

10 Lidar Datasets for Autonomous Driving

By [Tobias Cornille](#) on April 25th, 2022

Lidar sensors use laser beams to capture the world in 3D. The sensors output 3D point clouds, which are simply collections of points in 3D. Machine learning models can be used to detect and track objects in these point clouds, or even to classify every single point (segmentation). This enables autonomous vehicles to understand their surroundings, and can also be used to make cities smarter, to create AR/VR applications, and for indoor design/real estate applications.

In this article, we give an overview of 10 public labeled lidar datasets that you can use in your autonomous driving projects. The mentioned datasets contain either 3D bounding box (cuboid) labels or segmentation labels. We'll also show how you can create your own 3D point cloud dataset, in case the open datasets do not fit your use case or if their licenses are too restrictive (only 2 datasets can be used commercially).



3D point cloud driving datasets

1. KITTI

KITTI is a dataset of lidar sequences of street scenes in Karlsruhe, Germany. The dataset was launched in 2012 and different labels have been added over the years.

License: [Creative Commons Attribution-NonCommercial-ShareAlike 3.0](#)

Lidar sensor: Velodyne HDL-64E

Tasks

- [Object detection](#)

7481 labeled frames, 8 categories

[Information + download](#) (account required)

[Label format](#)

- [Multi-object tracking](#)

21 labeled sequences, 8 categories

[Information + download](#) (account required)

[Label format](#)



- [Semantic + panoptic segmentation \(SemanticKITTI\)](#)

11 labeled sequences, 28 categories

[Information + label format](#) (email required)

[Download](#)

2. nuScenes

nuScenes is a large-scale autonomous driving dataset consisting of urban street scenes captured in Singapore and Boston, U.S.

[Download](#) (account required)

License: [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International](#), or [acquire a commercial license](#)

Tasks

- [Multi-object tracking](#)

850 labeled sequences, 23 categories

[Information](#)

- [Semantic + panoptic segmentation \(nuScenes-lidarseg\)](#)

850 labeled sequences, 32 categories

[Information](#)

3. Waymo Open Perception Dataset

Waymo Open is a diverse autonomous driving dataset. It includes scenes captured in 6 U.S. areas in a wide variety of environments and weather conditions.

[Dataset information](#)

[Download](#) (Google account required)

License: [custom non-commercial license](#)

Lidar sensors: 1x mid-range + 4x short-range

Tasks

- [Multi-object tracking](#)

1200 labeled sequences, 4 categories

- [Semantic segmentation](#)

1150 labeled sequences, 23 categories

4. A2D2

A2D2 stands for Audi Autonomous Driving Dataset (A2D2). The data was captured in 3 German cities.

[Dataset information](#)

[Download](#)

License: [Creative Commons Attribution-NoDerivatives 4.0 International](#)

Lidar sensors: 5x sensor

Tasks

- [Multi-object tracking](#)

12,499 labeled frames, 14 categories

- [Semantic segmentation](#)



41,280 labeled frames, 38 categories

Labels obtained from 2D semantic segmentation on camera images

5. Argoverse 2 Sensor

Argoverse 2 is a collection of open-source autonomous driving data from six U.S. cities.

License: [Attribution-NonCommercial-ShareAlike 4.0 International](#)

Lidar sensors: 2x Velodyne VLP-32C

Tasks

- [Multi-object tracking](#)

1000 labeled sequences, 30 categories

[Information](#)

[Download](#)

6. ApolloScape

ApolloScape is an autonomous driving dataset created by Baidu research. The dataset was collected under various lighting conditions and traffic densities in Beijing, China.

License: academic use only

Lidar sensors: 2x Riegl VMX-1HA

Tasks

- [Multi-object tracking](#)

53 labeled sequences, 5 categories

[Information](#)

[Download](#) (Google/Baidu/Github account required)

7. PandaSet

PandaSet is a high-quality dataset for autonomous driving created by lidar producer Hesai. Its 100+ scenes are selected from two routes in Silicon Valley.

[Dataset information](#)

[Download](#) (account required)

License: [Creative Commons Attribution 4.0 International](#)

Lidar sensors: 1x Pandar64 + 1x PandarGT

Tasks

- [Multi-object tracking](#)

28 categories

- [Semantic segmentation](#)

37 categories

8. Winter Adverse Driving dataSet (WADS)

WADS is a dataset of 20 scenes for autonomous driving collected in severe winter weather in Michigan, U.S.

[Dataset information + download](#)

License: [Attribution-NonCommercial-ShareAlike 4.0 International](#)

Tasks



- [Semantic + panoptic segmentation](#)

22 categories

9. DENSE Seeing Through Fog

Seeing Through Fog is a driving dataset part of the DENSE project. The data includes different weather conditions like fog, snow, and rain and was captured in northern Europe.

[Dataset information](#)

[Download](#) (registration required)

License: [custom academic license](#)

Lidar sensor: Velodyne HDL-64E S3

Tasks

- [Object detection](#)

12000 labeled frames, 28 categories

10. Toronto-3D

Toronto-3D is a detailed dataset of 1km of road in Toronto, Canada.

[Dataset information](#)

[Download](#)

License: [Attribution-NonCommercial 4.0 International](#)

Lidar sensor: Teledyne Optech Maverick

Tasks

- [Semantic segmentation](#)

8 categories

11. Bonus: JackRabbit Dataset and Benchmark (JRDB)

JRDB is a dataset collected by a social robot called JackRabbit. It features sequences from different indoor and outdoor locations on the Stanford University campus. Since the robot's size is comparable to a human, the data has a different perspective than the other car-based datasets.

[Dataset information](#)

[Download](#) (account required)

License: [Attribution-NonCommercial-ShareAlike 3.0 Unported](#)

Lidar sensors: 2x Velodyne VLP-16

Tasks

- [Pedestrian tracking](#)

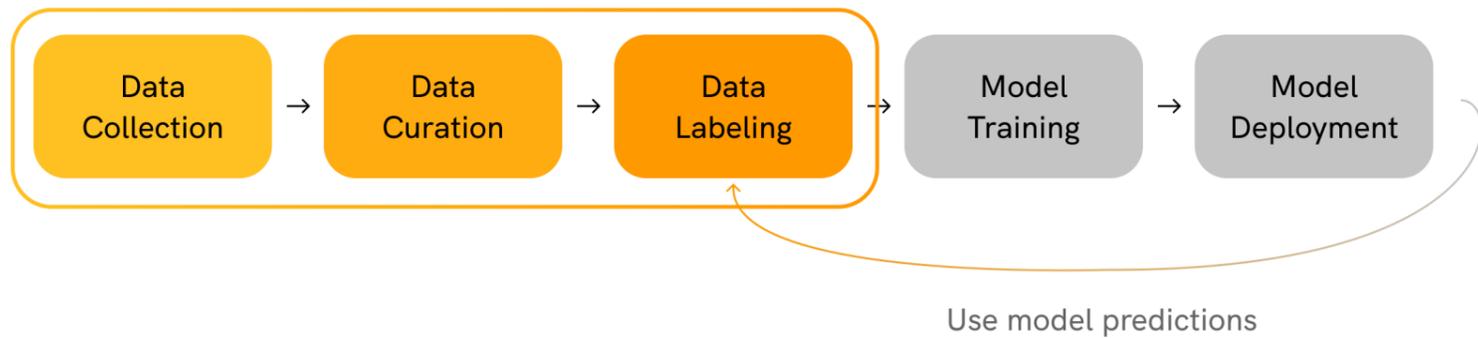
57600 labeled frames

[Label format](#)

Creating your own lidar dataset

Every machine learning system needs the right data to perform well. Public datasets can help you experiment quickly, but they often are not suited for training your final models. The ML models might perform worse when you deploy them, because the data in public datasets might be recorded in a different environment (country, weather condition), or because your sensor set-up is different. To avoid this performance loss, we'll show you how you can create your own 3D point cloud dataset.





Creating a dataset requires three steps:

1. Data collection

Data collection involves acquiring the right tool to capture new data, e.g. a vehicle with a lidar sensor, and going out and capturing the data. You should capture data in an environment that matches the real production environment as closely as possible.

2. Data selection/curation

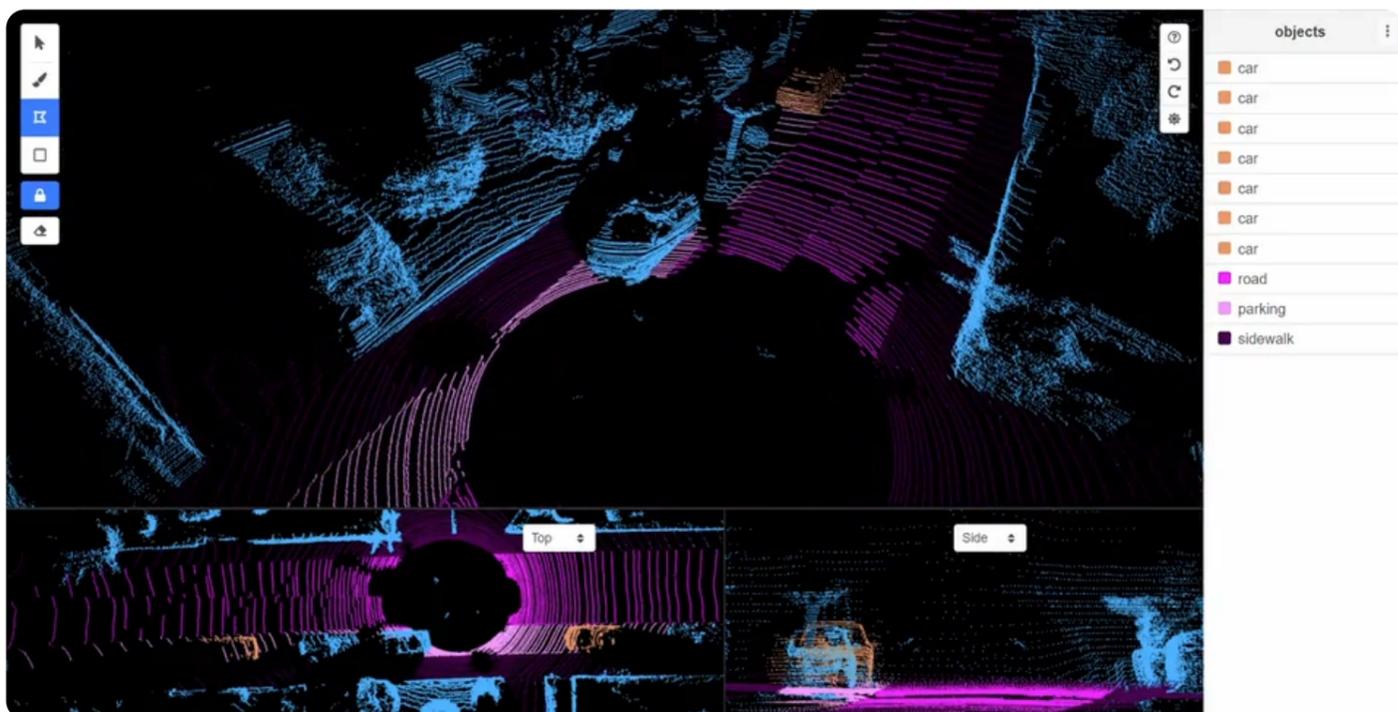
Next, you often have to select which captured data you want to include in the dataset, as it can be infeasible and inefficient to use all captured data. Here, it is important to choose diverse data that covers all the different scenarios you captured. Discarding boring data can speed up labeling and improve model performance.

3. Data labeling

Finally, you have to label your selected data. For object detection/tracking, this means drawing 3D bounding boxes (cuboids) around the objects you want to detect. For segmentation, you have to **annotate** the individual points in your point clouds. This can be a tedious and time-consuming process, but with the right tools you can speed up your labeling significantly.

[Segments.ai](#) has [dedicated labeling interfaces](#) for 3D point cloud data. If you work with sequential data, you can use our interpolation feature to label faster. To speed up your labeling even further, Segments.ai also allows you to set up [model-assisted workflows](#), where you train an initial model on a small set of labeled data, and then use the model to help label the complete dataset. Finally, you can choose whether to label the data in-house or [work with an external workforce](#).

Check out [SemanticKITTI](#) on Segments.ai. You can also [try the platform for free](#) for 14 days, or [book a demo](#). We're always happy to see if [Segments.ai](#) is the right fit for your use case, so do not hesitate to get in touch.



A showcase of Segments.ai's lidar interface

Conclusion



Autonomous vehicles use lidar sensors to see the world around them in 3D. To detect objects and understand the scene, we need 3D point cloud datasets. In this article, we highlighted 10 lidar datasets for autonomous driving. The datasets can be used for tasks such as 3D object detection, 3D MOTs, and 3D point cloud segmentation.

If you want to create a machine learning model for a different application, if you want to use different categories, or if your data in production differs from the data in the public datasets, you'll have to create your own dataset. For this, you need to collect, curate, and label data. For lidar data, [Segments.ai](#) is the best tool for labeling your data and managing the labeling workforce.

Hope this was useful! If you have any questions or suggestions, feel free to send us an email at support@segments.ai



TOBIAS CORNILLE
April 25th, 2022

SHARE:   

