Photometric Depth Super-Resolution – Supplementary Material

Bjoern Haefner*, Songyou Peng*, Alok Verma*, Yvain Quéau, and Daniel Cremers

1 ILL-POSEDNESS IN DEPTH SUPER-RESOLUTION AND SHAPE-FROM-SHADING (SECT. 2)

Figure 1 illustrates the ambiguities of depth super-resolution, and Figure 2 those of shape-from-shading.

Fig. 1: There exist infinitely many ways (dashed lines) to interpolate between low-resolution depth samples (rectangles). Our disambiguation strategy builds upon shape-from-shading applied to the companion high-resolution color image (c.f. Figure 2), in order to resurrect the fine-scale geometric details of the genuine surface (solid line).

Fig. 2: Shape-from-shading suffers from the concave / convex ambiguity: the genuine surface (solid line) and both the surfaces depicted by dashed lines produce the same image, if lit and viewed from above. We put forward low-resolution depth clues (c.f. Figure 1) for disambiguation.

2 SINGLE SHOT DEPTH SUPER-RESOLUTION USING SHAPE-FROM-SHADING (SECT. 3)

Figure 3 illustrates the synthetic datasets used for evaluation, which were generated using four different 3D-shapes (“Lucy”, “Thai Statue”, “Armadillo” and “Joyful Yell”), each of them rendered using three different albedo maps (“voronoi”, “rectcircle” and “bar”) and three different scaling factors (2, 4 and 8) for the low-resolution depth image.

Figure 4 illustrates the effect of each hyper-parameter on shape and reflectance estimation (these experiments were conducted on the “Joyful Yell” dataset, with the three proposed albedo maps and three different scaling factors).

Table 1 presents the quantitative results on all synthetic datasets, in comparison with other state-of-the-art methods.

Figure 5 presents insightful qualitative comparisons on four synthetic datasets. Note that in this visualisation we only show super-resolution using a scaling factor of 4 to make comparisons fair, as [4] only provides code to perform super-resolution at such an upsampling rate.

Figure 6 shows the qualitative comparison on real-world data captured with a RealSense D415 Camera. The input images I are shown in the main paper in Figure 2. Note that [3] seems to give good depth estimates wherever the underlying assumption (an edge in the RGB image coincides with an edge in the depth image) is met, cf. “Rucksack” dataset, but fails to result in detail preserving depth maps where reflectance is uniform or changes only slightly, as it only uses a sparse set of information from the RGB data to improve geometry, cf. “Android” and “Minion” dataset. [4] can not hallucinate surface details since it does not use the color image. [5] does a much better job at improving geometry, but it is largely overcome by shading-based super-resolution, as it uses information from a high-resolution RGB image.

Figure 7 shows four qualitative comparisons with state-of-the-art multi-view approaches on the publicly available datasets [6], [7]. “Augustus”, “Lucy” and “Relief” in column four are the results of [8], where data is captured using a PrimeSense RGB-D camera. “Gate” in the fourth column is the result of [9], where data is acquired using a Structure Sensor for an iPad. “Augustus”, “Relief” and “Gate” use an upsampling factor of 2, whereas “Lucy” provides RGB-D of [640 × 480 px²] for both I and z₀. Although our approach needs significantly less data compared to multi-view approaches, we are still able to recover fine geometry.

* Those authors contributed equally
Fig. 3: Illustration of synthetic data used for quantitative evaluation. $z^0$ with a scaling factor of 2 is shown here.

Fig. 4: Impact of the parameters $(\mu, \nu, \lambda)$ on the accuracy of the albedo and depth estimates.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Armadillo</td>
<td>bar</td>
<td>2</td>
<td>0.043643</td>
<td>38.6274</td>
<td>0.17865</td>
<td>45.6972</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0.051558</td>
<td>42.2277</td>
<td>0.17865</td>
<td>45.6972</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>0.072466</td>
<td>43.5649</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Joyful Yell</td>
<td>rectcircle</td>
<td>2</td>
<td>0.05089</td>
<td>29.1719</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0.066517</td>
<td>33.0843</td>
<td>0.084094</td>
<td>42.6111</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>0.10212</td>
<td>36.565</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Lucy</td>
<td>voronoi</td>
<td>2</td>
<td>0.057987</td>
<td>39.4714</td>
<td>0.50472</td>
<td>47.605</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0.06502</td>
<td>42.7169</td>
<td>0.50472</td>
<td>47.605</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>0.098713</td>
<td>46.4775</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Thai Statue</td>
<td></td>
<td>2</td>
<td>0.040821</td>
<td>42.8976</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>0.050296</td>
<td>47.1017</td>
<td>0.22363</td>
<td>49.9553</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>0.066515</td>
<td>49.8604</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

**TABLE 1:** Quantitative comparison between our results and three state-of-the-art methods, on all the synthetic datasets.

Close to the degree of detail of [6], [7]. Even with more complex lighting, cf. “Gate”, our approach can result in high-resolution depth maps with fine scale details and the depth does not seem to deteriorate.

Figure 8 shows additional qualitative comparison on data we captured with an Asus Xtion Pro Live camera and a scaling factor of 4, i.e. depth maps were acquired in \([320 \times 240 \text{ px}^2]\) resolution.
Fig. 5: Qualitative comparison of our results against state-of-the-art methods on four synthetic datasets using a scaling factor of 4.

Fig. 6: Qualitative comparison with other state-of-the-art methods on four real-world datasets captured with a RealSense D415 camera.

Fig. 7: Qualitative comparison against state-of-the-art multi-view approaches. The publicly available dataset [6] was captured with a PrimeSense camera, whereas [7] was acquired with a Structure Sensor for the iPad.
Fig. 8: Qualitative comparison of state-of-the-art single-view approaches on five real-world datasets captured with an Asus Xtion Pro Live camera at resolution $[1280 \times 960 \text{ px}^2]$ for the RGB images and $[320 \times 240 \text{ px}^2]$ for the low-resolution depth.
3 Depth Super-Resolution using Shape-from-Shading and Reflectance Learning (Sect. 4)

Figure 9 illustrates the lack of inter-class generalisation in learning-based methods: the approach of [11] (trained on Sintel [12] and MIT [13] datasets) performs poorly on the car image, because such an object was not present in the learning database. For the same reason, the alternative approach of [14] (trained on ShapeNet objects [15]) fails on the MIT object, and both approaches fail on the face image.

Input Image

Fig. 9: CNN-based albedo estimation applied to an object from the MIT database (first row), a car from the ShapeNet dataset (second row), and two images of human faces we generated with a renderer using ICT-3DRFE [16] database.

Figure 10 shows an example of a failure case for end-to-end learning approaches which simultaneously estimate reflectance and geometry. As soon as the scene to analyse contains unexpected deviation from the learned model, artifacts appear.

Fig. 10: Reconstruction results from SfSNet [17], which is an end-to-end deep learning based approach. It fails to account for small departures from usual face images, here fingers for example, and provides an erroneous normal estimation.

Figure 11 illustrates the rendering of synthetic human faces with extended non-directional light sources, emulating usual indoor light conditions. Geometry and reflectance are obtained from ICT-3DRFE database [16].

Fig. 11: Light sources used for rendering human faces.

Figure 12, illustrates the U-Net architecture used for albedo estimation. It essentially comprises of an initial convolution layer of kernel size 4, stride 2 and padding 1; after which there are repeated blocks of 8 ReLU-Conv-BatchNorm layers. This results in downsampling of a 512x512 resolution image to a 1x512 vector at the bottleneck of the “U”. Further, the 1D array is upsampled to input resolution with multiple ReLU-Transpose Convolution-BatchNorm layers. Dropout is also used in a few layers to allow for randomness while learning the mapping from input images to albedo maps. Finally, for the loss function we use the L1 loss, which favors sharper output compared to the L2 loss.

Figure 13 and Table 2 show several results of our approach on synthetic datasets, in comparison with two other state-of-the-art methods. We choose to compare against SIRFS [18] and Pix2Vertex [19], because the former is a completely prior-based approach with minimal learning while the latter is a deep neural networks-based approach. Our approach, which stands inbetween, inherits the strengths of both approaches. It reconstructs fine-scale details without extensive smoothing, and it can also easily reconstruct new
Fig. 12: The U-Net Architecture used for albedo estimation. The top two layers on the extreme left and right are the input and output respectively. The rest hidden layers are obtained by performing the operations mentioned for every color of arrow. The skip connections are shown as dotted lines which implies that the layers on the left are concatenated to the layers on the right.

geometries which were not present in the training database.

Figure 14 presents the qualitative comparison on real-world results with [18] and [17]. [18] attempts to provides reflectance which has minimal shading effects, but due to large number of priors on smoothness, parsimony and absolute color, the reflectance estimate is deceiving. [17] performs better than the pure prior-based approach, but is limited by the resolution \([128 \times 128 \text{ px}^2]\) and thus misses small-scale details. Our method provides high-resolution realistic albedo and depth maps directly out of the box.
<table>
<thead>
<tr>
<th>Subject (S)</th>
<th>Expression (E)</th>
<th>SF</th>
<th>[18]</th>
<th>[17]</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>RMSE</td>
<td>MAE</td>
<td>RMSE</td>
</tr>
<tr>
<td>0</td>
<td>2</td>
<td>0.1069</td>
<td>11.281</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.2281</td>
<td>23.441</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>0.1110</td>
<td>12.175</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.2285</td>
<td>30.548</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0.2025</td>
<td>28.975</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>0.1888</td>
<td>19.088</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.2837</td>
<td>30.395</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>5</td>
<td>2</td>
<td>0.3609</td>
<td>35.366</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>0.2205</td>
<td>27.965</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0.2697</td>
<td>24.925</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>0.2145</td>
<td>26.486</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0.1859</td>
<td>22.924</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0.2744</td>
<td>28.850</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>0.1044</td>
<td>12.791</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>0.1814</td>
<td>21.953</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>0.1151</td>
<td>13.656</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>0.1810</td>
<td>21.507</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Median</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>0.1590</td>
<td>15.542</td>
<td>-</td>
<td>-</td>
<td>0.1888</td>
<td>7.0528</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.6163</td>
<td>51.219</td>
<td>-</td>
<td>-</td>
<td>0.1245</td>
<td>7.9581</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>2</td>
<td>0.6791</td>
<td>15.978</td>
<td>-</td>
<td>-</td>
<td>0.1911</td>
<td>7.1025</td>
<td></td>
</tr>
</tbody>
</table>

|          |               |    |       |      |      |       |      |      |
|          | 4             | 0.6473 | 51.379 | 0.8703 | 15.0056 | 0.1365 | 8.3518 |

TABLE 2: Quantitative comparison between our method combining variational methods with machine learning, and two other state-of-the-art methods on two subjects from 3DRFE Dataset [20].
Fig. 13: Reconstruction results of the state-of-the-art and our combined variational and machine learning approach. SIRFS [18] (third column) provides smoothed out faces. Results of SfSNet [17] (fourth column) are shown as depth maps here, after integrating the output normals. Our method directly provides depth and is reliably able to do super-resolution and reconstruction of fine wrinkles of the face without any false enhancements.
Fig. 14: Qualitative comparison of our results against state-of-the-art methods, on four real-world subjects faces captured from Intel RealSense D415 camera.
4 Multi-Shot Depth Super-Resolution Using Photometric Stereo (Sect. 5)

Figure 15 illustrates the synthetic datasets used for evaluation. Same as the the experiments for shape-from-shading, four objects (“Lucy”, “Thai Statue”, “Armadillo” and “Joyful Yell”) are used to render with three different albedo maps (“ebsd”, “mandala” and “rectcircle”) and three different scaling factors (2, 4 and 8).

Figure 16 shows the impact of the number of images \( n \) on the accuracy of the albedo and depth estimates, as well as the runtime using our multi-shot photometric approach (\( \gamma = 0.01 \)). These experiments were conducted on the Joyful Yell dataset, with three different scaling factors and three different albedos.

Figure 17 illustrates the effect of the hyper-parameter \( \gamma \) on shape and reflectance estimation (\( n = 10 \)). Same as Figure 16, these experiments were conducted on the Joyful Yell dataset, with three different scaling factors and three different albedos.

Table 3 quantitatively compares various methods including ours (\( n = 20 \)). For [3], we randomly select one image out of 20. \( \gamma = 0.01 \) is used for ours in all experiments.

Figure 18 presents qualitative comparisons against three other methods on synthetic datasets shown in Figure 15.

Figure 19 shows four qualitative comparisons on real-world data captured with an Asus Xtion Pro Live camera against three other state-of-the-art methods. It can be seen that image-based depth super-resolution approach halluci- nates reflectance information as geometric information, since the underlying concept assumes to allow for larger depth variations where strong image gradients are present. Clearly, [21] suffers from the GBR problem, as geometry deteriorates in the uncabillated photometric stereo setup with a data-free depth prior, cf. “Tablet Case” and “Vase”. [3] provides better depth estimates, as it takes into account depth images from a depth sensor, but it mistakenly hallucinates albedo information, as it uses only a single image. This clearly shows the advantages of acquiring multiple images under different illumination to separate reflectance and geometry in a regularisation-free manner.

References


Fig. 15: Illustration of synthetic data used for quantitative evaluation in multi-shot depth super-resolution setup.
Fig. 16: Impact of the number of images $n$ on the accuracy of the albedo and depth estimates using our multi-shot photometric approach ($\gamma = 0.01$).

Fig. 17: Impact of the parameter $\gamma$ on the accuracy of the albedo and depth estimates using our multi-shot photometric approach ($n = 10$).
<table>
<thead>
<tr>
<th>Albedo</th>
<th>3D-shape</th>
<th>SF</th>
<th>Image Based depth SR</th>
<th>RMSE MAE</th>
<th>RMSE MAE</th>
<th>RMSE MAE</th>
<th>RMSE MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>2</td>
<td>0.031468 46.149</td>
<td>0.51996 17.325</td>
<td>0.4320 69.731</td>
<td>0.025366 2.883</td>
<td>0.037789 2.8391</td>
</tr>
<tr>
<td>mandala</td>
<td>Lucy</td>
<td>4</td>
<td>0.042467 43.5403</td>
<td>0.5427 18.5244</td>
<td>0.3948 63.7382</td>
<td>0.073928 2.9196</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thai Statue</td>
<td>8</td>
<td>0.088849 42.6184</td>
<td>0.52998 17.6553</td>
<td>0.5961 83.7853</td>
<td>0.051316 3.6142</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Joyful Yell</td>
<td>2</td>
<td>0.043889 46.4903</td>
<td>0.373797 15.7192</td>
<td>0.475 84.6189</td>
<td>0.036334 3.5842</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thai Statue</td>
<td>4</td>
<td>0.063575 44.0677</td>
<td>1.3022 75.8815</td>
<td>0.945 82.0516</td>
<td>0.051316 3.6142</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Joyful Yell</td>
<td>8</td>
<td>0.12668 42.8905</td>
<td>0.39393 15.9998</td>
<td>0.523 64.7317</td>
<td>0.084713 4.7644</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thai Statue</td>
<td>2</td>
<td>0.48487 45.0552</td>
<td>1.037 14.2423</td>
<td>0.375 70.2724</td>
<td>0.044198 3.3143</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Joyful Yell</td>
<td>4</td>
<td>0.690988 42.444</td>
<td>1.0524 13.9289</td>
<td>0.2985 55.6927</td>
<td>0.063392 3.6407</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Thai Statue</td>
<td>8</td>
<td>0.13103 40.0426</td>
<td>1.0625 14.0809</td>
<td>0.424 44.2549</td>
<td>0.1046 3.753</td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 3:** Comparison results on the various multi-shot depth super-resolution methods. $n = 20$ are used for this task. For [1], we randomly select one image out of 20. $\gamma = 0.01$ is used for Ours in all experiments.
Fig. 18: Qualitative comparison of our UPS results against state-of-the-art methods on four synthetic datasets using a scaling factor of 4.
Fig. 19: Comparison between the proposed multi-shot method and 3 state-of-the-art methods, on real-world datasets. These results confirm the conclusion of the synthetic experiments in Figure 19.