Vijay Jaisankar, Jaya Sreevalsan Nair

IMT2019525
Done as part of 20-credit project report

Literature Survey for the 20-credit project

Dec 10, 2023

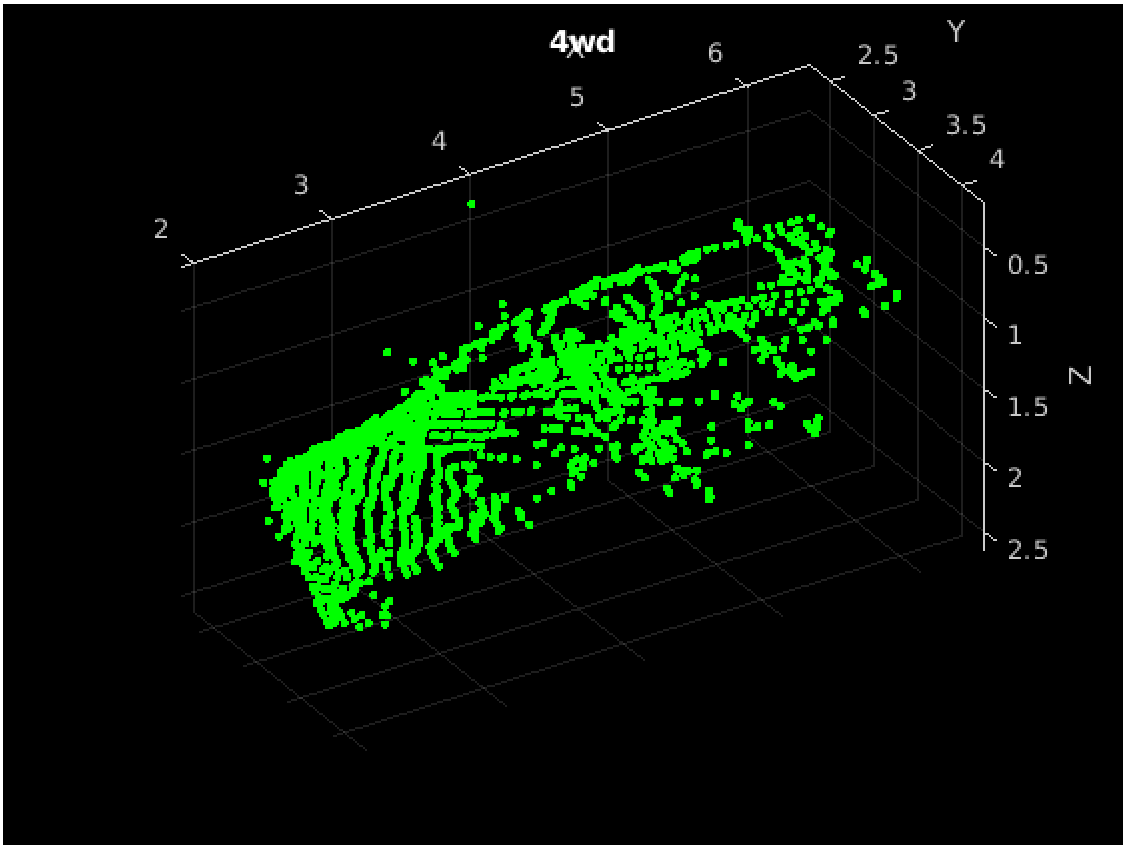
Abstract

In this report, we review existing research work in the areas of Point Cloud Understanding.

We summarise the key ideas behind these projects and note its relevance for the themes of our 20-credit project.

Point Clouds

3D data can be represented in different formats, like depth images, meshes, and point clouds. We choose to review point clouds as it does not involve discretisation and preserves the original geometric information of the scene it represents.



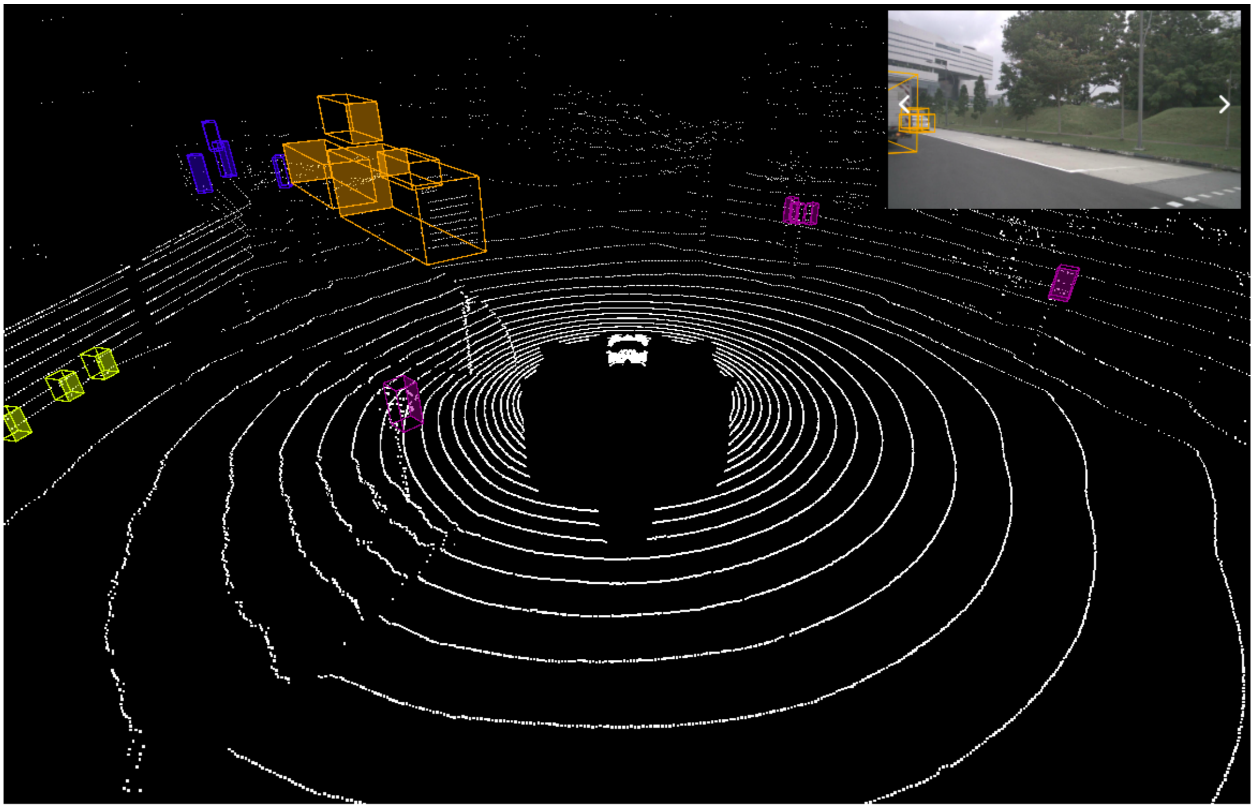
(Example of a point cloud with x, y, and z coordinates)

Datasets and Models

We now present our analysis of various datasets and papers in the field of point cloud analysis. These datasets pertain to the Automotive LIDAR domain.

NuScenes

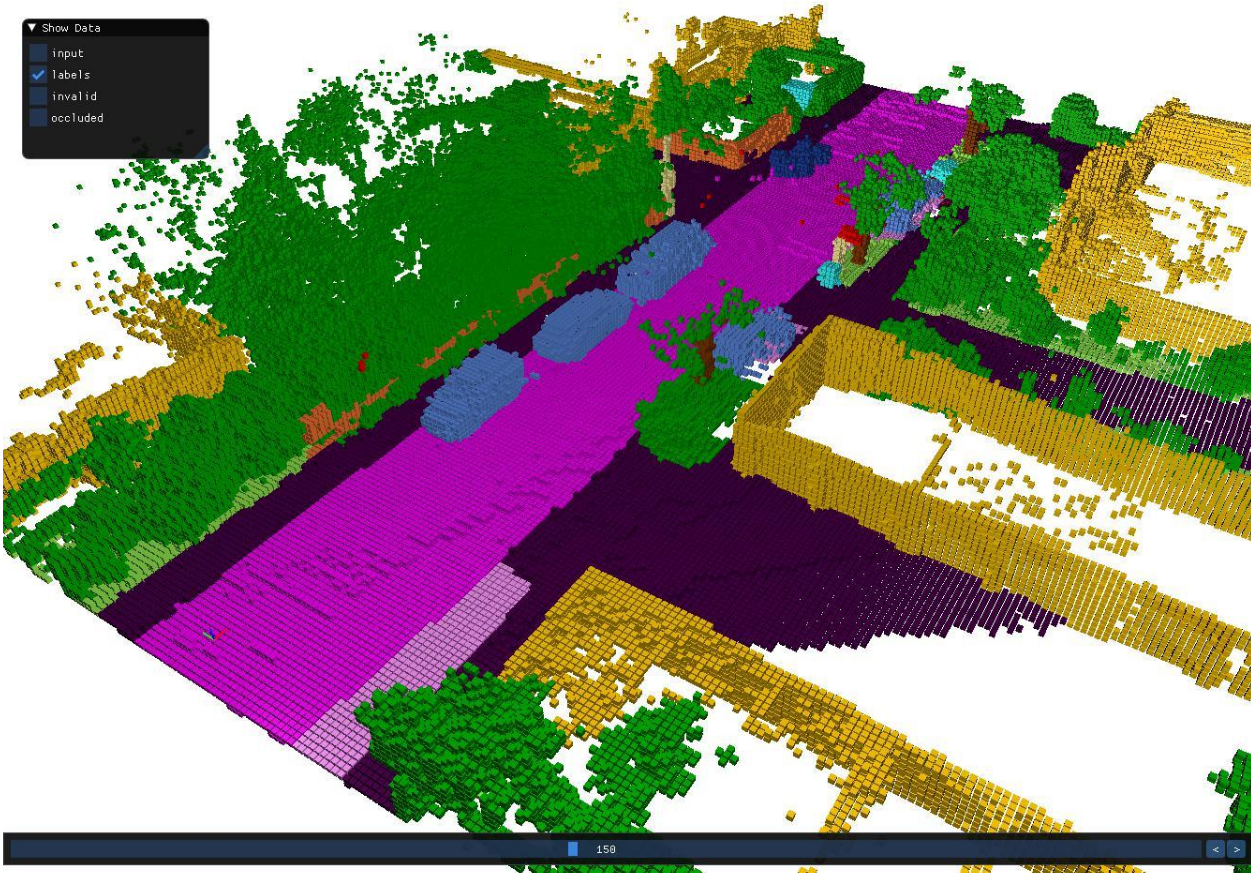
nuScenes is one of the gold standards in Autonomous driving applications. It comprises 1000 scenes from different cities, each 20 seconds in length.



(Visualisation of a frame from the nuScenes dataset)

SemanticKITTI

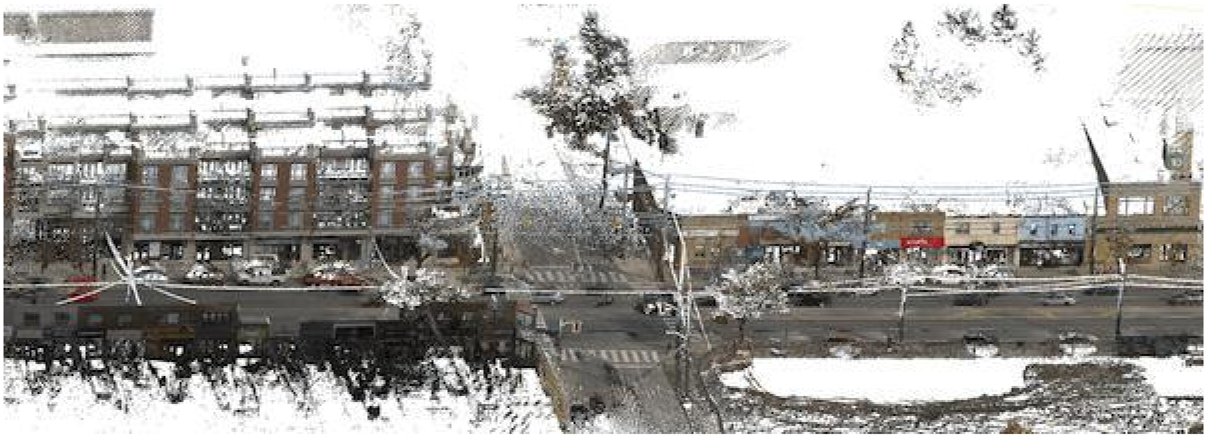
SemanticKITTI is a large-scale dataset for Semantic Segmentation applications. It consists of 23201 point clouds for training and 20351 point clouds for testing.



(Visualisation of a point cloud from the SemanticKITTI dataset)

Toronto3D

Toronto3D is a medium-scale dataset for Semantic Segmentation applications. It consists of ~73.8 points in total. Typically, researchers use 3 of the 5 scenes for training and allocate the remaining 2 for testing purposes.



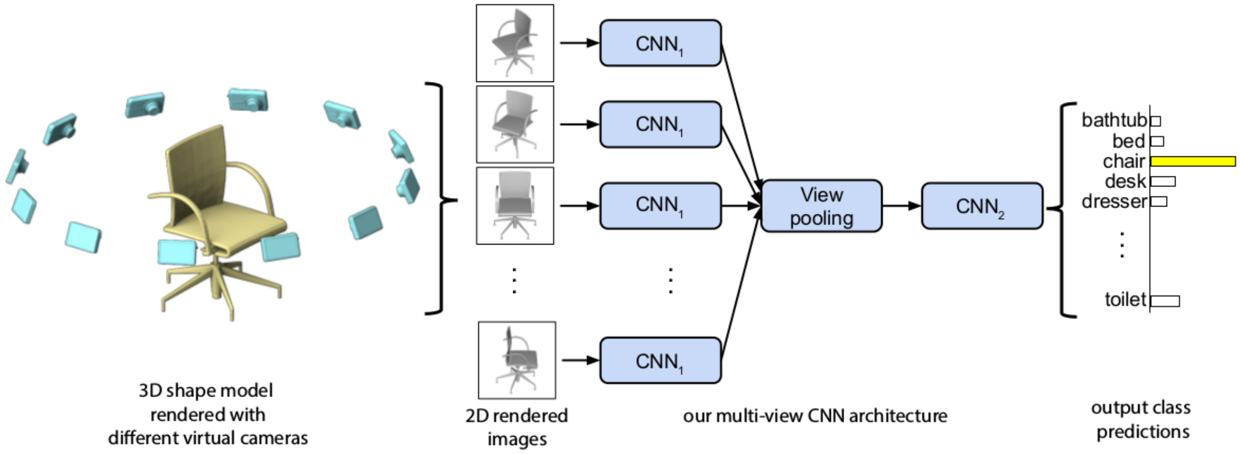
(Visualisation of a scene from the Toronto3D dataset)

MVCNN

Multi-view CNN (abbreviated as MVCNN) employs an image-processing approach to classify point clouds. In particular, given an object represented through its point cloud, it generates k renderings for it and calculates feature maps for each rendering through a CNN. These maps are then aggregated through a View Pooling layer (which computes element-wise maxpooling values across the renderings' outputs). This two-stage CNN network produces shape descriptors for the input 3D shape.

Relevance to our project

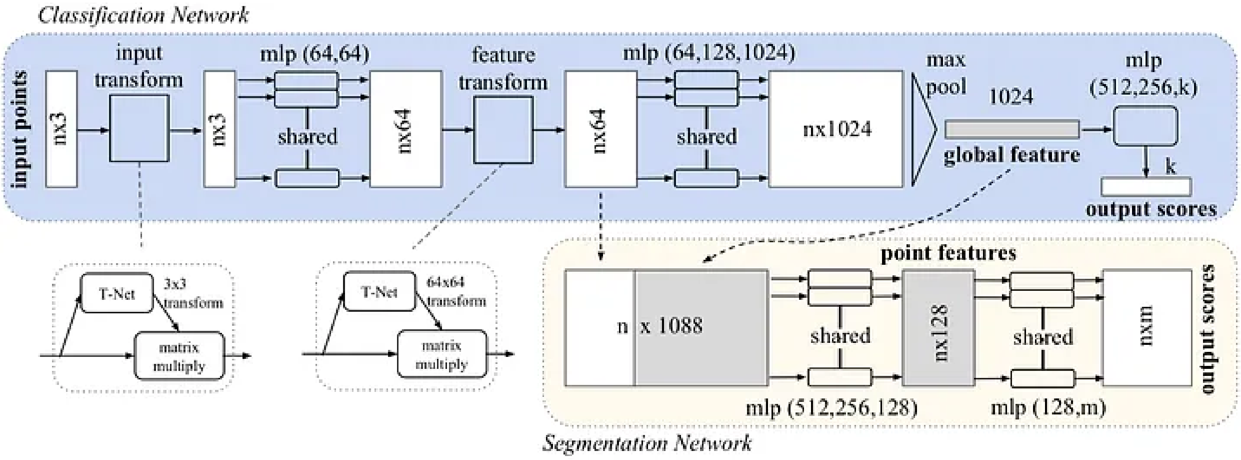
MVCNN is a light-weight architecture that can be used as a baseline for point cloud descriptors. However, this model does not work on point cloud inputs directly.



(MVCNN Architecture)

PointNet

Pointnet employs a simple architecture to ensure permutation (As point clouds are unordered sets, permutations like *{p1, p2, …, pn}* is equivalent to *{pn, …, p2, p1}*) and transformation invariance (classification and segmentation outputs should not vary with rotation and other transformations) in the processing of point clouds.



As shown in the architecture diagram, it employs a shared MLP to transform the input points to get a local point embedding. This is then maxpooled to get global features for the point clouds.

Relevance to our project

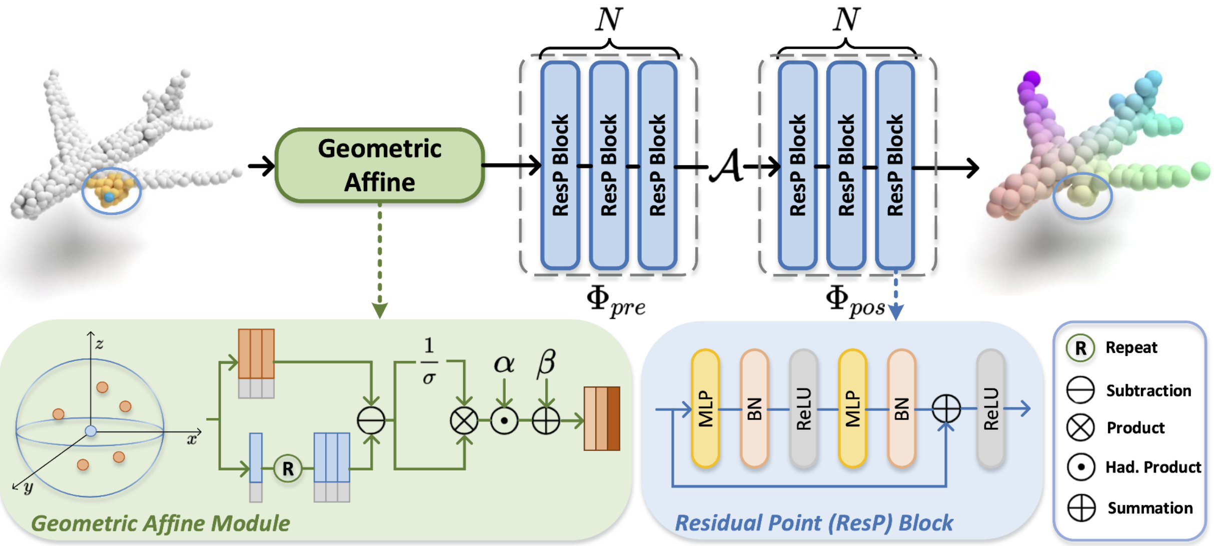
Pointnet is a light-weight architecture that can be used as a baseline for point cloud descriptors. We can use this either through the global feature vectors for a point cloud, or through the point features itself in semantic segmentation.

PointMLP

PointMLP uses a multi-stage architecture that successively increases the receptive field of the points processed. It employs an affine transformation module and an MLP block before and after aggregation of features. This progressively accounts for increasing inter-point dependencies.

Relevance to our project

PointMLP is easily implementable and can serve as rich feature vectors for point clouds as a whole.



(PointMLP architecture - note the MLP blocks before and after aggregation *A* - by stacking multiple such blocks, the receptive field is effectively increased.)

PointTransformer

The self-attention function, popularised by various transformer models in NLP and CV also holds the key property of permutation invariance owing to it being implemented as a set function. In this regard, PointTransformer proposes a special form of self-attention that can be used on point cloud data and shows strong performance on various benchmarks. For positional embeddings, a learnable MLP is projected on the difference of point position vectors. Then, successive downsampling (through FPS) of the point cloud cardinality is performed to enable more granular self-attention through applying it in a local neighbourhood of the points. In the event of semantic segmentation applications, there is a symmetric set of upsampling operations.

Point Transformer layer

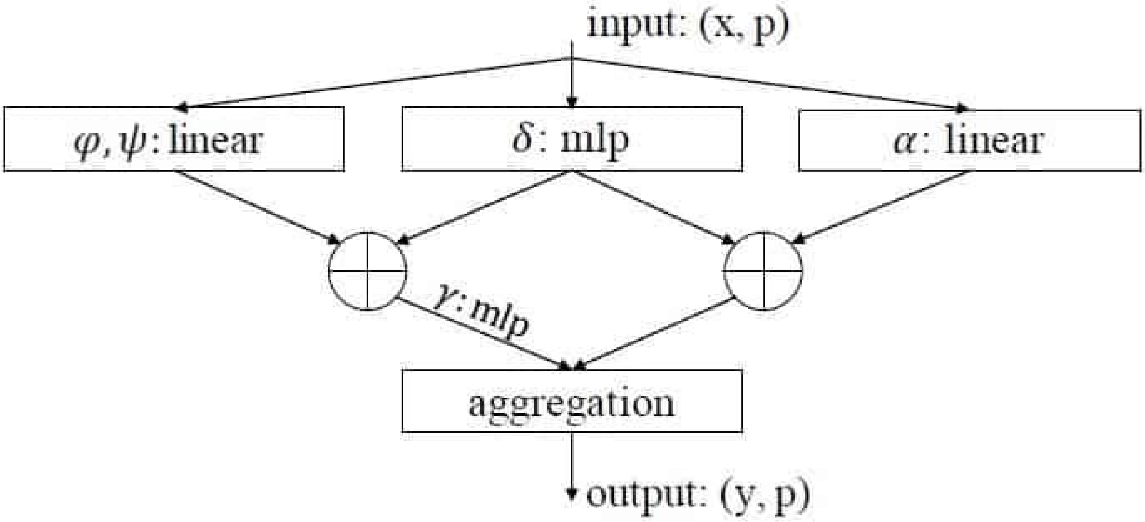


Image from the original paper

(Point transformer architecture with the Query, Key, and Value layers annotated - this block is successively applied to extract feature vectors and then projected into downstream tasks through MLP blocks.)

Relevance to our project

Point Transformer is a strong baseline for tasks like semantic segmentation and classification. However, training such models from scratch needs high computational resources owing to its deep architecture.

RandLANet

RandLANet is a semantic segmentation model that can operate on large point clouds in one-shot (as opposed to breaking it down and computing results on each chunk) without relying on time-consuming preprocessing or voxelisation steps. It uses Random Point Sampling (RPS) instead of other methods like FPS and IDIS; and local feature aggregation is done through a combination of spatial encoding (for a given point, it does a KNN search and then embeds the query point through an MLP concatenation of the neighbours' embeddings), attentive pooling (which aggregates the neighbouring feature sets, better than maxpooling which loses information), and dilated residual block(This is akin to skip connections, wherein the model can learn multiple levels of receptive fields). This 3D encoder-decoder architecture with skip connections is fast in practice and scales well to arbitrary-dimensional point clouds.

Relevance to our project

RandLANet is a fast and performant model for semantic segmentation and has proven performance in automotive applications.

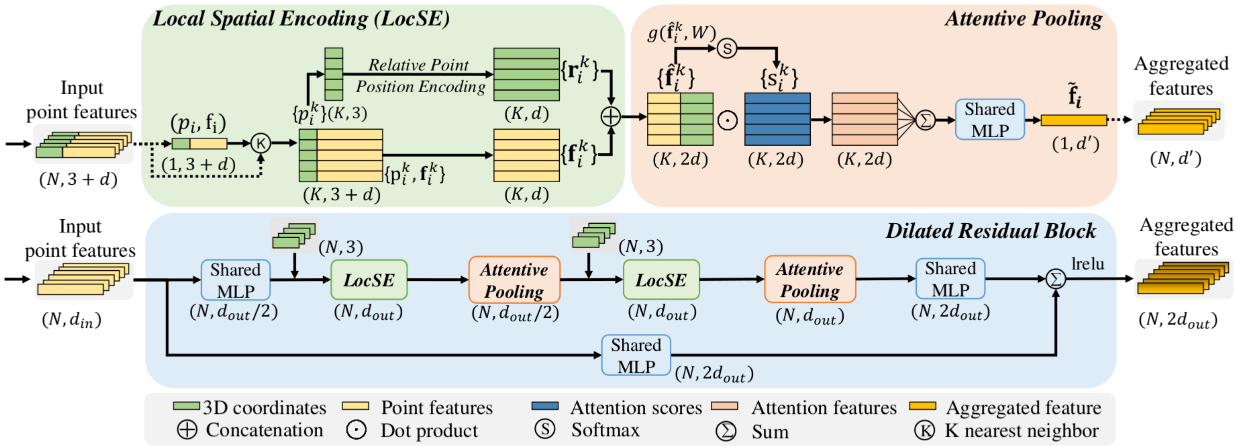


Figure 3. The proposed local feature aggregation module. The top panel shows the location spatial encoding block that extracts features, and the attentive pooling mechanism that weights the most important, based on the local context and geometry. The bottom panel shows how two of these components are chained together, to increase the receptive field size, within a residual block.

(Architecture of RandLANet: Figure from the paper)

DGCNN

Dynamic Graph CNNs involve transforming point clouds into graphs and then applying the principles of convolutions to graphs in lieu of local geometric information. This operation of {transforming local points to graph, applying graph convolution} can be stacked repeatedly to learn semantic relationships between groups of points. The embeddings for each point that are accumulated is able to capture semantic groupings across large distances.

Relevance to our project

DGCNN is an interesting application of Graph Neural Networks (GNNs) that scales well to increased model size. Using such architectures constitute interesting future work.

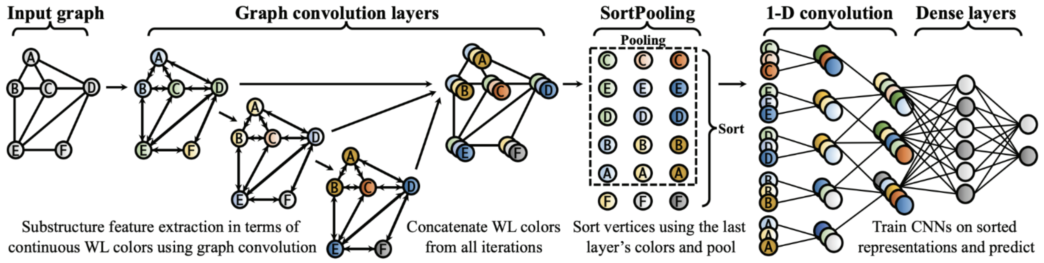

Input graph
Graph convolution layers
SortPooling
Pooling
Sort
1-D convolution
Dense layers
Substructure feature extraction in terms of continuous WL colors using graph convolution
Concatenate WL colors from all iterations
Sort vertices using the last layer's colors and pool
Train CNNs on sorted representations and predict

Figure 2: The overall structure of DGCNN. An input graph of arbitrary structure is first passed through multiple graph convolution layers where node information is propagated between neighbors. Then the vertex features are sorted and pooled with a SortPooling layer, and passed to traditional CNN structures to learn a predictive model. Features are visualized as colors.

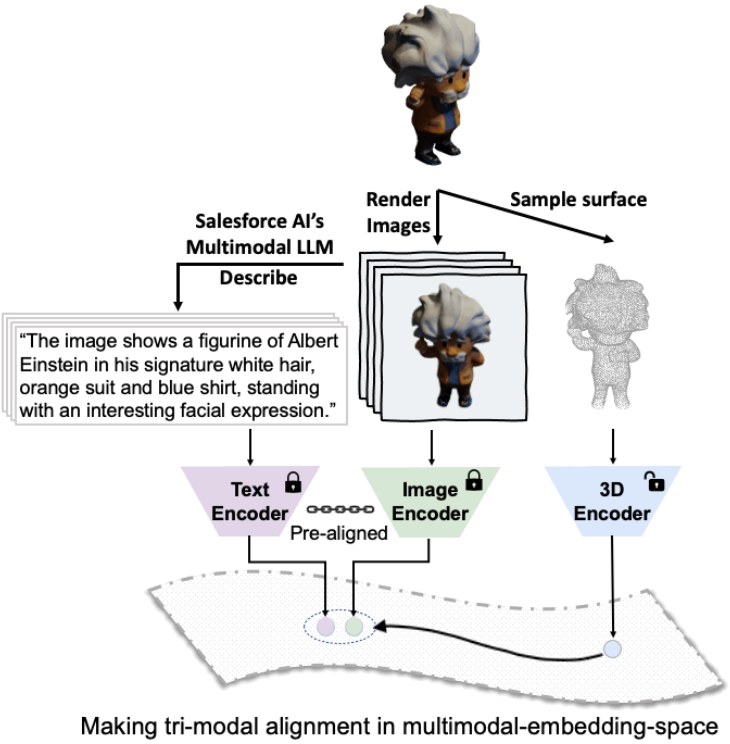
(Architecture of DGCNN: from the paper, the SortPooling operation reads the graph in a meaningful order that is transferable to other point clouds.)

ULIP

ULIP is a seminal work in the domain of Multimodal deep learning with point clouds. It learns a unified representation of text, images, and point clouds. It uses a backbone of a frozen pre-aligned vision-language model like CLIP that can already project images and text into a common feature space. Then, through a small set of {text, image, point cloud} triplets, it uses contrastive learning to align the 3D features with the texts and images.

Relevance to our project

ULIP can be used for cross-domain tasks like Point cloud <-> Image and to find candidates for point cloud editing through similarity search metrics.



"The image shows a figurine of Albert Einstein in his signature white hair, orange suit and blue shirt, standing with an interesting facial expression."

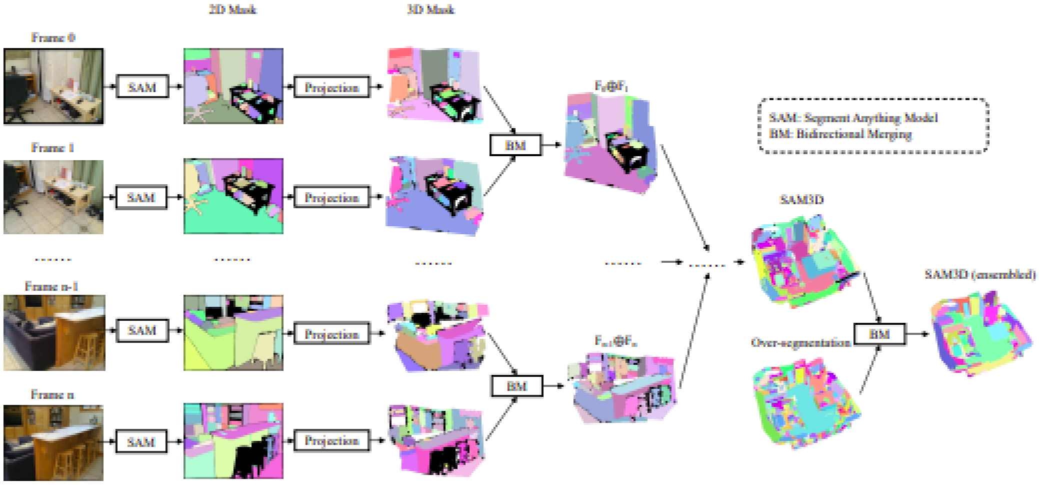
(ULIP architecture - the alignment is done through multi-modal contrastive learning through image/point-cloud and text/point-cloud alignment)

SAM3D

SAM3D extends the powerful Segment Anything Model (SAM) to 3D settings by transferring the SAM outputs of corresponding RGB images to 3D space. This model works on 3D scene videos and uses iterative merging to generate a segmentation output. The main steps of this pipeline are Generating Masks (through SAM) and then mapping them into 3D through depth information; Merging two adjacent point clouds (through their proposed Bidirectional group overlap algorithm) and then merging the scenes' point cloud outputs group-wise to get a semantic segmentation output.

Relevance to our project

SAM3D is an interesting application of using 2D priors for 3D segmentation, it can be used to evaluate the efficiency of such methods in the automotive domain.



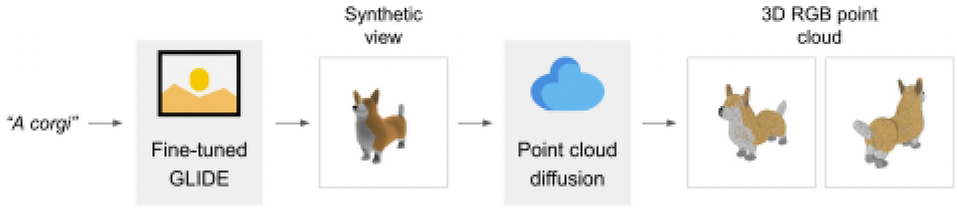
(Overview of SAM3D Architecture: from the paper)

PointE

PointE generates point clouds conditioned on texts and images. It follows a two-stage process, first generating a single synthetic view image using a text-to-image diffusion model (fine-tuned GLIDE), and then using another diffusion model to produce a point cloud conditioned on the synthetic view image. The second stage of point cloud generation is in itself a two-hop approach, wherein a permutation-invariant diffusion model generates a low-resolution point cloud (1024 points) and then a smaller diffusion model conditioned on the low-resolution point cloud upsamples it to a larger coloured point cloud (4096 points).

Relevance to our project

PointE is an efficient approach to convert images to point clouds. The checkpoints have multiple versions which vary in size, enhancing the flexibility of use cases in low-resource settings.



(PointE architecture: from the paper)

On Isometry Robustness of Deep 3D Point Cloud Models under Adversarial Attacks

This paper proposes an adversarial attack on point cloud models called Thompson Sampling with Restricted Isometry Property. Contrary to some other attacks, this attack invokes global geometry properties like robustness under Euclidean distance projection (||f(x) - f(y)|| = ||x-y||). Thompson sampling is a probability distribution over actions for an agent which balances the goals of exploration and exploitation. In this work, the actions involve sampling a parameterised isometry transformation in the local neighbourhood of a point cluster. This RL paradigm results in strong adversarial transfer to multiple point cloud models.

Relevance to our project

Isometry Robustness is an important property to benchmark the models we use. The fact that global geometry is preserved makes it especially relevant to automotive settings.

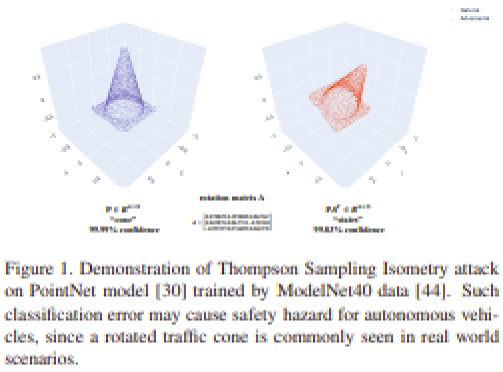
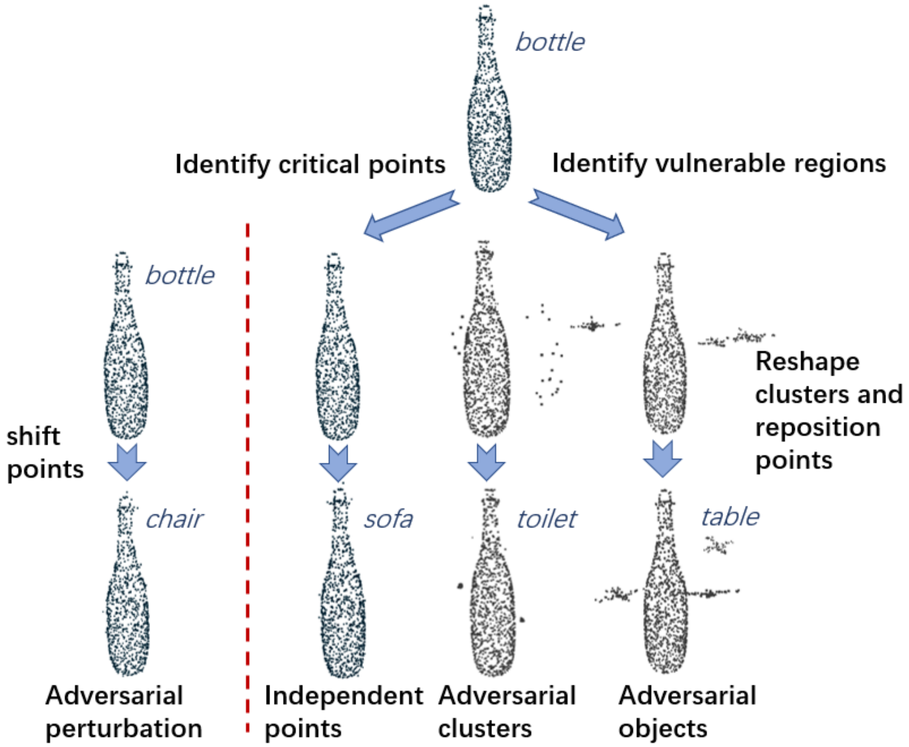


Figure 1. Demonstration of Thompson Sampling Isometry attack on PointNet model [30] trained by ModelNet40 data [44]. Such classification error may cause safety hazard for autonomous vehicles, since a rotated traffic cone is commonly seen in real world scenarios.

(Example of Thompson Sampling Attacks)

Generating 3D Adversarial Pont Clouds

This paper proposes a set of adversarial attacks on point cloud models - adversarial point perturbation (shifting existing points negligibly under the conditions of minimal changes to L2 norm) and adversarial point generation (placing a set of independent points or clusters of points that is invisible to the human eye - measured through Hausdorff distance norms). It formulates the same as gradient-based optimisation algorithms and also delivers stunning performance on Pointnet models.



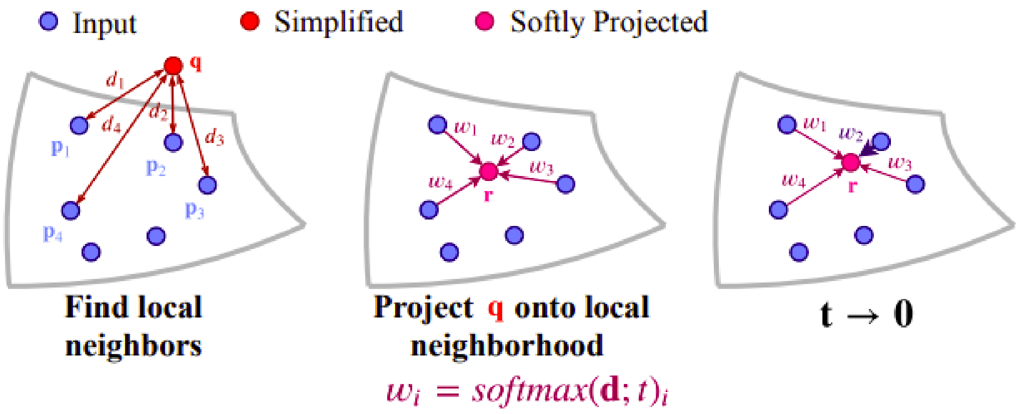
(Example of Adversarial perturbations and additions)

Relevance to our project

The distance norm measures proposed in this paper can help us validate other attacks and the dataset released can augment our work on multimodal alignment.

SampleNet

Sampling point clouds to extract the most relevant points can produce computational benefits. This paper notes that classical methods like FPS do not consider downstream tasks and hence propose Soft projection, a differentiable sampling strategy to extract a strong subset of points from the point cloud. As the temperature parameter of this soft projection goes to zero, the selected points optimising their soft projection scores with their neighbours approximate sampling through a non-differentiable step function.



(Architecture of SampleNet: from the paper)

Relevance to our project

Effective sampling of LIDAR data for semantic segmentation can reduce the memory burden on the models and increase training time. It can also enable more effective distillation experiments.

Conclusion

In this report, we have reviewed existing research work in the areas of Point Cloud Understanding.

References

[1912.12033] Deep Learning for 3D Point Clouds: A Survey

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, Oscar Beijbom; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11621-11631

SemanticKITTI

Toronto-3D: A Large-Scale Mobile LiDAR Dataset for Semantic Segmentation of Urban Roadways

PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation

[2202.07123] Rethinking Network Design and Local Geometry in Point Cloud: A Simple Residual MLP Framework

Point Transformer

RandLA-Net: Efficient Semantic Segmentation of Large-Scale Point Clouds

DCGNN: a single-stage 3D object detection network based on density clustering and graph neural network | Complex & Intelligent Systems

ULIP: Learning a Unified Representation of Language, Images, and Point Clouds for 3D Understanding

[2306.03908] SAM3D: Segment Anything in 3D Scenes

[2212.08751] Point-E: A System for Generating 3D Point Clouds from Complex Prompts

On Isometry Robustness of Deep 3D Point Cloud Models Under Adversarial Attacks

Generating 3D Adversarial Point Clouds

SampleNet: Differentiable Point Cloud Sampling