Fine-scale Surface Normal Estimation using a Single NIR Image

Youngjin Yoon*[1], Gyeongmin Choe*[1], Namil Kim[1],
Joon-Young Lee[2] and In So Kweon[1]

Korea Advanced Institute of Science and Technology (KAIST), South Korea[1]
Adobe Research, USA[2]
{yjyoon,gmchoe,nikim}@rcv.kaist.ac.kr,
jolee@adobe.com, iskweon@kaist.ac.kr

Abstract. We present surface normal estimation using a single near infrared (NIR) image. We are focusing on reconstructing fine-scale surface geometry using an image captured with an uncalibrated light source. To tackle this ill-posed problem, we adopt a generative adversarial network, which is effective in recovering sharp outputs essential for fine-scale surface normal estimation. We incorporate the angular error and an integrability constraint into the objective function of the network to make the estimated normals incorporate physical characteristics. We train and validate our network on a recent NIR dataset [1], and also evaluate the generality of our trained model by using new external datasets that are captured with a different camera under different environments.

Keywords: Shape from shading, near infrared image, generative adversarial network

1 Introduction

Estimating surface geometry is a fundamental problem in understanding the properties of an object and reconstructing its 3D information. There are two different approaches: geometric methods such as structure-from-motion and multi-view stereo, and photometric methods such as photometric stereo and shape-from-shading. The geometric methods are usually useful for metric reconstructions while the photometric methods are effective in estimating accurate per-pixel surface geometry.

Recently, with the massive use of commercial depth sensors, *e.g.*, Kinect and RealSense, many works have been proposed to enhance the depth quality of the sensors by fusing the photometric cues of the color image [2, 3] or the near infrared (NIR) image [4, 5]. Although these methods have proven their effectiveness in photometric shape estimation and have provided promising results, they rely highly on the sensors and usually require heavy computational time.

On the other hand, deep convolutional neural networks (CNN) have been broadly used for various computer vision tasks such as image classification [6, 7],

* The first and the second authors provided equal contributions to this work.

object detection [8, 9], segmentation [10, 11], and depth estimation [12, 13]. With its rich learning capability, deep CNN has shown state-of-the-art performances in many areas and has also made algorithms more practical with fast evaluation times. Lately, several works have also tried to solve depth or surface normal estimation using CNN [12, 13]. However, they have largely been focused on scene-level estimation [13] or context-aware methods [14], which generate rough surface normals and therefore they cannot generate the fine-scale surface details of the target object.

The goal of this paper is to propose a practical system that estimates fine-scale surface normals, not a scene-level structure, from an image captured with an uncalibrated light source. We solve this shape-from-shading problem by training a deep CNN on a recent NIR dataset [1]. This dataset consists of 101 objects, captured by an NIR camera with 9 different viewing directions and 12 lighting directions. It allows us to train a variety of textures such as fabrics, leaves, and papers. As shown in [1], the major benefits of using NIR images for estimating fine-scale geometry are that the albedo variation in NIR images is less prevalent than in visible band images and undesired ambient indoor lightings are filtered out. Therefore, this setting can simplify a light model and makes building a practical system easier. The proposed model for training the mapping between NIR intensity distributions and normal maps is a generative adversarial network (GAN). We design the objective function of the GAN model to consider photometric characteristics of the surface geometry by incorporating angular error and an integrability constraint. Since we train various object images captured from different lighting directions, our method estimates fine-scale surface normals without the need for calibrating the lighting direction. We verify that deep CNN is effective in handling the ill-posed, uncalibrated shape-from-shading problem without complex heuristic assumptions. Also, we evaluate the generality of our trained model by testing our own datasets, which are captured using different configurations from that of the training dataset. One example result of our method is shown in Figure 1.

The major contributions of our work are as follows:

- First work analyzing the relationship between an NIR image and its surface normal using a deep learning framework.
- Fine-scale surface normal estimation using a single NIR image where the light direction need not be calibrated.
- Suitable design of an objective function to reconstruct the fine details of a target object surface.

2 Related Work

Photometric Stereo and Shape from Shading Photometric stereo [15] is one of the well-studied methods for estimating surface normals. By taking at least 3 images captured under different lighting directions, photometric stereo can determine a unique set of surface normals of an object. Also, the usage of

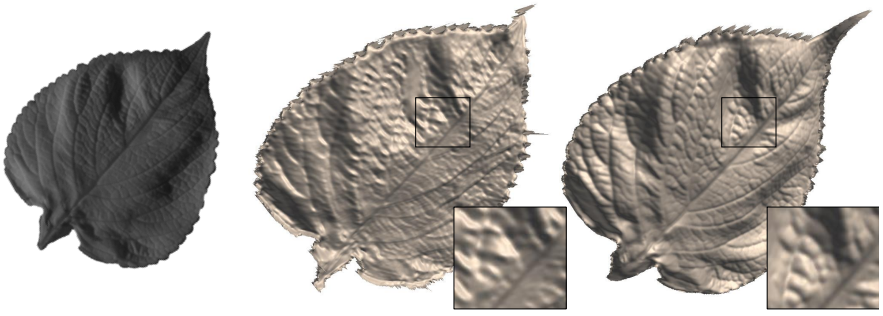

Fig. 1. Comparison of reconstruction results, left: Input NIR image, middle: Our reconstruction from a single NIR image, right: ground-truth reconstruction using NIR images captured under 12 different lighting directions.

more images makes the output increase in accuracy since it becomes an over-determined problem.

Shape from shading is a special case of photometric stereo, which predicts a shape from a single image. However, it is an ill-posed problem and needs to exploit many restrictions and constraints [16, 17]. Beginning with numerical SfS methods [18], many works have shown results based on the Lambertian BRDF assumption. Tsai*et al.* [19] use discrete approximation of surface normals. Lee and Kuo [20] estimate shape by using a triangular element surface model. We refer readers to [21] for better understanding regarding comparisons and evaluations of the classical SfS methods.

Shape from a NIR image has been recently studied in several literatures [4, 1]. They analyze the discriminative characteristics of NIR images and experimentally show the albedo (surface reflectance) simplicity in the NIR wavelength of various materials. In [4, 5], they propose the shape refinement methods using the photometric cues in NIR images. They show the high-quality shape recovery results, however they need an additional depth camera to obtain the results.

Although many conventional photometric approaches can work on NIR images and the albedo simplicity in the NIR image actually help robust estimation, estimating the surface normal from a single NIR image still have many limitations for practical uses, such as heavy computation time, heuristic assumptions, special system configuration, and the calibration of a light direction. To overcome those limitations, we study the mapping from NIR intensity distributions to surface normal vectors via a deep CNN framework. We combine a GAN [22] with the specially designed objective function. Through the adversarial training process, our network naturally encodes the photometric cues of a scene and produces fine surface normals.

Data-Driven Shape estimation There have been various studies on estimating the shape information from images via data-driven approaches. Saxena *et al.* [23] estimate depths using a discriminatively trained MRF model with multiple scales of monocular cues. Hoiem *et al.* [24] reconstruct rough surface orien-

tations of a scene by statistically modeling categories of coarse structures (*e.g.*, ground, sky and vertical). Ladicky *et al.* [25] incorporate semantic labels of a scene to predict better depth outputs.

One of the emerging directions for shape estimation is using deep CNN. In [26], Fouhey *et al.*try to discover the right primitives in a scene. In [14], Wang *et al.*explore the effectiveness of CNNs for the tasks of surface normal estimation. Although this work infers the surface normals from a single color image, it outputs scene-level rough geometries and is not suitable for object-level detailed surface reconstruction. To estimate the object shape and the material property, Rematas *et al.* [27] use the two different CNN architectures which predict surface normals directly and indirectly. The direct architecture estimates a reflectance map from an input image while the indirect architecture estimates a surface orientation map as an intermediate step towards reflectance map estimation. In [28], Liu *et al.*estimate depths from a single image using a deep CNN framework by jointly learning the unary and pairwise potentials of the CRF loss. In [29], Eigen *et al.*use a multi-scale approach which uses coarse and fine networks to estimate a better depth map.

Compared to the existing works, we focus on estimating fine-scale surface normals suing a deep CNN framework, therefore we bear in mind to design a network to produce photometrically meaningful outputs.

3 Method

3.1 Generative Adversarial Network

Generative adversarial network (GAN) [22] is a framework for training generative models which consists of two different models; a generative network G for modeling the data distribution and a discriminative network D for estimating the state of a network input. For our setup, G tries to generate a realistic surface normal map for the input NIR image and D tries to determine whether the input surface normal map is from G or from the dataset. Therefore, the generative network learns to generate more realistic images to fool the latter, while the discriminative network learns to correctly classify its input as a real image or a generated image. The two networks are simultaneously trained through a minimax optimization.

Given an input image of the discriminative network, an initial discriminative parameter θ_D is stochastically updated to correctly predict whether the input comes from a training image I or a generated image F. After that, while keeping the discriminative parameter θ_D fixed, a generative parameter θ_G is trained to produce the better quality of images, which could be misclassified by the discriminative network as real images. These procedures are repeated until they converge. This minimax objective is denoted as:

$$\min_{\Theta_G} \max_{\Theta_D} \mathbb{E}_{F \sim D_{desire}}[logD(I)] + \mathbb{E}_{Z \sim D_{input}}[log(1 - D(F))] \tag{1}$$

where D_{desire} is the distribution of images that we desired to estimate and D_{input} is that of the input domain. This objective function encourages D to be assigned

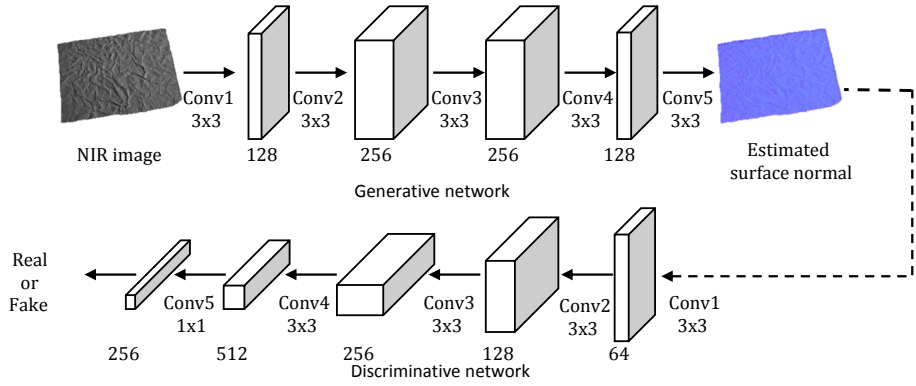

Fig. 2. Our network architecture. The proposed network produces surface normal map from a single NIR image. The generative model reconstructs surface normal map and the discriminative network predicts the probability whether the surface normal map comes from the training data or the generative model.

to the correct label for both real and generated images and make G generate a realistic output F from an input Z. In our method, both the generative and the discriminative model are based on convolutional networks. The former takes a single NIR image as an input and results in a three-dimensional normal image as an output. The latter classifies an input by using the binary cross-entropy to make the probability high when an input comes from the training data.

3.2 Deep Shape from Shading

Based on the generative adversarial network explained in Section 3.1, we modified the GAN model to be suitable for the shape-from-shading problem. Since shape-from-shading is the ill-posed problem, it is important to incorporate proper constraints to uniquely determine the right solution. Therefore, we combine angular error and integrability loss, which are shown to be effective in many conventional SfS methods, into the objective function of the generative network. Also, the existing GAN approaches typically take a random noise vector [22], pre-encoded vector [30], or an image [31, 32] as the input of their generative networks, and each generative model produces the output which lies in the same domain as its input. In this work, we apply the generative model to produce a three-dimensional normal map from a NIR image where both data lies in the different domains. Compared to the conventional SfS methods, we do not need to calibrate the lighting directions. To the best of our knowledge, our work is the first application of the adversarial training to estimate fine-scale geometry from a single NIR image.

Generative Networks We use a fully convolutional network to construct the generative network. This type of a convolutional model was recently adopted in

image restoration [33, 34] and was verified to have superior performance in the task. To keep the image size of the input and output constant, we pad zeros before the convolution operations. Through our experiments, we found that this strategy works well in reconstructing the normal map.

Our network architecture is depicted in Figure 2. We feed a 64 x 64 NIR patch to the generative network as an input. The network consists of 5 convolution layers (128-256-256-128-3 convolution filters at each of layers), each followed by ReLU except the last layer. Since the generative network is fully convolutional, the output of the network has same size as the input NIR image. We have empirically determined the number and sizes of filters for all networks.

Discriminative Networks Given the output of the generative network, a typical choice of the objectives function is the averaged L_1 or L_2 distance between ground-truth and generated output. However, such a choice has some limitations to be applied to our problem. L_2 distance produces blurry predictions because it assumes that the errors follow the Gaussian distribution. In L_1 distance, this effect could be diminished, but the estimated images would be the median of the set of equally likely intensities. We propose to add the discriminative network as a loss function with the distance metric. Recently, [31] proved that the combination of the distance, gradient and discriminative networks as a loss function provides the realistic and accurate output. Our discriminative model has a binary cross-entropy loss to make the high probability when the input is real images, and vice versa.

3.3 Training

We will explain how we iteratively train the generative model G and the discriminative model D. Let us consider a single NIR image $Z \in \{Z_1, Z_2, ..., Z_j\}$ from a training dataset and the corresponding ground truth normal map $Y \in \{Y_1, Y_2, ..., Y_j\}$. The training dataset covers various objects captured from diverse lighting directions, and we uniformly sampled the image from the dataset in terms of the balance of lighting directions.

Basically, we followed the procedure of the paper [30]. Given N paired image set, we first train D to classify the real image pair (Z, Y) into the class 1 and the generated pair $(Z, G(Z))$ into the class 0. In this step, we fixed the parameters (θ_G) of the generative network G to solely update the parameters (θ_D) of D. The objective function of the discriminative model is denoted as:

$$\mathcal{L}_D(Z, Y) = \sum_{i=1}^{N} \mathcal{D}_{bce}(Y_i, 1) + \mathcal{D}_{bce}(G(Z_i), 0), \tag{2}$$

where $\mathcal{D}_{bce}$ is the binary cross-entropy, defined as

$$\mathcal{D}_{bce}(Y_i, C) = -C_i log(Y_i) + (1 - C_i) log(1 - Y_i), \tag{3}$$

where C_i is the binary class label. We minimize the objective function so that the network outputs high probability scores for real images Y_i and low probability scores for generated images $G(Z_i)$.

After that, we keep the parameters of D fixed and train the generative model G. Many previous deep learning based image restoration and generation methods [33, 35] used the mean square error(MSE) loss function to minimize between the ground-truth images and output images. However, as studied in the conventional SfS works, estimating accurate surface normal maps requires the minimization of angular errors and the output normals satisfy the integrability constraint. Therefore, we modified the objective function of the GAN model to incorporate those photometric objective functions. By taking the objective functions, we can effectively remove angular error and estimate physically meaningful surface normals.

Specifically, to evaluate surface normal properly, we defined the objective function of our generative network as:

$$\mathcal{L}_G(Z, Y) = \sum_{i=1}^{N} \mathcal{D}_{bce}(G(Z_i), 1) + \lambda_{l_p} L_p + \lambda_{ang} L_{ang} + \lambda_{curl} L_{curl}. \tag{4}$$

Following the conventional L_1 or L_2 loss, the estimated normal map difference $\mathcal{L}_p$ is denoted as:

$$\mathcal{L}_p(Y, G(Z)) = ||Y - G(Z)||_p^p \tag{5}$$

where $p = 1$ or $p = 2$

To estimate the accuracy of photometric stereo, the angular error is often used in conventional photometric approaches because it describes more physically meaningful error than direct normal map difference. To minimize the angular error, we normalize both the estimated normals ($G(Z)$) and the ground-truth normals (Y), then simply apply the dot product between them as:

$$\mathcal{L}_{ang}(Y, G(Z)) = 1 - \langle Y, G(Z) \rangle = 1 - \frac{Y^T G(Z)}{||Y||||G(Z)||} \tag{6}$$

The angular error provides physically meaningful measures, however it averaged entire surface normals. In order to encourage the generative network to estimate photometrically correct surface normals, we also add the integrability constraint in local neighbors into the objective function, which is denoted as:

$$\mathcal{L}_{curl} = || \bigtriangledown \times G(Z) \rangle ||. \tag{7}$$

The integrability constraint enforces that the integral of normal vectors in a local closed loop must sum up to zero, meaning that angles are returned to the same height. The integrability constraint prevents a drastic change and guarantees estimated normals lie on the same surface in a local region.

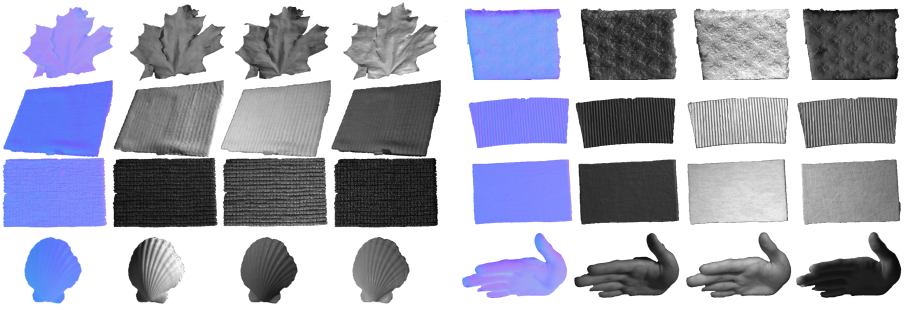

Fig. 3. Dataset [1] has various real-world object taken by 12 different lighting directions and 9 objects of view points. The leftmost is a normal map as the ground-truth and others are NIR images from different lighting directions. The Variety of lighting directions makes the same object appear vastly different.

4 Experiment

4.1 Dataset

To apply deep learning framework to our purpose, it is required to have a good quality dataset with numerous examples for training. However, most existing datasets are not large enough to train the network and are often inadequate for our tasks. Recently Choe *et al.* [1] opened a new NIR benchmark dataset, including 101 real-world objects such as fabrics, leaves and paper taken at 9 views and 12 lighting directions.

We used a pair of NIR as input and surface normal maps as target for ground truth. For fine-scale refinement, we augmented NIR images into 12 patches (64×64) within a single ground truth. For training, we used images from 91 objects and the remaining objects are for validation and test dataset. Note that we uniformly sampled validation and test samples according to the object category. When we trained the network, we normalized NIR images and normal maps to -1 and 1.

4.2 Training Parameters

We provide parameters used to train our proposed network. The configuration of the network is depicted in Table 1. Training used batches of size 32. For initializing weights, we assigned a Gaussian distribution with zero mean and a standard deviation of 0.02. We trained all experiments using the Adam optimizer [36] with momentum $\beta_1 = 0.5$. The learning rate started from 0.0002 and decreased by a factor of 0.95 every 5000 iterations. For balancing the scale of normalization, we set a hyperbolic tangent at the end of the generative network. Lastly, we

Layer	Number of filters	Filter size (w×h×ch)	Stride	Pad	Batch norm.	Activation function
Conv. 1	128	3×3×1	1	1	○	ReLU
Conv. 2	256	3×3×128	1	1	○	ReLU
Conv. 3	256	3×3×256	1	1	○	ReLU
Conv. 4	128	3×3×256	1	1	○	ReLU
Conv. 5	3	3×3×128	1	1	×	tanh

(a) Details of the Generative network.

Layer	Number of filters	Filter size (w×h×ch)	Stride	Pad	Batch norm.	Activation function
Conv. 1	64	3×3×3	2	0	×	L-ReLU
Conv. 2	128	3×3×64	2	0	○	L-ReLU
Conv. 3	256	3×3×128	2	0	○	L-ReLU
Conv. 4	512	3×3×256	2	0	○	L-ReLU
Conv. 5	256	1×1×512	1	0	×	sigmoid

(b) Details of the Discriminative network.

Table 1. Network Configuration.

used a 5×5 sliding window with 3 pixels overlap to compute the integrability. In the optimization procedure, we used a combined loss function including intensity(L_p), angular(L_{ang}), and integrability constraint(L_{curl}). Note that we did not tune the weighted parameters of each loss functions and set them with the same weights, $\lambda_p = \lambda_{ang} = \lambda_{curl} = 1$.

4.3 Experimental Result

We use Tensorflow[1] to implement and train the proposed network. The proposed network is a fully convolutional network, we apply the entire NIR image at evaluation. Computation time to estimate a surface normal is about 2 seconds with a Titan X, meanwhile the conventional shaped from shading method takes 10 minutes with Matlab implementation.

Quantitative Analysis. For the quantitative evaluation, firstly, we validate each terms of our cost functions. In this experiment, we tested our method using 3rd NIR direction among 12 lighting directions. To evaluate the performance of our method, we use three metrics; angular error, good pixel ratio and intensity error. In Table 2, all the quantitative errors are shown. Compared to case of using only intensity loss, when the angular cost function added, the performance is improved. This validates that our angular loss measures the physically meaningful error. The integrability term insures the continuity of the local normals. Although the integrability is satisfied for most of smooth surfaces, it does not

[1] https://www.tensorflow.org/

		Angular error(°) (Lower Better)		Good pixels(%) (Higher Better)			Intensity error (Abs Error)	
View points	Methods	Mean	Median	10°	15°	20°	Mean	Median
All views	L_1	16.42	16.18	17.10	38.23	72.60	0.14	0.09
	L_2	16.72	16.68	17.49	36.13	69.80	0.14	0.10
	$L_1 + L_{ang}$	15.88	15.81	19.40	37.13	73.08	0.13	0.09
	$L_2 + L_{ang}$	**15.55**	**15.30**	**20.26**	**49.84**	**74.45**	0.13	0.08
	$L_1 + L_{ang} + L_{curl}$	16.27	15.89	17.90	38.34	73.70	0.13	0.09
	$L_2 + L_{ang} + L_{curl}$	16.20	15.54	18.65	41.77	73.04	0.13	0.09
Single view	L_1	10.02	9.19	58.17	82.82	93.47	0.08	0.05
	L_2	8.76	8.37	67.14	90.97	97.44	0.07	0.05
	$L_1 + L_{ang}$	**7.35**	6.74	77.07	**93.90**	**98.59**	0.06	0.04
	$L_2 + L_{ang}$	7.70	6.82	73.36	91.91	98.36	0.07	0.04
	$L_1 + L_{ang} + L_{curl}$	10.46	8.92	57.19	80.84	91.27	0.09	0.05
	$L_2 + L_{ang} + L_{curl}$	7.52	**6.43**	**77.28**	92.41	97.59	0.06	0.04
Single view	Eigen *et al.* [13]	77.87	80.78	0.48	0.96	1.52	0.61	0.75

Table 2. Quantitative evaluation. We validate each terms of our cost functions with various error measures.

		Angular error(°) (Lower Better)		Good pixels(%) (Higher Better)			Intensity error (Abs Error)	
View points	Methods	Mean	Median	10°	15°	20°	Mean	Median
All views	L_1	4.68	3.96	90.33	97.03	98.87	0.06	0.03
	L_2	**3.34**	2.88	96.57	99.40	99.80	0.05	0.02
	$L_1 + L_{ang}$	3.47	2.98	**96.73**	**99.42**	**99.81**	0.05	0.02
	$L_2 + L_{ang}$	3.61	2.99	95.98	99.09	99.61	0.06	0.02
	$L_1 + L_{ang} + L_{curl}$	3.95	3.39	94.30	98.83	99.66	0.06	0.02
	$L_2 + L_{ang} + L_{curl}$	3.83	**2.77**	95.56	98.25	98.86	0.06	0.02
Single view	L_1	4.53	3.70	90.13	96.32	98.43	0.07	0.03
	L_2	**2.91**	**2.35**	97.29	99.27	99.69	0.06	0.03
	$L_1 + L_{ang}$	3.06	2.57	**97.55**	**99.51**	**99.83**	0.05	0.02
	$L_2 + L_{ang}$	3.61	2.73	96.39	98.82	99.31	0.06	0.02
	$L_1 + L_{ang} + L_{curl}$	3.62	3.00	95.82	99.00	99.64	0.06	0.02
	$L_2 + L_{ang} + L_{curl}$	4.23	2.51	94.80	97.21	97.95	0.07	0.02
Single view	SfS	5.09	4.14	88.25	97.19	99.27	0.06	0.03

Table 3. Quantitative evaluation on a detail map. In this evaluation, we subtract low-frequency geometry variations from the results to focus on fine-scale surface geometry.

guarantee performance improvement in some non-smooth surfaces. In our experiments, $L_2 + L_{ang}$ loss function shows the best performance for all views case, and $L_1 + L_{ang}$ achieves the lowest error for center view case. We compare our results with the conventional SfS method and we verified that our framework performs competitively. We also compare our method with the deep CNN-based surface normal estimation method [13]. Although this method estimates the surface normal, it is designed for reconstructing the scene-level low-frequency geometries and is not suitable for our purpose. We also measure errors for the single view which provides the best performance. Since extreme viewing directions are saturated or under-exposed in some cases, measuring the error of the single view results in lower errors. We found that estimated normal maps are distorted in extreme view points (error in low-frequency geometry). To evaluate the fine-scale (high-frequency) geometry, we define a detail map (M) based on the measure in [37]. This measure is computed as: $M = f(Y) + G(Z) - f(G(Z))$, where function f is smoothing function. Table 3 shows the result.

Qualitative Analysis Figure 4 and Figure 5 show the qualitative results of our network. Our network is able to estimate fine-scale textures of objects. Comparing between L_2 and $L_2 + L_{ang}$, we figure out that the angular loss provides more fine-scale textures than intensity loss. By adding the integrability constraint, the result produces a smoother surface. This demonstrates, therefore, that our network is trained to follow physical properties relevant to SfS.

4.4 Shape Estimation at Arbitrary Lighting Direction

We evaluate our network for the surface estimation with an arbitrary lighting direction. Without prior knowledge of the lighting directions, SfS becomes a more challenging problem. As shown in Figure 6, we captured several real-world objects. The glove has a complex surface geometry. Note that the bumpy surface and the stitches at the bottom are reconstructed. The cap has a 'C' letter on it and the geometry of this is reconstructed in mesh result.

5 Conclusion

In this paper, we have presented a generative adversarial network for estimating surface normal maps from a single NIR image. As far as we aware, this is the first work to estimate fine-scale surface geometry from a NIR images using a deep CNN framework. The proposed network shows competitive performance without any lighting information. We demonstrated that our photometrically-inspired object function improves the quality of surface normal estimation. We also applied our network to arbitrary NIR images which are captured under different configuration with the training dataset and have shown the promising results.

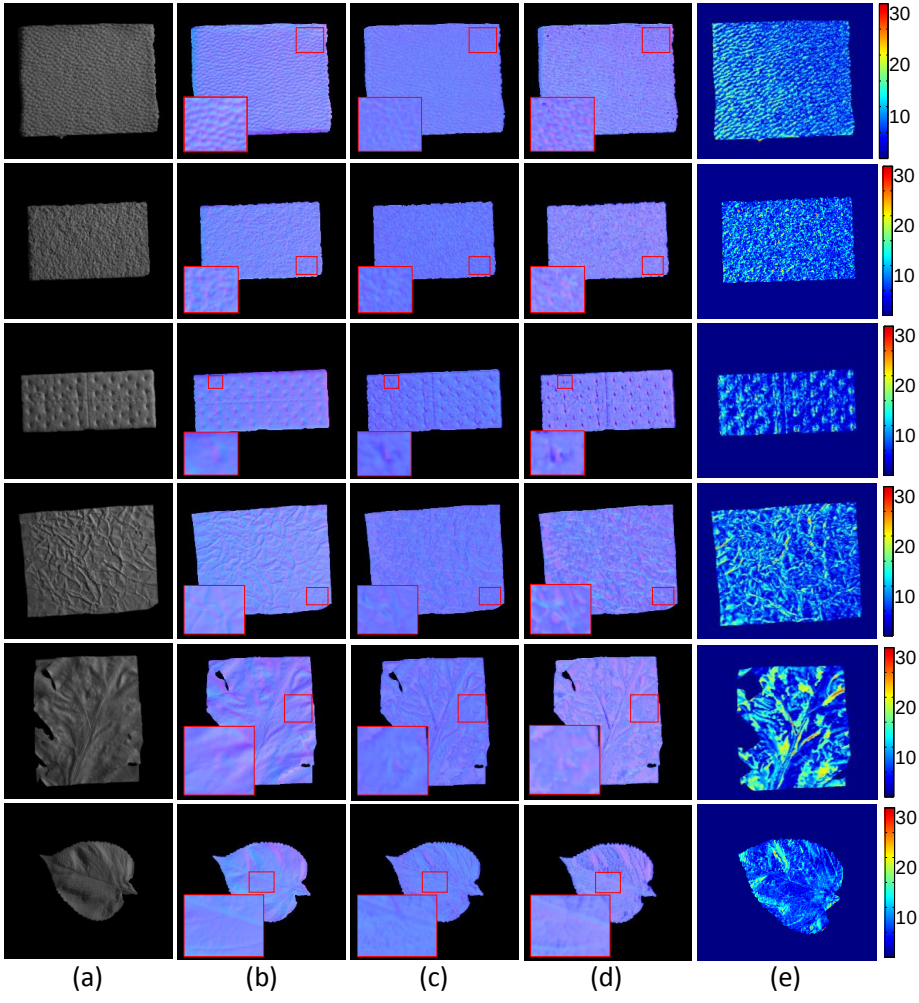

Fig. 4. Qualitative results of surface normal estimation using the proposed network. From left to right: (a) input NIR images, (b) ground-truth normal maps, (c) normal maps from L_2, (d) normal maps from $L_2 + L_{ang}$, (e) error maps of (d).

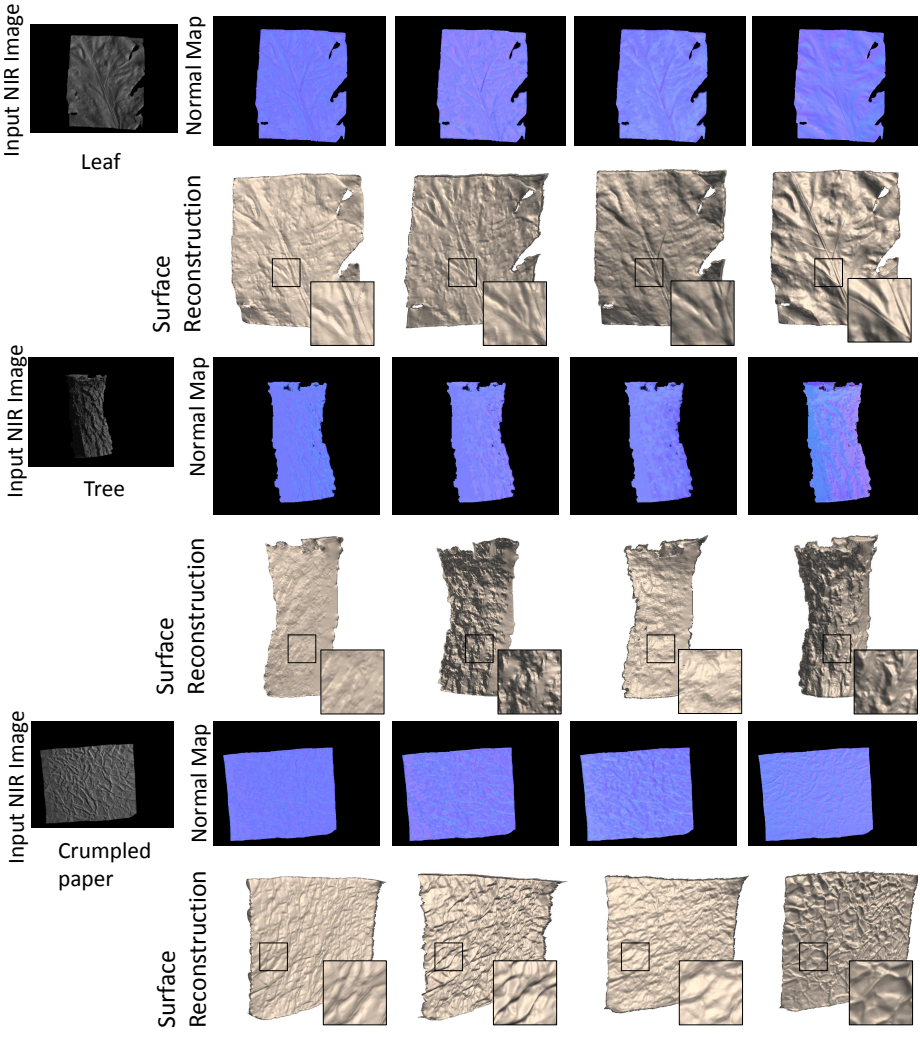

Fig. 5. Surface reconstruction results. From left to right: input, L_2, $L_2 + L_{ang}$, $L_2 + L_{ang} + L_{curl}$ and ground-truth. We compute a depth map from a surface normal map, then reconstruct a mesh. All three cases are visualized.

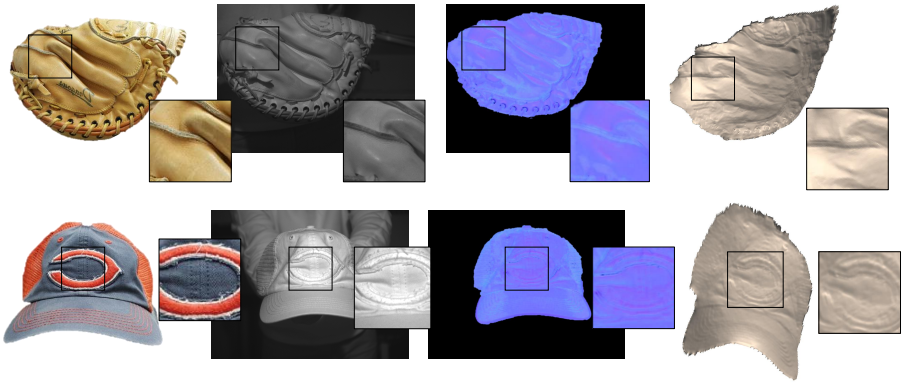

Fig. 6. Surface normal reconstruction results from an arbitrary lighting direction. From left to right, the columns show the RGB images, NIR images, estimated surface normals, and reconstructed 3D models.

Limitation and Future Work In our work, we did not take inter-reflections into account, which might produce inaccurate normals at concave regions. We also observed convexity/concavity ambiguity at some examples analogous to conventional SfS methods. Further study should be conducted to resolve this problem. Our reconstruction might suffer from distortions of low-frequency geometry as stated in Section 4. This is because we have relatively small amount of training data and we restrict our goal as estimating fine-scale geometry to train our network without overfitting to the limited training data. Despite we aimed reconstructing fine-scale surface geometry, we believe this can be further combined with various scene-level depth estimation techniques. Moreover, our network can be extended to estimate a lighting direction as well as surface normals, which can be a strong prior for conventional SfS methods.

Acknowledgements This research was supported by the Ministry of Trade, Industry & Energy and the Korea Evaluation Institute of Industrial Technology (KEIT) with the program number of 10060110.

References

1. Choe, G., Narasimhan, S.G., Kweon, I.S.: Simultaneous estimation of near ir brdf and fine-scale surface geometry. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016)
2. Han, Y., Lee, J.Y., Kweon, I.: High quality shape from a single rgb-d image under uncalibrated natural illumination. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 1617–1624
3. Yu, L.F., Yeung, S.K., Tai, Y.W., Lin, S.: Shading-based shape refinement of rgb-d images. In: Proceedings of the IEEE International Conference on Computer Vision. (2013)
4. Choe, G., Park, J., Tai, Y.W., Kweon, I.S.: Exploiting shading cues in kinect ir images for geometry refinement. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 3922–3929
5. Haque, S., Chatterjee, A., Govindu, V.: High quality photometric reconstruction using a depth camera. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 2275–2282
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. arXiv:1512.03385 (2015)
7. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
8. Yoo, D., Park, S., Lee, J.Y., Paek, A.S., Kweon, I.S.: Attentionnet: Aggregating weak directions for accurate object detection. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 2659–2667
9. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 580–587
10. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 3431–3440
11. Hong, S., Noh, H., Han, B.: Decoupled deep neural network for semi-supervised semantic segmentation. In: Advances in Neural Information Processing Systems. (2015) 1495–1503
12. Li, B., Shen, C., Dai, Y., van den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1119–1127
13. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 2650–2658
14. Wang, X., Fouhey, D., Gupta, A.: Designing deep networks for surface normal estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 539–547
15. Woodham, R.J.: Photometric method for determining surface orientation from multiple images. Optical engineering **19**(1) (1980) 191139–191139
16. Zheng, Q., Chellappa, R.: Estimation of illuminant direction, albedo, and shape from shading. In: Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on, IEEE (1991) 540–545

17. Barron, J.T., Malik, J.: Shape, albedo, and illumination from a single image of an unknown object. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 334–341
18. Ikeuchi, K., Horn, B.K.: Numerical shape from shading and occluding boundaries. Artificial intelligence **17**(1-3) (1981) 141–184
19. Ping-Sing, T., Shah, M.: Shape from shading using linear approximation. Image and Vision computing **12**(8) (1994) 487–498
20. Lee, K.M., Kuo, C.: Shape from shading with a linear triangular element surface model. IEEE Transactions on Pattern Analysis and Machine Intelligence **15**(8) (1993) 815–822
21. Zhang, R., Tsai, P.S., Cryer, J.E., Shah, M.: Shape-from-shading: a survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **21**(8) (1999) 690–706
22. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems. (2014) 2672–2680
23. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: Advances in Neural Information Processing Systems. (2005) 1161–1168
24. Hoiem, D., Efros, A.A., Hebert, M.: Automatic photo pop-up. ACM transactions on graphics (TOG) **24**(3) (2005) 577–584
25. Ladicky, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 89–96
26. Fouhey, D., Gupta, A., Hebert, M.: Data-driven 3d primitives for single image understanding. In: Proceedings of the IEEE International Conference on Computer Vision. (2013) 3392–3399
27. Rematas, K., Ritschel, T., Fritz, M., Gavves, E., Tuytelaars, T.: Deep reflectance maps. arXiv:1511.04384 (2015)
28. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 5162–5170
29. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: Advances in neural information processing systems. (2014) 2366–2374
30. Alec Radford, Luke Metz, S.C.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: arXiv:1511.06434. (2015)
31. Michael Mathieu, Camille Couprie, Y.L.: Deep multi-scale video prediction beyond mean square error. In: arXiv:1511.05440. (2015)
32. Denton, E.L., Chintala, S., Fergus, R., et al.: Deep generative image models using a laplacian pyramid of adversarial networks. In: Advances in Neural Information Processing Systems. (2015) 1486–1494
33. Jiwon Kim, Jung Kwon Lee, K.M.L.: Accurate image super-resolution using very deep convolutional networks. In: arXiv:1511.04587. (2015)
34. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. (2015)
35. John Flynn, Ivan Neulander, J.P.N.S.: Deepstereo: Learning to predict new views from the worlds imagery. In: arXiv:1506.06825. (2015)
36. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. arXiv:1412.6980 (2014)
37. Nehab, D., Rusinkiewicz, S., Davis, J., Ramamoorthi, R.: Efficiently combining positions and normals for precise 3d geometry. ACM transactions on graphics (TOG) **24**(3) (2005) 536–543