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: $\{CBR(n, 3, 1), CBR(2 * n, 3, 2)\}_{n=\{32,64,128,256\}}$, $\{CBR(512, 3, 1), CBR(512, 3, 2)\}_{Ksets}$, $\{FcBR(1024)$, $FcBR(512)$, $Fc(1)\}$, where, CBR(out channels, kernel size, stride) = Conv + BatchNorm2d + ReLU and FcBR(out nodes) = Fully conncected + 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.

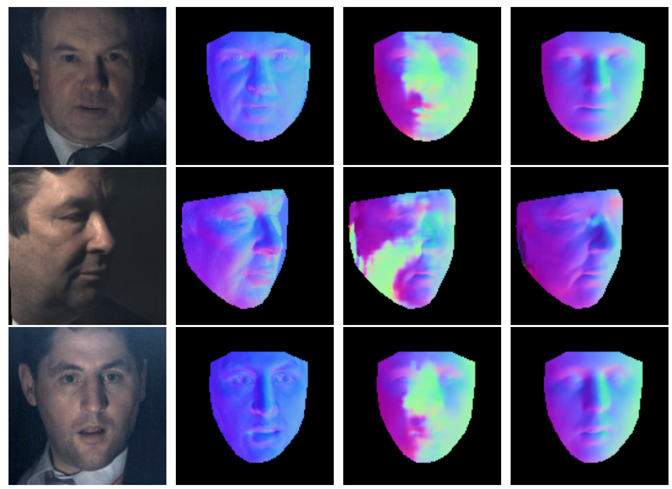
Face Normal Estimation We update the generator 3 times for each update of the discriminator, which in turn is updated 5 times internally as per [1, 3]. The generator learns from a new batch each time, while the discriminator trains on a single batch for 5 times.

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 $\{x_r^{sh}, x_s^{sh}\}$ look visually similar.

*Hao Zhou is currently at Amazon AWS.



(a) Input Image (b) GT (c) SfSNet[6] (d) SharinGAN

Figure 1: Additional Qualitative comparisons of our method with SfSNet on the examples from test set of the Photoface dataset [7]. Our method generalizes much better to unseen data during training.

Figure 1 depicts additional qualitative results of the predicted face normals for the test set of the Photoface dataset [7].

Algorithm	top-1%	top-2%	top-3%
SfSNet [6]	80.25	92.99	96.55
SharinGAN	**81.83**	**93.88**	**96.69**

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

Lighting Estimation The primary network estimates not only face normals but also lighting. We also evaluate this. Following a similar evaluation protocol as that of [6], Table 1 summarizes the light classification accuracy on the MultiPIE dataset [2]. Since we do not have the exact cropped

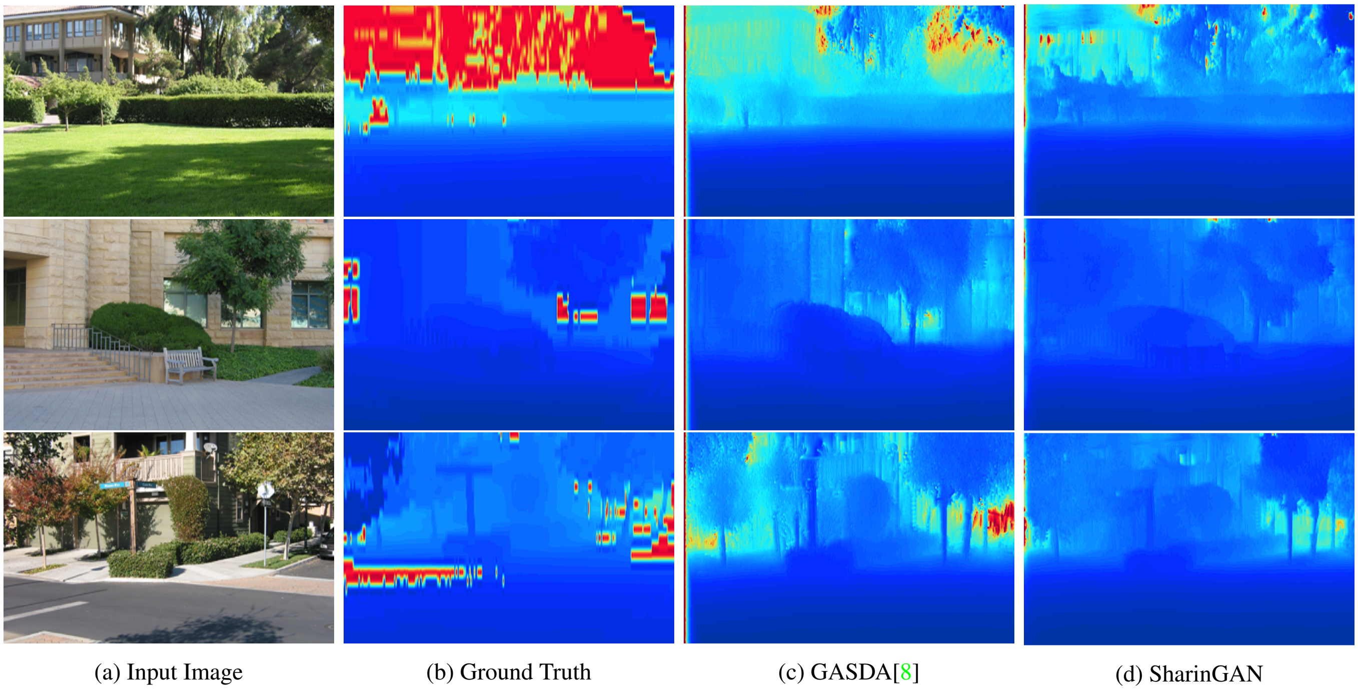


Figure 2: Additional Qualitative results on the test set of Make3D dataset [5]. Our method is able to capture better depth estimates compared to [8] for all the examples.

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

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *NeurIPS*, 2017. 1

[2] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807 – 813, 2010. Best of Automatic Face and Gesture Recognition 2008. 1

[3] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NeurIPS*. 2017. 1

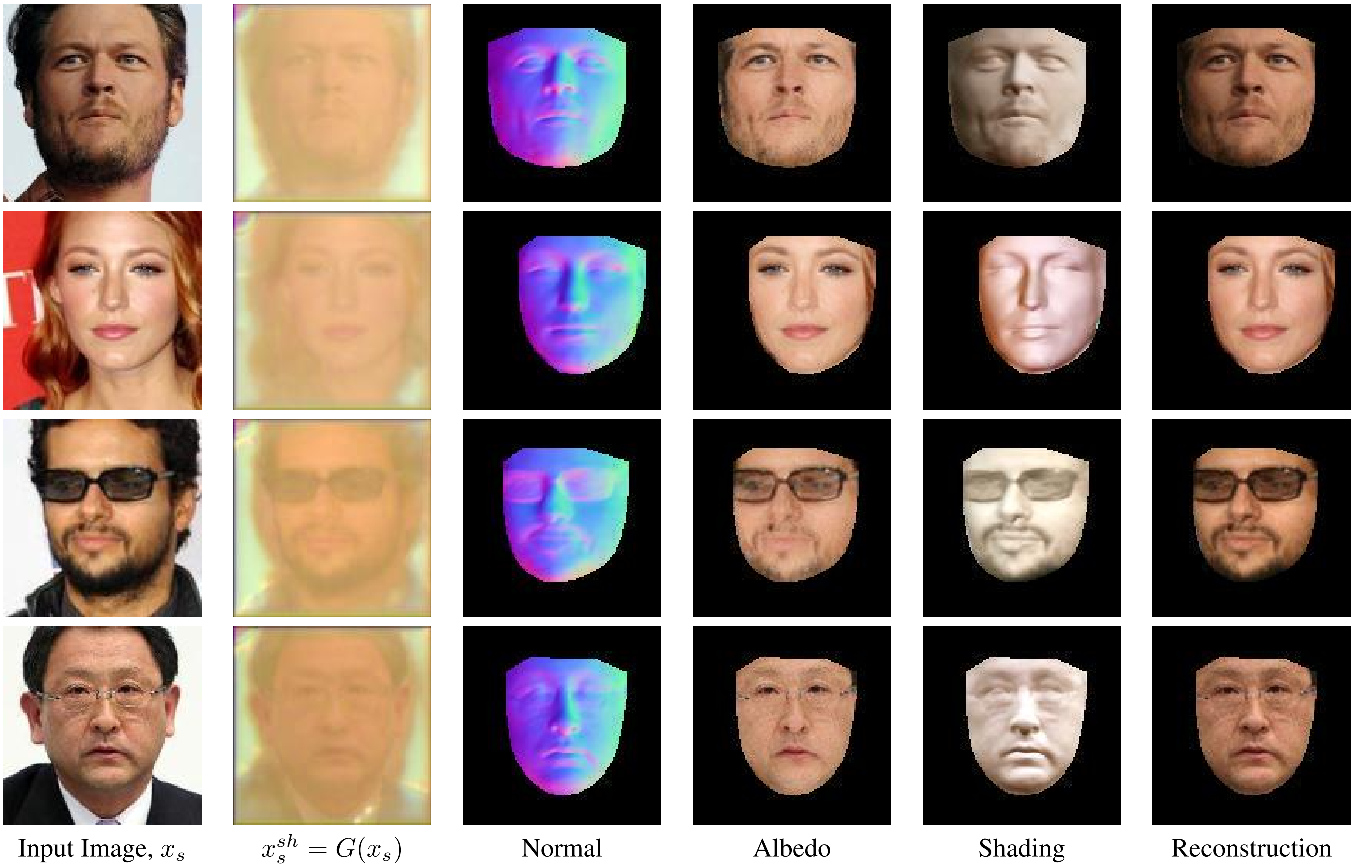
[4] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015. 1, 3

[5] Ashutosh Saxena, Min Sun, and Andrew Y. Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE Trans. PAMI*, 31(5):824–840, 2009. 1, 2

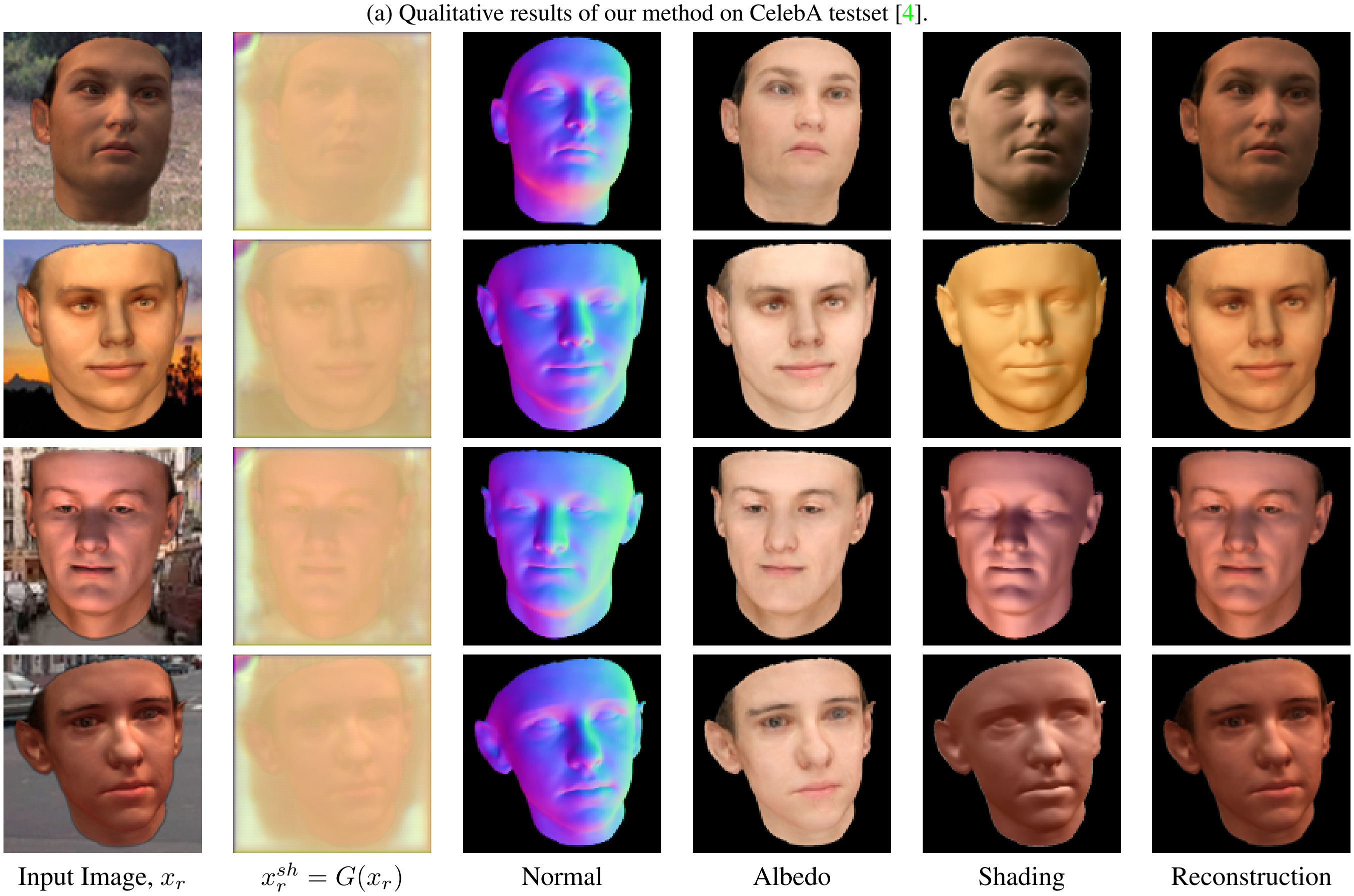
[6] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *CVPR*, 2018. 1, 2, 3

[7] Stefanos Zafeiriou, Mark F. Hansen, Gary A. Atkinson, Vasileios Argyriou, Maria Petrou, Melvyn L. Smith, and Lyndon N. Smith. The photoface database. In *CVPR Workshops*, 2011. 1

[8] Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *CVPR*, 2019. 1, 2



(a) Qualitative results of our method on CelebA testset [4].



(b) Qualitative results of our method on the synthetic data used in [6].

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