High Quality Monocular Depth Estimation via Transfer Learning

Constantinescu Maria-Ecaterina
Enache George-Vlad
Ghiorghiu Bianca-Alexandra

Objectives

- Replicate steps from "High Quality Monocular Depth Estimation via Transfer Learning" using DenseNet169 for building a depth estimation model.
- Implement an alternative depth estimation model using DenseNet121.
- Train both models on the same dataset for a fair comparison.
- Compare qualitative results using visualizations of depth maps and input images.
- Evaluate quantitative performance using metrics like root mean squared error, average relative error, average logarithmic error, and threshold accuracy.
- Analyze and draw conclusions on the effectiveness of DenseNet121 vs. DenseNet169 for monocular depth estimation via transfer learning.

Dataset

The **KITTI** dataset is a widely used computer vision dataset that provides images and sensor data collected from a car driving in urban environments.

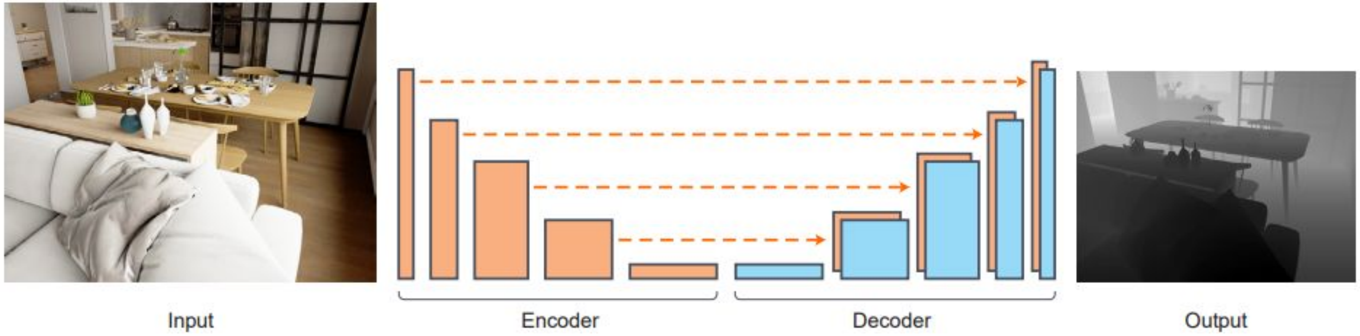
The dataset includes over 120k high-resolution images, lidar point clouds, and calibrated camera poses, and is primarily used for tasks such as object detection, object tracking, and depth estimation.

The **NYU Depth v2** dataset is a popular computer vision dataset that provides RGB-D images captured from a Microsoft Kinect camera. The dataset includes over 1449 densely labeled pairs of aligned RGB and depth images, along with segmentation masks for 27 object classes.

The dataset is a benchmark dataset for depth estimation research and has been used to evaluate many state-of-the-art depth estimation models.

Encoder-Decoder Model

- Encoder: pre-trained truncated DenseNet-121/DenseNet-169
- Decoder: basic blocks of convolutional layers applied on the concatenation of the 2× bilinear upsampling of the previous block with the block in the encoder with the same spatial size after upsampling
- Skip connections



Implementation Details

Original Model

- Encoder: Pre-trained truncated DenseNet169
- Decoder: Weights randomly initialized
- Batch size: 8
- Optimizer: ADAM with learning rate 0.0001 and β1 = 0.9, β2 = 0.999
- Epochs: 20

Our Model

- Encoder: Pre-trained truncated DenseNet121 and DenseNet169
- Decoder: Weights randomly initialized
- Batch size: 2 for DenseNet121 and 8 for DenseNet 169
- Optimizer: ADAM with learning rate 0.0001 and β1 = 0.9, β2 = 0.999
- Epochs: 5

Loss Function

$$L(y,\hat{y}) = \lambda L_{depth}(y,\hat{y}) + L_{grad}(y,\hat{y}) + L_{SSIM}(y,\hat{y})$$

$$L_{depth}(y,\hat{y}) = \frac{1}{n}\sum_{p}^{n}|y_p - \hat{y}_p|$$

$$L_{grad}(y,\hat{y}) = \frac{1}{n}\sum_{p}^{n}|\boldsymbol{g}_{\mathbf{x}}(y_p,\hat{y}_p)| + |\boldsymbol{g}_{\mathbf{y}}(y_p,\hat{y}_p)|$$

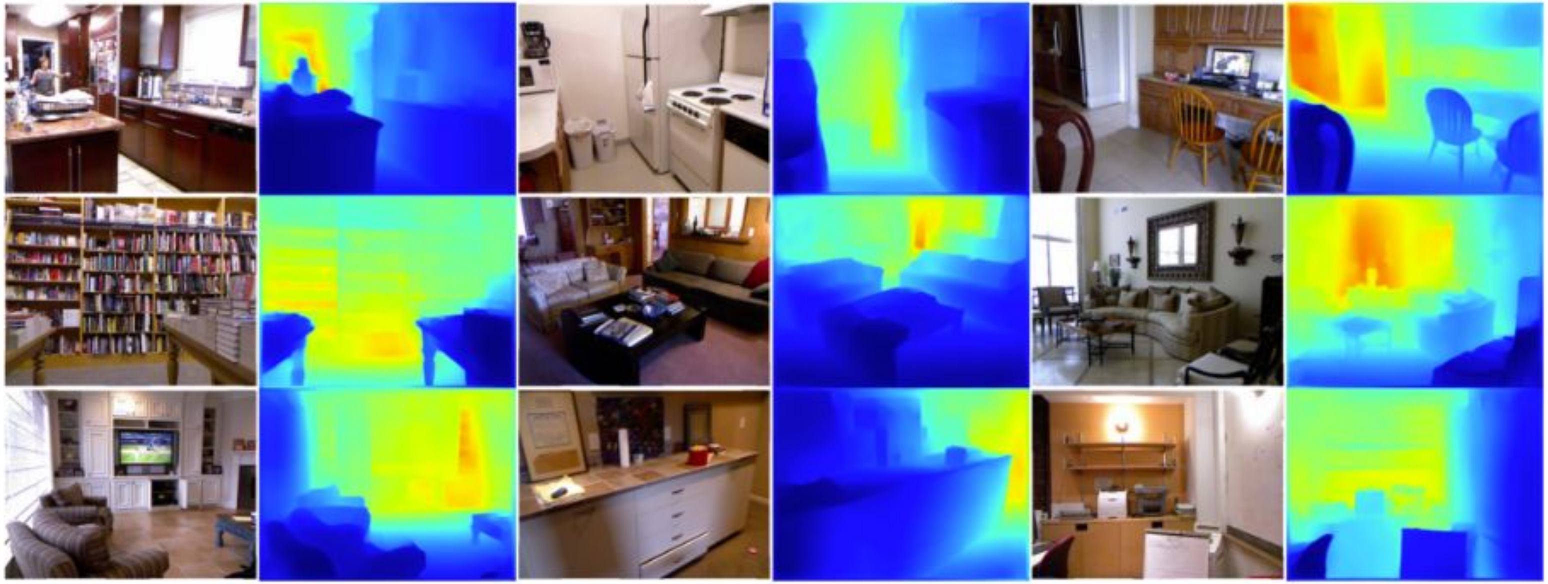
$$L_{SSIM}(y,\hat{y}) = \frac{1 - SSIM(y,\hat{y})}{2}$$

Evaluation

- average relative error (rel): $\frac{1}{n}\sum_{p}^{n}\frac{|y_p-\hat{y}_p|}{y}$;
- root mean squared error (rms): $\sqrt{\frac{1}{n}\sum_{p}^{n}(y_p-\hat{y}_p)^2)}$;
- average ($\log_{10}$) error: $\frac{1}{n}\sum_{p}^{n}|\log_{10}(y_p)-\log_{10}(\hat{y}_p)|$;
- threshold accuracy (δ_i): % of y_p s.t. $\max(\frac{y_p}{\hat{y}_p},\frac{\hat{y}_p}{y_p})=\delta<thr$ for $thr=1.25,1.25^2,1.25^3$;

Qualitative Evaluation - DenseNet121

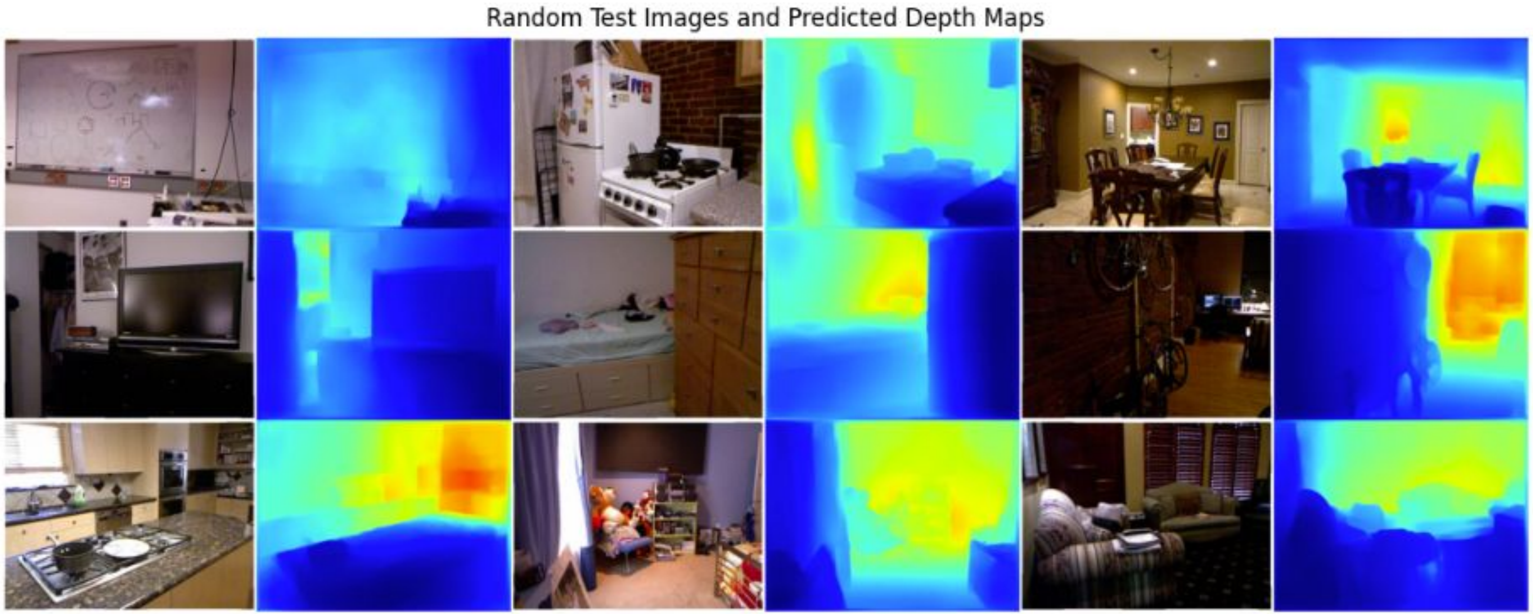
Random Test Images and Predicted Depth Maps



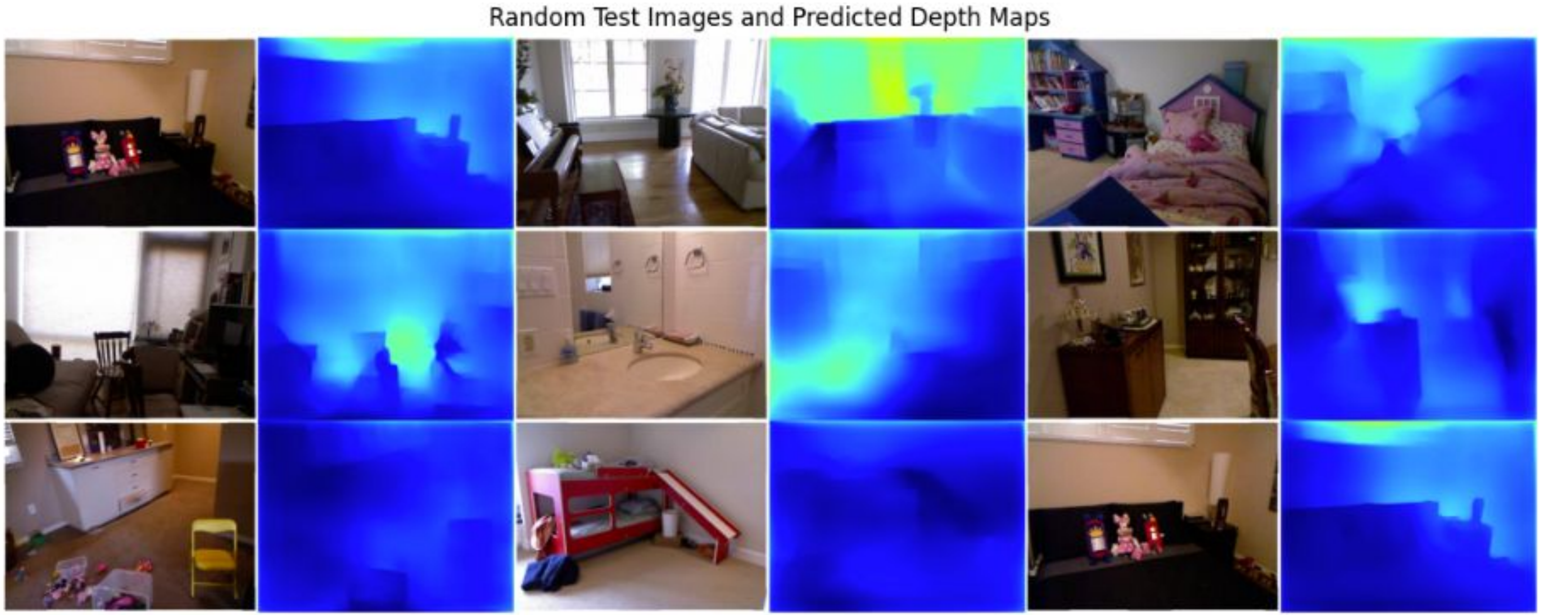
Qualitative Evaluation - DenseNet169

Random Test Images and Predicted Depth Maps

Qualitative Evaluation - NYU Pre-trained Model



Qualitative Evaluation - KITTI Pre-trained Model



Quantitative Evaluation

Metric	DenseNet169	DenseNet121	Paper NYU	Paper KITTI
$\boldsymbol{\delta}_1$	0.8033	0.8108	0.8407	0.0037
$\boldsymbol{\delta}_2$	0.9596	0.9642	0.9721	0.0170
$\boldsymbol{\delta}_3$	0.9912	0.9921	0.9721	0.0448
rel	0.1412	0.1396	0.1259	0.7246
rms	0.5246	0.5023	0.4712	2.3185
$\log_{10}$	0.0614	0.0597	0.0551	0.592

Conclusion

- The pre-trained NYU model shows superior results compared to our trained model with DenseNet169 on almost all metrics.
  - The superior performance can be attributed to the longer training duration (20 epochs vs 5 epochs).
- DenseNet121 performs better than DenseNet169.
  - When limited training resources are available it's better to use DenseNet121.
- The pre-trained KITTI model shows worst results.
  - The paper suggests that this result is influenced by the reduced quality of the provided depth maps in the KITTI dataset.

References

- [1812.11941] High Quality Monocular Depth Estimation via Transfer Learning
- Digging Into Self-Supervised Monocular Depth Estimation | Papers With Code
- [1702.02706] Semi-Supervised Deep Learning for Monocular Depth Map Prediction