Spatial and Channel Attention-based 3D Object Classification Research for 3D Point Clouds

by

Xikai Tang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2022

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Deep learning has been widely used in Two Dimensional (2D) computer vision and has led to the realization that machine learning techniques have become one of the key research directions for future scientific research. In 2D computer vision, CNN[49], RNN[34], SENet[40], Transformer[89], as well as many other algorithms show amazing results in 2D data. With the accelerating development of computer version technologies, the exploitation of 2D data is insufficient for machine learning research and researchers considering the transfer of 2D computer vision algorithms to Three Dimensional (3D) domain. Point clouds is an important expression of 3D data. The more detailed information found in 3D point cloud data compared to 2D point cloud data, it has accelerated research in recent years, which has led to significant breakthroughs in artificial intelligence, deep learning, autonomous driving, tracking, and other domains. There have been a large number of deep learning methods recently proposed based on point clouds. PointNet[72], P4Transformer[21], and SampleNet[47] show significant success in 3D domain. Disorder and sparse shape make a challenge in designing deep neural networks for point clouds processing.

In chapter one, we will introduce the background of point clouds, the existing public datasets and evaluation metrics, then investigate and analyze deep learning methods based on classification of point clouds. In chapter two, we will introduce generation of point clouds and analyse the existing methods based on classification and segmentation. Furthermore, we investigate attention mechanism in computer vision, includes background of attention mechanism, evolution of attention mechanism, spatial and channel attention in vision and point cloud-based attention model in deep learning. Based on the chapter one and two analyse and investigation, we found that this data type's ability to provide depth information, point sparsity and disorder pose a challenge in designing appropriate deep neural networks to process them and it is still challenging to explore local relationships in point clouds data. so, in chapter three, in order to better extract features and obtain geometric information we will propose a point attention (PointAT) model and propose attention value (AT value) model for feature fusion to apply geometric relationship to the data. Then, we propose a new spatial and channel attention-based network (SCA). The SCA is the overall structure of the network, and the main purpose is to connect PointAT and AT value model, then capturing meaningful geometric information by applying the geometric relationship between point clouds patches to the model, then propose an auto pooling framework to extract global features. In this work, we concentrate on learning geometric relationship between point cloud data. For this purpose, we introduce a point attention model based on spatial and channel attention to learn the geometric relationship

between point clouds, and further combine the geometric relationship with the point cloud data by the AT Value Model. Finally, we introduce an adaptive downsampling structure, Autopooling. This downsampling structure considers each point's importance weight and picking key points adaptively, which can be used with convolutional networks. Extensive experiments conducted on two benchmark datasets (ModelNet40[96] and ShapeNet[11]) clearly demonstrate the effectiveness of our SCA and SCA-Auto (SCAA with Auto pooling) methods. Finally, in chapter four, we summary our contribution, and significant of study findings and limitations of proposed methods. Then, we get future research directions based on our analyse and investigation.

Acknowledgements

I would like to thank those who helped me during my Master's study at the University of Waterloo. First, I would like to gratefully thank my supervisor Professor Dayan Ban, as well as Zhou Wang, who offered me the chance to study at the University of Waterloo and supported me during my research. This work is supported by Natural Science and Engineering Research Council (NSERC) of Canada and AIH Technology Inc. Second, I would like to again extend my sincere thanks to professor Dayan Ban and Zhou Wang, who gave me research suggestions and supported me by sharing their lab resource. Third, I would like to thank all group members of Prof. Dayan's group for sharing idea with me. Finally, I would like to thank my parents. The completion of my thesis would not be possible without their moral and spiritual support.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

2D Two Dimensional iii, 1, 5, 8–12, 15, 17, 18, 20–26, 28, 29, 44, 46, 47

3D Three Dimensional iii, 1–6, 8–15, 17, 18, 20–24, 26, 28, 37, 38, 40, 43–47

Acc. Overall Accuracy 37

AP Average Precision 2

CAD Computer Aided Design Software 5, 37

CMOS Sensor Complementary Metal-Oxide-Semiconductor Sensor 16

CNNs Convolutional Neural Network 25

ER Error Rate 2

F1-Score F1-Score 2

GRU Gate Recurrent Unit 28

IoU Intersection over Union 2

LADAR Laser Radar 14

LiDARs Light Detection and Ranging 14, 15

LSTM Long Short Term Memory networks 28

mAcc Mean Class Accuracy 2

mAP Mean Average Precision 2, 5

mIoU Mean Intersection over Union 2, 5, 41

ML Mostly Lost 2

MMML Multi-modal Machine Learning 12, 47

MOTA Multiple Object Tracking Accuracy 2

MOTP Multiple Object Tracking Precision 2

MT Mostly Tracked 2

NLP Nature Language Processing 24, 25

OA Overall Accuracy 2, 5

Pixel Accuracy (Global Acc) Pixel Accuracy (Global Acc) 5

RGB Pure, Red, Green, and Blue 6, 15

RGB-D RGB + Depth Map 14

RNNs Recurrent Neural Networks 25, 28

SGD Stochastic Gradient Descent 37

SOTA State of the Art 9, 23, 25, 30

TOF Time-Of-Flight 15, 16

Chapter 1

Introduction

1.1 Background

1.1.1 Purpose and Objective

With the accelerating development of computer version technologies, there are various deep learning methods for 2D data that are widely used. However 2D computer vision models lack the capacity to fully capture the structure they aim to represent. Research that considers transferring 2D computer vision algorithms to 3D fields rectify this lack. This data type can provide depth information; unfortunately, point sparsity and disorder pose a challenge when designing appropriate deep neural networks to process them. In addition to this, it is still challenging to explore local relationships in point cloud data. The purpose of this thesis is to examine the effect geometric information between point cloud data has on the model and to perform classification and segmentation tasks on the data. This thesis provides an overview of the current state of research and development regarding point cloud data under deep learning and summarizes the point cloud public dataset.

1.1.2 Public Datasets

With the emphasis on point cloud and deep learning, as well as the improvement of the data collection research, it is possible to use a large amount of point cloud public data. Based on the point cloud method for deep learning, this article is divided into three different

sections,datasets for classification, detection, tracking and segmentation (see Table 1.1), Data sources are summarized in Table 1.2.

Point cloud-based classification is divided into 3 branches: 3D point cloud, point cloud action recognition, and point cloud pose estimation. There are real 3D scenes, non-rigid objects, and rigid object datasets types, as shown in Table 1.1. Real 3D scene datasets contain a variety of common objects with different levels, and enviroment noise which make these datasets more difficult to pre-process. Rigid object datasets and non-rigid object datasets are synthetic datasets, and there is no environmental nor background noise. The non-rigid object datasets are more fine-grained, and have more point details and more boundaries for object depiction compared to rigid object datasets, which have a defined shape and proper particle orientation.

For point cloud-based detection and tracking in Table 1.1, there are three types of scenes features: driving scenes, room scenes, and indoor scenes. The driving scene datasets are used for autonomous driving, which includes the vehicle data suite. The room scene datasets are collected from the room. The indoor scene datasets include large-scale indoor areas, like the classroom and the office.

For point cloud-based segmentation in Table 1.1, the scene features under this category are divided into several subcategories: indoor scenes, outdoor scenes, and urban scenes. The urban scenes are sparse data; Unlike outdoor scenes, indoor scenes have corresponding depth and segmentation maps.

1.1.3 Evaluation Metrics

Different sections have different evaluation metrics, as shown in Table 1.3. For point cloud classification, Overall Accuracy (OA), Mean Class Accuracy (mAcc), MIoU, Error Rate (ER), and Mean Average Precision (mAP) are frequently used evaluation metrics. OA is the accuracy for the whole data set (irrespective of category), while MAcc is calculated for each category and then averaged. Error rate describes the percentage of misclassification by the classifier. Intersection over Union (IoU) measures the degree of overlap between the prediction frame and the true frame. Average Precision (AP) is the area under the curve drawn using different combinations of precision and recall points. MAP is the average of AP values for all categories. For point cloud detection and tracking, common evaluation metrics include MAP, Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), Mostly Tracked (MT), Mostly Lost (ML), and F1-Score (F1-Score). These metrics are used to evaluate the effectiveness of the model. For point cloud

Table 1.1: A summary of Existing Datasets for Point Cloud

Classification Dataset

3D Point Cloud Dataset

Datasets Name	Year	Samples	Classes	Training	Test	Object Feature
Sydney Urban Objects [17]	2013	588	14	-	-	Real 3D Scenes
ModelNet10 [95]	2014	4899	10	3991	605	Rigid Object
ModelNet40 [95]	2014	12311	40	9843	2468	Rigid Object
ShapeNet [11]	2015	51190	55	-	-	Non-rigid Object
SHREC15 [50]	2015	1200	50	-	-	Non-rigid Object
ScanObjectNN [88]	2019	2902	15	2321	581	Real 3D Scenes

Point Cloud Action Recognition Dataset

Datasets Name	Year	Samples	Classes	Training	Test	Object Feature
MSR-Action3D [53]	2010	567	20	-	-	Rigid Object
SUN RGB-D [83]	2015	10335	37	5285	5050	Real 3D Scenes
NTU RGB+D 60 [78]	2016	56880	60	40320	16560	Real 3D Scenes
ScanNet [16]	2017	12283	17	9677	2606	Real 3D Scenes
NTU RGB+D 120 [55]	2019	114,480	120	-	-	Real 3D Scenes

Point Cloud Pose Estimation Dataset

Datasets Name	Year	Samples	Classes	Training	Test	Object Feature
HPS [31]	2021	300K	-	-	-	Real 3D Scenes

Detection and Tracking Dataset

Datasets Name	Year	Samples	Classes	Training	Test	Scenes Feature
KITTI [25]	2012	80256	8	7481	7518	Driving
SUN RGB-D [83]	2015	10335	37	5285	5050	Room
S3DIS [2]	2016	-	13	-	-	Indoor & Room
ScanNetV2 [16]	2017	-	18	-	-	Room
RBO [64]	2018	-	–	-	-	Room
Argoverse [12]	2019	323557	15	205942	78143	Driving
H3D [71]	2019	-	8	-	-	Driving
Waymo Open [85]	2019	-	4	-	-	Driving
nuScenes [10]	2019	40157	10	28130	6008	Driving
Lyft L5 [38]	2020	-	9	-	-	Driving
CODD [3]	2021	-	-	-	-	Driving

Segmentation Dataset

Datasets Name	Year	Points Samples	Classes	Type	Scenes Feature	Sensors
Synthia [77]	2016	-	13	-	Outdoor	-
S3DIS [2]	2016	273M	20	Static point cloud	Indoor	Matterport
ScanNet [16]	2017	768000M	13	RGB-D	Intdoor	RGB-D
Semantic3D [32]	2017	4009M	8	Static point cloud	Outdoor	TLS
SemanticKITTI [8]	2019	4549M	28	sequential point cloud	Outdoor	MLS
Toronto-3D [87]	2020	78.3M	8	Static point cloud	Urban	MLS
SemanticPOSS [69]	2020	216M	14	sequential point cloud	Outdoor	-
SensatUrban [42]	2020	2847M	13	Static point cloud	Urban	UAV

Table 1.2: A Summary of Existing Datasets for Point Cloudn

Classification Dataset	
3D Point Cloud Dataset	
Datasets Name	Data Sources
Sydney Urban Objects [17]	https://www.acfr.usyd.edu.au/papers/SydneyUrbanObjectsDataset.shtml
ModelNet10 [95]	https://modelnet.cs.princeton.edu/
ModelNet40 [95]	https://modelnet.cs.princeton.edu/
ShapeNet [11]	https://shapenet.org/
SHREC15 [50]	https://www.cs.cf.ac.uk/shaperetrieval/shrec15/
ScanObjectNN [88]	https://hkust-vgd.github.io/scanobjectnn/
Point Cloud Action Recognition Dataset	
Datasets Name	Data Sources
MSR-Action3D [53]	https://sites.google.com/view/wanqingli/data-sets/msr-action3D
SUN RGB-D [83]	https://rgbd.cs.princeton.edu/
NTU RGB+D 60 [78]	https://rose1.ntu.edu.sg/dataset/actionRecognition/
ScanNet [16]	http://www.scan-net.org/
NTU RGB+D 120 [55]	https://rose1.ntu.edu.sg/dataset/actionRecognition/
Point Cloud Pose Estimation Dataset	
Datasets Name	Data Sources
Human POSEitioning System [31]	http://virtualhumans.mpi-inf.mpg.de/hps/
Detection and Tracking Dataset	
Datasets Name	Data Sources
KITTI [25]	http://www.cvlibs.net/datasets/kitti/
SUN RGB-D [83]	https://rgbd.cs.princeton.edu/
S3DIS [2]	http://buildingparser.stanford.edu/dataset.html
ScanNetV2 [16]	http://www.scan-net.org/
RBO [64]	https://zenodo.org/record/1036660/#.Yostf-zMJhE
Argoverse [12]	https://www.argoverse.org/
H3D [71]	https://usa.honda-ri.com/h3D
nuScenes [10]	https://www.nuscenes.org/
Lyft L5 [38]	https://level-5.global/
Waymo Open [85]	https://waymo.com/open/
CODD [3]	https://github.com/eduardohenriquearnold/CODD
Segmentation Dataset	
Datasets Name	Data Sources
ScanNet [16]	http://www.scan-net.org/
S3DIS [2]	http://buildingparser.stanford.edu/dataset.html
Synthia [77]	https://www.v7labs.com/open-datasets/synthia-dataset
Semantic3D [32]	https://www.semantic3D.net/
SemanticKITTI [8]	http://www.semantic-kitti.org/
Toronto-3D [87]	https://github.com/WeikaiTan/Toronto-3D
SemanticPOSS [69]	http://www.poss.pku.edu.cn/
SensatUrban [42]	https://github.com/QingyongHu/SensatUrban

Table 1.3: A Summary of Evaluation Metrics for Point Cloud Datasets Classification, Detection and Tracking, and Segmentation

Evaluation Metrics						
Point Cloud Classification	Overall Accuracy(OA)	mean Class Accuracy (mAcc)	mean Intersection Over Union (mlou)	Error Rate	mean Average Precision (mAP)	-
Point Cloud Detection & Tracking	Multiple Object Tracking Accuracy (MOTA)	Multiple Object Tracking Precision (MOTP)	Mostly Tracked (MT)	Mostly Lost (ML)	mean Average Precision (mAP)	Fl -Score
Point Cloud Segmentation	Pixel Accuracy (Global Acc)	mean Accuracy(mA)	mean Intersection over Union (mloU)	Overall Accuracy(OA)	-	-

segmentation, Pixel Accuracy (Global Acc) (Pixel Accuracy (Global Acc)), MAP, MIoU and OA are the most commonly used evaluation metrics.

1.2 Point Clouds

1.2.1 Point Cloud in Theory

We first briefly introduce relationship between point clouds and 3D images. A 3D image is a special form of information expression, and its features are expressed as data in 3Dof space. These expressions include: depth maps (expressing the object's distance from the camera in grayscale), geometric models (created by Computer Aided Design Software (CAD)), and point cloud models (all reverse engineering devices take objects and make them sample into point clouds). Compared to 2D images, 3D images can achieve natural object background decoupling due to the additional dimension of information. Point cloud data is the most common and basic 3D model. Point cloud models are often obtained directly from measurements, with each point matching a measured point without any processing measures, Therefore, they contain the largest amount of information. The information hidden in point clouds needs to be extracted by other extraction methods; the process of extracting information in point clouds is 3D image processing.

A point cloud is a data structure usually used to represent 3D geometries. It is a massive set of points in spatial coordinates, directly represented by extracting 3D information from a stereo-vision camera, as well as a depth map representation that can be generated using RGB-D [1]. Figure 1.1 is point cloud of an airplane.

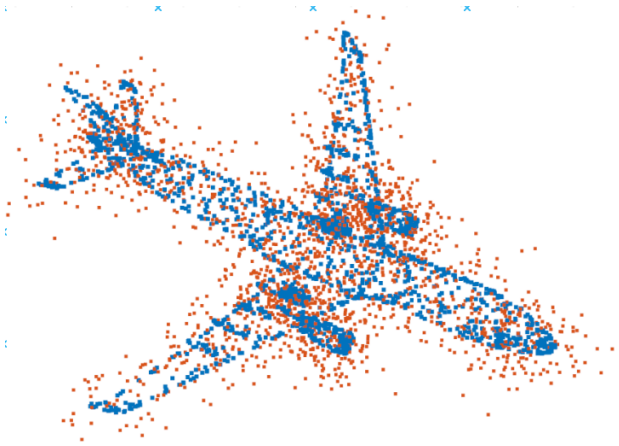

Figure 1.1: Point Cloud of an Airplane [95]

.

1.2.2 Point Cloud Characteristic

Point cloud data is different from traditional Pure, Red, Green, and Blue (RGB) images in that the data is not regularly arranged. It has two important characteristics: (1) displacement invariance (2) rotation invariance [72].

Displace invariance implies that the order of storage in the point cloud is independent of the point cloud, and disrupting the order of each row does not affect an original point cloud [72]. This feature can be found by dropping certain rows of the point cloud data in order, as shown in Figure 1.2, 1.3, 1.4, 1.5, 1.6, 1.7. It can be seen that deleting point cloud data sequentially does not delete a certain continuous region of the point cloud, but instead makes the original point cloud sparse.

Rotational invariance refers to how in 3D space, point cloud data is 3D, and 3D coordinate rotation does not change the structure or form of the original data. Therefore, two important characteristics of point cloud data arise: (1) displacement invariance (2) rotation invariance. These two characteristics determine that algorithms cannot be designed in the same way as traditional images.

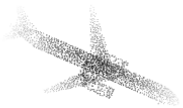

Figure 1.2: Remaining number of point clouds after deleting half of them in order: 8192

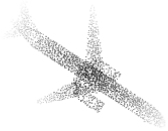

Figure 1.3: Remaining number of point clouds after deleting half of them in order: 4096

Figure 1.4: Remaining number of point clouds after deleting half of them in order: 2048

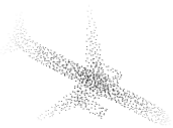

Figure 1.5: Remaining number of point clouds after deleting half of them in order: 1024

Figure 1.6: Remaining number of point clouds after deleting half of them in order: 512

Figure 1.7: Remaining number of point clouds after deleting half of them in order: 256

1.3 Object Classification

1.3.1 3D Object Classification on Point Clouds

Currently, 2D classification methods have become very prosperous. For example, ViT [19] has achieved excellent results by applying the Transformer method for classification by using the attention mechanism. Point cloud classification becomes challenging, due to the unstructured form of the data and the variation in the amount of data. [30].

The regular pipeline for point cloud-based classification involves extracting features from point cloud datasets to get the global features. By embedding the global features with several classifiers, as shown in Figure 1.8, labels can be achieved. In the early deep learning tasks for point clouds, models and methods are implemented based on classification tasks and priority. For example, MVCNN [84] performs the classification task by composing data from multiple views and performing aggregation of the data under each view.

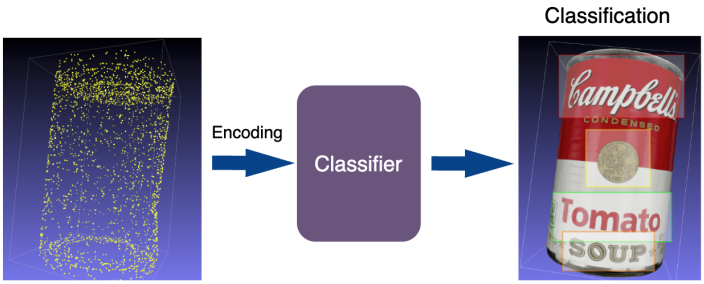

Figure 1.8: Regular Pipeline of Point Cloud Classification[9].

1.3.2 3D Object Classification Methods on Point Clouds

Based on Two Dimensional Methods

SqueezeSeg [93] was proposed by Bic hen Wu et al,. in 2017. A lightweight convolutional neural network was used to achieve real-time semantic, instance classification of 3D objects in point clouds. In order to easily process the 2D convolutional neural network, the point cloud data is first projected to obtain the front view, then the input image is feature extracted and using the SqueezeSeg [93] convolutional network, the output is optimized using the conditional Random Field (CRF) . Geometric shared network (GS-Net) [101]

can specifically and efficiently learn point descriptors, reduce losses due to image transformations, and improve the robustness of the model. The GSC module in GS-Net can effectively capture both local and global geometric features, making the model significantly more effective in classification tasks. The VIASeg [113] model, in which the authors convert color information to data-level embedded in point cloud data, was designed so that the colour information of the image provides rich visual information and improves the classification performance. This approach is a multi-modal (visual + point cloud) combination approach. LU-Net [48] extracts high-level 3D features for each point of the point cloud , then projects these features to a 2D multi-channel front view. This model benefits from high level 3D feature extraction and embeds 3D local features into the 2D image, which allows the model to be used effectively. RangeNet++ [65] was proposed by Andres Milioto et al,. in 2019. This model projects the 3D data to a 2D front view. This model has high classification performance on point cloud data and obtains the full label of the original point cloud.

Based on 3D Methods

JSNet [110] extracts features from the original point cloud data In order to improve the feature discrimination, a feature fusion method is incorporated to perform feature fusion at backbone for different levels. Using the joint instance semantic segmentation module, semantic features are subsequently embedded in the space and feature fusion is performed to improve feature robustness. Furthermore, the model aggregates the fused features into the feature space to improve the classification task. MPNet [82] was proposed by Tong He et al,. in 2020. Due to the complex 3D structure of point cloud data and the irregular distribution of point cloud data it become more difficult to process the distribution and learning of point cloud data, MPNet [82] utilizes a memory module to achieve small batch processing. Each small batch of samples is learned and memorized to achieve memory enhancement, to reduce the problem of category imbalance. SceneEncoder [98] was proposed as a scene encoding module to achieve enhanced global information. This module can filter categories that do not belong and perform classification tasks. A regional collinearity loss was designed to approach features and neighboring points of the same label to improve feature recognition. The SceneEncoder [98] achieved State of the Art (SOTA) experimental effects.

1.4 Significance and Contributions

As we can see in this subsection, because of the sparsity and irregularity of the point cloud data, CNN in 2D does not work directly on the 3D point cloud data, which requires the point cloud data to be transformed first. SqueezeSeg [93], Geometric Shared Network (GS-Net) [101], VIASeg [113], LU-Net [48] and RangeNet++ [65] project 2D point cloud data to the 2D plane, and use different methods to improve the projection. Thus, it can be inferred that the methods in the projection of 3D data to 2D will be more diverse, which is an interesting and intuitive idea since the projection process leads to loss of information. JSNet [110], MPNet [82], SceneEncoder [98], which are all based on 3D data, the point cloud data is directly input to the model and the feature information is extracted in 3D space. In this process, the point clouds data can be helped by adding feature fusion and attention mechanism. In future work, the missing information due to sparsity of point cloud data can be complemented by incorporating additional methods to provide richer information.

Point cloud data in its 3D representation is now widely used and referenced in various areas with great importance. Point cloud data can play an important role in the exploration of 3D space. In point cloud classification tasks, it provides important information for human-related tasks, the data it provides important information for human-related tasks, such as point cloud action recognition, P4Transformer [21], PSTNet [22], PoseNet [111], HandVoxNet [61], PVN3D [36], point cloud pose estimation PointContrast [97], HandVoxNet [61] and PVN3D [36]. In point cloud detection and tracking, Transformer3D-Det(T3D) [109], VoteNet [18], Channel-wise Transformer 3D Object Detection (CT3D) [79], Group-Free [59], 3DETR [66], Semantic Point Generation (SPG) [102],Range-Guided Cylindrical Network [74], RangeDet [23],PV-RCNN [81] and CenterPoint [103] can be applied to automatic driving, tracking, etc. In point cloud segmentation, SqueezeSeg [93], Geometric Shared Network (GS-Net) [101], VIASeg [113], LU-Net [48], RangeNet++ [65], JSNet [110], MPNet [82], and SceneEncoder [98] perform the segmentation task well. These models perform more accurate extraction of global to local features. Main methods include adding attention mechanism, adding more information to the original point cloud data, and mining sequence information in spatial and temporal dimensions.

Point cloud data can present objects well in 3D , which contains detailed information. Therefore, methods to mine information becomes critical. PointContrast [97] and MPNet [82] improved the model effect by extracting local features. We believe this is a fruitful and useful research direction. In the process of mining local information, we can avoid losing global information after extracting local features by performing joint local and global feature extraction.

3D point cloud data has been of great use for researchers to study 3D real information; however, due to non-dense expressions of point cloud data, which present a sparse state, there is no information in the surrounding areas of points. This has become a major area for research. By filling the missing information in the non-dense representation of point cloud, we think this is a very worthy research direction. Few studies have been conducted focusing on supplementing point cloud data with missing information in this context here. In HandVoxNet [61], researchers add depth maps to supplement the missing information of point cloud data. In Semantic Point Generation (SPG) [102], semantic data and original point cloud data are fused for data tolerance to reproduce the original point cloud missing data. VIASeg [113] embeds colour information into the point cloud data to provides rich visual information to the point cloud data

Data processing in 2D images have been very diverse, including data enhancement, data fusion, data inversion, and data transformation. There has been great progress in data processing for 2D data and many algorithms have been proposed. In 3D data, data processing of 3D data has also attracted attention. PointAugment [52] performed data fusion of the enhanced 2D data with 3D data. SampleNet [47] added a projection operation to transform the data and achieved a new point cloud sampling. LU-Net [48] extracted the point cloud data with high-level features and embedded 3D local features into 2D data to obtain the new point cloud data. RangeNet++ [65] researchers used projected the 2D data to get 3D data. JSNet [110] performed a fusion of different layers of backbone features using a data fusion method. The above papers have shown that data processing for 3D point cloud data is a very effective model enhancement method, increasing the robustness of the model and improving the model effectiveness by. There are still many data processing methods that are worth developing in order to obtain new data to increase the training data.

The attention mechanism is one of the most popular techniques in machine vision and has been used with amazing results in many computer vision tasks [46]. Its application in neural networks has also made it a prominent topic. In [46], researchers have observed that humans do not process a whole scene; instead a person's attention will select information the mind deems key to guide later decisions. This selective screening of the whole greatly eliminates invalid features and reduces complexity. Feature extraction can be performed accurately and efficiently for the object in the model. P4Transformer [21], Transformer3D-Det(T3D) [109], VoteNet [18], Channel-wise Transformer 3D Object Detection (CT3D) [79], Group-Free [59], 3DETR [66] applied the attention mechanism to the their respective models and achieved successful results on the whole effect of the model. From these papers, the main focus of many existing models based on attention mechanisms is on feature extraction and analysis of weights for feature tasks which has have the advantage

of improving the overall performance effect of the model in a very direct way. However, there is little attention to the problem of localization accuracy, Thus, localizing attention regions is a worthwhile study and interest point for the future. In particular, in 3D point cloud data, there is a problem of overlap in the range of attention regions, which makes it possible to design new attention models in three-point cloud data as the focus of future attention.

Many loss functions have been proposed to improve the capability of networks and further optimize models. The common loss functions are Cross-Entropy Loss, Center Loss, Triplet Loss, etc. In 3D point cloud, researchers have also designed new loss functions for point cloud data, such as PointInfoNCE Loss [97], to further better optimize the model for higher performance. The effect of different loss functions on point cloud data is a very important direction.

Point cloud models are built using two streams, 2D-based and 3D-based methods. The 2D-based method mainly converts 3D point cloud data into 2D maps for various tasks. The 3D-based approach inputs 3D point cloud data directly into the model. There are now also methods that combine 2D and 3D methods. Using fused image information to help with 3D tasks is also an interesting idea.

From this, modality can be represented as a source or expression of information. The purpose of Multi-modal Machine Learning (MMML) is to perform via machine learning thus lending to the ability to process and understand multi-modal information [6]. In machine vision, multi-modal learning can be formed between image, video, audio, and semantics. Researchers can complement point cloud data by adding information from different modalities to the point cloud data. Some researchers have tried to use fluid mechanics to help improve overall model performance. For example, by considering hydrodynamic natural flow phenomenon, AdvectiveNet[[35] processed point cloud data, PointContrast [97] added depth differences to estimate depth complementary information, and HandVoxNet [61] added depth maps for pose estimation. Multi-modal point cloud data remains a worth problem for future research as it can be used for data complementation or to help the overall model.

There is a novel approach which integrates point cloud data tasks into the pipeline of other related tasks. Many multi-channel methods are already proposed. For example, point cloud-based classification models PointContrast [97], point cloud-based pose estimation PoseNet [111], HandVoxNet [61], PVN3D [36], and point cloud-based 3D methods JSNet [110]. In the subsequent tasks, exploring how to better utilize and implement the multi-channel approach and use other tasks to better facilitate effective mining of point cloud data is an important future research direction.

1.5 Summary

In chapter one, we introduced the background of point clouds, the existing public datasets and evaluation metrics, then investigate and analyze deep learning methods based on classification of point clouds. Point cloud data in its 3D representation is now widely used and referenced in various areas with great importance. Point cloud data can play an important role in the exploration of 3D space. In chapter two, we will illustrate what is the point cloud data, the generation of point clouds, then we will analyse the point cloud-based methods by analysing classification task, segmentation task and attention in vision.

For the above research content, our research intends to focus on the following three aspects: (1) an overview study about point clouds. (2) proposing the problems faced by point clouds based on deep learning according to the research; (3) proposing solutions and model SCA according to the problems. The specific technical route is shown in 1.9.

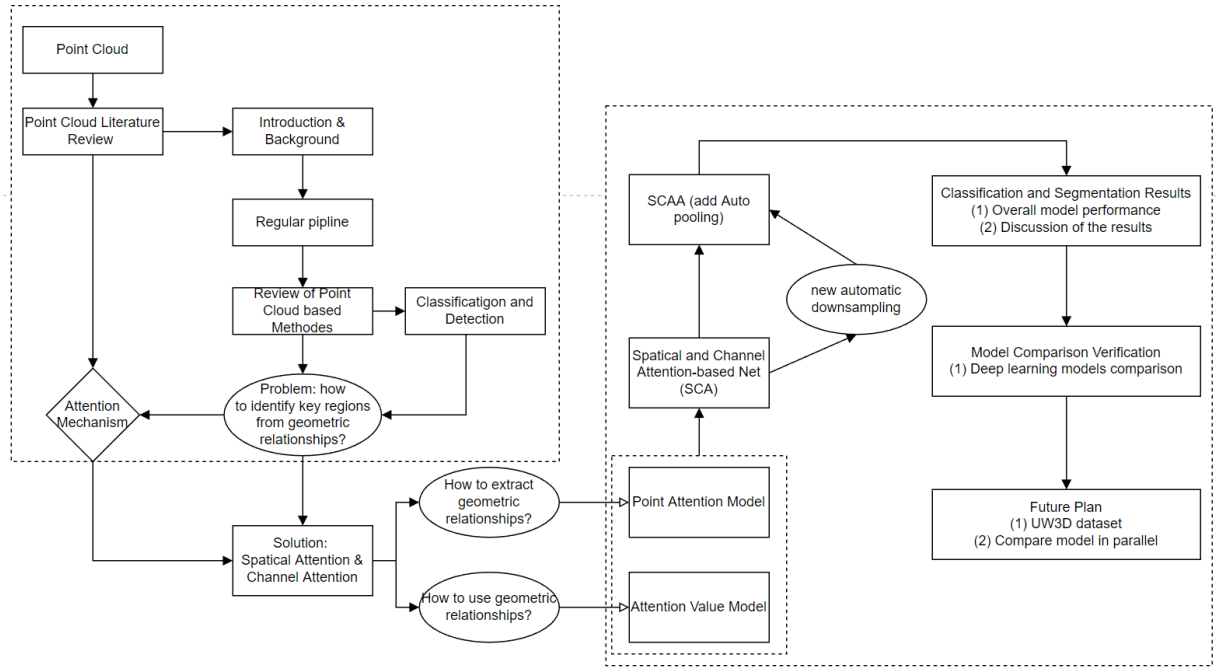

Figure 1.9: The image of technical route

.

Chapter 2

Literature Review

2.1 Generation of Point Clouds

2.1.1 LIDAR

With the emergence of cutting-edge 3D acquisition technologies, 3D sensors, such as LiDARs, and RGB + Depth Map (RGB-D) cameras [14], are becoming increasingly affordable and available. 3D data acquired by these sensors can provide rich geometric, shape, and scale information, which usually needs be represented with different formats including point cloud, mesh, volumetric and multi-view images [30]. LiDARs, also known as Laser Radar (LADAR) (Laser Detection and Ranging), is an active remote sensing device that uses lasers as the emitting light source and optoelectronic detection technology. It analyzes the reflected energy level, the amplitude, frequency, and phase of the reflected wave spectrum on the target surface by measuring the distance travelled from the laser emitted from the sensor to the target surface, and then precisely solves target information and displays accurate 3D structure information of the target. The laser point cloud data, obtained from a vehicle-mounted laser scanning system, emits laser signals to the surrounding area and collects the reflected laser signals. Through field data collection and combining navigation and point cloud solving, it is possible to calculate accurate spatial information of these points. Unlike cameras and radar, LiDARs can operate in any light condition, which is necessary for self-driving cars. Cameras, radars, and other technologies can help vehicles visualize surroundings, as shown in 2.1. However, once it is dark or raining, camera technology cannot provide the high-resolution images needed for cars to accurately see and distinguish between people and other objects, so LiDARs are still the sensor that offers

the highest range accuracy and best angular resolution.

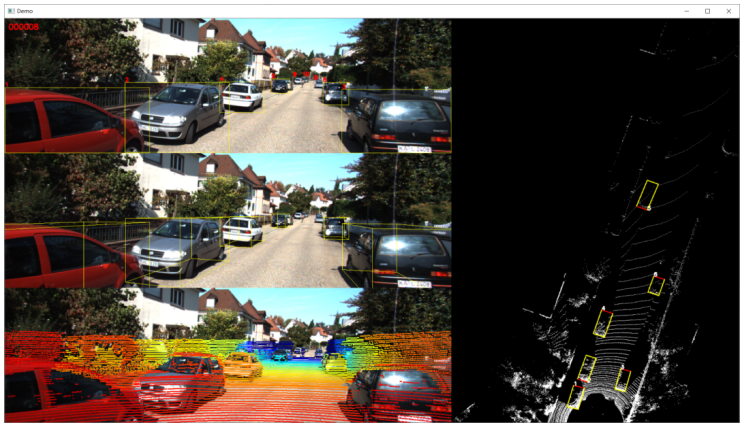

Figure 2.1: Illustration of Structured LiDARs [24].

2.1.2 Depth Camera (3D Camera)

With the rapid development of machine vision and artificial intelligence, and other technologies that use scene modeling, object recognition, and environment recognition are increasingly used. In contrast to traditional 2D cameras, depth cameras can obtain depth-of-field information by shooting space, so as to obtain target 3D information and later build 3D models. Depth cameras include Structured-light cameras, Stereo-vision cameras, and Time-Of-Flight (TOF) cameras. Structured-light cameras use a near-infrared laser to project light with certain structural characteristics to a subject, which is captured by a special infrared camera. This structured light (generally streak structured light based on the encoding pattern enshape, encoding as Mantis Vision, Realsense (F200) structured light, with a scattered structured light apple (primesense), will be collected depending on depth information of a subject. Computing units will convert the structured light pattern into depth information to obtain the 3D structure, as shown in Figure 2.2.

Stereo-vision cameras, based on parallax principles, uses an imaging device to acquire two images of an object from different locations to obtain 3D object geometry information by calculating position deviations between corresponding points of images. RGB stereo-vision cameras are very dependent on pure image feature matching, so they work very poorly in low light or over-exposure situations, and it is also difficult to extract and match features if there is a lack of texture in the subject scene itself.

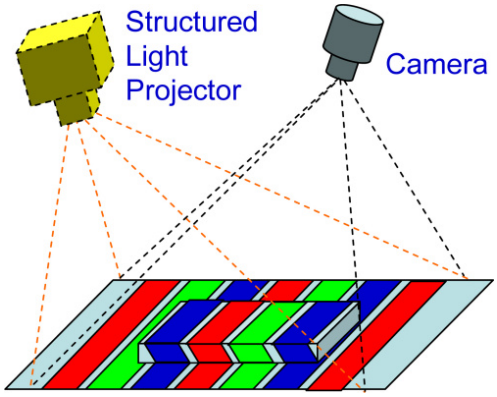

Figure 2.2: Illustrate of Structured Light [27].

TOF cameras transmit modulated light pulses through an infrared emitter, reflect them when they encounter an object, receive the reflected light pulses with a receiver, and calculate distance to an object based on round-trip time of light pulses. This modulation method requires high requirements for transmitter and receiver, and light speed is so fast that there is an extremely high accuracy requirement for time measurement. In practical applications, it is usually modulated into a pulse wave (usually a sine wave), and when it encounters an obstacle with diffuse reflection, the reflected sine wave is then received by a specially designed Complementary Metal-Oxide-Semiconductor Sensor (CMOS Sensor). When the waveform produces a phase shift, and the distance from the object to the depth camera can be calculated by the phase shiftas shown in Figure 2.3.

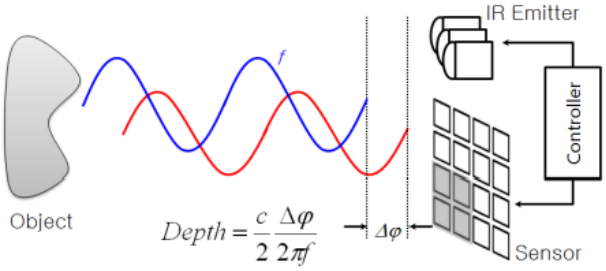

Figure 2.3: Illustrate of TOF [33].

2.2 Classification

There are three different classifications of point cloud-based methods: 3D point cloud-based methods, point cloud with action recognition methods, and point cloud with pose estimation methods. In each of these subsections, we will review some well-known deep network models that have been widely acknowledged and used repeatedly in point cloud-based deep learning, as shown in Figure 2.4.

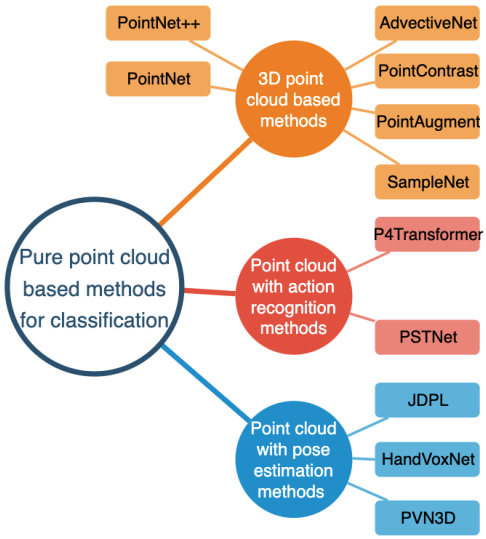

Figure 2.4: Point Cloud Based Methods for Classification.

2.2.1 3D Point Cloud Based Methods

Currently, there are several models directly consuming the raw point cloud datasets without losing information. These frameworks advance the research of 3D scene by enable better learning of feature representations for point cloud processing, such as PointWeb [107], Mo-Net [44], SRINet [86].

PointNet [72] is the first method that inputs point cloud data directly to deep learning models as shown in Figure 2.5. The paper details the theoretical basis of point cloud data, performs further analysis, and proposes the PointNet [72] model structure based on the proposed theoretical basis. The basic process of PointNet [72] is to form an $n \times 3$ 2D tensor (where n represents the number of point clouds) by taking the whole set of

point cloud data of one frequency as input to ensure the model-specific spatial transformation invariance. Global features are then extracted after feature extraction of the data by using multilayer perceptron, and classification and the segmentation task can be finally performed. PointNet [72] performs classification well by extracting global features, but local feature extraction capabilities are poor, making it difficult to analyze complex scenes. PointNet++ [73] modifies the feature extraction based on PointNet [72] to extract local features, and the deep features are extracted using a multilayer network structure, as shown in Figure 2.6. AdvectiveNet [35] applies the natural flow phenomena in fluid dynamics to the point cloud processing, and builds a new deep learning method to process point cloud. The Eulerian-Lagrangian representation of physics is introduced to implement mining particle features, giving a new perspective to view and solve the classification task of point cloud. PointContrast [97] proposes an unsupervised training method, which combines the idea of 3D representation to build a new loss function called PointInfoNCE Loss [97]. The point cloud is adjusted (rotated, scaled) and local features are extracted so that the new loss function may be constructed by contrasting matching features with non-matching features in the point cloud. PointAugment [52] considers characteristics of the 3D data space domain, and the 2D data augmentation method to obtain new samples from the original sample data. After data augmentation, new samples are sent to the classifier for training and then the classifier results are to the augmenter for training to achieve an iterative process. This is the first attempt to augment 3D data. This attempt considers the characteristics of point cloud data. SampleNet [47] perform a differentiable acquisition to approximate point cloud sampling and incorporate a soft projection operation to change the representation by using the local position weight value coordinates in the initial point cloud. [72], [73], [52] use point cloud feature to mine local feature. [35], [47] perform novel feature extraction methods to achieve more effect results. Point-BETR [104] combine point cloud and Transformer structure from natural language processing to encode points, build masked language modeling for self-supervised training, and translate point clouds into language-like words, as shown in Figure 2.7. GLRV [20] proposes an unsupervised adaptation method to modelling internal structures in point clouds and marking target samples using voting

2.2.2 Point Cloud with Action Recognition Methods

In the subsection, We introduce famous and important methods. In these methods, point clouds are used for action recognition which is a critical part of classification tasks. P4Transformer [21], proposed in 2021, uses point space-time(PST) [22] to extract features by using new convolution allowing for the encoding of the time-space structure to be effec-

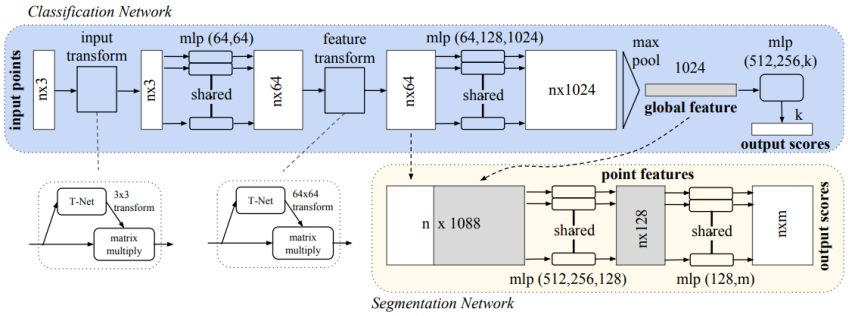

Figure 2.5: The structure of PointNet [72].

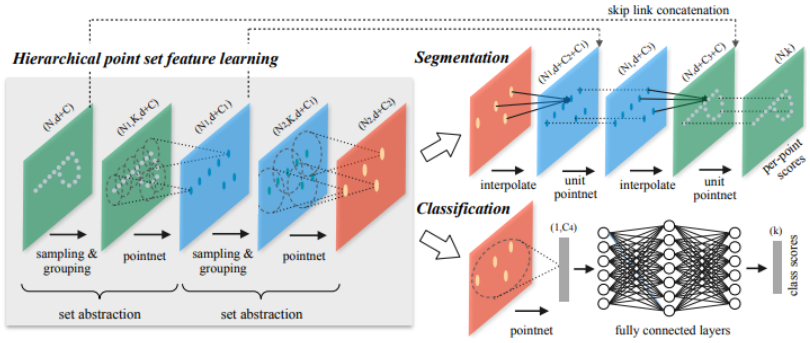

Figure 2.6: The structure of PointNet++ [73].

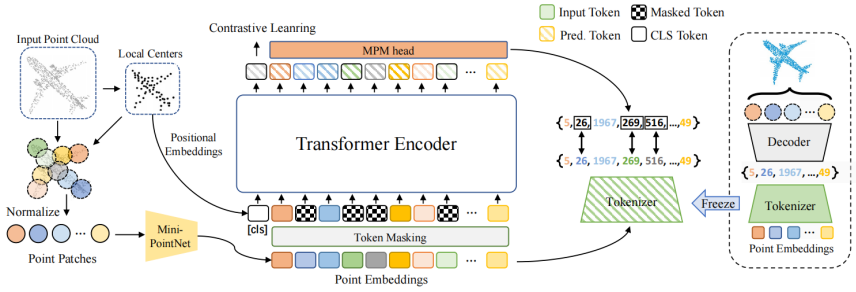

Figure 2.7: The structure of Point-BETR [104].

tively processed. In order to avoid difficult point tracing and easy failure in the absence of colour information caused by point tracking the Transformer is applied to process point cloud data by using the P4Transformer model [21] to simulate the original point cloud video. P4Transformer [21] includes 4D convolution and Transformer model [89], which are respectively are used to display the spatial structure and capture the information of the point cloud video. The attention mechanism is introduced for self-focusing on local features, and the weights obtained from Transformer [89] are used to determine the conditions for merging similar parts. PSTNet [22] mine the spatial and temporal dimensions of point cloud data and propose point space-time (PST) [22] convolution for the sequence information representation of point cloud data. The PST convolution is used to separate the spatial and temporal dimensions, allowing spatial and temporal convolutions are used to capture the local structure and simulate the space dynamics, respectively.

2.2.3 Point Cloud with Pose Estimation Methods

In these methods, point cloud are used for pose estimation task. Joint Depth-Pose Learning without PoseNet [111] improves the traditional method of using the learning ability of the model to obtain priories information from large amounts of data for training in monocular estimation and position estimations. Proposing Joint Depth-Pose Learning without PoseNet [111] solves the scale geometric problem. HandVoxNet [61] implements pose estimation from a single depth map. It estimates the joint coordinates by inputting a single voxelized depth map, then estimates the heat map of the joint points using V2V-PoseNet [68]. Finally, it will combine both of them to form the resultant 3D hand shape. The PVN3D[36] model first performs feature extraction using the feature extraction module, inputs the features into the 3D keypoint detection module and instance semantic segmentation module to predict each key point, centroid and semantic segmentation respectively, and then utilizes a clustering algorithm.

2.2.4 Summary and Analyze

In the above reviewed and analyzed papers on point cloud classification, it can be found that the algorithms utilize the features and data structure of point cloud data well. Because of the special characteristics of point cloud data in 3D space, the methods of analyzing point cloud data are more diverse. PointNet [72], PointNet++ [73], PointAugment [52], and PSTNet [22] use 3D structural features of point cloud data to mine and propose new models, such as the distinction between spatial and temporal dimensions of point cloud

data, and the combination of 3D data spatial dimensions and 2D data augmentation. AdvectiveNet[[35], SampleNet [47], P4Transformer [21], PSTNet [22], HandVoxNet [61], and PVN3D [36] use the feature extraction method to extract local features more accurately in point clouds. Joint Depth-Pose Point Contrast [97] creates a new loss function to reduce point overlap in point clouds.

2.3 Segmentation

Point cloud based segmentation uses point cloud datasets as the input directly, using the regular pipeline for point cloud based segmentation as shown in Figure 2.8. In this section, point cloud based segmentation is divided into two different categories, 2D based and 3D based methods, this shown in Figure 2.9. The 2D based method projects the point cloud to a 2D plane and applies traditional 2D semantic segmentation network to do the processing. Alternatively, the 3D based method extracts the feature information directly in 3D space.

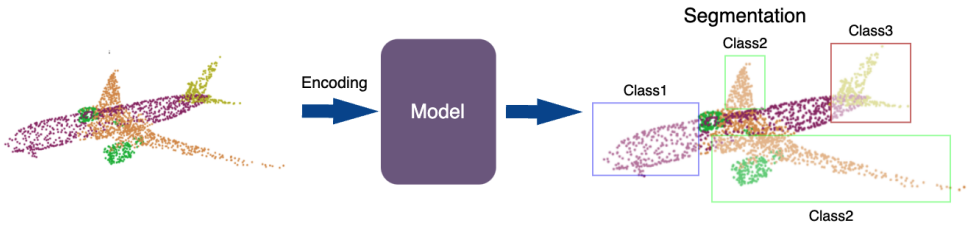

Figure 2.8: Regular Pipeline for Point Cloud Segmentation

2.3.1 Based on Two Dimensional Methods

SqueezeSeg [93] was proposed by Bic hen Wu et al,. in 2017. In this study, a lightweight convolutional neural network was used to achieve real-time semantic, instance segmentation of 3D objects in point cloud. In order to easily process the 2D convolutional neural network, the point cloud data is first projected to obtain the front view, then the input image is feature extracted and segmented using the SqueezeSeg [93] convolutional network. Finally the output is further optimized using the conditional Random Field (CRF), as shown in Figure 2.10. Geometric Shared Network (GS-Net) [101] can specifically and efficiently learn point descriptors, reduce losses due to image transformations, and improve the robustness of the model. The GSC module in GS-Net can effectively capture both local and global

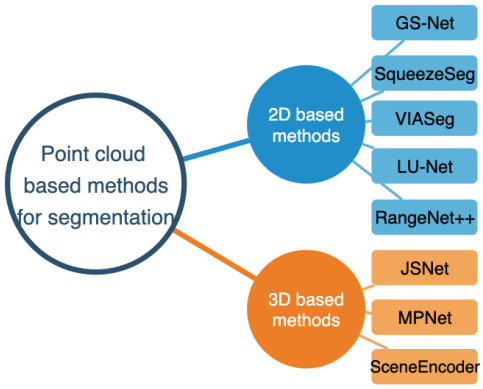

Figure 2.9: Point Cloud Based Methods for Segmentation.

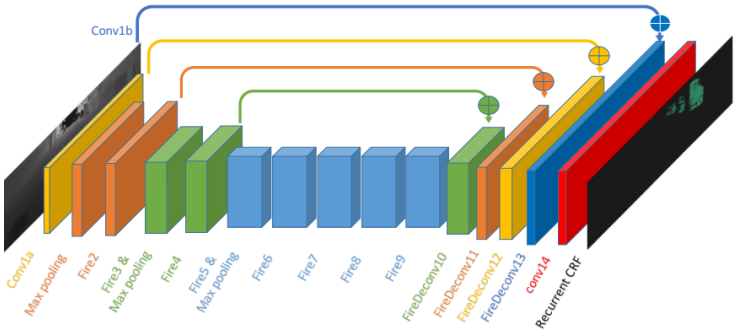

Figure 2.10: The structure of SqueezeSeg [93].

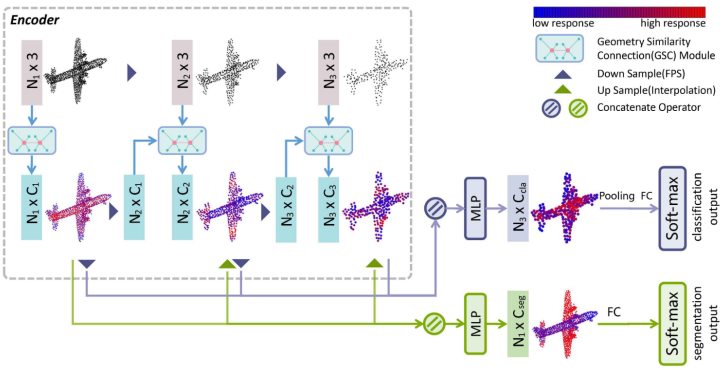

Figure 2.11: The structure of Geometric Shared Network (GS-Net) [93].

geometric features, making the model significantly more effective in segmentation tasks as shown in Figure 2.11. In the VIASeg [113], the authors convert color information to data-level embedded in point cloud data. This is done so that the color information of the image provides rich visual information and improves the semantic segmentation performance. This approach is a multimodal (visual + point cloud) combination approach. LU-Net [48] extracts high-level 3D features for each point of the point cloud , then projects these features to a 2D multi-channel front view. This model benefits from high level 3D feature extraction and embeds 3D local features into the 2D image, which allows the model to be used effectively. RangeNet++ [65] was proposed by Andres Milioto et al,. in 2019. This model projects the 3D data to a 2D front view. Despite this, the model performs well the full semantic segmentation of the point cloud data and obtains the full semantic label of the original point cloud.

2.3.2 Based on 3D Methods

JSNet [110] extracts features from the original point cloud data. In order to improve the feature discrimination, a feature fusion method is incorporated to perform feature fusion at backbone for different levels. Then, using the joint instance semantic segmentation module, semantic features are embedded in the space and feature fusion is performed to improve feature robustness. Furthermore, the model also aggregates the fused features into the feature space to improve the semantic segmentation. MPNet [82] was proposed by Tong He et al,. in 2020. The structure of point cloud data is 3D, which is complex. Also, the distribution of point cloud data is irregular, which makes it more difficult to process the distribution and learning of point cloud data. Then leads MPNet [82] to introduces the method of memory module to achieve small batch processing. Each small batch of samples is learned and memorized to achieve memory enhancement, achieving to reduce the problem of category imbalance. SceneEncoder [98] was proposed, where a scene encoding module is used to achieve enhanced global information. The module can filter the categories that do not belong and perform semantic segmentation. On the other hand, a regional collinearity loss was designed to approach features and neighboring points of the same label to improve feature recognition. The experimental effect of SceneEncoder [98] also achieved SOTA.

2.3.3 Summary and Analysis

As we can see in this subsection, because of the sparsity and irregularity of point cloud data, CNNs in 2D does not work directly on 3D point cloud data. This requires the point

cloud data to be transformed first. The 2D-based methods, SqueezeSeg [93], Geometric Shared Network (GS-Net) [101], VIASeg [113], LU-Net [48], and RangeNet++ [65] project the point cloud data to the 2D plane, and add different methods to improve the projection effect during the projection process. Thus, the methods in the projection of 3D data to 2D will be more diverse. This is an interesting and intuitive idea since the projection process leads to loss of information. The other class of models, based on 3D processing, including JSNet [110], MPNet [82], SceneEncoder [98] the point cloud data is directly input to the model and the feature information is extracted in 3D space. In this process the semantic segmentation of point cloud can be helped by adding feature fusion, attention mechanism. In future work, the missing information due to sparsity of point cloud data can be compensated for incorporating more effective methods to provide richer information.

2.4 Attention in Vision

2.4.1 What is the Attention Mechanism

The Attention Mechanism is obtained from intuition, based in the limited attention resources used to efficiently filter out high-value information from a large amount of information. Deep learning has been widely used in different types of tasks such as Nature Language Processing (NLP)[39], image classification, and speech recognition, and has achieved remarkable results. Therefore, understanding the working principle of the attention mechanism principles crucial. Figure 2.12 shows the evolution of attention mechanisms.

2.4.2 Evolution of Attention Mechanisms

In 2014, the Google Mind team published the paper "Recurrent Models of Visual Attention" [67] leading researchers to pay attention to the Attention mechanism, and this paper proposed to use it on an Recurrent Neural Networks (RNNs) model for image classification, resulting in a good performance. Subsequently, Bahdanau et al. published the paper "Neural Machine Translation by Jointly Learning to Align and Translate" [5], which proposed using the Attention mechanism on machine translation tasks to perform translation and alignment becoming the first publication to apply the Attention mechanism in NLP. Then, in the paper "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention" [99], published by Xu et al. successfully applied the Attention mechanism to the Image Caption domain. Thereafter, the Attention mechanism has been widely used

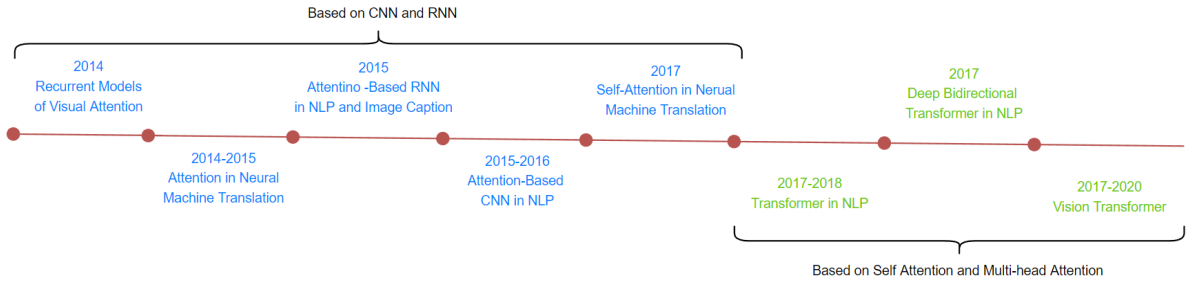

Figure 2.12: The Evolution of Attention Mechanisms.

in various deep learning tasks based on RNNs neural network models. In 2017, Google published a paper "Attention is all you need" [89], which proposed to use self-attention mechanism to learn text representation on machine translation. Figure 2.12 shows the general progress trend of Attention mechanism research. [100] first introduced visual attention in 2015, which presents soft and hard, pioneering the attention model for 2D visual processing. [60] proposed two modified versions of attention, known as global attention and local attention. With wide success use of attention models, various attention models such as RNN Encoder-Decoder [15], GATs [90], and MA-CNN [112] have further extended the attention model in the network framework. However, Convolutional Neural Network (CNNs) and RNNs are still used as the base models as location information cannot be modeled. [89] uses a novel model Transformer, and proposes self attention and multi-head attention, which can ignore sequence problem of position information, which are already widely used for various vision task. Recently, [19] presented a Vision Transformer, ViT, based on a structure of encoder to decoder in Transformer, and implements Transformer in computer vision and achieved SOTA results on public datasets.

However, in point cloud data, point clouds are sparse and disordered, each point does not hold semantic information, which still leads to a huge challenge in processing point cloud data.

2.4.3 Spatial and Channel Attention in Vision

Spatial attention and channel attention was already used in many computer vision tasks. [70] proposed a Bottleneck Attention Module(BAM) which employs spatial or channel attention improving network characterization effectively. [92] presented a CBAM model, a simple and effective feedforward convolutional neural network attention module, which

achieves effective determination for high focus points of images. [41] introduced a novel input weight distribution for attention model SENet, which applies channel attention in vision, as shown in Figure 2.13. [43] designed STN, which used channel attention for feature extracting. [13] presented SCA-CNN structure, which combine the spatial and channel attention in CNN.

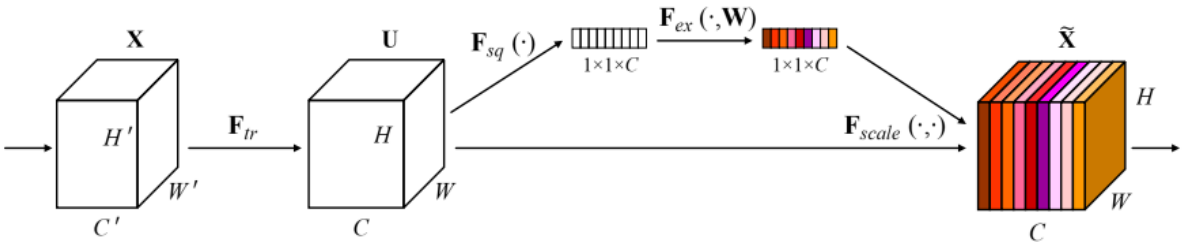

Figure 2.13: The structure of Squeeze-and-Excitation block [41].

Inspired from the image patch generation structure used in ViT [19], and the weighting relationship in spatial and channel attention, we propose a point weight relationship module (PointAT), which splits the point clouds data into patches (point clouds patches), uses spatial and channel attention to extract local features from the point clouds patches and obtains the intensity of attention relationship between the point clouds patches.

2.4.4 Point Cloud-based Attention Model in Deep Learning

Using the attention mechanism has become a trend in many models, as well as in point cloud data. Because of the irregularity and disorder of point clouds data, it is not easy to process point clouds data directly by using convolutional networks. In contrast, attention models are a sort invariant, computation non dependant on point connections. [28] introduced point cloud transformer (PCT), which use the attention model for data processing and capture point cloud features. The model is made up of an encoder and decoder structure in the PCT, giving model with good global attention feature learning ability. [28], [108] are good to apply a 2D Transformer model to 3D point clouds data. To contrast the PCT, this implementation applies vector attention, which is the main conceptual difference between the two models. Nevertheless, the spatial and channel dimensions in point clouds data still show that local relationships in point clouds data are not effectively used, and it is still challenging to explore local relationships in point clouds data.

P4Transformer [21] split points cloud data and extracts local features using a four-dimensional convolution of points with encodable spatio-temporal local structure information. Voxel Transformer [63] applies local Attention and dilated Attention in a multi-head attention mechanism, which is used to explore local information and gradually increases the search step to expand the search range, achieving information extraction from local to global, respectively. Finally, CT3D [80] includes region proposal and channel-wise Transformer, where spatial context modeling of key points is performed and using channel-wise reweighting to enrich the information.

Large numbers of methods also employ attention and Transformer model. [59] proposed Group-Free, which include self-attention model and cross attention model. [66] introduced an end-to-end Transformer model, which combines feature points of local features, then uses a transformer to encode and decode.

In contrast to the above models, our network SCA is based on the spatial and channel attention mechanism to explore the geometric relationship between point clouds, and applies the geometric relationship strength to the model to fully and effectively use the attention model to explore local features of point clouds.

2.5 Summary

In chapter two, we introduced generation of point clouds and analyse the existing methods based on classification and segmentation. Furthermore, we investigate attention mechanism in computer vision, includes background of attention mechanism, evolution of attention mechanism, spatial and channel attention in vision and point cloud-based attention model in deep learning. Based on the chapter one and two analyse and investigation, we found that this data type's ability to provide depth information, point sparsity and disorder pose a challenge in designing appropriate deep neural networks to process them and it is still challenging to explore local relationships in point clouds data. so, in chapter three, in order to better extract features and obtain geometric information we will propose a point attention (PointAT) model and propose attention value (AT value) model for feature fusion to apply geometric relationship to the data. Then, we will propose a new spatial and channel attention-based network (SCA). The SCA is the overall structure of the network, and the main purpose is to connect PointAT and AT value model, then capturing meaningful geometric information by applying the geometric relationship between point clouds patches to the model, then propose an auto pooling framework to extract global features.

Chapter 3

SCA-Net: Spatial and Channel Attention-based Network for 3D Point Clouds

The disordered and sparse form of point clouds data in 3D space makes it challenging to explore internal relationships in point clouds data. In contrast to 2D data, the disorder and sparse form of point clouds data make it more difficult for neural networks to process point clouds data [29]. Point 4d transformer (P4Transformer) [21], Point cloud transformer (PCT) [29] has pioneered applying the attention mechanism to point clouds data processing. PCT [29] extracts local attention of point clouds by inputting point cloud data directly into the attention module, PCT ensures local attention of point clouds with rotation and order invariance, and benefits from attention mechanism in Transformer [89] which can ignore the disorder of point clouds data and well use attention model for feature extraction. P4Transformer [21] splits images to perform effective local feature extraction of point clouds.

In recent years, Attention Mechanism [100] has been widely used in computer vision tasks and obtained excellent results. Attention mechanism [100] was proposed by Xu et al. in 2015. In contrast to traditional neural network Long Short Term Memory networks (LSTM) [37], Gate Recurrent Unit (GRU) [15], Attention Mechanism [100] overcomes the sequence alignment and model transformation problems. The inherent data location information in RNNs [106] networks (e.g., LSTM, GRU) is quite important for training models, but limits batching capability of samples and reduces the training capability of neural networks. Thus far, the predominant methods of employing attention are spatial and channel attention. [41].

Spatial attention focuses on which part of the spatial dimension is meaningful. For the input feature map, it performs average pooling and max pooling in one channel dimension, splices the two obtained feature maps in channel dimension, downscales them by using a convolutional layer, then generates spatial weight coefficients by the activation function, finally multiplies weight coefficients with the original feature map to get a final output feature map, as shown in Figure 3.1.

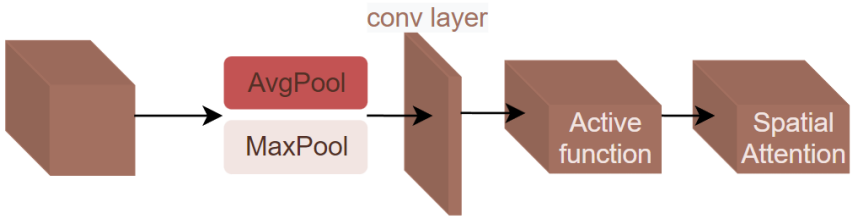

Figure 3.1: The structure of spatial attention.

Channel attention focuses on which channel features are meaningful, input feature map and perform average pooling and max pooling respectively to get two feature maps, giving them deeper meaning and sharing parameters through the neural network, then share the parameters and add the two feature maps together. Finally, the weight coefficients are multiplied with the original feature map to obtain a final output feature map as shown in Figure 3.2.

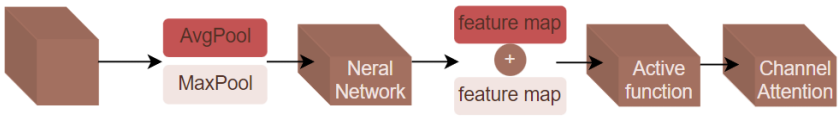

Figure 3.2: The structure of channel attention.

Inspired by the Attention model's success in 2D computer vision tasks, we propose a novel framework SCA for point cloud attention learning based on the spatial attention and channel attention. In order to achieve this, we propose to employ attention to uncover the geometric relationship between different point cloud patches, and we define this relationship as geometric relationship. From here, we capture this geometric relationship for geometric features that are meaningful for point cloud understanding.

In our model, we find that there is an geometric relationship between different regions of the point cloud data. As shown in Figure 3, this geometric is inherent to point cloud data,

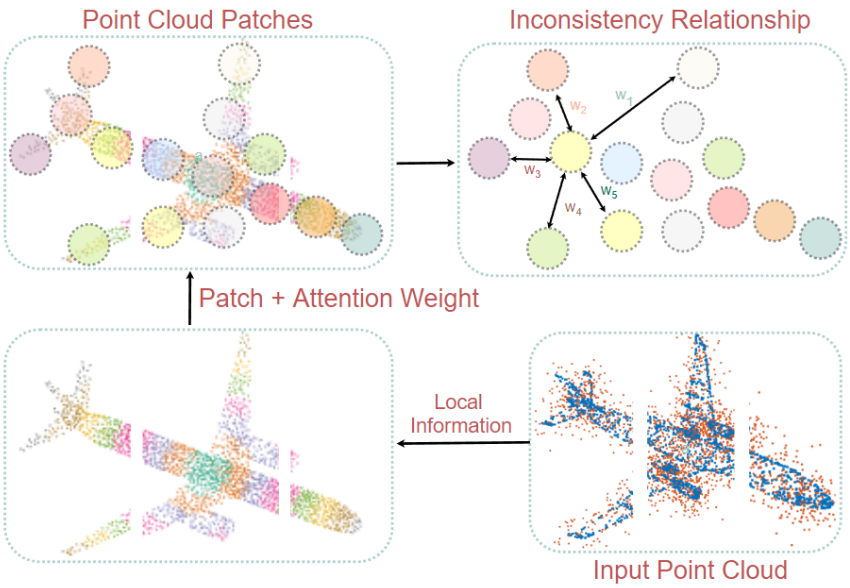

Figure 3.3: The structure of geometric relationship.

there are different degrees of correlations between different distribution regions in point clouds, and this correlation is invariant to data transformation. Given this geometric, we create a model that takes advantage of this, Our model uses the geometric relationship between point clouds to improve model efficiency. The main advantages and contributions of this paper are summarized as follows:

1. We propose a model leveraging the attention mechanism to learn the geometric relationship between point clouds in point cloud data. This is done through a proposed attention feature fusion method for exploration of relationship correlations.

2. We propose a new attention-based mechanism for point clouds learning framework, named SCA, which is well suited to handle disordered, unstructured point clouds data.

3. Through extensive experiments, the proposed framework is shown to achieve SOTA performance in classification and segmentation tasks.

Besides, we also introduce an adaptive downsampling structure we name Autopooling. This model considers each point's importance weight and picking key points adaptively. This structure preserves the original information and order. Additionally, this structure can be easily migrated to other networks.

3.1 Spatial and Channel with Attention network in Point Cloud (SCA)

In contrast to the above models, the SCA is based on the spatial and channel attention mechanism to explore the geometric relationship between point clouds, and apply the geometric relationship strength to the model to fully and effectively use the attention model to explore local features of point clouds.

In this section, we first illustrate Spatial and channel with attention network (SCA) framework, and how Point Attention (PointAT) establishes relationship strength between point cloud features to the model and successfully explores the relationship strength between point clouds features data with different weights. We also shows how to establish relationship strength and model combination in attention value model(AT value model) to achieve feature fusion, and eventually implement it for classification and segmentation tasks. Firstly, we introduce the concept of Spatial and channel with attention network (SCA) separately, and apply the geometric relationship between point clouds to the model. Then, we present how point attention (PointAT) and AT value model can be used to extract geometric relationships between point clouds and apply feature fusion to the model.

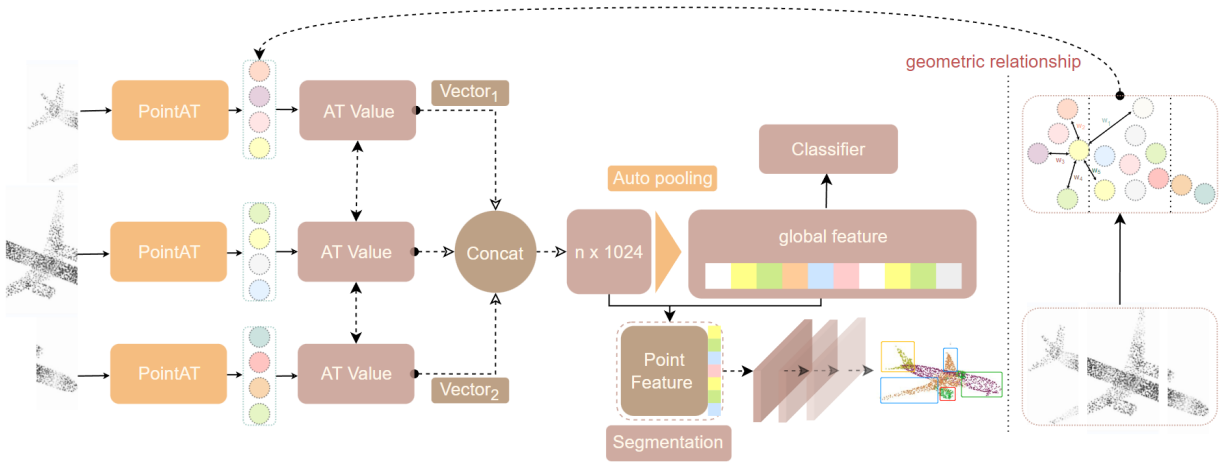

Figure 3.4: The structure of Spatial and channel with Attention network (SCA).

A general structure of the Spatial and channel with Attention network (SCA) is shown in Figure 3.4. The SCA is the overall structure of the network, and the main purpose is to connect PointAT and AT value model, apply the geometric relationship between point clouds to the model, then propose a auto pooling framework to extract to global features.

Firstly, it divides input data into three local patch partitioned contains plentiful geometric information to better extract local features and prepare for subsequent exploration of geometric relationships. Following the splitting, each point clouds are input into the PointAT model, which contains the spatial and channel attention (described below), and the attention weights are effectively obtained between the point cloud, and we find that they have different correlations, which indicates that there are geometric relationships between the point cloud. We input this geometric relationship obtained from PointAT into the AT value model, then perform feature fusion to obtain new data between the geometric relationship and the original point cloud. The newly formed point clouds data consequently has richer information and explore the geometric relationship strength between point clouds data, learning the relationship strength information of each points. From here, we propose a new global feature extraction method, Auto pooling. Because all points contain geometric relationship information, we further employ the attention mechanism in the pooling layer to get better global features.

The model accepts as input a point cloud point $\in \mathbb{R}^{N\times d}$ with N points each having d-dimensional feature vector. We divide point clouds into several local point cloud patches. $p_i \in \mathbb{R}^{N_i\times d}(i \in \mathrm{N})$ as several local point cloud patches are input to PointAT, and geometric Intensity relationships between point clouds are learned using the PointAT module to obtain point cloud weights $w_i(i \in \mathrm{N})$. From here, the weights are connected in feature space by AT value model for feature fusion, and then linear transformation is performed.

$$\text{Vector }_a = w_i p_i, a = \mathrm{N}, (1)$$

$$F_v = \text{concat}\left(\text{ Vector }_1, \text{ Vector }_2 \ldots \text{ Vector }_a\right), v = \mathrm{N}, (2)$$

$$F_o = f_{\text{Autopooling }}\left(F_v\right), o = \mathrm{N}, (3)$$

Where Vector $_a$ represents a weight vector obtained after the AT value model, F_v uses concat to splice vectors in preparation for subsequent Auto pooling. In order to extract effective global features and apply the learned geometric relationship between point clouds data to the model, we propose a novel auto pooling layer to extract global features, and F_o is global feature which is obtained from F_v after auto pooling.

3.2 Point Attention (PointAT)

In our model, we find that there is an geometric relationship between point cloud data as shown in Figure 3.5. We use PointAT model to explore the geometric relationship between point cloud data by adding temporal and spatial attention mechanisms to the

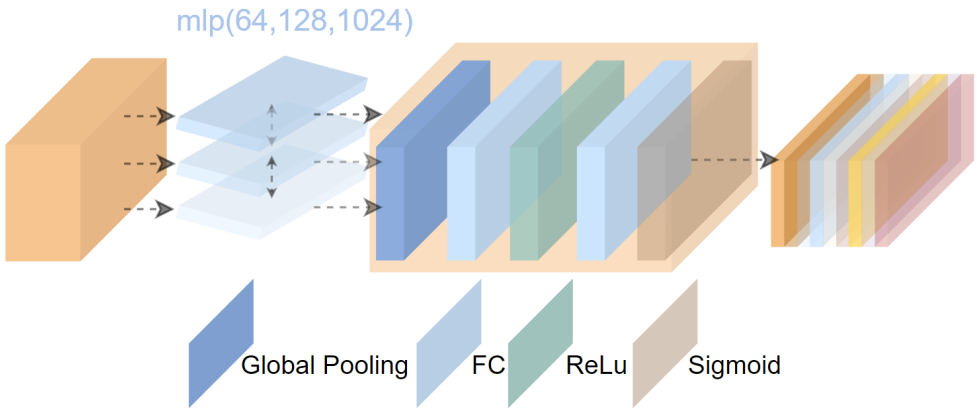

Figure 3.5: The structure of Point Attention (PointAT).

model in Figure 5. First, we consider point embedding in the model by splitting original data into patches. $p_i \in \mathbb{R}^{(N_i \times d)} (i \in N)$ is used as new point cloud data input to PointAT. A shared neural network is used here and the point data has 128 dimensions in multi-layer perception. And spatial attention is added to extract the point cloud geometric relations at spatial dimension. The combined features of cross-spatial dimensions by global pooling. generate a 1×1×C matrix, combining across spatial dimension features by global pooling and generate a 1×1×C matrix, Subsequently add weight values to each output to form weight features and add two fully-connected layers. Then calculate weights between point cloud data by sigmoid function, the output method in the following way. Where w_i is the weight value of the PointAT output, and $F_{\text{global pooling}}\ (p_i)$ is the maximum downsampling of input data. The above process is only used to extract the point cloud weight values without changing and reducing the number of point clouds.

$$\begin{aligned} w_i &= F_{\text{PointaT}}\ \left(F_{\text{global pooling}}\ (p_i), W\right) \\ &= F_{\text{Pointat}}\ \left(F_{FC}\left(F_{\text{global pooling}}\ (p_i), W\right)\right), \end{aligned} \tag{4}$$

$$F_{\text{global pooling}}\ (p_i) = \frac{1}{N_i \times d} \sum_{N_1}^{N_i} \sum_{1}^{d} p_i \left(N_i \times d\right), \tag{5}$$

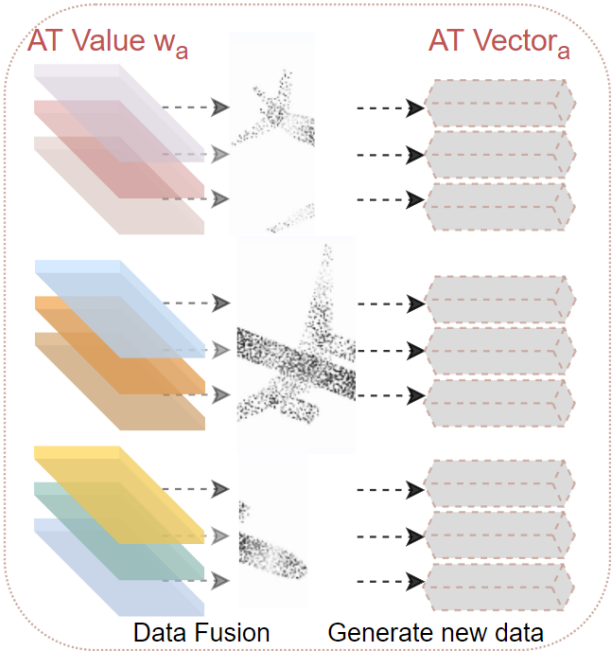

Figure 3.6: The structure of Attention Value Model (AT value model).

3.3 Attention Value Model (AT value model)

The main approaches of data fusion are early-fusion, data-level fusion, late-fusion or decision-level fusion and intermediate-fusion. Data fusion often accompanies insufficient complementary among multiple data points and generates a large amount of redundant information. In this paper, we adopt inputting point clouds data into models separately to get their weight values, effectively avoiding generating a lot of redundant information. And the generated errors are independent and do not affect each other, it does not lead to further error accumulation. Attention Value model performs data fusion on point cloud data and weights to get a new weight vector in Figure 3.6. By leveraging the point cloud geometric relationship and the original data together, each points has different weight coefficients within it, which significantly improves model feature extraction effectiveness. Perform data fusion operation on point cloud data $p_i \in \mathbb{R}^{(N_i \times d)} (i \in N)$ and geometric intensity relations $w_i (i \in N)$.

$$\text{Vector}_{\ a} = w_i p_i, a = \text{N}, \quad (6)$$

3.4 Auto Pooling

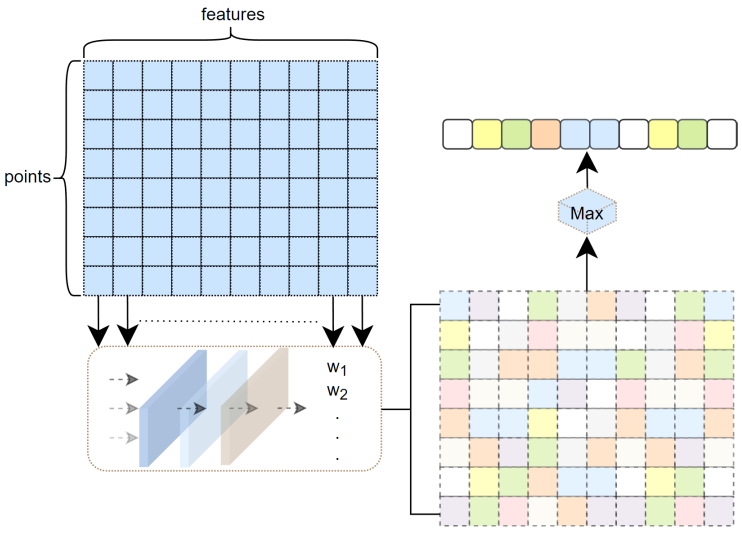

Figure 3.7: The structure of Attention Value Model (AT value model).

In this section, we propose a new automatic downsampling method that can be used for deep neural networks, as shown in Figure 3.7. A new solution for feature extraction, this method for weight selection of feature vectors in arbitrary neural networks. By considering each point's importance without changing the original data, we effectively retain important points in the point cloud and reduce unnecessary points. Moreover, this downsampling layer can be used together with convolutional networks. The main structure of this adaptive downsampling layer is as follows.

$$F_v = \text{concat}\left(\text{ Vector }_1,\ \text{Vector }_2 \ldots\ \text{Vector }_a\right), (2)$$
$$F_o = f_{\text{Autopooling}}\ \left(F_v\right), o = \text{N}, (3)$$
$$f_{\text{Autopooling}}\left(x\right) = f_{\text{attention}}\ \left(f_{\text{maxpooling}}\left(x\right)\right), (7)$$
$$f_{\text{attention}}\left(y\right) = F_{FC}\left(F_{\text{global pooling}}\left(y\right), W\right), (8)$$

3.5 SCA Networks for classification and Segmentation

Classification: In the point cloud classification task, the SCA classification network and details graph as shown in Figure 3.4. In order to classify the point cloud data $p_i \in \mathbb{R}^{(N_i \times d)} (i \in N)$ we provide the global features F_o to the classifier. In order to make a better comparison, the classifier does not add a new model, and consists of two feedforward neural networks (linear layers, BatchNorm, and ReLu layers), finally adding linear layers for classification and predicting final results, marking the highest scoring categories.

Segmentation: Segmentation falls under the umbrella of classification, where we have to each point's label is predicted. We first connect the global features F_o and F_v. As shown in Figure 3.4, the architecture of segmentation networks is basically similar to the classification net- works, with the main difference being that each point's classification results are predicted.

3.6 Experiments and Evaluation

Now we evaluate SCA and SCA-Auto (SCAA with Auto pooling) performance on two publicly available datasets, ModelNet40 [96] and ShapeNet [11], and perform a comprehensive comparative analysis of these two models with other methods, including target classification and segmentation. Each set was trained with the same negative log-likelihood loss

Table 3.1: Comparison with state-of-the-art methods on the ModelNet40 classification dataset. Acc. means overall accuracy. All results quoted are taken from the cited papers.

Method	#points	Acc.
PointNet [72]	1K	89.2%
PointNet++ [73]	1K	90.5%
NPCT [29]	1K	91.0%
DGCNN [91]	1K	91.84%
PointNet++ [73]	5K	91.9%
SO-Net [51]	1K	92.5%
PointCNN [54]	1K	92.5%
Point2Sequence [56]	1K	92.6%
Point Transformer [108]	1K	92.8%
DensePoint [57]	1K	92.8%
RSCNN [58]	1K	92.9%
SCA(this work)	1K	92.3%
SCAA(this work)	1K	93.4%

function with a learning rate of 0.001. We use Stochastic Gradient Descent (SGD) optimizer, momentum weight decay set to 0.9 and 0.0008. Other training parameters including batchsize, and epoch, will be given in subsequent datasets.

3.6.1 Classification on ModelNet40 Dataset

The Modelnet40 dataset is a 3D image classification dataset that contains all hand-drawn CAD point cloud images. ModelNet40 [96] contains 12,311 meshed CAD models in 40 different categories. For better comparison, 9,843 models are used for training and 2468 models are used for testing. For the base model we used Pointnet [72] as the base model, 1024 points were sampled uniformly. For data augmentation, we perform random scaling and rotation for the data in 3D space. In the training process, random translation in [-0.2, 0.2] was used, and we trained 250 epochs with an initial learning rate set to 0.001 and a batchsize set to 8. The experimental results are shown in Table 3.1. The overall accuracy of SCA was 92.3%, compared with PointNet [72] and PointNet++ [73], SCA improves by 3.1% and 1.8%, respectively. After adding Auto pooling module, SCAA improves by 1.1% compared to SCA. Note that slicing original data as input is being done, which could in principle further improve network performance.

3.6.2 Attention on Value Task

We perform pre-training with the PointAT model and fine-tune the model to evaluate the generalization ability of learning representations. We visualize feature weights in Figure 3.8. In Figure 3.8(a) and 3.8(b), visualized feature weights are obtained from Figure 3.8(a), and after performing fine-tuning before and after fine-tuning ModelNet40 to obtain 3.8(b), using PointAT model to calculate point cloud data weight values. Figures 3.8(a) and 3.8(b) show that we have chunked data before it into PointAT model to better feature extraction and improve the fine granularity. In Figure 3.8(b) after fine-tuning, we get the point cloud feature weights, and it is obvious to see that each point cloud has different feature weights, which indicates that feature is well filtering data under the attention mechanism.

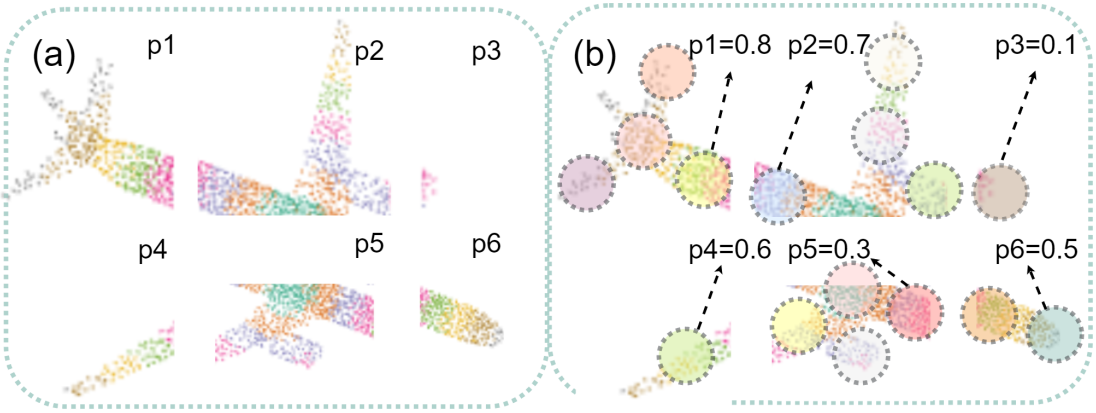

Figure 3.8: The attention weights are visualized as attention values. We show modelNet40 data weight values for each part of features after PointAT.

3.6.3 PointAT Task

We use PointNet as the baseline. PointAT is used as a comparison experiment, as shown in Figure 3.9 and 3.10. It can be seen that our method can separate features from different classes well. In Figure 3.8 we can see that each point cloud data after chunking gets geometric feature weight coefficients, which we can easily find that PointAT significantly improves model learning ability from the experiments in Figure 3.9 and 3.10, and further verifies that point feature weights help models learn more 3D objects' information.

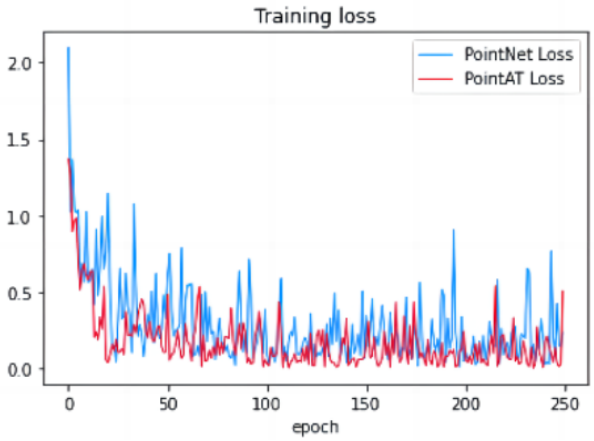

Figure 3.9: PointAT is used as a comparison experiment.

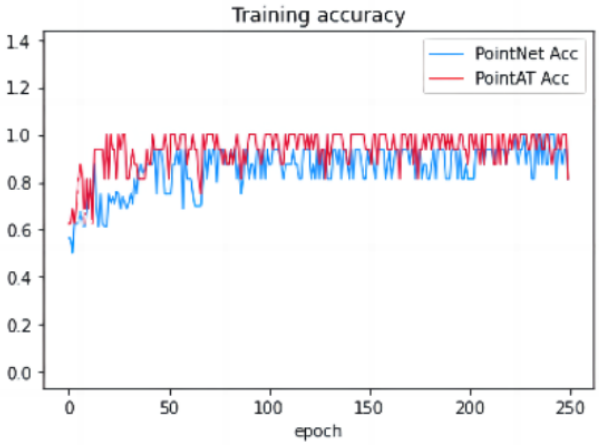

Figure 3.10: PointAT is used as a comparison experiment.

3.6.4 Auto Pooling

We propose the novel adaptive downsampling layer Auto pooling, we use PointAT as a baseline for comparison experiments of SCA and SCAA respectively, SCAA changes max pooling in model SCA to auto pooling as shown in Figure 3.11 and Figure 3.12. It can be seen that auto pooling significantly improves convergence speed of the model and improves model performance on datasets in both accuracy and speed.

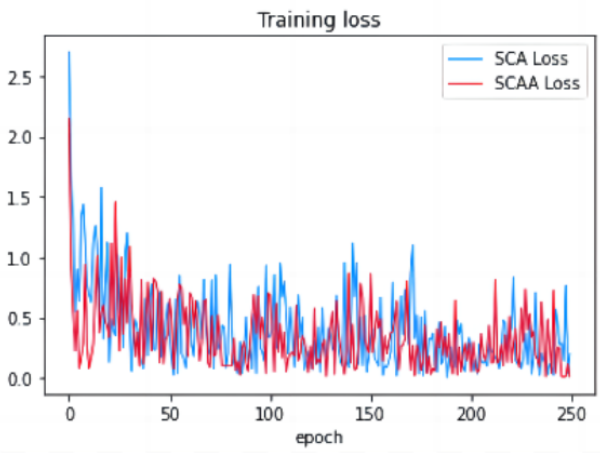

Figure 3.11: SCAA changes max pooling in model SCA to auto pooling.

3.6.5 Segmentation task on ShapeNet dataset

Point cloud segmentation is always a challenging task, which aims to segment a 3D model into multiple meaningful parts, that could be considered as a special form of classification. We perform experiment evaluations on the ShapeNet dataset [11]. ShapeNetPart dataset [11] consists 16 classes selected from the ShapeNetCore dataset and annotated with semantic information for semantic segmentation task. ShapeNet Part consists of 16 classes, 50 parts, and a total 16846 samples. This sample set exhibits unbalanced characteristics. Each sample contains more than 2000 points, which is a small dataset. There are 12137 training samples, 1870 validation sets, and 2874 test sets in this dataset, totaling 16881.

In training task, after Point AT, data was downsampled to 2048 points, retaining point state part, using random panning in [-0.2,0.2] to increase input data, and training 250 epochs with initial learning rate set to 0.001 and batchsize to 8. In the test SCAA and

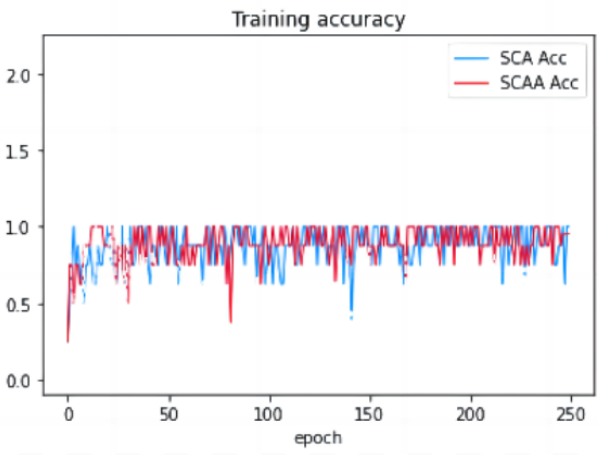

Figure 3.12: SCAA changes max pooling in model SCA to auto pooling.

Table 3.2: Comparison on the ShaperNet part segmentation dataset. MIoU means part-average Intersection-over-Union. All results quoted are taken from the cited papers.

Methods	mIoU	airp lane	bag	cap	car	chair	ear phone	guitar	knife	lamp	laptop	motor	mug	pistol	rocket	skate board	table
Kd-Net[45]	82.3	80.1	74.6	74.3	70.3	88.6	73.5	90.2	87.2	81	94.9	57.4	86.7	78.1	51.8	69.9	80.3
PointNet [72]	83.7	83.4	78.7	82.5	74.9	89.6	73	91.5	85.9	80.8	95.3	65.2	93	81.2	57.9	72.8	80.6
PointNet++ [73]	85.1	82.4	79	87.7	77.3	90.8	71.8	91	85.9	83.7	95.3	71.6	94.1	81.3	58.7	76.4	82.6
PCNN [4]	85.1	82.4	80.1	85.5	79.5	90.8	73.2	91.3	86	85	95.7	73.2	94.8	83.3	51	75	81.8
DGCNN [91]	85.2	84	83.4	86.7	77.8	90.6	74.7	91.2	87.5	82.8	95.7	66.3	94.9	81.1	63.5	74.5	82.6
P2Sequence [56]	85.2	82.6	81.8	87.5	77.3	90.8	77.1	91.1	86.9	83.9	95.7	70.8	94.6	79.3	58.1	75.2	82.8
Point-BERT [105]	85.6	84.3	84.8	88	79.8	91	81.7	91.6	87.9	85.2	95.6	75.6	94.7	84.3	63.4	76.3	81.5
PointConv [94]	85.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PointCNN [54]	86.1	84.1	86.5	86	80.8	90.6	79.7	92.3	88.4	85.3	96.1	77.2	95.2	84.2	64.2	80	83
PCT [29]	86.4	85	82.4	89	81.2	91.9	71.5	91.3	88.1	86.3	95.8	64.6	95.8	83.6	62.2	77.6	83.7
SCA(this work)	85.5	85.2	85.4	86.3	79	91	81.5	92.4	87.4	85.4	95.3	75.1	94.6	84.3	61.7	77	84
SCAA(this work)	**86.5**	**85.4**	85.4	**88.5**	80.4	91.6	**81.7**	**92.4**	88	86.8	95.9	75	94.8	84.2	62	78.8	84.1

SCAA models were trained with the same batch size, training epoch and learning rate settings as for the normal estimation task. Table 3.2 shows segmentation results. The mean IoU across all instance categories for each category was used as the evaluation metric, which was given for each object category at the same time. Our SCA and SCAA achieved 85.5 and 86.5 mIoU, respectively. For each object category the results show that our SCAA improved by 2.8% and 1% over PointNet [72] and SCA, respectively.

Chapter 4

Conclusion

4.1 Significance of Study Findings

Point cloud is an important expression form of 3D data. It has enjoyed a continuous development and attracted increasing attention due to its wide applications in many areas, such as artificial intelligence, deep learning, autonomous driving, tracking. In order to better use point cloud data for analysis and to explore future research directions, this paper presents a comprehensive review of existing methods and publicly available datasets, with a focus on the methods and research status of using point cloud data as direct input. Despite this data type's ability to provide depth information, point sparsity and disorder pose a challenge in designing appropriate deep neural networks to process them and it is still challenging to explore local relationships in point clouds data. Then, based on these review, we propose a new Spatial and Channel Attention network (SCA). The SCA is the overall structure of the network, and the main purpose is to connect PointAT and AT value model, then capturing meaningful geometric information by applying the geometric relationship (inconsistent relationship) between point clouds patches to the model, then propose a auto pooling framework to extract to global features. In this work, we concentrate on learning inconsistency relationship between point cloud data. For this purpose, we introduce a point attention model based on spatial and channel attention to learn the inconsistency relationship between point clouds, and further combine the inconsistency relationship with the point cloud data by an Attention Value Model. Besides, we introduc a adaptive downsampling structure, Autopooling. This downsampling structure consider each point's importance weight and picking key points adaptively, which can be used together with convolutional networks. Extensive experiments conducted on two benchmark datasets

(ModelNet40 and ShapeNet) clearly demonstrate the effectiveness of our SCA method.

4.2 Limitations of Proposed Method

The attention-based mechanism model already shows strong model learning and generalization capabilities in the 3D field, however, there are limitations in a number of publicly available point cloud data, which are very limited compared to other available data at present. In addition, better data splitting and fusion methods can be performed for the input point cloud data, which can improve the granularity of data and increase the integrity of data. Moreover, the existing public point cloud data does not include various influencing factors in realistic scenes and has limited data availability. On the other hand, with the attention mechanism added, models will have information missing and the information order cannot be aligned, we will continue to improve the models' completeness continuously. In the future, we plan to compare models in parallel in more diverse datasets to improve model generalization ability.

4.3 Future Research Directions

4.3.1 More Effective Data Processing Methods

Data processing in 2D images have been very diverse, including data enhancement, data fusion, data inversion, and data transformation. There has been great progress in data processing for 2D data, and many algorithms have been proposed. In 3D data, data processing of 3D data has also attracted attention. PointAugment [52] performed data fusion of the enhanced 2D data with the 3D data. SampleNet [47] added a projection operation to transform the data and achieved a new point cloud sampling. LU-Net [48] extracted the point cloud data with high-level features and embedded 3D local features into 2D to obtain the new point cloud data. RangeNet++ [65] researcher use projection method to project the 3D data to get 2D data. JSNet [110] performs fusion of different layers of backbone features using data fusion method. The above papers have shown that data processing for 3D point cloud data is a very effective model enhancement method, increasing the robustness of the model and improving the model effect by effective data transformation of the point cloud data. There are still many data processing methods that are worth designing new algorithms to obtain new data based on effective data processing methods to increase the training data.

4.3.2 Attention Mechanism

In recent years, attention has become the focus of research in machine vision [46]. The attention mechanism is one of the most popular techniques nowadays and has been used with amazing results in many computer vision tasks. Its application in neural networks has also made it a hot topic. In [46], researchers indicate that humans are not very good at processing the whole scene at once, but a person's attention will choose selectively some of the important information while observing the whole scene to guide the later decisions. This selective screening of the whole greatly eliminates invalid features and reduces complexity. Feature extraction can be performed accurately and efficiently for the object in the model. P4Transformer [21], Transformer3D-Det(T3D) [109], VoteNet [18], Channel-wise Transformer 3D Object Detection (CT3D) [79], Group-Free [59], 3DETR [66] applied the attention mechanism to the model and achieved quite good results on the whole effect of the model. From these papers, it is easy to find that the main focus of many existing models based on attention mechanisms is on feature extraction and analysis of weights for feature tasks, which has the advantage of improving the overall performance effect of the model in a very direct way. However, there is little attention to the problem of localization accuracy, and how to accurately localize attention regions is a worthwhile study and interest point in the future. In particular, in 3D point cloud data, there is a problem of overlap in the range of attention regions, which makes it possible to design new attention models in three-point cloud data as the focus of future attention.

4.4 Research Summary

Point cloud data as a form of 3D data representation is now widely used in various areas with great importance and references. Point cloud data can play an important role in the exploration of 3D space. In point cloud classification tasks, it provides important information for human-related tasks, such as point cloud action recognition P4Transformer [21], PSTNet [22], PoseNet [111], HandVoxNet [61], PVN3D [36], point cloud pose estimation PointContrast [97], HandVoxNet [61], PVN3D [36]. In point cloud detection and tracking, Transformer3D-Det(T3D) [109], VoteNet [18], Channel-wise Transformer 3D Object Detection (CT3D) [79], Group-Free [59], 3DETR [66], Semantic Point Generation (SPG) [102],Range-Guided Cylindrical Network [74], RangeDet [23],PV-RCNN [81], CenterPoint [103] can be applied to automatic driving, tracking, etc. In point cloud segmentation, SqueezeSeg [93], Geometric Shared Network (GS-Net) [101], VIASeg [113], LU-Net [48], RangeNet++ [65], JSNet [110], MPNet [82], SceneEncoder [98] perform the segmentation

task well, these models perform more accurate extraction from global to local by performing feature extraction on point cloud data, main methods are adding attention mechanism, adding more information to the original point cloud data, and mining sequence information in spatial and temporal dimensions. In the remaining part, we will propose several interesting directions for future work on point cloud.

Point cloud data can present objects well in 3D, which contain detailed information, so it becomes a key point that how to mine the key information. PointContrast [97], MPNet [82] improved the model effect by extracting local features, and it is easy to find that mining the local information of the point cloud and designing the algorithm from global to local is an intuitive and effective idea. Based on the local information for more accurate feature extraction, from how to perform accurate extraction for analysis and mining more local information. We think this is also a good and useful research direction. In the process of mining local information, we can avoid losing global information after extracting local features by performing joint local and global feature extraction.

In recent years, 3D point cloud data has been of great use for researchers to study 3D real information, but because the expression of point cloud data is not dense, presenting a sparse state, there is no information in the surrounding areas of points, which has become a breakthrough for research. By filling the missing information in the non-dense representation of point cloud, we think this is a very worthy research direction. There have been some studies focused on supplementing point cloud data with missing information here. In HandVoxNet [61] researchers add depth maps to supplement the missing information of point cloud data. In Semantic Point Generation (SPG) [102] semantic data and original point cloud data are fused for data tolerance to reproduce the original point cloud missing data. VIASeg [113] embedding color information into the point cloud data provides rich visual information to the point cloud data.

In recent years, many loss functions have been proposed to improve the capability of networks and further optimize models. The common loss functions are Cross-Entropy Loss, Center Loss, Triplet Loss, etc. In 3D point cloud, researchers have also designed new loss functions for point cloud data, such as PointInfoNCE Loss [97], to further improve the model effect to optimize the model performance. The effect of different loss functions on point cloud data is a very important direction.

Point cloud models are built using two streams, 2D-based and 3D-based methods. The 2D-based method mainly converts 3D point cloud data into 2D maps for various tasks. The 3D-based approach inputs 3D point cloud data directly into the model. There are now also methods that combine 2D and 3D methods, and using fused image information to help with 3D tasks is also an interesting idea.

Modality can be represented as a source or expression of information. MMML, the purpose of MMML is to perform through machine learning thus the ability to process and understand multi-modal information [6]. In machine vision, multi-modal learning can be formed between image, video, audio, and semantics. Researchers can complement point cloud data by adding information from different modalities to the point cloud data. On the other hand, some researchers have tried to use knowledge in fluid mechanics to help improve overall model performance. For example, AdvectiveNet[[35] used hydrodynamic natural flow phenomena to process point cloud data, PointContrast [97] added depth differences to estimate depth complementary information, and HandVoxNet [61] added depth maps for pose estimation. Therefore, how to use this information for multi-modal point cloud data complementation or to help the overall model to further explore deep relationships remains a worthy problem for future research.

There is also a novel approach which integrates point cloud data tasks into the pipeline of other related tasks. Many multi-channel methods are already proposed. For example, point cloud-based classification models PointContrast [97], point cloud-based pose estimation PoseNet [111], HandVoxNet [61], PVN3D [36], and point cloud-based 3D methods JSNet [110]. In the subsequent tasks, exploring how to better utilize and implement the multi-channel approach and use other tasks to better facilitate effective mining of point cloud data is an important future research direction.

In this paper, we review various aspects of point cloud models based on classification. To the best of our knowledge, the present work provide the first comprehensive review of point cloud approaches with a focus on deep learning-based methods. Specifically, we introduce the background and basics of point cloud data, then explain the difficulties faced by point cloud data, the public dataset, and the evaluation criteria. After that, we analyze the algorithms from classification and segmentation. And under each classification, we perform subdivisions. Then some popular neural networks are analyzed and reviewed, these models are widely used, then we analyze these algorithms based on point cloud from different perspectives. In classification module, it includes 3D point cloud, point cloud action recognition, and point cloud pose estimation. In segmentation, it includes 2D and 3D based models. Then we briefly describe each of the related tasks. On the other hand, based on the above review, we introduce a novel paradigm for 3D point cloud methods, using attentional model pre-training to learn inter-feature relationships between point clouds to collect structural information for synthesizing new point cloud data. Experiments on classification and segmentation of 3D point cloud tasks show that there are inconsistent relationships between point clouds, and enable the model to learn and generalize substantially better. In the meantime, Auto pooling, an unique adaptive downsampling framework, was introduced. Figure 3.11 and Figure 3.12 illustrates the enhanced performance by

demonstrating how auto pooling considers each point's importance and adaptively chooses significant points while preserving the information and order of the original data. We look forward to potential application of inconsistency relationships between point clouds to the model and will continue to investigate the model.

References

[1] Saifullahi Aminu Bello, Shangshu Yu, and Cheng Wang. deep learning on 3d point clouds. *arXiv e-prints*, pages arXiv–2001, 2020.

[2] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[3] Eduardo Arnold. Cooperative driving dataset (codd), November 2021.

[4] Matan Atzmon, Haggai Maron, and Yaron Lipman. Point convolutional neural networks by extension operators. *arXiv preprint arXiv:1803.10091*, 2018.

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[6] Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *CoRR*, abs/1705.09406, 2017.

[7] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Juergen Gall. A dataset for semantic segmentation of point cloud sequences. *CoRR*, abs/1904.01416, 2019.

[8] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019.

[9] Yasemin Bekiroglu, Naresh Marturi, Máximo A. Roa, Komlan Jean Maxime Adjigble, Tommaso Pardi, Cindy Grimm, Ravi Balasubramanian, Kaiyu Hang, and Rustam Stolkin. Benchmarking protocol for grasp planning algorithms. *IEEE Robotics and Automation Letters*, 5(2):315–322, 2020.

[10] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. *CoRR*, abs/1903.11027, 2019.

[11] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015.

[12] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. *CoRR*, abs/1911.02620, 2019.

[13] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5659–5667, 2017.

[14] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. *CoRR*, abs/1611.07759, 2016.

[15] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[16] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas A. Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. *CoRR*, abs/1702.04405, 2017.

[17] Mark De Deuge, Alastair Quadros, Calvin Hung, and Bertrand Douillard. Unsupervised feature learning for classification of outdoor 3d scans. In *Australasian Conference on Robitics and Automation*, volume 2, page 1. University of New South Wales Kensington, Australia, 2013.

[18] Zhipeng Ding, Xu Han, and Marc Niethammer. Votenet: A deep learning label fusion method for multi-atlas segmentation. *CoRR*, abs/1904.08963, 2019.

[19] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[20] Hehe Fan, Xiaojun Chang, Wanyue Zhang, Yi Cheng, Ying Sun, and Mohan Kankanhalli. Self-supervised global-local structure modeling for point cloud domain adaptation with reliable voted pseudo labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6377–6386, June 2022.

[21] Hehe Fan, Yi Yang, and Mohan Kankanhalli. Point 4d transformer networks for spatio-temporal modeling in point cloud videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14204–14213, 2021.

[22] Hehe Fan, Xin Yu, Yuhang Ding, Yi Yang, and Mohan Kankanhalli. Pstnet: Point spatio-temporal convolution on point cloud sequences. In *International conference on learning representations*, 2020.

[23] Lue Fan, Xuan Xiong, Feng Wang, Naiyan Wang, and Zhaoxiang Zhang. Rangedet: In defense of range view for lidar-based 3d object detection. *CoRR*, abs/2103.10039, 2021.

[24] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[25] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[26] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012.

[27] Jason Geng. Structured-light 3d surface imaging: a tutorial. *Advances in Optics and Photonics*, 3(2):128–160, 2011.

[28] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R Martin, and Shi-Min Hu. Pct: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, 2021.

[29] Meng-Hao Guo, Jun-Xiong Cai, Zheng-Ning Liu, Tai-Jiang Mu, Ralph R. Martin, and Shi-Min Hu. PCT: Point cloud transformer. *Computational Visual Media*, 7(2):187–199, apr 2021.

[30] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *CoRR*, abs/1912.12033, 2019.

[31] Vladimir Guzov, Aymen Mir, Torsten Sattler, and Gerard Pons-Moll. Human poseitioning system (HPS): 3d human pose estimation and self-localization in large scenes from body-mounted sensors. *CoRR*, abs/2103.17265, 2021.

[32] Timo Hackel, Nikolay Savinov, Lubor Ladicky, Jan Dirk Wegner, Konrad Schindler, and Marc Pollefeys. Semantic3d.net: A new large-scale point cloud classification benchmark. *CoRR*, abs/1704.03847, 2017.

[33] Miles Hansard, Seungkyu Lee, Ouk Choi, and Radu Patrice Horaud. *Time-of-flight cameras: principles, methods and applications*. Springer Science & Business Media, 2012.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[35] Xingzhe He, Helen Lu Cao, and Bo Zhu. Advectivenet: An eulerian-lagrangian fluidic reservoir for point cloud processing. *CoRR*, abs/2002.00118, 2020.

[36] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. PVN3D: A deep point-wise 3d keypoints voting network for 6dof pose estimation. *CoRR*, abs/1911.04231, 2019.

[37] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[38] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. *CoRR*, abs/2006.14480, 2020.

[39] Dichao Hu. An introductory survey on attention mechanisms in nlp problems. In *Proceedings of SAI Intelligent Systems Conference*, pages 432–448. Springer, 2019.

[40] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. *CoRR*, abs/1709.01507, 2017.

[41] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[42] Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges. *CoRR*, abs/2009.03137, 2020.

[43] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[44] Mor Joseph-Rivlin, Alon Zvirin, and Ron Kimmel. Mo-net: Flavor the moments in learning to classify shapes. *CoRR*, abs/1812.07431, 2018.

[45] Roman Klokov and Victor Lempitsky. Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In *Proceedings of the IEEE international conference on computer vision*, pages 863–872, 2017.

[46] M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23, 2010.

[47] Itai Lang, Asaf Manor, and Shai Avidan. Samplenet: Differentiable point cloud sampling. *CoRR*, abs/1912.03663, 2019.

[48] Sarah Leclerc, Erik Smistad, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Thomas Grenier, Carole Lartizien, Pierre-Marc Jodoin, et al. Lu-net: a multi-task network to improve the robustness of segmentation of left ventriclular structures by deep learning in 2d echocardiography. *arXiv preprint arXiv:2004.02043*, 2020.

[49] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[50] Bo Li, Yijuan Lu, Chunyuan Li, Afzal Godil, Tobias Schreck, Masaki Aono, Martin Burtscher, Qiang Chen, Nihad Karim Chowdhury, Bin Fang, Hongbo Fu, Takahiko Furuya, Haisheng Li, Jianzhuang Liu, Henry Johan, Ryuichi Kosaka, Hitoshi Koyanagi, Ryutarou Ohbuchi, Atsushi Tatsuma, Yajuan Wan, Chaoli Zhang, and Changqing Zou. A comparison of 3d shape retrieval methods based on a large-scale benchmark supporting multimodal queries. *Comput. Vis. Image Underst.*, 131:1–27, 2015.

[51] Jiaxin Li, Ben M Chen, and Gim Hee Lee. So-net: Self-organizing network for point cloud analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9397–9406, 2018.

[52] Ruihui Li, Xianzhi Li, Pheng-Ann Heng, and Chi-Wing Fu. Pointaugment: an auto-augmentation framework for point cloud classification. *CoRR*, abs/2002.10876, 2020.

[53] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *2010 IEEE computer society conference on computer vision and pattern recognition-workshops*, pages 9–14. IEEE, 2010.

[54] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *Advances in neural information processing systems*, 31, 2018.

[55] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding. *CoRR*, abs/1905.04757, 2019.

[56] Xinhai Liu, Zhizhong Han, Yu-Shen Liu, and Matthias Zwicker. Point2sequence: Learning the shape representation of 3d point clouds with an attention-based sequence to sequence network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8778–8785, 2019.

[57] Yongcheng Liu, Bin Fan, Gaofeng Meng, Jiwen Lu, Shiming Xiang, and Chunhong Pan. Densepoint: Learning densely contextual representation for efficient point cloud processing. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5239–5248, 2019.

[58] Yongcheng Liu, Bin Fan, Shiming Xiang, and Chunhong Pan. Relation-shape convolutional neural network for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8895–8904, 2019.

[59] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. *CoRR*, abs/2104.00678, 2021.

[60] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[61] Jameel Malik, Ibrahim Abdelaziz, Ahmed Elhayek, Soshi Shimada, Sk Aziz Ali, Vladislav Golyanik, Christian Theobalt, and Didier Stricker. Handvoxnet: Deep voxel-based network for 3d hand shape and pose estimation from a single depth map. *CoRR*, abs/2004.01588, 2020.

[62] Jameel Malik, Ahmed Elhayek, Fabrizio Nunnari, Kiran Varanasi, Kiarash Tamaddon, Alexis Héloir, and Didier Stricker. Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth. *CoRR*, abs/1808.09208, 2018.

[63] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jiashi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3d object detection. *CoRR*, abs/2109.02497, 2021.

[64] Roberto Martín-Martín, Clemens Eppner, and Oliver Brock. The RBO dataset of articulated objects and interactions. *CoRR*, abs/1806.06465, 2018.

[65] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet ++: Fast and accurate lidar semantic segmentation. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4213–4220, 2019.

[66] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. *CoRR*, abs/2109.08141, 2021.

[67] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. *Advances in neural information processing systems*, 27, 2014.

[68] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. *CoRR*, abs/1711.07399, 2017.

[69] Yancheng Pan, Biao Gao, Jilin Mei, Sibo Geng, Chengkun Li, and Huijing Zhao. Semanticposs: A point cloud dataset with large quantity of dynamic instances. *CoRR*, abs/2002.09147, 2020.

[70] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. *arXiv preprint arXiv:1807.06514*, 2018.

[71] Abhishek Patil, Srikanth Malla, Haiming Gang, and Yi-Ting Chen. The H3D dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes. *CoRR*, abs/1903.01568, 2019.

[72] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016.

[73] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J. Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *CoRR*, abs/1706.02413, 2017.

[74] Meytal Rapoport-Lavie and Dan Raviv. It's all around you: Range-guided cylindrical network for 3d object detection. *CoRR*, abs/2012.03121, 2020.

[75] Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015.

[76] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[77] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[78] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. NTU RGB+D: A large scale dataset for 3d human activity analysis. *CoRR*, abs/1604.02808, 2016.

[79] Hualian Sheng, Sijia Cai, Yuan Liu, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, and Min-Jian Zhao. Improving 3d object detection with channel-wise transformer. *CoRR*, abs/2108.10723, 2021.

[80] Hualian Sheng, Sijia Cai, Yuan Liu, Bing Deng, Jianqiang Huang, Xian-Sheng Hua, and Min-Jian Zhao. Improving 3d object detection with channel-wise transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2743–2752, 2021.

[81] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. PV-RCNN: point-voxel feature set abstraction for 3d object detection. *CoRR*, abs/1912.13192, 2019.

[82] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *CoRR*, abs/2004.09297, 2020.

[83] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[84] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. *CoRR*, abs/1505.00880, 2015.

[85] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. *CoRR*, abs/1912.04838, 2019.

[86] Xiao Sun, Zhouhui Lian, and Jianguo Xiao. Srinet: Learning strictly rotation-invariant representations for point cloud classification and segmentation. *CoRR*, abs/1911.02163, 2019.

[87] Weikai Tan, Nannan Qin, Lingfei Ma, Ying Li, Jing Du, Guorong Cai, Ke Yang, and Jonathan Li. Toronto-3d: A large-scale mobile lidar dataset for semantic segmentation of urban roadways. *CoRR*, abs/2003.08284, 2020.

[88] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. *CoRR*, abs/1908.04616, 2019.

[89] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[90] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[91] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *Acm Transactions On Graphics (tog)*, 38(5):1–12, 2019.

[92] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018.

[93] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. Squeezeseg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3d lidar point cloud. *CoRR*, abs/1710.07368, 2017.

[94] Wenxuan Wu, Zhongang Qi, and Li Fuxin. Pointconv: Deep convolutional networks on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9621–9630, 2019.

[95] Zhirong Wu, Shuran Song, Aditya Khosla, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets for 2.5d object recognition and next-best-view prediction. *CoRR*, abs/1406.5670, 2014.

[96] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[97] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas J. Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. *CoRR*, abs/2007.10985, 2020.

[98] Jiachen Xu, Jingyu Gong, Jie Zhou, Xin Tan, Yuan Xie, and Lizhuang Ma. Sceneencoder: Scene-aware semantic segmentation of point clouds with A learnable scene descriptor. *CoRR*, abs/2001.09087, 2020.

[99] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[100] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach and David Blei, editors,

Proceedings of the 32nd International Conference on Machine Learning, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR.

[101] Mingye Xu, Zhipeng Zhou, and Yu Qiao. Geometry sharing network for 3d point cloud classification and segmentation. *CoRR*, abs/1912.10644, 2019.

[102] Qiangeng Xu, Yin Zhou, Weiyue Wang, Charles R. Qi, and Dragomir Anguelov. SPG: unsupervised domain adaptation for 3d object detection via semantic point generation. *CoRR*, abs/2108.06709, 2021.

[103] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. *CoRR*, abs/2006.11275, 2020.

[104] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. *CoRR*, abs/2111.14819, 2021.

[105] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen 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*, pages 19313–19322, 2022.

[106] Wojciech Zaremba, Ilya Sutskever, and Oriol Vinyals. Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*, 2014.

[107] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. Pointweb: Enhancing local neighborhood features for point cloud processing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5565–5573, 2019.

[108] Hengshuang Zhao, Li Jiang, Jiaya Jia, Philip HS Torr, and Vladlen Koltun. Point transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16259–16268, 2021.

[109] Lichen Zhao, Jinyang Guo, Dong Xu, and Lu Sheng. Transformer3d-det: Improving 3d object detection by vote refinement. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(12):4735–4746, 2021.

[110] Lin Zhao and Wenbing Tao. Jsnet: Joint instance and semantic segmentation of 3d point clouds. *CoRR*, abs/1912.09654, 2019.

[111] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. *CoRR*, abs/2004.01314, 2020.

[112] Heliang Zheng, Jianlong Fu, Tao Mei, and Jiebo Luo. Learning multi-attention convolutional neural network for fine-grained image recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 5209–5217, 2017.

[113] Zhibin Zhong, Chi Zhang, Yuehu Liu, and Ying Wu. Viaseg: Visual information assisted lightweight point cloud segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1500–1504, 2019.