A NOVEL MONOCULAR DISPARITY ESTIMATION NETWORK WITH DOMAIN TRANSFORMATION AND AMBIGUITY LEARNING

Juan Luis Gonzalez Bello Munchurl Kim
Department of Electrical Engineering
Korea Advanced Institute of Science and Technology

CONTRIBUTIONS
Our novel network architecture outperforms the unsupervised monocular baseline [1] by:
• Accounting for ambiguities (occluded, complex or cluttered image areas)
• Efficient fusion between encoder (left domain) and decoder (left-right domain) features via rectangular 5x3 convolutions and domain transformation blocks
• Full disparity estimation in a single pass
• 50% parameter reduction

EXPERIMENTAL RESULTS ON THE KITTI DATASET
Monodepth (pp) [1] rdispnet_m rdispnet_dtm

<table>
<thead>
<tr>
<th>Model</th>
<th>D</th>
<th>R</th>
<th>F</th>
<th>abs rel</th>
<th>sq rel</th>
<th>rmse</th>
<th>log rmse</th>
<th>a1</th>
<th>a2</th>
<th>a3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monodepth</td>
<td>0.149</td>
<td>2.565</td>
<td>6.645</td>
<td>0.245</td>
<td>0.849</td>
<td>0.936</td>
<td>0.969</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Monodepth pp</td>
<td>x 0.114</td>
<td>1.138</td>
<td>5.452</td>
<td>0.204</td>
<td>0.859</td>
<td>0.946</td>
<td>0.977</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rdispnet_m</td>
<td>x 0.111</td>
<td>1.031</td>
<td>5.416</td>
<td>0.199</td>
<td>0.860</td>
<td>0.948</td>
<td>0.978</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rdispnet_m</td>
<td>x</td>
<td>x 0.113</td>
<td>1.114</td>
<td>5.364</td>
<td>0.195</td>
<td>0.866</td>
<td>0.951</td>
<td>0.981</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rdispnet_dtm</td>
<td>x</td>
<td>x</td>
<td>x 0.112</td>
<td>1.038</td>
<td>5.304</td>
<td>0.198</td>
<td>0.863</td>
<td>0.950</td>
<td>0.979</td>
<td></td>
</tr>
<tr>
<td>rdispnet_dtm</td>
<td>x</td>
<td>x</td>
<td>x 0.105</td>
<td>0.949</td>
<td>5.174</td>
<td>0.190</td>
<td>0.866</td>
<td>0.952</td>
<td>0.981</td>
<td></td>
</tr>
</tbody>
</table>

TAKING INTO ACCOUNT AMBIGUITIES
The network predicts the ambiguity masks for both left and right disparities. The ambiguity masks weight most terms of the total loss function consisting of photometric reconstruction (1 + SSIM), edge preserving smoothness, perceptual, ambiguity penalty, and left-right consistency terms:

\[ L = a_{reconst} + a_{dp} \frac{1}{N-1} \sum_{l=1}^{N} d_l + a_{aep} + a_{alr} + a_{lrr} \]

FULL DISPARITY ESTIMATION IN A SINGLE PASS
Ambiguity learning allows for full depth estimation in a single pass.

RESIDUAL AND DOMAIN TRANSFORMATION BLOCKS
For better feature extraction with less filter channels, and content alignment between left and right domain features

PROPOSED NETWORK ARCHITECTURE