Single-Shot Analysis of Refractive Shape Using Convolutional Neural Networks

Stets, Jonathan Dyssel; Li, Zhengqin; Frisvad, Jeppe Revall; Chandraker, Manmohan

Published in:
Proceedings of IEEE Winter Conference on Applications of Computer Vision

Link to article, DOI:
10.1109/WACV.2019.00111

Publication date:
2019

Document Version
Peer reviewed version

Single-Shot Analysis of Refractive Shape Using Convolutional Neural Networks

Jonathan Dyssel Stets
Technical University of Denmark
stet@dtu.dk

Zhengqin Li
University of California, San Diego
zhl378@eng.ucsd.edu

Jeppe Revall Frisvad
Technical University of Denmark
jerf@dtu.dk

Manmohan Chandraker
University of California, San Diego
mkchandraker@eng.ucsd.edu

Abstract

The appearance of a transparent object is determined by a combination of refraction and reflection, as governed by a complex function of its shape as well as the surrounding environment. Prior works on 3D reconstruction have largely ignored transparent objects due to this challenge, yet they occur frequently in real-world scenes. This paper presents an approach to estimate depths and normals for transparent objects using a single image acquired under a distant but otherwise arbitrary environment map. In particular, we use a deep convolutional neural network (CNN) for this task. Unlike opaque objects, it is challenging to acquire ground truth training data for refractive objects, thus, we propose to use a large-scale synthetic dataset. To accurately capture the image formation process, we use a physically-based renderer. We demonstrate that a CNN trained on our dataset learns to reconstruct shape and estimate segmentation boundaries for transparent objects using a single image, while also achieving generalization to real images at test time. In experiments, we extensively study the properties of our dataset and compare to baselines demonstrating its utility.

1. Introduction

Light refracts and reflects at an interface between two materials. For transparent objects composed of material such as glass, ice or some plastics, scattering and absorption are negligible, so that light passes through the material and we observe the effect of surface refraction and reflections. Thereby, the appearance of a refractive object is a distorted image of its surroundings. This means that we cannot reconstruct the shape of a refractive object using a local model, which makes solid refractive objects particularly challenging to handle in computer vision. We pick up this challenge and consider shape analysis of refractive objects made of homogeneous glass with a smooth surface, that is, glass without air bubbles, significant absorption or surface scratches. Such glass objects are common in human-made environments (windows, glasses, drinking cups, plastic containers and so on), which makes it desirable to design a computer vision system able to determine their geometric properties based on a single image.

The physics of refraction and reflection observed for transparent objects is well-understood. Refraction occurs when light passes from one medium to another, the angle of refraction depends on the propagation speed of the light wave in the two media as described by Snell's law, while the relative fractions of reflection and refraction are described by Fresnel's equations. The amount of reflection increases with greater angle of incidence. When light is incident on an optically thinner medium at a grazing angle less than the critical angle, total internal reflection occurs. Due to these different types of interaction, the light path undergoes significant deviation from a straight line when passing through a glass object and its appearance becomes a complex combination of light incident from the entire surrounding environment. Thus, it is a challenging task for conventional shape estimation methods to cope with glass objects.

Recent years have seen convolutional neural networks (CNNs) perform well across a variety of computer vision tasks. This includes shape estimation, where encouraging results have been observed for point cloud, depth or normal estimation for opaque objects and diffuse scenes. A few works consider the challenges of complex reflectance, but also for opaque objects. Consequently, we consider the question of whether similar CNN-based approaches are applicable for transparent objects too. But despite the versatility of CNNs in adapting to appearance variations in several computer vision problems, our experiments demonstrate that the gap between opaque and transparent image formation is too vast. This is not surprising, since estimating transparent shape requires decoupling a highly complex interaction between depth, normals and environment map.

Stets, J. D., Li, Z., Frisvad, J. R., and Chandraker, M. Single-shot analysis of refractive shape using convolutional neural networks. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV 2019)*, pp. 995-1003. January 2019.
https://doi.org/10.1109/WACV.2019.00111

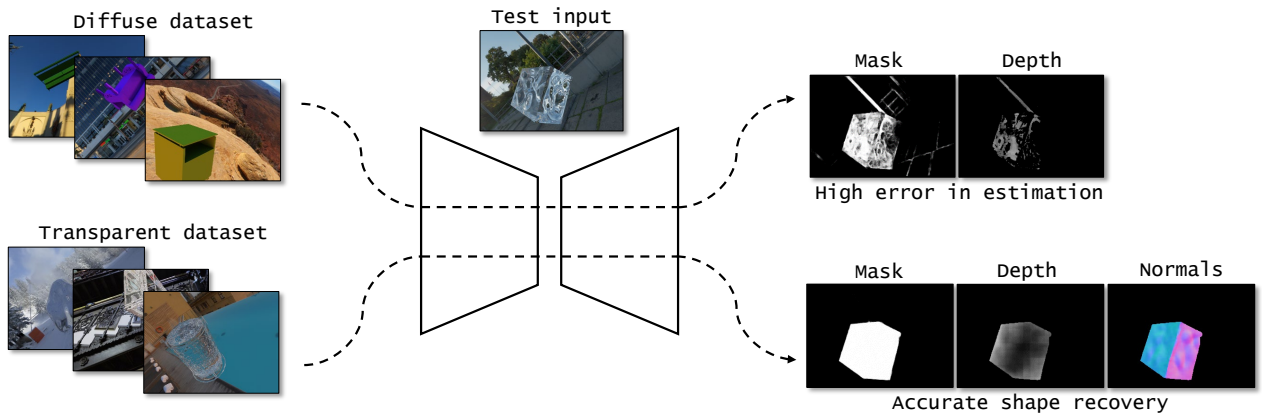


Figure 1: Illustration of our approach. We render a large-scale physically accurate dataset of transparent shapes observed under arbitrary environment maps. Using this dataset to train a CNN yields accurate shape recovery results, as compared to using a dataset of Lambertian images. We demonstrate both the need for our dataset, as well as the ability of our network to solve this challenging inverse rendering problem.

Thus, a dataset is needed to specifically train CNNs for estimating the shape of transparent objects. But acquiring such a dataset requires significant expense, as indicated in Section 2. On the other hand, shape estimation CNNs trained on synthetic datasets have been demonstrated to generalize well to real opaque scenes [5, 25]. Thus, in Section 3.1, we present our first contribution, which is a large-scale dataset of glass objects rendered for a variety of shapes under several different environment maps. Unlike several synthetic datasets, the physical accuracy required for accurately representing refractions is very high, thus, we use a GPU-accelerated physically-based renderer. Given such a dataset, it is still an open question whether an end-to-end trained CNN can disentangle the complex factors of image formation to estimate shape. In Section 3.2, we present our second contribution, which is a CNN for estimating depth, normal maps, and segmentation masks that achieves low prediction errors. But more importantly, we demonstrate in Section 4 that this CNN trained on our dataset also generalizes well to images of real transparent objects.

Figure 1 illustrates our approach and contributions, which are summarized as follows:

- A novel synthetic dataset for transparent objects generated with a high-quality physically-based renderer.
- A demonstration that CNNs trained on our dataset succeed at estimating depths, normals and segmentation.
- Empirical justification of the diversity and quality of our rendered dataset through generalization to real transparent objects at test time.

2. Related Work

Transparent shape acquisition Methods for acquiring the shape of a glass object often require an elaborate instrumental setup, such as a CT scanner or a laser range scanner [10]. Simpler setups have been presented in more recent work [9, 34]. Nevertheless, such methods cannot be used "in the wild".

Transparent shape reconstruction The shape of a single refractive surface, like a water surface or a glass object resting on a diffuse base, can be recovered through shape-from-distortion techniques. Such techniques are usually based on optical flow calculations, a known background, or a CNN trained on a specific dataset [10, 32, 20, 26, 16]. We work with the more challenging case of multiple refractions and reflections in a solid transparent object. One way to deal with this case is using texture mapping operations [37, 36] or light path triangulation [12]. However, those techniques only work for up to two refractions or reflections and either require more than one view [37, 12] or user markup of the input image [36]. Two ray-surface interactions are too few to deal with cases of total internal reflection, for example. Techniques based on optical flow also cannot deal with partial reflection and refraction (Fresnel effects) and total internal reflection [1]. Since the shape of glass objects is notoriously difficult to acquire, state of the art in 3D reconstruction of scenes with glass objects is limited to planar glass surfaces [33].

Shape estimation with CNNs As RGB-D cameras are readily available, providing easy access to color images with associated depth images, deep networks have been trained to predict a depth image from a single RGB image [27, 7]. Using techniques for estimating surface normals from depth and for semantic labeling based on depth and normals, this deep learning technique has been extended to predicting both depth, normals, and labels [6]. Different improvements

are available for the prediction of depth and normal maps using CNNs [19, 31, 14, 13]. However, these techniques are all based on databases built from images captured with an RGB-D sensor. Existing RGB-D sensors do not provide reliable depth information for transparent objects [2]. As a consequence, the databases tend to avoid transparent objects, which in turn means that existing techniques for shape estimation with CNNs neglect the existence of transparent objects. CNN-based techniques have recently been presented for material estimation too [24, 8, 17]. While such techniques might also be useful for labeling transparent objects, existing methods do not consider transparent materials.

Synthetic datasets A large-scale dataset is necessary for a data-driven method. However, manual labeling of data is in many cases too expensive or even impossible. Thus, use of high quality synthetic data to train a model is increasing in popularity. Previous methods for object level single-shot shape reconstruction have utilized large-scale shape repositories such as ShapeNet [3]. However, this only contains objects with opaque materials and therefore cannot be used directly for transparent object shape reconstruction. In addition, the category-specific bias of the dataset may limit the model's generalization ability. For scene level reconstruction, Song et al. [29] proposed the SUNCG dataset which contains $45{,}622$ indoor scenes designed by artists. Li et al. [15] propose another indoor scene dataset, with a custom path tracer to model complex light transport effects such as interreflections and shadows. We also write our own custom GPU-based path tracer to create our dataset efficiently.

All the datasets mentioned above were initially designed for 3D reconstruction. Meanwhile, Georgoulis et al. [8] use synthetic datasets for material and lighting estimation, utilizing the 3D shapes from ShapeNet and rendering the images by randomly sampling materials and lighting. Li et al. [17] render images by applying materials from the Adobe Stock dataset to a planar surface for spatially-varying reflectance estimation. Xu et al. [35] create a synthetic dataset for image based relighting by procedurally generating random shapes to create complex scenes. A similar shape generation strategy has been used in our method to create a diverse dataset with images of glass objects.

3. Method

We aim to estimate the shape of transparent objects in the wild. Inspired by the recent success of deep learning in vision and graphics, we train a deep network on a large-scale synthetic dataset generated by a photorealistic rendering engine. In the next section, we will first discuss our synthetic dataset and then compare different network design choices for solving this challenging problem.

3.1. Dataset

Training a deep network to recover the shape of transparent objects in real-world environments necessitates a large, representative dataset. However, building such a dataset with real objects is not tractable. Firstly, it is difficult to collect a large number of transparent objects with diverse enough shapes. Secondly, there are no existing methods to acquire shapes of transparent objects accurately, efficiently and cheaply. Thus, we generate our synthetic dataset by rendering transparent shapes under real environment maps. Figure 2 shows a few examples from our dataset.

Previous methods for single image shape reconstruction have utilized category-specific 3D repositories such as ShapeNet [3] for training. However, while incorporating category-level semantic may yield information to hallucinate shapes, it may decrease the generalizability of the learned model. Therefore, in addition to the 3D shapes in ShapeNet, we enrich our dataset by procedurally generating random shapes (cube, ellipse and cylinder) and then apply a randomly generated depth map onto it. Xu et al. [35] and Li et al. [18] use a similar strategy to create a large dataset for training a deep network to do relighting and joint shape and reflectance estimation. Our dataset contains 600 shapes in total, 300 shapes from the ShapeNet repository and 300 procedurally generated shapes. A total of $80{,}000$ images are rendered of the 600 shapes from randomly chosen viewpoints. We render the depths, normals and segmentation masks as ground truth. One may argue that using both normal map and depth for supervision will cause redundancy, but they have been shown to be complementary and capture different aspects of information for shape recovery [21].

Photorealistic rendering of transparent materials is challenging since visually significant paths can be very long when we account for both refraction and reflection. Based on observations in the seemingly first study with quantitatively verified photorealistic rendering of glass objects [30], we use full path tracing with deep paths. In path tracing, refraction or reflection is chosen probabilistically based on Fresnel reflectance, which depends on the angle of incidence of the light [23]. We neglect absorption in our renderings, therefore, we cannot stop our paths probabilistically in rare cases (when light is trapped by total internal reflections inside a glass object). Thus, we set a maximum trace depth (number of bounces) of 100 in our path tracer. To efficiently render a large-scale dataset, we implemented our path tracing of transparent objects in a GPU ray tracer based on NVIDIA OptiX [22]. It takes from 5 to 20 seconds to render a 480×640 RGB image in our dataset on a GeForce GTX Titan X GPU, depending on the complexity of the scene.

3.2. Network Architecture

We now describe our CNN architecture for shape recovery of a transparent object using a single image. The overall

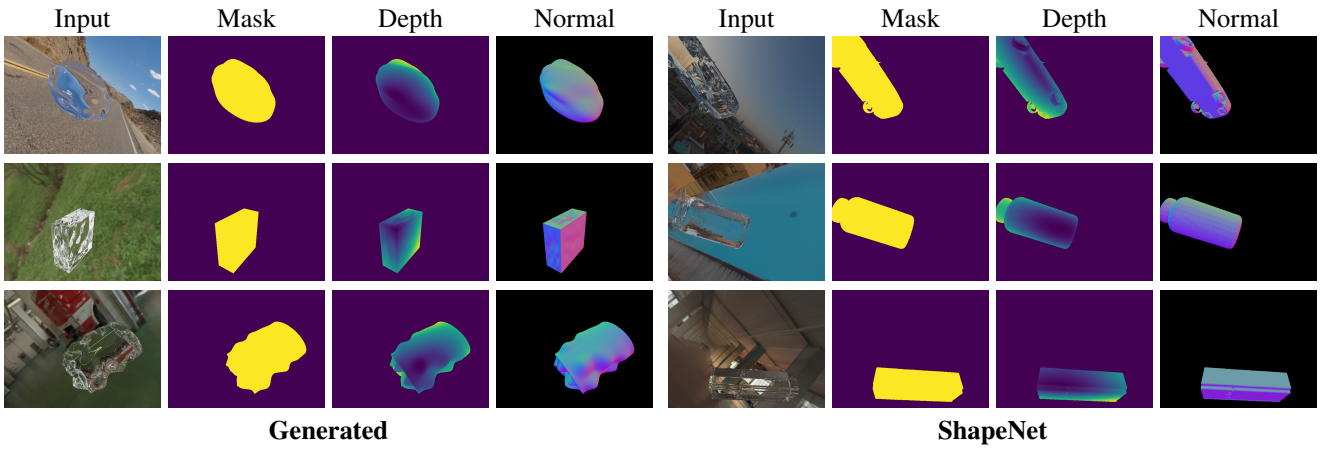


Figure 2: Rendered images of glass objects, label images, relative depth images and normal maps in our dataset. The left set uses procedurally generated shapes and the right set uses objects from ShapeNet.

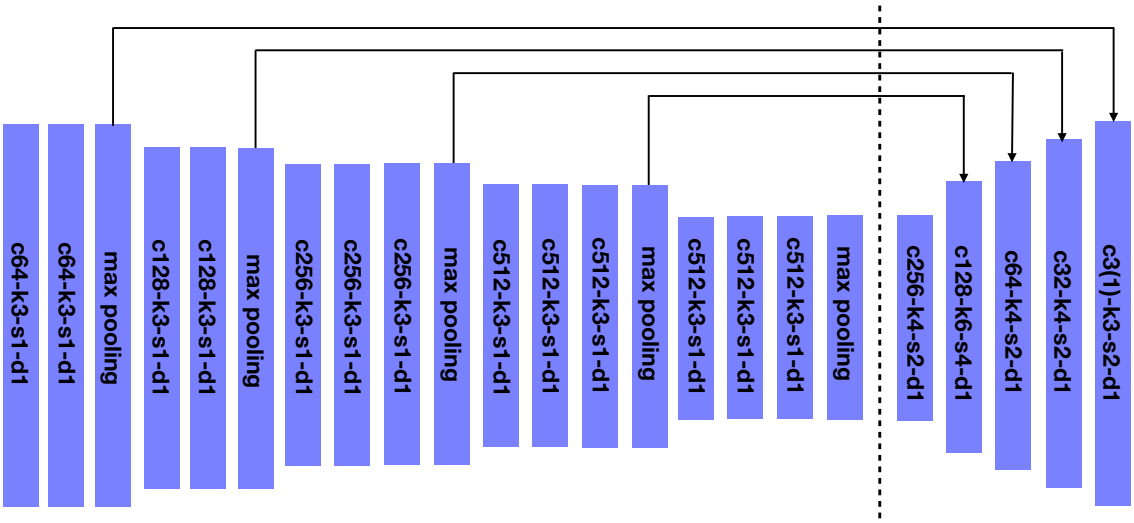


Figure 3: The encoder-decoder architecture of our network. The encoder is on the left side of the dotted line while the decoder is on the right side. Here $cX_1 - kX_2 - sX_3 - dX_4$ represents a convolutional layer (encoder) or a transposed convolutional layer (decoder) with channel X_1, kernel size X_2, stride X_3 and dilation X_4. The number of channels of the last transposed convolutional layer is 3 for normal prediction and 1 for depth and segmentation mask prediction.

structure of our network is shown in Figure 3 and follows the basic encoder-decoder network architecture for image translation. The input to our network is an image of a transparent object under environment illumination. We use the VGG16 network [28] pretrained on ImageNet [4] as the backbone of our network, which consists of a series of convolutional and max-pooling layers, followed by 3 fully connected layers for image classification. We replace the 3 fully connected layers with 5 transposed convolutional layers with stride two as our decoder to get an output image of the same size as the input. Skip links are added to preserve details of the reconstruction results. The output of our network is a segmentation mask, which separates the transparent object from the background, a normal map, and a depth map. We train a separate encoder-decoder for each task.

Loss functions We use different loss functions associated with each shape recovery task. We use L2 loss for the segmentation mask. Let $\mathcal{P}$ be the set of pixels in the image, M_p and $\hat{M}_p$ be the predicted and ground-truth segmentation mask of pixel p. Here, $M_p = 1$ represents the foreground while $M_p = 0$ stands for the background. The loss function for the segmentation mask is defined as

$$\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} (M_p - \hat{M}_p)^2 \,. \tag{1}$$

Next, for depth estimation, global scale shift is a well known ambiguity [7]. Unlike indoor scene depth estimation [6], since we estimate the shape of a single object, there is no context information for the network to resolve the scale ambiguity. Therefore, we use a scale invariant L2 loss for depth estimation. We normalize both the ground-truth and the predicted depth map to be in the range from 0 to 1. Let D_p and $\hat{D}_p$ be the predicted and ground-truth depth after normalization. Since we only wish to recover the depths of the foreground pixels $\mathcal{P}_f$, the loss function can be written as:

$$\frac{1}{|\mathcal{P}_f|} \sum_{p \in \mathcal{P}_f} (D_p - \hat{D}_p)^2 \,. \tag{2}$$

We use the same loss function as Eigen et al. [6] for normal prediction, that is, the dot product of the ground-truth normal vector $\hat{N}_p$ and the predicted normal vector N_p:

$$-\frac{1}{|\mathcal{P}_f|} \sum_{p \in \mathcal{P}_f} N_p \cdot \hat{N}_p \,. \tag{3}$$

Training and testing details We first compute the mean and variance of pixel values across the whole dataset and normalize the input images so that the distribution of pixel values has zero mean and unit variance. The same mean and variance are used to normalize the test images too. The input images are scaled to half of their original height and width and cropped to a size of 224×320 before sending to the network. We use Adam optimizer to train our network [11]. We set the learning rate to be 10^{-3} at the beginning and decrease the learning rate by half after every epoch. We train the networks for 7 epochs, which is observed to suffice for convergence.

4. Experiments and Results

In this section, we demonstrate the effectiveness of our CNN for single image reconstruction of transparent shapes. Experiments on both real and synthetic data show that by training a deep network on our large-scale dataset specifically designed for this problem, our model can generalize across different transparent shapes and even produce promising results for real transparent objects.

4.1. Experiments on Synthetic Data

Training on Lambertian objects We first verify the necessity of building a large-scale dataset with transparent materials for shape recovery. For this, we consider our shape repository but render the images with Lambertian material. We train a CNN with the same architecture and test on images rendered with Lambertian and transparent materials. The diffuse colors of Lambertian materials are chosen randomly. The test errors are reported in the first two rows of Table 1. We observe that errors for transparent test objects are significantly higher than for Lambertian ones, showing that simply training the network with opaque materials may not generalize to shape reconstruction of transparent objects.

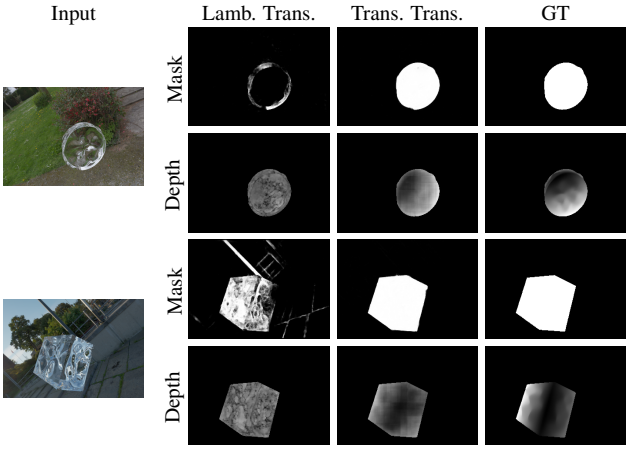


Figure 4: Comparison of mask and depth predictions by the Lambertian network trained with Lambertian objects versus the refractive network trained with transparent objects. The input images contain transparent objects. The Lambertian network makes large errors in mask and depth predictions, while the refractive network produces accurate outputs. The depths are only shown for the valid pixels in the image.

In the third row of Table 1, we report test errors of the network trained and tested on rendered images of transparent objects. We observe significant improvements and the performance is comparable to the network trained and tested on opaque objects. Thus, we conclude that single image shape reconstruction of a transparent object in an uncontrolled environment is a feasible problem for our deep network and that a large-scale, representative dataset rendered with transparent material is significantly useful.

Qualititave comparison in Figure 4 indicates that the network trained with Lambertian material has significant errors when predicting the segmentation mask of a transparent object, and the estimated depth is often incorrect. This matches the intuition that the network may confuse background and foreground when predicting the shape of transparent objects. Surprisingly, such ambiguity can be successfully solved after training our network on our dataset with transparent objects.

Shape reconstruction on synthetic dataset In Figures 5, 6, and 7, we illustrate the details of segmentation mask, depth, and normal estimates obtained by our network. We observe high quality reconstruction results for segmentation mask, depth map, and normal map. Figure 8 shows several examples of shape reconstruction results on an unseen test set. In all the examples showed, the segmentation mask successfully covers the whole transparent object, separating it from the background. Our depth map captures the

	Train	Test	Mask	Depth	Normal
VGG16	Lamb.	Lamb.	0.0012	0.0455	-
VGG16	Lamb.	Trans.	0.0807	0.0744	-
VGG16	Trans.	Trans.	0.0017	0.0349	-0.9181

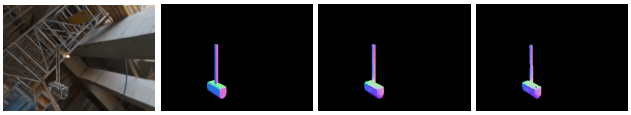
Table 1: Comparison of networks trained on Lambertian and transparent materials. The errors are calculated according to the loss functions. Higher errors are observed for a Lambertian network tested on transparent objects, as compared with the refractive network, which shows the need for our dataset. Errors for the refractive network tested on transparent objects are similar to those for the Lambertian network tested on Lambertian images, which indicates that our network is able to handle the single image reconstruction problem for transparent objects.

(a) Image (b) GT (c) Pred. (d) Pred.$>$ .9

Figure 5: Prediction of segmentation masks: (a) input image, (b) ground truth mask used for comparison, (c) predicted mask with decimal values, (d) binary mask using values with certainty above 0.9.

(a) Image (b) GT (c) Pred.* (d) Pred.**

Figure 6: Prediction of depth maps: (a) input image, (b) ground truth depth map used for comparison (c,d) depth maps predicted by the network masked with the ground truth mask of Figure 5b denoted * and the predicted mask of Figure 5d denoted **.



(a) Image (b) GT (c) Pred.* (d) Pred.**

Figure 7: Prediction of normal maps: (a) input image, (b) ground truth normals used for comparison, (c,d) normals predicted by the network masked with the ground truth mask of Figure 5b denoted * and the predicted mask of Figure 5d denoted **.

coarse shape of the objects accurately, the majority of errors are observed around the occlusion boundaries. The normal estimation is quite accurate. We observe that the network has the tendency to over-flatten the normals, which might be caused by the inherent ambiguity of transparent object reconstruction.

Generalization to a different refractive index The transparent objects in our dataset are rendered with fixed index of refraction (IOR) of 1.5. However, different transparent materials may have different IORs, thus, it is important to verify whether the network trained on materials with fixed IOR can generalize to other transparent materials. We test our trained model on shapes rendered with a range of IORs and reported the errors in Figure 9. We picked 100 shapes and rendered sets of 20 images with different IOR values evenly distributed in the range $]1, 2]$. From Figure 9, we see that even though our network performs the best on the test set rendered with the same IOR as the training set, the test results on images rendered with other IORs still achieve reasonably low errors. As we demonstrate next, this robustness suffices to recover shapes for real objects with unknown refractive indices.

4.2. Experiments on Real Data

To verify the generalization ability of our network on real transparent objects imaged in uncontrolled environments, we show results on real images of transparent glass objects. Notice that our network is trained completely on synthetic data and we do not fine tune our network for specific devices or environments. We show several shape reconstruction examples in Figure 10. In two of these examples, the transparent objects are placed in front of a checkerboard background so that the distortion caused by refraction can be more clearly observed. In the other three examples, the capturing environment is completely uncontrolled. While ground truth shapes are not available, we observe that the network produces reasonable shape reconstruction results, for images captured in front of checkerboard background as well as in uncontrolled settings. However, our network does perform better when the distortion can be easily observed, as noted by comparing the spheres in Figure 10 images with the checkerboard or in-the-wild. An interesting phenomena is that our network successfully reconstructs the shape of the pitcher in the fifth row. In this example, most rays are refracted four times before reaching the camera while in the majority of examples in our training set, rays most often refract only twice.

5. Discussion and Future Work

Utilizing the vast progress in deep learning, we make a first attempt to tackle the problem of single image shape reconstruction of transparent objects in an uncontrolled environment. Such a problem is often avoided in computer vision research, but should have a major impact on real world applications since transparent objects such as windows and cups are very common in daily life. Previous datasets for

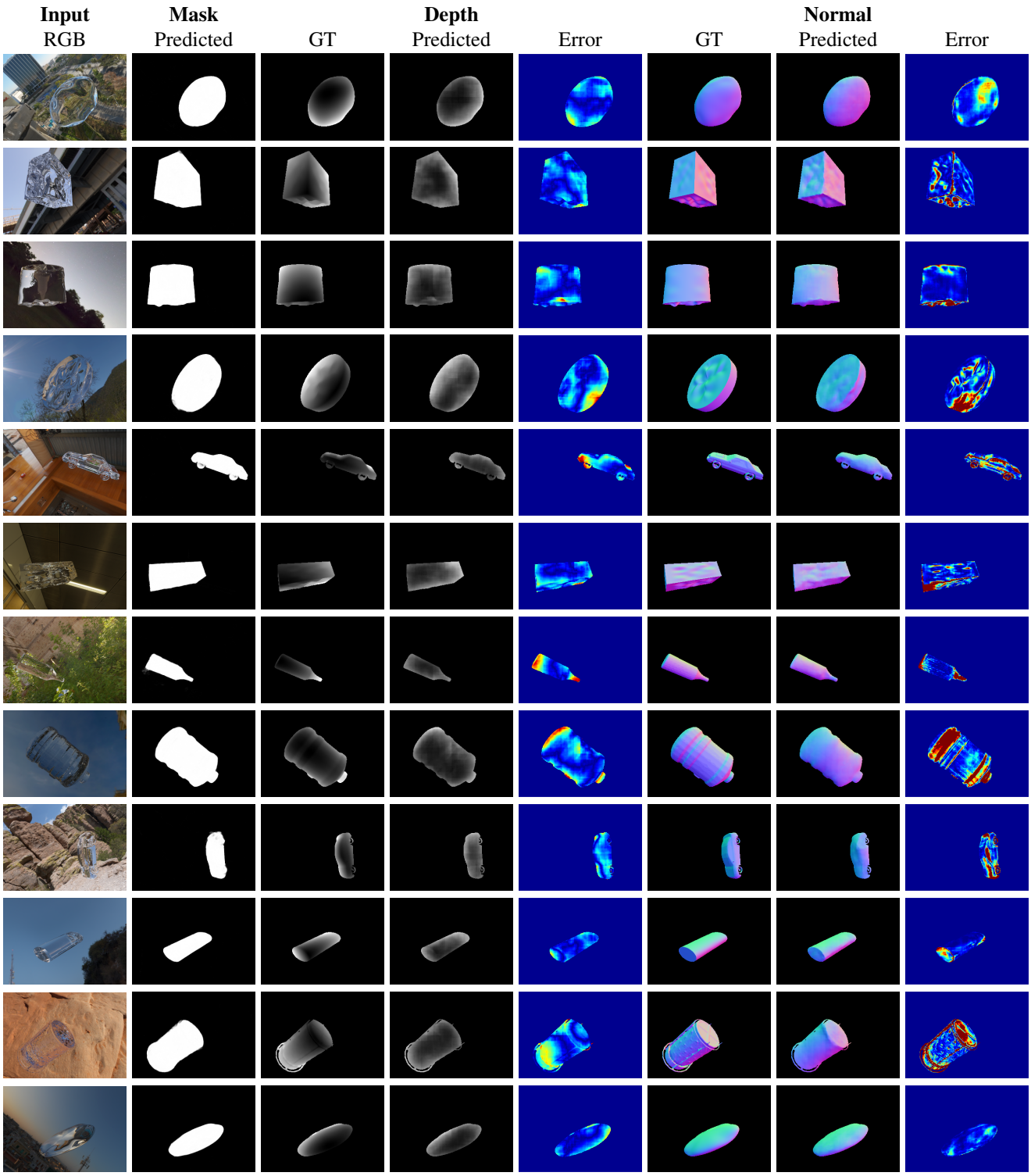


Figure 8: A selection of results from the test set. The first column has the input images, the second column has the predicted masks. Columns 3-5 have the depth estimation results, and columns 6-8 have the normal estimation results. The predicted depths and normals are masked with the ground truth masks to only show the valid regions. The error plots show pixel-wise loss in the valid region, according to the loss functions stated in Section 3.2, but scaled from 0 to 1.

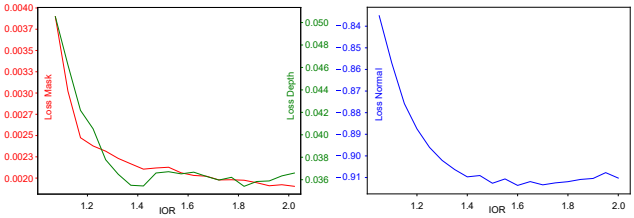


Figure 9: The mean prediction error when varying the refractive index. We observe that our network achieves low test errors across a wide range of refractive indices.

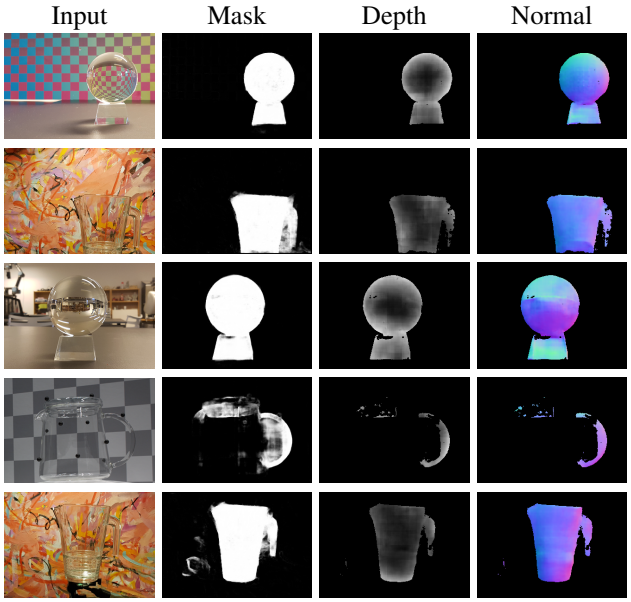


Figure 10: Predictions on photographs of real glass objects. The depth and normal predictions are based on a predicted mask with a 0.9 threshold.

shape reconstruction are only rendered with opaque materials, therefore, we present an extensive dataset that consists of photorealistic rendering of transparent objects. Our experiments verify that such a dataset is necessary to solve this challenging problem. We show that by training on our large-scale dataset, our model generalizes well to transparent materials with different IORs, while handling arbitrary, complex shapes of even real transparent objects captured in uncontrolled environments. We still observe errors for real-world transparent object reconstruction, which suggests that we should either use more aggressive data augmentation when creating the dataset or design a physically-inspired network architecture which has better generalization power. In particular, we will consider enriching the dataset by rendering transparent objects with different IORs and reconstructing the shapes of the objects that are not directly visible to the camera. Since the shape of the invisible part influences the appearance of the image of a transparent object, we expect to achieve better reconstruction signals for reconstructing the back of transparent objects using a single image, as compared to the opaque problem where such information must be recovered purely based on data-driven priors.

Acknowledgements

The environment maps are all free for any non-commercial purposes and downloaded from Bernhard Vogl, hdrlabs.com, hdrihaven.com, and hdrmaps.com. Some of the models used for our dataset are from ShapeNet.org.

References

[1] S. Agarwal, S. P. Mallick, D. Kriegman, and S. Belongie. On refractive optical flow. In *Proceedings of European Conference on Computer Vision (ECCV 2004)*, pages 483–494. Springer, 2004.

[2] N. Alt, P. Rives, and E. Steinbach. Reconstruction of transparent objects in unstructured scenes with a depth camera. In *Proceedings of IEEE International Conference on Image Processing (ICIP 2013)*, pages 4131–4135, September 2013.

[3] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. ShapeNet: An information-rich 3D model repository. arXiv:1512.03012 [cs.GR], 2015.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 248–255, 2009.

[5] A. Dosovitskiy, P. Fischery, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. v. d. Smagt, D. Cremers, and T. Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of IEEE International Conference on Computer Vision (ICCV 2015)*, pages 2758–2766, 2015.

[6] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of IEEE International Conference on Computer Vision (ICCV 2015)*, pages 2650–2658, 2015.

[7] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems (NIPS 2014)*, volume 27, pages 2366–2374, 2014.

[8] S. Georgoulis, K. Rematas, T. Ritschel, E. Gavves, M. Fritz, L. V. Gool, and T. Tuytelaars. Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):1932–1947, August 2018.

[9] K. Han, K.-Y. K. Wong, and M. Liu. A fixed viewpoint approach for dense reconstruction of transparent objects. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 4001–4008, 2015.

[10] I. Ihrke, K. N. Kutulakos, H. P. A. Lensch, M. Magnor, and W. Heidrich. Transparent and specular object reconstruction. *Computer Graphics Forum*, 29(8):2400–2426, 2010.

[11] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representations (ICLR 2015)*. arXiv:1412.6980 [cs.LG], 2015.

[12] K. N. Kutulakos and E. Steger. A theory of refractive and specular 3D shape by light-path triangulation. *International Journal of Computer Vision*, 76(1):13–29, 2008.

[13] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In *Proceedings of International Conference on 3D Vision (3DV 2016)*, pages 239–248. IEEE, 2016.

[14] B. Li, C. Shen, Y. Dai, A. Van Den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 1119–1127, 2015.

[15] W. Li, S. Saeedi, J. McCormac, R. Clark, D. Tzoumanikas, Q. Ye, Y. Huang, R. Tang, and S. Leutenegger. InteriorNet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In *Proceedings of British Machine Vision Conference (BMVC 2018)*. arXiv:1809.00716 [cs.CV], 2018.

[16] Z. Li, Z. Murez, D. Kriegman, R. Ramamoorthi, and M. Chandraker. Learning to see through turbulent water. In *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV 2018)*, pages 512–520, 2018.

[17] Z. Li, K. Sunkavalli, and M. Chandraker. Materials for masses: SVBRDF acquisition with a single mobile phone image. In *Proceedings of European Conference on Computer Vision (ECCV 2018)*, pages 72–87, 2018.

[18] Z. Li, Z. Xu, R. Ramamoorthi, K. Sunkavalli, and M. Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Transactions on Graphics*, 37(6), 2018. To appear.

[19] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 5162–5170, 2015.

[20] N. J. W. Morris and K. N. Kutulakos. Dynamic refraction stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1518–1531, 2011.

[21] D. Nehab, S. Rusinkiewicz, J. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. *ACM Transactions on Graphics*, 24(3):536–543, July 2005.

[22] S. G. Parker, J. Bigler, A. Dietrich, H. Friedrich, J. Hoberock, D. Luebke, D. McAllister, M. McGuire, K. Morley, A. Robison, and M. Stich. OptiX: A general purpose ray tracing engine. *ACM Transactions on Graphics*, 29(4):66:1–66:13, 2010.

[23] M. Pharr, W. Jakob, and G. Humphreys. *Physically Based Rendering: From Theory to Implementation*. Morgan Kaufmann/Elsevier, third edition, 2017.

[24] K. Rematas, T. Ritschel, M. Fritz, E. Gavves, and T. Tuytelaars. Deep reflectance maps. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 4508–4516, 2016.

[25] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez. The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, pages 3234–3243, 2016.

[26] Q. Shan, S. Agarwal, and B. Curless. Refractive height fields from single and multiple images. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, pages 286–293, 2012.

[27] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from RGBD images. In *Proceedings of European Conference on Computer Vision (ECCV 2012)*, pages 746–760. Springer, 2012.

[28] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 [cs.CV], 2014.

[29] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 190–198, 2017.

[30] J. D. Stets, A. Dal Corso, J. B. Nielsen, R. A. Lyngby, S. H. N. Jensen, J. Wilm, M. B. Doest, C. Gundlach, E. R. Eiriksson, K. Conradsen, A. B. Dahl, J. A. Bærentzen, J. R. Frisvad, and H. Aanæs. Scene reassembly after multimodal digitization and pipeline evaluation using photorealistic rendering. *Applied Optics*, 56(27):7679–7690, September 2017.

[31] X. Wang, D. Fouhey, and A. Gupta. Designing deep networks for surface normal estimation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, pages 539–547, 2015.

[32] G. Wetzstein, D. Roodnick, W. Heidrich, and R. Raskar. Refractive shape from light field distortion. In *Proceedings of IEEE International Conference on Computer Vision (ICCV 2011)*, pages 1180–1186, 2011.

[33] T. Whelan, M. Goesele, S. J. Lovegrove, J. Straub, S. Green, R. Szeliski, S. Butterfield, S. Verma, and R. Newcombe. Reconstructing scenes with mirror and glass surfaces. *ACM Transactions on Graphics*, 37(4):102:1–102:11, August 2018.

[34] B. Wu, Y. Zhou, Y. Qian, M. Cong, and H. Huang. Full 3D reconstruction of transparent objects. *ACM Transactions on Graphics*, 37(4):103:1–103:11, August 2018.

[35] Z. Xu, K. Sunkavalli, S. Hadap, and R. Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (TOG)*, 37(4):126, 2018.

[36] S.-K. Yeung, C.-K. Tang, M. S. Brown, and S. B. Kang. Matting and compositing of transparent and refractive objects. *ACM Transactions on Graphics*, 30(1):2:1–2:13, 2011.

[37] D. E. Zongker, D. M. Werner, B. Curless, and D. H. Salesin. Environment matting and compositing. In *Proceedings of SIGGRAPH 1999*, pages 205–214. ACM Press/Addison-Wesley, 1999.