Multi Image Depth from Defocus Network with Boundary Cue for Dual Aperture Camera

Gwangmo Song, Yumee Kim, Kukjin Chun, Kyoung Mu Lee
Seoul National University

Presenter : Gwangmo Song
Introduction

• Depth estimation problem
• Double defocused images (Different Depth-of-Field)
Previous Works

- **MRF-based vs. DNN-based**
- Depth from Defocus (DFD) problem [1,2]
- Dual DFD [3]

Introduction

• Contributions

Boundary Cue
- Edge information
- Improve accuracy

Real dataset
- Collect new dataset
- Limited circumstance
Dual DFD Network

Dual aperture Input

Shallow DoF  Deep DoF


CONV  CONV  CONV  CONV

N Res. Blocks  N Res. Blocks  N Res. Blocks

Add.

CONV

Depth map

Res. Block

CONV  ReLU  CONV

Sub.
Dual DFD Network

• Boundary Cue
  • Homogeneous region has less effect of blur
  • Subtraction highlights around the edge of the object
Dual DFD Network

Dual aperture Input

Shallow DoF  Deep DoF

Sub.

Add.

CONV  CONV  CONV  Add.

N Res. Blocks

Add.

CONV  CONV  CONV  Add.

N Res. Blocks

Add.

CONV  CONV  CONV  Add.

N Res. Blocks

CONV  ReLU  CONV

Res. Block

CONV

Depth map

Dual aperture Input

Shallow DoF  Deep DoF

Sub.

Add.

CONV  CONV  CONV  Add.

N Res. Blocks

Add.

CONV  CONV  CONV  Add.

N Res. Blocks

Add.

CONV  CONV  CONV  Add.

N Res. Blocks

CONV  ReLU  CONV

Res. Block

CONV

Depth map
Dual DFD Network

• Main Network
  • EDSR \[1\] based
  • Baseline network \[2\]

Dual DFD Network

Dual aperture Input

Shallow DoF  Deep DoF


N Res. Blocks  N Res. Blocks  N Res. Blocks

Depth map

Res. Block

CONV  CONV  CONV

Sub.

CONV  CONV  CONV

CONV  RELU  CONV

CONV  CONV  CONV

CONV  CONV  CONV
Datasets

• Synthetic dataset [1]
  • Dual defocused dataset
  • NYU-v2 dataset
  • Using thin lens model ($F_\# = 2, 14$)
    \[
    \sigma = \frac{1}{\sqrt{2}} \frac{p F_\# d_{IFP}}{d} \left( 1 - \frac{d_{IFP}}{d} \right)
    \]
  • 795 training data and 654 test data

Datasets

• Real dataset
  • Tunable aperture camera
  • F\# = 1.8, 4.0
  • Static scene
  • LIDAR align
  • 199 training data and 100 test data
Datasets

• Real dataset
  • Parking lot
  • Maximum distance : 70m
  • 3 types of car, 3 types of pedestrian
Experiments

- Synthetic dataset
  - *NYU v2*-based
- *rel*
  - Average relative error
- *log10*
  - Average log_{10} error
- *rms*
  - Root mean square error

<table>
<thead>
<tr>
<th>Method</th>
<th>rel</th>
<th>log10</th>
<th>rms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single DFD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anwar <em>et al.</em> [1]</td>
<td>0.094</td>
<td>0.039</td>
<td>0.347</td>
</tr>
<tr>
<td>D3-Net [2]</td>
<td>0.068</td>
<td>0.028</td>
<td>0.274</td>
</tr>
<tr>
<td>D3-Net* [2]</td>
<td>0.036</td>
<td>0.016</td>
<td>0.144</td>
</tr>
<tr>
<td>D3-Net** [2]</td>
<td>0.056</td>
<td>0.024</td>
<td>0.244</td>
</tr>
<tr>
<td>Dual DFD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3-Net** [2]</td>
<td>0.030</td>
<td>0.013</td>
<td>0.164</td>
</tr>
<tr>
<td>Song <em>et al.</em> [3]</td>
<td>0.028</td>
<td>0.012</td>
<td>0.154</td>
</tr>
<tr>
<td>Our</td>
<td>0.026</td>
<td>0.011</td>
<td>0.139</td>
</tr>
</tbody>
</table>

* Using dataset of [1]
** Using dataset of [3]

Experiments

- Ablation Study
- Effect of boundary cue

<table>
<thead>
<tr>
<th>Method</th>
<th>rel</th>
<th>log10</th>
<th>rms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.031</td>
<td>0.013</td>
<td>0.162</td>
</tr>
<tr>
<td>Boundary Cue</td>
<td>0.028</td>
<td>0.012</td>
<td>0.146</td>
</tr>
<tr>
<td>Boundary Cue + Skip Conn.</td>
<td>0.026</td>
<td>0.011</td>
<td>0.139</td>
</tr>
</tbody>
</table>
Experiment

- Qualitative

[Image: Images and depth maps of two different images with annotations for Gwangmo Song and Kyoung Mu Lee's work on depth estimation for dual defocused images.]

## Experiments

- **Real dataset**
- **Patch-size**
  - 224 vs. 48
  - Homogeneous region
- **GT depth map**
  - Sparse LIDAR point
  - Boundary is **not clean**
  - Boundary cue ↓

<table>
<thead>
<tr>
<th>Method</th>
<th>rel</th>
<th>log10</th>
<th>rms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single DFD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3-Net [1]</td>
<td>0.027</td>
<td>0.012</td>
<td>2.070</td>
</tr>
<tr>
<td>Dual DFD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3-Net [1]</td>
<td>0.027</td>
<td>0.012</td>
<td>1.202</td>
</tr>
<tr>
<td>Song <em>et al.</em> [2]</td>
<td>0.019</td>
<td>0.008</td>
<td>1.400</td>
</tr>
<tr>
<td><strong>Our</strong></td>
<td><strong>0.018</strong></td>
<td><strong>0.008</strong></td>
<td><strong>1.320</strong></td>
</tr>
</tbody>
</table>


Experiment

- Qualitative

<table>
<thead>
<tr>
<th>Image ($F_# = 1.8$)</th>
<th>Image ($F_# = 4$)</th>
<th>GT depth map</th>
<th>Our result</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image5.png" alt="Image" /></td>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
</tr>
<tr>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
<td><img src="image11.png" alt="Image" /></td>
<td><img src="image12.png" alt="Image" /></td>
</tr>
</tbody>
</table>
Experiments

- Object-based measure
  - Ignore sky, road
  - Center of object
  - Mean value of bounding box

<table>
<thead>
<tr>
<th>Method</th>
<th>rel</th>
<th>log10</th>
<th>rms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single DFD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3-Net [1]</td>
<td>0.045</td>
<td>0.020</td>
<td>1.064</td>
</tr>
<tr>
<td>Dual DFD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3-Net [1]</td>
<td>0.041</td>
<td>0.018</td>
<td>0.943</td>
</tr>
<tr>
<td>Song et al. [2]</td>
<td>0.040</td>
<td>0.016</td>
<td>0.887</td>
</tr>
<tr>
<td>Our</td>
<td>0.031</td>
<td>0.013</td>
<td>0.718</td>
</tr>
</tbody>
</table>

• Create **boundary cue** through dual defocused images
• Proposal of **DFD network** structure using boundary cue
• Dataset collection using tunable aperture camera
• Record **SOTA** in synthetic and real dataset