Monocular Depth Estimation using Deep Neural Networks

Student: Yuheng Li
Project Leader: Dr Changjae OH
Introduction of background

- **Depth estimation**: estimate the distance of each pixel in the image relative to the shooting source
- **Object**: investigate a depth estimation model for **UAV** (unmanned aerial vehicle) based on deep neural networks
Introduction of background

• Binocular VS. Monocular depth estimation
  - Binocular method —— require more devices, not light!
  - Monocular method

adjacent frames | stereo pairs

Occluded pixel | Good match

Frame -1, Frame 0, Frame+1 | left image, predicted right image
Introduction of background

- **Supervised VS. Unsupervised monocular depth estimation**

- **Supervised method** —— cost high to acquire ground truth!

  ![Supervised Method Diagram]

  - **Input image**
  - **CNN**
  - **Output: prediction depth**
  - **Target: ground truth**

- **Unsupervised method**

  ![Unsupervised Method Diagram]

  - **Input images 0, -1, +1**
  - **CNN**
  - **Prediction depth + relative pose**
  - **Output: reconstructed image**
  - **Target: input image 0**
Introduction of background

• Baseline model—Monodepth2

Monodepth2 on KITTI dataset

Monodepth2 on MidAir dataset

Strengths of Monodepth2
• Unsupervised monocular depth estimation model
• State-of-art model in KITTI dataset

Limitations of Monodepth2
• Almost can't improve the effect of high-resolution inputs
• Inaccurate in predicting the depth of large gradient areas
Proposed model SS-MDE

- Idea for improvement—bilinear interpolation

<table>
<thead>
<tr>
<th>High Resolution</th>
<th>Ground Truth</th>
<th>Bilinear Interpolation Low Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accurate</strong></td>
<td>1 1 10 10</td>
<td>1 3.25 7.75 10</td>
</tr>
<tr>
<td></td>
<td>1 1 10 10</td>
<td>1 3.25 7.75 10</td>
</tr>
<tr>
<td></td>
<td>1 1 10 10</td>
<td>1 3.25 7.75 10</td>
</tr>
<tr>
<td></td>
<td>1 1 10 10</td>
<td>1 3.25 7.75 10</td>
</tr>
<tr>
<td><strong>Inaccurate</strong></td>
<td>1 1 10 10</td>
<td>3.25 4.93 8.31 10</td>
</tr>
<tr>
<td></td>
<td>1 1 10 10</td>
<td>7.75 8.31 9.43 10</td>
</tr>
<tr>
<td></td>
<td>1 1 10 10</td>
<td>10 10 10 10</td>
</tr>
<tr>
<td><strong>Inaccurate</strong></td>
<td>1 10 10 10</td>
<td>10 9.43 8.31 9.43 10</td>
</tr>
<tr>
<td></td>
<td>1 10 10 10</td>
<td>10 8.31 4.94 8.31 10</td>
</tr>
<tr>
<td></td>
<td>1 10 10 10</td>
<td>10 8.31 4.94 8.31 10</td>
</tr>
<tr>
<td></td>
<td>1 10 10 10</td>
<td>10 9.43 8.31 9.43 10</td>
</tr>
</tbody>
</table>

Abs Rel: Absolute relative error

LR>HR 0.6

LR>HR 0.3

LR = HR

SS-MDE: Self-Supervised Monocular Depth Estimation
HR: High Resolution
LR: Low Resolution
Abs Rel: Absolute relative error
Proposed model SS-MDE

- Depth estimation network-U-net structure

Results of the experiment

• Qualitative results (tested on MidAir)

Input images

Monodepth2-MidAir

SS-MDE

- no detection error
- clearer semantic objects
- shaper edges
# Results of the experiment

## Quantitative results

<table>
<thead>
<tr>
<th>Model</th>
<th>Train</th>
<th>Abs Rel</th>
<th>Sq Rel</th>
<th>RMSE</th>
<th>RMSE log</th>
<th>$\delta &lt; 1.25$ ↑</th>
<th>$\delta &lt; 1.25^2$ ↑</th>
<th>$\delta &lt; 1.25^3$ ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang</td>
<td>KITTI</td>
<td>0.241</td>
<td>5.532</td>
<td>12.599</td>
<td>0.368</td>
<td>0.648</td>
<td>0.831</td>
<td>0.911</td>
</tr>
<tr>
<td>Monodepth</td>
<td>KITTI</td>
<td>0.314</td>
<td>8.713</td>
<td>13.595</td>
<td>0.438</td>
<td>0.678</td>
<td>0.828</td>
<td>0.895</td>
</tr>
<tr>
<td>ST-CLSTM</td>
<td>KITTI</td>
<td>0.404</td>
<td>6.390</td>
<td>13.685</td>
<td>0.438</td>
<td>0.751</td>
<td>0.865</td>
<td>0.911</td>
</tr>
<tr>
<td>Monodepth2-KITTI</td>
<td>KITTI</td>
<td>0.717</td>
<td>37.164</td>
<td>74.552</td>
<td>0.882</td>
<td>0.281</td>
<td>0.425</td>
<td>0.521</td>
</tr>
<tr>
<td>Monodepth2-MidAir</td>
<td>MidAir</td>
<td>0.135</td>
<td>2.500</td>
<td>13.214</td>
<td>0.222</td>
<td>0.720</td>
<td>0.910</td>
<td>0.996</td>
</tr>
<tr>
<td>M4Depth</td>
<td>MidAir</td>
<td>0.143</td>
<td>3.680</td>
<td><strong>8.864</strong></td>
<td>0.246</td>
<td><strong>0.840</strong></td>
<td>0.924</td>
<td>0.959</td>
</tr>
<tr>
<td>SS-MDE</td>
<td>MidAir</td>
<td><strong>0.114</strong></td>
<td><strong>1.742</strong></td>
<td>10.766</td>
<td><strong>0.173</strong></td>
<td>0.791</td>
<td><strong>0.967</strong></td>
<td><strong>0.998</strong></td>
</tr>
</tbody>
</table>

Abs Rel: Absolute Relative Error  
Sq Rel: Squared Relative Error  
RMSE: Root Mean Squared Error  
RMSE log: Root Mean Squared Logarithmic Error  
$\delta$: Standard Deviation
Summary of my final project

- Learned deep learning and completed literature survey for monocular depth estimation.
- Implemented the baseline Monodepth2.
- Performed a self-supervised monocular depth estimation in UAV data.
- Designed the algorithm to improve baseline model: dense skip connections in U-Net structure contributing to predicting more accurate depth maps.
- Implemented the algorithm and experiments validated that proposed model SS-MDE achieved state-of-art performance in UAV dataset.
- Realized the application of proposed model SS-MDE with real UAV video and got expected results.

<table>
<thead>
<tr>
<th>Literature Survey</th>
<th>Implemented Baseline</th>
<th>Designed Improved Algorithm</th>
<th>Implemented Algorithm</th>
<th>Achieved Good Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>2021.12.15</td>
<td>2022.1.15</td>
<td>2022.2.20</td>
<td>2022.3.5</td>
<td>2022.3.25</td>
</tr>
</tbody>
</table>
Thank you