Neuron Selectivity for Efficient Monocular Depth Estimation

Federico Siciliani

Computer Vision | a.y 2024/2025

Outline

1

Problem Statement

The "Black Box" problem

2

State of the Art

MDE models and explainability

3

Proposed Method

Neuron Selectivity

4

Dataset

NYU Depth V2

5

Experimental Setup

Framework and training parameters

6

Results

Model Evaluation and Trade-off Analysis

7

Conclusion

Key findings and future work

Problem Statement

Monocular Depth Estimation: used in technologies like autonomous driving, robotics, and augmented reality

The "Black Box" Problem: Deep learning models for MDE are highly accurate, but their decision-making process is not clear

Why Explainability is important:

- **Trust:** Essential for safety-critical applications
- **Debugging:** Helps identify models errors

The idea: Is it possible to make lightweight MDE models explainable without a significant loss in performance?

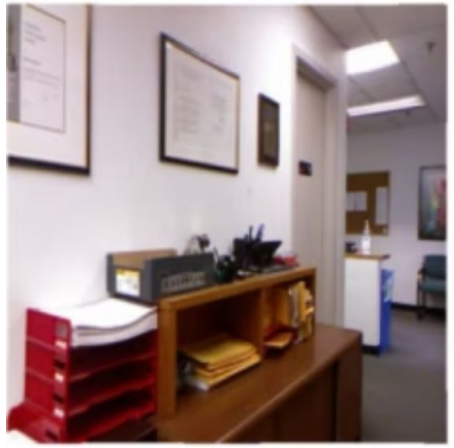


Image RGB

Model

Depth Estimation

State of the Art: MDE & Explainability

Heavyweight Models

Architectures like ResNet-101 provide high accuracy but are computationally expensive

Lightweight Models

Architectures like MobileNet are designed for efficiency on mobile devices

Explainability

Neuron Selectivity: specific neurons are forced to become "specialists" for pre-defined concepts. In MDE, this means training neurons to activate for specific depth ranges

Proposed Method: Integrating Selectivity

Objective: Apply and evaluate the neuron selectivity strategy on a lightweight MDE model.

Chosen Architecture: MobileNetSkipAdd, from fastdepth library

- **Baseline Model Training:**
  - Trained with a standard depth loss to establish a performance benchmark: L_depth = L1Loss(pred_depth, gt_depth)

Selectivity Approach: Key Steps

1

Choosing the layer

Select a late stage decoder layer (decode_conv5), with 32 output units, close to the final output, where the spatial resolution is higher

2

Logarithmic Depth Binning

Discretize continuous depth range (0-10m) into 27 logarithmic bins for more precise resolution in the near depths, and to create a more balanced distribution of pixels across the bins

3

Valid Bin Assignment

Assign pixels to their corresponding depth bins, using a validity mask to ignore pixels without ground truth depth, in this way 16 valid bins are obtained

4

Combined Loss Function

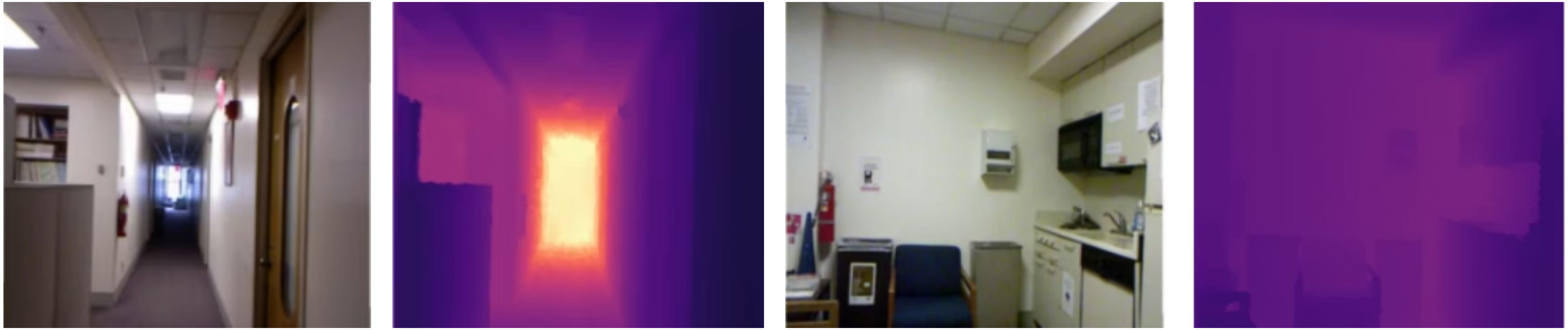
Train the model with a weighted sum of depth loss and the new selectivity loss (L_total = L_depth + λ * L_sel), The selectivity loss (L_sel) is designed to maximize a **Depth Selectivity Score** . This score compares a neuron's activation on its target bin against its average activation on all other bins

Dataset: NYU Depth V2

Indoor MDE: the most used benchmark for this task

Content: Over 50.000 RGB images and their corresponding depth maps

- **Preprocessing:**
  - Images resized to 224x224, to match the input size of MobileNetSkipAdd
  - Standard data augmentations (color jitter)



Examples of images in NYU Depth V2 and their depth maps

Experimental Setup

Framework & Hardware

- **Framework**: PyTorch
- **Hardware**: GPU NVIDIA RTX 4060

Training Details

- **Optimizer**: AdamW
- **Learning Rate**: 1e-3
- **Batch size**: 16
- **Training epochs**: 20

Selectivity Parameters

- **Number of Depth Bins**: 16 valid
- **Selectivity weight (λ)**: 0.1

Evaluation Metrics

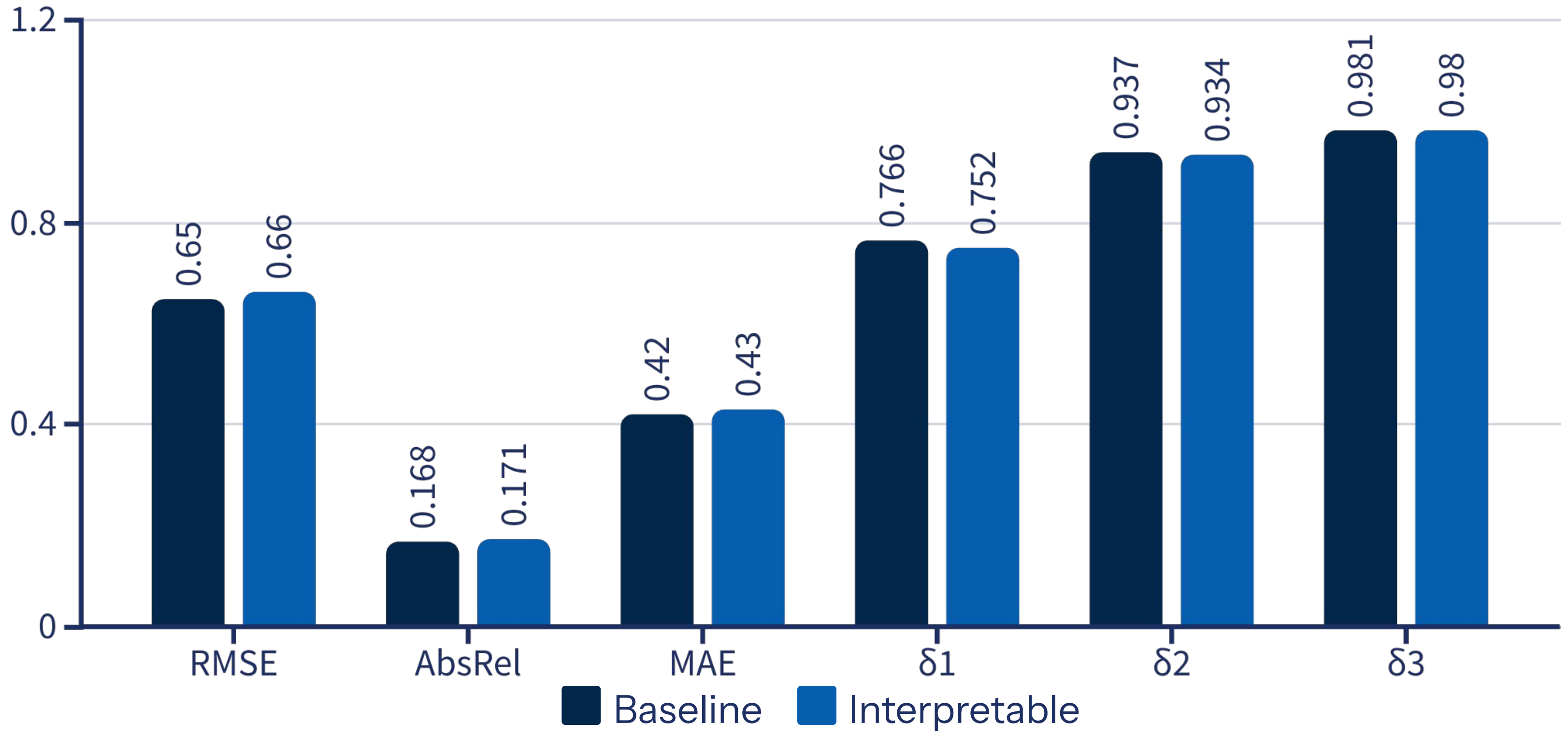
- **Performance (Depth Error)**: RMSE, AbsRel , MAE, δ1, δ2, δ3
- **Interpretability (Selectivity)**: Depth Score, Target Depth Score, Target Accuracy

Results: Depth Performance

The selectivity model maintains competitive performance with only a little and acceptable drop in the accuracy compared to the baseline

Error metrics: RMSE, AbsRel, MAE

Accuracy metrics: δ1, δ2, δ3



Depth performance for both models

Qualitative Results: Visual Comparison

Depth Prediction

The predicted depth maps from the interpretable model are visually almost identical to the baseline, indicating high quality depth predictions are maintained

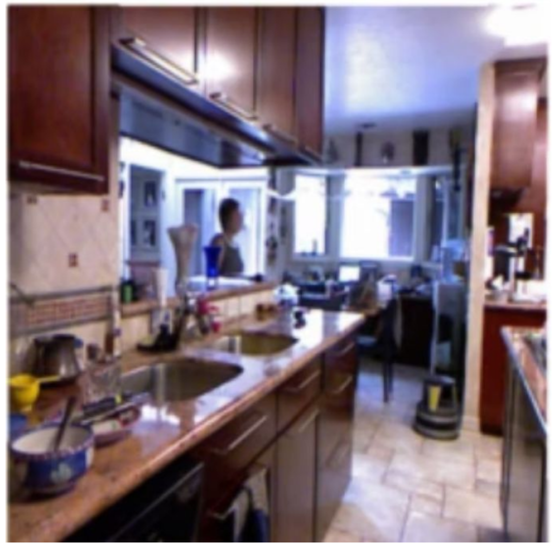
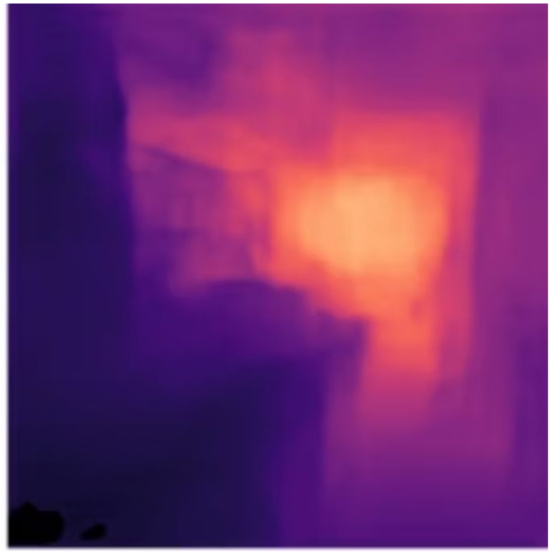


Image RGB

Ground truth

Baseline model prediction



Interpretable model prediction

Results: Selectivity Performance

Overall Selectivity: Measures how a neuron prefers a single depth bin.

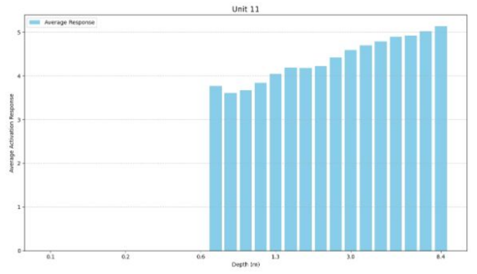
Targeted Selectivity: measures how strongly a neuron prefers its assigned depth bin.

The baseline's **negative score (-0.21)** proves its neurons activate randomly with respect to depth

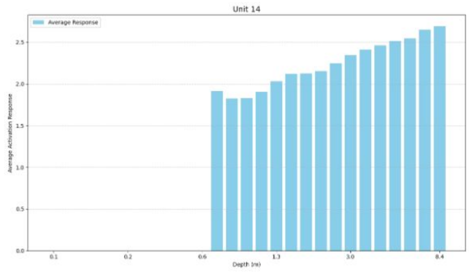
Interpretable model's score of **+0.78** is a huge improvement, confirming that neurons are now correctly specialized.

Target Accuracy: The percentage of neurons whose strongest activation is on target

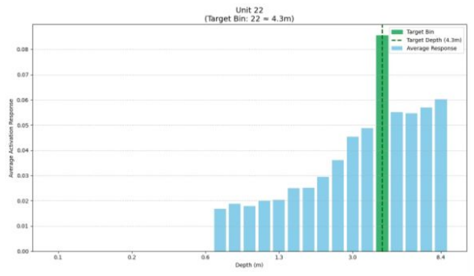
Model	DS score	Ds score target	Peak on target
Baseline	0.71	-0.21	0.06
Interpretable	**0.87**	**0.78**	**0.59**



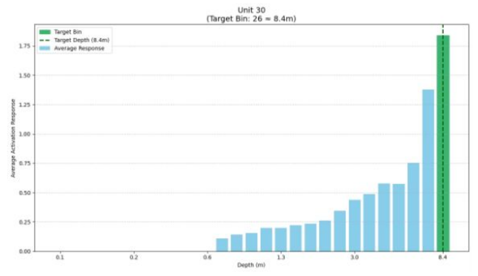
Unit 11 response, baseline model



Unit 14 response, baseline model



Unit 22 response, interpretable model



Unit 30 response, interpretable model

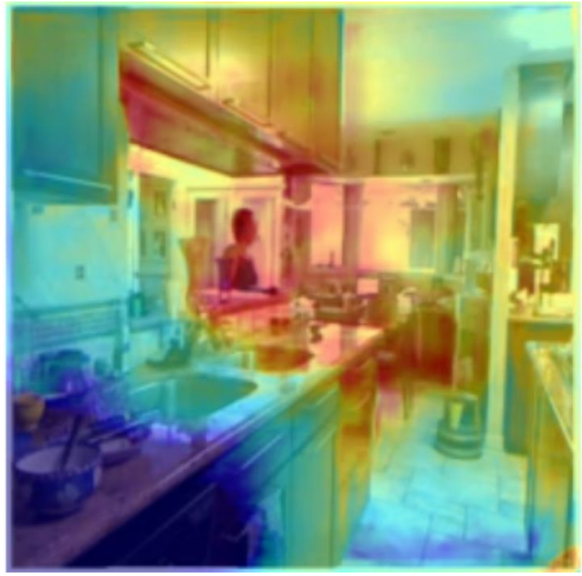
Qualitative Results: Visual Comparison

Activation maps

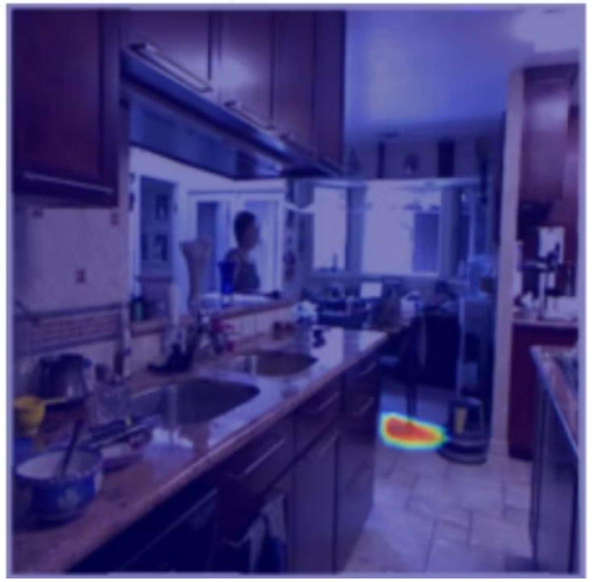
Baseline Model: Neuron activations lack a clear correlation with any specific depth, making them impossible to interpret

Interpretable Model: Neuron activations are sparse, localized, and directly correspond to their assigned depth ranges.

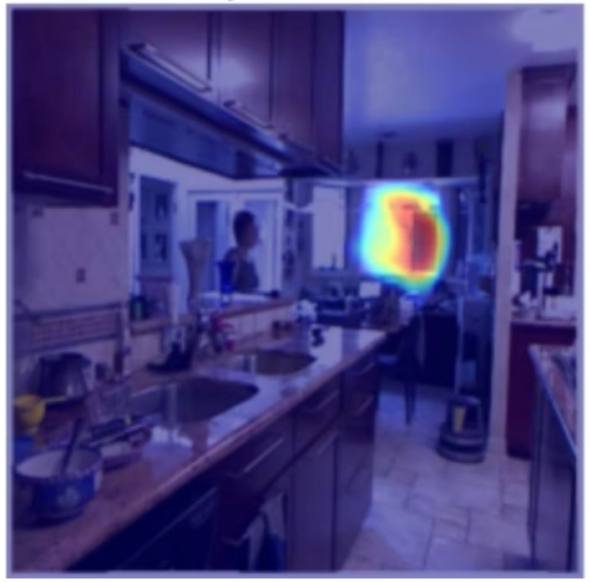
Unit 11, activation map, baseline model



Unit 14 activation map, baseline model



Unit 22 activation map, interpretable model



Unit 30 activation map, interpretable model

Conclusion

Key Findings

Success of the Method: This project confirms that neuron selectivity can be applied successfully to lightweight MDE models

Trade-Off: an increase in model's interpretability was achieved with only an low decrease in depth estimation accuracy

Limitations

Specialization: While significantly improved, neuron specialization is not perfect. Some neurons were suppressed (became inactive) during training, and others showed activations outside their target depth bin.

Future Work

Application to other datasets: Apply and evaluate this technique on outdoor datasets

Optimal binning strategies: Develop a method to assigning bins in a optimal way, rather assigning them manually

References

[1] You, Z., Tsai, Y.-H., Chiu, W.-C., and Li, G. (2021). Towards Interpretable Deep Networks for Monocular Depth Estimation. arXiv.

[2] C. Schiavella, L. Cirillo, L. Papa, P. Russo, and I. Amerini, (2023). Optimize vision transformer architecture via efficient attention modules: a study on the monocular depth estimation task. In: International Conference on Image Analysis and Processing, Cham: Springer Nature Switzerland, pp. 383-394.

[3] Papa, L., Russo, P., and Amerini, I. (2023). METER: A Mobile Vision Transformer Architecture for Monocular Depth Estimation. IEEE Transactions on Circuits and Systems for Video Technology, 33(10), 5882–5893

[4] D. Wofk, F. Ma, T.-J. Yang, S. Karaman, and V. Sze, "Fast-Depth: Fast Monocular Depth Estimation on Embedded Systems