Project 9d：Monocular Depth Estimation via Transfer Learning

Supervisor: Yongbo Chen

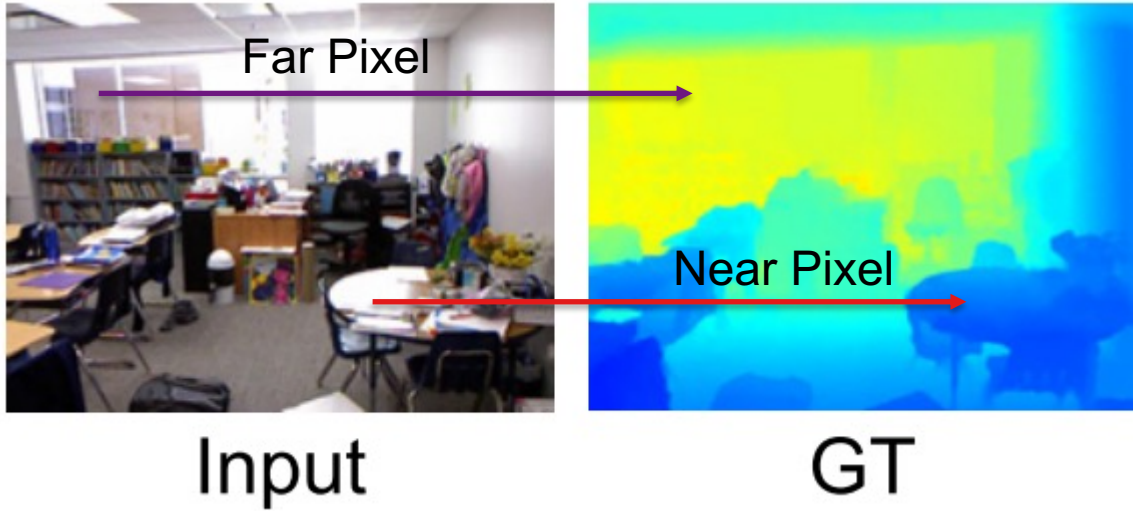
Who are we?
Students from Robotics at TU Delft and Autonomous Systems at DTU

Weihao Xuan, Ruijie Ren, Siyuan Wu, Liangchen Sui, Shaohang Han

Presentation by Liangchen Sui

Task Description:

- Pixel to pixel regression[1]



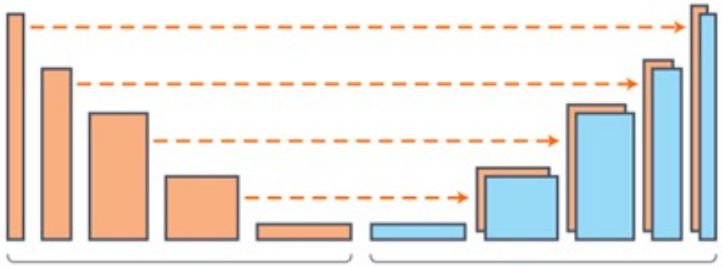
[1]Alhashim, Ibraheem, and Peter Wonka. "High quality monocular depth estimation via transfer learning." *arXiv preprint arXiv:1812.11941* (2018).

Encoder & Decoder Design:

- Encoder
  - Backbone of DenseNet-169 transferred from ImageNet classification

- Decoder
  - Successive bilinear upsampling blocks followed by shallow convolutional layers to realize pixel-pixel regression(high complexity proved to backfire)
  - 1x1 Convs to establish inter-channel connections
  - Skip connections to retrieve information and add high-resolution details

Encoder & Decoder Design:



LAYER	OUTPUT	FUNCTION
INPUT	480 × 640 × 3	
CONV1	240 × 320 × 64	DenseNet CONV1
POOL1	120 × 160 × 64	DenseNet POOL1
POOL2	60 × 80 × 128	DenseNet POOL2
POOL3	30 × 40 × 256	DenseNet POOL3
...	...	...
CONV2	15 × 20 × 1664	Convolution 1 × 1 of DenseNet BLOCK4

UP1	30 × 40 × 1664	Upsample 2 × 2
CONCAT1	30 × 40 × 1920	Concatenate POOL3
UP1-CONVA	30 × 40 × 832	Convolution 3 × 3
UP1-CONVB	30 × 40 × 832	Convolution 3 × 3
UP2	60 × 80 × 832	Upsample 2 × 2
CONCAT2	60 × 80 × 960	Concatenate POOL2
UP2-CONVA	60 × 80 × 416	Convolution 3 × 3
UP2-CONVB	60 × 80 × 416	Convolution 3 × 3
UP3	120 × 160 × 416	Upsample 2 × 2
CONCAT3	120 × 160 × 480	Concatenate POOL1
UP3-CONVA	120 × 160 × 208	Convolution 3 × 3
UP3-CONVB	120 × 160 × 208	Convolution 3 × 3
UP4	240 × 320 × 208	Upsample 2 × 2
CONCAT3	240 × 320 × 272	Concatenate CONV1
UP2-CONVA	240 × 320 × 104	Convolution 3 × 3
UP2-CONVB	240 × 320 × 104	Convolution 3 × 3
CONV3	240 × 320 × 1	Convolution 3 × 3

Encoder

Decoder

Loss Function Design

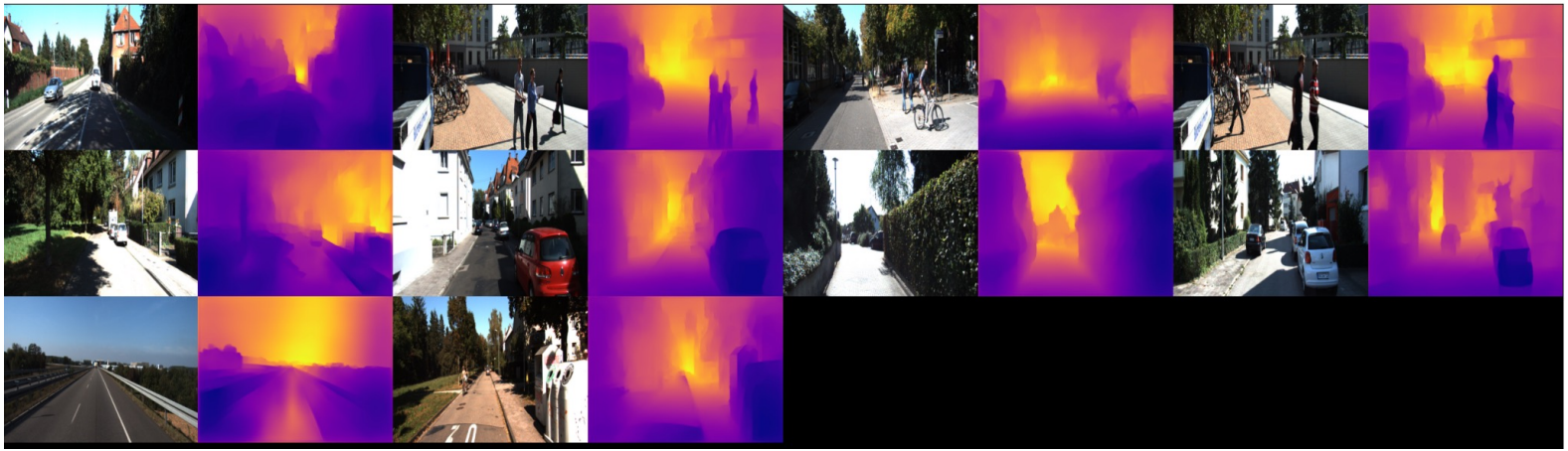
- L1 Loss(Depth + Gradient) + SSIM ➔ High Granularity
- L2 Loss(Depth + Gradient) + SSIM ➔ Fast Convergence
- Berhu Loss(Depth + Gradient) + SSIM ➔ Make a Compromise

Data Augmentation

- To achieve higher generalizations, series of operations simulating natural deformations are implemented on the dataset
  - Random crops
  - Horizontal flips
  - Vertical flips
  - Rotations
  - Affine transformations

Results

- RGB Inputs & Depth Predictions(fine-tunned on KITTI)



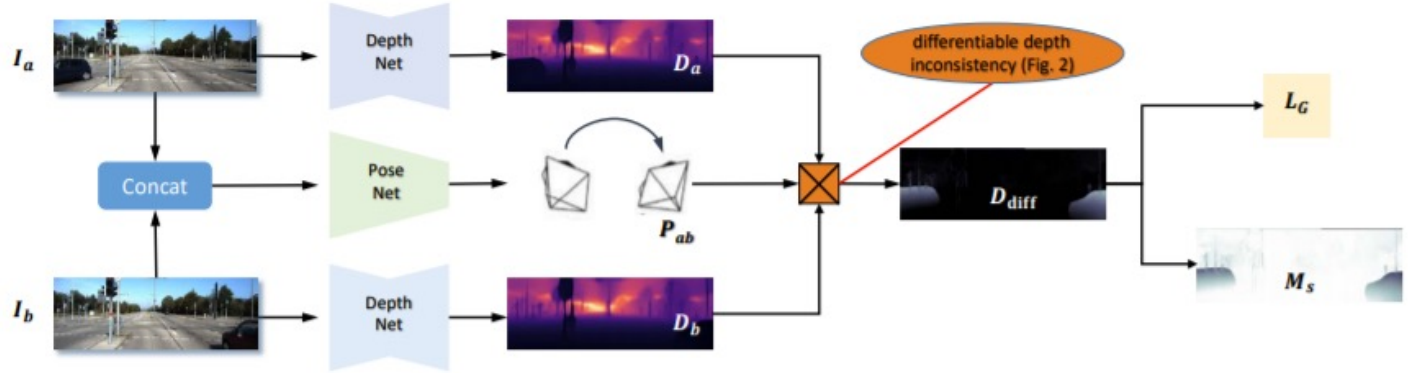
- Evaluation metrics are regretfully not available due to the key files missing in the benchmark

Future Work

- An interesting discovery: specific scenarios such as **outdoors** and **over exposures** are not considerately collected in NYU Depth v2 dataset, which could jeopardize the **robustness** and **generalization** of the model under some extreme circumstances, especially when the **depth values are large and light intensities are high.**

Future Work

- The structure of ideal model based on supervised learning[1] could be used as Depth Net component in unsupervised scale-consistent depth learning from video, which tends to serve as an important part in visual SLAM to provide depth information.



[1]Bian, Jia-Wang, et al. "Unsupervised Scale-consistent Depth Learning from Video." *International Journal of Computer Vision* (2021): 1-17.

Thank you for listening!