Mizzou INformation and Data FUsion Lab (MINDFUL)

Title: Ignorance is Bliss: Flawed Assumptions in Simulated Ground Truth

Authors: Andrew R. Buck, Derek T. Anderson, Joshua Fraser,
Jeffrey Kerley, and Kannappan Palaniappan

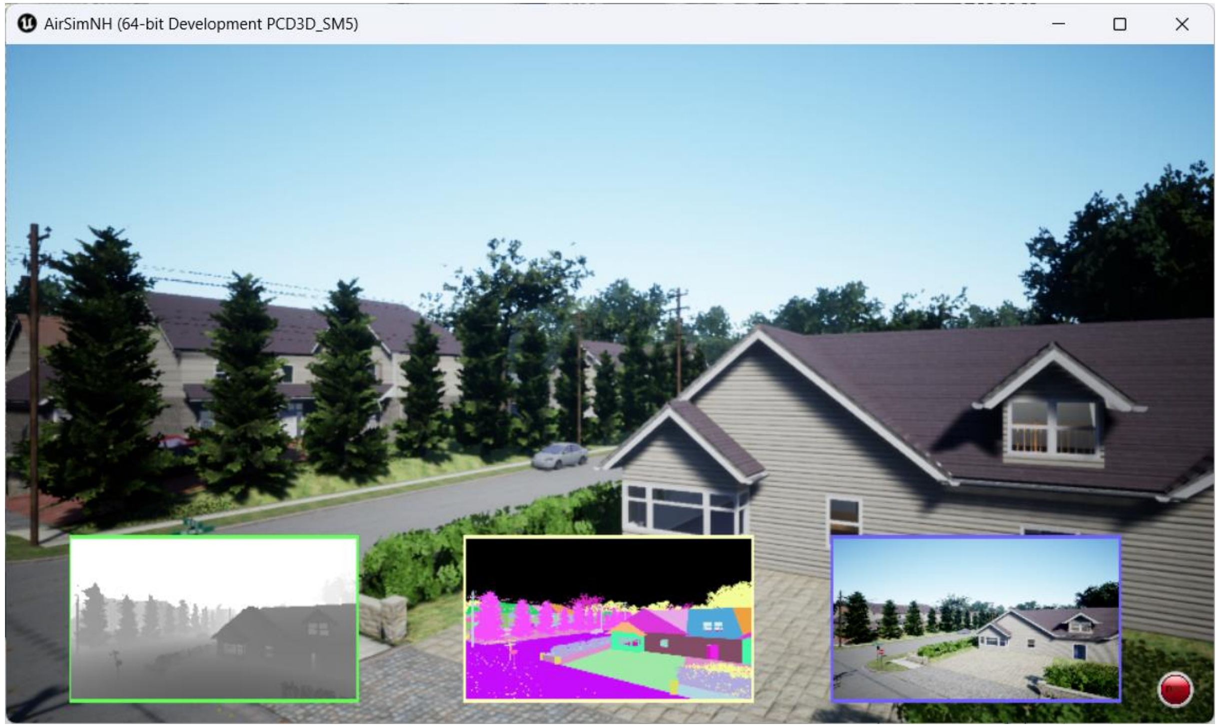
University of Missouri

Department of Electrical Engineering and Computer Science

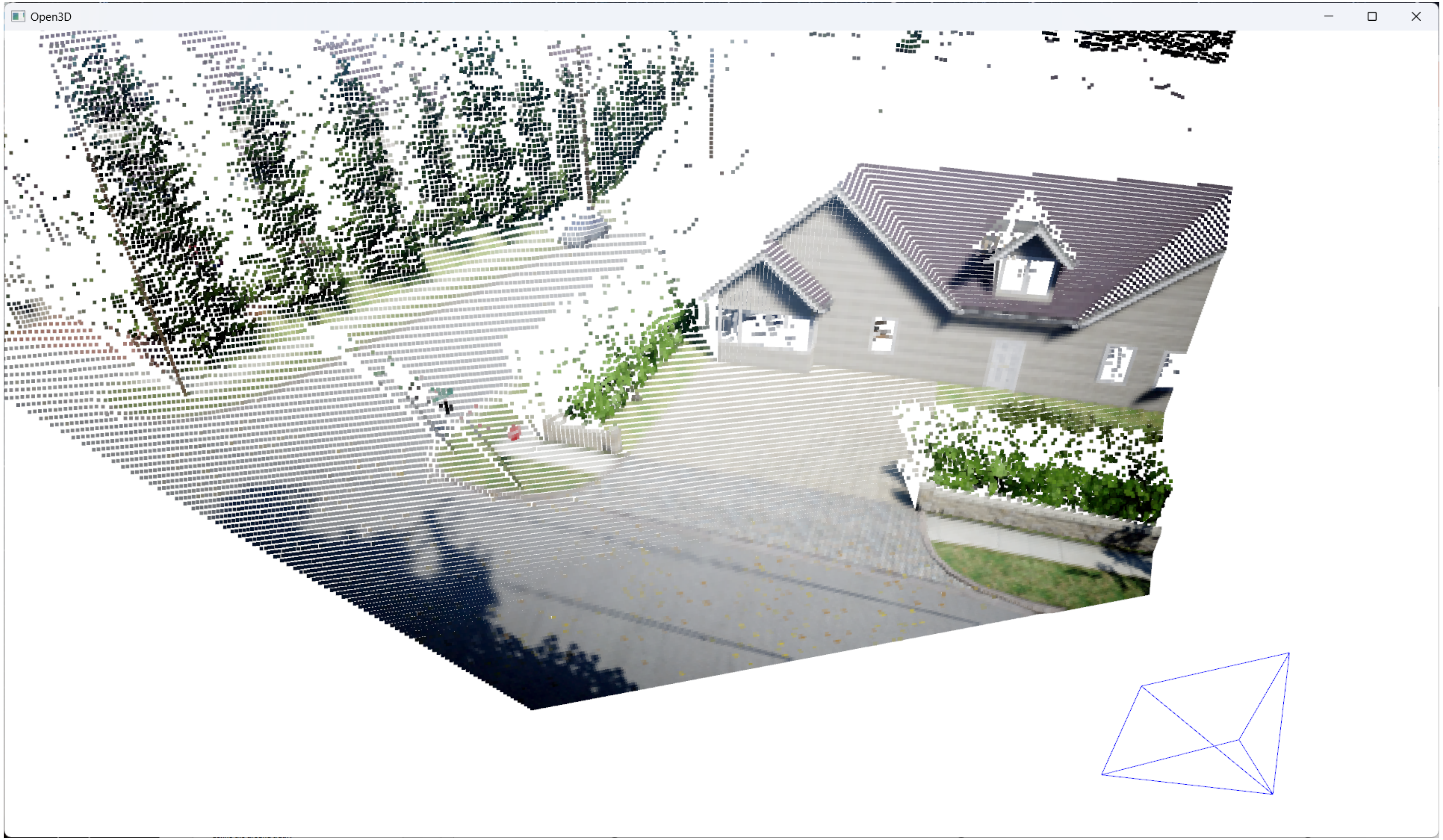
May 1st, 2023

Why Are We Doing This?

- We want a 3D simulator for generating synthetic data with ground truth.



Why Are We Doing This?



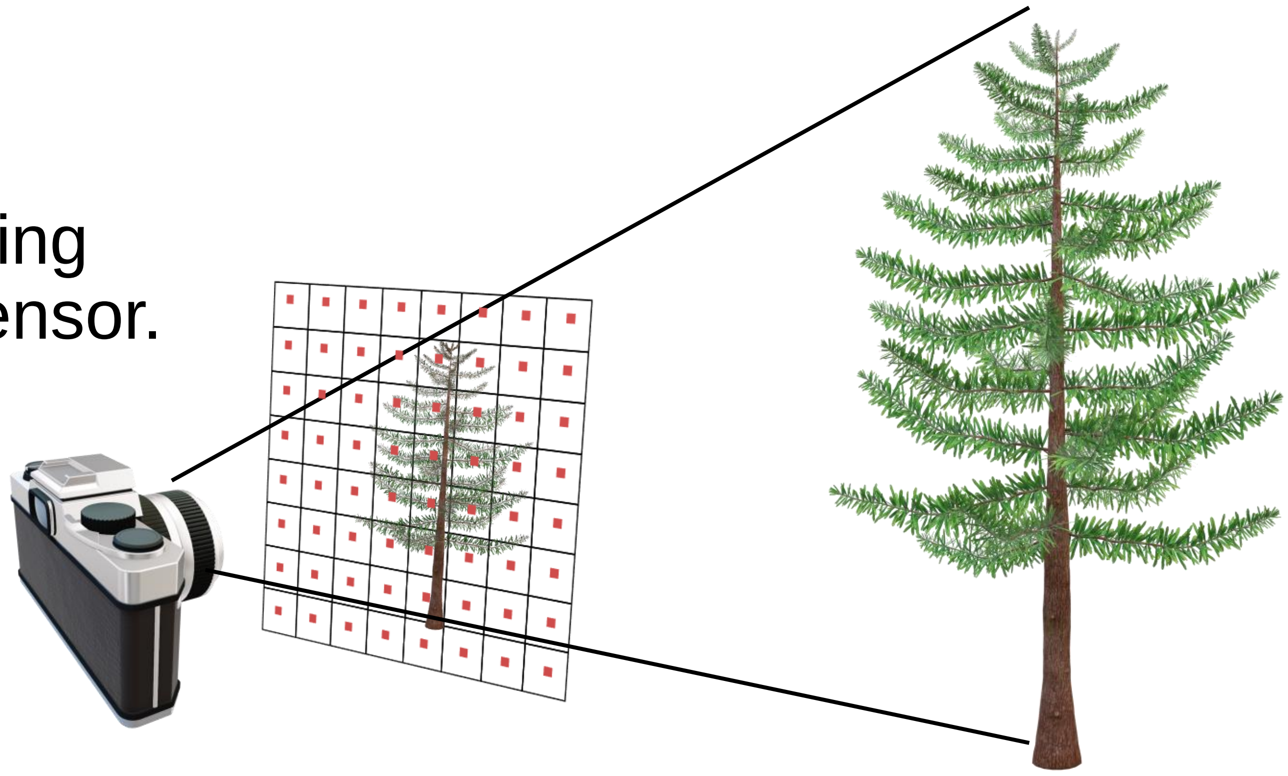
Introduction

- **What is “ground truth?”**
  - From Wikipedia: “Ground truth is information that is known to be real or true, provided by direct observation and measurement (i.e. empirical evidence) as opposed to information provided by inference.”

- **Where does it come from?**
  - Depends on the application and context
  - In remote sensing, it refers to what actually exists in the world for each pixel in an image.

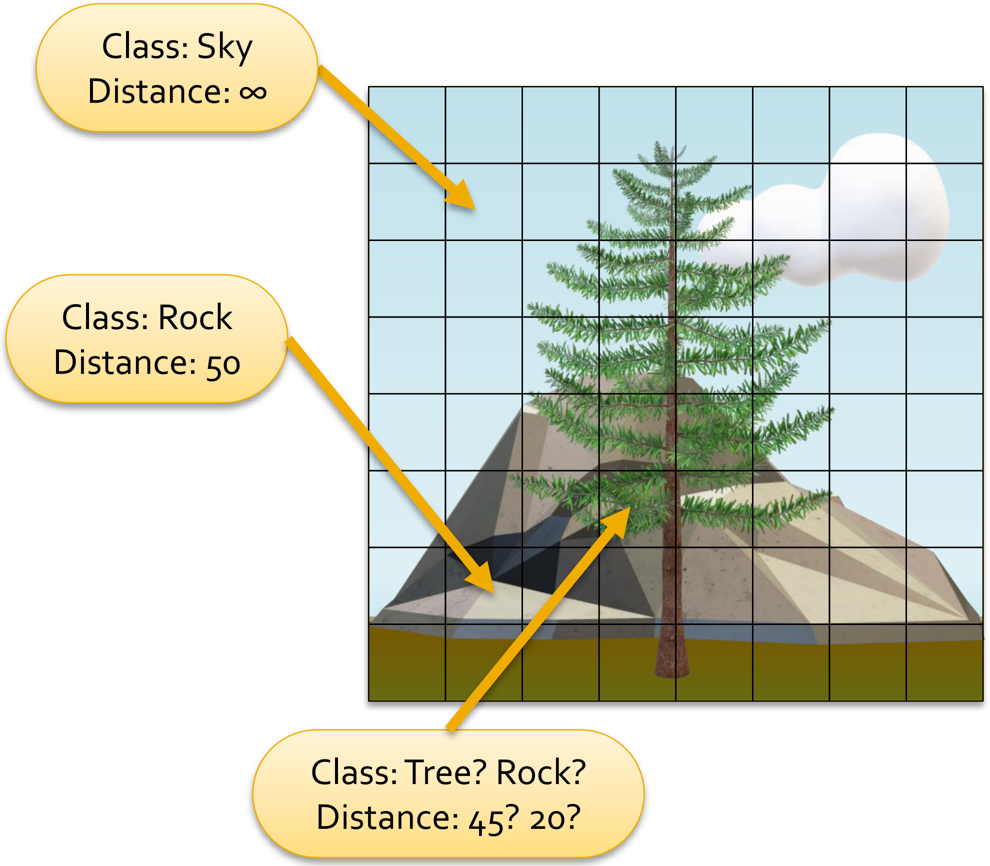
The Meaning of a Pixel

- **What is a pixel?**
  - “Not a little square!” – Alvy Ray Smith
  - Sampled points on a grid
- **In photography,**
  - Each pixel is a discrete sampling of the light that reaches the sensor.
  - Pixels aggregate all this information into a single scalar value.
  - Color (and other features) can be represented with multiple image channels.



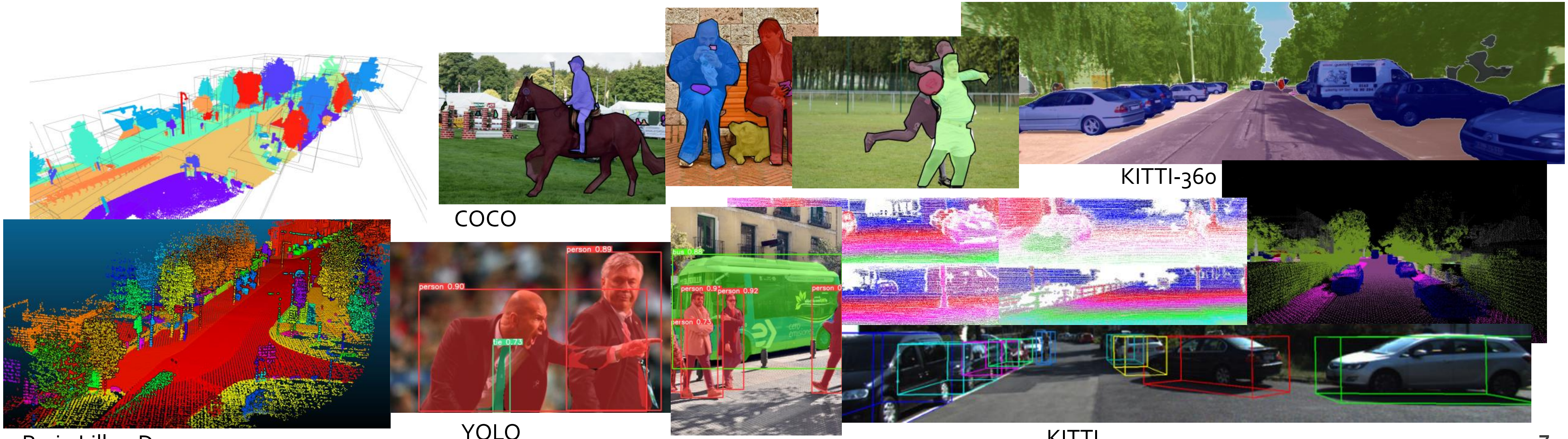
What Is Truth?

- **Because pixels aggregate information, how do we define the ground truth?**
  - Each pixel only gets one value
    - Class label
    - Depth
  - However, sometimes it's not clear what value to assign.
  - We can increase resolution, but this doesn't solve the underlying problem.



Hand Annotation

- **A lot of effort can go into hand-labeling data**
  - But how accurate is it?
  - Pixel-level accuracy is hard to come by.
  - We often use coarse labels (e.g. bounding boxes, image classes)



COCO

KITTI-360

Paris-Lille-3D

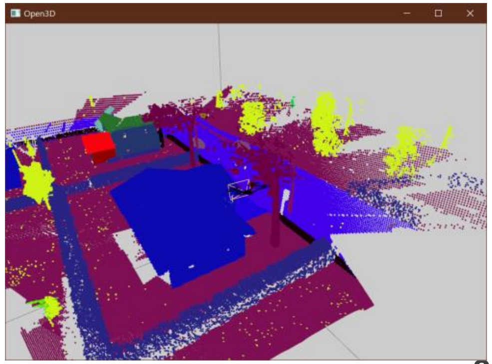
YOLO

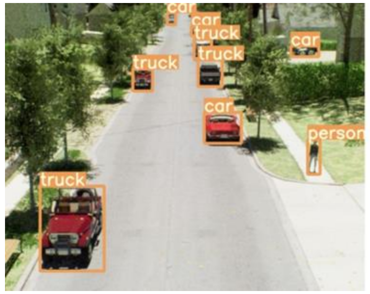
KITTI

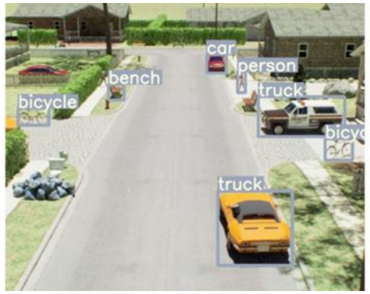
Using Simulated Data

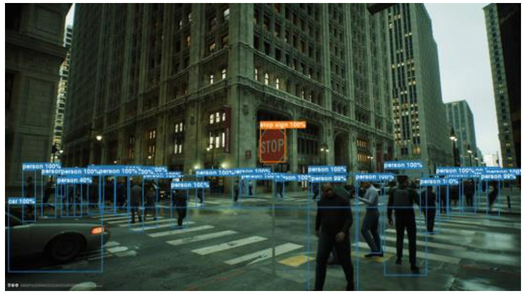
- **Synthetic data can provide “ground truth”**
  - Automatically generated alongside data
    - Object detections
    - Semantic labels
    - Depth
- **However, even simulated ground truth isn’t perfect.**







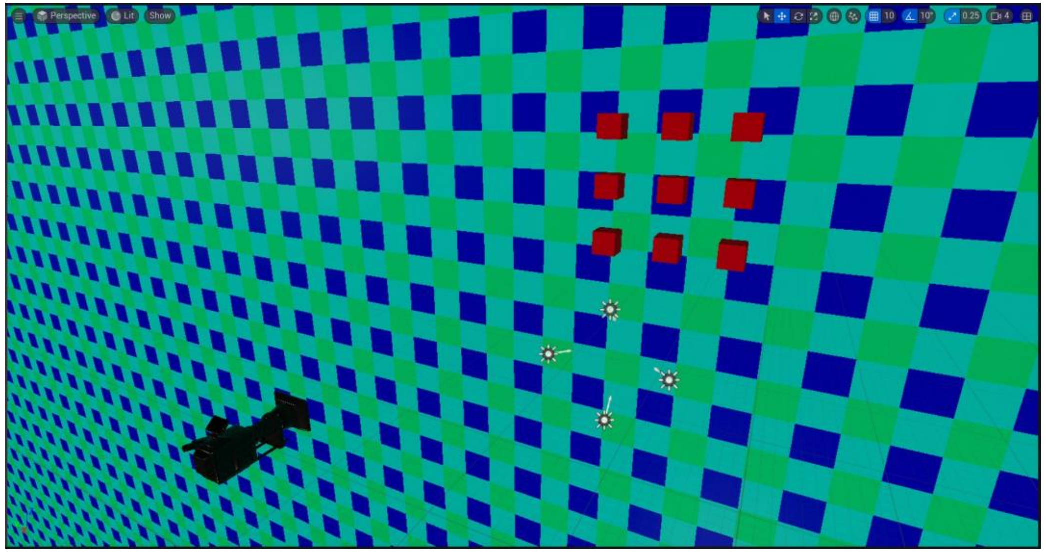




Experiments

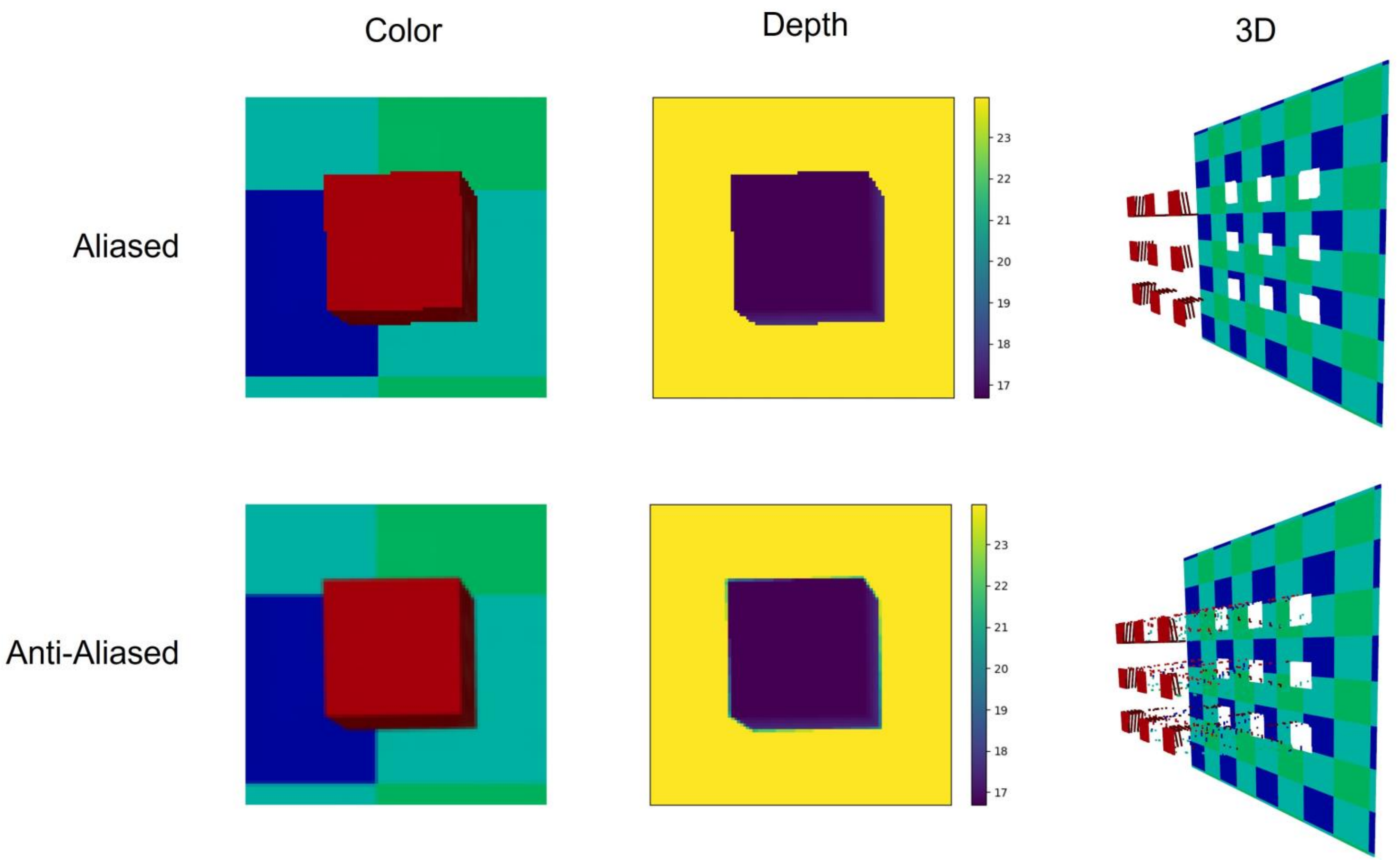
- **We designed a series of experiments to study the issues associated with simulated ground truth.**
  - Focus on single image depth estimation
  - Simple dataset to understand fundamentals (nothing fancy)

- Scene consists of rotating cubes in front of a flat plane
  - Cubes are red. Background has green/blue checkerboard pattern.
    - Should be able to learn that red=near and blue/green=far
  - Background plane is at various depths.
    - Want to learn how cube size relates to depth
  - Collect 40 images at 24 different background depths. (960 images total)



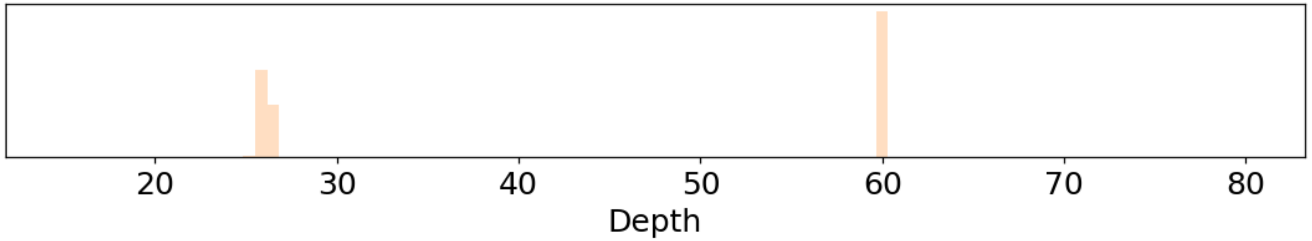
Aliased vs Anti-aliased

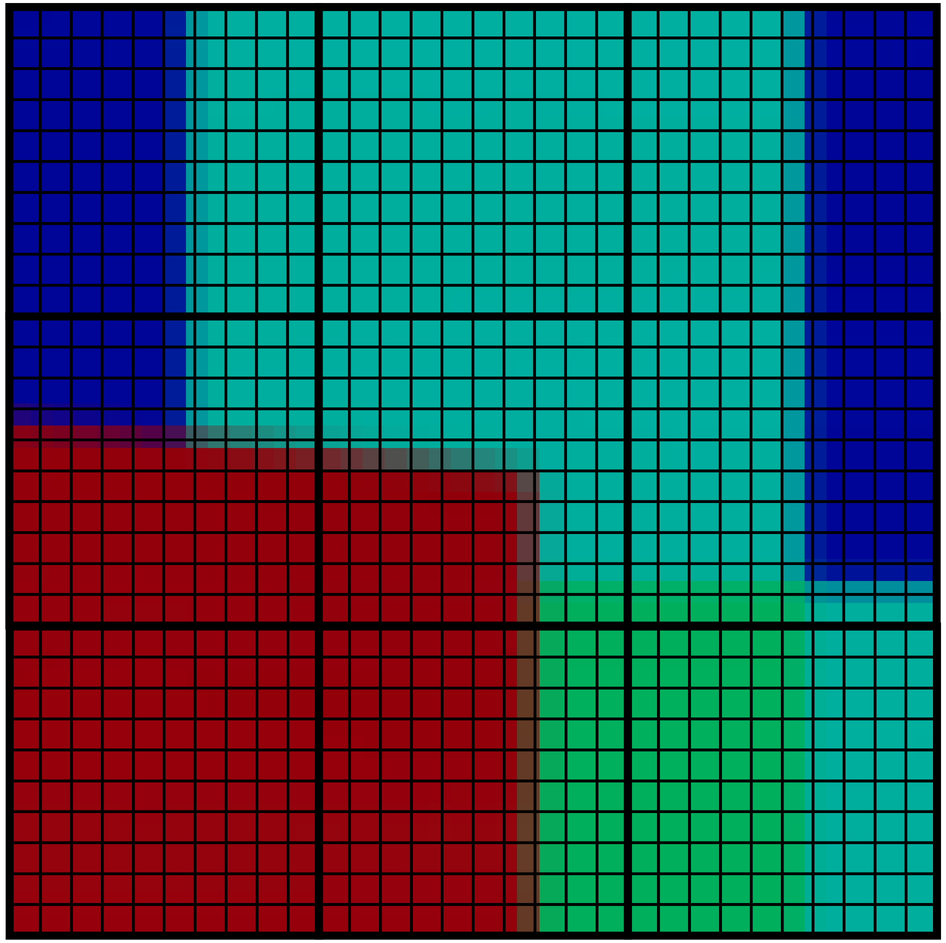
- We collected both aliased and anti-aliased imagery



Bundled Depth

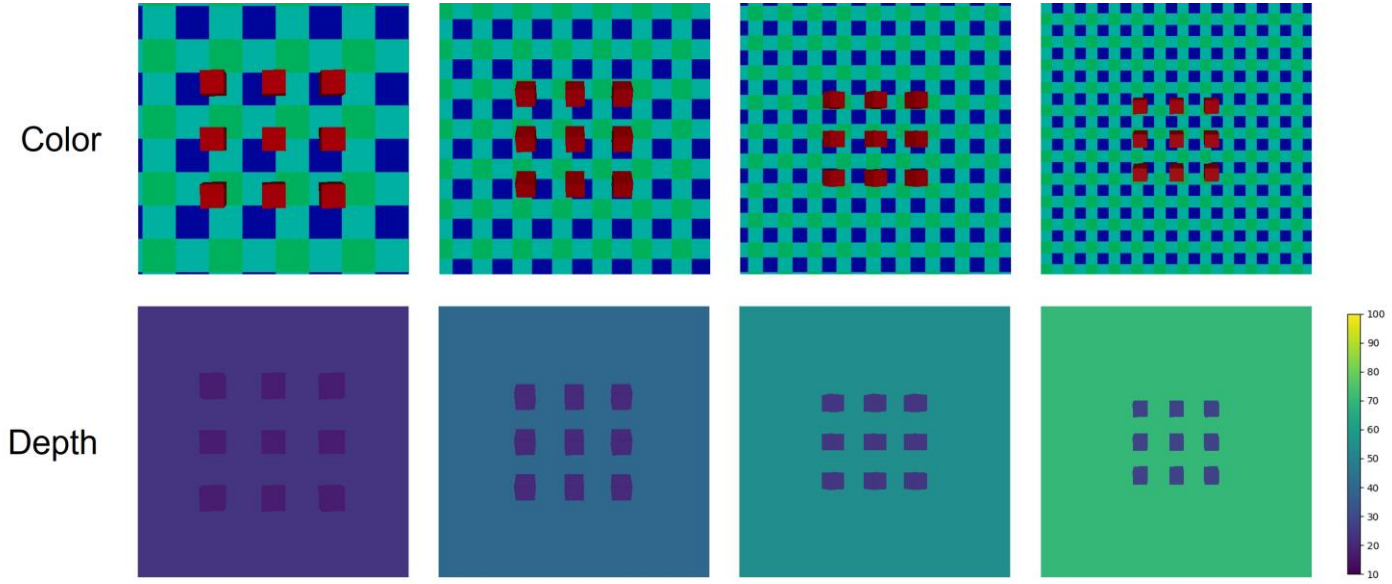
- **We also collect a high-resolution image**
  - Upscaled 10x
  - Each pixel now has 100 depth samples
  - We store these as an array of values for each pixel
  - This is an alternative to aliased or anti-aliased imagery





Depth Estimation Model

- **We use a Resnet18 depth network from Monodepth2**
  - Train/test on interleaved sets (even/odd)
  - Trained for 30 epochs
  - Output is mapped to a fixed range between 10 and 100 meters

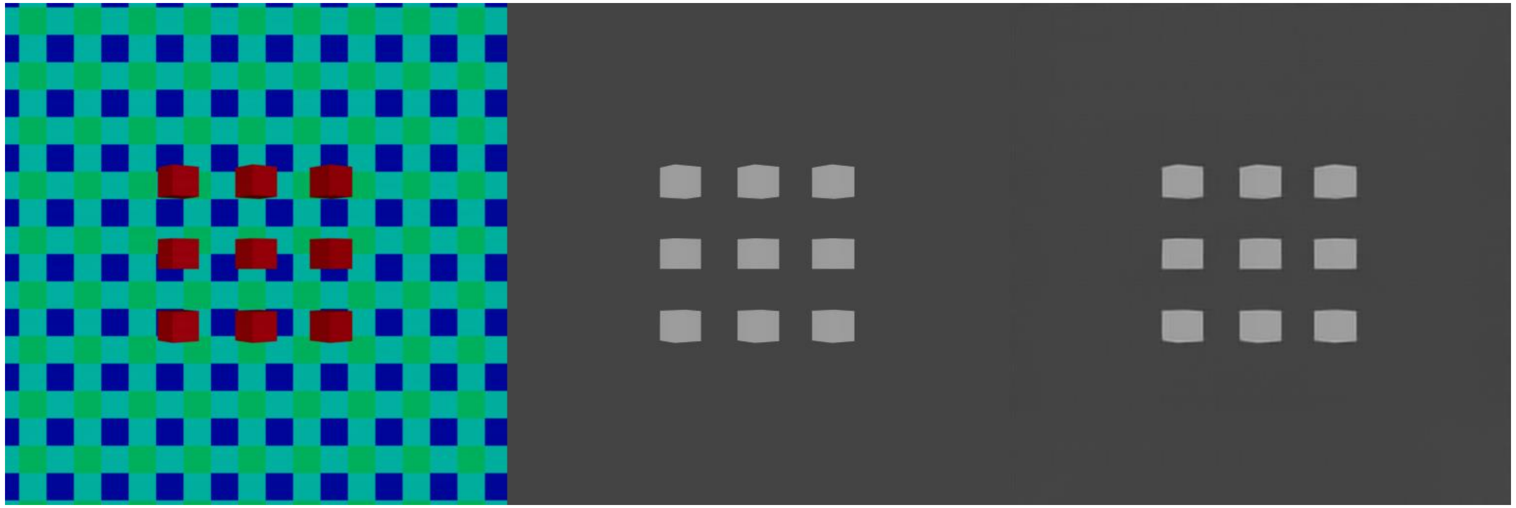


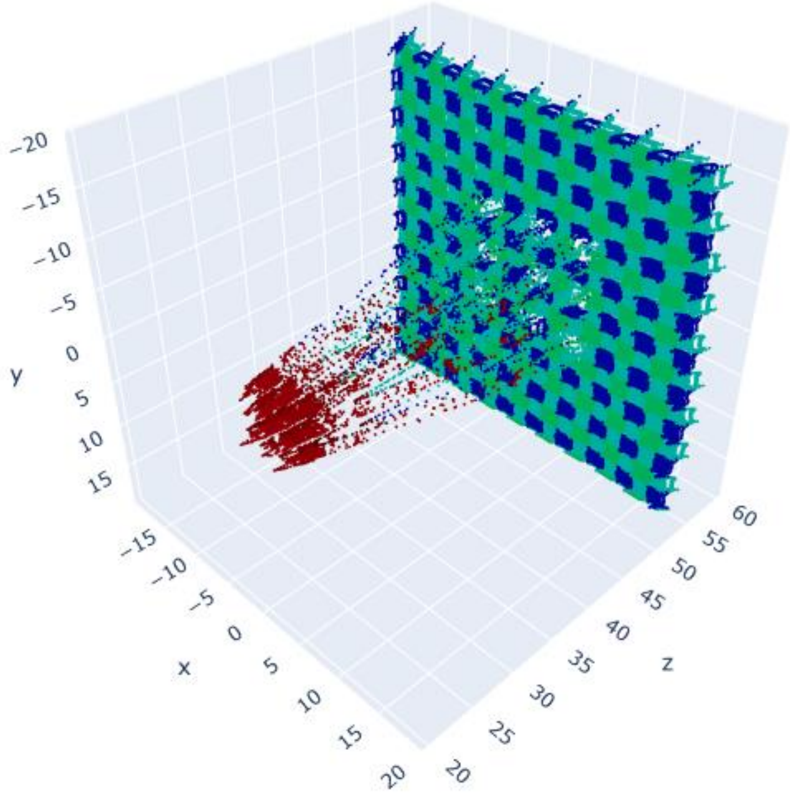
Ex. 1: Lie in the Data

- **GT is clearly wrong**
  - Anti-aliased color
  - Anti-aliased depth

$$L(X,Y) = \frac{1}{N}\sum_i (\log(Y_i) - \log(X_i))^2$$

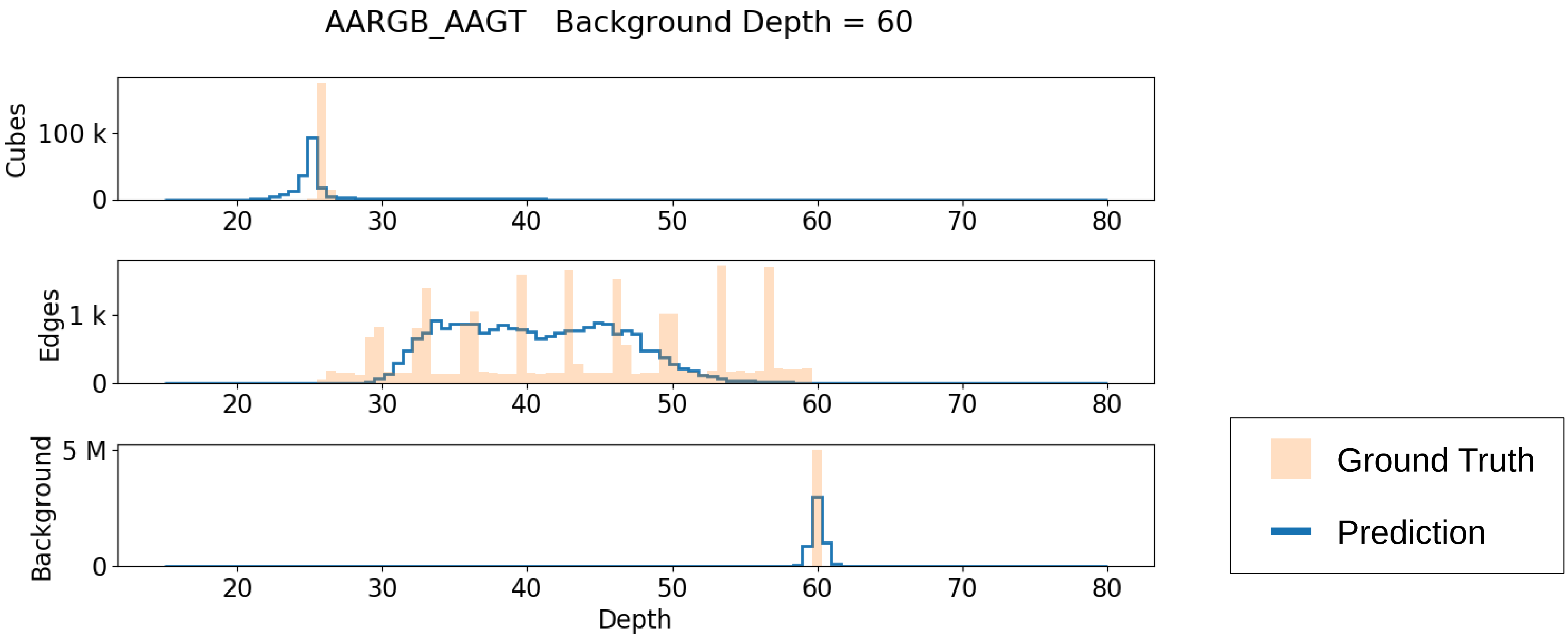
Input Color Image | Ground Truth Depth | Predicted Depth





Ex. 1: Lie in the Data

- Machine learns to match the wrong depth GT

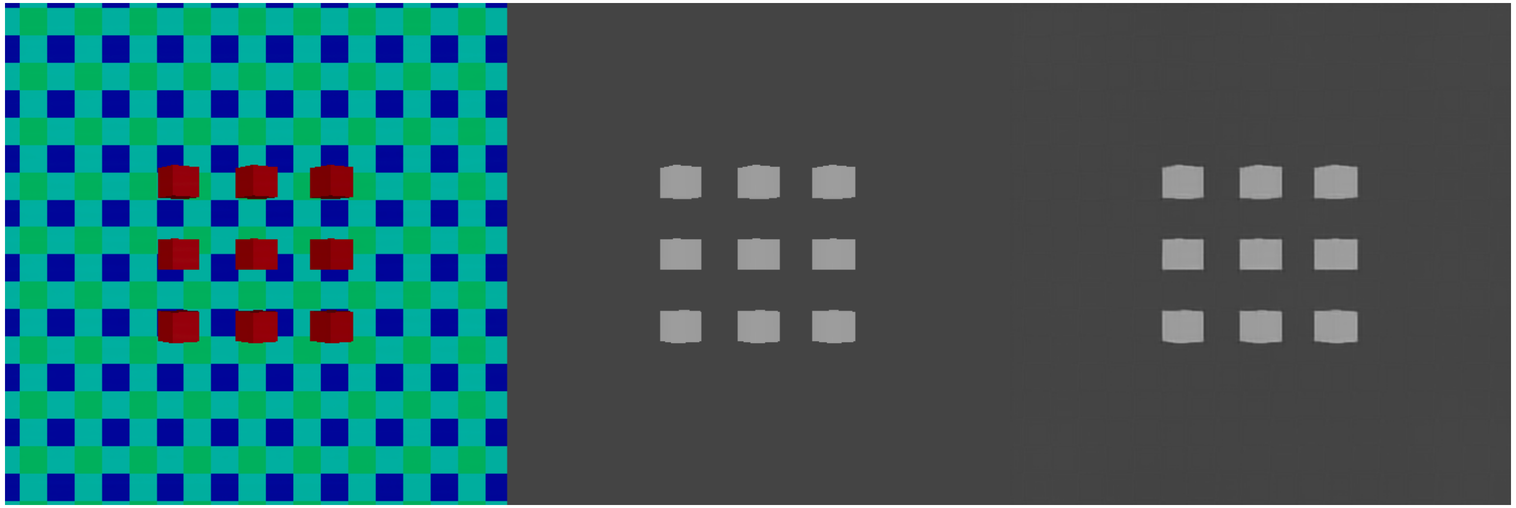


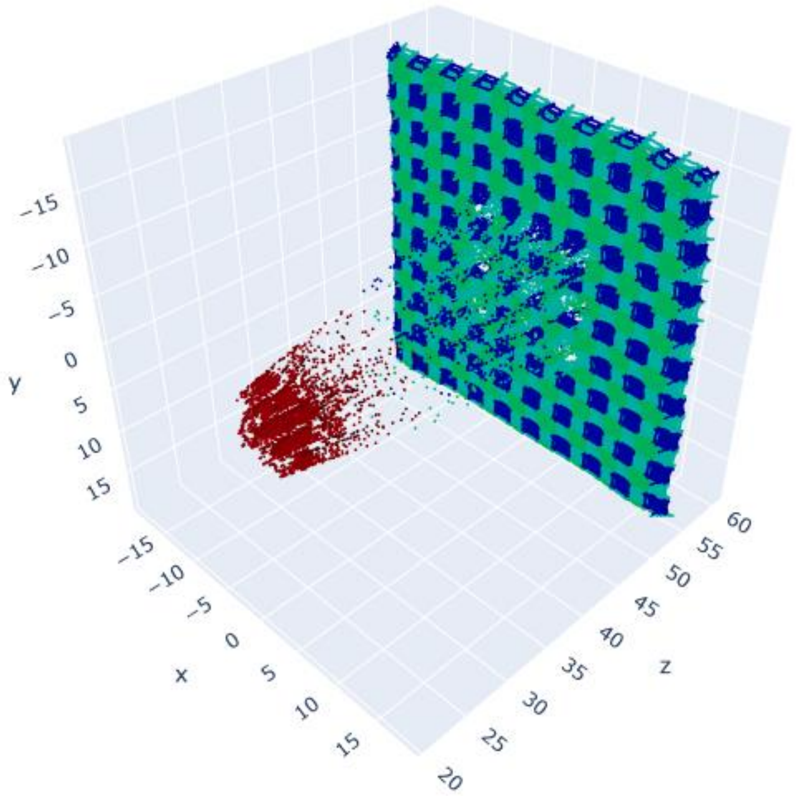
Ex. 2: Multiple True States

- **GT could be near or far**
  - Aliased color
  - Aliased depth

$$L(X,Y) = \frac{1}{N}\sum_i (\log(Y_i) - \log(X_i))^2$$

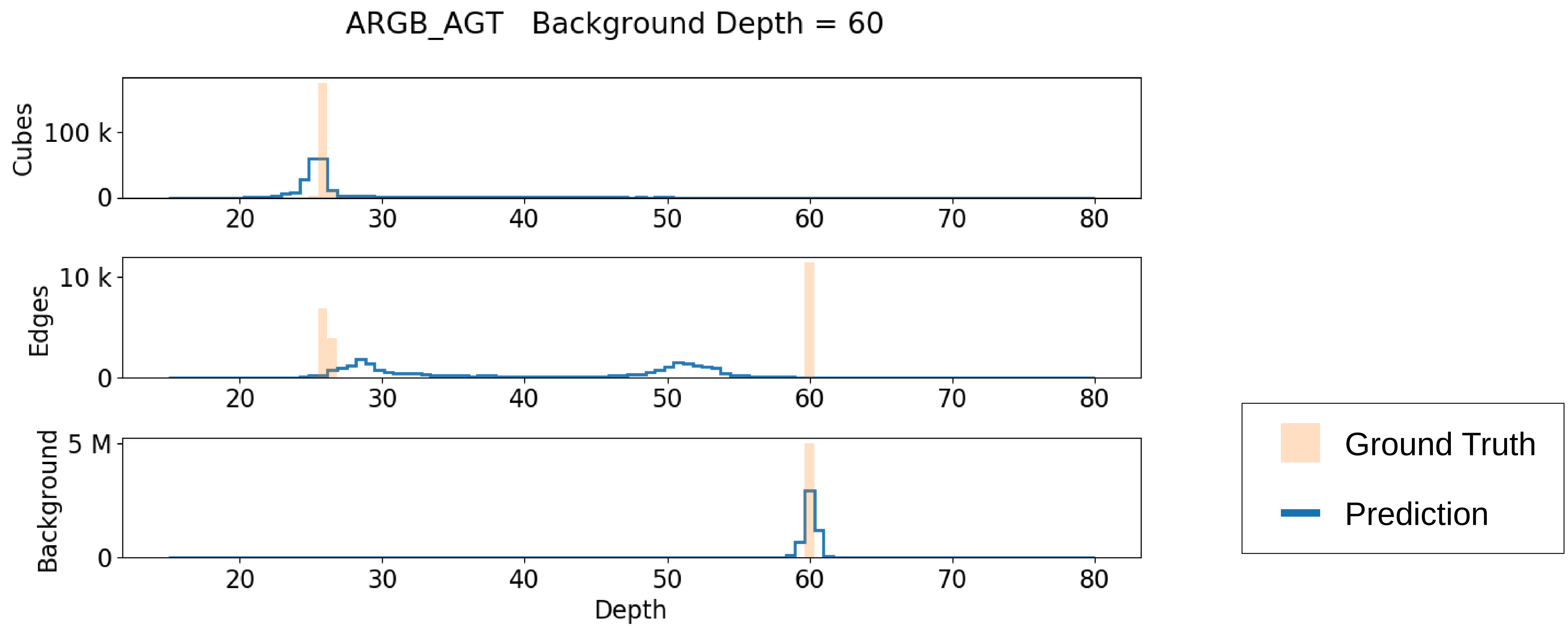
Input Color Image | Ground Truth Depth | Predicted Depth





Ex. 2: Multiple True States

- Machine picks one or the other (bimodal distribution)

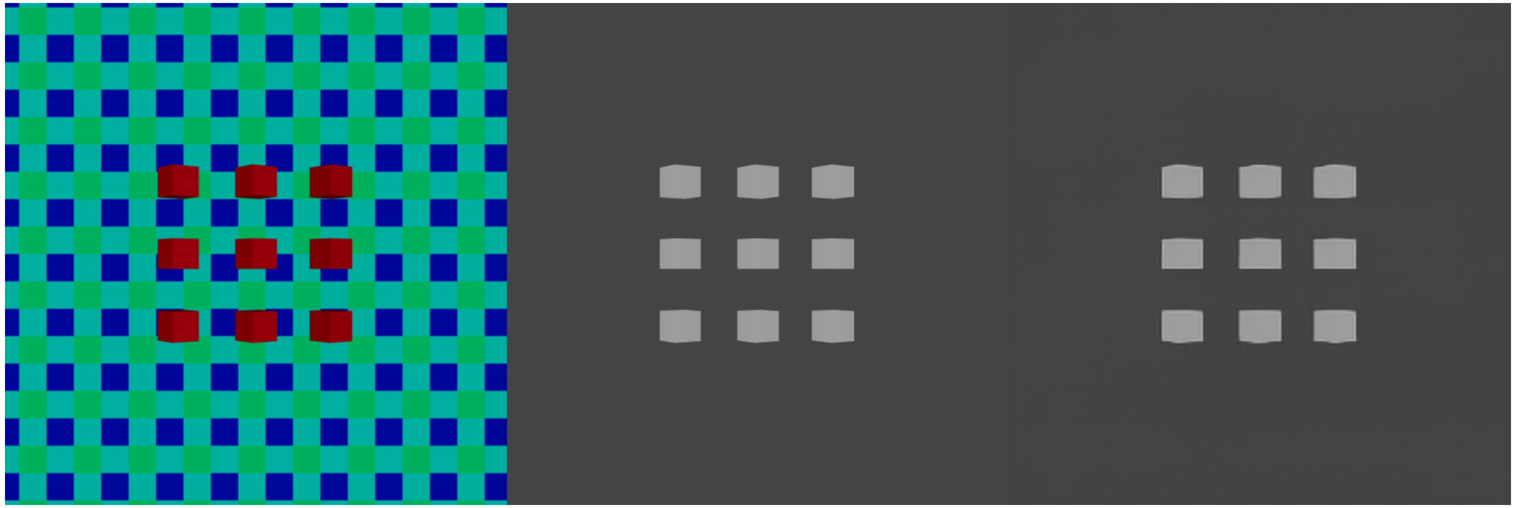


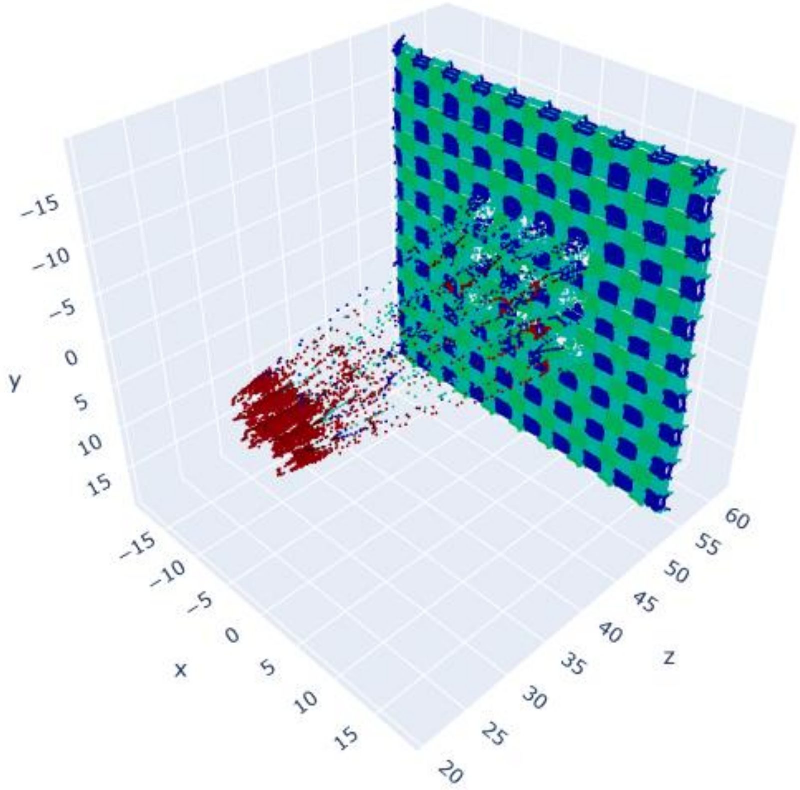
Ex. 3: You Can't Handle The Truth!

- **Many possible truths**
  - Anti-aliased color
  - Bundle depth

$$L(X,\hat{Y}) = \frac{1}{N}\sum_i \min_{y_i \in Y_i}(\log(y_i) - \log(X_i))^2$$

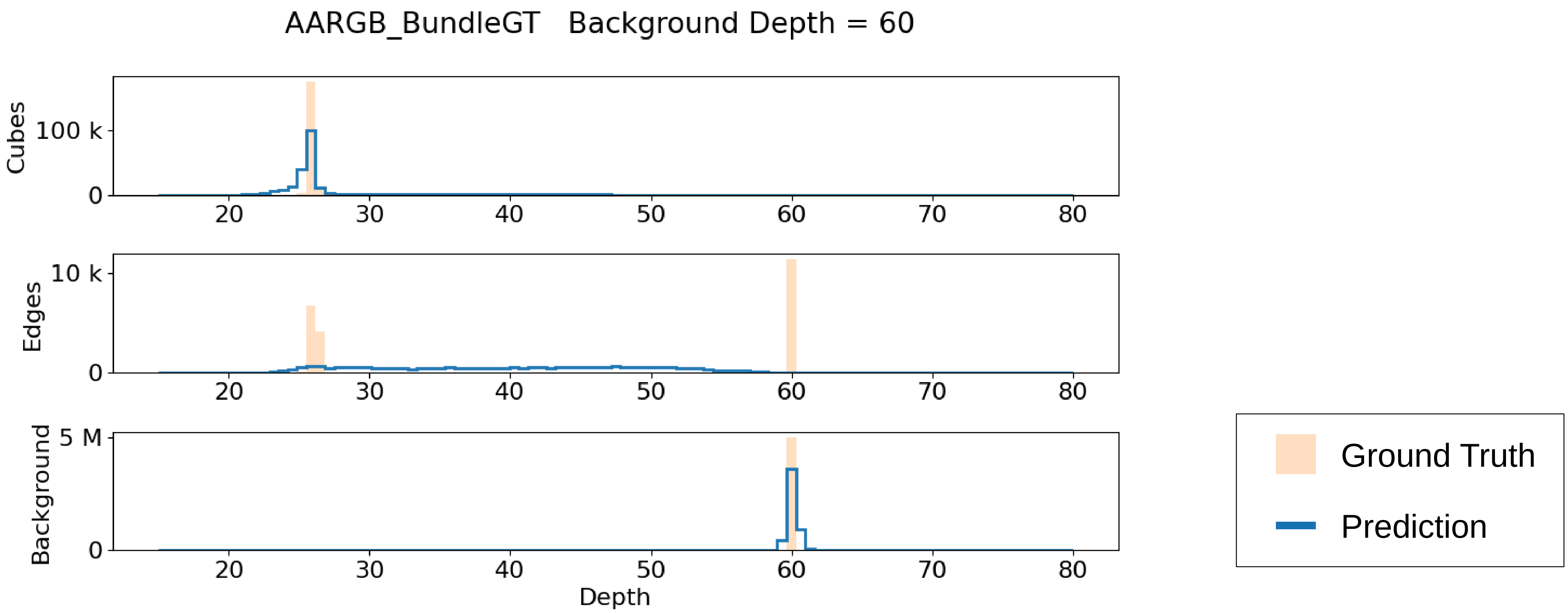
Input Color Image | Ground Truth Depth | Predicted Depth





Ex. 3: You Can't Handle The Truth!

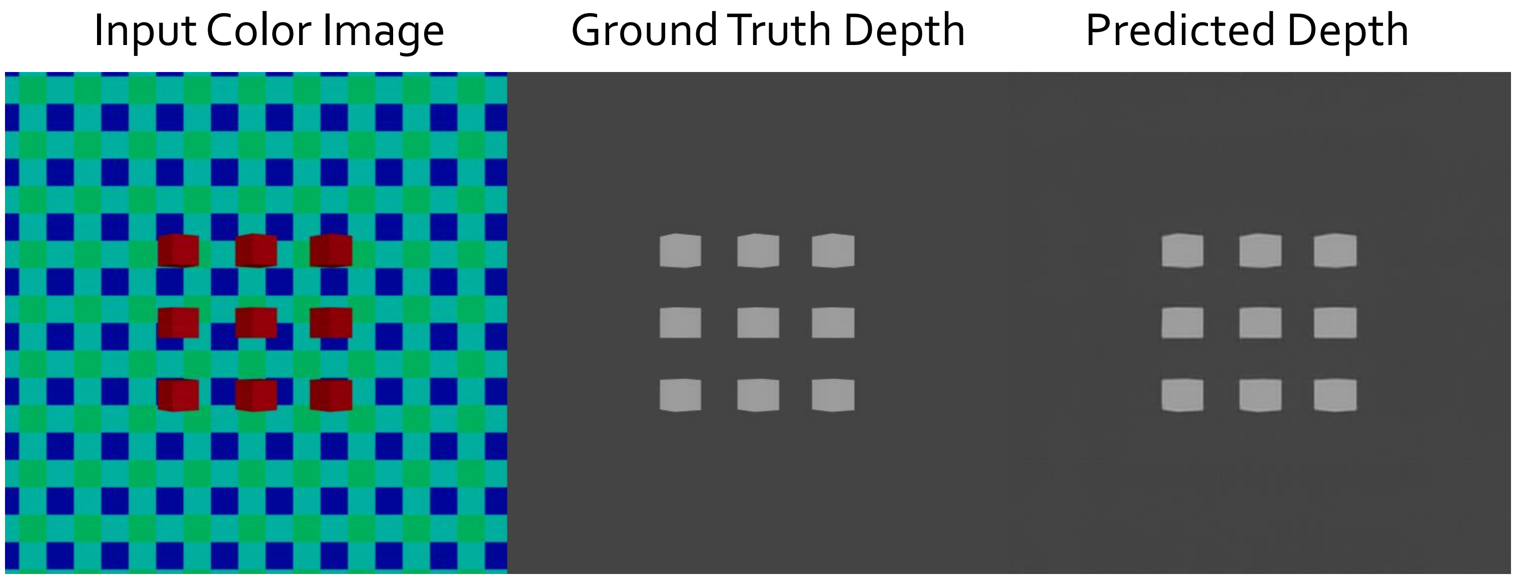
- Machine can't decide what value to pick

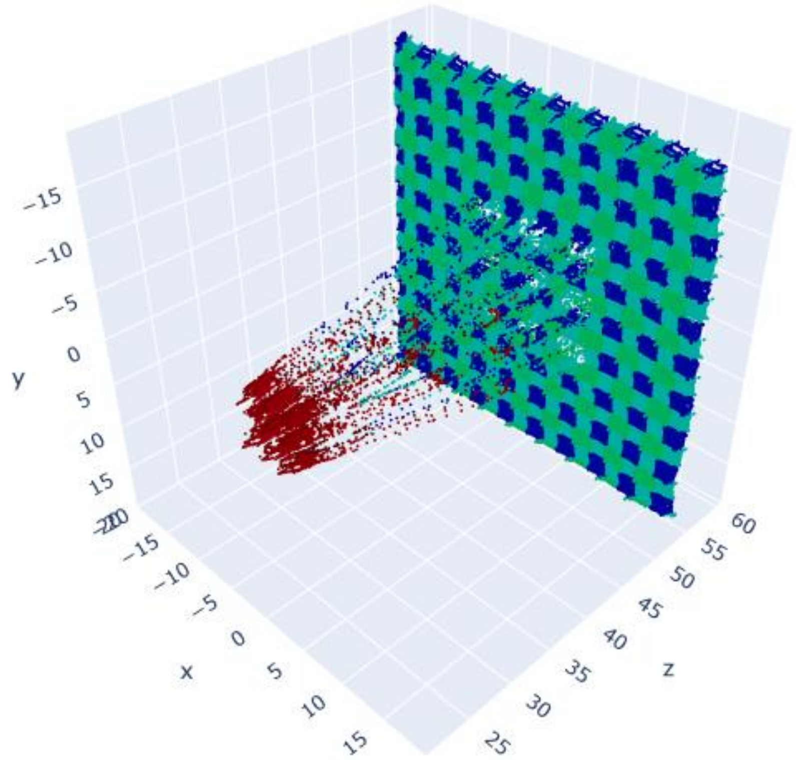


Ex. 4: Add Some Bias

$$L(X,\hat{Y}) = \frac{1}{N}\sum_i \left[\min_{y_i \in Y_i}(\log(y_i) - \log(X_i))\right]^2$$

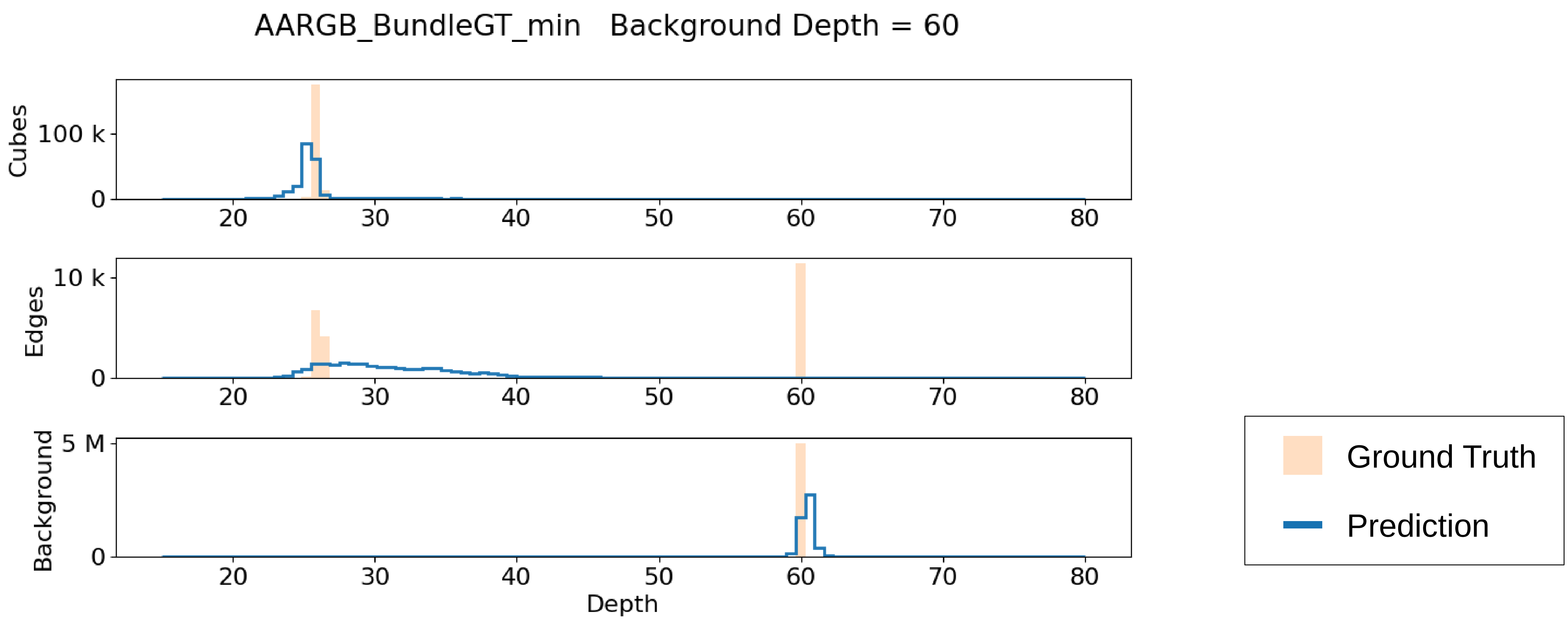
- **Not all values are equal**
  - Same as Ex. 3 but change the loss
  - Now prefers closer points





Ex. 4: Add Some Bias

- Machine now tends to learn edges as foreground



Conclusions

- **Simulated data can help train AI algorithms, but care should be taken when using as ground truth.**
  - May be better to think in terms of a “gold standard”
- **Anti-aliased depth images can cause an algorithm to learn a false average depth.**
- **Aliasing in the ground truth is also problematic.**
  - Network cannot tell if a feature should map to near or far
- **Bundled depth is one mitigation strategy.**
  - May be able to optimize in future work