Vote3Deep: Fast Object Detection in 3D Point Clouds Using Efficient Convolutional Neural Networks

Martin Engelcke, Dushyant Rao, Dominic Zeng Wang, Chi Hay Tong, Ingmar Posner

Abstract— This paper proposes a computationally efficient approach to detecting objects natively in 3D point clouds using convolutional neural networks (CNNs). In particular, this is achieved by leveraging a feature-centric voting scheme to implement novel convolutional layers which explicitly exploit the sparsity encountered in the input. To this end we examine the trade-off between accuracy and speed for different architectures and additionally propose to use an $\mathcal{L}_1$ penalty on the filter activations to further encourage sparsity in the intermediate representations. To the best of our knowledge, this is the first work to propose sparse convolutional layers and $\mathcal{L}_1$ regularisation for efficient large-scale processing of 3D data. We demonstrate the efficacy of our approach on the KITTI object detection benchmark and show that *Vote3Deep* models with as few as three layers outperform the previous state of the art in both laser and laser-vision based approaches across the board by margins of up to 40% while remaining highly competitive in terms of processing time.

I. INTRODUCTION

3D point cloud data is ubiquitous in mobile robotics applications such as autonomous driving, where efficient and robust object detection is pivotal for planning and decision making. Recently, computer vision has been undergoing a transformation through the use of convolutional neural networks (CNNs) (e.g. [1], [2], [3], [4]). Methods which process 3D point clouds, however, have not yet experienced a comparable breakthrough. We attribute this lack of progress to the computational burden introduced by the third spatial dimension. The resulting increase in the size of the input and intermediate representations renders a naive transfer of CNNs from 2D vision applications to native 3D perception in point clouds infeasible for large-scale applications. As a result, previous approaches tend to convert the data into a 2D representation first, where nearby features are not necessarily adjacent in the physical 3D space – requiring models to recover these geometric relationships.

In contrast to image data, however, typical point clouds encountered in mobile robotics are spatially sparse, as most regions are unoccupied. This fact was exploited in [5], where the authors propose *Vote3D*, a feature-centric *voting algorithm* leveraging the sparsity inherent in these point clouds. The computational cost is proportional only to the number of *occupied* cells rather than the total number of cells in a 3D grid. [5] proves the equivalence of the voting scheme to a dense convolution operation and demonstrates its effectiveness by discretising point clouds into 3D grids and performing exhaustive 3D sliding window detection with

Authors are from the Mobile Robotics Group, University of Oxford. {firstname}@robots.ox.ac.uk
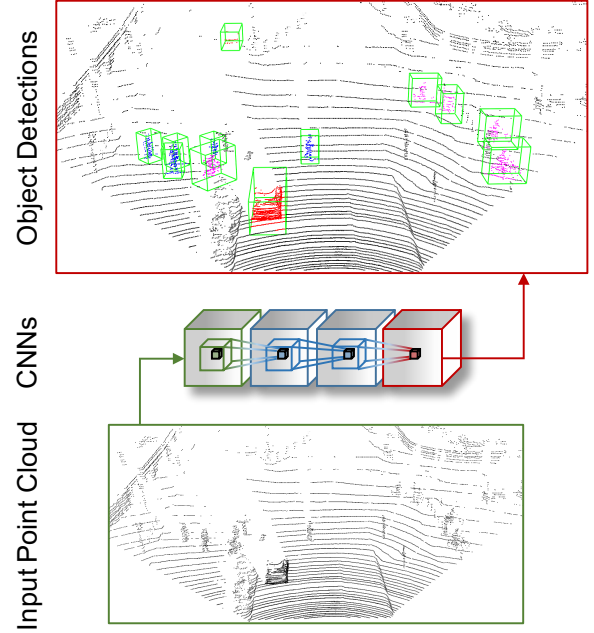
(a) 3D point cloud detection with CNNs

(b) Reference image

Fig. 1. The result of applying *Vote3Deep* to an unseen point cloud from the KITTI dataset, with the corresponding image for reference. The CNNs apply sparse convolutions natively in 3D via *voting*. The model detects cars (red), pedestrians (blue), and cyclists (magenta), even at long range, and assigns bounding boxes (green) sized by class. Best viewed in colour.

a linear Support Vector Machine (SVM). Consequently, [5] achieves the current state of the art in both performance and processing speed for detecting cars, pedestrians and cyclists in point clouds on the object detection task from the popular KITTI Vision Benchmark Suite [6].

Inspired by [5], we propose to exploit feature-centric voting to build efficient CNNs to detect objects in point clouds *natively* in 3D – that is to say without projecting the input into a lower-dimensional space first or constraining the search space of the detector (Fig. 1). In addition, in order to leverage the computational benefits associated with sparse inputs throughout the entire CNN stack, we encourage sparsity

in the inputs to intermediate layers by imposing an $\mathcal{L}_1$ model regulariser. This enables our approach, named *Vote3Deep*, to use sparse convolutions for learning high-capacity, non-linear models while providing constant-time evaluation at test-time, in contrast to non-parametric methods.

To the best of our knowledge, this is the first work to propose sparse convolutional layers based on voting and $\mathcal{L}_1$ regularisation for efficient processing of 3D data at scale. In particular, the contributions of this paper can be summarised as follows:

1) the construction of efficient convolutional layers as basic building blocks for CNN-based point cloud processing by leveraging a voting mechanism exploiting the inherent sparsity in the input data;
2) the use of rectified linear units and an $\mathcal{L}_1$ sparsity penalty to specifically encourage *data* sparsity in intermediate representations in order to exploit sparse convolution layers throughout the entire CNN stack.

We demonstrate that *Vote3Deep* models with as few as three layers achieve state-of-the-art performance amongst purely laser-based approaches across all classes considered on the popular KITTI object detection benchmark. *Vote3Deep* models exceed the previous state of the art in 3D point cloud based object detection in average precision by a margin of up to 40% while remaining competitive in terms of processing time.

II. RELATED WORK

A number of works have attempted to apply CNNs in the context of 3D point cloud data. A CNN-based approach in [7] obtains comparable performance to [5] on KITTI for car detection by projecting the point cloud into a 2D depth map, with an additional channel for the height of a point from the ground. The model predicts detection scores and regresses to bounding boxes. While the CNN is a highly expressive model, the projection to a specific viewpoint discards information, which is particularly detrimental in crowded scenes. It also requires the network filters to learn local dependencies with regards to depth by brute force, information that is readily available in a 3D representation and which can be efficiently extracted with sparse convolutions.

Dense 3D occupancy grids computed from point clouds are processed with CNNs in [8] and [9]. With a minimum cell size of 0.1m, [8] reports a speed of 6ms on a GPU for their slowest model to classify a single crop with a grid-size of $32 \times 32 \times 32$ cells. In addition, it takes up to 0.5s to convert 200,000 points into an occupancy grid. When restricting point clouds from the KITTI dataset to the field of view of the camera, a total of 20,000 points are typically spread over 2×10^6 grid cells with a resolution of 0.2m as used in this work. Naively evaluating the classifier of [8] at all possible locations would therefore approximately take $6 \times 10^{-3} \times 2 \times 10^6/8 = 1,500$s, accounting for the reduction in resolution and ignoring speed ups from further parallelism on a GPU. Similarly, a processing time of up to 5ms per m^3 for detecting landing zones is reported in [9].

A alternative approach that takes advantage of sparse representations can be found in [10] and [11] who apply sparse convolutions to relatively small 2D and 3D crops respectively. While the convolutional kernels are only applied at sparse feature locations, the presented algorithm still has to consider neighbouring values which are either zeros or constant biases, leading to unnecessary operations and memory consumption. Another method for performing sparse convolutions is introduced in [12] who use "permutohedral lattices" and bilateral filters with trainable parameters.

CNNs have also been applied to dense 3D data in biomedical image analysis (e.g. [13], [14], [15]). A 3D equivalent of residual networks [4] is utilised in [13] for brain image segmentation. A cascaded model with two stages is proposed in [14] for detecting cerebral microbleeds. A combination of three CNNs is suggested in [15]. Each CNN processes a different 2D image plane and the three streams are joined in the last layer. These systems run on relatively small inputs and in some cases take more than a minute for processing a single frame with GPU acceleration.

III. METHODS

Vote3Deep performs efficient, large-scale, multi-instance object detection with CNNs natively in 3D point clouds. The first step is to convert the point cloud to a discrete 3D representation. In our work, a point cloud is discretised into a 3D grid as in [5]: For each cell that contains a non-zero number of points, a feature vector is extracted based on the statistics of the points in the cell. The feature vector holds a binary occupancy value, the mean and variance of the reflectance values and three shape factors. Cells in empty space are not stored which leads to a sparse representation.

We employ the voting scheme from [5] to perform a sparse convolution across this native 3D representation, followed by a ReLU non-linearity, which returns a new sparse 3D representation. This process can be repeated and stacked as in a traditional CNN. Finally, the output layer predicts confidence scores that indicate the presence of an object.

As in [5], to handle objects at different orientations, the CNN is run over a point cloud at N different angular orientations in N parallel threads. This allows objects with arbitrary pose to be handled at a minimal increase in computation time. Duplicate detections are pruned with non-maximum suppression (NMS) in 3D space. NMS in 3D is better able to handle objects that are behind each other as the 3D bounding boxes overlap less than their projections into 2D.

A. Sparse Convolutions via Voting

When running a dense 3D convolution across a discretised point cloud, most of the computation time is wasted as the majority of operations are multiplications by zero. The additional third spatial dimension makes this process even more computationally expensive compared to 2D convolutions, which form the basis of image-based CNNs. Using the insight that meaningful computation only takes place where the 3D features are non-zero, [5] introduce a feature-centric voting scheme. The basis of this algorithm is the idea of
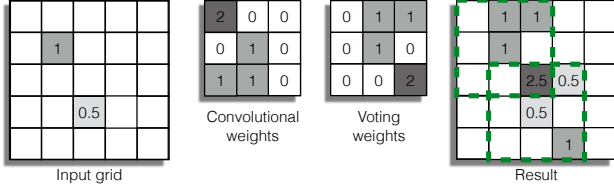
Fig. 2. An illustration of the voting procedure on a 2D example sparse grid. The voting weights are obtained by flipping the convolutional weights about each dimension. Whereas a standard convolution applies the filter at every location in the input, the equivalent voting procedure only needs to be applied at each non-zero location and obtains an identical result. While this illustration is in 2D for just one feature map, the actual voting procedure is on a 3D grid with several feature maps. For a full mathematical justification, the reader is referred to [5]. Best viewed in colour.

letting each non-zero input feature vector cast a set of votes, weighted by the filter weights, to its surrounding cells in the output layer, as defined by the receptive field of the filter. The weights used for voting are obtained by flipping the convolutional filter kernel along each spatial dimension. The final convolution result is then simply obtained by accumulating the votes falling into each cell of the output layer (Fig. 2).

This procedure can be formally stated as follows. Without loss of generality, assume we have one 3D convolutional filter in network layer c with odd-valued side lengths, operating on a single input feature, with the filter weights denoted by $\mathbf{w}^c \in \mathbb{R}^{(2I+1)\times(2J+1)\times(2K+1)}$. Then, for an input grid $\mathbf{h}^{c-1} \in \mathbb{R}^{L\times M\times N}$, the convolution result at location (l, m, n) is given by:

$$z_{l,m,n}^c = \sum_{i=-I}^{I} \sum_{j=-J}^{J} \sum_{k=-K}^{K} w_{i,j,k}^c \, h_{l+i,m+j,n+k}^{c-1} + b^c \quad (1)$$

where b^c is a bias value applied to all cells in the grid. This operation needs to be applied to all $L \times M \times N$ locations in the input grid for a regular dense convolution. In contrast to this, given the set of cell indices for all of the non-zero cells $\Phi = \{(l, m, n) \; \forall \; h_{l,m,n}^{c-1} \neq 0\}$, the convolution can be recast as a feature-centric voting operation, with each input cell casting votes to increment the values in neighbouring cell locations according to:

$$z_{l+i,m+j,n+k}^c = z_{l+i,m+j,n+k}^c + w_{-i,-j,-k}^c \, h_{l,m,n}^{c-1} \quad (2)$$

which is repeated for all tuples $(l, m, n) \in \Phi$ and where $\{i, j, k \in \mathbb{Z} \mid i \in [-I, I], j \in [-J, J], k \in [-K, K]\}$.

The voting output is passed through a ReLU non-linearity which discards non-positive features as described in the next subsection. The biases are constrained to be non-positive as a single positive bias would return an output grid in which almost every cell is occupied with a non-zero feature vector, hence completely eliminating sparsity. The bias b^c therefore only needs to be added to each non-empty output cell.

With this sparse voting scheme, the filter only needs to be applied to the occupied cells in the input grid, rather than convolved over the entire grid. The full algorithm is described in more detail in [5], including formal proof

that feature-centric voting is equivalent to an exhaustive convolution.

B. Maintaining Sparsity with ReLUs

When stacking multiple sparse 3D convolution layers to build a deep neural network, it is necessary to maintain sparsity in the intermediate representations. With additional convolutional layers, however, the receptive field of the network grows with each layer. This means that an increasing number of cells receive votes which progressively decreases sparsity higher up in the representation hierarchy. A simple way to counteract this behaviour is to follow a sparse convolution layer by a rectified linear unit (ReLU) as advocated in [16], which can be written as:

$$h^c = \max(0, z^c) \quad (3)$$

with z^c being the input to the ReLU non-linearity in layer c as typically computed by a sparse convolution, and h^c being the output, denoting the *hidden* activations in the sparse intermediate representations.

In this case, only features that have a value greater than zero will be allowed to cast votes in the next sparse convolution layer. In addition to enabling a network to learn non-linear function approximations, ReLUs effectively perform a thresholding operation by discarding negative feature values which helps to maintain sparsity in the intermediate representations. Lastly, another advantage of ReLUs compared to other non-linearities is that they are fast to compute.

IV. TRAINING

Based on the premise that bounding boxes in 3D space are similar in size for object instances of the same class, we simply assume a fixed-size bounding box for each class. A set of fixed 3D bounding box dimensions is selected for each class, based on the 95th percentile ground truth bounding box size over the training set. The receptive field of a network should be at least as large as this bounding box, but not excessively large so as to waste computation. We therefore train three separate networks which can be run in parallel at test time, each with a different receptive field, and specialised for detecting a certain class. It is possible, however, to share computation and features in the lower layers followed by a class-specific output layer; a task left for future work.

Fixed-size bounding boxes imply that networks can be straightforwardly trained on 3D crops of positive and negative examples whose dimensions equal the receptive field size of a network. While processing point clouds with several angular bins allows us to handle objects with different poses to some degree, we augment the training data by randomly rotating the original front-facing positive training examples by an angle that is smaller than the resolution of the angular bins. Similarly, we also augment the training data by randomly translating positive training examples by a distance smaller than the 3D grid cell size to account for discretisation effects. Negative training examples are obtained by performing hard negative mining periodically, after a fixed number of training epochs.

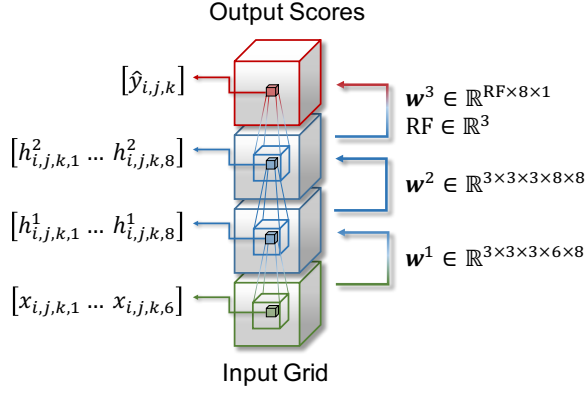| Model	Layer 1	Layer 2	Layer 3
A	RF	-	-
B	3x3x3	RF	-
C	5x5x5	RF	-
D	3x3x3	3x3x3	RF
E	5x5x5	3x3x3	RF

Fig. 3. Illustration of the "Model D" architecture from Table I. The input $\mathbf{x}$ (green) and the intermediate representations $\mathbf{h^c}$ (blue) for layer c are sparse 3D grids, where each occupied spatial location holds a feature vector (solid cubes). The sparse convolutions with the filter weights $\mathbf{w^c}$ are performed natively in 3D to compute the predictions (red). Best viewed in colour.

The class-specific networks are binary classifiers so we choose a linear hinge loss for training due to its maximum margin property. The hinge loss, $\mathcal{L}_2$ weight decay and in some cases an $\mathcal{L}_1$ sparsity penalty are used to train the networks with stochastic gradient descent. Both the $\mathcal{L}_2$ weight decay as well as the $\mathcal{L}_1$ sparsity penalty serve as regularisers. The sparsity penalty in addition encourages the network to learn sparse intermediate representations to reduce the computation cost.

A. Linear Hinge Loss

Given an output detection score $\hat{y} \in \mathbb{R}$ and a class label $y \in \{-1, 1\}$ distinguishing between positive and negative samples, the hinge loss is formulated as:

$$L(\theta) = \max(0, 1 - \hat{y} \cdot y) \qquad (4)$$

where θ denotes the parameters of the network.

The loss in Eq. 4 is zero for positive samples that score over 1 and negative samples that score below -1. As such, the hinge loss drives sample scores away from the margin given by the interval $[-1, 1]$. As with standard CNNs, the $\mathcal{L}_1$ hinge loss can be backpropagated through the network to compute the gradients with respect to the weights, and subgradients can be computed at the discontinuities.

B. $\mathcal{L}_1$ Sparsity Penalty

The ability to perform fast voting is predicated on the assumption of sparsity in the input to each layer. While the input point cloud is sparse, the regions of non-zero cells are dilated in each successive layer, approximately by the receptive field size of the corresponding convolutional filters. It is therefore prudent to encourage sparsity in each layer, such that the model only utilises features if they are absolutely necessary for the detection task.

The $\mathcal{L}_1$ loss has been shown to result in sparse representations in which several values are exactly zero [17], which is precisely the requirement for this model. Whereas the sparsity of the output layer can be tuned with a detection threshold, we encourage sparsity in the intermediate layers by incorporating a penalty term using the $\mathcal{L}_1$ norm of each feature activation.

V. EXPERIMENTS

A. Dataset

We use the well-known KITTI Vision Benchmark Suite [6] for training and evaluating our detection models. The dataset consists of synchronised stereo camera and lidar frames recorded from a moving vehicle with annotations for eight different object classes, showing a wide variety of road scenes with different appearances. We only use the 3D point cloud data to train and test the models. There are 7,518 frames in the KITTI test set whose labels are not publicly available. The labelled training data consist of 7,481 frames which we split into two sets for training and validation (80% and 20% respectively). The object detection benchmark considers three classes for evaluation: cars, pedestrians and cyclists with 28,742; 4,487; and 1,627 training labels, respectively.

B. Architectures

A range of fully convolutional architectures with up to three layers and different filter configurations are compared (Table I). The "Model D" architecture is illustrated as an example in Fig. 3. To exploit context around an object, the architectures are designed so that the *total* receptive field is slightly larger than the class-specific bounding boxes. Small $3 \times 3 \times 3$ and $5 \times 5 \times 5$ kernels are used in the lower layers, followed by a ReLU non-linearity. The network output is computed by a linear layer which is implemented as a convolutional filter whose kernel size gives the desired receptive field for the network for a given class.

C. Training

The networks are trained on 3D crops of positive and negative examples. The number of positives and negatives is initially balanced with negatives being extracted randomly from the training data at locations that do not overlap with any of the positives. Hard negative mining is performed every ten epochs by running the current model across the full point clouds in the training set. In each round of hard negative mining, up to 10,000 of the highest scoring false positives are added to the training set.

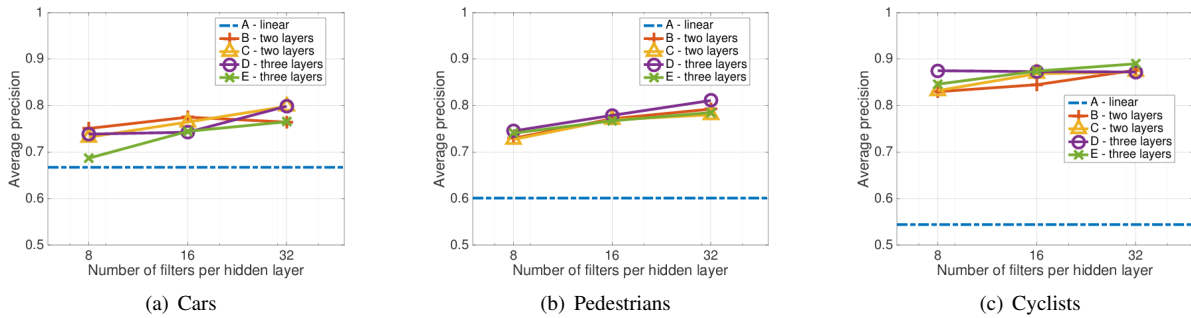|		
(a) Cars	(b) Pedestrians	(c) Cyclists

Fig. 4. Model comparison for the architecture in Table I, showing the average precision for the moderate difficulty level. The non-linear models with two or three layers consistently outperform the linear baseline model our internal *validation set* by a considerable margin for all three classes. The performance continues to improve as the number of filters in the hidden layers is increased, but these gains are incremental compared to the large margin between the linear baseline and the smallest multi-layer models. Best viewed in colour.
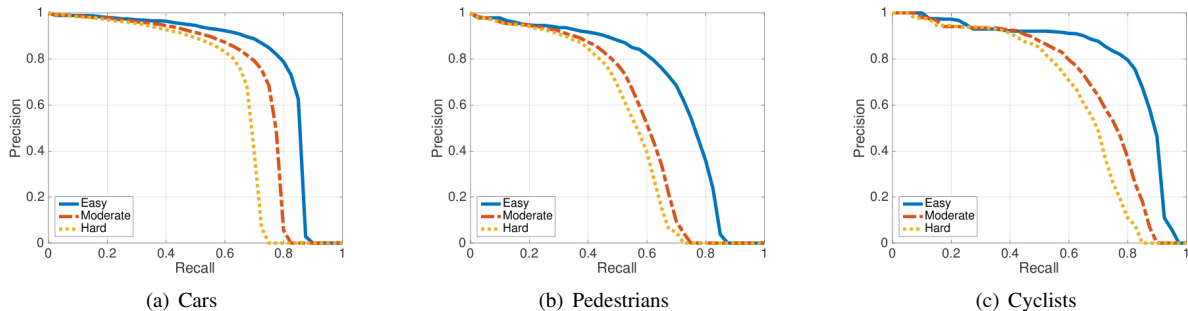

(a) Cars	(b) Pedestrians	(c) Cyclists

Fig. 5. Precision-Recall curves for the evaluation results on the KITTI *test set*. "Model B" for cars and "Model D" for pedestrians and cyclists, all with eight filters in the hidden layers and trained without sparsity penalty, are used for the submission to the official test server. Best viewed in colour.

The weights are initialised as proposed in [18] and trained with stochastic gradient descent with momentum of 0.9 and $\mathcal{L}_2$ weight decay of 10^{-4} for 100 epochs with a batch size of 16. The model from the epoch with the best average precision on the validation set is selected for the model comparison and the KITTI test submission in Sections V-E and V-F, respectively. Other hyperparameters are tuned on the validation set.

We implemented a custom C++ library for training and testing. For the largest models, training takes about three days on a cluster CPU node with 16 cores where each example in a batch is processed in a separate thread.

D. Evaluation

The official benchmark evaluation on the KITTI test server is performed in 2D image space. We therefore project our 3D detections into the 2D image plane using the provided calibration files and discard any detections that fall outside of the image. The KITTI benchmark differentiates between easy, moderate and hard test categories depending on the bounding box size, object truncation and occlusion. An average precision score is independently reported for each difficulty level and class. The easy test examples are a subset of the moderate examples, which are in turn a subset of the hard examples. The official KITTI rankings are based on the performance on the moderate cases. Results are obtained for a variety of models on the validation set, and selected models for each class are submitted to the KITTI test server.

E. Model Comparison

Fast detection speeds are particularly important in the context of urban transport. As larger, more expressive models come at a higher computational cost and consequently run at slower speeds, this section investigates the trade-off between detection performance and model capacity on the *validation set*. Five architectures are benchmarked against each other with up to three layers and different numbers of filters in the hidden layers (Fig. 4). These models are trained without the $\mathcal{L}_1$ penalty which is discussed later in Section V-G.

The non-linear, multi-layer networks clearly outperform the linear baseline, which is comparable to [5]. First and foremost, this demonstrates that increasing the complexity and expressiveness of the models is extremely helpful for detecting objects in point clouds.

Even though performance improves with the number of convolutional filters in the hidden layers, the resulting gains are comparatively moderate. Similarly, increasing the receptive field of the filter kernels, while keeping the total receptive field of the networks the same, does not improve the performance. It is possible that these larger models are not sufficiently regularised. Another potential explanation is that the easy interpretability of 3D data enables even these relatively small models to capture most of the variation in the input representation which is useful for solving the task.

TABLE II

AVERAGE PRECISION IN % ON THE KITTI TEST SET FOR METHODS ONLY USING POINT CLOUDS

	Processor	Speed	Cars			Pedestrians			Cyclists		
			Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
Vote3Deep	4-core 2.5GHz CPU	1.5s	**76.79**	**68.24**	**63.23**	**68.39**	**55.37**	**52.59**	**79.92**	**67.88**	**62.98**
Vote3D [5]	4-core 2.8GHz CPU	**0.5s**	56.80	47.99	42.56	44.48	35.74	33.72	41.43	31.24	28.60
VeloFCN [7]	2.5GHz GPU	1.0s	60.34	47.51	42.74	-	-	-	-	-	-
CSoR	4-core 3.5GHz CPU	3.5s	34.79	26.13	22.69	-	-	-	-	-	-
mBoW [19]	1-core 2.5GHz CPU	10s	36.02	23.76	18.44	44.28	31.37	30.62	28.00	21.62	20.93

TABLE III

AVERAGE PRECISION IN % ON THE KITTI TEST SET FOR METHODS UTILISING BOTH POINT CLOUDS AND IMAGES AS INDICATED BY *

	Processor	Speed	Cars			Pedestrians			Cyclists		
			Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
Vote3Deep	4-core 2.5GHz CPU	**1.5s**	76.79	68.24	**63.23**	68.39	55.37	**52.59**	**79.92**	**67.88**	**62.98**
MV-RGBD-RF* [20]	4-core 2.5GHz CPU	4s	76.40	**69.92**	57.47	**73.30**	**56.59**	49.63	52.97	42.61	37.42
Fusion-DPM* [21]	1-core 3.5GHz CPU	30s	-	-	-	59.51	46.67	42.05	-	-	-

F. Test Results

From Table I, we select the "Model B" for cars, and the "Model D" for pedestrians and cyclists, with 8 filters per hidden layer and trained without sparsity penalty for evaluation on the KITTI *test* set. These models are selected for their high performance at a relatively small number of parameters. The PR curves for *Vote3Deep* on the KITTI test set are shown in Figure 5.

The performance of *Vote3Deep* is compared against the other leading approaches for object detection in point clouds (at the time of writing) in Table V-E. *Vote3Deep* establishes new state-of-the-art performance in this category for all three classes and all three difficulty levels. The performance boost is particularly significant for cyclists with a margin of almost 40% on the easy test case, in some cases more than doubling the average precision. *Vote3Deep* currently runs on CPU and is about three times slower than [5] and 1.5 times slower than [7] with the latter relying on GPU acceleration. We expect that a GPU implementation of *Vote3Deep* will further improve the detection speed. Compared to the very deep networks commonly used in vision (e.g. [2], [3], [4]), these relatively shallow networks trained without any of the recently developed tricks are expressive enough to achieve significant performance gains.

We also compare *Vote3Deep* against methods that utilise both point cloud and image data in Table III. Despite only using point cloud data, *Vote3Deep* still performs better than these ([20], [21]) in the majority of test cases and only slightly worse in the remaining ones at a considerably faster detection speed. For all three classes, *Vote3Deep* achieves the highest average precision on the hard test cases, which consider the largest number of positive ground truth objects.

Interestingly, cyclist detection benefits the most from the expressiveness of CNNs even though this class has the least number of training examples. We conjecture that cyclists are have a more distinctive shape in 3D compared to pedestrians and cars, which can be more easily confused with poles or vertical planes, respectively, and that the *Vote3Deep* models are able exploit this complexity particularly well, despite the small amount of training data.

G. Timing and Sparsity

The three models from the previous subsection are also trained with different values for the $\mathcal{L}_1$ sparsity penalty to examine the effect of the penalty on detection speed and performance on the moderate difficulty cases (Table IV). Larger penalties than those presented in the table tend to push all the activations to zero. We found that selecting the models from the epoch with the largest average precision on the validation set tends to favour models with a comparatively low sparsity in the intermediate representations. Thus, the networks are all trained for 100 epochs and the models after the final epoch are used for evaluation in order to enable a fair comparison. The mean and standard deviation of the detection time per frame are measured on 100 frames from the KITTI validation set.

Unsurprisingly, pedestrians have the fastest detection time as the receptive field of the networks is smaller compared to the other two classes. The two-layer "Model B" is used for cars during testing, as opposed to the three-layer "Model D" for the other two classes, which explains why the car detector runs faster than cyclist detector even though cars require a larger receptive field than cyclists. The sparsity penalty improves the detection speed by about 12% and 6% for cars and cyclists, respectively, at a negligible difference in average precision. For pedestrians, the two models trained without sparsity penalty run slower and perform better than the baseline. Notably, the benefit of the sparsity penalty increases with the receptive field size of the network. We conjecture that pedestrians are too small to learn representations with a significantly higher sparsity through the sparsity penalty, and that the drop in performance for the baseline model is a consequence of the model selection process.

	Cars		Pedestrians		Cyclists	
Penalty	Run-time	AP	Run-time	AP	Run-time	AP
0	884±256	0.76	**670±169**	0.7	1543±510	0.86
10^{-4}	**786±208**	0.76	707±171	0.74	1505±492	0.83
10^{-3}	809±217	0.76	681±170	0.73	**1451±459**	0.85

VI. CONCLUSIONS

This work performs object detection in point clouds at fast speeds with CNNs constructed from sparse convolution layers based on the voting scheme introduced in [5]. With the ability to learn hierarchical representations and non-linear decision boundaries, a new state of the art is established on the KITTI benchmark for detecting objects in point clouds. *Vote3Deep* also outperforms other methods that utilise information from both point clouds and images in most test cases. Possible future directions include a more low-level input representation as well as a GPU implementation of the voting algorithm.

ACKNOWLEDGMENT

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances In Neural Information Processing Systems*, pp. 1–9, 2012.

[2] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *ICLR*, pp. 1–14, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June, 2015, pp. 1–9.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *arXiv preprint arXiv:1512.03385*, vol. 7, no. 3, pp. 171–180, 2015. [Online]. Available: http://arxiv.org/pdf/1512.03385v1.pdf

[5] D. Z. Wang and I. Posner, "Voting for Voting in Online Point Cloud Object Detection," *Robotics Science and Systems*, 2015.

[6] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2012, pp. 3354–3361.

[7] B. Li, T. Zhang, and T. Xia, "Vehicle Detection from 3D Lidar Using Fully Convolutional Network," *arXiv preprint arXiv:1608.07916*, 2016. [Online]. Available: https://arxiv.org/abs/1608.07916

[8] D. Maturana and S. Scherer, "VoxNet: A 3D Convolutional Neural Network for Real-Time Object Recognition," *IROS*, pp. 922–928, 2015.

[9] ——, "3D Convolutional Neural Networks for Landing Zone Detection from LiDAR," *International Conference on Robotics and Automation*, no. Figure 1, pp. 3471–3478, 2015.

[10] B. Graham, "Spatially-sparse convolutional neural networks," *arXiv Preprint arXiv:1409.6070*, pp. 1–13, 2014. [Online]. Available: http://arxiv.org/abs/1409.6070

[11] ——, "Sparse 3D convolutional neural networks," *arXiv preprint arXiv:1505.02890*, pp. 1–10, 2015. [Online]. Available: http://arxiv.org/abs/1505.02890

[12] V. Jampani, M. Kiefel, and P. V. Gehler, "Learning Sparse High Dimensional Filters: Image Filtering, Dense CRFs and Bilateral Neural Networks," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[13] H. Chen, Q. Dou, L. Yu, and P.-A. Heng, "VoxResNet: Deep Voxelwise Residual Networks for Volumetric Brain Segmentation," *arXiv preprint arXiv:1608.05895*, 2016. [Online]. Available: http://arxiv.org/abs/1608.05895

[14] Q. Dou, H. Chen, L. Yu, L. Zhao, J. Qin, D. Wang, V. C. Mok, L. Shi, and P. A. Heng, "Automatic Detection of Cerebral Microbleeds From MR Images via 3D Convolutional Neural Networks," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1182–1195, 2016. [Online]. Available: http://ieeexplore.ieee.org

[15] A. Prasoon, K. Petersen, C. Igel, F. Lauze, E. Dam, and M. Nielsen, "Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8150 LNCS, no. PART 2, 2013, pp. 246–253.

[16] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Neural Networks," *AISTATS*, vol. 15, pp. 315–323, 2011.

[17] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. MIT press, 2012, ch. 13, pp. 423–480.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," *arXiv preprint arXiv:1502.01852*, pp. 1–11, 2015. [Online]. Available: https://arxiv.org/abs/1502.01852

[19] J. Behley, V. Steinhage, and A. B. Cremers, "Laser-based segment classification using a mixture of bag-of-words," in *IEEE International Conference on Intelligent Robots and Systems*, 2013, pp. 4195–4200.

[20] A. Gonzalez, G. Villalonga, J. Xu, D. Vazquez, J. Amores, and A. M. Lopez, "Multiview random forest of local experts combining RGB and LIDAR data for pedestrian detection," in *IEEE Intelligent Vehicles Symposium, Proceedings*, vol. 2015-Augus, 2015, pp. 356–361.

[21] C. Premebida, J. Carreira, J. Batista, and U. Nunes, "Pedestrian detection combining RGB and dense LIDAR data," in *IEEE International Conference on Intelligent Robots and Systems*, 2014, pp. 4112–4117.