Multi-Level Context Ultra-Aggregation for Stereo Matching

Guang-Yu Nie¹, Ming-Ming Cheng², Yun Liu², Zhengfa Liang³,
Deng-Ping Fan², Yue Liu¹,⁴, and Yongtian Wang¹,⁴

¹ Beijing Institute of Technology        ² TKLNDST, CS, Nankai University
³ National Key Laboratory of Science and Technology on Blind Signal Processing
⁴ AICFVE, Beijing Film Academy
Depth from images is a very intuitive ability

- Given two images of a scene from (slightly) different viewpoints, we are able to infer depth

Can we do the same using computers?

- Yes

Depth from Stereo

Geometry in stereo

- Think of images as projections of 3D points (in the real world) onto a 2D surface (image plane)

- $X_A$ is the projection of $X$, $X_1$, $X_2$, $X_3$, .... onto the left image

- $X$, $X_1$, $X_2$, $X_3$ will also project onto the right image

Depth from Stereo

Geometry in stereo

- Projections of $X_1$, $X_2$, $X_3$ on right image all lie on a line
- This line is known as an **epipolar line**
  - Projections of cameras’ optical centers $O_A$, $O_B$ onto the images
  - Points $e_A$, $e_B$ are known as **epipoles**
  - All epipolar lines will intersect at epipoles
  - Left image has corresponding epipolar line

Depth from Stereo
Geometry in stereo

What does this give us?

- All 3D points that could have resulted in $X_A$ must have a projection on the right image, and must be on the epipolar line $e_B - x_B$

- Given just the left/right images and $X_A$, you can search on the corresponding epipolar line in the right image. If you can find the corresponding match $X_B$, you can uniquely determine the 3D position of $X$.

Depth from Stereo

Geometry in stereo

- Epipolar lines can be made parallel through a process called **rectification**
- Simplifies the process of finding a match and calculating the 3D point

Depth from Stereo
Geometry in stereo

Problem statement, reformulated:
Find the disparity for every pixel in the left (or right) image by finding matches in the right (or left) image

\[
\text{disparity} = x - x' = \frac{Bf}{Z} \quad \frac{x - x'}{O - O'} = \frac{f}{Z}
\]

Source: https://docs.opencv.org/3.0-beta/doc/py_tutorials/py_calib3d/py_depthmap/py_depthmap.html
Related Research
Basic stereo matching algorithm

1. If necessary, **rectify** the two stereo images to transform epipolar lines into scanlines

2. For each pixel $x$ in the first image:
   - Find corresponding **epipolar scanline** in the right image
   - Search the scanline and pick the best match $x'$
   - Compute disparity $x-x'$ and set $\text{depth}(x) = \frac{Bf}{(x-x')}$

Source: [https://courses.cs.washington.edu/courses/cse455/16wi/notes/11_Stereo.pdf](https://courses.cs.washington.edu/courses/cse455/16wi/notes/11_Stereo.pdf)
Related Research

Failures of correspondence search

Textureless surfaces

Occlusions, repetition

Non-Lambertian surfaces, specularities

Source: https://courses.cs.washington.edu/courses/cse455/16wi/notes/11_Stereo.pdf
Related Research

Learning-Based Stereo Matching

End-to-end training network
Related Research

GC-Net by Kendall et al.

End-to-End Learning of Geometry and Context for Deep Stereo Regression (ICCV’17)

Figure 1: Our end-to-end deep stereo regression architecture, GC-Net (Geometry and Context Network).
Related Research

PSM-Net by Chang et al.

Pyramid Stereo Matching Network (CVPR’18)
Related Research
Learning-Based Stereo Matching

End-to-end training network
Related Research
Different aggregation patterns

(a) DenseNets

(b) Deep Layer Aggregation

Intra-Level combination
Method/MCUA

Multi-Level Context Ultra-Aggregation

![Diagram with receptive field and 2-D feature symbols]

IN

F_0 \rightarrow F_1 \rightarrow F_2 \rightarrow F_3 \rightarrow F_4 \rightarrow F_5 \rightarrow F_6 \rightarrow F_7 \rightarrow F_8 \rightarrow OUT

\(1 \times\)

\(\frac{1}{2} \times\)

\(\frac{1}{4} \times\)
MCUA Intra-Level Combination

Receptive field
2-D feature
Connections

(a) $F_0 \xrightarrow{1 \times} F_1 \xrightarrow{1 \times} F_2 \xrightarrow{1 \times} F_3 \xrightarrow{1 \times} F_4 \xrightarrow{1 \times} F_5 \xrightarrow{1 \times} F_6 \xrightarrow{1 \times} F_7 \xrightarrow{1 \times} F_8 \xrightarrow{1 \times} \text{OUT}$
MCUA  Inter-Level Combination

Independent child module

AvgPool

Share parameters

IN

OUT

1x

F0

F1

F2

F3

F4

F5

F6

F7

F8

1/2

1/4

1/4

1/4

1/2

Receptive field

2-D feature

Connections
**MCUA Inter-Level Combination**

Independent child module

- **AvgPool**
  - $\frac{1}{2} \times $ P_0

- **IN**

- **F_0**
  - $\frac{1}{2} \times $

- **F_1**
  - $\frac{1}{4} \times $

- **F_2**
  - $\frac{1}{4} \times $

- **F_3**

- **F_4**
  - $\frac{1}{4} \times $

- **F_5**

- **F_6**

- **F_7**

- **F_8**

- **OUT**

**Icons:**
- Green squares: Receptive field
- White boxes: 2-D feature
- Dotted arrows: Connections

*IGTA2019*
MCUA Inter-Level Combination

Independent child module

Share parameters

Receptive field
2-D feature
Connections

IN

OUT

AvgPool

F₀

F₁

F₂

F₃

F₄

F₅

F₆

F₇

F₈

\[
\begin{align*}
F₀ &\rightarrow F₁ \rightarrow F₂ \rightarrow F₃ \\
\frac{1}{2} \times &\quad \frac{1}{4} \times &\quad \frac{1}{4} \times
\end{align*}
\]

IGTA2019
MCUA Inter-Level Combination

Independent child module

AvgPool $F_0$ $F_1$ $F_2$ $F_3$

IN

$\frac{1}{2} \times$ $\frac{1}{4} \times$

Share parameters

2-D feature

Connections

Receptive field

$F_0$ $F_1$ $F_2$ $F_3$ $F_4$ $F_5$ $F_6$ $F_7$ $F_8$

OUT

$1 \times$ $\frac{1}{2} \times$ $\frac{1}{4} \times$

$\frac{1}{2} \times$ $\frac{1}{4} \times$ $\frac{1}{4} \times$

$\frac{1}{2} \times$

$\frac{1}{4} \times$
MCUA Dense Connection
MCUA Dense Connection

1x1 \arrow{F_0} \times \frac{1}{2} \rightarrow F_1 \rightarrow F_2 \rightarrow F_3 \rightarrow F_4 \rightarrow F_5 \rightarrow F_6 \rightarrow F_7 \rightarrow F_8 \rightarrow \text{OUT}

AvgPool \rightarrow \text{Receptive field} \rightarrow \text{2-D feature} \rightarrow \text{Connections}
MCUA

1x1

\[
\begin{pmatrix}
\frac{1}{2} & \frac{1}{2} \\
\end{pmatrix}
\]

Receptive Field

\[
\begin{pmatrix}
2W \\
2H \\
\end{pmatrix}
\]

Capture more area

IGTA2019
MCUA Stereo Matching

Stereo Images

Unary Features Learning

Cost Volume

Cost Volume Regularization

Disparity Map

1x

1/2

1/4

Scale

Information Flow

2-D Features

3-D Features

Element-wise Summation

Concatenation

Output

Residual

Warped Map

Initial Map
EMCUA Stereo Matching

Stereo Images

Unary Features Learning

Cost Volume

Cost Volume Regularization

Disparity Map

Scale

Information Flow

2-D Features

3-D Features

Element-wise Summation

Concatenation

Output

Residual

Warped Map

Initial Map

1x

1/2 x

1/4 x

1x

1/4 x

1/2 x

1/4 x
Scene Flow dataset:
FlyingThings3D, Driving, Monkaa

>39000 (35454/4370 train/test) stereo frames
960 × 540 pixel resolution

KITTI2015/2012 datasets
Left view
Right view
Disparity map (Ground truth)

KITTI2015: 200/200 train/test stereo images
KITTI2012: 194/200 train/test stereo images
1242 × 375 pixel resolution

Source:
https://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html
http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo
http://www.cvlibs.net/datasets/kitti/eval_stereo_flow.php?benchmark=stereo
Experiment

Train on a lot of data:
- Scene Flow datasets
- Finetuning on KITTI

Test on Flying Things and KITTI

Implementation Details

Input: 256×512 pixel resolution
Optimizer: Adam

The training process of EMCUA contains two steps:
- Train MCUA:
  - 20+50 epochs on SF dataset (lr=0.01)
  - 600 (lr=0.001) + 400 (lr=0.0001) epochs on KITTI datasets
- Train EMCUA (+ Residual module)
  - 1 epoch on SF dataset (lr=0.01)
  - 600 (lr=0.001) + 400 (lr=0.0001) epochs on KITTI datasets
### Performance on KITTI2015 dataset

**Table 2. KITTI2015 Results**

<table>
<thead>
<tr>
<th>Mod.</th>
<th>All (%)</th>
<th>Noc (%)</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D1-bg</td>
<td>D1-fg</td>
<td>D1-all</td>
<td>D1-bg</td>
<td>D1-fg</td>
<td>D1-all</td>
</tr>
<tr>
<td>SegStereo</td>
<td>1.88</td>
<td>4.07</td>
<td>2.25</td>
<td>1.76</td>
<td>3.70</td>
<td>2.08</td>
</tr>
<tr>
<td>iResNet</td>
<td>2.25</td>
<td>3.40</td>
<td>2.44</td>
<td>2.07</td>
<td><strong>2.76</strong></td>
<td>2.19</td>
</tr>
<tr>
<td>CRL</td>
<td>2.48</td>
<td>3.59</td>
<td>2.67</td>
<td>2.32</td>
<td>3.12</td>
<td>2.45</td>
</tr>
<tr>
<td>GC-Net [9]</td>
<td>2.21</td>
<td>6.16</td>
<td>2.87</td>
<td>2.02</td>
<td>5.58</td>
<td>2.61</td>
</tr>
<tr>
<td>PSM-Net</td>
<td>1.86</td>
<td>4.62</td>
<td>2.32</td>
<td>1.71</td>
<td>4.31</td>
<td>2.14</td>
</tr>
<tr>
<td>MCUA</td>
<td>1.69</td>
<td>4.38</td>
<td>2.14</td>
<td>1.55</td>
<td>3.90</td>
<td>1.93</td>
</tr>
<tr>
<td><strong>EMCUA</strong></td>
<td><strong>1.66</strong></td>
<td><strong>4.27</strong></td>
<td><strong>2.09</strong></td>
<td><strong>1.50</strong></td>
<td><strong>3.88</strong></td>
<td><strong>1.90</strong></td>
</tr>
</tbody>
</table>

“**All**” and “**Noc**” : percentage of outliers averaged over ground truth pixels of all/non-occluded regions. “**D1-bg**”, “**D1-fg**”, and “**D1-all**”: percentage of outliers averaged only over background regions, foreground regions, and all ground truth pixels.

Sample output:

(a) EMCUA

(b) PSM-Net
### Performance KITTI2012 dataset

<table>
<thead>
<tr>
<th>Mod</th>
<th>&gt; 2px Noc</th>
<th>&gt; 3px Noc</th>
<th>&gt; 4px Noc</th>
<th>&gt; 5px Noc</th>
<th>ME(px) Noc</th>
<th>&gt; 2px All</th>
<th>&gt; 3px All</th>
<th>&gt; 4px All</th>
<th>&gt; 5px All</th>
<th>ME(px) All</th>
<th>AN</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SegStereo</td>
<td>2.66</td>
<td>3.19</td>
<td>1.68</td>
<td>2.03</td>
<td>1.25</td>
<td>1.52</td>
<td>1.00</td>
<td>1.21</td>
<td>0.5</td>
<td>0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>iResNet</td>
<td>2.69</td>
<td>3.34</td>
<td>1.71</td>
<td>2.16</td>
<td>1.30</td>
<td>1.63</td>
<td>1.06</td>
<td>1.32</td>
<td>0.5</td>
<td>0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>GC-Net</td>
<td>2.71</td>
<td>3.46</td>
<td>1.77</td>
<td>2.30</td>
<td>1.36</td>
<td>1.77</td>
<td>1.12</td>
<td>1.46</td>
<td>0.6</td>
<td>0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSM-net</td>
<td>2.44</td>
<td>3.01</td>
<td>1.49</td>
<td>1.89</td>
<td>1.12</td>
<td>1.42</td>
<td>0.90</td>
<td>1.15</td>
<td>0.5</td>
<td>0.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCUA</td>
<td>2.07</td>
<td>2.64</td>
<td>1.30</td>
<td>1.70</td>
<td>0.98</td>
<td>1.29</td>
<td>0.80</td>
<td>1.04</td>
<td>0.5</td>
<td>0.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>EMCUA</strong></td>
<td><strong>2.02</strong></td>
<td><strong>2.56</strong></td>
<td><strong>1.26</strong></td>
<td><strong>1.64</strong></td>
<td><strong>0.95</strong></td>
<td><strong>1.24</strong></td>
<td><strong>0.76</strong></td>
<td><strong>0.99</strong></td>
<td><strong>0.4</strong></td>
<td><strong>0.5</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

“Noc” and “All”: percentage of erroneous pixels in non-occluded areas, and in total. “AN” and “AA”: average disparity/end-point error in non-occluded areas, and in total. “ME”: mean error.

Sample output:

(a) EMCUA

(b) PSM-Net
Residual module is mainly used to improve the performance of the accuracy of the foreground.

### Table 2. KITTI2015 Results

<table>
<thead>
<tr>
<th>Mod.</th>
<th>All (%)</th>
<th>Noc (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D1-bg</td>
<td>D1-fg</td>
</tr>
<tr>
<td>SegStereo</td>
<td>1.88</td>
<td>4.07</td>
</tr>
<tr>
<td>iResNet</td>
<td>2.25</td>
<td><strong>3.40</strong></td>
</tr>
<tr>
<td>CRL</td>
<td>2.48</td>
<td>3.59</td>
</tr>
<tr>
<td>GC-Net [9]</td>
<td>2.21</td>
<td>6.16</td>
</tr>
<tr>
<td>PSM-Net</td>
<td>1.86</td>
<td>4.62</td>
</tr>
<tr>
<td>MCUA</td>
<td>1.69</td>
<td>4.38</td>
</tr>
<tr>
<td>EMCUA</td>
<td><strong>1.66</strong></td>
<td>4.27</td>
</tr>
</tbody>
</table>

“All” and “Noc”: percentage of outliers averaged over ground truth pixels of all/non-occluded regions. “D1-bg”, “D1-fg”, and “D1-all”: percentage of outliers averaged only over background regions, foreground regions, and all ground truth pixels.

### Table 3. KITTI2012 Results

<table>
<thead>
<tr>
<th>Mod</th>
<th>&gt; 2px Noc</th>
<th>&gt; 2px All</th>
<th>&gt; 3px Noc</th>
<th>&gt; 3px All</th>
<th>&gt; 4px Noc</th>
<th>&gt; 4px All</th>
<th>&gt; 5px Noc</th>
<th>&gt; 5px All</th>
<th>ME(px) AN</th>
<th>ME(px) AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SegStereo</td>
<td>2.66</td>
<td>3.19</td>
<td>1.68</td>
<td>2.03</td>
<td>1.25</td>
<td>1.52</td>
<td>1.00</td>
<td>1.21</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>iResNet</td>
<td>2.69</td>
<td>3.34</td>
<td>1.71</td>
<td>2.16</td>
<td>1.30</td>
<td>1.63</td>
<td>1.06</td>
<td>1.32</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>GC-Net</td>
<td>2.71</td>
<td>3.46</td>
<td>1.77</td>
<td>2.30</td>
<td>1.36</td>
<td>1.77</td>
<td>1.12</td>
<td>1.46</td>
<td>0.6</td>
<td>0.7</td>
</tr>
<tr>
<td>PSM-net</td>
<td>2.44</td>
<td>3.01</td>
<td>1.49</td>
<td>1.89</td>
<td>1.12</td>
<td>1.42</td>
<td>0.90</td>
<td>1.15</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td>MCUA</td>
<td>2.07</td>
<td>2.64</td>
<td>1.30</td>
<td>1.70</td>
<td>0.98</td>
<td>1.29</td>
<td>0.80</td>
<td>1.04</td>
<td>0.5</td>
<td>0.5</td>
</tr>
<tr>
<td>EMCUA</td>
<td><strong>2.02</strong></td>
<td><strong>2.56</strong></td>
<td><strong>1.26</strong></td>
<td><strong>1.64</strong></td>
<td><strong>0.95</strong></td>
<td><strong>1.24</strong></td>
<td><strong>0.76</strong></td>
<td><strong>0.99</strong></td>
<td><strong>0.4</strong></td>
<td><strong>0.5</strong></td>
</tr>
</tbody>
</table>

“Noc” and “All”: percentage of erroneous pixels in non-occluded areas, and in total. “AN” and “AA”: average disparity/end-point error in non-occluded areas, and in total. “ME”: mean error.
**Performance**

**Scene Flow Datasets**

Table 4. Performance comparison on Scene Flow test set

<table>
<thead>
<tr>
<th>Mod.</th>
<th>EPE</th>
<th>Mod.</th>
<th>EPE</th>
<th>Mod.</th>
<th>EPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCUA</td>
<td>0.56</td>
<td>PSM-Net [2]</td>
<td>1.09</td>
<td>StereoNet [10]</td>
<td>1.10</td>
</tr>
</tbody>
</table>

Mod.: model; EPE: end-point-error;

Sample output

Inputs | Ground truth | MCUA | PSM-Net
Discussion

Different aggregation schemes

- Dense connection
- Deep Layer Aggregation
- MCUA

<table>
<thead>
<tr>
<th>Mod.</th>
<th>Scene Flow</th>
<th>KITTI2015</th>
<th>Para.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt; 1px</td>
<td>&gt; 3px</td>
<td>&gt; 5px</td>
</tr>
<tr>
<td>PSM-Net</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>DenseNets</td>
<td>8.526</td>
<td>3.329</td>
<td>2.286</td>
</tr>
<tr>
<td>DLA</td>
<td>8.586</td>
<td>3.337</td>
<td>2.280</td>
</tr>
<tr>
<td>MCUA</td>
<td>7.885</td>
<td>3.108</td>
<td>2.148</td>
</tr>
</tbody>
</table>

Compare of architecture components

<table>
<thead>
<tr>
<th></th>
<th>Scene Flow</th>
<th>KITTI2015</th>
<th>Para.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt; 1px</td>
<td>&gt; 3px</td>
<td>&gt; 5px</td>
</tr>
<tr>
<td>UChi</td>
<td>8.185</td>
<td>3.153</td>
<td>2.147</td>
</tr>
<tr>
<td>Chi</td>
<td>8.133</td>
<td>3.242</td>
<td>2.226</td>
</tr>
<tr>
<td>DenPool</td>
<td>8.187</td>
<td>3.187</td>
<td>2.179</td>
</tr>
<tr>
<td>MCUA</td>
<td>7.885</td>
<td>3.108</td>
<td>2.148</td>
</tr>
</tbody>
</table>

> tpx: EPE; VE: three-pixel-error; Para.: number of parameters.
**Discussion**

### Effect of MCUA

![Diagram of MCUA](image)

#### Table 5. Ablation study

<table>
<thead>
<tr>
<th>Mod.</th>
<th>Scene Flow</th>
<th>KITTI2015</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt; 1px</td>
<td>&gt; 3px</td>
</tr>
<tr>
<td>PSM-Net</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>DenseNets</td>
<td>8.526</td>
<td>3.329</td>
</tr>
<tr>
<td>DLA</td>
<td>8.586</td>
<td>3.337</td>
</tr>
<tr>
<td>MCUA</td>
<td><strong>7.885</strong></td>
<td><strong>3.108</strong></td>
</tr>
</tbody>
</table>

**Compare of aggregation patterns**

**Compare of architecture components**

<table>
<thead>
<tr>
<th>Mod.</th>
<th>Scene Flow</th>
<th>KITTI2015</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt; 1px</td>
<td>&gt; 3px</td>
</tr>
<tr>
<td>UChi</td>
<td>8.185</td>
<td>3.153</td>
</tr>
<tr>
<td>Chi</td>
<td>8.133</td>
<td>3.242</td>
</tr>
<tr>
<td>DenPool</td>
<td>8.187</td>
<td>3.187</td>
</tr>
<tr>
<td>MCUA</td>
<td><strong>7.885</strong></td>
<td><strong>3.108</strong></td>
</tr>
</tbody>
</table>

> *tx*: EPE; VE: three-pixel-error; Para.: number of parameters.
## Discussion

### Effect of MCUA

![Diagram of MCUA](image)

<table>
<thead>
<tr>
<th>Mod.</th>
<th>Scene Flow</th>
<th>KITTI2015</th>
<th>Para.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt; 1px</td>
<td>&gt; 3px</td>
<td>&gt; 5px</td>
</tr>
<tr>
<td>PSM-Net</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>DenseNets</td>
<td>8.526</td>
<td>3.329</td>
<td>2.286</td>
</tr>
<tr>
<td>DLA</td>
<td>8.586</td>
<td>3.337</td>
<td>2.280</td>
</tr>
<tr>
<td>MCUA</td>
<td><strong>7.885</strong></td>
<td><strong>3.108</strong></td>
<td><strong>2.148</strong></td>
</tr>
</tbody>
</table>

**Compare of aggregation patterns**

<table>
<thead>
<tr>
<th>Mod.</th>
<th>Scene Flow</th>
<th>KITTI2015</th>
<th>Para.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt; 1px</td>
<td>&gt; 3px</td>
<td>&gt; 5px</td>
</tr>
<tr>
<td>PSM-Net</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>DenseNets</td>
<td>8.526</td>
<td>3.329</td>
<td>2.286</td>
</tr>
<tr>
<td>DLA</td>
<td>8.586</td>
<td>3.337</td>
<td>2.280</td>
</tr>
<tr>
<td>MCUA</td>
<td><strong>7.885</strong></td>
<td><strong>3.108</strong></td>
<td><strong>2.148</strong></td>
</tr>
</tbody>
</table>

**Compare of architecture components**

<table>
<thead>
<tr>
<th>Mod.</th>
<th>Scene Flow</th>
<th>KITTI2015</th>
<th>Para.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt; 1px</td>
<td>&gt; 3px</td>
<td>&gt; 5px</td>
</tr>
<tr>
<td>UChi</td>
<td>8.185</td>
<td>3.153</td>
<td><strong>2.147</strong></td>
</tr>
<tr>
<td>Chi</td>
<td>8.133</td>
<td>3.242</td>
<td>2.226</td>
</tr>
<tr>
<td>DenPool</td>
<td>8.187</td>
<td>3.187</td>
<td>2.179</td>
</tr>
<tr>
<td>MCUA</td>
<td><strong>7.885</strong></td>
<td><strong>3.108</strong></td>
<td><strong>2.148</strong></td>
</tr>
</tbody>
</table>

* > tpx: EPE; VE: three-pixel-error; Para.: number of parameters.*
Discussion

Effect of MCUA

Table 5. Ablation study

<table>
<thead>
<tr>
<th>Mod.</th>
<th>Scene Flow</th>
<th>KITTI2015</th>
<th>Para.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$&gt;1px$</td>
<td>$&gt;3px$</td>
<td>$&gt;5px$</td>
</tr>
<tr>
<td>PSM-Net</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>DenseNets</td>
<td>8.526</td>
<td>3.329</td>
<td>2.286</td>
</tr>
<tr>
<td>DLA</td>
<td>8.586</td>
<td>3.337</td>
<td>2.280</td>
</tr>
<tr>
<td>MCUA</td>
<td>7.885</td>
<td>3.108</td>
<td>2.148</td>
</tr>
</tbody>
</table>

Compare of aggregation patterns

<table>
<thead>
<tr>
<th>Mod.</th>
<th>Scene Flow</th>
<th>KITTI2015</th>
<th>Para.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$&gt;1px$</td>
<td>$&gt;3px$</td>
<td>$&gt;5px$</td>
</tr>
<tr>
<td>UChi</td>
<td>8.185</td>
<td>3.153</td>
<td>2.147</td>
</tr>
<tr>
<td>Chi</td>
<td>8.133</td>
<td>3.242</td>
<td>2.226</td>
</tr>
<tr>
<td>DenPool</td>
<td>8.187</td>
<td>3.187</td>
<td>2.179</td>
</tr>
<tr>
<td>MCUA</td>
<td>7.885</td>
<td>3.108</td>
<td>2.148</td>
</tr>
</tbody>
</table>

> $tx$: EPE; VE: three-pixel-error; Para.: number of parameters.
**Discussion**

Effect of MCUA

![Diagram showing the effect of MCUA](image)

<table>
<thead>
<tr>
<th>Mod.</th>
<th>Scene Flow</th>
<th>KITTI2015</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt; 1px</td>
<td>&gt; 3px</td>
</tr>
<tr>
<td>PSM-Net</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>DenseNets</td>
<td>8.526</td>
<td>3.329</td>
</tr>
<tr>
<td>DLA</td>
<td>8.586</td>
<td>3.337</td>
</tr>
<tr>
<td>MCUA</td>
<td><strong>7.885</strong></td>
<td><strong>3.108</strong></td>
</tr>
</tbody>
</table>

**Table 5. Ablation study**

Compare of aggregation patterns

<table>
<thead>
<tr>
<th>Mod.</th>
<th>Scene Flow</th>
<th>KITTI2015</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&gt; 1px</td>
<td>&gt; 3px</td>
</tr>
<tr>
<td>UChi</td>
<td>8.185</td>
<td>3.153</td>
</tr>
<tr>
<td>Chi</td>
<td>8.133</td>
<td>3.242</td>
</tr>
<tr>
<td>DenPool</td>
<td>8.187</td>
<td>3.187</td>
</tr>
<tr>
<td>MCUA</td>
<td><strong>7.885</strong></td>
<td><strong>3.108</strong></td>
</tr>
</tbody>
</table>

> tpx: EPE; VE: three-pixel-error; Para.: number of parameters.
Conclusion

- We propose a general feature aggregation scheme, MCUA, which contains both intra- and inter-level feature aggregation, while DenseNets and DLA contain only intra-level aggregation.

- We use an independent child module to introduce inter-level aggregation, which enlarges the receptive fields and captures more context information.
Future work

- Dataset bias (Stereo matching, Depth estimation)
- Real-time stereo matching
Future work

Datasets

Scene Flow dataset:
FlyingThings3D, Driving, Monkaa

>39000 (35454/4370 train/test) stereo frames
960 × 540 pixel resolution

KITTI2015/2012 datasets

Left view

Right view

Disparity map (Ground truth)

KITTI2015: 200/200 train/test stereo images
KITTI2012: 194/200 train/test stereo images
1242 × 375 pixel resolution

Source: https://lmb.informatik.uni-freiburg.de/resources/datasets/SceneFlowDatasets.en.html
http://www.cvlibs.net/datasets/kitti/eval_scene_flow.php?benchmark=stereo
http://www.cvlibs.net/datasets/kitti/eval_stereo_flow.php?benchmark=stereo
Future work

- Dataset bias (Stereo matching, Depth estimation)
- Real-time stereo matching

Matching cost: SSD, SAD, or normalized correlation

\[
SSD(x,y,d) = \sum_{(x,y)\in w} \left| l_i(x,y) - l_r(x-d,y) \right|^2
\]

Future work

- Dataset bias (Stereo matching, Depth estimation)
- Real-time stereo matching

StereoNet architecture (ECCV'18)


Qualitative results on the FlyingThings3D test set
Thanks for your watching.
Q&A

Guang-Yu Nie
guyuneeeee@outlook.com

IGTA2019
04/19-20/2019