# Vision (Monocular) Depth Estimation

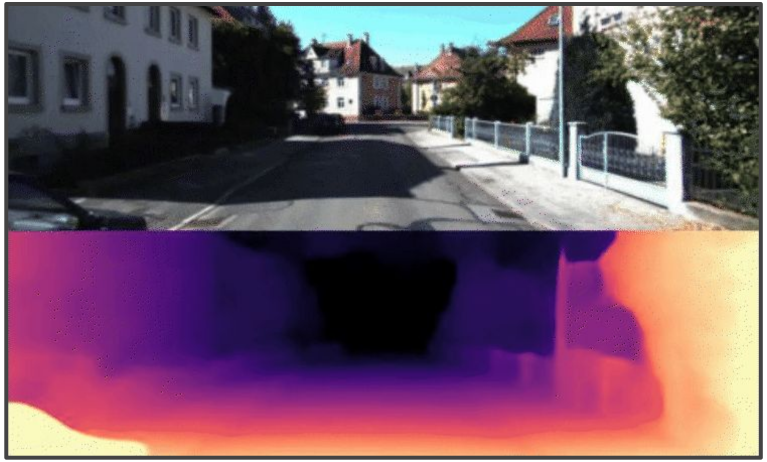
Group 31: Siyuan Zhu & Linus Grasel

# Background

**Monocular Depth Estimation:** The task of estimating scene depth using a single image

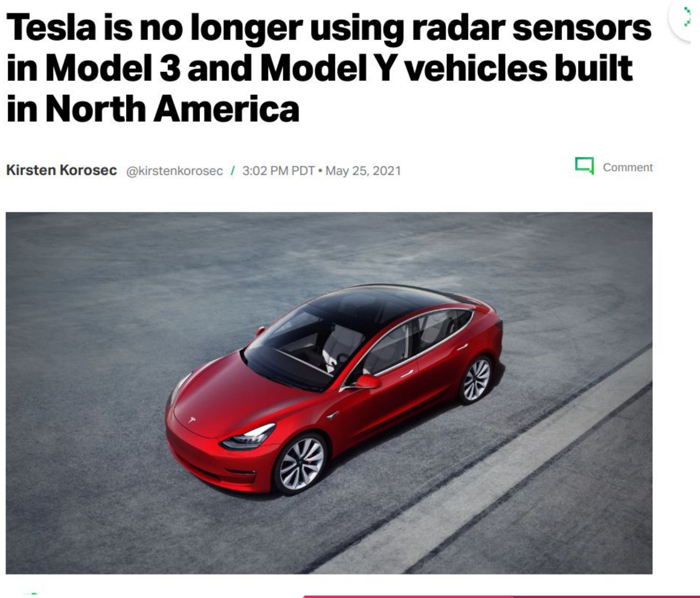
**Importance**:

- Autonomous Driving
- Robotics
- Drones
- Power Consumption Reduction



# Background

**Existing** depth estimation methods:
- LiDAR & RGB-D Cameras
*Pros*: Accuracy & Reliability
*Cons*: Energy consumption, cost and sparsity on prediction

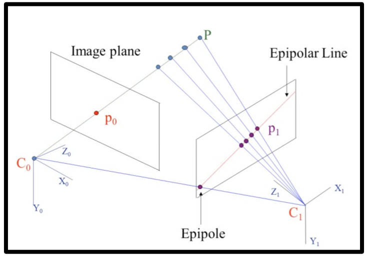

Tesla is no longer using radar sensors in Model 3 and Model Y vehicles built in North America

Kirsten Korosec @kirstenkorosec / 3:02 PM PDT • May 25, 2021

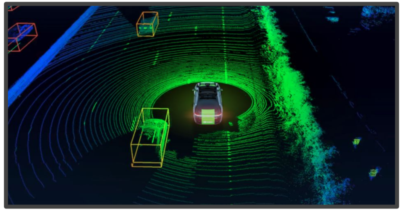
Comment

# Literature

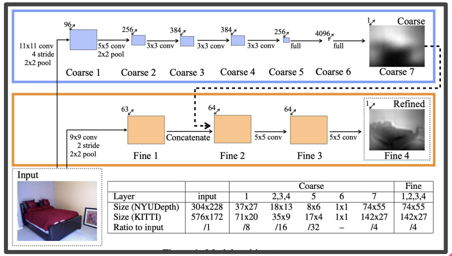
- Structure from motion(SfM) [1]  Stereo vision matching [2]
  - Feature correspondences and geometric constraints between images
  - Need calibrated camera/stereo camera setup
  - Sparse depth map



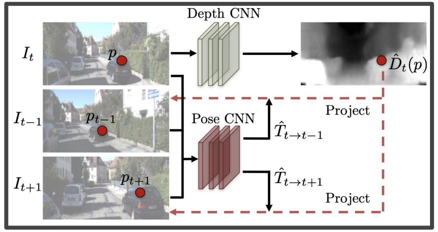

- Depth Sensors
  - Large size, high power consumption
  - RGB-D: Limited measure range, light condition sensitivity
  - LIDAR: Sparse depth map



- Depth map prediction from a single image [3]
  - Supervised methods, regression problem
  - Single Image
  - CNN
  - Dense depth map

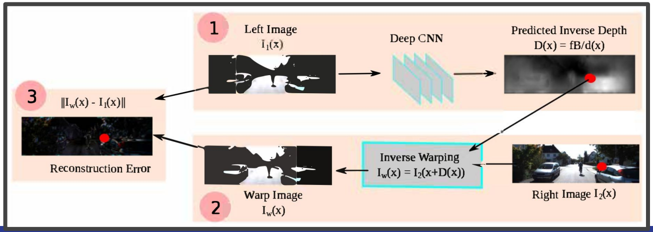


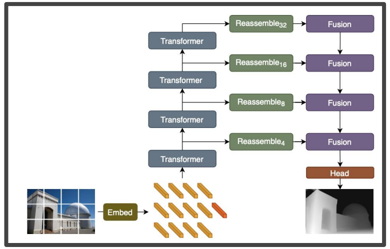

| | | Coarse | | | | | Fine |
|---|---|---|---|---|---|---|---|
| Layer | input | 1 | 2,3,4 | 5 | 6 | 7 | 1,2,3,4 |
| Size (NYUDepth) | 304x228 | 37x27 | 18x13 | 8x6 | 1x1 | 74x55 | 74x55 |
| Size (KITTI) | 576x172 | 71x20 | 35x9 | 17x4 | 1x1 | 142x27 | 142x27 |
| Ratio to input | /1 | /8 | /16 | /32 | – | /4 | /4 |




- Depth and Ego-Motion from Video [4]
  - Unsupervised method
  - Sequence of images
  - Image reconstruction
  - Dense depth map

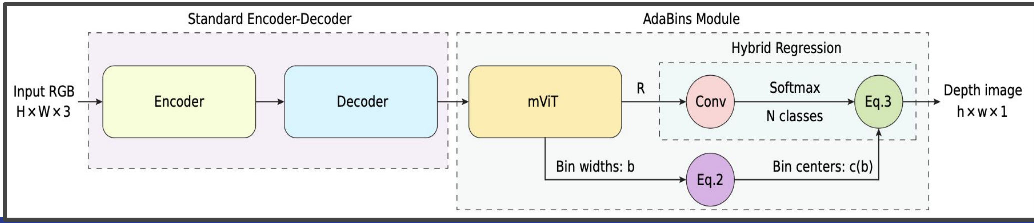
- Geometry to the Rescue [5]
  - Semi-supervised method
  - Stereo image pairs
  - Image reconstruction
  - Dense depth map



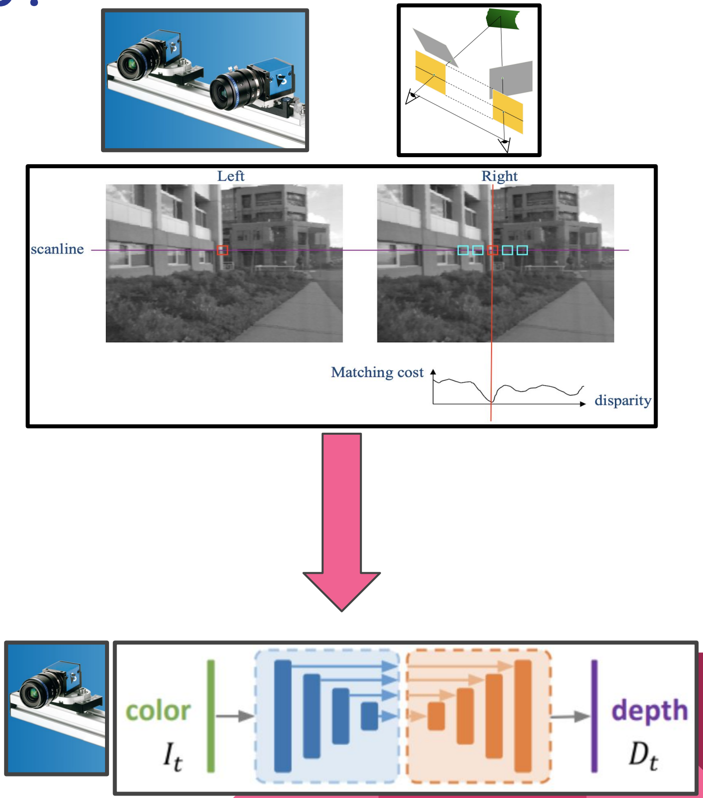

- DPT (Vision Transformer for Dense Prediction) [6]
  - Supervised learning
  - Attention mechanism
  - Strong global receptive fields
  - Dense depth map

- AdaBins (Depth Estimation using Adaptive Bins) [7]
  - Supervised learning
  - Attention mechanism: Simplified Vision transformer
  - Quantization technique
  - Modularized depth prediction structure
  - Dense depth map



# How does Machine Learning help?

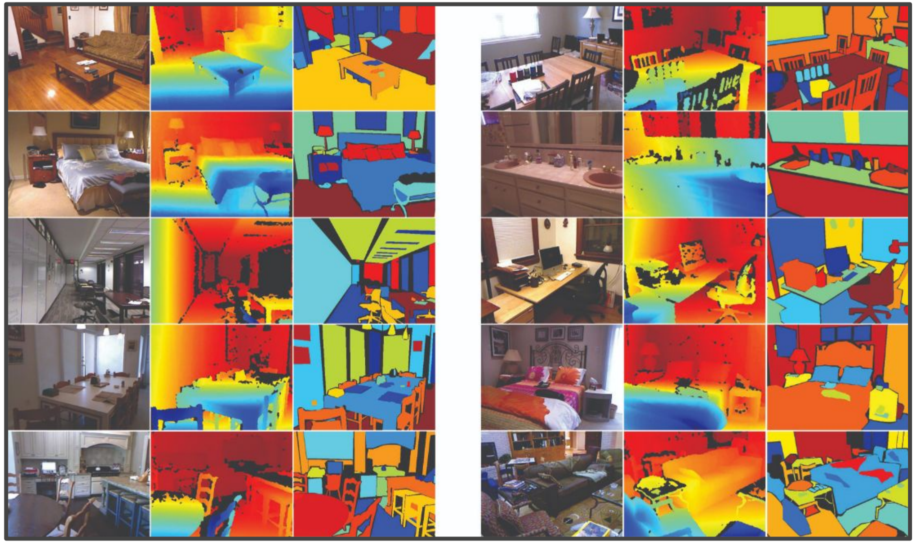
- Simpler setup
  - Smaller size
  - Low energy consumption
  - Don’t need camera intrinsic parameters
- Denser depth prediction
- End-to-end pipeline
  - Faster prediction
  - Simpler architecture
- Data-driven
  - Prior knowledge encoding
  - Fusion of various representations



# Dataset: NYU-Depth-V2

- Focuses on Indoor Environments
  - Basements, bathrooms, bedrooms, kitchens, offices etc.

- Collects ground truth depth by RGB-D camera
  - Vs. LIDAR



NYU v2 consists of:

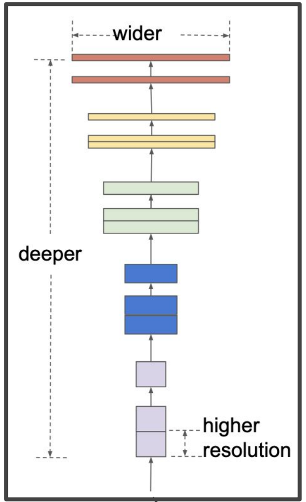
1. Labeled Dataset (Fine depth details) (2.8 GB)

2. Raw Dataset (Less fine details) (428 GB)

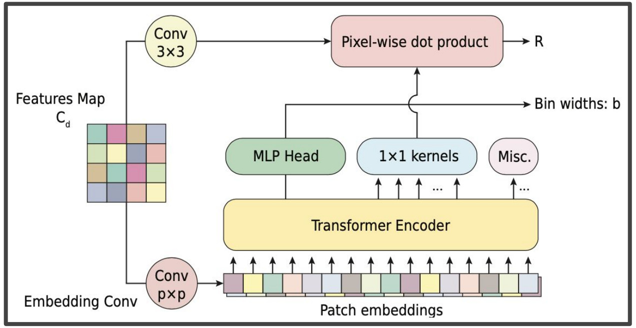
- 200x Greater than Labeled

# Feature Extraction

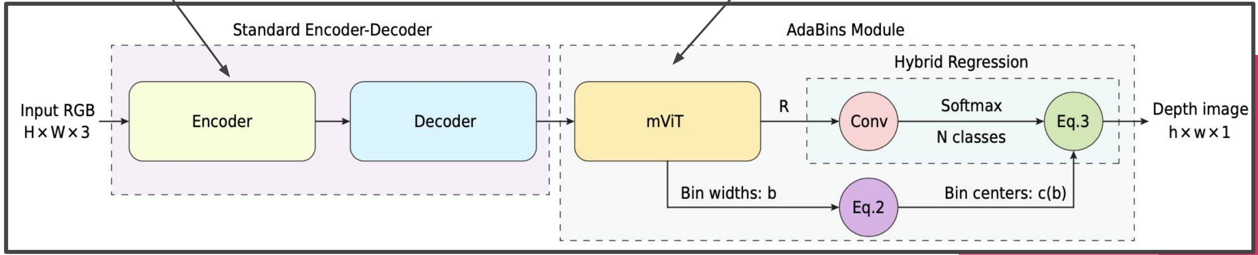
EfficientNet B5




Mini Vision Transformer

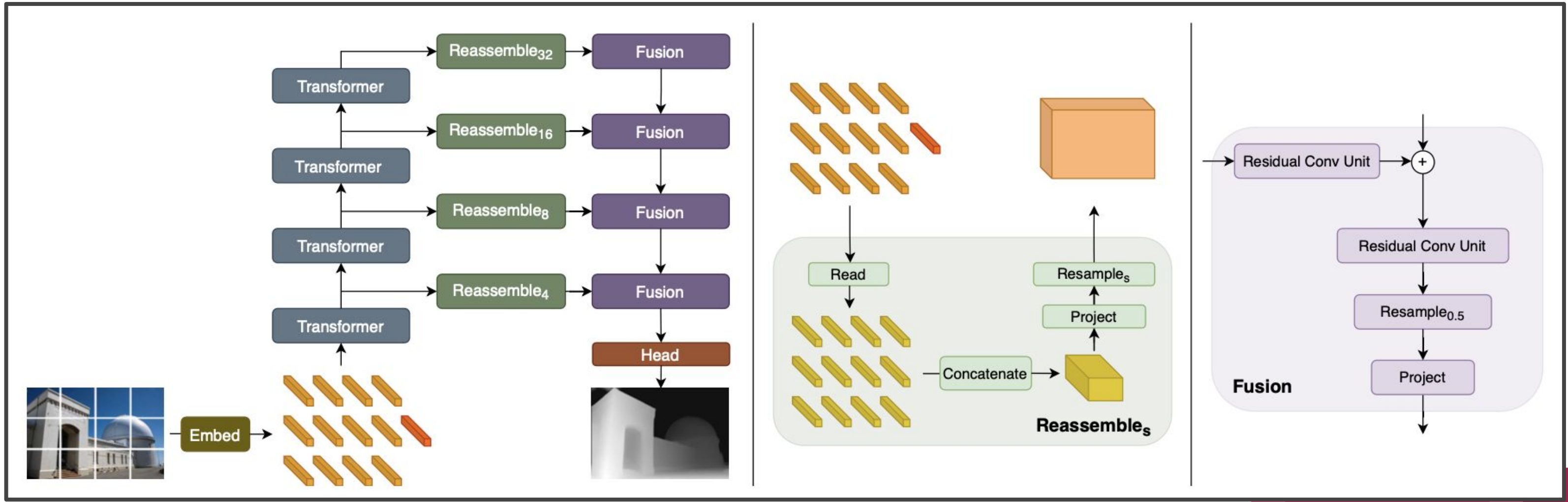



AdaBins Model



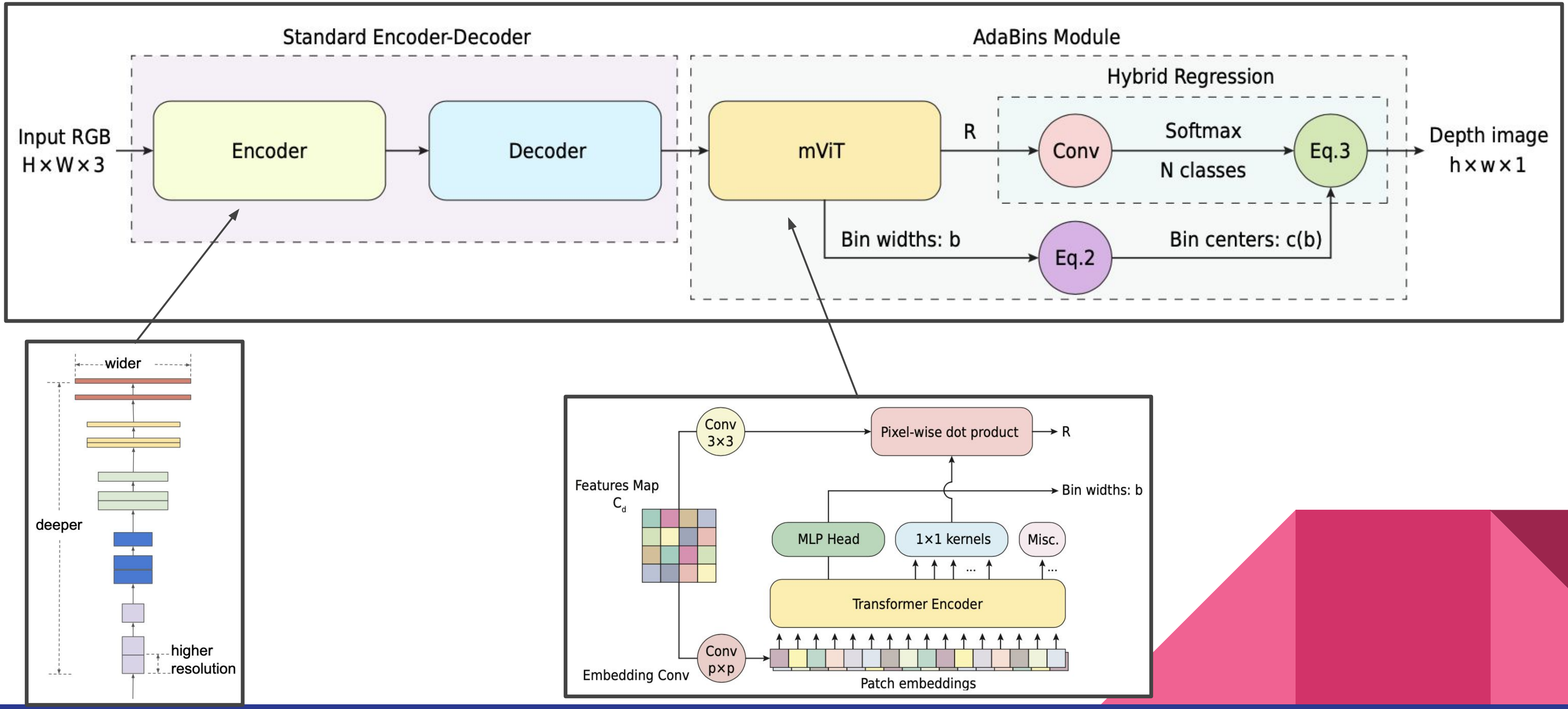
# Models (DPT Model)

- Details:



# Models (AdaBins)

- Details:



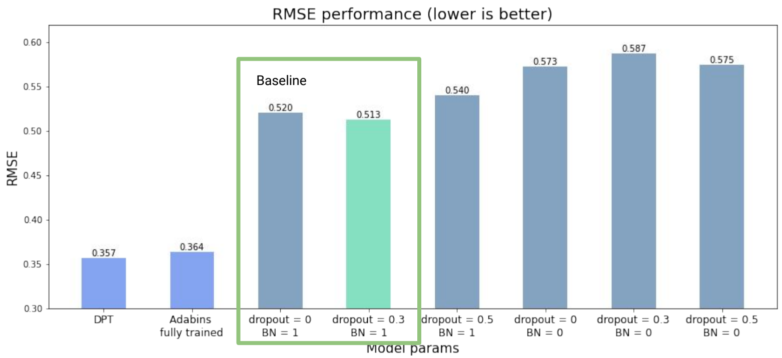
# Models

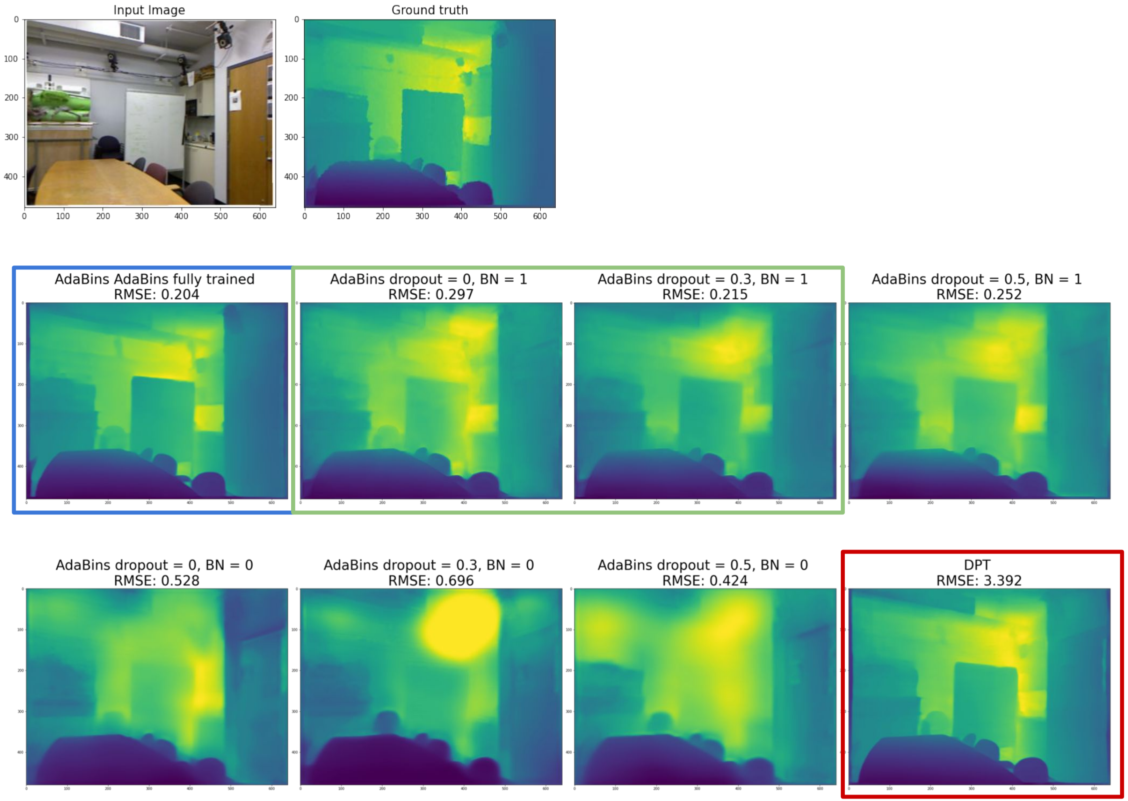
- Model tweaking at 4 upsampling decoding layers

| | #0 (baseline) | #1 | #2 | #3 | #4 | #5 |
|---|---|---|---|---|---|---|
| Drop Out Rate | 0.0 | 0.3 | 0.5 | 0.0 | 0.3 | 0.5 |
| Batch Normalization | True | True | True | False | False | False |

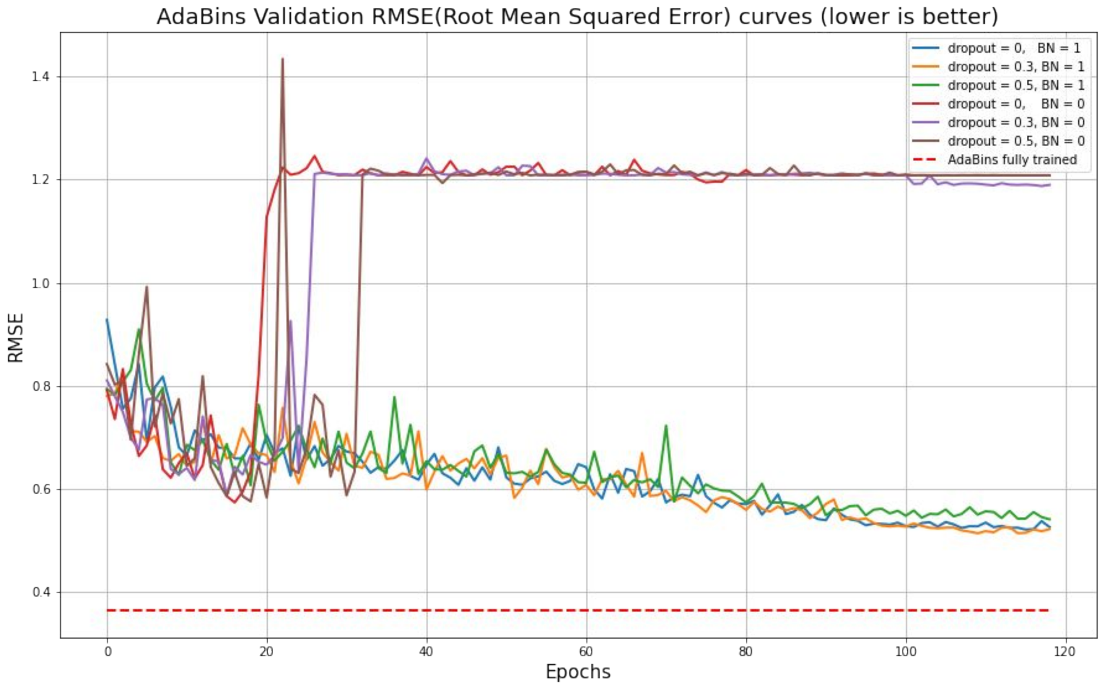
- Compared vs. AdaBins fully trained & DPT

# Results/Observations

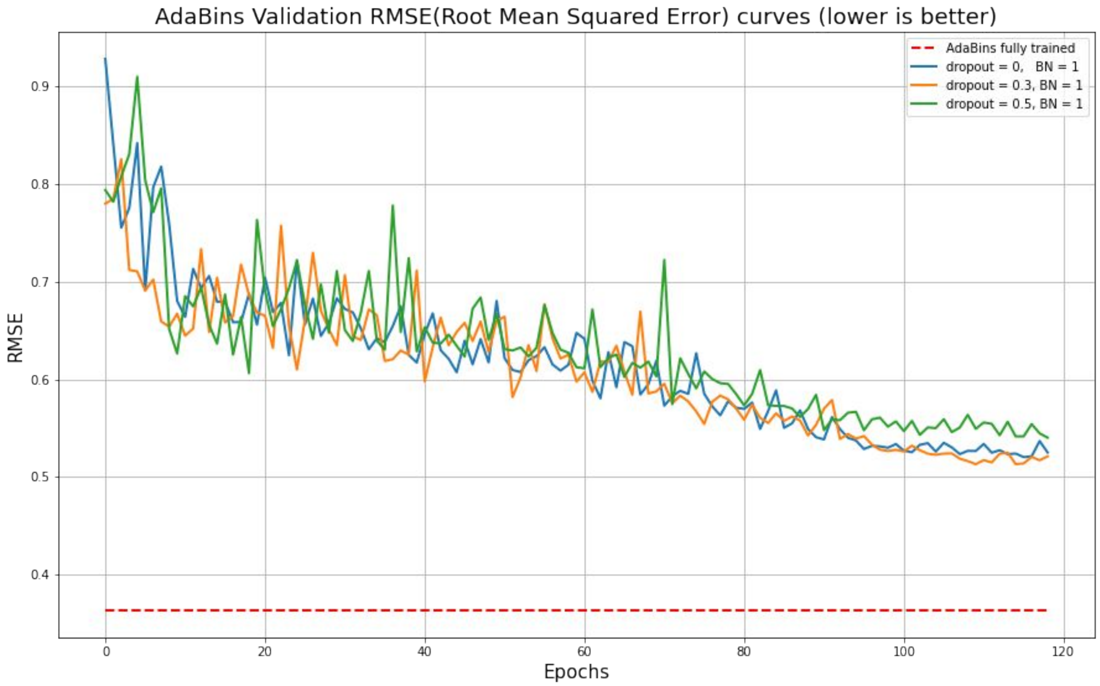


# Results/Observations



# Results/Observations



# Next Steps

- Train on more data
- Train on more concentrated data (i.e. Living Rooms only)

# References


[1] Rene Ranftl, Alexey Bochkovskiy and Vladlen Koltun. Vision Transformers for Dense Prediction. [cs.CV] 24 Mar 2021.

[2] S. Ullman, “The interpretation of structure from motion,” Proceedings of the Royal Society of London. Series B. Biological Sciences, vol. 203, no. 1153, pp. 405–426, 1979.

[3] “Cnn-slam, keisuke and tombari, federico and laina, iro and navab, nassir,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6243–6252.

[4] J. M. Facil, B. Ummenhofer, H. Zhou, L. Montesano, T. Brox, and J. Civera, “Cam-convs: camera-aware multi-scale convolutions for single-view depth,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 11 826–11 835.

[5] R. Wang, S. M. Pizer, and J.-M. Frahm, “Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5555–5564.

[6] P. Chakravarty, P. Narayanan, and T. Roussel, “Gen-slam: Generative modeling for monocular simultaneous localization and mapping,” in 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 147–153.

[7] C.Godard,O.MacAodha, and G.J.Brostow, “Unsupervised Monocular depth estimation with left-right consistency,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 270–279.

[8] Chaoqiang Zhao, Qiyu Sun, Chongzhen Zhang, Yang Tang and Feng Qian. Monocular Depth Estimation Based On Deep Learning: An Overview. [cs.CV] 3 Jul 2020.

[9] OpenAI, GPT3, Retrieved from {\it https://openai.com/blog/gpt-3-apps/}

[10] OpenAI, Dall-E, Retrieved from {\it https://openai.com/blog/dall-e/}

[11] NYU Depth Dataset V2, Indoor Segmentation and Support Inference from RGBD Images, Retrieved from {\it https://cs.nyu.edu/~silberman/datasets/nyu_depth_v2.html}

[12] Bhat, S. F., Alhashim, I., and Wonka, P. Adabins: Depth estimation using adaptive bins. arXiv:2011.14141, 2020.