SharinGAN: Combining Synthetic and Real Data for Unsupervised Geometry Estimation
Supplementary Material

Koutilya PNVR
koutilya@terpmail.umd.edu
Hao Zhou*
hzhou@cs.umd.edu
David Jacobs
djacobs@umiacs.umd.edu
University of Maryland, College Park, MD, USA.

1. More Implementation details
   The discriminator architecture we used for this work is:
   \[
   \{CBR(n, 3, 1), CBR(2 \times n, 3, 2)\}_{n=\{32, 64, 128, 256\}},
   \{CBR(512, 3, 1), CBR(512, 3, 2)\}_{K=sets},
   \{FcBR(1024), FcBR(512), Fc(1)\},
   \]
   where, \(CBR\) (out channels, kernel size, stride) = Conv + BatchNorm2d + ReLU and \(FcBR\) (out nodes) = Fully connected + BatchNorm1D + ReLU and \(Fc\) is a fully connected layer. For face normal estimation, we do not use batchnorm layers in the discriminator. We use the value \(K = 2\) for MDE and \(K = 1\) for FNE.

Face Normal Estimation
We update the generator 3 times for each update of the discriminator, which in turn is update 5 times internally as per [1, 3]. The generator learns from a new batch each time, while the discriminator trains on a single batch for 5 times.

2. Experiments
   Monocular Depth Estimation
   We provide more qualitative results on the test set of the Make3D dataset [5]. Figure 2 further demonstrates the generalization ability of our method compared to [8].

   Face Normal Estimation
   Figure 3 depicts the qualitative results on the CelebA [4] and Synthetic [6] datasets. The translated images corresponding to synthetic and real images look similar in contrast to the MDE task (Figure 4 of the paper). We suppose that for the task of MDE, regions such as edges are domain specific, and yet hold primary task related information such as depth cues, which is why SharinGAN modifies such regions. However, for the task of FNE, we additionally predict albedo, lighting, shading and a reconstructed image along with estimating normals. This means that the primary network needs a lot of shared information across domains for good generalization to real data. Thus the SharinGAN module seems to bring everything into a shared space, making the translated images \(\{x_{fr}^{sh}, x_{fr}^{sh}\}\) look visually similar.

*Hao Zhou is currently at Amazon AWS.

Table 1: Light classification accuracy on MultiPIE dataset [2]. Training with the proposed SharinGAN also improves lighting estimation along with face normals.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>top-1%</th>
<th>top-2%</th>
<th>top-3%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SfSNet [6]</td>
<td>80.25</td>
<td>92.99</td>
<td>96.55</td>
</tr>
<tr>
<td>SharinGAN</td>
<td>81.83</td>
<td>93.88</td>
<td>96.69</td>
</tr>
</tbody>
</table>

Figure 1 depicts additional qualitative results of the predicted face normals for the test set of the Photoface dataset [7]. Our method generalizes much better to unseen data during training.
dataset that [6] used, we used our own cropping and resizing on the original MultiPIE data: centercrop 300x300 and resize to 128x128. For a fair comparison, we used the same dataset to re-evaluate the lighting performance for [6] and reported the results in Table 1. Our method not only outperforms [6] on the face normal estimation, but also on lighting estimation.

References


Figure 3: Qualitative results of our method on face normal estimation task. The translated images $x_r^{\text{nh}}, x_s^{\text{nh}}$ look reasonably similar for our task which additionally predicts albedo, lighting, shading and reconstructed image along with the face normal.