling large amounts of noise and sparseness, no approaches tackle clutter specifically. The second major gap is the lack of employing different classifications for different levels of detail.

With respect to the Core sub-theme, major gaps in the literature are again identified. In the Functionalities category, it is seen that only one approach has shown capabilities of providing different levels of detail. However, the detail of their reconstructions

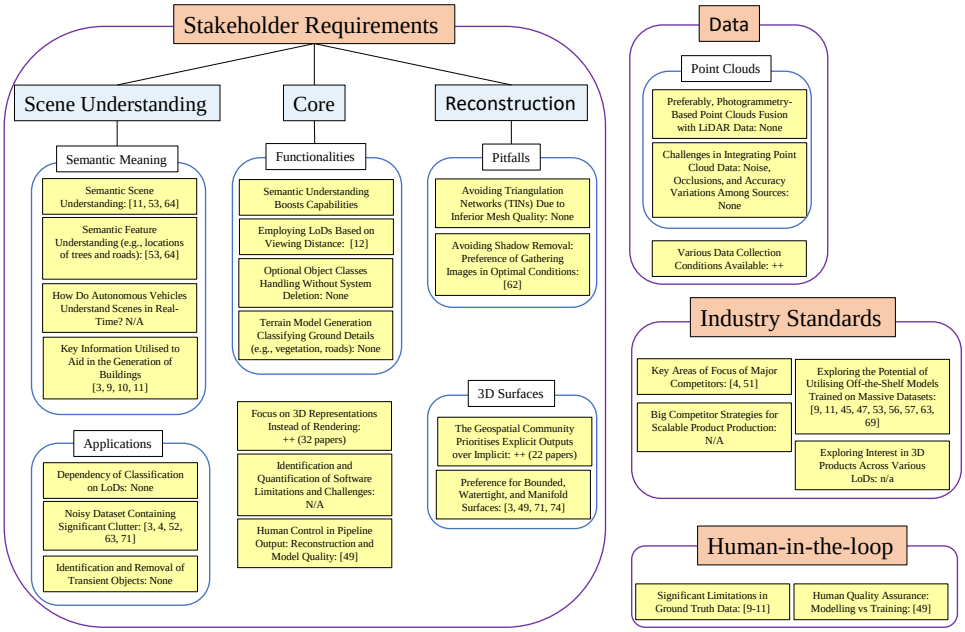


Figure 5.5: Literature mapped to the relevant labels, N/A means that the statement is irrelevant to our search, None indicates that no literature addresses that label, showing gaps between our stakeholder needs and the literature surveyed, ++ indicates that there is vast literature addressing that label.

solely depends on their data, and their approaches are not used in conjunction. The lack of semantic capable reconstructions is also reflected in the requirement of optional class handling. In addition, no techniques were found which were concerned with the reconstruction of a terrain model. In regard to our stakeholder's requirement of focusing on 3D representations, it is seen that most approaches provide quantifiable explicit methods of reconstruction in their methods. A great limitation is found in that most approaches do not incorporate any human intervention in their techniques, where only one author was seen to accomplish this.

In the Reconstruction sub-theme, we have found that the Pitfalls instructed by our stakeholder are indeed avoided by the recent literature. Where no authors are utilising TINs as their final reconstruction or focusing on shadow removal. Most interestingly, an approach was found that leveraged the presence of shadows to aid their reconstructions. In regard to the 3D Surfaces, it was shown that the most common reconstruction modality is meshes. Moreover, it was shown that meshes are almost always generated from an

off-the-shelf technique, such as marching cubes and variants. In addition, whilst only four approaches directly report watertight, bounded or manifold surfaces, it can be inferred that this is the outcome for most.

Moving to the data theme, we identify that no approach utilises multi-sourced point clouds. Thus, no methods for integration of multi-sourced point clouds were identified either. However, our stakeholder is able to provide data under different conditions in both point cloud and image formats, which are the main input modalities used by the literature.

It is obvious that the survey presented is not appropriate for appropriate exploration of the Industry Standards theme. The funding of different approaches was investigated, but when mentioned, it is either academic or governmental. Exceptions are two approaches funded by NVIDIA. Regarding off-the-shelf models, several techniques are utilising pre-trained and well-known backbones to enable their approaches.

Finally, we have found an important gap in terms of the Human-in-the-loop theme. Only one approach was found to use interactivity to supervise their reconstruction quality. However, three approaches were seen to identify building geometries and footprints, which is a promising approach to overcome the limitations with ground truth data.

Chapter 6

Discussion

The methodology of the Affinity Diagram based qualitative analysis approach was demonstrated through the iterative process shown in Figures 5.1 and 5.2. Accounting for the group reductions that were performed, the bottom-up methodology of the Affinity Diagram was not strictly adhered to. However, this was expected due to the non-rational and nonlinear elements that constitute the Affinity Diagrams. Through this Affinity Diagram, key areas of analysis for the literature survey are presented. Following this analysis, the literature analysed was visualised on our Affinity Diagram in order to get a more holistic picture of how recent literature compares to our stakeholder's needs.

With respect to the Scene Understanding sub-theme, a critical limitation is identified in that very few approaches incorporate both scene reconstruction and semantic capabilities. Out of these three approaches that are able to carry out these tasks simultaneously, two incorporate multi-class capabilities but are restricted to small-scale indoor scenes. The other approach is able to output large-scale outdoor scenes, but they can only use semantic tasks to identify and reconstruct buildings. In addition to the building-related semantic tasks, other approaches were seen to utilise building-specific geometric information to identify and reconstruct the footprints of buildings. It can be inferred that the state-of-the-art cannot yet tackle the considerably increased complexity of identifying and reconstructing multi-class objects on such a large scale.

In the Applications category, gaps are identified in the lack of scene classification depending on levels of detail and in the identification and removal of transient objects, where no techniques are seen to address either task. In addition, whilst some approaches

mention that they can address noisy and sparse datasets, no approach addresses clutter explicitly, as it is a more complex problem.

The limitations already discussed are also reflected in the Core sub-theme. One approach was seen to be able to provide different LoDs, however, as already explained, their reconstructions are not synchronised. With the identified limited semantic capabilities within reconstruction approaches, it was found that optional class handling is also a big gap. In addition, it can be understood that city semantic classification and class handling is a large field with a different focus than 3D reconstruction. Literature specialising in that task promises to show much more capable semantic understanding. Furthermore, the lack of terrain model generation might also indicate the need for a specialised technique.

Different results are seen with the Reconstruction sub-subtheme. The pitfalls our stakeholder instructed us to avoid have not been seen in any of the literature. With respect to the nature of the reconstructions, it is shown that most approaches employing implicit representations are also capable of extracting explicit modalities. In addition, whilst only four approaches explicitly report watertight, manifold and bounded surfaces, it is understood that marching cubes and related techniques are also capable of producing such surfaces. Whilst our stakeholder has shown a disinterest in implicit methods, it is seen that they dominate the state-of-the-art. By utilising the off-the-shelf meshing techniques, many authors extract meshes which are quantifiable and measurable modalities.

In the Data theme, it is identified that the point fusion of LiDAR data is not something preferred by the 3D reconstruction community as no approaches utilise multi-modal data inputs. Thus, the challenges relating to the integration of multi-sourced point clouds have not been addressed either. That being said, it is seen that healthy amounts of literature utilise both point cloud and image data inputs. In fact, neural representations, which are a very popular 3D reconstruction approach, are seen to prefer multi-view image inputs. It would be particularly interesting to investigate the combination of image and point cloud modalities fusion.

It is identified that the Industry Standards theme is not really relevant to the survey presented, as we are focusing on academic literature. Approaches referencing funded research often cite academic institutions or government sources. Exceptions are [4, 51], which is work funded by NVIDIA. In this theme, it is also identified that there are multiple techniques relying on backbones to guide their reconstructions. As already

stated, there is a need to explore how to utilise a semantic classifier and potentially be the backbone of the reconstruction pipeline.

It has been demonstrated in Chapter 3 that human-in-the-loop incorporation is very rare in the literature. The approaches shown incorporating text guidance did not involve any interactivity in the sense that the user can guide the software to deliver more accurate reconstructions. Only one approach was seen to be able to better their reconstructions through human involvement, which is further proof that the state-of-the-art concentrates on technical matters and neglects human-in-the-loop initiatives. However, the approach is limited as they are directly augmenting octree structures. Regarding ground truth limitations, some approaches were seen to be able to provide footprints of printings. Whilst not involving people, they give the possibility of alleviating problems with ground truth for guiding the building reconstructions.

Chapter 7

Limitations and Future Work

As expected, the themes identified and used to guide the undertaken analysis are multi-faceted and cover a broad range of fields. To this extent, it is difficult to cover in detail all these aspects within the short period of time allocated for this research project. Furthermore, the literature on this topic is vast, and it is impossible to cover it extensively. That said, the literature analysed can be considered a good representation of the state-of-the-art and will shape future research. In addition, the literature survey will be repeated in a more comprehensive manner and with the scope modified to address large-scale urban scenes.

Owing to the iterative procedure of the categorisation of the Affinity Diagram, some labels may not accurately represent the statements that our stakeholders intended. Whilst non-rationality is in the nature of the Affinity Diagrams, there is a need to verify the findings of this dissertation with our stakeholders to ensure that the objective of future research is correctly characterised. In addition, due to the rules of this dissertation, the Affinity Diagram was carried out only by the author. For future work, this process is to be repeated with the contribution of other researchers to effectively carry out the method. Finally, more qualitative data are to be gathered by our stakeholder.

A significant limitation was also found in technical implementation. Computer graphics is a very hardware-demanding area, and state-of-the-art techniques can be very computationally expensive and extremely difficult to use. Whilst an attempt has been made to implement some of the techniques analysed to assess their practical utilities and performance, the lack of appropriate hardware has greatly obstructed this goal. This limitation is expected to be alleviated early in future research. A standardised

workflow will be developed using Docker, which will enable the execution and evaluation of multiple techniques with consistency.

In addition to 3D reconstructions, two major areas for future research were identified: semantic scene understanding and Human-in-the-loop. In our upcoming research, we plan to focus on identifying how to enable semantic scene understanding. The literature surveyed has shown very limited capabilities in comparison with what our stakeholder is aiming for. It has been recognised that a suitable semantic backbone must be designed to support our reconstruction pipeline. Whilst many approaches are seen to utilise strong convolutional backbones, we believe that a task-specific semantic backbone can robustly support the reconstruction pipeline.

In terms of 3D reconstruction, neural representations are seen to be very popular among the literature analysed. Several approaches are able to achieve high-quality and detailed reconstructions. Many techniques not directly incorporating neural representations are seen to also utilise implicit representations. For our research, it would be interesting to attempt to learn multi-view features from the point clouds in order to compute neural representations to extract meshes. A possible architecture to accomplish this might include a generative model supported by the semantic backbone, which learns novel denoised projections of the original points. Approaches have shown that points cloud projections can be represented as occupancy fields and hence extract meshes through marching cubes-based methods.

Finally, it was shown that a large gap in Human-in-the-loop approaches is evident. Similar to the semantic understanding conclusions, this area needs to be investigated in depth to identify how people could be integrated effectively to enhance the pipeline. The combined gaps in semantic and human-in-the-loop capabilities might suggest that an approach that combines both could lead to achieving multi-class reconstructions, which are the ultimate target of our stakeholder.

Chapter 8

Conclusions

The transformation of two-dimensional maps to 3D offers to support many upcoming technologies such as smart cities, digital twins and autonomous vehicles. However, data derived from real sensors are unavoidably noisy and susceptible to the environmental conditions in which these data have been collected. The 3D reconstruction literature is seen to focus on producing faithful reconstructions, but these techniques are seen to be sensitive to thin structures, noise and shadows.

Qualitative data derived from stakeholder engagement activities have been organised using affinity diagramming to identify our stakeholder's needs. The needs identified formed the themes that guided our literature survey. The themes identified relate to our stakeholder's technical requirements, data availability and obstacles that are associated with these modalities, industry standards of 3D map reconstruction and human-in-the-loop initiative. Finally, to solidify the strengths and gaps identified, the literature analysed is mapped onto the Affinity Diagram.

Three main areas for future contribution have been identified. The 3D reconstructions themselves, semantic understanding and scene reconstruction capabilities, and human-in-the-loop.

Bibliography

[1] [Online]. Available: https://www.ordnancesurvey.co.uk/about/history

[2] L. Nan and P. Wonka, "Polyfit: Polygonal surface reconstruction from point clouds," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2353–2361.

[3] J. Huang, J. Stoter, R. Peters, and L. Nan, "City3d: Large-scale building reconstruction from airborne lidar point clouds," *Remote Sensing*, vol. 14, no. 9, p. 2254, 2022.

[4] J. Huang, Z. Gojcic, M. Atzmon, O. Litany, S. Fidler, and F. Williams, "Neural kernel surface reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4369–4379.

[5] R. Scupin, "The kj method: A technique for analyzing data derived from japanese ethnology," *Human organization*, vol. 56, no. 2, pp. 233–237, 1997.

[6] A. Khatamian and H. R. Arabnia, "Survey on 3d surface reconstruction." *Journal of Information Processing Systems*, vol. 12, no. 3, 2016.

[7] B. K. Horn, "Shape from shading: A method for obtaining the shape of a smooth opaque object from one view," 1970.

[8] L. We, "Marching cubes: A high resolution 3d surface construction algorithm," *Comput Graph*, vol. 21, pp. 163–169, 1987.

[9] J. Chen, Y. Qian, and Y. Furukawa, "Heat: Holistic edge attention transformer for structured reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3866–3875.

[10] F. Zhang, X. Xu, N. Nauata, and Y. Furukawa, "Structured outdoor architecture reconstruction by exploration and classification," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 12 427–12 435.

[11] W. Li, L. Meng, J. Wang, C. He, G.-S. Xia, and D. Lin, "3d building reconstruction from monocular remote sensing images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 548–12 557.

[12] N. Poliarnyi, "Out-of-core surface reconstruction via global tgv minimization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5641–5650.

[13] [Online]. Available: https://earthengine.google.com/

[14] "Building footprints," Feb 2020. [Online]. Available: https://www.microsoft.com/en-us/maps/bing-maps/building-footprints

[15] [Online]. Available: https://www.qgis.org/en/site/

[16] [Online]. Available: https://www.esri.com/en-us/arcgis/products/arcgis-online/overview

[17] E. Grilli, F. Menna, and F. Remondino, "A review of point clouds segmentation and classification algorithms," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 42, pp. 339–344, 2017.

[18] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[19] K. K. Y. Wong, "Towards a national 3d mapping product for great britain," Ph.D. dissertation, UCL (University College London), 2018.

[20] G. Venturi and J. Troost, "Survey on the ucd integration in the industry," in *Proceedings of the third Nordic conference on Human-computer interaction*, 2004, pp. 449–452.

[21] S. Breen, T. Enkerud, and K. Husøy, "Applying usability engineering in abb," in *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*, 2006, pp. 491–492.

[22] B. Szabó and K. Hercegfi, "User-centered approaches in software development processes: Qualitative research into the practice of hungarian companies," *Journal of Software: Evolution and Process*, vol. 35, no. 2, p. e2501, 2023.

[23] D. C. Rose, C. Parker, J. Fodery, C. Park, W. J. Sutherland, and L. V. Dicks, "Involving stakeholders in agricultural decision support systems: Improving user-centred design," *International Journal of Agricultural Management*, vol. 6, no. 1029-2019-924, pp. 80–89, 2018.

[24] G. Venturi, J. Troost, and T. Jokela, "People, organizations, and processes: An inquiry into the adoption of user-centered design in industry," *International Journal of Human-Computer Interaction*, vol. 21, no. 2, pp. 219–238, 2006.

[25] J. Harding, "Usability of geographic information–factors identified from qualitative analysis of task-focused user interviews," *Applied Ergonomics*, vol. 44, no. 6, pp. 940–947, 2013.

[26] D. E. Kosnik and L. J. Henschen, "Design and interface considerations for web-enabled data management in civil infrastructure health monitoring," in *Human-Computer Interaction. Applications and Services: 15th International Conference, HCI International 2013, Las Vegas, NV, USA, July 21-26, 2013, Proceedings, Part II 15*. Springer, 2013, pp. 107–116.

[27] T. Jokela and A. Lucero, "Mixednotes: a digital tool to prepare physical notes for affinity diagramming," in *Proceedings of the 18th International Academic MindTrek Conference: Media Business, Management, Content & Services*, 2014, pp. 3–6.

[28] Z. He and H. Peng, "Research on user experience design based on affinity diagram assisting user modeling–taking music software as an example," in *International Conference on Human-Computer Interaction*. Springer, 2023, pp. 516–526.

[29] A. Lucero, "Using affinity diagrams to evaluate interactive prototypes," in *Human-Computer Interaction–INTERACT 2015: 15th IFIP TC 13 International Conference, Bamberg, Germany, September 14-18, 2015, Proceedings, Part II 15*. Springer, 2015, pp. 231–248.

[30] D. Mohamedally and P. Zaphiris, "Categorization constructionist assessment with software-based affinity diagramming," *Intl. Journal of Human–Computer Interaction*, vol. 25, no. 1, pp. 22–48, 2009.

[31] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric environment*, vol. 32, no. 14-15, pp. 2627–2636, 1998.

[32] L. Atlas, T. Homma, and R. Marks, "An artificial neural network for spatio-temporal bipolar patterns: Application to phoneme classification," in *Neural Information Processing Systems*, 1987.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[35] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[36] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[37] A. Van Den Oord, O. Vinyals *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[38] A. Vahdat, F. Williams, Z. Gojcic, O. Litany, S. Fidler, K. Kreis *et al.*, "Lion: Latent point diffusion models for 3d shape generation," *Advances in Neural Information Processing Systems*, vol. 35, pp. 10 021–10 039, 2022.

[39] L. Melas-Kyriazi, C. Rupprecht, and A. Vedaldi, "Pc2: Projection-conditioned point cloud diffusion for single-image 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 923–12 932.

[40] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[41] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," *arXiv preprint arXiv:2106.10689*, 2021.

[42] F. Williams, Z. Gojcic, S. Khamis, D. Zorin, J. Bruna, S. Fidler, and O. Litany, "Neural fields as learnable kernels for 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 500–18 510.

[43] O. Elharrouss, Y. Akbari, N. Almaadeed, and S. Al-Maadeed, "Backbones-review: Feature extraction networks for deep learning and deep reinforcement learning approaches," *arXiv preprint arXiv:2206.08016*, 2022.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[45] P. Mittal, Y.-C. Cheng, M. Singh, and S. Tulsiani, "Autosdf: Shape priors for 3d completion, reconstruction and generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 306–315.

[46] Y.-C. Cheng, H.-Y. Lee, S. Tulyakov, A. G. Schwing, and L.-Y. Gui, "Sdfusion: Multimodal 3d shape completion, reconstruction, and generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4456–4465.

[47] Z. Huang, V. Jampani, A. Thai, Y. Li, S. Stojanov, and J. M. Rehg, "Shapeclipper: Scalable 3d shape learning from single-view images via geometric and clip-based consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 912–12 922.

[48] P. Zins, Y. Xu, E. Boyer, S. Wuhrer, and T. Tung, "Multi-view reconstruction using signed ray distance functions (srdf)," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 696–16 706.

[49] C. H. Koneputugodage, Y. Ben-Shabat, and S. Gould, "Octree guided unoriented surface reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 717–16 726.

[50] Y. Ren, T. Zhang, M. Pollefeys, S. Süsstrunk, and F. Wang, "Volrecon: Volume rendering of signed ray distance functions for generalizable multi-view reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 685–16 695.

[51] Z. Li, T. Müller, A. Evans, R. H. Taylor, M. Unberath, M.-Y. Liu, and C.-H. Lin, "Neuralangelo: High-fidelity neural surface reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8456–8465.

[52] Z. Wang, S. Zhou, J. J. Park, D. Paschalidou, S. You, G. Wetzstein, L. Guibas, and A. Kadambi, "Alto: Alternating latent topologies for implicit 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 259–270.

[53] T. Chu, P. Zhang, Q. Liu, and J. Wang, "Buol: A bottom-up framework with occupancy-aware lifting for panoptic 3d scene reconstruction from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4937–4946.

[54] Y. Wang, X. He, S. Peng, H. Lin, H. Bao, and X. Zhou, "Autorecon: Automated 3d object discovery and reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 382–21 391.

[55] X. Xu, P. Guerrero, M. Fisher, S. Chaudhuri, and D. Ritchie, "Unsupervised 3d shape reconstruction by part retrieval and assembly," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8559–8567.

[56] Z. Zhou and S. Tulsiani, "Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 588–12 597.

[57] F. Wimbauer, N. Yang, C. Rupprecht, and D. Cremers, "Behind the scenes: Density fields for single view reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9076–9086.

[58] L. Melas-Kyriazi, I. Laina, C. Rupprecht, and A. Vedaldi, "Realfusion: 360deg reconstruction of any object from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8446–8455.

[59] R. A. Rosu and S. Behnke, "Permutosdf: Fast multi-view reconstruction with implicit surfaces using permutohedral lattices," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8466–8475.

[60] B. Cai, J. Huang, R. Jia, C. Lv, and H. Fu, "Neuda: Neural deformable anchor for high-fidelity implicit surface reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8476–8485.

[61] P. Li, J. Guo, X. Zhang, and D.-M. Yan, "Secad-net: Self-supervised cad reconstruction by learning sketch-extrude operations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 816–16 826.

[62] J. Ling, Z. Wang, and F. Xu, "Shadowneus: Neural sdf reconstruction by shadow ray supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 175–185.

[63] A. Boulch and R. Marlet, "Poco: Point convolution for surface reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6302–6314.

[64] H. Guo, S. Peng, H. Lin, Q. Wang, G. Zhang, H. Bao, and X. Zhou, "Neural 3d scene reconstruction with the manhattan-world assumption," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5511–5520.

[65] S. Duggal and D. Pathak, "Topologically-aware deformation fields for single-view 3d reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1536–1546.

[66] B. Ma, Y.-S. Liu, M. Zwicker, and Z. Han, "Surface reconstruction from point clouds by learning predictive context priors," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6326–6337.

[67] C. Sun, M. Sun, and H.-T. Chen, "Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5459–5469.

[68] S. Song, Z. Cui, and R. Qin, "Vis2mesh: Efficient mesh reconstruction from unstructured point clouds of large scenes with learned virtual view visibility," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6514–6524.

[69] L. Jin, S. Qian, A. Owens, and D. F. Fouhey, "Planar surface reconstruction from sparse views," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 991–13 000.

[70] D. Chen, Z. Tang, Z. Xu, Y. Zheng, and Y. Liu, "Gaussian fusion: Accurate 3d reconstruction via geometry-guided displacement interpolation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5916–5925.

[71] Y. Siddiqui, J. Thies, F. Ma, Q. Shan, M. Nießner, and A. Dai, "Retrievalfuse: Neural 3d scene reconstruction with a database," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 568–12 577.

[72] J. Zhang, Y. Yao, and L. Quan, "Learning signed distance field for multi-view surface reconstruction. 2021 ieee," in *CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6505–6514.

[73] D. Wang, X. Cui, X. Chen, Z. Zou, T. Shi, S. Salcudean, Z. J. Wang, and R. Ward, "Multi-view 3d reconstruction with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5722–5731.

[74] M. Yavartanoo, J. Chung, R. Neshatavar, and K. M. Lee, "3dias: 3d shape reconstruction with implicit algebraic surfaces," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 446–12 455.

[75] Y. Furukawa, B. Curless, S. M. Seitz, and R. Szeliski, "Manhattan-world stereo," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1422–1429.

[76] C. Liu, K. Kim, J. Gu, Y. Furukawa, and J. Kautz, "Planercnn: 3d plane detection and reconstruction from a single image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4450–4459.

[77] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[78] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics (ToG)*, vol. 41, no. 4, pp. 1–15, 2022.

[79] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996, pp. 303–312.

[80] S. Schaefer and J. Warren, "Dual marching cubes: Primal contouring of dual grids," in *12th Pacific Conference on Computer Graphics and Applications, 2004. PG 2004. Proceedings.* IEEE, 2004, pp. 70–76.

[81] J. Huang, H. Su, and L. Guibas, "Robust watertight manifold surface generation method for shapenet models," *arXiv preprint arXiv:1802.01698*, 2018.

[82] P. Labatut, J.-P. Pons, and R. Keriven, "Robust and efficient surface reconstruction from range data," in *Computer graphics forum*, vol. 28, no. 8. Wiley Online Library, 2009, pp. 2275–2290.

[83] D. Stutz and A. Geiger, "Learning 3d shape completion under weak supervision," *International Journal of Computer Vision*, vol. 128, pp. 1162–1181, 2020.

[84] M. Kazhdan and H. Hoppe, "Screened poisson surface reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 32, no. 3, pp. 1–13, 2013.

[85] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 2016, pp. 501–518.

[86] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[87] O. Michel, R. Bar-On, R. Liu, S. Benaim, and R. Hanocka, "Text2mesh: Text-driven neural stylization for meshes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 492–13 502.

Appendix A

Supplementary Data

A.0.1 Presentation 1

Scalable AI-Assisted 3D City Model Reconstruction from Multiple Sourced Systems
MSc Progress Presentation

Supervision team:
Dr Gary K.L Tam
Dr Sean Walton
Dr Nicholas Micallef

Student:
Andreas Christodoulides

MSc Research Component and Goals

- MSc research Submission: End of September
- MSc research goals:
  - ❖ Deep understanding of State-of-the-Art (SOTA) literature
    - ➢ Identify trends and gaps
  - ❖ Deep understanding of OS's needs and requirements
  - ❖ Implementation of at least one suitable SOTA approach and start human-in-the-loop approach to gather insights

MSc work so far: Literature preliminary insights

- Signed Distance Functions (SDF) are the most popular rendering approach
  - Strengths:
    - Low computational requirements compared to other rendering techniques (e.g., NeRF)
    - Can be easily combined with other techniques (e.g., SDF-NeRF, SDF-NV)
  - Weaknesses
    - High computational costs when rendering multiple objects/large scenes at high resolutions
    - Struggles with occlusion removal when employed by itself

- Gaps in the literature so far:
  - High focus in literature on enhancing digital arts or knowledge progression on general object reconstruction – a specialised technique promises better results and robustness
  - Very little literature is focused on how to fuse different modalities (examples found: sRGB with metadata, multi-view images with camera poses)
  - Literature is primarily interested in single object reconstruction instead of scene

MSc work so far: Best performing SOTA techniques

LION: Latent Point Diffusion Model [NeurIPS2022]

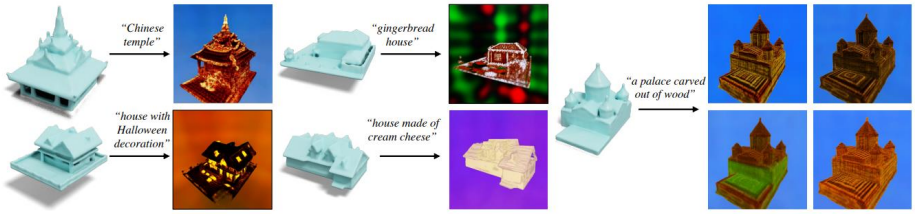
High quality reconstructions:

Can be combined with text-driven shape generation, or texture stylising methods (i.e., Text2Mesh)

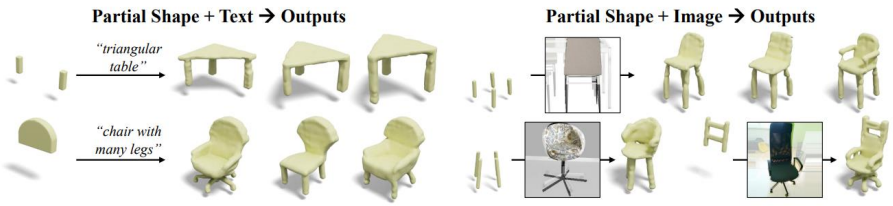
Shape Interpolation capabilities:

Possible interactive methods (borrowed from literature)

Example 1: SDFusion text guided texturing [CVPR2023]

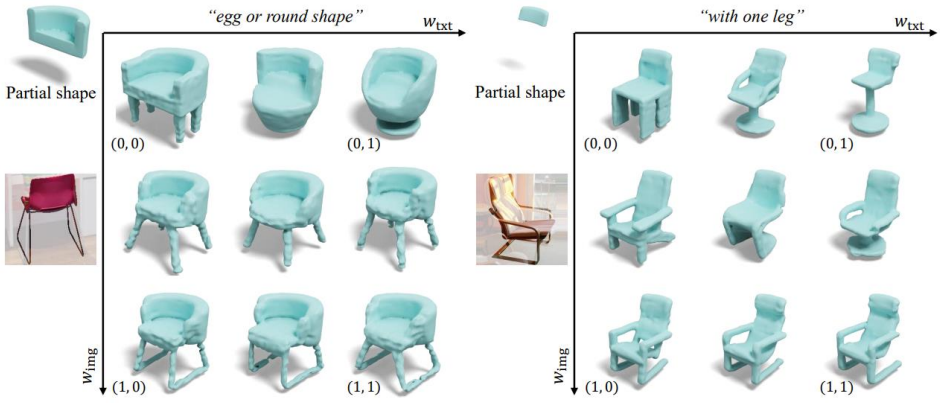


Example 2: SDFusion Conditional generation, partial shape with text guidance



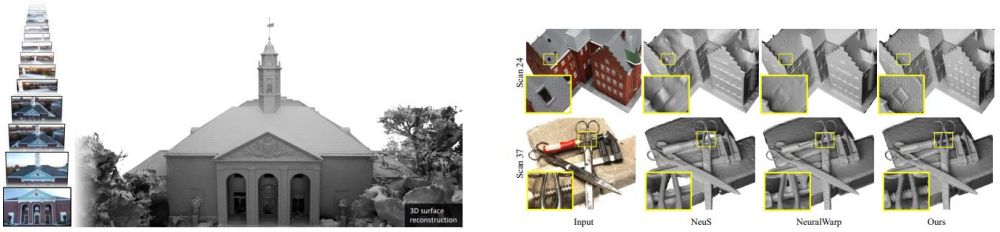
Possible interactive methods (borrowed from literature)

Example 3: SDFusion Multiple conditioning variables with adjustable weight control



Possible interactive methods (borrowed from literature)

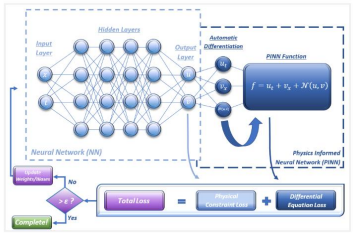
Example 4: Neuralangelo Interactive selection of region of interest from point cloud [CVPR2023]



Areas which may benefit OS

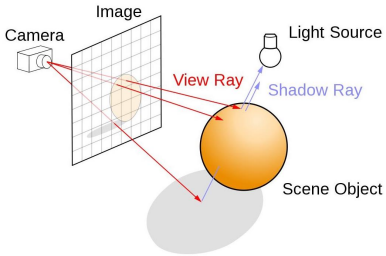
Physics informed Neural Networks (PINNs)

- Building defect detection
- Smart city foundation



Material detection through ray tracing

- Leverage on simulated light interactions



Understanding of OS's needs and requirements

- Human-in-the-loop, Scalability, Data Fusion… What else?
- Further extend needs and requirements through:
  - ❖ Management consultation
  - ❖ User centred studies involving OS's experts (possibly interviews)
    - ➢ Who are the experts we need to contact for interviews and what are their roles? (different roles and backgrounds will be primarily concerned with different aspects, need to account for variability in expertise in the studies)
    - ➢ Need understanding of the manual process from start to finish (data types, intermediate inputs/outputs, finalised representation)

Scalable AI-Assisted 3D City Model Reconstruction from Multiple Sourced Systems
MSc Progress Presentation

Thank you for your attention!

A.0.2 Presentation 2

Scalable AI-Assisted 3D City Model Reconstruction from Multiple Sourced Systems
MSc Progress Presentation

Supervision team:
Dr Gary K.L Tam
Dr Sean Walton
Dr Nicholas Micallef

Student:
Andreas Christodoulides

MSc Research Component (Fast Track)

- MSc research Submission: End of September
- MSc research goals:
  - ❖ Deep understanding of State-of-the-Art (SOTA) literature
    - ➢ Identify trends and gaps
  - ❖ Deep understanding of OS's needs and requirements
  - ❖ Implementation of at least one suitable SOTA approach and start using human-in-the-loop approaches to gather insights

PhD Research Component (Slow Track)

- PhD Duration: September 2023 – September 2026
- PhD research plan (Year 1):
  - ❖ Further extend literature review → Literature Survey
    - ➢ Objects of interest: Noisy point cloud processing, 3D scene understanding, 3D reconstruction (explicit surface output)
  - ❖ Implementation of several suitable SOTA techniques (Qualitative and Quantitative analysis of techniques)
  - ❖ Co-evolutionary approach: Work packages of shorter durations that explore OS's needs
    - ➢ Regular meetings with OS team to shape research undertaken

Semantics of Geospatial Information

How can semantic information be included in 3D datasets? (Borrowed from Kelvin's PhD Dissertation)

- Building Installations (Used inside buildings with external components e.g., AC units)
- City Furniture (e.g., street lamps, benches)
- Generic City Object (e.g., monuments, fences)
- Door (entry and exit points of buildings)
- Ground Surface (ground level terrain)
- Land Use (e.g., residential, industrial, etc.)
- Plant Cover (e.g., trees, parks)
- Relief Feature (elevation)
- Road Types
- Roof Surface (Roof types)
- Wall Surface (External walls of buildings)
- Water Body
- Water Surface
- Window

Semantics of Geospatial Information

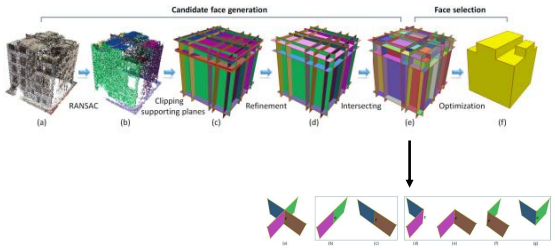
How can semantic information be included in 3D datasets? (Borrowed from Kelvin's PhD Dissertation)

- Building Installations (Used inside buildings with external components e.g., AC units)
- City Furniture (e.g., street lamps, benches)
- Generic City Object (e.g., monuments, fences)
- Door (entry and exit points of buildings)
- Ground Surface (ground level terrain)
- Land Use (e.g., residential, industrial, etc.)
- Plant Cover (e.g., trees, parks)
- Relief Feature (elevation)
- Road Types
- Roof Surface (Roof types)
- Wall Surface (External walls of buildings)
- Water Body
- Water Surface
- Window

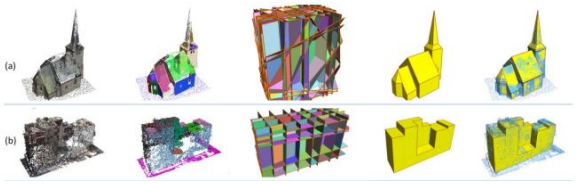
A. Supplementary Data

Explicit Surface Representation example: PolyFit[ICCV2017]

PolyFit procedure:



PolyFit: Examples



MSc work so far: Best performing SOTA technique

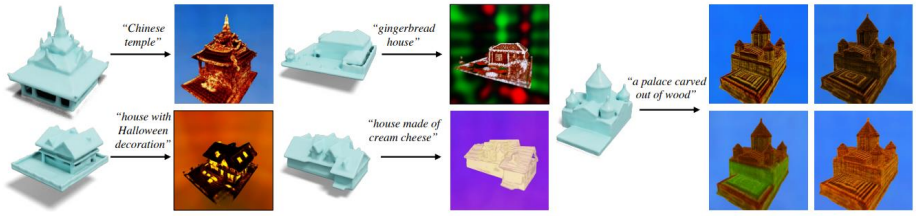
LION: Latent Point Diffusion Model [NeurIPS2022]
High quality reconstructions:

Can be combined with text-driven shape generation, or texture stylising methods (i.e., Text2Mesh)

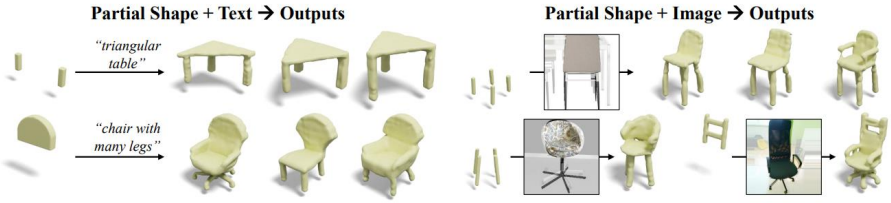
Shape Interpolation capabilities:

Possible interactive methods (borrowed from literature)

Example 1: SDFusion text guided texturing [CVPR2023]

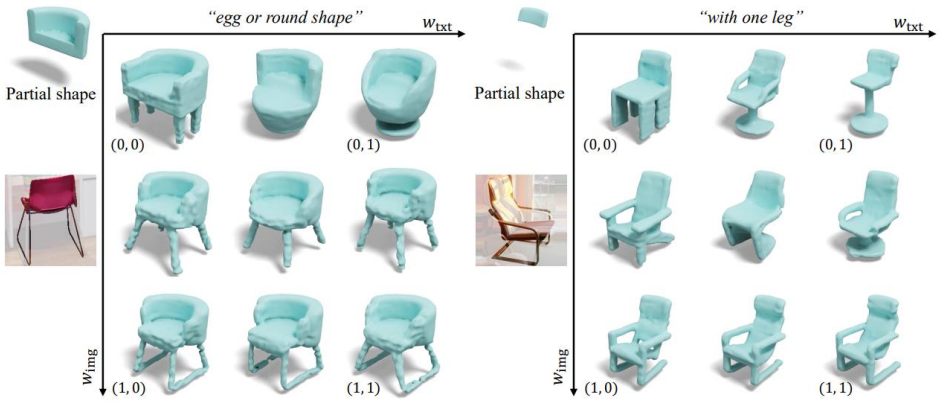


Example 2: SDFusion Conditional generation, partial shape with text guidance



Possible interactive methods (borrowed from literature)

Example 3: SDFusion Multiple conditioning variables with adjustable weight control



Possible interactive methods (borrowed from literature)

Example 4: Neuralangelo Interactive selection of region of interest from point cloud [CVPR2023]

Possible interactive methods (borrowed from literature)

Example 4: Neuralangelo Interactive selection of region of interest from point cloud [CVPR2023]

Scalable AI-Assisted 3D City Model Reconstruction from Multiple Sourced Systems
MSc Progress Presentation

Thank you for your attention!