The University of Melbourne

THE UNIVERSITY OF MELBOURNE

Research Proposal

Segment Anything for 3D Outdoor Scenes

Masters of Computer Science

Author:
Arya Araban

Supervisors:
A/Prof Kourosh Khoshelham
Cipher Zhang

April 2024

Contents

1 Introduction

3D point cloud data has gained significant traction as a powerful representation for capturing the rich geometric structure of real-world environments. Acquired through various sensing technologies, such as light detection and ranging (LiDAR) and photogrammetry, point clouds offer a direct and detailed depiction of the 3D geometry, preserving crucial depth information and enabling precise localization and mapping of objects and surfaces [1]. Segmentation, a process that partitions the point cloud into distinct regions or objects, plays a pivotal role in this context, enabling the extraction of meaningful information and facilitating the identification and analysis of specific structures within the data. This is crucial in fields where accurate object identification and spatial understanding are key to safe and efficient operation.

Despite point cloud data offering a wealth of detailed geometric information, extracting meaningful representations and segmenting objects of interest within this unstructured type of data remains a formidable challenge. Traditional approaches often rely on handcrafted feature descriptors or shallow learning models, which struggle to capture the intricate patterns and long-range dependencies inherent in complex 3D scenes [2]. While deep learning techniques have shown remarkable success in 2D image segmentation tasks, adapting these models to the unique characteristics of 3D point clouds poses several challenges.

One of the main challenges in point cloud segmentation is the lack of structured data. Unlike 2D images, where the pixels are neatly arranged in a grid of rows and columns, points in a 3D point cloud are unstructured and unordered, making it difficult to apply conventional convolutional operations. Furthermore, point clouds are often sparse and cover large 3D spaces, making it challenging to capture the long-range dependencies between points. These challenges necessitate the development of novel methods that can effectively handle the unique characteristics of 3D point cloud data.

Building on these challenges, various methodologies exist for point cloud segmentation, each with their own unique characteristics. Conventionally, semantic segmentation, which combines object segmentation and classification into a single process, has been widely used across applications such as autonomous driving and robotics [3]. However, this approach relies on predefined class labels, which may not cover all objects or scenarios in dynamic real-world environments where the scene is constantly changing. An alternative approach is class-agnostic segmentation, which involves segmenting arbitrary objects within the point cloud without the need for predefined class labels. Following segmentation, a separate classification step may be applied. This approach is advantageous as it enables open set recognition, which is the ability to recognize objects that were not part of the training set, making it more suitable for dynamic environments.

Recent advancements in class-agnostic segmentation for 2D images, exemplified by Meta's release of the Segment Anything Model (SAM) [4], have yielded highly promising outcomes in achieving their objective, opening new opportunities for similar breakthroughs in 3D point cloud data. SAM's strength lies in its ability to segment arbitrary objects in an image by leveraging the self-attention mechanism. This mechanism allows the model to capture long-range dependencies and learn discriminative pixel-wise

features, enabling it to distinguish between different objects in an image, even when they are not predefined. The success of SAM in 2D image segmentation presents an intriguing possibility: could a similar approach be applied to 3D point cloud data?

The potential benefits of applying a SAM-like architecture to 3D point cloud data are immense. For instance, in outdoor driving scenes, accurately segmenting objects within the point cloud data can significantly improve the performance of autonomous driving systems and enable downstream tasks such as object tracking and behavior prediction. By being able to segment arbitrary objects, such as pedestrians, vehicles, and buildings, without relying on predefined class labels, a SAM-like model could provide a more flexible and robust solution for 3D point cloud segmentation.

2 Problem Statement

Building upon the challenges and opportunities outlined in the previous section, there is lack of effective and efficient methods capable of handling the diverse and intricate nature of objects encountered in real-world 3D environments for class-agnostic point cloud segmentation. Despite the increasing interest in this area, addressing the unique challenges posed by point clouds remains a formidable task. These challenges arise from the unstructured and irregular nature of point clouds, their sparsity, and the absence of inherent order.

Our research aims to bridge this gap by formulating a pioneering architecture specifically tailored for class-agnostic 3D point cloud segmentation. While the remarkable success of transformer-based architectures, especially SAM, have demonstrated the potential for accurate segmentation of arbitrary objects in 2D images, extending these approaches to the 3D domain necessitates the development of novel architectural designs and techniques to handle the challenges posed by point cloud data.

In pursuit of this objective, we are guided by two fundamental research questions:

- **RQ1:** Is it possible to successfully adapt the key components of the transformer-based architecture in SAM to the 3D domain by identifying suitable equivalents for encoding, self-attention, and decoding modules that can effectively process and segment 3D point cloud data in a class-agnostic manner?
- **RQ2:** Transformer-based models generally necessitate substantial volumes of labeled data for training. Is it possible to utilize synthetic 3D driving scene data to pre-train a high-performing model for class-agnostic 3D point cloud segmentation in real-world scenarios, thereby mitigating the requirement for extensive real-world labeled data?

To summarize, the principle goal of our research is to develop a customized transformer-based architecture for class-agnostic segmentation of arbitrary objects in 3D point clouds, with a focus on outdoor driving scenes. We also explore the utilization of synthetic data for effective pre-training and transfer learning.

The following objectives and hypotheses are associated with the research questions:

Objective 1: Develop a transformer-based architecture that can effectively encode

and process 3D point cloud data, leveraging self-attention mechanisms and specialized encoding modules to capture long-range dependencies and learn discriminative point-wise features.

Hypothesis 1: By adapting the key components of the architecture of SAM to the 3D domain, it is possible to create a model that can segment arbitrary objects within 3D point clouds in a class-agnostic manner, overcoming the challenges posed by the inherent characteristics of point clouds.

Objective 2: Investigate the feasibility of utilizing substantial volumes of synthetic 3D driving scene data for pre-training the proposed architecture, mitigating the need for extensive real-world labeled data.

Hypothesis 2: Given the vast diversity and complexity of real-world driving scenes, pre-training the proposed model on large-scale synthetic data can provide a valuable initialization, enabling the model to learn generalizable representations that can be effectively fine-tuned on real-world data.

Objective 3: Evaluate the performance and generalization capabilities of the proposed architecture on real-world outdoor driving scenes, and compare it with existing approaches.

Hypothesis 3: The proposed architecture, pre-trained on synthetic data and fine-tuned on limited real-world data, can outperform existing supervised and unsupervised approaches in class-agnostic segmentation of arbitrary objects in real-world 3D point clouds, demonstrating superior performance and generalization capabilities.

By fulfilling these objectives and confirming the hypotheses, this research will provide a robust solution for class-agnostic 3D point cloud segmentation in outdoor driving scenes, paving the way for more flexible and adaptable 3D perception systems across various domains.

3 Literature Review

In this section, we explore 3D data acquisition techniques and deep learning methods for point cloud segmentation, setting the foundation for our research on class-agnostic segmentation of arbitrary objects in 3D point clouds. We discuss techniques like LiDAR, photogrammetry, and synthetic data generation for capturing 3D data. We also briefly cover the various deep learning approaches for 3D semantic segmentation, and finally highlight the emerging field of 3D class-agnostic segmentation inspired by SAM.

3.1 Point Clouds and Their Acquisition

Point clouds are a fundamental data structure which represent a set of points, denoted as $\{P_i\}_{i=1}^{n}$ in a three-dimensional coordinate system. Each point is defined by its spatial coordinates (X, Y, Z) and can also have additional attributes such as color values (R, G, B), surface normal vectors (n_x, n_y, n_z), or intensity information [5]. Point clouds provide a versatile and efficient way to represent and analyze real-world objects and environments, making them invaluable in a wide range of applications.

Exhibiting strengths in capturing intricate geometric details and accurately representing complex 3D structures, point clouds find extensive utilization in diverse domains such as 3D reconstruction, object recognition, and scene understanding. They offer a direct and efficient means to digitize real-world environments. However, point clouds also have limitations. Their unstructured nature makes processing and analysis challenging using traditional image processing techniques. Additionally, point clouds can be sparse or non-uniform, leading to information loss or incomplete representations. Furthermore, the lack of semantic information makes understanding object relationships difficult without additional processing or labeling.

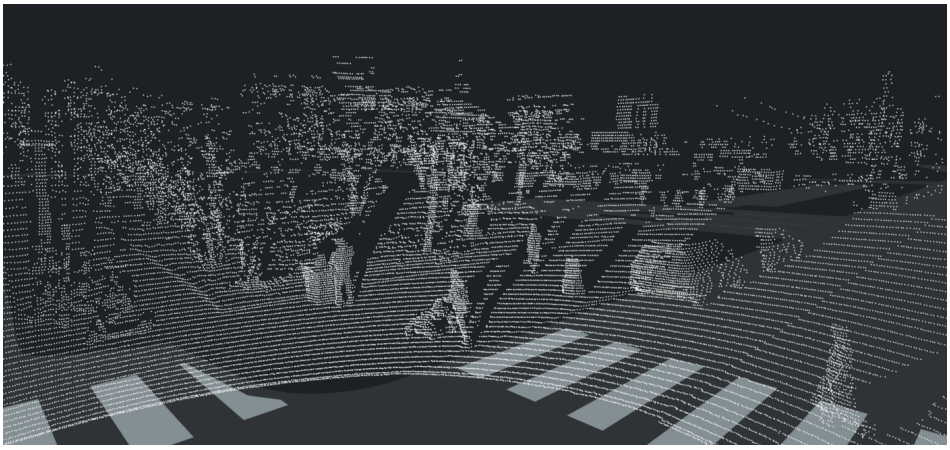

Figure 1: Example of a point cloud [6]

Point cloud data can be acquired through various techniques, broadly categorized into two main approaches: active sensing and passive sensing [7]. Active sensing methods actively emit and detect signals, such as light or sound waves, to measure the distance and position of points. Passive sensing methods, on the other hand, rely on capturing and analyzing existing environmental conditions, such as light reflections or images, to reconstruct the 3D structure of objects and scenes.

Light Detection and Ranging (LiDAR) stands out as a prominent active sensing technique utilized for acquiring point cloud data. In LiDAR systems, a laser scanner emits light pulses, and the time elapsed before the pulses return after bouncing off objects in the surrounding environment is measured. This time-of-flight (ToF) measurement is then employed to construct a precise 3D point cloud representation of the environment. LiDAR systems find application across various domains, including surveying, mapping, and industrial inspection.

There are various types of LiDAR systems, each with their own advantages and applications:

- **Airborne LiDAR**: Mounted on aircraft, airborne LiDAR efficiently maps vast areas, making it ideal for topographic mapping and landscape modeling due to its high vantage point [8].
- **Terrestrial LiDAR**: Ground-based, terrestrial LiDAR produces detailed 3D models of buildings and infrastructure with high resolution, valuable for architectural surveys, construction monitoring, and heritage documentation [9].

- **Mobile LiDAR**: Mounted on vehicles or robots, mobile LiDAR combines broad coverage with precise data capture by operating while in motion. It's suitable for swiftly modeling urban environments and transportation infrastructure, including roads, bridges, and tunnels [10].

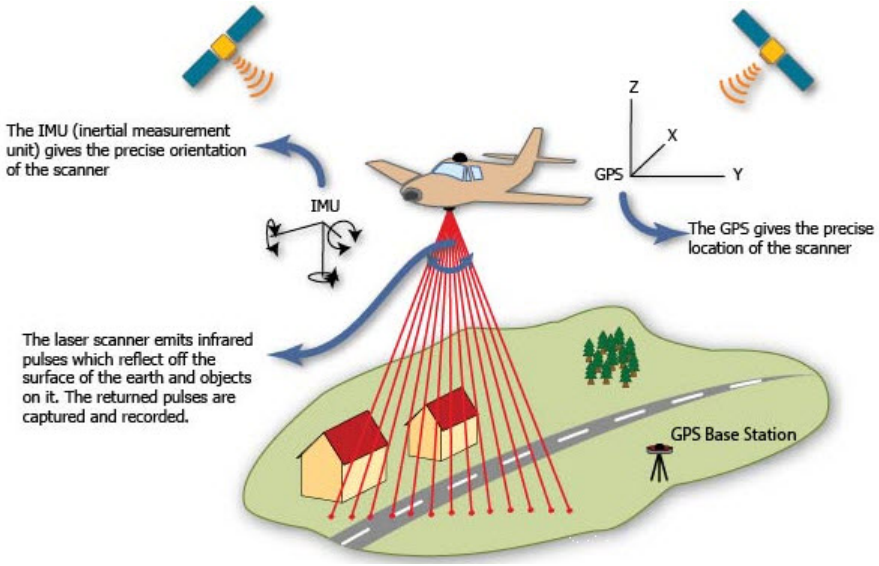

Figure 2: Working principle of an airbourne LiDAR system [11]

Shifting focus from active to passive sensing techniques, Photogrammetry emerges as a significant method. This passive technique extracts 3D information from overlapping 2D images or video frames. By analyzing the geometrical relationships between multiple images captured from different viewpoints, photogrammetry can reconstruct the 3D structure of objects or scenes [12].

The versatility of photogrammetry allows it to be performed using various imaging platforms, including aerial photography from manned aircraft or unmanned aerial vehicles (UAVs/drones), terrestrial close-range photography, or satellite imagery [13]. The choice of platform depends on factors like the desired spatial resolution, coverage area, and accessibility of the target environment.

There are two main approaches to photogrammetry:

- **Stereoscopic Photogrammetry:** Uses stereovision to calculate 3D coordinates by matching points in images captured from different viewpoints. Effective in controlled environments or with stationary objects Effective in controlled environments or with stationary objects due to its similarity to human depth perception [14].
- **Structure from Motion (SfM):** Combines computer vision algorithms with bundle adjustment methods. It detects and matches features across overlapping images to estimate camera positions and orientations, then refines these estimates to reconstruct 3D geometry. Particularly powerful as it can utilize images from arbitrary viewpoints, including handheld cameras or drones. [15].

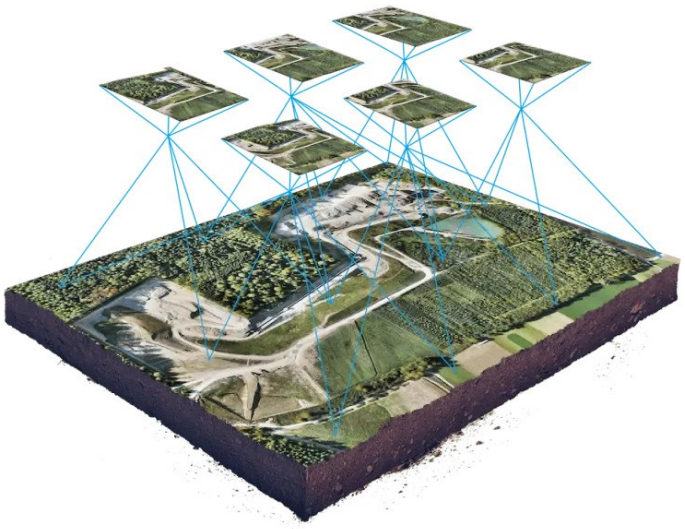

Figure 3: Photogrammetry illustration [16]

LiDAR and photogrammetry are both effective methods for capturing detailed information about the environment, but they have different strengths and weaknesses. LiDAR excels at capturing detailed geometric information and penetrating dense vegetation, while photogrammetry is cost-effective and offers high-resolution color information [17]. The choice between the two often depends on factors such as desired accuracy, environmental complexity, budget constraints, and specific application requirements.

While these techniques possess valuable strengths, the capture of highly detailed point clouds often comes with significant costs, leading to a limited availability of high-quality datasets. The emergence of synthetic data generated through advanced simulation software, however, offers a viable solution to overcome these challenges [18]. This approach not only allows for the creation of large volumes of data, but also enables the seamless incorporation of additional information, such as color, into the simulated point clouds. By leveraging synthetic data, researchers and developers can accelerate algorithm development for various applications, including autonomous driving, robotics, and environmental modeling, without being hindered by data scarcity [19].

3.2 3D Semantic Segmentation: An Overview

While our research endeavors to address the challenge of class-agnostic segmentation of arbitrary objects in 3D point clouds, it is important to understand the substantial and relevant body of work dedicated to semantic segmentation, which involves assigning class labels to individual points or groups of points within a 3D point cloud. While our task differs from semantic segmentation in that we do not rely on predefined object classes, we nonetheless benefit from the insights and techniques developed in this field.

Point cloud semantic segmentation has seen considerable progress with the introduction of deep learning techniques, enabling automatic feature extraction and outperforming traditional methods. This has led to the development of various deep learning-based

approaches, which can be categorized into three paradigms: image-based, voxel-based, and point-based methods [20].

Figure 4: Segmentation example from Mapillary Vistas dataset [21]

Image-based and voxel-based methods address the challenge of unstructured and disordered input data inherent to point clouds by transforming the original point cloud into a structured format that is readily amenable to processing by deep learning networks, enabling effective feature extraction and subsequent semantic segmentation. In contrast, point-based methods directly utilize the raw point cloud data and extract features utilized in segmentation.

This section delves into an exploration of all three paradigms, providing an overview of their key functionalities and characteristics.

3.2.1 Image-based Segmentation

Image-based methods project the 3D point cloud onto 2D image planes, leveraging the well-established deep learning techniques for image processing. These methods can render the point cloud from multiple viewpoints, generating depth and RGB images, which are then processed by convolutional neural networks (CNNs) to extract features for semantic segmentation.

One of the initial works in this domain is Multi-View CNN (MVCNN) [22], whcih projected the initial point cloud from various perspectives, generating diverse 2D images. These images were processed using an innovative network that extracted features and combined them in a pooling layer. The aggregated features were seamlessly applied to the point cloud, resulting in successful segmentation. Building upon MVCNN, Boulch et al. introduced SnapNet [23]. SnapNet utilizes an MVCNN architecture to project the point cloud onto multiple virtual camera views, generating depth and RGB images. These images are then processed by separate CNN streams, and the resulting features are fused for semantic segmentation.

Although Image-based methods leverage the well-established techniques of 2D computer vision and can benefit from the rich feature representations learned by CNNs, they suffer from information loss during the projection step and struggle with occlusion and viewpoint variations, especially in the case of sparse point clouds.

3.2.2 Voxel-based Segmentation

Voxel-based methods, on the other hand, represent the 3D point cloud as a structured 3D voxel grid, where each voxel encodes the occupancy or density information of the points within its volume. This voxelized representation can then be processed by 3D CNNs or other deep learning architectures for semantic segmentation.

An example of such an approach is VoxNet [24], which utilized a 3D CNN for object classification and detection using voxelized point cloud data. By encoding occupancy or density information within each voxel, VoxNet facilitated spatial feature extraction. However, challenges arose when dealing with large-scale or high-resolution point clouds due to computational complexity and memory requirements. To tackle these challenges, OctNet [25] introduced an innovative octree-based encoding scheme. This hierarchical structure efficiently handled sparse voxel grids by allocating computational resources to non-empty voxels while bypassing empty regions. OctNet's adaptive partitioning scheme played a crucial role in enhancing the segmentation of complex scenes by capturing features at multiple scales.

Despite enabling direct application of 3D CNNs and capturing spatial context by encoding occupancy or density information, voxel-based methods may suffer from quantization artifacts, high computational complexity, and substantial memory requirements, especially when dealing with large-scale or high-resolution point clouds.

3.2.3 Point Based Segmentation

In contrast to image-based and voxel-based methods, point-based methods circumvent the need for intermediate representations. These methods aim to develop deep learning architectures that can effectively extract features and perform segmentation directly from the unstructured point cloud information, offering several advantages, including the ability to preserve the inherent spatial relationships within the point cloud data. There are various approaches to point-based methods, which can be categorized as follows:

- **Multi-Layer Perceptron (MLP) based Methods:** MLP-based methods employ multi-layer perceptron networks, which are fully connected neural networks, to process the point cloud data. These methods operate on individual points or local neighborhoods, extracting features and performing segmentation tasks.

  A groundbreaking work in this domain is PointNet [26], which revolutionized the field by directly operating on raw point cloud data without intermediate transformations. Its architecture uses MLPs to extract features for each point, which are then combined into a single global feature vector. PointNet's design ensures that the order in which the points are input doesn't affect the result, so it can capture the underlying geometry of the scene regardless of the point order. Building on this success, Qi et al. introduced PointNet++ [27], which incorporates hierarchical feature learning and local neighborhood information. PointNet++ breaks down the point cloud into smaller areas and combines features from each area to capture both fine-grained details and overall context, leading to improved segmentation performance.

- **Graph Convolutional Network (GCN) based Methods:** GCN-based methods represent the point cloud as a graph, where each point is treated as a node, and the relationships between points are encoded as edges. Graph convolutional operations are then applied to aggregate and propagate features across the graph, capturing both local and global contextual information.

  Dynamic graph CNN (DGCNN) [28] exemplifies a GCN-based approach which represents point clouds as graphs. It introduces edge-conditioned convolution to capture intricate local geometric features, taking into account relative positions and feature variances among neighboring points. This facilitates the extraction of rich local features crucial for accurate segmentation. Furthermore, DGCNN incorporates graph max-pooling, enabling effective aggregation of features across the graph to capture global characteristics. By integrating both local and global information, DGCNN achieves robust segmentation of point clouds.

- **Transformer-based Methods:** Transformer-based methods, inspired by the success of transformers in natural language processing and computer vision tasks, have recently gained traction in point cloud segmentation. These methods employ self-attention mechanisms to capture long-range dependencies and global contextual information.

  A significant advancement employing this approach is Point Transformer [29], which applies the transformer architecture [30], to point cloud data. It utilizes a vector self-attention layer that adaptively modulates feature channels within local neighborhoods. This layer enables the Point Transformer to capture and aggregate local geometric features effectively. In addition, the model incorporates a position encoding function to capture 3D coordinates by transforming the original 1D positional encoding [30] into a 3D positional encoding scheme, further enhancing its ability to understand the spatial relationships within the point cloud data. By stacking multiple layers of the Point Transformer, the model achieves superior performance compared to previous state-of-the-art models in semantic segmentation. Its success has inspired many subsequent works [31, 32, 33], which further refine and extend the transformer-based approach for point cloud processing.

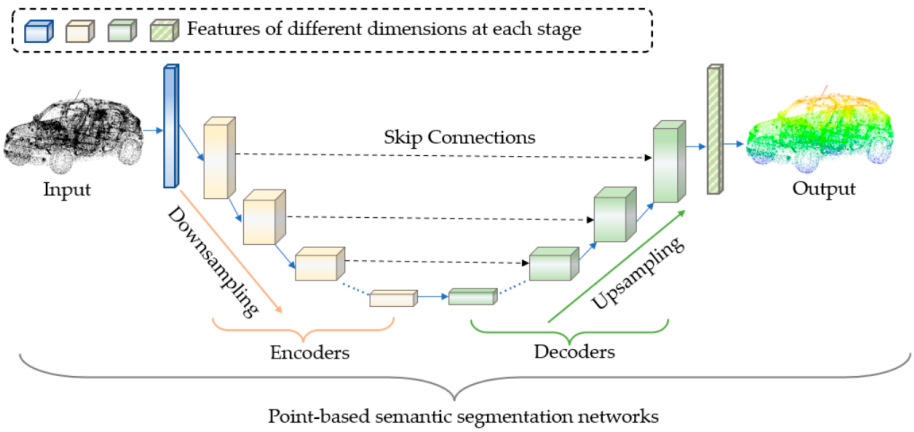

Figure 5: General structure of point-based networks [20]

While point-based methods are effective, they are not without limitations. MLP-based methods are efficient but struggle to capture long-range dependencies and global context. GCN-based methods, on the other hand, capture local and global context well but are computationally expensive and struggle with complex graph structures. Transformer-based methods, despite their computational costs and data requirements, excel in capturing both global and local information, achieving state-of-the-art segmentation results, making them become increasingly popular.

3.3 3D Class Agnostic Segmentation

Building upon the extensive research in 3D semantic segmentation, it is evident that significant strides have been made in developing deep learning architectures capable of accurately classifying and segmenting predefined object categories within point cloud data. However, the realm of class-agnostic segmentation, which aims to segment arbitrary objects without relying on predefined class labels, has garnered relatively less attention, despite its immense potential and practical applications, particularly in complex environments such as driving scenes where the diversity and unpredictability of objects present unique challenges and opportunities for advanced segmentation techniques.

Despite its relatively low profile, there has been research efforts to advance this field. LRGNet [34] proposes a learnable region growing method. Their approach can segment objects of any class using a single deep neural network, without assumptions about their shapes and sizes. Another notable work is Region Transformer [35], which introduces a novel region-based transformer model, employing a region-growth approach and self-attention mechanism to iteratively expand or contract a region by adding or removing points. These studies have demonstrated the feasibility and value of segmenting arbitrary objects within 3D point clouds.

Motivated by these findings, recent advancements in 3D class-agnostic segmentation have been greatly influenced by SAM, which revolutionized 2D image segmentation with its network leveraging transformer-based architectures and self-attention mechanisms. This success has sparked interest in adapting and extending these principles to tackle the unique challenges of 3D point cloud data. Researchers are exploring how SAM's remarkable capabilities in segmenting arbitrary objects can be applied to point clouds, with the potential to unlock new frontiers in class-agnostic segmentation. To understand SAM's potential for point cloud segmentation, it's important to grasp its inner workings and how it's being utilized for this purpose.

Unlike traditional models that require extensive training on specific tasks, SAM takes a more adaptable approach. Its game-changing impact lies in its zero-shot inference capabilities, meaning that SAM can accurately segment images without prior specific training, a task that traditionally requires tailored models. This leap forward is due to the influence of foundation models in natural language processing (NLP), including the GPT [36] and BERT [37] models. These models revolutionized how machines understand and generate human language by learning from vast data, allowing them to generalize across various tasks. SAM applies a similar philosophy to computer vision, using a large dataset to understand and segment a wide variety of images.

Figure 6: Example of SAM in action

The architecture of SAM consists of three main components: the image encoder, prompt encoder, and mask decoder. The image encoder is based on Vision Transformer (ViT) [38], a model that treats an image as a sequence of patches and applies transformer mechanisms to it. This process is time-consuming due to the complexity of the transformer model, but it allows for a more detailed understanding of the image content.

The prompt encoder transforms user prompts, such as points and bounding boxes, into vector embeddings, enabling the model to comprehend and identify the designated object of interest. For instance, a point is represented as the sum of a positional encoding of its location and a learned embedding indicating whether it belongs to the foreground (object of interest) or background. This process is carried out by passing the prompt through a small neural network that maps the prompt elements to their corresponding vector representations.

Finally, The mask decoder is a module that efficiently combines the image embedding from the encoder and the prompt embeddings to produce an output mask. It takes inspiration from Transformer segmentation models, featuring a modified Transformer decoder with cross-attention mechanisms. The decoder updates both the image embedding and prompt tokens through cross-attention, allowing for effective integration of the two sources of information. After upscaling the updated image embedding, the output token is used to dynamically predict the final mask through a small multi-layer perceptron.

SAM has been trained on a large and diverse dataset of 11 million images and 1.1 billion masks, generated using its own data engine, which enables it to demonstrate strong zero-shot performance on various segmentation tasks, accurately segmenting objects in new images it has never seen before.

It is important to note that user prompts, such as points and boxes, are not part of the dataset, and are dynamically created based on the ground truth mask while training the model. For inference, users can input their own prompts to segment specific objects of interest. However, if the goal is to segment everything in the image, SAM overlays a grid of point prompts on the image to attempt to segment all existing objects.

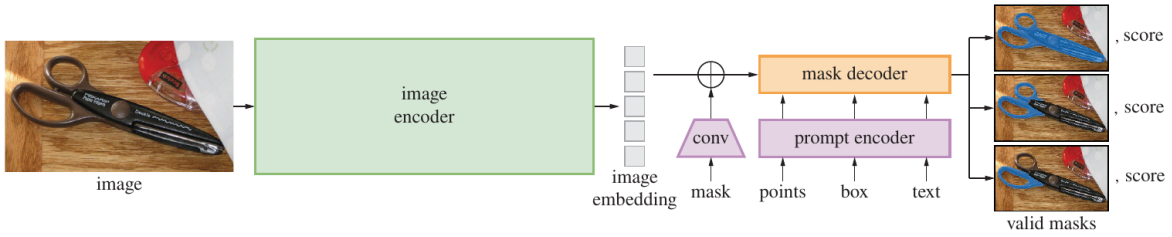

Figure 7: SAM overview: A heavyweight image encoder generates a compact image embedding that can be efficiently searched by input prompts to generate object masks [4]

SAM's capabilities have been extended to 3D point cloud segmentation, with SAM3D [39] being a prime example. In the context of a 3D scene, complete with a collection of RGB images and corresponding pose data, SAM3D leverages SAM to generate segmentation masks for each individual RGB image. These 2D masks are then projected onto the 3D point cloud, resulting in a series of 3D masks. In the final step, these 3D masks are merged together using a bottom-up approach, iteratively combining the masks to produce a comprehensive, final mask for the entire point cloud.

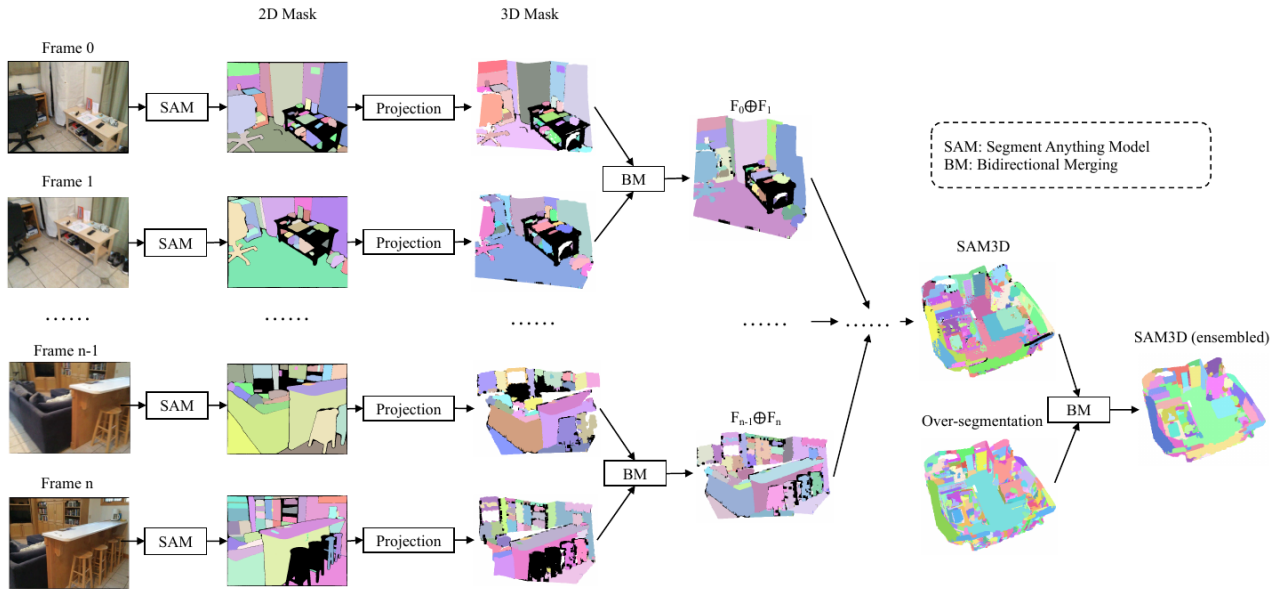

Figure 8: SAM3D Procedure

Segment3D [40] is another application that uses SAM for 3D segmentation. It employs a two-stage training approach to generate class-agnostic 3D segmentation masks without manual annotations. Initially, Segment3D pre-trains a 3D segmentation model on partial RGB-D point clouds using SAM-generated 2D masks as pseudo ground-truth labels. Then, the pre-trained model is fine-tuned on complete 3D point clouds in a self-supervised manner, using its own high-confidence mask predictions as the training signal. This approach avoids the merging process of SAM3D and directly infers

segmentation masks for entire 3D scenes using a native 3D model.

While existing methods demonstrate potential, they encounter the domain shift challenge when transitioning from 2D to 3D. This challenge arises due to the inherent differences in the data distribution and spatial relationships between 2D images and 3D point clouds. This issue opens an avenue for exploration: recreating SAM's architecture in 3D and training it within the same domain, potentially circumventing the domain shift problem and unlocking new possibilities in 3D image segmentation.

4 Methodology

The proposed methodology aims to develop a transformer-based architecture for class-agnostic segmentation of objects in 3D point clouds, with a focus on outdoor driving scenes. It is inspired by SAM's success in 2D image segmentation and aims to adapts its key components for 3D point cloud data. The methodology consists of data collection, model development, training, and evaluation.

4.1 Data Collection and Preprocessing

In this stage, we address the challenge of limited real-world labeled data for 3D point cloud segmentation by leveraging synthetic data generated from the CARLA (Car Learning to Act) [18] simulation environment. CARLA offers a realistic and diverse virtual environment specifically designed for autonomous driving scenarios. Within this environment, we create a wide range of driving scenes that vary in environmental conditions, vehicle types, and object configurations. The goal is to capture the complexity and diversity of real-world driving scenarios.

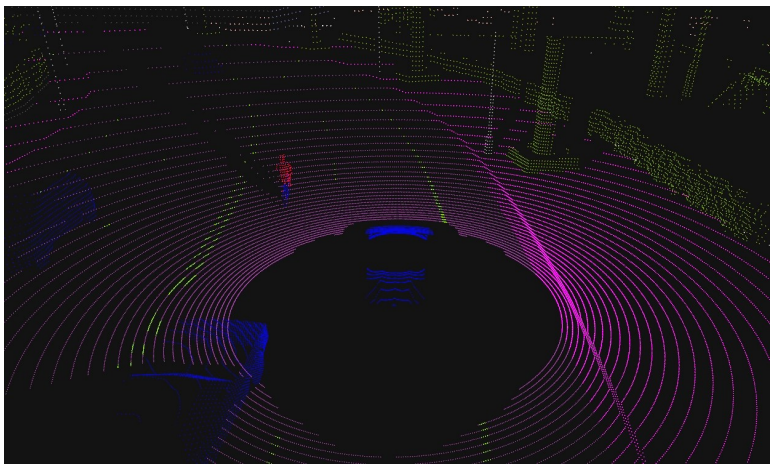

Figure 9: LiDAR Scan Example in CARLA

To obtain accurate 3D point cloud representations of driving scenes, we utilize a simulated LiDAR sensor mounted on a vehicle within the CARLA environment. This method involves casting rays from the sensor and recording the points where these rays intersect with the virtual environment. By simulating this process, we can capture the scene's geometric information in a manner that closely resembles real-world LiDAR

data, enabling us to generate detailed point cloud data that faithfully represents the spatial characteristics of the scenes.

CARLA provides comprehensive annotations for all objects within the scenes, including instance segmentation masks. These annotations serve as the foundation for generating ground truth segmentation masks, which are crucial for our class-agnostic segmentation task. However, to ensure compatibility with our proposed architecture, this raw point cloud data obtained from CARLA undergoes preprocessing.

To start the preprocessing, a region of interest (RoI) filtering step is applied, focusing the point cloud on the immediate surroundings of the vehicle. This reduces computational burden and directs the model's attention to critical areas for autonomous driving. Additionally, data augmentation techniques, including random rotations, translations, and scaling, are utilized to augment the point cloud data and segmentation masks. This enhances training data diversity and improves model generalization.

4.2 Model Architecture Development

Our proposed architecture, dubbed PointSAM (Point Cloud Segment Anything Model), is a transformer-based model designed for class-agnostic segmentation of 3D point clouds. similar to the base SAM, PointSAM consists of three main components: a point cloud encoder, a prompt encoder, and a segmentation mask decoder.

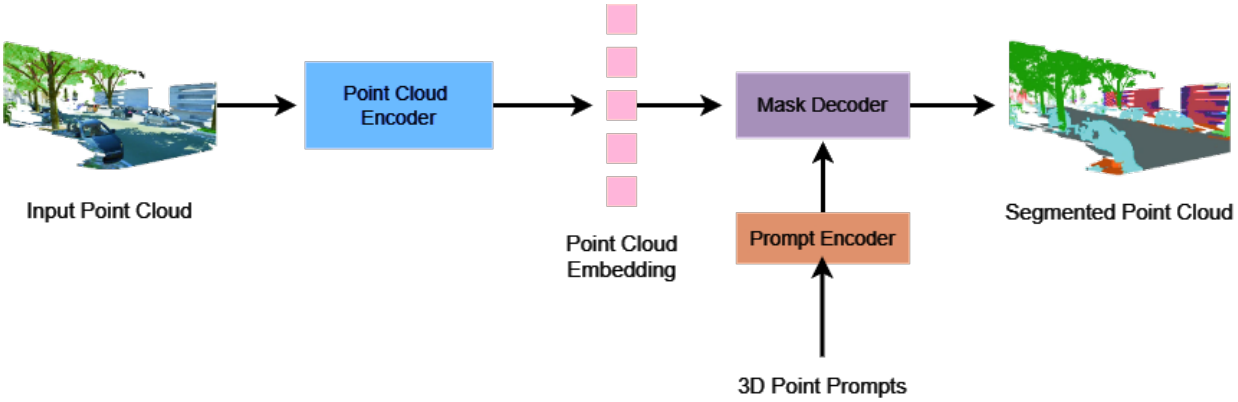

Figure 10: PointSAM architecture overview

4.2.1 Point Cloud Encoder

The point cloud encoder is responsible for embedding the 3D point cloud data into a latent representation that captures the geometric and structural information. To address the challenges posed by the irregularity and sparsity of point clouds, we utilize a pre-trained transformer architecture, such as Point Transformer [29], or other suitable alternatives [31, 41, 42].

Even though such architectures were originally designed for semantic segmentation, it is feasible to utilize them for class-agnostic segmentation. The architecture itself is not inherently tied to a specific task, and can be trained for class-agnostic segmentation given the appropriate training setup. These architectures are powerful at capturing geometric

and structural information from point clouds, which is crucial for the segmentation task at hand.

These point-based neural network architectures typically consist of encoding and decoding stages. During training, the encoder maps the input point cloud to a latent representation, capturing long-range dependencies and global context. The decoder then attempts to reconstruct the original point cloud from this latent representation. By training the model to accurately reconstruct the input, the encoder learns to extract rich geometric features that encode the underlying structure of the point cloud data.

For our purposes, we can leverage the pre-trained encoding stage to obtain high-level feature representations that encapsulate both local and global geometric information from the input point cloud.

4.2.2 Prompt Encoder

The prompt encoder plays a crucial role in translating the point prompts, which serve as the guidance for the segmentation task, into a meaningful representation that can be effectively integrated with the encoded point cloud features. In contrast to the 2D domain, where prompts are typically provided as points or bounding boxes on an image, our approach relies solely on 3D point prompts within the point cloud.

To generate these point prompts, we employ a strategic sampling strategy during training. Leveraging the ground truth segmentation masks provided by the synthetic CARLA data, we sample a predetermined number of points from the regions corresponding to the objects of interest. This approach ensures that the point prompts are representative of the objects to be segmented, providing valuable localization cues to the model.

To encode these point prompts, we explore the use of 3D convolutional operations. This approach allows us to transform the 3D coordinates of the prompts into spatially-aware features that incorporate local neighborhood information. By applying 3D convolutions on the prompt coordinates, we can capture the intricate patterns and relationships between points in the structured 3D space.

During inference, we strategically sample point prompts within the region of interest (RoI) surrounding the vehicle, ensuring coverage of relevant areas for autonomous driving applications, rather than relying on user-provided prompts. To balance computational efficiency and segmentation performance, we explore various sampling strategies and prompt densities, carefully controlling the number of prompts to provide sufficient localization cues while avoiding excessive computational burden.

By effectively encoding the point prompts, our model can leverage the localization cues provided by the prompts to guide the segmentation process.

4.2.3 Mask Decoder

The mask decoder is the core component responsible for integrating the encoded point cloud features and the prompt embeddings, ultimately generating accurate segmenta-

tion masks for arbitrary objects within the 3D point cloud. This module plays a pivotal role in bridging the gap between the two distinct representations and leveraging their complementary information to produce comprehensive segmentation results.

The mask decoder utilizes a transformer-based architecture with self-attention layers to capture long-range dependencies between the encoded point features and prompts. This enables the model to learn discriminative representations by attending to relevant context within the point cloud and prompts. Additionally, cross-attention layers integrate the point cloud features and prompt embeddings, allowing the model to guide the segmentation process based on localization cues provided by the prompts. The point cloud embedding is iteratively updated through cross-attention from the prompt embeddings, incorporating localization information from the prompts into the point cloud representation.

To generate the final segmentation masks, the updated point cloud embedding is upsampled to the original point cloud resolution. This upsampling process involves leveraging the decoder component of the pre-trained point cloud transformer architecture. By utilizing the same decoder structure, we ensure compatibility between the learned representations and the upsampling operation, enabling a seamless reconstruction of the segmentation masks at the original point cloud resolution.

4.3 Training Procedure

The training of PointSAM is divided into two stages: a preliminary stage and a pretraining stage.

In the preliminary stage, we train PointSAM on a small subset of the synthetic data from CARLA. This stage serves as a proof of concept, allowing us to verify the functionality of the proposed architecture and ensure that the model can effectively segment arbitrary objects within the point clouds. We employ a simple training setup with a limited number of epochs and a smaller batch size, and use a binary cross-entropy loss to optimize the model.

After validating the model's performance in the preliminary stage, we proceed to the pretraining stage, where PointSAM is trained on the entire synthetic dataset generated from CARLA. This stage aims to leverage the vast diversity and complexity of the synthetic data to enable the model to learn generalizable representations and segmentation capabilities. During this stage, we optimize the model using a linear combination of losses, including segmentation loss functions (e.g., cross-entropy loss, dice loss [43]) and regularization terms to encourage smooth and consistent segmentation masks. We also employ a more extensive training setup, with a larger number of epochs, a larger batch size, and carefully tuned hyperparameters.

To comprehensively evaluate the performance of PointSAM, we adopt a rigorous testing procedure. Initially, we assess the model's generalization capabilities on 10% of the CARLA dataset that was held out during pretraining. This step allows us to gauge the model's ability to segment unseen synthetic data. Subsequently, we evaluate PointSAM on the nuScenes [44] dataset, a widely-used real-world autonomous driving dataset capturing challenging urban environments. This evaluation is conducted with-

out any fine-tuning or adaptation, providing insights into the model's out-of-the-box performance on real-world data. Finally, we fine-tune the pretrained PointSAM model on the nuScenes training set, leveraging the real-world data distribution to further refine the model's segmentation capabilities. The fine-tuned model is then evaluated on the nuScenes test set, allowing us to assess its performance on unseen real-world data after adaptation.

4.4 Evaluation Metrics

To comprehensively assess the performance of PointSAM in class-agnostic segmentation of 3D point clouds, we employ two complementary evaluation metrics: mean Intersection over Union (mIoU), and precision/recall [35].

- **Mean Intersection over Union (mIoU):** The Intersection over Union (IoU) is a widely adopted metric for evaluating segmentation tasks, as it quantifies the overlap between predicted and ground truth segmentations. For the class-agnostic setting, we calculate the mean IoU across all instances, regardless of class labels. The formula for mIoU is given by:

$$\text{mIoU} = \frac{1}{N} \sum_{i=1}^{N} \frac{|P_i \cap GT_i|}{|P_i \cup GT_i|} \tag{1}$$

  Here, N represents the total number of instances, while P_i and GT_i denote the predicted and ground truth point sets for the i-th instance respectively.

- **Precision / Recall:** Precision measures the ratio of correctly predicted segments to all predicted segments, while recall measures the ratio of correctly predicted segments to all ground truth segments providing a comprehensive understanding of the model's performance. Instances in which the IoU between the predicted point cloud segment and the ground truth point cloud segment exceeds 50% are considered as true positive detections.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{3}$$

In addition to quantitative metrics, we perform qualitative evaluations by visually inspecting the predicted segmentation masks overlaid on the input point clouds. This qualitative assessment allows for a comprehensive understanding of the model's performance, aiding in the identification of potential areas for improvement or further optimization.

By employing these evaluation metrics, we can thoroughly assess the proposed architecture's performance for class-agnostic segmentation of arbitrary objects in 3D point clouds. The results of these evaluations provide valuable insights into the model's strengths, limitations, and potential areas for refinement.

4.5 Time Plan

Stage	Timeline (Semester/Week)
Preliminary research and proposal writing	S1/W1 - S1/W8
Preliminary data collection and preprocessing	S1/W9 - S1/W12
Preliminary model development, training, and evaluation	S1/W13 - S1/W16
Review and refinement of preliminary model	S2/W1 - S2/W2
Data collection and preprocessing for pretraining (full dataset)	S2/W3 - S2/W6
Pretraining of PointSAM on synthetic data	S2/W7 - S2/W8
Evaluating pretrained model on synthetic and real-world data	S2/W9 - S2/W10
Fine-tuning of PointSAM and comparative analysis	S2/W11 - S2/W13
Thesis writing and assembly	S2/W6 - S2/W14

Table 1: Proposed Time Plan

References

[1] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3d point clouds: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 12, pp. 4338–4364, 2021.

[2] 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 (CVPR)*, July 2017.

[3] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[4] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollar, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

[5] M. Weinmann, *Reconstruction and Analysis of 3D Scenes*. Cham: Springer International Publishing, 2016.

[6] Shvicha, "Point cloud of yandex sdg proprietary lidar." https://commons.wikimedia.org/w/index.php?curid=115546168, 2021.

[7] S. Park, S. Ju, M. H. Nguyen, S. Yoon, and J. Heo, "Automated Point Cloud Registration Approach Optimized for a Stop-and-Go Scanning System," *Sensors*, vol. 24, p. 138, Jan. 2024. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[8] B. Höfle and M. Rutzinger, "Topographic airborne LiDAR in geomorphology: A technological perspective," *Zeitschrift für Geomorphologie, Supplementary Issues*, vol. 55, pp. 1–29, Apr. 2011.

[9] M. Dassot, T. Constant, and M. Fournier, "The use of terrestrial LiDAR technology in forest science: application fields, benefits and challenges," *Annals of Forest Science*, vol. 68, pp. 959–974, Aug. 2011.

[10] I. Puente, H. González-Jorge, J. Martínez-Sánchez, and P. Arias, "Review of mobile mapping and surveying technologies," *Measurement*, vol. 46, p. 2127–2145, Aug 2013.

[11] T. Agarwal, "LIDAR System (Light Detection And Ranging) Working and Applications — elprocus.com." https://www.elprocus.com/lidar-light-detection-and-ranging-working-application/. [Accessed 27-03-2024].

[12] T. Luhmann, S. Robson, S. Kyle, and J. Boehm, *Close-range photogrammetry and 3D imaging*. Walter de Gruyter GmbH & Co KG, 2023.

[13] T. Schenk, "Introduction to photogrammetry," *The Ohio State University, Columbus*, vol. 106, no. 1, 2005.

[14] F. Remondino and S. El-Hakim, "Image-based 3d modelling: a review," *The photogrammetric record*, vol. 21, no. 115, pp. 269–291, 2006.

[15] M. J. Westoby, J. Brasington, N. F. Glasser, M. J. Hambrey, and J. M. Reynolds, "'structure-from-motion'photogrammetry: A low-cost, effective tool for geoscience applications," *Geomorphology*, vol. 179, pp. 300–314, 2012.

[16] G. Torres, "Photogrammetry vs. lidar: What sensor to choose for a given application," Jan 2023.

[17] F. Remondino, M. G. Spera, E. Nocerino, F. Menna, and F. Nex, "State of the art in high density image matching," *The photogrammetric record*, vol. 29, no. 146, pp. 144–166, 2014.

[18] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*, pp. 1–16, PMLR, 2017.

[19] A. H. B. Zaydie, M. Y. H. Low, and W. Lin, "Perception pipeline of autonomous mobile robots using 3d point cloud classification," in *2023 IEEE International Conference on Service Operations and Logistics, and Informatics (SOLI)*, pp. 1–7, IEEE, 2023.

[20] R. Zhang, Y. Wu, W. Jin, and X. Meng, "Deep-learning-based point cloud semantic segmentation: A survey," *Electronics*, vol. 12, no. 17, p. 3642, 2023.

[21] G. Neuhold, T. Ollmann, S. Rota Bulo, and P. Kontschieder, "The mapillary vistas dataset for semantic understanding of street scenes," in *Proceedings of the IEEE international conference on computer vision*, pp. 4990–4999, 2017.

[22] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE international conference on computer vision*, pp. 945–953, 2015.

[23] A. Boulch, J. Guerry, B. Le Saux, and N. Audebert, "Snapnet: 3d point cloud semantic labeling with 2d deep segmentation networks," *Computers & Graphics*, vol. 71, pp. 189–198, 2018.

[24] D. Maturana and S. Scherer, "Voxnet: A 3d convolutional neural network for real-time object recognition," in *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 922–928, IEEE, 2015.

[25] G. Riegler, A. Osman Ulusoy, and A. Geiger, "Octnet: Learning deep 3d representations at high resolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3577–3586, 2017.

[26] 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*, pp. 652–660, 2017.

[27] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.

[28] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[29] H. Zhao, L. Jiang, J. Jia, P. H. Torr, and V. Koltun, "Point transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 16259–16268, 2021.

[30] 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.

[31] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-bert: Pre-training 3d point cloud transformers with masked point modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19313–19322, 2022.

[32] X. Lai, J. Liu, L. Jiang, L. Wang, H. Zhao, S. Liu, X. Qi, and J. Jia, "Stratified transformer for 3d point cloud segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8500–8509, 2022.

[33] D. Robert, H. Raguet, and L. Landrieu, "Efficient 3d semantic segmentation with superpoint transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17195–17204, October 2023.

[34] J. Chen, Z. Kira, and Y. K. Cho, "Lrgnet: learnable region growing for class-agnostic point cloud segmentation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2799–2806, 2021.

[35] D. Gyawali, J. Zhang, and B. B. Karki, "Region-transformer: Self-attention region based class-agnostic point cloud segmentation," *arXiv preprint arXiv:2403.01407*, 2024.

[36] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[37] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[38] 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.

[39] Y. Yang, X. Wu, T. He, H. Zhao, and X. Liu, "Sam3d: Segment anything in 3d scenes," *arXiv preprint arXiv:2306.03908*, 2023.

[40] R. Huang, S. Peng, A. Takmaz, F. Tombari, M. Pollefeys, S. Song, G. Huang, and F. Engelmann, "Segment3d: Learning fine-grained class-agnostic 3d segmentation without manual labels," *arXiv preprint arXiv:2312.17232*, 2023.

[41] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "Pct: Point cloud transformer," *Computational Visual Media*, vol. 7, pp. 187–199, 2021.

[42] X. Wu, L. Jiang, P.-S. Wang, Z. Liu, X. Liu, Y. Qiao, W. Ouyang, T. He, and H. Zhao, "Point transformer v3: Simpler, faster, stronger," *arXiv preprint arXiv:2312.10035*, 2023.

[43] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*, pp. 565–571, Ieee, 2016.

[44] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11621–11631, 2020.