
DSCnet: Replicating Lidar Point Clouds with Deep Sensor Cloning

Paden Tomasello*, Sammy Sidhu, Anting Shen, Matthew W. Moskewicz,
Nobie Redmon, Gayatri Joshi, Romi Phadte, Paras Jain, Forrest Iandola
DeepScale, Inc.
Mountain View, CA, USA
*paden@deepscale.ai

Abstract

Convolutional neural networks (CNNs) have become increasingly popular for solving a variety of computer vision tasks, ranging from image classification to image segmentation. Recently, autonomous vehicles have created a demand for depth information, which is often obtained using hardware sensors such as Light detection and ranging (LIDAR). Although it can provide precise distance measurements, most LIDARs are still far too expensive to sell in mass-produced consumer vehicles, which has motivated methods to generate depth information from commodity automotive sensors like cameras.

In this paper, we propose an approach called Deep Sensor Cloning (DSC). The idea is to use Convolutional Neural Networks in conjunction with inexpensive sensors to replicate the 3D point-clouds that are created by expensive LIDARs. To accomplish this, we develop a new dataset (DSDepth) and a new family of CNN architectures (DSCnets). While previous tasks such as KITTI depth prediction use an interpolated RGB-D images as ground-truth for training, we instead use DSCnets to directly predict LIDAR point-clouds. When we compare the output of our models to a \$75,000 LIDAR, we find that our most accurate DSCnet achieves a relative error of 5.77% using a single camera and 4.69% using stereo cameras.

1 Introduction and Related Work

1.1 Introduction

Convolutional neural networks have become quintessential for solving a variety of computer vision tasks such as classification [8], semantic segmentation [9], and depth prediction [2]. Most previous attempts of using CNNs for depth prediction attempt to predict depth for each pixel of the image, and the ground truth comes from either a virtual dataset as ShapeNet [1], or a dataset that interpolates missing values from a LIDAR as in the KITTI dataset [4]. In this paper, we propose a method called Deep Sensor Cloning, which directly regresses on the output of LIDAR using a variety of sensor inputs.

1.2 Motivation

The promise of autonomous vehicles has inspired numerous advancements in both perception systems and automotive sensors. Perception systems create three-dimensional representation of their environment, so autonomous vehicles can safely and effectively navigate and control themselves. Any effective perception system must provide accurate depth information, which has traditionally been obtained using a sensor like radar or LIDAR. Radars, which are already in mass-produced vehicles, only provide coarse depth measurements for objects like cars or motorcycles. While this

may be sufficient for adaptive cruise control or emergency braking, full autonomy will require more fine grained depth information. LIDAR provides perception systems with precise distance measurements of its environment in the form of a three-dimensional point cloud. Given these advantages, LIDAR (particularly Velodyne HDL-32 and HDL-64) is now used in prototype autonomous vehicles developed by Uber, Zoox, and Cruise Automation.

While accomplishing the task of depth prediction well these Velodyne LIDARs are very expensive, which means that LIDAR is now a barrier of entry for autonomous vehicles to the mass market, because of their cost, power consumption, and complexity. For example, Velodyne’s HDL-64 LIDAR, a model similar to the one used in DARPA Grand Challenge, consumes 60W of power and is estimated to cost \$75,000, limiting their use to vehicles in robo-taxi services. Velodyne’s cheapest model the VLP-16 contains $\frac{1}{4}$ of the number lasers as the HDL-64, and is still estimated to cost \$8,000, a price still too expensive for all but the most expensive consumer vehicles. When trading notes with others who have done field work in autonomous driving, we have heard criticism that the Velodyne VLP-16 and HDL-64 LIDARs rely on precisely calibrated internal components, which can make them prone to be less accurate or break from standard wear and tear.

More and more companies are developing LIDAR products, and a comparison of some currently-available LIDAR devices is shown in Table 1¹. As can be seen in this table, there is a tradeoff between price and resolution. The Ibeo Scala LIDAR has a price-point low enough for certain mass-produced vehicles, but its resolution is not high enough to support advanced autonomous capabilities. In Section 2.3, we present a method to fuse data from a low-cost, low-resolution LIDAR with camera images, to clone a high-cost, high-resolution LIDAR at a fraction of its price.

Manufacturer	Model	Number of Lasers	Data Rate (pts/sec)	Power (Watts)	Cost (USD)
Velodyne	VLS-128	128	9,600,000	Unknown	Unknown
Velodyne	HDL-64	64	1,300,000	60	\$75,000
Velodyne	HDL-32	32	700,000	12	\$30,000
Velodyne	VLP-16	16	300,000	8	\$8,000
Robosense	RS-LIDAR-32	32	640,000	13.5	\$16,800
Ouster	OS-1	64	1,310,720	Unknown	\$12,000
Ibeo	Lux	4-8	Unknown	7-10	\$10,000-20,000
Ibeo	Scala	4	Unknown	7	\$600

Table 1: Comparison of some of the LIDAR sensors that can be purchased today. In most cases, the number of lasers is the vertical resolution; for example, a 4-laser LIDAR has just 4 pixels of vertical resolution. And, the Data Rate is equivalent to $framerate * resolution$.

Thanks to Woodside Capital Partners for this table.

There are numerous companies attempting to improve the LIDAR hardware, including Luminar, Quanergy, and Innoviz, which are not shown in Table 1 because (a) they have not announced their sensor’s specifications, and/or (b) because their sensors are not available for purchase yet. In the future, some of these attempts may produce a durable and high-resolution LIDAR that is cheap enough for the mass market. At the moment however, there does not appear to be a solution which provides the detail needed for an advanced perception system, at a cost needed for mass-market production. As the industry waits for further development, deep learning has recently provided alternative methods to predict the depth from various cheaper sensors.

1.3 Related Work

Depth estimation creates a dense depth map, or an RGB-D image, given no explicit input information about depth. From the 1980s until around 2014, the most widely discussed (and probably most widely used) approach for depth estimation from cameras was *stereo-matching* [11] [17]. Stereo-matching identifies point-correspondences across two cameras and then uses relative position

¹LIDAR information was obtained through http://www.woodsidecap.com/wp-content/uploads/2018/04/Yole_WCP-LiDAR-Report_April-2018-FINAL.pdf

of the two cameras to reconstruct the depth of each point in the image. However, the principal weakness of stereo-matching is the robustness of the point-correspondence algorithms. While better feature engineering (e.g. the invention of SIFT features in 1999 [10, 18]) has led to incremental improvements in the accuracy of point-correspondence algorithms, stereo algorithms still frequently fail to find point-correspondences in numerous situations, such as featureless walls, vegetation, scenes with significant scale changes, and so on [5]. In 2014, Eigen *et al.* published one of the first in a series of papers on using convolutional neural networks to directly regress depth from an single input image, dismissing notions that stereo disparity is essential for depth prediction [2]. Since 2014, further innovations in CNN architecture and loss-functions have yielded additional improvements in depth estimation [12, 22, 19].

A key challenge for depth estimation tasks is collecting the training and evaluation data. Two popular datasets used for depth estimation are KITTI Depth and Make3D, which provide synchronized camera images and dense depth-maps that are derived from interpolating sparse LIDAR point-clouds [2, 16]. ShapeNet is a dataset that uses simulation to generate 3D imagery, with ground-truth 3D information stored in voxels [1]. We compare and contrast these datasets in Table 2. Notably, none of these allow researchers to train a CNN to clone a real sensor; rather, each of these datasets provides ground-truth based on (a) sensor data that is postprocessed with interpolation that hallucinates that data that doesn't exist, or (b) simulation of an imaginary sensor. In contrast to this, our dataset presented in Section 2.2 enables CNNs to be trained to directly clone a \$75,000 LIDAR sensor.

	Inference Data	Ground Truth (GT) Training Data				
Name	Inference Sensors	GT Sensors	GT Data Type	GT Coordinate System	GT Uses Interpolation?	Number of samples
Kitti Depth [4]	Cameras	Velodyne HDL-64 LIDAR	Depth Map	Cartesian	Yes	93,000
Make3D [16]	Mono Camera	custom LIDAR	Depth Map	Cartesian	Yes	400
ShapeNet [1]	Simulation	Simulation	Voxels	Cartesian	No	53,000
DSDepth (ours)	Cameras and Scala	HDL-64	PCDM	Polar	No	78,968

Table 2: Comparison of depth datasets. DSDepth will be explained in Section 2.2.

1.4 Key Contributions

Unlike the previous approaches, we present a solution called Deep Sensor Cloning (DSC), which regresses the depth output of the LIDAR directly. In the process of developing and evaluating DSC, we have made the following key contributions:

1. A novel method of regressing depth using a Point Cloud as a Dense Matrix (PCDM) output format
2. A template for leveraging sensor fusion in CNNs
3. A new set of metrics for evaluating depth estimation in the context of autonomous driving

The rest of this paper is organized as follows. In Section 2, we present our approach to collecting and representing multi-sensor data, and we introduce the DSCnet family of CNN architectures. Next, Section 3 describes our approach for training and evaluating our CNNs. Then, in Section 4, we present qualitative and quantitative results on how well our CNNs can predict depth using only low-cost sensors. We conclude in Section 5.

2 Approach

In this section, we explain our approach for using deep neural networks to ingest data from inexpensive sensors and output a point-cloud that is similar to what is produced by a \$75,000 Velodyne LIDAR. To do this, we develop a custom data format (called a PCDM), create a new dataset (called DSDepth),

design a new family of CNNs (called DSCnets), and propose a loss function (called Sparse L2 Loss). We devote the rest of this section to explaining these concepts.

2.1 PCDM Data Format

Traditional datasets utilizing LIDAR such as KITTI [4], store LIDAR point-cloud data in a cartesian coordinate system, where each point in the scan is represent as a triplet of (x, y, z) . This format can be difficult to use in neural networks, since it is both sparse and three dimensional. VoxelNet uses this format as an input to a 3D object detector by dividing the point cloud in 3D voxels and then transforming each voxel using a voxel feature encoding [23]. While this showed promising results using a sparse input to a CNN, sparse outputs present new and different challenges. To overcome to challenges, We created a novel format for storing point-cloud we call Point Cloud as a Dense Matrix, or PCDM.

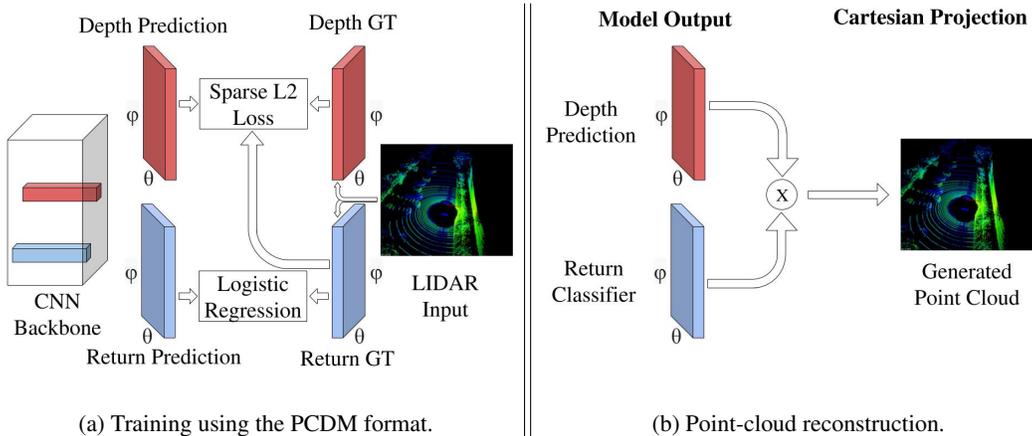


Figure 1: Training and inference methodology using the PCDM format.

The PCDM format is composed of a "Depth" matrix and a "Return" matrix. In the "Depth" matrix, each column corresponds to an angular position of the LIDAR, and each row corresponds to a azimuth positioning of a laser, and the value at each point is the measured distance from the LIDAR.

A LIDAR will not usually have a distance measurement for all locations in its scan, because an object may be too far away, or not have enough reflectance for the LIDAR to receive any reflected signal. For any of scan that does not have a return, the LIDAR will mark its distance as 0. If we simply tried to regress on the sparse "Depth" matrix, CNNs would have difficulty learning the difference between a small distance measurement and a non-return scan. In order to encode this additional information, we create a separate "Return" matrix, which stores a binary value representing whether the LIDAR had a return or not. As can be scene in Figure 1a, when we train our CNN, we mask the gradients for any pixels that do not have a LIDAR return, so the network can ignore any non-returns.

In addition to using the "Return" ground truth to block the gradient, we also create a classifier head for a network to try to replicate its values. By doing so, we can create realistic LIDAR point clouds by masking our depth prediction by our "Return" classifier, as can be seen in Figure 1b.

2.2 DSDepth Dataset

Now that we have defined data format, we describe how we collected a new dataset called DSDepth. In designing this dataset, our goal is to accurately represent two sets of hardware:

- **Expensive Sensors:** Hardware that can be deployed on an expensive (but small) group of cars that are used for data-collection and R&D.
- **Inexpensive Sensors:** Hardware that can be deployed on millions of reasonably low-cost cars that are used every day by consumers and fleet-operators.

In DSDepth, the sole "expensive sensor" is a Velodyne HDL-64 LIDAR (Figure 2a). With a price of \$75,000, the HDL-64 is the one of the most expensive LIDARs in Table 1. While autonomous

R&D vehicles often have other sensors such as cameras and radars, the LIDAR is often the go-to sensor for depth sensing. Given that our goal in this paper is to produce depth estimates (in the form of point-clouds), we think it is reasonable to use the HDL-64 as the sole "expensive sensor" for the purposes of this paper.

In DSDepth, we have also have a set of three "inexpensive sensors." Two of these sensors are cameras, which we have mounted side-by-side on the roof of the car. The third "inexpensive" sensor is an Ibeo Scala LIDAR (Figure 2b). While the \$75,000 Velodyne HDL-64 is too expensive to be deployed on mass-produced cars, the sub-\$1000 Ibeo Scala LIDAR has been deployed on mass-produced vehicles such as the Audi A8.² Note that, while the Scala and HDL-64 are both LIDARs, the Scala has $\frac{1}{16}$ the vertical resolution of the HDL-64, and the Scala's vertical field of view is almost an order-of-magnitude narrower than the HDL-64's vertical field of view.³ In Figure 3, we show how these and other sensors are integrated onto our data-collection car.



(a) Velodyne HDL-64 LIDAR (\$75,000). This is part of our "Expensive" sensor set.



(b) Ibeo Scala LIDAR (under \$1000). This is part of our "Inexpensive" sensor set.

Figure 2: LIDAR sensors mounted on DeepScale's data collection vehicle. Note that the Velodyne has 16x more vertical resolution than the Scala.



(a) Front view

(b) Side view

Figure 3: DeepScale's data collection vehicle.

2.2.1 Data Capture Implementation

We synchronized our sensors using a triggering system, which captures an image when the Velodyne LIDAR was pointed toward the front of the vehicle. We then collected the Velodyne LIDAR data samples from the previous 180 degrees, and next 180 degrees, and we transformed them into a PCDM using the strategy mentioned in Section 2.1. In our training procedure, we crop the PCDM so that all points are included in the field of field of our input images. Further, we capture the most-recent full scan from the inexpensive Scala LIDAR.

²<https://www.businesswire.com/news/home/20180705005220/en/Global-In-vehicle-LiDAR-Industry-Outlook-2022-Expected>

³The HDL-64 as a 26.9-degree vertical field-of-view, and the Scala has only a 3.2-degree vertical field-of-view.

2.3 DSCnet Model Architectures

In contrast to traditional stereo vision algorithms, we propose a family of Deep Sensor Cloning models (called *DSCnets*), which do not require any information of the camera intrinsics or extrinsics for registering with the LIDAR. Rather, we allow the model to learn how to best leverage multiple sensors using end-to-end training. We additionally designed a sensor fusion template that allows us to quickly experiment with various sensor configurations.

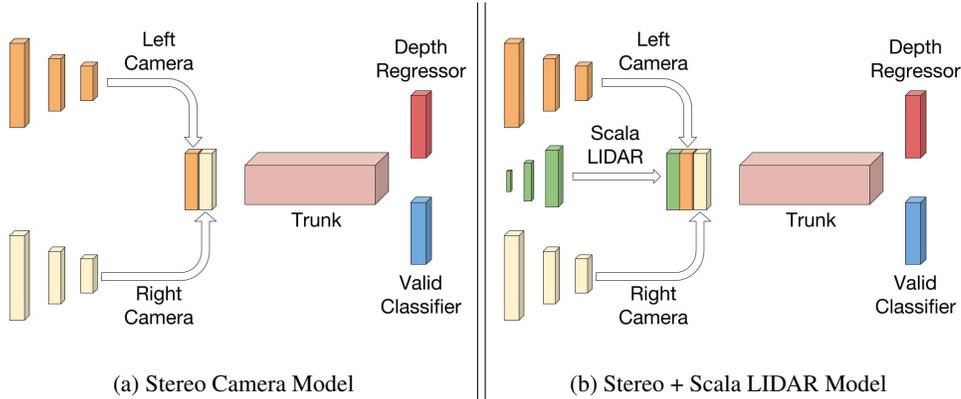


Figure 4: Two examples of our DSCnet architectures

One of our goals in designing the DSCnet family of CNNs was to enable ourselves to quickly experiment with various input sensor sets, and we accomplished that as follows. For each sensor, we created an independent branch of convolutions or deconvolution to both resize the data and learn features that are relevant to that particular sensor. Downstream of these sensor-specific networks, we add a "Trunk", which assumes an input of equal size to the Velodyne LIDAR PCDM, 64 by 256 in our experiments. For our experiments, our "Trunk" is a V-net architecture (inspired by [13]). Figure 4a shows an example of a model which fuses images from a left and right camera, and Table 3 shows the layer parameters of the camera's "resize" branch.

# of Units	Channels	Kernel Dimensions	Stride	Input Size	Output Size
2	8	(3,3)	(1,1)	(576, 768)	(576, 768)
1	16	(5,5)	(3,3)	(576, 768)	(192, 256)
2	16	(3,3)	(1,1)	(192, 256)	(192, 256)
1	32	(5,5)	(3,1)	(192, 256)	(64, 256)
2	32	(3,3)	(1,1)	(64, 768)	(64, 256)

Table 3: Image Resize Branch of DSCnet

The benefit of this approach is that adding additional sensors only requires creating a new branch into the concatenation operator of the network. Additionally, it's much easier to compare sensor configuration, because the trunk network backbone is held constant. An example architecture utilizing the Scala 4 Beam LIDAR is shown in Figure 4b, and Table 4 shows the layer parameters of Scala "resize" branch.

As dicussed in Section 1.2, the Scala LIDAR is a much cheaper but lower resolution LIDAR compared to Velodyne HDL-64. Using this model, we can create a point-cloud with the same resolution as the HDL-64, but at a fraction of its cost.

2.4 Sparse L2 Loss

When training DSCnet, we use separate loss functions for (a) regressing the distance measurements and (b) classifying the valid return in the PCDM. For the classifier, we use a logistic regression loss. As mentioned above, we do not backprop the gradient for scanned points that have no return, so we utilize a modified Least Squares Error, which we call Sparse L2 Loss. It is defined as:

# of Units	Channels	Kernel Dimensions	Stride	Input Size	Output Size
2	64	(3,3)	(1,1)	(4, 192)	(4, 192)
1	128	(3,3)	(1,3)	(4, 192)	(4, 64)
2	128	(3,3)	(1,1)	(4, 64)	(4, 64)
1	64	(3,3)	($\frac{1}{2}, \frac{1}{2}$)	(4, 64)	(8, 128)
2	64	(3,3)	(1,1)	(8, 128)	(8, 128)
1	32	(3,3)	($\frac{1}{2}, \frac{1}{2}$)	(8, 128)	(16, 256)
2	32	(3,3)	(1,1)	(16, 256)	(16, 256)
1	16	(3,3)	($\frac{1}{2}, 1$)	(16, 256)	(32, 256)
2	16	(3,3)	(1,1)	(32, 256)	(32, 256)
1	16	(3,3)	($\frac{1}{2}, 1$)	(32, 256)	(64, 256)
2	16	(3,3)	(1,1)	(64, 256)	(64, 256)

Table 4: Scala LIDAR Resize Branch of DSCnet

$$L = \frac{1}{N} \sum_{i=0}^n ((D_i - f(X_i) * V_i)$$

where D_i is Depth Ground Truth, $f(X_i)$ is depth prediction, and V_i is the "Return" mask, as described in section 2.1.

Finally, we block the gradient from the classifier one layer before the loss function, so the trunk network is only trained to predict distance. Empirically, we found this to be sufficient to train the classifier head.

3 Training and Evaluation Methodology

3.1 Training Routine

Our DSCnet models was trained on 58853 training samples and evaluated on 20115 validation samples. We use a stochastic gradient descent optimizer with a learning rate of 0.013, momentum of 0.9, and weight decay of 0.0005, and decrease the learning rate by a factor of 0.2 every 60,000 iterations. We use a batch size of 48 and train across 3 Nvidia Titan Xp GPUs.

3.2 Evaluation Metrics

A number of metrics have been established in the research community to evaluate the correctness of depth-estimation algorithms. These metrics include:

1. Relative absolute error (percent): $\frac{1}{N} \sum_y \frac{|y-y^*|}{y^*}$
2. Relative squared error (percent): $\frac{1}{N} \sum_y \frac{(y-y^*)^2}{y^{*2}}$
3. Root mean squared error of inverse depth [1/km]: $\sqrt{\frac{1}{N} \sum_y \left| \frac{1}{y} - \frac{1}{y^*} \right|^2}$
4. Scale invariant logarithmic error [1/km]: $\frac{1}{N} \sum_i d_i^2 - \frac{1}{N^2} (\sum_i d_i)^2$ where $d_i = \log y_i - \log y_i^*$

where y is the predicted distance, and y^* is the distance ground truth.

In Table 5, we show a snapshot of the current state-of-the-art results on the leaderboard for the KITTI Depth challenge [2]. Out of the metrics shown on Table 5, we find Absolute Relative Error (absErrorRel) to be particularly intuitive. When driving a car, when we encounter an object that is 1 meter away, we care far more about 1-meter error than we do for an object that is 100 meters away. The absErrorRel metric takes this into account – a 1 meter error on a 100-meter-away object is worth the same absErrorRel penalty as a 0.01-meter error on a 1-meter-away object.

Dataset	CNN	Model Input	absErrorRel	sqErrorRel	iRMSE	SILog	Return Classifier Error	GFLOP	Parameter Size (MB)
KITTI Depth	DORN [3]	Mono camera	8.78	2.23	12.98	11.77	N/A	N/A	N/A

Table 5: Snapshot of the top result on the KITTI Depth leaderboard as of November 2018 [2].

We have also added a few new metrics to our evaluation that are not included in the KITTI leaderboard. In particular, since we are training DSCnet to mimic the sparsity pattern of an expensive Velodyne LIDAR, we report the accuracy of our return-classifier (see Section 2.4). Further, for our experiments in the next section, we will report the model’s resource utilization in terms of computation (GFLOP per inference) and parameter file size (in megabytes).

3.3 Metric Zones

In autonomous driving applications, some areas are more critical than others to have accurate depth information. In adaptive cruise control for example, the distance measurements directly in front of the vehicle are much more important than those to the side. In order to create a better evaluation our models, we designed metric zones for a few different autonomous vehicle applications, and calculated the above metrics for each of these zones. The metric zones are defined in Table 6.

Name	Min Distance (m)	Max Distance (m)	Horizontal Field of View (degrees)
Parking Assist	0	10	44
Adaptive Cruise Control (Highway)	0	100	11.06
Collision Detection (Urban)	0	30	27.66

Table 6: Automotive Metric Zones

4 Results

4.1 Quantitative Results

Dataset	CNN	Model Input	absErrorRel	sqErrorRel	iRMSE	SILog	Return Classifier Error	GFLOP	Parameter Size (MB)
DSDepth	DSCnet	Mono camera	5.77	4.16	8.04	11.22	4.79	8.79	82.21
DSDepth	DSCnet	Stereo camera	4.69	2.91	6.90	9.21	4.54	11.26	82.36
DSDepth	DSCnet	Stereo + Scala	4.37	2.77	6.86	8.89	4.61	12.62	85.24

Table 7: DSCnet results with different sets of input sensors

In Table 7, we show results from various input sensor configurations across our evaluation metrics on the DSDepth test set. As can be seen, with only a monocular camera as input, DSCnet achieves under 6% absolute relative error. Also, with each additional input, our model improves across all evaluation metrics other than the return classifier error. This result is particularly exciting because of the minimal amount of effort required to incorporate new sensors.

Dataset	CNN	Model Input	Parking Assist	Collision Detection	Adaptive Cruise Control	Overall
DSDepth	DSCnet	Mono camera	3.49	4.47	5.29	5.77
DSDepth	DSCnet	Stereo camera	3.00	3.61	4.30	4.69
DSDepth	DSCnet	Stereo + Scala	2.99	3.52	4.13	4.37

Table 8: Relative Error (absErrorRel) for DSCnet in the automotive metric zones

4.1.1 Automotive Metric Zones

As discussed in Section 3.3, we also evaluated our models using metrics designed for automotive use cases in Table 8. As can be seen, our models achieved significantly lower relative error in each of the automotive metric zones when compared error across the entire point cloud. This result shows areas that are both far away and in the vehicle’s periphery account for the majority of the error.

4.1.2 DSCnet-lite

Dataset	CNN	Model Input	absErrorRel	sqErrorRel	iRMSE	SILog	Return Classifier Error	GFLOP	Parameter Size (MB)
DSDepth	DSCnet	Stereo Camera	4.69	2.91	6.90	9.21	4.54	11.26	82.36
DSDepth	DSCnet-lite	Stereo Camera	6.42	6.51	11.31	14.44	5.50	2.30	1.83

Table 9: Evaluation of our DSCnet-lite model for embedded devices

As we mentioned earlier in the paper, CNNs have yielded a dramatic improvement of the state-of-the-art error-rate on a variety of computer vision tasks including image classification, object detection, semantic segmentation, and depth estimation. However, CNNs often require far more resources (e.g. computation, memory, time, and energy) than previous computer vision methods. This is of particular concern when deploying CNNs on embedded platforms such as smartphones, security cameras, and low-cost automotive-grade processors. To mitigate this issue, researchers have developed resource-efficient CNNs such as SqueezeNet [7] and MobileNet [6] for image classification; YOLO [15] and SqueezeDet [21] for object detection; and ENet [14] and SqueezeNet-based models [20] for semantic segmentation. But, resource-efficient CNNs for depth estimation is a relatively untapped field. To begin to address this opportunity, we have created a resource-efficient version of DSCnet called DSCnet-lite.

For brevity, we omit the precise dimensions of DSCnet-lite. But, in order to reduce the number of parameters and floating point operations, one of the techniques behind DSCnet-lite is to replace dense convolutions with depthwise separable convolutions, similar to [6].

We show a quantitative evaluation of DSCnet-lite in Table 9. Going from DSCnet to DSCnet-lite, we have reduced the computational cost by 4.9x (to 2.3 GFLOP per inference), and we have reduced the quantity of parameters by 45x (to 1.83 MB). This yields a modest increase in the error-rate (from 4.6% absolute relative error for DSCnet to 6.4% absolute relative error in DSCnet-lite). With our own CNN framework running on a garden-variety 4-core ARM A72 processor (found in millions of smartphones today) and without using any type of GPU or accelerator, we can routinely run CNN inference at 12.5 GFLOP/s, which implies that we should be able to run DSCnet-lite at over 5 inferences-per-second⁴ on a generic smartphone processor. Further, many of today’s server GPUs are able to run CNNs at much more than 1 TFLOP/s, but if we conservatively envision the case of running on a GPU at 1 TFLOP/s, we could do over 400 inferences-per-second with DSCnet-lite.

⁴We say inferences-per-second instead of frames-per-second, because we are talking about two-camera input in this example.

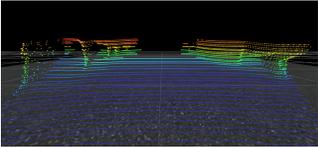
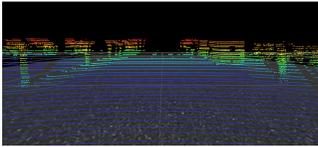
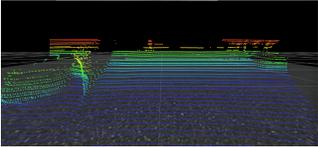
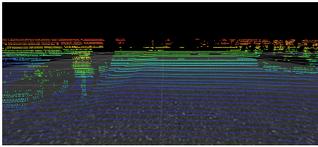
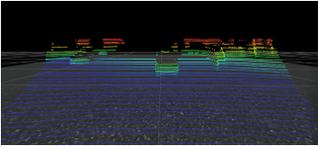
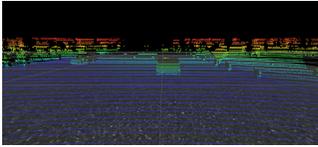
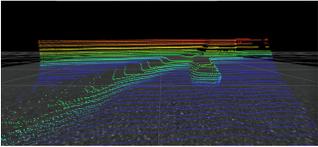
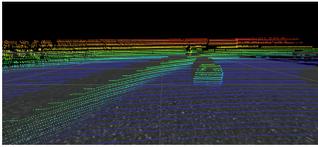
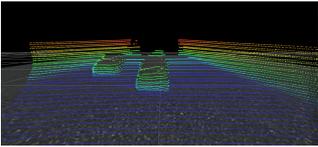
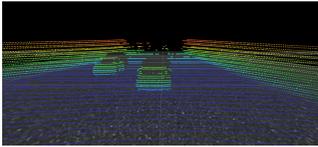
Example	DSCnet Output	HDL-64 Output	Input Image
1			
2			
3			
4			
5			

Figure 5: Qualitative examples of DSCnet results compared to ground-truth data from a Velodyne HDL-64 LIDAR. In these results, the inputs to DSCnet are stereo cameras and Scala data.

4.2 Qualitative Results

In Figure 5, we visualize the generated point-cloud from our of our DSCnet model (using two cameras and Scala data as inputs to DSCnet).

In Examples 1 and 2, you can see that DSCnet’s generated point cloud looks similar to the ground truth LIDAR, and the model is able to distinguish the depths of objects such as cars, trees and traffic light poles and signs, as well as the ground plane.

Example 3 shows our depth prediction near the beginning of a construction site. While our DSCnet model performs well on both the cars and the ground plane, DSCnet does not correctly predict the depth of the orange traffic cones along the right side of the road.

In Examples 4 and 5, we visualize DSCnet results for predicting depth on the highway. In both examples, DSCnet is able to perceive the rough depth for the cars in front of the ego vehicle, as well as the road boundaries.

5 Conclusion

Expensive sensors such as Velodyne HDL-64 LIDAR are commonly used in autonomous vehicle research. However, due principally to their high cost, these expensive LIDARs are difficult to deploy in mass-market vehicles that are manufactured in the millions of units per year. In this work, we have created a family of neural network architectures called DSCnet, which can be trained to "clone" expensive LIDAR while using only low-cost sensors as input. We defined new metric zones for calculating distance predictions for the use of autonomous driving, and showed our DSCnet models could help perform certain perception tasks at a fraction of the price. While LIDAR may still be needed for fully-autonomous driving, we feel that DSCnets running on low-cost sensors can provide high-quality real-time 3D data for semi-automation, or as a backup solution to systems relying on LIDAR. Finally, we are interested to see how the emerging research field of Deep Sensor Cloning will impact the cost, quality, and reliability of autonomous vehicles and other applications.

References

- [1] Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., and Yu, F. (2015). ShapeNet: An information-rich 3d model repository. *arXiv:1512.03012*.
- [2] Eigen, D., Puhrsch, C., and Fergus, R. (2014). Depth map prediction from a single image using a multi-scale deep network. In *NIPS*.
- [3] Fu, H., Gong, M., Wang, C., Batmanghelich, K., and Tao, D. (2018). Deep Ordinal Regression Network for Monocular Depth Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [4] Geiger, A., Lenz, P., and Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- [5] Howard, A. (2008). Real-time stereo visual odometry for autonomous ground vehicles. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.
- [6] Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., and Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv:1704.04861*.
- [7] Iandola, F. N., Han, S., Moskewicz, M. W., Ashraf, K., Dally, W. J., and Keutzer, K. (2016). SqueezeNet: Alexnet-level accuracy with 50x fewer parameters and <0.5mb model size. *arXiv:1602.07360*.
- [8] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*.
- [9] Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [10] Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision (ICCV)*.
- [11] Lucas, B. D. and Kanade, T. (1981). An iterative image registration technique with an application to stereo vision. In *IJCAI*.
- [12] Memisevic, R. and Conrad, C. (2014). Stereopsis via deep learning. In *NIPS*.
- [13] Milletari, F., Navab, N., and Ahmadi, S. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. *arXiv:1606.04797*.
- [14] Paszke, A., Chaurasia, A., Kim, S., and Curcuriello, E. (2016). Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv:1606.02147*.
- [15] Redmon, J., Divvala, S. K., Girshick, R. B., and Farhadi, A. (2015). You only look once: Unified, real-time object detection. *arXiv:1506.02640*.
- [16] Saxena, A., Sun, M., and Ng, A. Y. (2009). Make3D: Learning 3d scene structure from a single still image. *IEEE Trans. Pattern Anal. Mach. Intell.*
- [17] Scharstein, D. and Szeliski, R. (2002). A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*.
- [18] Se, S., Lowe, D., and Little, J. (2001). Vision-based mobile robot localization and mapping using scale-invariant features. In *IEEE International Conference on Robotics and Automation (ICRA)*.
- [19] Sinz, F. H., Candela, J. Q., Bakır, G. H., Rasmussen, C. E., and Franz, M. O. (2004). Learning depth from stereo. In *Joint Pattern Recognition Symposium*.
- [20] Treml, M., Arjona-Medina, J., Unterthiner, T., Durgesh, R., Friedmann, F., Schuberth, P., Mayr, A., Heusel, M., Hofmarcher, M., Widrich, M., Bodenhofer, U., Nessler, B., and Hochreiter, S. (2016). Speeding up semantic segmentation for autonomous driving. In *NIPS MLITS Workshop*.
- [21] Wu, B., Iandola, F., Jin, P. H., and Keutzer, K. (2016). SqueezeDet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. *arXiv:1612.01051*.
- [22] Yamaguchi, K., Hazan, T., McAllester, D., and Urtasun, R. (2012). Continuous markov random fields for robust stereo estimation. In *ECCV*.
- [23] Zhou, Y. and Tuzel, O. (2017). Voxelnet: End-to-end learning for point cloud based 3d object detection. *arXiv:1711.06396*.