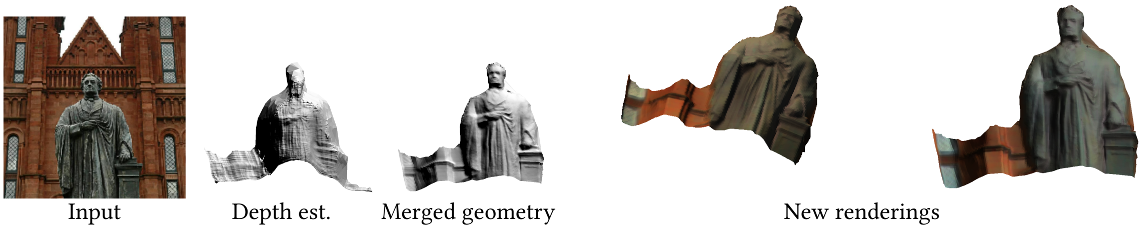
Depth estimation meets inverse rendering for single image novel view synthesis

Ye Yu
Department of Computer Science
University of York
yy1571@york.ac.uk

William A. P. Smith
Department of Computer Science
University of York
william.smith@york.ac.uk



Input Depth est. Merged geometry New renderings

Figure 1: Given a single RGB image, we perform depth estimation using the MegaDepth Network [Li and Snavely 2018c] and inverse rendering using InverseRenderNet [Yu and Smith 2019]. The geometry is then computed by merging depth and normal estimations. Finally we texture triangulated meshes with albedo estimates and re-render the scene with novel lighting and viewpoint.

ABSTRACT

In this paper we propose a method for estimating geometry, lighting and albedo from a single image of an uncontrolled outdoor scene. To do so, we combine state-of-the-art deep learning based methods for single image depth estimation and inverse rendering. The depth estimate provides coarse geometry that is refined using the inverse rendered surface normal estimates. Combined with the inverse rendered albedo map, this provides a model that can be used for novel view synthesis with both viewpoint and lighting changes. We show that, on uncontrolled outdoor images, our approach yields geometry that is qualitatively superior to that of the depth estimation network alone and that the resulting models can be re-illuminated without artefacts.

CCS CONCEPTS

• **Computing methodologies → Rendering**;

KEYWORDS

inverse rendering, depth estimation, novel view synthesis, relighting

ACM Reference format:
Ye Yu and William A. P. Smith. 2019. Depth estimation meets inverse rendering for single image novel view synthesis. In *Proceedings of the 16th ACM SIGGRAPH European Conference Visual Media Production, London, UK, Dec. 17–18 (CVMP 2019)*, 7 pages.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Over the past 5 years there has been tremendous progress on classical low level vision problems. This has been made through the use of image-to-image deep learning architectures that can be trained end-to-end to predict pixel-wise quantities. Two specific examples include depth estimation and inverse rendering.

Depth estimation from a single image is usually learnt in a supervised fashion, where the training examples come from a depth sensor [Vasiljevic et al. 2019] or multiview stereo reconstructions from image collections [Li and Snavely 2018c]. These methods are now capable of providing robust performance on real world images and the recovered depth often correctly estimates ordinal relationships between objects and components within a scene. By texture-mapping the depth map with the original image it is possible to synthesise new images with novel viewpoint. However, the depth maps estimated by these approaches do not capture finescale local detail and they do not provide any reflectance or lighting estimates. So, the output cannot be used for rendering novel illumination conditions, i.e. for *relighting*.

In a different direction, inverse rendering (or *intrinsic image decomposition*) provides a decomposition of the image into shading and reflectance, perhaps with shading further decomposed into a normal map and lighting estimate. Here, the challenge is that no existing method can provide training examples for real world scenes since inverse rendering in the wild is still an open problem. For this reason, state-of-the-art methods use self-supervision [Yu and Smith 2019]. Estimated surface normal and albedo maps are sufficient for relighting and often capture finescale, high frequency shape details. However, with no absolute geometric information, viewpoint cannot be edited and cast shadows cannot be predicted.

CVMP 2019, Dec. 17–18, London, UK
2019. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

In this paper, we seek to combine the advantages of these two divergent strands of research. Using state of the art deep CNNs for each task, we merge the coarse depth map and high quality normal map yielding a high quality geometric model. Combined with the estimated albedo map, this provides a textured, relightable 3D model that can be used for novel view synthesis.

2 RELATED WORK

Deep depth prediction. Direct estimation of shape alone using deep neural networks has attracted a lot of attention. Eigen and Fergus [2015]; Eigen et al. [2014] were the first to apply deep learning in this context. Subsequently, performance gains were obtained using improved architectures [Laina et al. 2016], post-processing with classical CRF-based methods [Liu et al. 2015; Wang et al. 2015; Xu et al. 2017] and using ordinal relationships for objects within the scenes [Chen et al. 2016; Fu et al. 2018; Li and Snavely 2018c]. Zheng et al. [2018] use synthetic images for training but improve generalisation using a synthetic-to-real transform GAN. However, all of this work requires supervision by ground truth depth. An alternative branch of methods explore using self-supervision from augmented data. For example, binocular stereo pairs can provide a supervisory signal through consistency of cross projected images [Garg et al. 2016; Godard et al. 2017; Kendall et al. 2017]. Alternatively, video data can provide a similar source of supervision [Vijayanarasimhan et al. 2017; Wang et al. 2018; Zhou et al. 2017]. Some of other work built from specific ways were proposed recently. Tulsiani et al. [2017] use multiview supervision in a ray tracing network. While all these methods take single image input, Ji et al. [2017] tackle the MVS problem itself using deep learning.

Deep intrinsic image decomposition. Intrinsic image decomposition is a partial step towards inverse rendering. It decomposes an image into reflectance (albedo) and shading but does not separate shading into shape and illumination. Even so, the lack of ground truth training data makes this a hard problem to solve with deep learning. Recent work either uses synthetic training data and supervised learning [Bi et al. 2018; Fan et al. 2018; Han et al. 2018; Lettry et al. 2018; Narihira et al. 2015] or self-supervision/unsupervised learning. Very recently, Li and Snavely [2018b] used uncontrolled time-lapse images allowing them to combine an image reconstruction loss with reflectance consistency between frames. This work was further extended using photorealistic, synthetic training data [Li and Snavely 2018a]. Ma et al. [2018] also trained on time-lapse sequences and introduced a new gradient constraint which encourage better explanations for sharp changes caused by shading or reflectance. Baslamisli et al. [2018] applied a similar gradient constraint while they used supervised training. Shelhamer et al. [2015] propose a hybrid approach where a CNN estimates a depth map which is used to constrain a classical optimisation-based intrinsic image estimation.

Deep inverse rendering. To date, this topic has not received much attention. One line of work simplifies the problem by restricting to a single object class, e.g. faces [Tewari et al. 2017], meaning that a statistical face model can constrain the geometry and reflectance estimates. This enables entirely self-supervised training. Shu et al. [2017] extend this idea with an adversarial loss. Sengupta et al. [2017] on the other hand, initialise with supervised training on synthetic data, and fine-tuned their network in an unsupervised fashion on real images. Aittala et al. [2016] restrict geometry to almost planar objects and lighting to a flash in the viewing direction under which assumptions they can obtain impressive results. Kanamori and Endo [2018] focus on human bodies and infer not only geometry and reflectance but also estimate light transport so that occlusion can be modelled. More general settings have been considered including natural illumination [Li et al. 2017]. Kulkarni et al. [2015] show how to learn latent variables that correspond to extrinsic parameters allowing image manipulation. The only prior work we are aware of that tackles the full inverse rendering problem requires direct supervision [Janner et al. 2017; Li et al. 2018; Liu et al. 2017]. Hence, it is not applicable to scene-level inverse rendering, only objects, and relies on synthetic data for training, limiting the ability of the network to generalise to real images. This drawback was addressed by InverseRenderNet [Yu and Smith 2019] which uses uncontrolled outdoor images for training. The key insight is to apply multiview stereo image collections with wide illumination variations and to use the estimated geometry to cross project images between views, essentially simulating having fixed viewpoint/varying lighting images. In addition, they proposed to estimate only albedo and normals and infer lighting directly by solving in a least squares sense, restricted to statistical subspace of natural illumination.

Merging depth and normals. Different shape estimation techniques deliver shape in different representations. For example, photometric methods naturally estimate surface orientation and hence deliver a surface normal map. Stereo methods directly compute scene depth and so deliver a depth map. Moreover, these techniques often have complimentary strengths and weaknesses, for example photometric methods often recover finescale detail but contain low frequency bias whereas multiview techniques better capture gross structure but contain high frequency noise. For this reason, there has been interest in techniques that can merge position and surface normal information. Nehab et al. [2005] proposed an efficient method based on linear least squares that can work with both depth maps and meshes. They also proposed a low pass correction procedure, similar to Zivanov et al. [2009] who pose the merging process as a nonlinear optimisation problem. These approaches were extended to multiple viewpoints by Berkiten et al. [2014] who also avoid the linearisation assumptions, though at the cost of an optimisation problem of increased complexity. In deep depth prediction, the idea of separately estimating both depth and normals within the same network has been considered [Eigen and Fergus 2015]. Here, parts of the network for predicting the two representations are shared but the geometric relationship between them is never explicitly enforced.

3 OVERVIEW

An overview of our approach is shown in Figure 2. We use a state-of-the-art network for single image depth estimation (MegaDepth [Li and Snavely 2018c]) and for inverse rendering (InverseRenderNet [Yu and Smith 2019]). Our key insight is that merging the two shape estimates from these complimentary techniques, using a variant of the method of Nehab et al. [2005], yields high quality

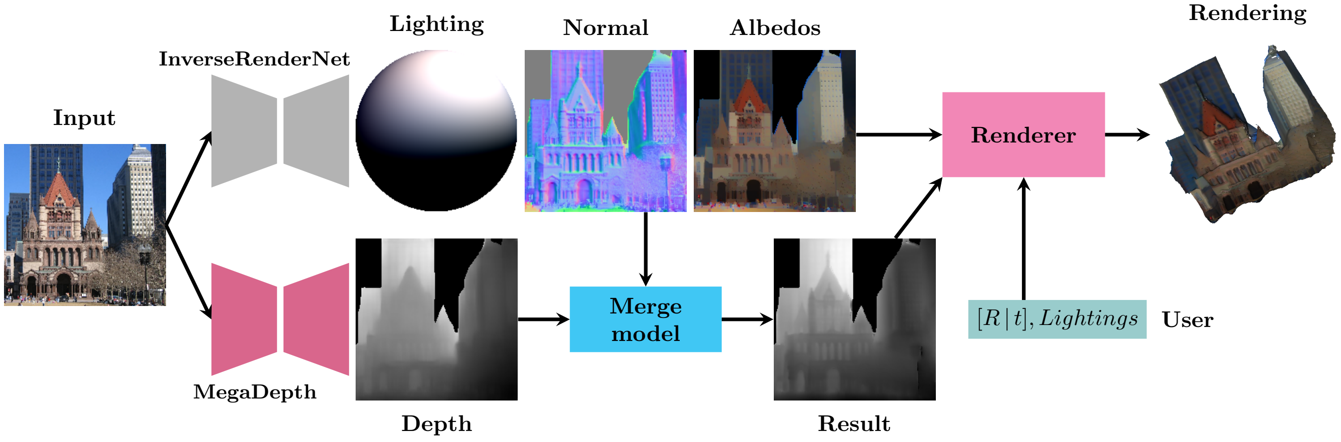


Figure 2: Overview of our proposed process for merging depth predictions and inverse rendering results for novel view synthesis.

geometry that can be re-lit using the albedo estimated by the inverse rendering network. In the remainder of the paper, we begin by introducing necessary notation, review the depth estimation and inverse rendering networks, describe the merging process and then present our results.

4 PERSPECTIVE GEOMETRY

We begin by introducing required notations and concepts from single view perspective geometry. We work in the coordinate system of the camera and parameterise the scene by the unknown depth function $Z(\mathbf{u})$, where $\mathbf{u} = (x, y)$ is a location in the image. The 3D coordinate at $\mathbf{u}$ is given by:

$$P(\mathbf{u}) = \begin{bmatrix} \frac{x-x_0}{f} Z(\mathbf{u}) \\ \frac{y-y_0}{f} Z(\mathbf{u}) \\ Z(\mathbf{u}) \end{bmatrix}, \tag{1}$$

where f is the focal length of the camera and (x_0, y_0) is the principal point.

The tangent vectors to the surface are given by:

$$\frac{\partial P(\mathbf{u})}{\partial x} = \begin{bmatrix} -\frac{1}{f}\left((x-x_0)\frac{\partial Z(\mathbf{u})}{\partial x} + Z(\mathbf{u})\right) \\ -\frac{1}{f}(y-y_0)\frac{\partial Z(\mathbf{u})}{\partial x} \\ \frac{\partial Z(\mathbf{u})}{\partial x} \end{bmatrix}, \tag{2}$$

$$\frac{\partial P(\mathbf{u})}{\partial y} = \begin{bmatrix} -\frac{1}{f}(x-x_0)\frac{\partial Z(\mathbf{u})}{\partial y} \\ -\frac{1}{f}\left((y-y_0)\frac{\partial Z(\mathbf{u})}{\partial y} + Z(\mathbf{u})\right) \\ \frac{\partial Z(\mathbf{u})}{\partial y} \end{bmatrix}. \tag{3}$$

Note that these tangent vectors are linear functions of the surface depth.

The direction of the outward pointing surface normal is defined as the cross product of the tangent vectors, themselves the partial derivatives of the position function:

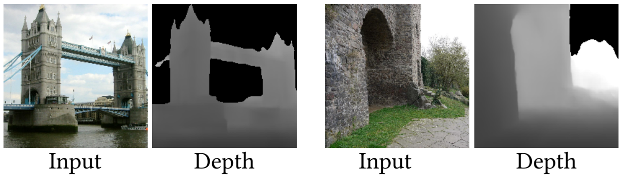


Figure 3: Sample output from MegaDepth [Li and Snavely 2018c]. Dark is closer to viewer.

$$\mathbf{n}(\mathbf{u}) = \frac{\partial P(\mathbf{u})}{\partial x} \times \frac{\partial P(\mathbf{u})}{\partial y} = k \begin{bmatrix} -f\frac{\partial Z(\mathbf{u})}{\partial x} \\ -f\frac{\partial Z(\mathbf{u})}{\partial y} \\ (x-x_0)\frac{\partial Z(\mathbf{u})}{\partial x} + (y-y_0)\frac{\partial Z(\mathbf{u})}{\partial y} + Z(\mathbf{u}) \end{bmatrix}, \tag{4}$$

where k is an arbitrary scale factor. Note that the magnitude of the surface normal vector is not important, only its direction. Also note that linearly scaling the depth function does not change the direction of the surface normal vector. We denote by $\bar{\mathbf{n}}(\mathbf{u}) = \mathbf{n}(\mathbf{u})/\|\mathbf{n}(\mathbf{u})\|$, the unit length surface normal.

5 SINGLE IMAGE DEPTH ESTIMATION

The goal of single image depth estimation is to compute a depth value, $Z(\mathbf{u})$, for each pixel in the image, collectively known as a depth map. Note that from (1), this cannot be transformed into positions in world units without knowing camera calibration information. In low accuracy applications, the principal point is usually assumed to be the centre of the image and we make this assumption. The focal length however is typically unknown. However, often a good estimate can be made from image metadata and a database of sensor sizes. We take this approach allowing us to assume the focal length is known. However, estimating absolute depth from a monocular image is highly ambiguous. For this reason, depth prediction networks usually estimate depth only up to an unknown global

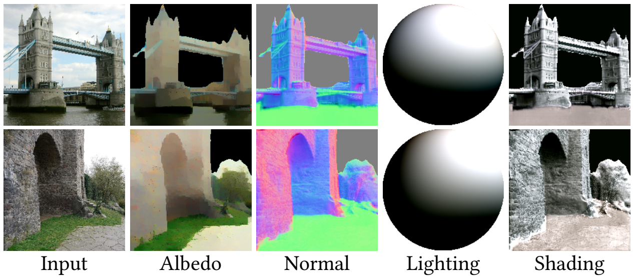


Figure 4: Sample outputs from InverseRenderNet [Yu and Smith 2019].

scale $s > 0$ such that $\tilde{Z}(\mathbf{u}) = sZ(\mathbf{u})$. This does not affect the surface normals and hence the merging process described later. However, for the purposes of computing scene geometry for rendering an absolute scale must be chosen and we leave this as a parameter for the user to select.

We use the depth predication network of Li and Snavely [2018c] (known as MegaDepth). This is trained in a supervised fashion using depth maps computed from scene geometry recovered using multiview stereo applied to community photo collections. The supervised training loss function is scale invariant so that the output scale of the depth map is arbitrary as discussed above. We show sample output in Figure 3.

6 SINGLE IMAGE INVERSE RENDERING

We use the inverse rendering network of Yu and Smith [2019] (referred to as InverseRenderNet). This is an image to image network that predicts surface normal and albedo maps from which optimal lighting (spherical harmonic coefficients) can be computed in closed form using linear least squares. The network is trained using a rather complex combination of partial direct supervision and self supervision. Again, the training data is based on multiview stereo models. This geometry provides direct supervision for the surface normal estimation and enables cross projection between views. This allows a loss measuring the consistency of photometric invariants (albedo) between views with different lighting. There is also a self supervised rendering loss which requires that the rerendered image is close to the original and various priors to ensure stable training.

InverseRenderNet directly predicts two components of the surface normal, $n_1(\mathbf{u})$ and $n_2(\mathbf{u})$, such that the surface normal direction is given by $\mathbf{n}_{\text{IRN}}(\mathbf{u}) = [n_1(\mathbf{u}), n_2(\mathbf{u}), 1]^T$. This avoids the need for any camera calibration information in the subsequent use of the surface normal vectors for rendering. We show sample output in Figure 4.

7 MERGING DEPTH AND NORMALS

Our process for merging surface normal and depth map is based on that of Nehab et al. [2005]. However, we include the centre of projection in our derivation and provide an explicit matrix formulation, decomposing the original formulation in terms of matrices for computing tangent vectors, numerical derivatives and dot products. This makes reproducing our approach much more straightforward and we make a Matlab implementation publicly available (https://github.com/waps101/MergePositionNormals). Unlike Nehab et al. [2005], we do not begin by removing low frequency bias from the estimated surface normals. In practice, we do not find that the surface normals delivered by InverseRenderNet are subject to low frequency bias in the same way that normals from photometric methods could be. We believe this is because of the direct normal map supervision during training.

The approach of Nehab et al. [2005] is to form a linear system of equations in the merged depth that seeks to satisfy two constraints. The first seeks to preserve the gross structure of the original estimated depth map by penalising in a least squares sense any deviations. The second seeks to encourage the surface normals of the refined depth map to align with the target normals. The key observation is that this second error function can be formulated to be linear in the surface depth. This is achieved by encouraging the tangent vectors of the refined surface to be perpendicular to the target normals, i.e., have a zero dot product.

First, we extend (2) and (3) to the whole image. Consider an image with N foreground pixels whose unknown depth values are vectorised in $\mathbf{z} \in \mathbb{R}^N$. We define matrices:

$$\mathbf{T}_x = \begin{bmatrix} \frac{-1}{f}\mathbf{X} & \frac{-1}{f}\mathbf{I} \\ \frac{-1}{f}\mathbf{Y} & \mathbf{0}_{N\times N} \\ \mathbf{I} & \mathbf{0}_{N\times N} \end{bmatrix} \begin{bmatrix} \mathbf{D}_x \\ \mathbf{I} \end{bmatrix}, \quad \mathbf{T}_y = \begin{bmatrix} \frac{-1}{f}\mathbf{X} & \mathbf{0}_{N\times N} \\ \frac{-1}{f}\mathbf{Y} & \frac{-1}{f}\mathbf{I} \\ \mathbf{I} & \mathbf{0}_{N\times N} \end{bmatrix} \begin{bmatrix} \mathbf{D}_y \\ \mathbf{I} \end{bmatrix}, \quad (5)$$

such that tangent vectors for the whole image concatenated into a vector can be computed by post-multiplication with the vector of depth values:

$$\mathbf{T}_x\mathbf{z} = \text{vec}\left(\begin{bmatrix} \frac{\partial P(\mathbf{u}_1)}{\partial x} & \dots & \frac{\partial P(\mathbf{u}_N)}{\partial x} \end{bmatrix}^T\right), \quad (6)$$

$$\mathbf{T}_y\mathbf{z} = \text{vec}\left(\begin{bmatrix} \frac{\partial P(\mathbf{u}_1)}{\partial y} & \dots & \frac{\partial P(\mathbf{u}_N)}{\partial y} \end{bmatrix}^T\right), \quad (7)$$

where $\mathbf{I}$ is the $N \times N$ identity matrix and $\mathbf{X} = \text{diag}(x_1 - x_0, \dots, x_N - x_0)$ and $\mathbf{Y} = \text{diag}(y_1 - y_0, \dots, y_N - y_0)$. $\mathbf{D}_x, \mathbf{D}_y \in \mathbb{R}^{N\times N}$ compute numerical approximations to the derivative of Z in the x and y directions respectively. In practice, we use forward finite differences. Hence $\mathbf{D}_x, \mathbf{D}_y$ have two non-zero values per row.

We are now ready to write the linear system of equations:

$$\begin{bmatrix} \lambda\mathbf{I} \\ \mathbf{N}\mathbf{T}_x \\ \mathbf{N}\mathbf{T}_y \end{bmatrix} \mathbf{z} = \begin{bmatrix} \mathbf{z}_{\text{MD}} \\ \mathbf{0}_{2N\times 1} \end{bmatrix}, \quad (8)$$

where $\mathbf{z}_{\text{MD}} \in \mathbb{R}^N$ contains the coarse depth estimates delivered by MegaDepth and

$$\mathbf{N} = \begin{bmatrix} \text{diag}\left(n^x_{\text{IRN}}(\mathbf{u}_1), \dots, n^x_{\text{IRN}}(\mathbf{u}_N)\right) \\ \text{diag}\left(n^y_{\text{IRN}}(\mathbf{u}_1), \dots, n^y_{\text{IRN}}(\mathbf{u}_N)\right) \\ \text{diag}\left(n^z_{\text{IRN}}(\mathbf{u}_1), \dots, n^z_{\text{IRN}}(\mathbf{u}_N)\right) \end{bmatrix}^T, \quad (9)$$

is a $N \times 3N$ matrix formed by concatenating diagonal matrices containing the x, y and z components of the target normals delivered by InverseRenderNet. Hence, each row contains one of the target surface normal vectors. The linear system of equations in (8) is large but sparse and can be solved efficiently. We do so using

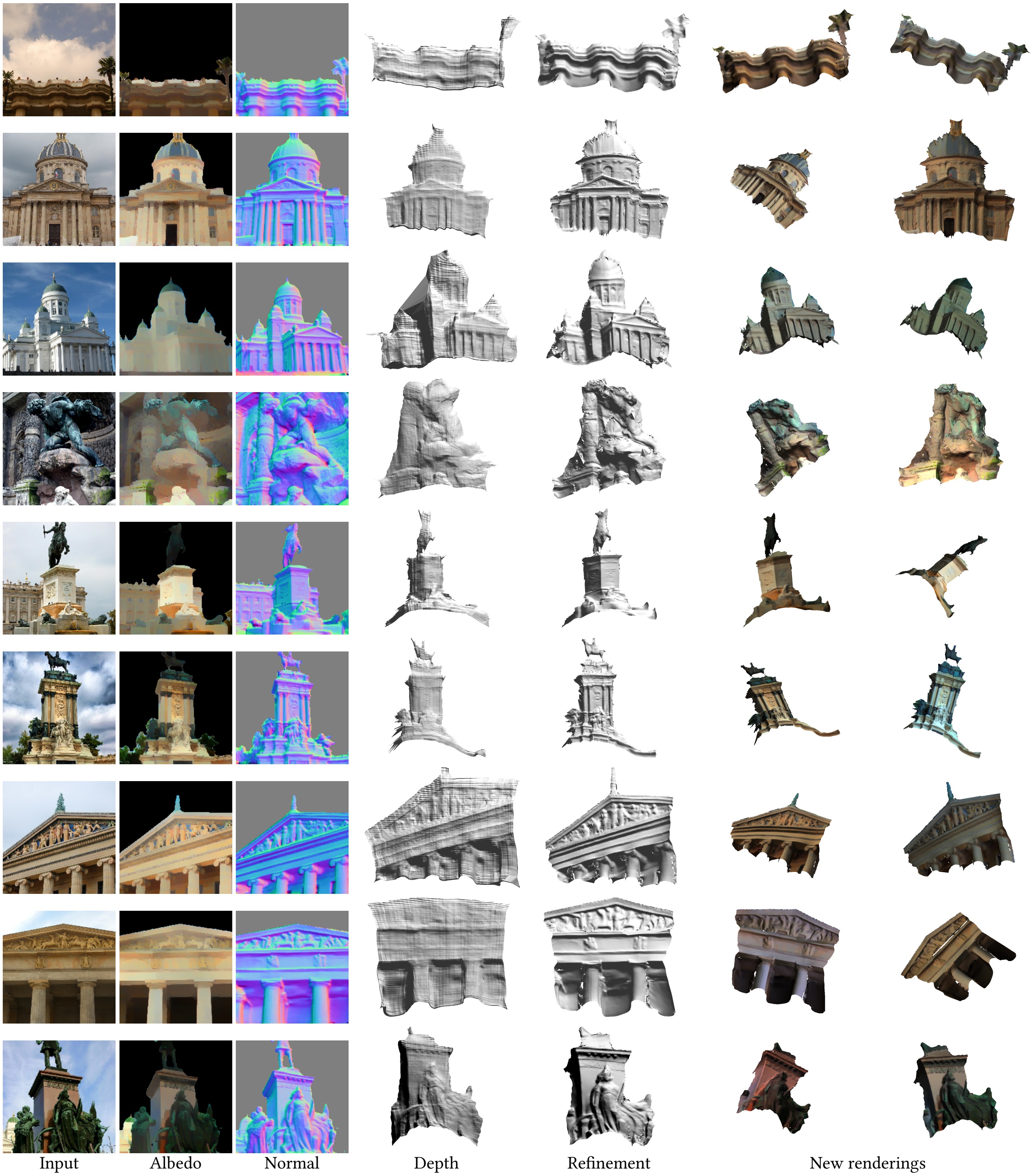


Figure 5: Results of applying our method to images from the MegaDepth dataset [Li and Snavely 2018c]. Col. 1: input image. Col. 2 and 3: albedo and surface normal maps estimated by InverseRenderNet [Yu and Smith 2019]. Col. 4: rendering of the geometry provided by the depth prediction network. Col. 5: refined geometry after merging with the surface normals. Col. 6 and 7: novel views under two different lighting conditions.

a QR solver as implemented in Matlab's `mldivide` function. The parameter λ balances the influence of the two constraints. When λ is large, the refined depth will stay close to the original estimate. When it is small, the surface normals have greater influence.

8 EXPERIMENTS

The overview in Figure 2 shows an example output from MegaDepth and InverseRenderNet. The striking feature is how the depth prediction is significantly improved when merged with the surface normal estimates. We show further results in Figure 5 for uncontrolled outdoor scenes. Here, we render the geometry as a mesh (by triangulating the depth map). Again, it is evident that the geometry is much improved, adding detail but also correcting gross structures. The textured models provide plausible appearance under large illumination and viewpoint changes.

9 CONCLUSIONS

In this paper we have presented an approach to combine the benefits of two recent advances in computer vision: deep single image depth estimation and deep single image inverse rendering. Using a simple techniques based on classical geometry and optimisation to merge the depth and normal maps, we obtain high quality geometry from a single image under demanding conditions. These models can be used for novel view synthesis.

We believe this is only a first step in an exciting direction. The most obvious future work is to train the two networks simultaneously. Since the merging process involves only the solution of a linear least squares system, this could be done within the network during training. Alternatively, one could consider estimating only depth but using inverse rendering losses such that the surface normals of the depth map better capture high frequency detail. This would require a method for estimating calibration parameters however. Another obvious direction would be to introduce adversarial networks to enhance the quality of the synthesised views. Clearly, our results lack background and sky which a GAN may be able to synthesise realistically, while retaining semantic control over the pose and lighting of the scene.

ACKNOWLEDGMENTS

W. Smith is supported by a Royal Academy of Engineering/The Leverhulme Trust Senior Research Fellowship. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp used for this research.

REFERENCES

Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2016. Reflectance modeling by neural texture synthesis. *ACM Transactions on Graphics (TOG)* 35, 4 (2016), 65.

Anil S. Baslamisli, Hoang-An Le, and Theo Gevers. 2018. CNN Based Learning Using Reflection and Retinex Models for Intrinsic Image Decomposition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sema Berkiten, Xinyi Fan, and Szymon Rusinkiewicz. 2014. Merge2-3D: Combining Multiple Normal Maps with 3D Surfaces. *IEEE International Conference on 3D Vision (3DV)* (Dec. 2014), 440–447.

Sai Bi, Nima Khademi Kalantari, and Ravi Ramamoorthi. 2018. Deep Hybrid Real and Synthetic Training for Intrinsic Decomposition. In *Eurographics Symposium on Rendering - Experimental Ideas & Implementations*, Wenzel Jakob and Toshiya Hachisuka (Eds.). The Eurographics Association. https://doi.org/10.2312/sre.20181172

Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. 2016. Single-image depth perception in the wild. In *Advances in Neural Information Processing Systems*. 730–738.

David Eigen and Rob Fergus. 2015. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE International Conference on Computer Vision*. 2650–2658.

David Eigen, Christian Puhrsch, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*. 2366–2374.

Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. 2018. Revisiting deep intrinsic image decompositions. In *Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8944–8952.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. 2018. Deep Ordinal Regression Network for Monocular Depth Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2002–2011.

Ravi Garg, Vijay Kumar BG, Gustavo Carneiro, and Ian Reid. 2016. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European Conference on Computer Vision*. Springer, 740–756.

Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. 2017. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, Vol. 2. 7.

Guangyun Han, Xiaohua Xie, Jianhuang Lai, and Wei-Shi Zheng. 2018. Learning an Intrinsic Image Decomposer Using Synthesized RGB-D Dataset. *IEEE Signal Processing Letters* 25, 6 (2018), 753–757.

Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Josh Tenenbaum. 2017. Self-supervised intrinsic image decomposition. In *Advances in Neural Information Processing Systems*. 5936–5946.

Mengqi Ji, Juergen Gall, Haitian Zheng, Yebin Liu, and Lu Fang. 2017. SurfaceNet: an end-to-end 3d neural network for multiview stereopsis. *arXiv preprint arXiv:1708.01749* (2017).

Yoshihiro Kanamori and Yuki Endo. 2018. Relighting humans: occlusion-aware inverse rendering for full-body human images. *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia)* 37, 6 (2018), 270.

Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. 2017. End-to-end learning of geometry and context for deep stereo regression. *CoRR, vol. abs/1703.04309* (2017).

Tejas D Kulkarni, William F Whitney, Pushmeet Kohli, and Josh Tenenbaum. 2015. Deep convolutional inverse graphics network. In *Advances in neural information processing systems*. 2539–2547.

Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. 2016. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 239–248.

Louis Lettry, Kenneth Vanhoey, and Luc Van Gool. 2018. DARN: a deep adversarial residual network for intrinsic image decomposition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1359–1367.

Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2017. Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 45.

Zhengqi Li and Noah Snavely. 2018a. CGIntrinsics: Better Intrinsic Image Decomposition through Physically-Based Rendering. In *European Conference on Computer Vision (ECCV)*.

Zhengqi Li and Noah Snavely. 2018b. Learning Intrinsic Image Decomposition from Watching the World. In *Computer Vision and Pattern Recognition (CVPR)*.

Zhengqi Li and Noah Snavely. 2018c. MegaDepth: Learning Single-View Depth Prediction from Internet Photos. In *Computer Vision and Pattern Recognition (CVPR)*.

Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. 2018. Learning to reconstruct shape and spatially-varying reflectance from a single image. In *SIGGRAPH Asia 2018 Technical Papers*. ACM, 269.

Fayao Liu, Chunhua Shen, and Guosheng Lin. 2015. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5162–5170.

Guilin Liu, Duygu Ceylan, Ersin Yumer, Jimei Yang, and Jyh-Ming Lien. 2017. Material editing using a physically based rendering network. In *Proceedings of the IEEE International Conference on Computer Vision*. 2261–2269.

Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. 2018. Single Image Intrinsic Decomposition without a Single Intrinsic Image. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 201–217.

Takuya Narihira, Michael Maire, and Stella X Yu. 2015. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *Proceedings of the IEEE international conference on computer vision*. 2992–2992.

Diego Nehab, Szymon Rusinkiewicz, James Davis, and Ravi Ramamoorthi. 2005. Efficiently combining positions and normals for precise 3D geometry. *ACM Transactions on Graphics (Proceedings of SIGGRAPH)* 24, 3 (2005), 536–543.

Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. 2017. SfSNet: Learning Shape, Reflectance and Illuminance of Faces âĂŸin the wildâĂŹ. *arXiv preprint arXiv:1712.01261* (2017).

Evan Shelhamer, Jonathan T Barron, and Trevor Darrell. 2015. Scene intrinsics and depth from a single image. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 37–44.

Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. 2017. Neural face editing with intrinsic image disentangling. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*. IEEE, 5444–5453.

Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian Theobalt. 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, Vol. 2. 5.

Shubham Tulsiani, Tinghui Zhou, Alexei A Efros, and Jitendra Malik. 2017. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, Vol. 1. 3.

Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. 2019. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *CoRR* abs/1908.00463 (2019). http://arxiv.org/abs/1908.00463

Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. 2017. Sfm-net: Learning of structure and motion from video. *arXiv preprint arXiv:1704.07804* (2017).

Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. 2018. Learning Depth from Monocular Videos using Direct Methods. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2022–2030.

Peng Wang, Xiaohui Shen, Zhe Lin, Scott Cohen, Brian Price, and Alan L Yuille. 2015. Towards unified depth and semantic prediction from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2800–2809.

Dan Xu, Elisa Ricci, Wanli Ouyang, Xiaogang Wang, and Nicu Sebe. 2017. Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In *Proceedings of CVPR*, Vol. 1.

Ye Yu and William AP Smith. 2019. InverseRenderNet: Learning single image inverse rendering. In *Proc. CVPR*. 3155–3164.

Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2018. T2Net: Synthetic-to-Realistic Translation for Solving Single-Image Depth Estimation Tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 767–783.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. 2017. Unsupervised learning of depth and ego-motion from video. In *CVPR*, Vol. 2. 7.

Jasenko Zivanov, Pascal Paysan, and Thomas Vetter. 2009. Facial normal map capture using four lights–an effective and inexpensive method of capturing the fine scale detail of human faces using four point lights. In *Proc. GRAPP*.