J. Vis. Commun. Image R. 100 (2024) 104111

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Full length article

Adapting projection-based LiDAR semantic segmentation to natural domains☆

Kelian J.L. Massa *, Hans Grobler

Department of Electrical, Electronic and Computer Engineering, University of Pretoria, Hatfield, Pretoria, 0028, South Africa

ARTICLE INFO

MSC:
62F15
68T05
68T10
68T45
68U10

Keywords:
Semantic analysis
Semantic segmentation
LiDAR
Natural data
Projection
Fusion

ABSTRACT

In this paper, an approach to the semantic segmentation of 3D LiDAR point clouds obtained from natural scenes is introduced. Using a state-of-the-art projection-based semantic segmentation model as the core segmentation network, several recent advances in projection-based 3D semantic segmentation methods are aggregated into a single model. These adaptions include: scan unfolding, soft-kNN post-processing, and multi-projection fusion. A novel Naïve Bayesian approach to multi-projection fusion which weights class probabilities based on the outputs of the base classifiers is proposed to further increase robustness.

Quantitative and qualitative evaluations on several datasets, including scenes from both urban and natural environments; show that aggregating these adaptions into a single model can further improve the accuracy of state-of-the-art projection-based approaches. Finally, it is demonstrated that the novel Naïve Bayesian approach to multi-projection fusion addresses a number of the challenges inherent to natural data while also improving results on urban data.

1. Introduction

Semantic analysis is a novel and fast-growing field of computer vision which automates the extraction of image and scene descriptions according to human perception. This is essential to broaden the uses of computer vision to human-like tasks, providing more meaningful descriptions than the traditional low-level properties and features of the scene. Semantic segmentation is a crucial component of semantic analysis as it predicts class labels for each individual sensory data point, providing a rich analysis of scene semantics.

Researchers have made significant progress in solving the problem of semantic segmentation in both two-dimensional (2D) images and three-dimensional (3D) scenes. However, most of the work on semantic segmentation of 3D scenes has been performed on artificial and urban scenes [1], containing large amounts of man-made structures. Off-road scenes with large amounts of natural scenery and vegetation have seen far less usage in comparison. Urban scenery tends to contain distinct objects with structured features, clearly defined boundaries, and relatively even distributions of different classes. While most datasets contain some vegetation and natural scenery, these are often grouped into a few broad categories such as terrain and vegetation, making them easy to distinguish from the surrounding urban scenery.

On the other hand, natural scenes typically exhibit a degree of randomness and fluidity, marked by an irregular distribution of classes where a few classes dominate the scene's overall composition. An off-road setting, such as a natural reserve or farm, often consists of diverse and uneven terrain along with a wide variety of objects, typically in limited quantities. This results in a markedly imbalanced distribution of class labels. Coupled with the inherent variability and irregularity of natural objects, achieving high intersection over-union scores becomes notably more demanding. Consequently, the problem of 3D semantic segmentation of natural scenes is yet to be solved with reasonable accuracy [2,3].

Despite the inherent complexities, humans are able to easily perceive natural environments. Expanding the work done in 3D semantic segmentation specifically to natural scenes would provide for significantly more flexible and adaptable computer vision based systems. This would largely have applications in robotics that requires some interpretation of surroundings in natural environments. For instance, autonomous vehicles such as self-driving cars operating in off-road environments, autonomous tractors used in farming, and surveillance drones operating in natural environments, could benefit from the ability to interpret and understand their surroundings more accurately [4,5]. By enabling these systems to perceive and analyse natural scenes

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail address: u17000841@tuks.co.za (K.J.L. Massa).

https://doi.org/10.1016/j.jvcir.2024.104111
Received 18 August 2023; Received in revised form 3 February 2024; Accepted 2 March 2024
Available online 5 March 2024

more effectively, a solution to semantic segmentation of natural data would contribute to the development of more robust and reliable computer vision-based technologies.

A number of remote sensing approaches to collecting 3D data exist, some of the most popular approaches include light detection and ranging (LiDAR), photogrammetry, and structured light scanning. It should be noted that 3D semantic segmentation datasets in outdoor environments are most commonly obtained through either LiDAR or photogrammetry. 3D information obtained through LiDAR is generally highly accurate and is not distorted by variance in illumination. LiDAR also adds reflectance information which helps to distinguish object surfaces. In contrast, cameras used in photogrammetry obtain dense colour and texture information, providing fine-grained semantic details which are highly useful in semantic segmentation. Photogrammetry-based systems are also generally more accessible and less costly to implement. This research limits the scope of 3D data representations to LiDAR-based data due to the greater availability of LiDAR-based outdoor datasets. The large amounts of vegetation in natural data also create complex 3D structures with minimal variance in colour and contrast, properties which would significantly impact the accuracy of 3D information obtained through photogrammetry. Furthermore, limiting the scope to a single remote sensing method allows for the development of more advanced methods tailored to LiDAR's unique format.

Semantic segmentation of point clouds is particularly challenging due to the unique properties of point clouds such as sparsity, randomness and lack of structure. While deep learning has led to the biggest developments in semantic segmentation of images, the general irregularity of point clouds means these approaches must be adapted significantly in order to be applied directly to point clouds. Several deep learning approaches have been developed to solve this problem, these can be divided into two broader categories: point-wise approaches and projection-based approaches [1], otherwise known as direct and indirect approaches. Point-wise approaches are applied directly to the raw point cloud data with no prior transformation, while projection-based approaches pre-process the raw 3D data using some method of projection to obtain a discrete and regularised representation of the original point cloud. Most commonly projection formats are some form of 3D grid representation (voxel grids) or 2D images.

While point-wise approaches have achieved state-of-the-art accuracy [6], they lack a means to efficiently scale-up to large point sets. Conversely, projection-based approaches are inherently more scalable and have been shown to achieve near state-of-the-art accuracy while running significantly faster than other approaches [7]. Spherical projection-based approaches in particular have shown the most progress in solving their inherent problems, and have achieved better results than direct approaches in urban and natural semantic segmentation [2,8,9]. They are also particularly well suited to LiDAR-based data, as this method of projection essentially aims to reproduce the scan in the original LiDAR range image format. Furthermore, researchers have made several adaptions to the pre-processing (projection) and post-processing pipelines which reduce unwanted artefacts and increase robustness in projection-based approaches [10–12]. These adaptions have yet to be aggregated into a single model, indicating a clear gap in research. Multi-projection fusion is one such adaption. Initially limited to a variant of weighted majority voting for ensembling, it has shown promising results in enhancing accuracy and robustness in LiDAR semantic segmentation [11,13]. While this adaption led to improvements in accuracy and robustness; more complex approaches to ensemble classification exist which provide for more accurate results than weighted majority voting [14]. Notably, recent advances like AMVNet's late fusion [15] and GFNet's geometric approach [16] in LiDAR semantic segmentation demonstrate the potential of advanced fusion techniques, particularly beneficial for resource-constrained robotics like autonomous vehicles due their underlying usage of resource efficient projection-based approaches. These projection-fusion approaches underscore the potential for advanced fusion techniques to significantly elevate the performance of projection-based LiDAR semantic segmentation beyond traditional methods.

In this work a number of adaptions are made to the current state of the art in projection-based 3D semantic segmentation with the aim of creating a model optimised for robustness and segmentation of natural scenery. Several of the latest advances in projection-based approaches are aggregated into a single semantic segmentation model, including: scan unfolding, soft-kNN post-processing, and multi-projection fusion. Furthermore, a novel Naïve Bayesian approach to multi-projection fusion is introduced. The implemented model uses SalsaNext [8] as the core segmentation network, as it is one of the few semantic segmentation approaches which has been evaluated on natural data [2] . However, adaptions are designed to be applied to the pre-processing and post-processing pipelines of any projection-based segmentation model. Quantitative and qualitative experiments are conducted on a number of datasets which cover both urban and natural scenery, including SemanticKITTI [17], SemanticPOSS [18], and RELLIS-3D [2].

2. Related work

In this section, a brief survey of recent work developing projection-based approaches to semantic segmentation of 3D point cloud data will be provided. Since this research aims to enhance existing network designs of projection-based approaches, emphasis is placed on prior work which improves the pre-processing and post-processing pipelines of projection-based approaches independent of network design.

2D projection-based approaches represent 3D data with 2D descriptors to directly apply existing image-based approaches to the projected point cloud. One of the simplest methods of 2D projection is multi-view, where the 3D point cloud is projected from multiple different views onto 2D images. Researchers have developed a variety of approaches based on multi-view which have shown success in the problem of 3D semantic segmentation [19–21]. However, these approaches are generally poorly suited to complex scenes with multiple objects due to a loss of spatial information.

Many 3D projection-based approaches discretise the point cloud through voxel-based approaches. Voxel-based approaches project the unstructured point cloud into a more regular volumetric occupancy grid consisting of a number of equally sized 3D cells (voxels). This regularised 3D representation of the point cloud is similar in properties to a 2D image, as it is essentially a 3D grid. Researchers have used these properties to directly apply 3D CNN to the voxel-grid for semantic analysis [22,23]. However, this approach leads to computational inefficiencies due to the sparsity of point clouds. The large amounts of empty space inherent to point clouds will lead to numerous empty voxels which are still allocated resources. Researchers have attempted to address this by dividing the voxel-grid with a computational graph using the kd-tree (Kd-Net [24]) or octree (Oct-Net [25]) structures, and applying the 3D convolutions level by level. This allows the model to exploit the sparsity of 3D data, as resources may be dynamically allocated based on the data density of different regions. However, these methods significantly increase the complexity of data structures and are thus difficult to implement efficiently.

The most popular method of projection to address sparsity in point cloud data is to project the 3D point cloud to a 2D range image either from a top-down bird's eye view (BEV) perspective [26,27] or a spherical (panoramic) perspective [8,9,28]. These projection methods provide for dense regularised representations that lead to more efficient computations, thus addressing point cloud sparsity. This has led to a number of projection-based approaches which have achieved near state-of-the-art results [7,15,16]; as well as approaches designed specifically for efficiency and real-time implementation [29–31]. One

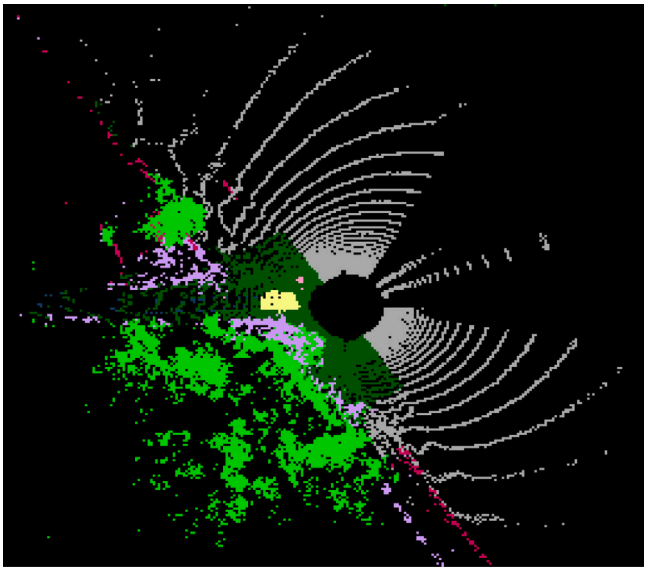


Fig. 1. An example of the BEV projection result on a ground-truth labelled RELLIS-3D scan. Colours are assigned by label.

consequence of an image-based representation is point occlusions, both from BEV and panoramic perspectives.

Post-processing in projection-based approaches generally aims to correct the errors in the model's predictions that the projection method introduced, such as information loss and re-projection error. This was initially performed in 2D with conditional random field (CRF) post-processing [28] and later in 3D with KNN post-processing [8,9] as well as nearest label assignment (NLA) [30]. Methods using 3D post-processing however still exhibit significant re-projection error. Other recent works have addressed this with adaptions such as systematic scan unfolding [10,32]; multi-projection fusion [11,13,15,16]; and the addition of 3D point-wise learnable components [12].

3. Method

In this section a detailed description of the method used to generate results is provided. The method is divided into a number of steps, including: projection to 2D; the core segmentation model; post-processing; and multi-projection fusion.

3.1. Projection methods

The 3D LiDAR data is represented in 2D through projection. Since this work makes use of multi-projection fusion, multiple methods of projections are utilised. These include Cartesian BEV projection, spherical projection, and scan unfolding.

3.1.1. BEV projection

BEV was implemented similarly to [26], where the point cloud is first collapsed along the z axis so that a flattened 2D version of the point cloud is produced. It is then discretised using a fixed rectangular grid to produce the resultant 2D image. This results in a multi-channel 2D image of the point cloud as seen from above. The main drawback of this is top-down point occlusions. Some approaches attempt to address this with additional channels such as average intensity and elevation of stacked points [27]. These additional channels were found to provide little to no benefit for this problem, thus no additional channels were included. Fig. 1 shows an example of BEV projection on the RELLIS-3D dataset.

3.1.2. Spherical projection

Spherical projection was implemented as in [28]. The 3D LiDAR point cloud is projected onto a spherical surface to generate a range image similar to the LiDAR's native range image. The similarity of this approach to LiDAR's native format makes this approach particularly well suited to projecting 3D LiDAR scans. Each raw 3D LiDAR point (x, y, z) is projected a 2D image coordinate (u, v) as

$$\begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} \frac{1}{2}\left[1 - \arctan(y,\, x)\,\pi^{-1}\right] w \\ \left[1 - (\arcsin(z,\, r^{-1}) + f_{down})\, f^{-1}\right] h \end{pmatrix}$$

where h and w of the desired height and width of the projected image, r is the range to each point $r = \sqrt{x^2 + y^2 + z^2}$, and f is the LiDAR sensor's vertical field of view $f = |f_{down}| + |f_{up}|$.

3.1.3. Scan unfolding

Traditional spherical projection back-projects the motion corrected LiDAR data to a spherical surface. This has the drawback of mutual point occlusions due to motion correction. Scan unfolding is a similar approach which can transform the raw LiDAR data (no motion correction) to its original range image format with minimal point occlusions [10]. Fig. 2 shows a comparison of the two approaches on the same scan. This is only possible by exploiting the distinct data representations in certain datasets, in this work it is used with both SemanticKITTI and SemanticPOSS. The 3D scan is unfolded as in Algorithm 1.

Algorithm 1: Scan Unfolding

Data: $N \times 3$ array of $points$, thresh = 0.3
Result: 2D projected points with shape $H \times W$
$depth \leftarrow \sqrt{points_x^2 + points_y^2 + points_z^2}$
$rows, columns \leftarrow unfold(points)$
sort $columns$ and $rows$ by decreasing $depth$
$projection \leftarrow$ array of zeros with shape $H \times W$
$projection[columns, rows] = depth$
Function *Unfold (points)*:
  $\phi \leftarrow \text{atan2}(points_x, points_y)$
  $jump \leftarrow \phi[1:] - \phi[:-1] > thresh$
  $rows \leftarrow$ cumulative sum over $jump$
  $columns \leftarrow W(\pi - \phi)/2\pi$

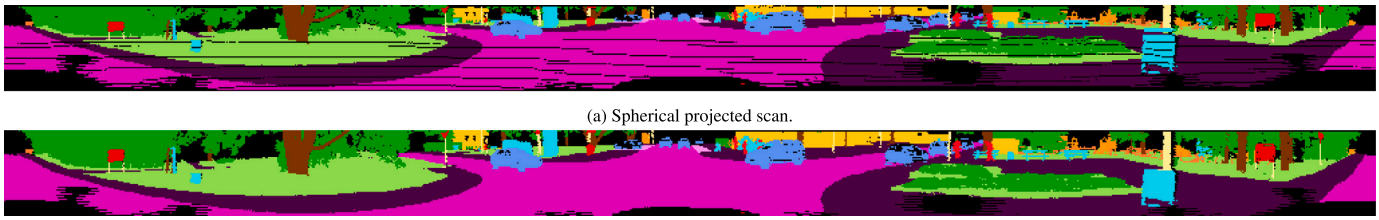
While this algorithm is effective in reverse engineering the original sensor range image, interestingly, certain sensors provide direct access to the sensor image. This means that no projection is necessary in these cases, providing for more efficient implementations.

3.2. Segmentation model

This research makes use of *SalsaNext* [8] as the core segmentation model, a state-of-the-art projection-based approach based on *SalsaNet* [27]. *SalsaNext* is based on an encoder–decoder architecture where the encoder consists of a residual dilated convolution stack with gradually increasing receptive fields, followed by a decoder which upsamples and fuses the features extracted by the encoder.

3.3. Post-processing

Post-processing in this context refers to processing performed directly on the output of the segmentation network. It is necessary to make this distinction as this work includes multi-projection fusion, which is performed after this post-processing step. Consequently, it is necessary to use post-processing which estimates label probabilities rather than a single label. Alnaggar et al. address this with non-sparse nearest neighbour post-processing using soft-voting [11]. However, their approach only detects nearest neighbours using the 2D projected coordinates, so that points that are nearby in 2D but distant in



(a) Spherical projected scan.

(b) Unfolded scan.

Fig. 2. Comparison of spherical-projected range image and scan unfolded range image for an example SemanticKITTI scan. The results show occlusion lines in the spherical projected image (a) but not in the scan unfolded image (b).

3D have equal weighting. The current state of the art in projection-based approaches generally makes use of k-nearest neighbours (KNN) post-processing with hard-voting [8,9]. This method of KNN takes an efficient approach to detecting nearest neighbours in 3D by first detecting the nearest-neighbours in 2D and then filtering them using the depth channel. Thus, a soft-voting variant of this method of KNN post-processing was used in this work.

3.4. Multi-projection fusion

In this work each projection method is assigned to a separate base classifier, so that the fusion problem is to combine the post-processed probabilities of a BEV classifier with a spherical or scan unfolded classifier to obtain a single set of predicted labels. The standard approach is a form of majority voting, where the outputs of both models are simply added together and the most probable label selected [13]. This approach is inherently limited as it assumes equal performance of the base classifiers. Weighted majority voting addresses this to some extent by weighting model predictions according to measured performance, however more advanced probabilistic approaches generally perform better for this purpose [14]. As with many classification problems, making the Naïve Bayes (NB) assumption of conditional independence of the features given the class can greatly reduce the complexity of the problem while providing a reasonably accurate estimate of the relevant posteriors. This work uses a NB-ensembling approach for fusion similar to that described in [14], with the main difference being the calculation of the likelihood term with the full set of probabilities for each label rather than only using the most probable label. This section begins with formulating the standard approach to NB ensembling, after which the probability based variant is formulated. The problem in the standard NB approach is to estimate the posterior $P(\omega_k|\mathbf{s})$ where $\mathbf{s}$ is the vector of class predictions by each classifier $\mathbf{s} = s_1, \dots, s_L$. Using Bayes theorem and removing the denominator as a normalising factor ($P(\mathbf{s})$ is independent of the class label) the posterior may be rewritten as

$$P(\omega_k|\mathbf{s}) \propto P(\mathbf{s}|\omega_k)P(\omega_k) \tag{1}$$

where $P(\mathbf{s}|\omega_k)$ is the likelihood describing the probability of the model predicting a label given the true class label, and $P(\omega_k)$ is the prior describing the probability of a class with no evidence. The prior can be estimated as the number of observations of a class N_k over the total number of observations N. However, estimating the priors in this manner can reduce segmentation accuracy due to the differences in class distributions between training and test sets, a problem inherent to most outdoor segmentation datasets. Since there is no other relevant evidence for estimating the priors, equal priors can be assumed using the principle of indifference. Making this assumption, the likelihood $P(\mathbf{s}|\omega_k)$ can be expanded as

$$P(\mathbf{s}|\omega_k) = \prod_{i=1}^{L} P(s_i|\omega_k) \tag{2}$$

by making the Naïve Bayes assumption of conditional independence of the model predictions on the true class label. A $c \times c$ confusion matrix CM_i is computed for each model D_i by applying it to the training set. The notation $cm^i_{k,s}$ then denotes the (k,s)th entry of the confusion matrix, which is the number of elements with true class label ω_i that were assigned the predicted label ω_s by classifier D_i. Using this notation $P(s_i|\omega_k)$ can be estimated as

$$P(s_i|\omega_k) = \frac{cm^i_{k,s}}{N_k} \tag{3}$$

However, this means that a single estimate of $P(s_i|\omega_k)$ as 0 will result in the entire $P(\omega_k|\mathbf{s})$ evaluating to 0. To avoid this, it is instead calculated as

$$\hat{P}(s_i|\omega_k) = \frac{cm^i_{k,s_i} + \frac{1}{c}}{N_k + 1} \tag{4}$$

so that there are zeros when estimating $P(s_i|\omega_k)$. Once training has been complete, a bespoke confusion matrix C_i is obtained for each classifier D_i where the (k,s)th entry is instead calculated as

$$C_i(k,s) = \frac{cm^i_{k,s} + \frac{1}{c}}{N_k + 1} \tag{5}$$

It should however be noted that this approach estimates the likelihood of only the single most probable label of each base classifier. This leads to significant information loss, since each model outputs a meaningful set of probabilities which is being discarded for the most probable label. This is addressed with a novel approach to NB fusion where the likelihood of the full output layer of each base classifier is instead estimated. The predicted labels of each model $\mathbf{s}$ in the previous posterior $P(\mathbf{s}|\omega_k)$ are replaced with the post-processed output layers of each model $\mathbf{O}$ so that the new posterior is $P(\mathbf{O}|\omega_k)$. The likelihood term for a base classifier D_i predicting its output layer O_i is then calculated as the sum of the likelihoods for each neuron in the output layer weighted by their activations

$$\hat{P}(O_i|\omega_k) = \sum_{j=1}^{c} C_i(k, s_j)O_i(j) \tag{6}$$

where s_j is the class label and $O_i(j)$ is the post-processed output layer of classifier D_i for class s_j. The final likelihood term is then calculated in the same manner as $P(\mathbf{s}|\omega_k)$ (2):

$$P(\mathbf{O}|\omega_k) = \prod_{i=1}^{L} P(O_i|\omega_k) \tag{7}$$

While the additional summation in (6) significantly increases the number of computations required, the posterior can be efficiently computed through matrix multiplications. The NB combiner algorithm with soft voting can then be described as

1. For each base classifier D_i, $i = 1, \dots, L$ obtain the output layer of Softmax probabilities O_i.
2. For each class ω_k, $k = 1, \dots, c$:
   (a) Set the prior $P(\omega_k)$ to $\frac{N_k}{N}$, or $\frac{1}{c}$ in the case of equal priors.
   (b) For $i = 1, \dots, L$, compute $P(\omega_k)$ as

$$P(\omega_k) \leftarrow P(\omega_k) \times \sum_{j=1}^{c} C_i(k, s_j)O_i(j)$$

Table 1
Quantitative comparison on Semantic-KITTI test set (sequences 11 to 21). All scores are given as percentages.

Approach	car	bicycle	motorcycle	truck	other-vehicle	person	bicyclist	motorcyclist	road	parking	sidewalk	other-ground	building	fence	vegetation	trunk	terrain	pole	traffic-sign	mIoU	mA
RangeNet++ [9]	91.4	25.7	34.4	25.7	23.0	38.3	38.8	4.8	91.8	65.0	75.2	27.8	87.4	58.6	80.5	55.1	64.6	47.9	55.9	52.2	
KPRNet [12]	95.5	54.1	47.9	23.6	42.6	65.9	65.0	16.5	93.2	**73.9**	80.6	30.2	91.7	68.4	85.7	69.8	71.2	58.7	64.1	63.1	
FIDNet [30]	93.9	54.7	48.9	27.6	23.9	62.3	59.8	23.7	90.6	59.1	75.8	26.7	88.9	60.5	84.5	64.4	69.0	53.3	62.8	59.5	
CENet [29]	91.9	58.6	50.3	40.6	42.3	68.9	65.9	43.5	90.3	60.9	75.1	31.5	91.0	66.2	84.5	69.7	70.0	61.5	**67.6**	64.7	
RangeVIT [33]	95.4	55.8	43.5	29.8	42.1	63.9	58.2	38.1	93.1	70.2	80.0	32.5	92.0	69.0	85.3	70.6	71.2	60.8	64.7	64.0	
RangeFormer [7]	**96.7**	**69.4**	**73.7**	59.9	**66.2**	**78.1**	75.9	**58.1**	92.4	73.0	78.8	**42.4**	92.3	**70.1**	86.6	**73.3**	72.8	**66.4**	66.6	**73.3**	
AMVNet [15]	96.2	59.9	54.2	48.8	45.7	71.0	65.7	11.0	90.1	71.0	75.8	32.4	**92.4**	69.1	85.6	71.7	69.6	62.7	67.2	65.3	
GFNet [16]	94.2	49.7	63.2	**74.9**	32.1	69.3	**83.2**	0.0	**95.7**	53.8	**83.8**	0.2	91.2	62.9	**88.5**	66.1	**76.2**	64.1	48.3	63.0	
SalsaNext [8]	91.9	48.3	38.6	38.9	31.9	60.2	59.0	19.4	91.7	63.7	75.8	29.1	90.2	64.2	81.8	63.6	66.5	54.3	62.1	59.5	
SalsaSU	93.1	54.6	47.6	33.2	36.6	63.4	58.9	8.7	92.1	65.3	77.3	31.2	89.2	61.9	82.4	61.0	67.9	48.4	60.2	59.6	90.1
SalsaBEV	82.3	12.4	11.5	12.9	25.7	19.8	41.6	11.4	88.5	41.3	67.2	6.1	87.8	52.0	74.1	45.6	53.4	33.9	39.0	42.5	85.2
SalsaFusedNB	92.7	51.2	44.6	30.9	38.6	63.5	59.2	8.1	92.3	65.9	76.9	30.5	91.2	65.7	83.0	66.0	66.5	53.5	62.9	60.2	90.5

3. Assign the object label k^*, where

$$k^* = \arg\max_{k=1}^{c} P(\omega_k)$$

Given a segmentation problem with c classes and N points in a data item, the more efficient method of calculating the $c \times N$ matrix of likelihoods $\mathbf{P}$ for a single model D_i using matrix computations is then

$$\mathbf{P} = \mathbf{C}_i \times \mathbf{O}_i \tag{8}$$

where $\mathbf{C}_i$ is the $c \times c$ bespoke confusion matrix of base classifier D_i and $\mathbf{O}_i$ is its $c \times N$ output layer. Assuming equal priors so that priors can be removed as a normalising factor, step 2 can be replaced with:

$$\mathbf{P}(\omega|\mathbf{O}) = \prod_{i=1}^{L} \mathbf{C}_i \times \mathbf{O}_i \tag{9}$$

where $\mathbf{P}(\omega|\mathbf{O})$ is the $c \times N$ matrix of posterior probabilities for each point. Using matrix computations rather than iterative computations allows for more efficient implementations due to the high degree of parallelism in matrix computations. This is particularly advantageous when utilising GPU-based processing.

4. Experiments

In this work both quantitative and qualitative experiments are conducted to evaluate the implemented approach. All training, validation and test splits were kept identical to those originally published by the authors of each dataset.

4.1. Evaluation metrics

The overall accuracy (OA) is the simplest metric used to measure accuracy in semantic segmentation, and is obtained by calculating the percentage of correctly classified points. This can be calculated as

$$OA = \frac{TP + TN}{N} \tag{10}$$

where N is the total number of points in a scene, TP is the total number of true positives, and TN is the total number of true negatives. However, OA can provide misleading results when the class representation is small in a single scene. For this reason, the popular mean intersection over union (mIoU) metric is used. The intersection over union (IoU) of an individual class is calculated as

$$IoU = \frac{T \cap P}{T \cup P} \tag{11}$$

where $T \cap P$ is the intersection of the ground truth positive and predicted positive sets and $T \cup P$ is the union of the ground truth positive and predicted positive sets. The mIoU is then calculated as

$$mIoU = \frac{\sum_{c=1}^{N} IoU_c}{N} \tag{12}$$

where c is the class index, IoU_c is the IoU of class c and N is the number of classes.

4.2. Quantitative results

Tables 1–3 show a comparison of the obtained quantitative results for various projection methods including projection-fusion on the SemanticKITTI, RELLIS-3D, and SemanticPOSS datasets respectively. For the benchmark datasets (SemanticKITTI and SemanticPOSS), a thorough comparison to the current state of the art in projection and projection-fusion approaches is included. The scores for individual labels are the IoU scores, and the highest scoring approach for each label and overall metric is bolded for convenience. The baselines and corresponding projection-fusion approaches developed in this research (those based on SalsaNext) are provided separately. In cases where the SalsaNext variants do not have the highest scoring approach overall, the highest scoring SalsaNext variant is highlighted to provide a comparison between the implemented variants.

It should be noted that the SemanticKITTI test-set results can only be obtained through the SemanticKITTI competition site, which provides a mean accuracy (mA) instead of overall accuracy (OA). Thus, mA is provided for the SemanticKITTI dataset instead of the OA provided for RELLIS-3D and SemanticPOSS. Furthermore, the included papers which make up the current state of the art do not report on this metric. Thus, OA and mA were only included for the variants developed and tested in this research — since it was still found to be of value particularly when assessing the strengths and weaknesses of the BEV model. In the approaches, SalsaSU refers to SalsaNext adapted to use scan unfolding and SalsaBEV to SalsaNext adapted to use Cartesian BEV projection. SalsaFusedMV is the fused scan-unfolding approach (if available) with the BEV approach using traditional majority-voting (MV) based fusion, while SalsaFusedNB fuses the same base approaches using the proposed novel Naïve Bayesian ensembling approach. Only NB-based fusion is evaluated on the SemanticKITTI test-set as a limited number of submissions are available.

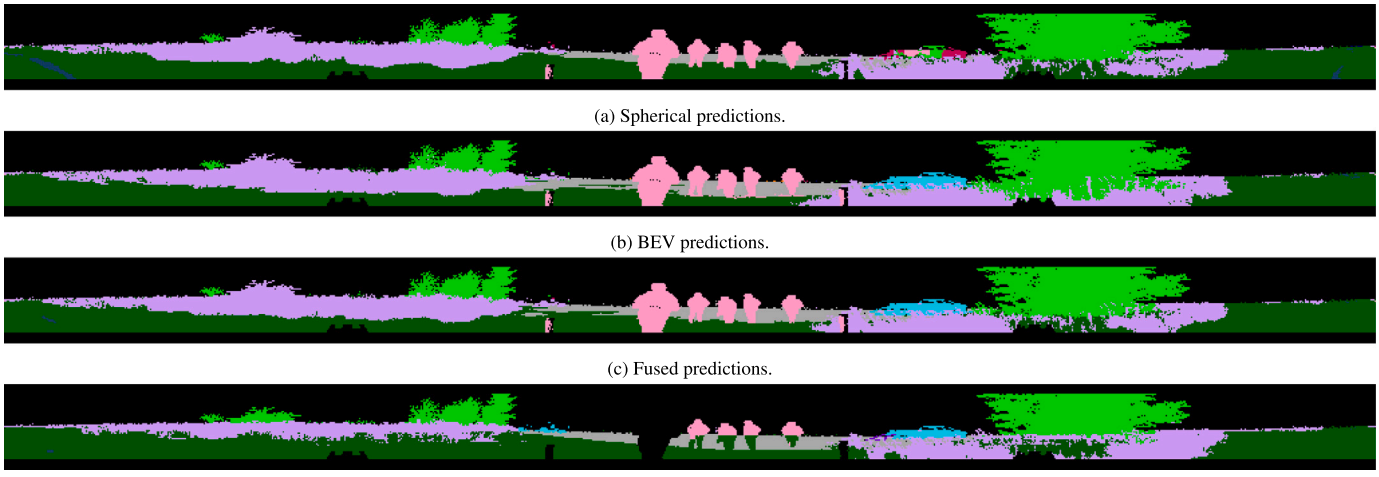
The discussion begins with an examination of the various developed SalsaNext variants in comparison to each other, followed by a broader comparison to the current state of the art in projection and projection-fusion approaches. Surprising observations emerged from our experiments. Scan unfolding, despite visibly reducing occlusions, only marginally outperforms spherical projection in the SemanticKITTI dataset, challenging the notion that occlusion is a significant source of error in spherical projection-based semantic segmentation approaches. Notably, SalsaBEV applied to the RELLIS-3D dataset achieves the highest overall accuracy (OA) score, primarily due to its excellent performance on the high-representation classes like bush, grass, and puddle. This observation suggests that BEV projection may be better suited to natural data, where broader and flatter features, such as grass and

Table 2
Quantitative comparison on SemanticPOSS test set. All scores are given as percentages.

Approach	person	rider	car	bike	truck	vegetation	traffic sign	pole	trash can	building	cone/stone	fence	ground	mIoU	OA
SqueezeSeg [28]	14.2	1.0	13.2	31.0	10.4	28.0	5.1	5.7	2.3	43.6	0.2	15.6	75.0	18.9	
SqueezeSeg + CRF [28]	6.8	0.6	6.7	18.5	4.0	2.5	9.1	1.3	0.4	37.1	0.2	8.4	72.1	12.9	
SqueezeSegV2 [34]	48.0	9.4	48.5	36.1	11.3	50.1	6.7	6.2	14.8	60.4	5.2	22.1	71.3	30.0	
SqueezeSegV2 + CRF [34]	43.9	7.1	47.9	35.3	18.4	40.9	4.8	2.8	7.4	57.5	0.6	12.0	71.3	26.9	
RangeNet53 [9]	55.7	4.5	34.4	28.3	13.7	57.5	3.7	6.6	23.3	64.9	6.1	22.2	72.9	30.3	
RangeNet53 + KNN [9]	57.3	4.6	35.0	28.6	14.1	58.3	3.9	6.9	24.1	66.1	6.6	23.4	73.5	30.9	
MINet [31]	61.8	12.0	63.3	44.5	22.2	68.1	16.3	29.3	28.5	74.6	25.9	31.7	76.4	42.7	
MINet + KNN [31]	62.4	12.1	63.8	44.9	22.3	68.6	16.7	30.1	28.9	75.1	28.6	32.2	76.3	43.2	
FIDNet-Point [30]	71.6	22.7	71.7	50.3	22.9	67.7	21.8	27.5	15.8	72.7	31.3	40.4	79.5	45.8	
FIDNet-Point + KNN [30]	72.2	23.1	72.7	50.3	23.0	68.0	22.2	28.6	16.3	73.1	**34.0**	40.9	79.1	46.4	
CENet [29]	74.9	21.8	77.0	51.7	25.3	72.0	18.0	30.9	46.9	75.9	26.1	47.5	**80.7**	49.9	
CENet + KNN [29]	**75.5**	22.0	**77.6**	51.4	25.3	72.2	18.2	31.5	**48.1**	76.3	27.7	47.7	80.3	50.3	
SalsaSU	54.13	38.16	46.63	52.72	**42.79**	76.91	38.03	37.75	21.45	75.04	18.43	49.54	80.07	48.59	83.87
SalsaBEV	35.61	27.52	46.00	42.79	6.83	72.76	19.74	18.11	0.09	75.07	29.45	35.39	75.81	37.32	79.82
SalsaFusedMV	54.76	38.91	47.87	55.29	31.06	77.65	34.52	30.96	1.53	77.31	10.53	52.42	**80.74**	45.66	84.43
SalsaFusedNB	54.42	**40.51**	57.02	**58.24**	38.92	**80.12**	**40.11**	**39.92**	15.65	**82.76**	27.28	**57.99**	80.49	**51.8**	**86.47**

Table 3
Quantitative comparison on RELLIS-3D test set. All scores are given as percentages.

Approach	grass	tree	pole	water	vehicle	log	person	fence	bush	concrete	barrier	puddle	mud	rubble	mIoU	OA
SalsaNext [8]	65.65	**78.91**	55.93	0.00	23.26	**19.83**	83.85	16.04	74.06	75.27	76.18	22.70	11.17	5.44	43.45	83.06
SalsaBEV	**71.14**	68.52	10.52	0.00	55.34	0.81	73.46	15.60	**78.42**	82.89	23.92	**55.90**	7.77	0.46	38.91	**85.32**
SalsaFusedMV	67.13	78.57	52.55	0.00	25.01	14.53	83.74	17.91	75.43	76.75	73.34	24.13	12.09	6.80	43.43	83.95
SalsaFusedNB	68.40	77.33	**66.36**	0.00	**64.13**	7.14	**83.91**	**18.28**	75.58	**83.04**	**81.13**	37.83	**13.73**	**16.51**	**49.53**	84.69



(a) Spherical predictions.

(b) BEV predictions.

(c) Fused predictions.

(d) Ground-truth labels.

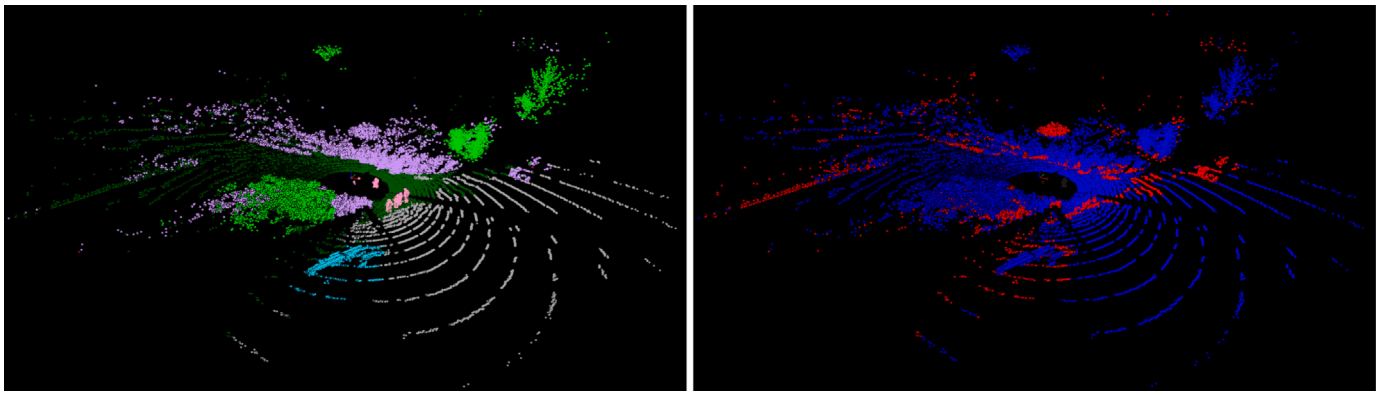
Fig. 3. A qualitative comparison of the fusion method with individual projection methods on natural data.

puddles, benefit from the BEV perspective. However, SalsaBEV's performance on flat urban classes, like road and sidewalk, is less impressive, possibly due to the denser urban scenes requiring higher resolution.

Our novel SalsaFusedNB approach clearly outperforms traditional MV-based fusion (SalsaFusedMV) and other non-fused baseline approaches in both mean IoU (mIoU) and mean accuracy/overall accuracy (mA/OA). It achieves the highest mIoU on RELLIS-3D and the highest mIoU and mA/OA on both SemanticKITTI and SemanticPOSS amongst our SalsaNext variants. Notably, SalsaFusedMV experiences a performance decrease when fusing SalsaSU with SalsaBEV, falling short of SalsaSU's performance alone. This highlights the robustness of the NB projection-fusion approach used in SalsaFusedNB, as it optimally combines individual models, favouring more accurate predictions when one model underperforms. This, coupled with SalsaBEV's strong performance in natural scenes, underscores the suitability of our approach for 3D semantic segmentation of natural data. Furthermore, we observed a substantial mIoU increase compared to the baseline SalsaNext model (6.08%) when evaluating it on an entirely natural dataset.

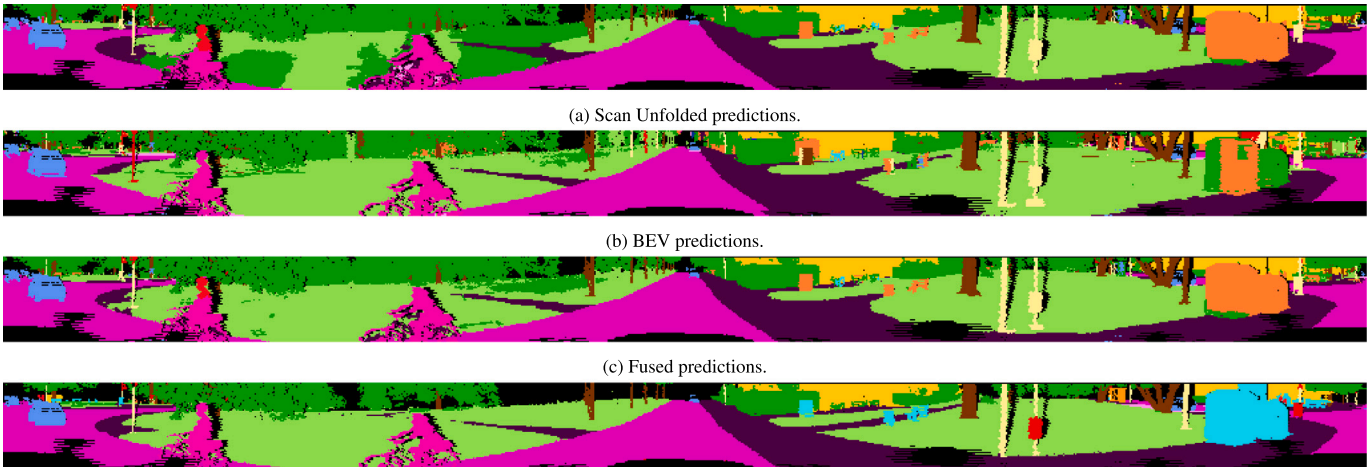
In comparing our results to the current state of the art, particularly on the SemanticKITTI dataset, it is evident that our primary aim was to enhance performance on natural data while maintaining competitiveness on urban datasets. The current state of the art outperforms our approaches on SemanticKITTI, primarily dominated by urban scenes. Nevertheless, it is worth noting that our projection fusion-based approach still shows promise by mitigating degradation on natural data while maintaining an improvement over our baseline methods.

Conversely, on SemanticPOSS, our fused variant exhibits superior performance compared to the existing state of the art in most cases. Due the campus setting of SemanticPOSS, it contains significantly more



(a) Predicted labels. (b) Binary label comparison.

Fig. 4. A qualitative comparison of final 3D predictions on natural data.



(a) Scan Unfolded predictions.

(b) BEV predictions.

(c) Fused predictions.

(d) Ground-truth labels.

Fig. 5. A qualitative comparison of the fusion method with individual projection methods on urban data.

natural scenery than SemanticKITTI. This underscores the adaptability of our approach to environments with a more natural element, demonstrating its potential to excel in scenarios beyond urban settings.

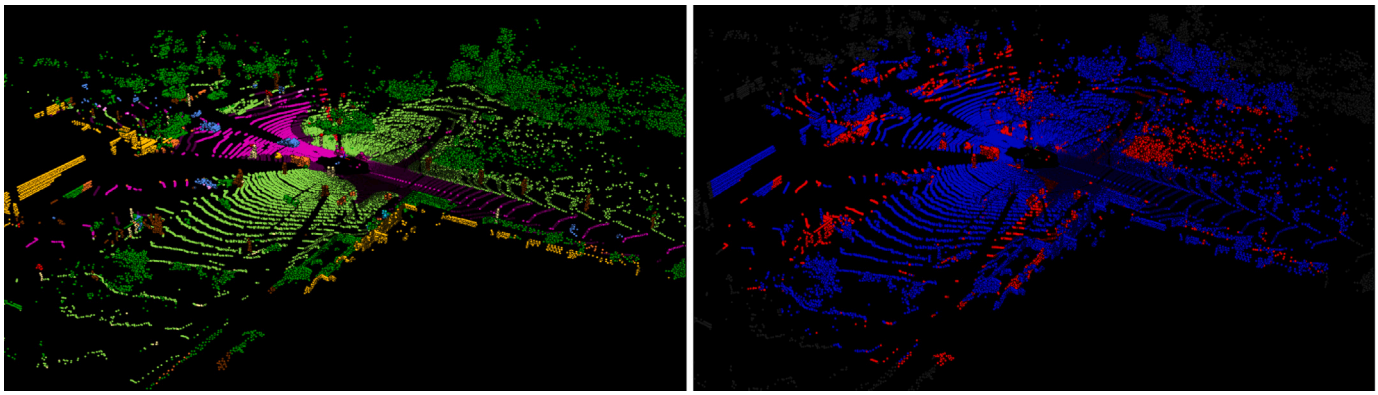
It is important to highlight that while our approach relies on efficient projection-based methods for base classifiers, its inherent nature of fusing multiple base classifiers leads to a trade-off in speed compared to other state-of-the-art projection-based approaches. The usage of A Naïve Bayesian approach does make the fusion approach itself relatively simple and resource-efficient, however it still necessitates independent inferences from multiple base classifiers, resulting in increased computational overhead.

4.3. Qualitative results

The qualitative evaluation is utilised to provide deeper insights to the measured quantitative results for the base classifiers and fused predictions. This evaluation makes use of samples selected to showcase scenarios where each base classifier produces different predictions while remaining consistent with other scans in the test set. Since the approach was developed for both natural and urban data, a natural scene was selected from RELLIS-3D and an urban scene from SemanticKITTI. To provide a direct comparison of the individual base classifier with the final NB-fused predictions, spherical-projected or scan unfolded (in the case of SemanticKITTI) results are shown at each step along with the projected ground truth labels. Finally, a view of the fused 3D results in point cloud format are also shown, as some artefacts may not be visible from a projected view. Since the 3D point cloud is relatively large and the predictions are highly similar to the ground truth, it can be difficult to see differences between the predictions and the ground truth. Thus, a binary label comparison was used instead of the ground truth which shows correctly predicted points in blue and incorrect points in red.

The evaluation begins on natural data from RELLIS-3D, Figs. 3 and 4 show the spherical-projected and 3D results respectively on a sample scan from RELLIS-3D. Label noise is clearly evident in the RELLIS-3D dataset sample, and was observed throughout the dataset, particularly with the ‘person’ labels (pink). In the provided sample, regions that clearly form a person are unlabelled. Additionally, the lower parts of some individuals are mislabelled as ‘grass’ (green). This noise, commonly encountered in natural data, emphasises the need for classifiers that are robust against such inconsistencies. It is worth noting that the tested classifiers offer a more accurate segmentation than the ground truth in this respect, particularly concerning the consistent labelling of people. This inconsistency in the ground truth impacts the reliability of the quantitative results, especially given the sparse representation of the ‘person’ class.

Both primary classifiers seem to produce accurate segmentation maps, though with minor artefacts. The Naïve Bayesian Fusion method, as demonstrated in the spherical-projected results, is beneficial. In this scan, the spherical classifier misses a vehicle on the right, and the BEV classifier slightly misclassifies the dimensions of certain objects. However, fusion manages to rectify most of these discrepancies. Still, some anomalies remain, like a foreground obstruction mistaken



(a) Predicted labels. (b) Binary label comparison.

Fig. 6. A qualitative comparison of final 3D predictions on urban data.

for a person. This particular error, inherent to both base classifiers, highlights a limitation of fusion-based methodologies, which is that they cannot account for errors that occur in all base classifiers. An interesting classification challenge arises with a large tree identified as a bush. Despite a clear height differential for this object, suggesting the presence of a tree, the BEV model still misclassifies the tree, suggesting that the BEV model might not prioritise its depth image adequately. While this is an issue with the BEV model, fusion still appears to optimise the available predictions.

Continuing with an evaluation of the approaches on urban data, Figs. 5 and 6 show the scan-unfolded and 3D results respectively on a sample scan from SemanticKITTI. The predicted label images closely resemble the ground truth images, supporting the conclusion that the approach taken does not degrade segmentation accuracy on urban data. Understandably, the scan-unfolded model predicts the head of a cyclist as a person instead of cyclist. The scan unfolded model also incorrectly labels the grass (labelled light green) around the cyclists as vegetation (labelled dark green). In contrast, the BEV model correctly predicts these regions as grass, supporting the earlier observation that a BEV perspective is better suited to broader and flatter features. An exception to this is the sidewalks, which are flat, this is likely due to sidewalks being narrow and more likely to be obstructed from a top-down view by surrounding objects. The BEV models limitations become more evident when identifying tall, thin objects such as poles, tree trunks, and signs, for which only a small cross-section would be visible from a top-down view. The fusion method often corrects these anomalies, highlighting its strength in merging the most accurate predictions from both models.

Regarding the final 3D point cloud results, the projected labels are largely congruent with the ground truth, but some anomalies are still present, like regions of grass misclassified as trees, a reflection of the BEV model's tendency to neglect the elevation image, as these should be easily identified using elevation. Furthermore, the surroundings of a misclassified vehicle reveal other classification errors, hinting at the model's challenges with unusual contexts, such as cars parked on sidewalks. The fusion method likely could not correct these discrepancies due to both models misidentifying the subject. While these errors might seem trivial, they substantially affect mIoU values. Future enhancements to the fusion approach, possibly integrating neighbourhood context, could potentially rectify such issues.

5. Future work

One inherent gap in any problem addressing 3D semantic segmentation of natural scenes is the lack of data. While the experiments in this work include a significant number of off-road scenes, there is comparatively far more urban data simply due its availability. The class imbalances and lack of structure inherent to natural environments means any real-world application in natural domains would likely require far more natural data than was used in this work. Aside from the development of more natural datasets, this could be addressed through domain adaption techniques to make use of urban data for training the natural segmentation model; or through more advanced data augmentation techniques to expand the effective size of the available datasets.

Despite a lack of natural data, this work clearly demonstrates that more advanced probabilistic ensembling methods to fusion can increase robustness and accuracy. While the novel Naïve Bayesian approach to fusion was effective in this case, there exists a number of alternative approaches to ensembling that could provide further value. Ensemble learning in general is a well-researched field that has seen numerous advancements, and many of these advancements have yet to be applied to semantic segmentation in a meaningful way. Conveniently, ensemble learning is highly synonymous with multi-projection fusion, indicating a clear avenue for future progress.

Another notable gap in the conducted research was the lack of contextual neighbourhood information in fusion. While this research already expanded existing fusion methods to make intelligent usage of the full probabilities from each base classifier, neighbourhood context has been shown to be a crucial feature of any segmentation approach. Thus, future research efforts to incorporate neighbourhood information into the Naïve Bayesian fusion approach could significantly enhance its utility.

Finally, it was noted that our approach, which relies on two separate base classifiers for independent inferences before projection-fusion, inherently results in higher computational costs compared to the generally efficient nature of other projection-based methods. In contrast, projection-fusion models like AMVNet [15] and GFNet [16] adopt more resource-efficient late fusion techniques and frequency-domain operations, respectively, while maintaining competitive performance. Thus, another avenue for future research could include exploring methods to adapt the NB-fusion approach to avoid making independent sets of inferences before fusion could improve computational efficiency.

6. Conclusion

In this research it has been demonstrated that state-of-the-art projection-based approaches can be improved through the aggregation of a number of recent developments to projection-based semantic segmentation into a single model, these include: scan unfolding, soft-kNN post-processing and multi-projection fusion. It was further demonstrated that multi-projection fusion can significantly increase robustness through the introduction of multiple perspectives, thus addressing a number of the challenges inherent to natural data. This was shown to be the case when more advanced probabilistic ensembling methods such as Naïve Bayesian ensembling are used for fusion instead of the traditional voting-based approaches. This approach improves the

mIoU scores of the core model the approach is based on (SalsaNext [8]) for the SemanticKITTI [17] and SemanticPOSS [18] datasets; and achieves state-of-the-art results on RELLIS-3D [2], a dataset specifically developed for semantic segmentation of natural data.

CRediT authorship contribution statement

Kelian J.L. Massa: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Conceptualization. **Hans Grobler:** Writing – review & editing, Supervision, Resources, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] J. Zhang, X. Zhao, Z. Chen, Z. Lu, A review of deep learning-based semantic segmentation for point cloud, IEEE Access 7 (2019) 179118–179133, http://dx.doi.org/10.1109/ACCESS.2019.2958671.

[2] P. Jiang, P. Osteen, M. Wigness, S. Saripalli, RELLIS-3D dataset: Data, benchmarks and analysis, 2020, arXiv preprint arXiv:2003.01174.

[3] D. Maturana, P.-W. Chou, M. Uenoyama, S. Scherer, Real-time semantic mapping for autonomous off-road navigation, in: Proceedings of 11th International Conference on Field and Service Robotics, FSR'17, 2017, pp. 335–350.

[4] E. Uhlemann, Autonomous vehicles have entered the off-road market [connected and automated vehicles], IEEE Veh. Technol. Mag. 16 (2) (2021) 15–19, http://dx.doi.org/10.1109/MVT.2021.3065792.

[5] Y. Akbari, N. Almaadeed, S. Al-ma'adeed, O. Elharrouss, Applications, databases and open computer vision research from drone videos and images: A survey, Artif. Intell. Rev. 54 (5) (2021) 3887–3938, http://dx.doi.org/10.1007/s10462-020-09943-1.

[6] C.R. Qi, L. Yi, H. Su, L.J. Guibas, PointNet++: Deep hierarchical feature learning on point sets in a metric space, in: Advances in Neural Information Processing Systems, Vol. 30, 2017, pp. 1–14.

[7] L. Kong, Y.-C. Liu, R. Chen, Y. Ma, X. Zhu, Y. Li, Y. Hou, Y. Qiao, Z. Liu, Rethinking range view representation for LiDAR segmentation, in: Proceedings of IEEE/CVF International Conference on Computer Vision, ICCV, 2023.

[8] T. Cortinhal, G. Tzelepis, E. Aksoy, SalsaNext: Fast, uncertainty-aware semantic segmentation of LiDAR point clouds, in: Advances in Visual Computing, 2020, pp. 207–222.

[9] A. Milioto, I. Vizzo, J. Behley, C. Stachniss, RangeNet ++: Fast and accurate LiDAR semantic segmentation, in: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2019, pp. 4213–4220, http://dx.doi.org/10.1109/IROS40897.2019.8967762.

[10] L. Triess, D. Peter, C. Rist, J. Zollner, Scan-based semantic segmentation of LiDAR point clouds: An experimental study, in: 2020 IEEE Intelligent Vehicles Symposium, IV, 2020, pp. 1116–1121, http://dx.doi.org/10.1109/IV47402.2020.9304631.

[11] Y.A. Alnaggar, M. Afifi, K. Amer, M. Elhelw, Multi projection fusion for real-time semantic segmentation of 3D LiDAR point clouds, in: 2021 IEEE Winter Conference on Applications of Computer Vision, WACV, 2021, pp. 1799–1808.

[12] D. Kochanov, F.K. Nejadasl, O. Booij, KPRNet: Improving projection-based LiDAR semantic segmentation, 2020, arXiv preprint arXiv:2007.12668.

[13] M. Kellner, B. Stahl, A. Reiterer, Fused projection-based point cloud segmentation, Sensors 22 (3) (2022) 1139.

[14] L. Kuncheva, Combining Pattern Classifiers: Methods and Algorithms, second ed., Wiley, London, 2014.

[15] V.E. Liong, T.N.T. Nguyen, S.A. Widjaja, D. Sharma, Z.J. Chong, AMVNet: Assertion-based multi-view fusion network for LiDAR semantic segmentation, 2020, ArXiv arXiv:2012.04934.

[16] H. Qiu, B. Yu, D. Tao, GFNet: Geometric flow network for 3D point cloud semantic segmentation, Trans. Mach. Learn. Res. (2022) http://dx.doi.org/10.48550/arXiv.2207.02605.

[17] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, J. Gall, SemanticKITTI: A dataset for semantic scene understanding of LiDAR sequences, in: 2019 IEEE/CVF International Conference on Computer Vision, ICCV, 2019, pp. 9296–9306, http://dx.doi.org/10.1109/ICCV.2019.00939.

[18] Y. Pan, B. Gao, J. Mei, S. Geng, C. Li, H. Zhao, SemanticPOSS: A point cloud dataset with large quantity of dynamic instances, 2020, arXiv preprint arXiv:2002.09147.

[19] H. Su, S. Maji, E. Kalogerakis, E. Learned-Miller, Multi-view convolutional neural networks for 3D shape recognition, in: 2015 IEEE International Conference on Computer Vision, ICCV, 2015, pp. 945–953, http://dx.doi.org/10.1109/ICCV.2015.114.

[20] A. Boulch, J. Guerry, B. Saux, N. Audebert, SnapNet: 3D point cloud semantic labeling with 2D deep segmentation networks, Comput. Graph. 71 (2017) 189–198, http://dx.doi.org/10.1016/j.cag.2017.11.010.

[21] J. Guerry, A. Boulch, B. Le Saux, J. Moras, A. Plyer, D. Filliat, SnapNet-R: Consistent 3D multi-view semantic labeling for robotics, in: 2017 IEEE International Conference on Computer Vision Workshops, ICCVW, 2017, pp. 669–678, http://dx.doi.org/10.1109/ICCVW.2017.85.

[22] D. Maturana, S. Scherer, VoxNet: A 3D convolutional neural network for real-time object recognition, in: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2015, pp. 922–928, http://dx.doi.org/10.1109/IROS.2015.7353481.

[23] L. Tchapmi, C. Choy, I. Armeni, J. Gwak, S. Savarese, SEGCloud: Semantic segmentation of 3D point clouds, in: 2017 International Conference on 3D Vision (3DV), 2017, pp. 537–547, http://dx.doi.org/10.1109/3DV.2017.00067.

[24] R. Klokov, V. Lempitsky, Escape from cells: Deep kd-networks for the recognition of 3D point cloud models, in: 2017 IEEE International Conference on Computer Vision, ICCV, 2017, pp. 863–872, http://dx.doi.org/10.1109/ICCV.2017.99.

[25] G. Riegler, A.O. Ulusoy, A. Geiger, OctNet: Learning deep 3D representations at high resolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2017, pp. 6620–6629, http://dx.doi.org/10.1109/CVPR.2017.701.

[26] C. Elich, F. Engelmann, T. Kontogianni, B. Leibe, 3D Bird's-Eye-View instance segmentation, in: Pattern Recognition, 2019, pp. 48–61.

[27] E.E. Aksoy, S. Baci, S. Cavdar, SalsaNet: Fast road and vehicle segmentation in LiDAR point clouds for autonomous driving, in: 2020 IEEE Intelligent Vehicles Symposium, IV, 2020, pp. 926–932, http://dx.doi.org/10.1109/IV47402.2020.9304694.

[28] B. Wu, A. Wan, X. Yue, K. Keutzer, SqueezeSeg: Convolutional neural nets with recurrent CRF for real-time road-object segmentation from 3D LiDAR point cloud, in: 2018 IEEE International Conference on Robotics and Automation, ICRA, 2018, pp. 1887–1893.

[29] H.-X. Cheng, X.-F. Han, G.-Q. Xiao, CENet: Toward concise and efficient LiDAR semantic segmentation for autonomous driving, in: 2022 IEEE International Conference on Multimedia and Expo, ICME, 2022, pp. 01–06, http://dx.doi.org/10.1109/ICME52920.2022.9859693.

[30] Y. Zhao, L. Bai, X. Huang, FIDNet: LiDAR point cloud semantic segmentation with fully interpolation decoding, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, 2021, pp. 4453–4458, http://dx.doi.org/10.1109/IROS51168.2021.9636385.

[31] S. Li, X. Chen, Y. Liu, D. Dai, C. Stachniss, J. Gall, Multi-scale interaction for real-time LiDAR data segmentation on an embedded platform, IEEE Robot. Autom. Lett. 7 (2) (2022) 738–745, http://dx.doi.org/10.1109/LRA.2021.3132059.

[32] P. Biasutti, A. Bugeau, J.-F. Aujol, M. Brédif, RIU-Net: Embarrassingly simple semantic segmentation of 3D LiDAR point cloud, 2019, arXiv preprint arXiv:1905.08748.

[33] A. Ando, S. Gidaris, A. Bursuc, G. Puy, A. Boulch, R. Marlet, RangeViT: Towards vision transformers for 3D semantic segmentation in autonomous driving, in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR, 2023, pp. 5240–5250, http://dx.doi.org/10.1109/CVPR52729.2023.00507.

[34] B. Wu, X. Zhou, S. Zhao, X. Yue, K. Keutzer, SqueezeSegV2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a LiDAR point cloud, in: 2019 International Conference on Robotics and Automation, ICRA, 2019, pp. 4376–4382, http://dx.doi.org/10.1109/ICRA.2019.8793495.