1. Sky mask generation for RGB-D data postprocessing

As the sky region often has erroneous disparity predictions from the optical flow algorithm, we use a sky segmentation model to detect sky regions and set the disparity to the minimum disparity value (farthest away) on a map.

We use two models: 1) a scene parsing model from [16] and 2) a sky segmentation model. The scene parsing model can more robustly detect sky regions if there is any, and the sky segmentation can provide accurate sky region segmentation. The sky segmentation model adopts a densenet [4] backbone and a two-branch structure described in [15]. The model is trained on sky regions from COCO-stuff plus an internally collected dataset of 2K high-res sky images.

For a given image to be post-processed, we run the scene parsing model and the sky segmentation model to get two sky region masks. If their IOU is above a threshold of 0.75, we will use the output from the sky segmentation model as the final sky mask of the image. Otherwise, we choose the output from the scene parsing model. Example sky segmentation results are shown in Fig. 1.

![Figure 1. Examples of our sky masks. The sky regions are shown by green masks.](image)

2. Qualitative results of monodepth models

As shown in Fig. 2, Fig. 3, and Fig. 4, we note that our method has the most accurate depth discontinuities when compared to related work.
Figure 2. Additional qualitative results of single image depth prediction methods applied to different datasets.
Figure 3. Additional qualitative results of single image depth prediction methods applied to different datasets.
Figure 4. Additional qualitative results of single image depth prediction methods applied to different datasets.
3. Quantitative comparisons of monodepth models

In addition to ordinal error (see Table 1 in our main paper), we also report metric depth error scores (i.e., \( \text{rel} \) and \( \delta > 1.25 \)) in Table 1. Although we use less training data, our full model still achieves competitive results under these metrics. To demonstrate the effectiveness of our loss, we also report the scores of a baseline model with the affine-invariant loss [6]. One can observe that the model trained with our loss performs better under this setting.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Training Datasets</th>
<th>Ibims ( \delta &gt; 1.25 ) rel</th>
<th>TUM ( \delta &gt; 1.25 ) rel</th>
<th>Sintel ( \delta &gt; 1.25 ) rel</th>
<th>NYUDv2 ( \delta &gt; 1.25 ) rel</th>
<th>KITTI ( \delta &gt; 1.25 ) rel</th>
<th>DIODE ( \delta &gt; 1.25 ) rel</th>
<th>Avg. Ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIW [2]</td>
<td>DIW</td>
<td>39.30 0.232</td>
<td>37.42 0.270</td>
<td>56.21 0.405</td>
<td>36.85 0.210</td>
<td>51.45 0.306</td>
<td>42.25 0.307</td>
<td>10.00</td>
</tr>
<tr>
<td>DL [3]</td>
<td>ID</td>
<td>34.75 0.211</td>
<td>25.26 0.203</td>
<td>48.20 0.407</td>
<td>32.71 0.196</td>
<td>45.32 0.371</td>
<td>40.04 0.311</td>
<td>8.50</td>
</tr>
<tr>
<td>RW [14]</td>
<td>RW</td>
<td>30.46 0.220</td>
<td>25.16 0.200</td>
<td>45.46 0.410</td>
<td>28.86 0.178</td>
<td>31.32 0.207</td>
<td>38.27 0.320</td>
<td>6.92</td>
</tr>
<tr>
<td>MD [8]</td>
<td>MD</td>
<td>31.31 0.200</td>
<td>26.86 0.226</td>
<td>53.56 0.422</td>
<td>29.69 0.182</td>
<td>36.32 0.238</td>
<td>39.03 0.323</td>
<td>8.83</td>
</tr>
<tr>
<td>YT3D [3]</td>
<td>RW+DIW+YT3D</td>
<td>26.02 0.174</td>
<td>26.36 0.230</td>
<td>47.50 0.329</td>
<td>23.13 0.153</td>
<td>30.20 0.185</td>
<td>36.48 0.279</td>
<td>5.25</td>
</tr>
<tr>
<td>MC [7]</td>
<td>MC</td>
<td>21.55 0.152</td>
<td>26.06 0.204</td>
<td>44.85 0.476</td>
<td>23.70 0.159</td>
<td>48.02 0.280</td>
<td>39.29 0.337</td>
<td>6.58</td>
</tr>
<tr>
<td>MiDaS [6]</td>
<td>RW+MD+MV</td>
<td>21.51 0.153</td>
<td>20.44 0.201</td>
<td>39.73 0.341</td>
<td>21.38 0.148</td>
<td>26.84 0.175</td>
<td>35.12 0.296</td>
<td>1.93</td>
</tr>
<tr>
<td>Ours †</td>
<td>HRWSI</td>
<td>29.16 0.199</td>
<td>23.58 0.209</td>
<td>44.46 0.414</td>
<td>27.90 0.174</td>
<td>34.69 0.220</td>
<td>37.96 0.316</td>
<td>6.50</td>
</tr>
<tr>
<td>Ours</td>
<td>HRWSI</td>
<td>23.09 0.170</td>
<td>21.24 0.197</td>
<td>47.93 0.450</td>
<td>25.71 0.165</td>
<td>28.45 0.192</td>
<td>36.40 0.341</td>
<td>5.58</td>
</tr>
</tbody>
</table>

Table 1. Zero-shot cross-dataset evaluation. We use \( \delta > 1.25 \) and \( \text{rel} \) as our additional metrics for model evaluation. The lowest error is boldfaced and the second lowest is underlined.

4. Qualitative results of different sampling strategies

![Figure 5. Additional qualitative evaluation of different sampling strategies and the affine-invariant loss. Best viewed zoomed in on-screen. Our full model trained with a combination of the structure-guide ranking loss and the multi-scale gradient matching loss generates a globally consistent depth map with sharp depth boundaries and detailed depth structures (e.g., the basket, chair, and head).](image)

References


