mindkosh

Whitepaper

1-Click Cuboid Annotation using adaptive clustering

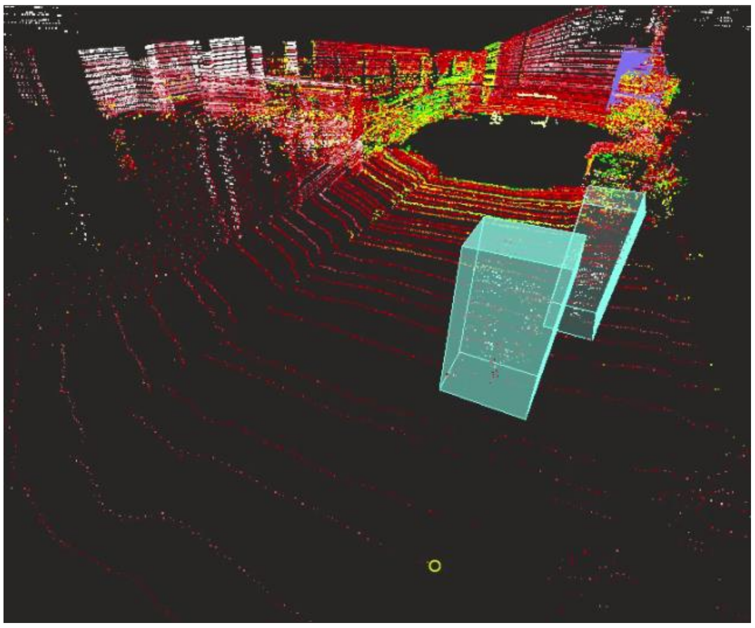
May 2025

Contents

The annotation problem in 3D point clouds

High-quality 3D object detection relies heavily on annotated pointcloud data. However, creating such labeled datasets is expensive, error-prone, and time-consuming, particularly when annotating cuboids in sparse LiDAR scans. While manually labeling 2D images with bounding boxes is relatively intuitive — objects are visible, well-defined, and anchored to familiar visual context, labeling 3D pointclouds with cuboids is far more complex.

First, accurately drawing the cuboid in 3D, adjusting its dimensions and rotating it to fit the object is time consuming.
Second, due to a variety of factors including sparsity of points, occlusion and noise, labelers often need to infer the full 3D shape and orientation from partial geometry. This requires guess work and a clear understanding of the entire scene.



Img-1. Labeling point clouds with cuboids can be difficult, specially in areas that are sparse or occluded.

Our solution

We introduce a point-and-click tool for 3D cuboid generation using an unsupervised clustering method, designed for fast, scalable annotation. It works on datasets from a variety of use-cases including, self-driving, medical imaging and logistics. Here is a brief overview of how it works:

Input: User clicks on a point belonging to the target object in the point cloud.

Local Clustering: Nearby points are clustered into a group.

Ground removal: A lightweight ground segmentation model filters out ground points.

Cuboid fit: A cuboid is fit covering the points.

Shape correction: The resulting cuboid is adjusted to class-specific default dimensions.

This entire process is extremely fast, and depending on the size of the object, requires very little waiting time from the user. In our experiments, we have found the response times to be between 0.2 to 0.8 seconds, depending on the hardware.

Why unsupervised?

Supervised Machine Learning models trained on large amounts of data have been at the fore-front of the rapid progress of AI systems in the last decade. However, in the context of 3D point clouds, particularly when using them to help annotation, supervised methods suffer from severe limitations.

Data Scarcity

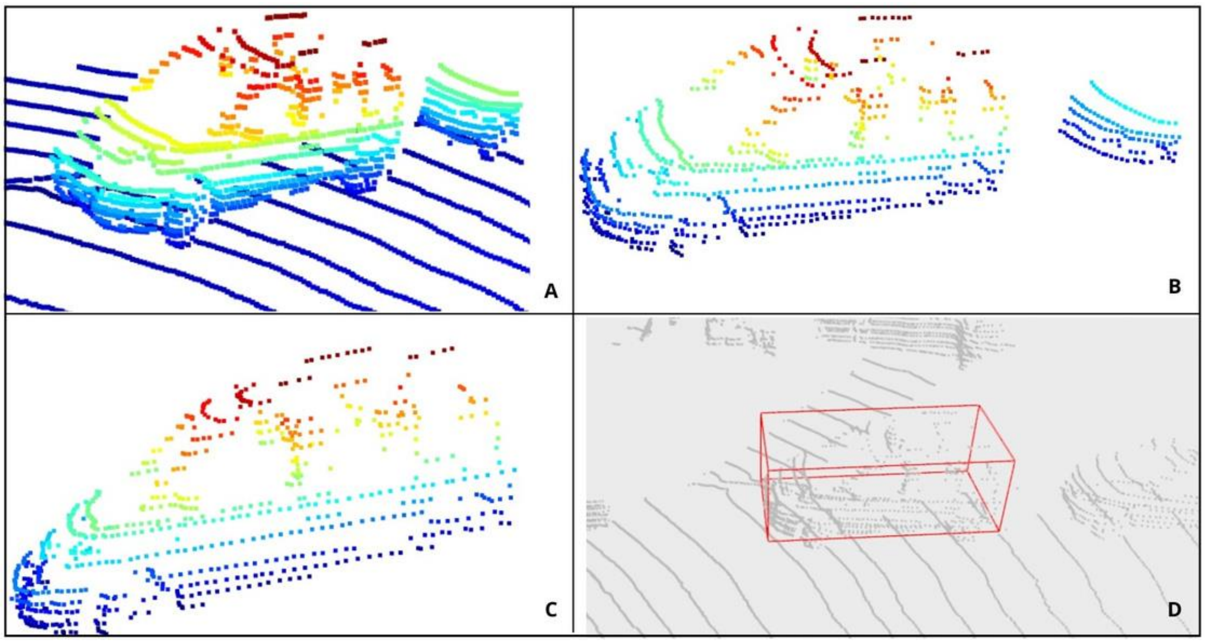
Labeled 3D datasets are limited in size and scope. They are also much more resource intensive to collect and then label. In addition, ML models trained on one dataset do not generalize well beyond their specific domains or environments.

Class and Context dependence

ML models trained on specific object types (e.g., cars) in one environment often perform poorly when deployed in different contexts or with unseen classes. They are also limited to classes they have seen, leading to limited applicability to new datasets.

Our methodology

Our approach is inspired by the clustering method proposed in Latte, but we make several key enhancements to improve its robustness, accuracy, and usability in real-world annotation pipelines. These are explained below.



Img-2 Steps in our clustering method. Figure A shows the raw point-cloud around the target point. Figure B shows the result of our local ground-removal method. Figure C shows points clustered from the non-ground points returned in B. Finally Figure D shows a cuboid fitting the clustered points

Clustering without pre-segmented ground points

Latte requires the point clouds to be pre-segmented into ground and non-ground points. This adds an extra layer of complexity to the process. In addition, segmenting entire point clouds into ground/non-ground points is not trivial and can lead to poor results. Most ground segmentation methods try to fit the ground points into a 2D plane. However, this approach does not work well if the ground is uneven. Our method does not require any preprocessing step, resulting in better adaptation across point-clouds with different densities and terrains.

Adaptive parameters for varying Pointcloud densities

In order to make our method suitable across different types of datasets, we introduce adaptive thresholding for clustering based on local point density. This allows our method to work well across scans captured from sensors with varying resolutions and under diverse environmental conditions.

Simple ground detection for object isolation

Rather than requiring the ground segmentation of the entire scene, we use a lightweight, local ground detection model tailored to the target object. This simplifies the architecture and vastly increases the overall speed. Ground points below individual objects can be easily modeled with a plane, which results is more consistent ground point removal.

Cuboid correction using default class dimensions

A big issue when using clustering to annotate objects, is when only a few points belonging to the object are visible. This can happen if the objects far away from the sensor, if the object is occluded, or only one of the sides is visible.
To address these issues, we apply a correction step that fits a cuboid to resemble default dimensions for the object's class. These default dimensions act as a prior and help regularize the cuboid fitting process, especially when data is sparse or occluded.

Evaluation Strategy

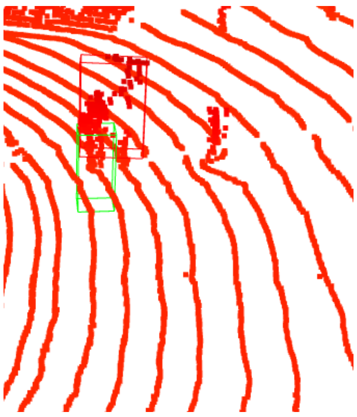
Traditional evaluation of interactive annotation tools like ours often relies on measuring the labeling time or gathering user feedback. Latte utilizes the same approach. However, these metrics are subjective, time consuming and hard to standardize. Instead, we take an objective approach based on measuring the geometric accuracy of predicted annotations against known ground truth datasets.

Evaluation Datasets

We use three well-established datasets for evaluation:

KITTI

We use the 3D object detection evaluation dataset found here. Our evaluation is done against the train dataset, which consists of 7481 point cloud files with ~80K labeled objects. Note that the KITTI dataset only contains labels for objects that are visible in the reference camera images. Also note that the cuboids are labeled in 2D only. The object's height and it's placement in z-direction are not accurate. Both of these points however, do not change our evaluation, since we use BEV method to calculate the IoU and only run our system against cuboids present in the dataset.



Img-3 The red cuboid shows predicted cuboid around the pedestrian, while the green cuboid shows the annotated object from KITTI. As can be seen, the GT label is going into the ground and does not cover the entire pedestrian in the z-axis.

nuScenes

We also test our system against the nuScenes dataset. We run our evaluation on the *Trainval* dataset, which contains annotated objects across 700 training and 150 validation scenes. Compared to KITTI, the nuscenes point clouds are much sparser, resulting in a lot of occlusion and missing points. To alleviate this issue, we only consider labeled objects with points > 10 and visibility > 20%.

ONCE

Finally we test our system against the ONCE dataset. Our evaluation is run over the test split of the dataset which contains 15K fully annotated scenes with 5 classes - Car, Bus, Truck, Pedestrian and Cyclist. Like nuScenes, the point clouds in ONCE are also sparse. We also note that pedestrians are labeled with loose cuboids, resulting in an overall lower IoU compared to cuboids predicted by our method.

Methodology

Simulation of the Annotation process

To simulate the user’s "1-click" action, we use the center of each ground truth cuboid and find the closest point in the raw point-cloud that falls within the cuboid. This selected point acts as the input to our system.

Evaluation metric

We use the BEV (Bird's Eye View) Intersection-over-Union (IoU) metric to compare predicted cuboids with ground truth. Following KITTI conventions, we report recall at a 70% IoU threshold for cars/vehicles and 50% for other classes. Because each Ground Truth cuboid is guaranteed to yield one simulation input and our system outputs only one cuboid per input (if the method is successful), precision is not applicable in our evaluation framework.

Default dimensions

Default class dimensions are computed as the average size of cuboids for each class in the respective dataset. On our annotation platform, users can input expected default dimensions, or let the platform keep a running average based on existing annotations. Thus, this process is transferable and customizable to real-world use cases.

Results and comparisons

In this section, we compare our system's performance with the original Latte method. The key results are highlighted in Table 1. Note that Latte is only tested against KITTI, so we only report the figures mentioned in their paper.

Dataset	Avg IoU (Ours)	Avg IoU (Latte)	Recall (Ours)	Recall (Latte)	Recall w/o Correction
KITTI	0.8316	82.9	**0.8607**	84.8	0.8434
nuScenes	0.7987	NA	**0.8213**	NA	0.7952
ONCE	0.7687	NA	**0.7943**	NA	0.7812

Table 1 - The Recall column reports recall values using the entirety of our system. The last column mentions recall values when not using the last step of our system, where we adjust the predicted cuboid to fit a cuboid with default dimensions.

There are several reasons for the disparity in performance of our system on KITTI compared to the other datasets. One, the point clouds in KITTI are much denser compared to the other two datasets. Two, as mentioned in the KITTI paper, annotators were instructed to only label objects that were clearly visible. Per the paper:

"The annotators were asked to draw bounding boxes for instances for which they feel confident, for example instances of vehicles where at least two complete edges are visible. Instances that are far away tend to be sparse or occluded and are therefore not annotated."

This results in annotated objects that always have a good number of points, and precludes objects that are highly occluded or very far away. Since these are the major failure cases for our system, not having them in the test set results in higher accuracy values.

Qualitative Observations

In qualitative analysis, we observe several patterns:

Annotation of pedestrians

Pedestrians in nuScenes are often labeled with loose-fitting cuboids, reducing the reported IoU despite visually reasonable predictions. Pedestrians in KITTI, on the other hand, are only labeled only in X and Y axes, ignoring the Z dimension completely.

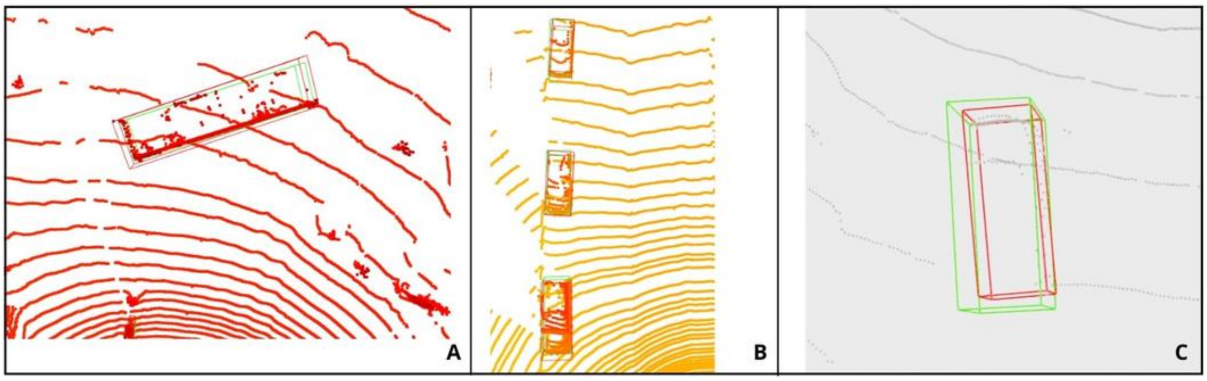
Impact of point density

Our system performs best when an object is well-scanned, i.e., with a dense point distribution. Performance degrades for distant or partially occluded objects. As already mentioned earlier, this is why the performance of our system is better on the KITTI dataset compared to other datasets.

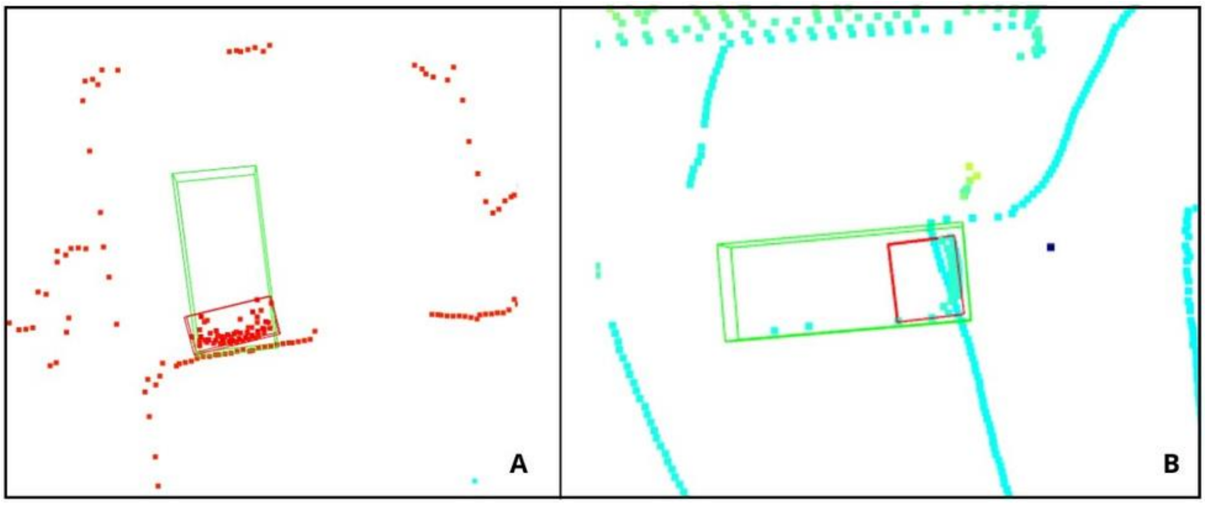
Effectiveness of dimension correction

The correction step using class-specific default dimensions substantially improves both the shape and alignment of the generated cuboids, especially for classes with consistent size (e.g., cars, trucks). This is reflected in the reported values in table 1 above.

Example predictions



Img-4. Examples from KITTI (Figures A and B) and Nuscenes (Figure C). Red cuboids are predictions from our system, while green cuboids are ground truth cuboids.



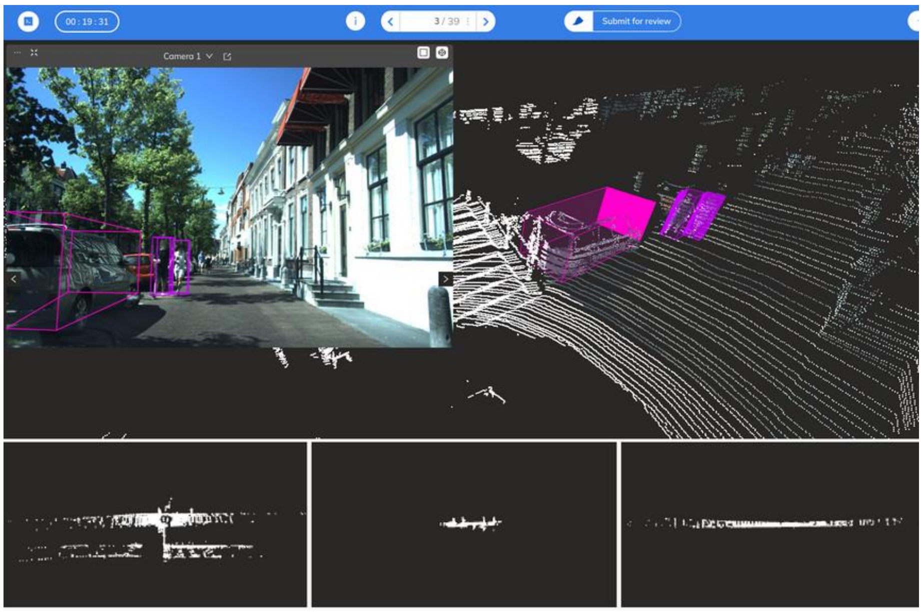
Img-5. Examples from KITTI (Figures A) and Nuscenes (Figure C). Due to only the front side of the vehicles being visible, our method (sans adjustment from default dimensions) only creates a cuboid around the visible points in close proximity. The last step in our method helps alleviate this very issue. Red cuboids are predictions from our system, while green cuboids are ground truth cuboids.

Built into Mindkosh

Our 1-click cuboid generation system is fully integrated into our annotation platform, designed for large scale datasets with multiple sensors.

Single-click interaction: Annotators simply click on a point within an object — the system clusters nearby points, fits a cuboid, and adjusts its shape using predefined dimensions.

Custom class dimensions: For each object class (e.g., car, truck, pedestrian), users can optionally define expected default dimensions. These help ensure reliable cuboid fits, even in sparse or partially occluded scans.



Labeling the KITTI dataset on Mindkosh.

Want to see Mindkosh in action? Let us know!

+91-6388-797751
hello@mindkosh.com
mindkosh.com