Supplementary Material for Through the Looking Glass: Neural 3D Reconstruction of Transparent Shapes

Zhengqin Li* Yu-Ying Yeh* Manmohan Chandraker
University of California, San Diego
{zhl378, yuyeh, mkchandraker}@eng.ucsd.edu

1. Real Data Evaluation with Ground Truth

We first present quantitative comparisons and then more qualitative ones on real data. We use four objects (*mouse*, *dog*, *pig* and *monkey*). All objects are reconstructed from 10 views under natural environment maps, except *monkey*, which needs 12 views since the shape is much more complex. To obtain the ground truth geometry, we paint each object with diffuse white paint and scan using a high-quality 3D scanner. All code and data will be publicly released.

Quantitative results We manually align ground-truth shapes with the predicted shapes using ICP method [1] and then uniformly sample 20000 points on the both shapes to compute the four error metrics (CD, CDN-mean, CDN-med, Metro). The quantitative numbers are summarized in Table 1. For all the 4 objects, our method consistently outperforms the visual hull baseline, which again demonstrates the effectiveness of our transparent shape reconstruction framework.

	Views	CD(10^{-4})		CDN-mean(°)		CDN-med(°)		Metro(10^{-3})	
		vh	Rec	vh	Rec	vh	Rec	vh	Rec
monkey	12	3.99	**3.94**	21.2	**16.4**	14.8	**11.9**	20.7	**13.9**
mouse	10	8.04	**5.35**	19.0	**16.3**	11.4	**12.0**	16.6	**13.0**
pig	10	5.58	**4.87**	19.0	**18.3**	**14.0**	14.6	13.0	**7.4**
dog	10	2.25	**1.86**	14.5	**12.4**	11.4	**10.3**	4.1	**4.0**
mean	10.5	4.97	**4.00**	18.4	**15.9**	12.9	**12.2**	13.6	**9.6**

Table 1. Quantitative comparisons of transparent shape reconstruction on real data. We observe that our reconstruction achieves lower average errors than the visual hull method on all the metrics.

Qualitative results and videos Figure 1 shows both the ground-truth transparent shapes and our reconstructed shapes rendered under different lighting and materials. Even though the shapes are complex and we use very limited inputs, our reconstructions still closely match the ground truth appearance. This demonstrates the efficacy of our physically-motivated network that models complex light paths induced by refractions and reflections. To better visualize the quality of our 3D reconstruction outputs, we create a video by rotating both the ground-truth shapes and the reconstructed shapes under different natural environment maps. The video is included in the supplementary material. A higher resolution video is available at this link.

*These two authors contributed equally

2. Sensitivity Analysis for Index of Refraction

As mentioned in Sec. 4.1 of the main paper, we perform a sensitivity analysis on the influence of a different test-time IoR on the shape reconstruction accuracy. We re-render our synthetic transparent testing set with the same shapes and environment maps. However, instead of rendering with a fixed IoR value of 1.4723, we randomly sample 5 different IoRs ranging from 1.3 to 1.7 for each shape. Figure 2 shows an example of the same shape rendered under different IoRs. We then test our network trained with a fixed IoR value of 1.4723 on the new test set with variable IoR. During testing, the IoR used by the rendering layer is kept fixed at 1.4723. The quantitative comparisons have been summarized in Tables 1 and 2 of the main paper.

Figure 3 and 4 show trends in the normal and shape reconstruction errors across varying IoRs in the test set. As expected, the errors are relatively smaller for IoRs close to the training set value of 1.4723. In particular, this trend is more explicitly visible in normal estimation, since the model leverages the features from the rendering layer and cost volume which require known IoRs. However, the overall variation in error is small across this range of IoRs.

The above plots further support the analysis in Tables 1 and 2 of the main paper. Even though the predicted normals and the final reconstructed mesh are expectedly more accurate in the known IoR case, the quantitative errors increase gracefully and not too much across a range even with unknown IoR. This suggests that our network is relatively robust to the IoR value. As stated in the main paper, our future work will consider simultaneously reconstructing the transparent shape and predicting its IoR.

3. Further Ablation Studies

Different number of views Table 2 summarizes the normal predictions from 5 and 20 views. Similar to the 10-view case, our entire method `wr+cv+op` outperforms all other baselines on all the five metrics. In particular, we find the cost volume (`cv`) and the optimization of the latent vector

Figure 1. Results on 3D reconstruction for four real transparent objects. All shapes are reconstructed from 10 views, except the *monkey* in the last row that uses 12 views. We first present reconstruction results from two input views (columns 1-6). From left to right, the odd rows show the input image and the reconstructed shapes under different lighting and materials. The corresponding outputs using the ground-truth shapes rendered from the same view are shown in the even rows. We also render the reconstructed shapes and ground-truth shapes from a novel view direction that has not been used to build the visual hull (columns 7-8). In each instance, we observe that the reconstructions are close to the ground truth despite the challenging shapes, complex light paths and small number of views used for 3D reconstruction.

Figure 2. Appearance changes for same shape geometry under various index of refraction (IoRs). IoRs range from 1.3 to 1.7.

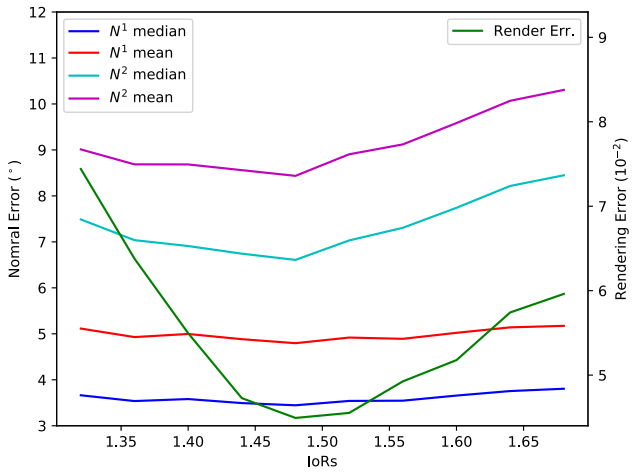

Figure 3. The mean normal estimation errors across varying IoRs in the test set, using the fixed training set IoR value for prediction.

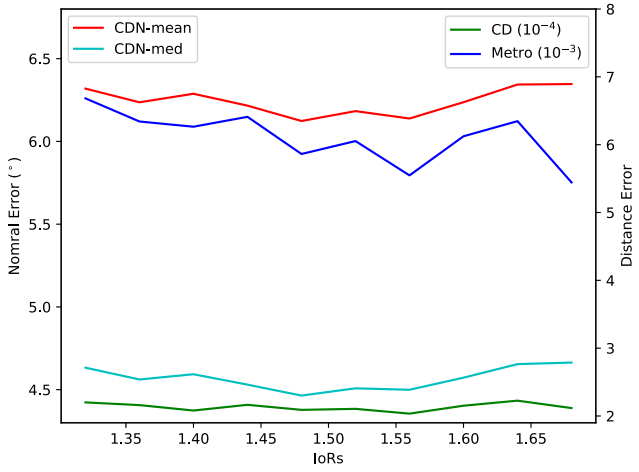

Figure 4. The mean shape reconstruction errors across varying IoRs in the test set, using the fixed training set IoR value for prediction.

(op) bring the largest improvements on normal reconstruction accuracy. This justifies our intuition that utilizing the correspondence between the input image and the captured environment map by modeling the image formation process within the network can lead to better normal reconstruction results. Figure 5 shows two normal reconstruction results on our synthetic dataset. For both examples, our physically-based network performs significantly better than the classical visual hull method for 5, 10 and 20 views.

5 views normal reconstruction	vh5	basic	wr	wr+cv	wr+cv +op
N^1 median (°)	12.7	6.1	6.0	6.0	**5.9**
N^1 mean (°)	15.3	7.8	7.9	7.8	**7.7**
N^2 median (°)	18.3	10.7	10.7	10.5	**10.0**
N^2 mean (°)	20.9	12.5	12.5	12.3	**11.9**
Render Err.(10^{-2})	9.7	5.9	5.8	5.9	**4.1**
20 views normal reconstruction	vh20	basic	wr	wr+cv	wr+cv +op
N^1 median (°)	2.5	2.2	2.2	2.2	**2.2**
N^1 mean (°)	4.6	3.4	3.4	3.3	**3.3**
N^2 median (°)	5.2	4.7	4.6	4.6	**4.3**
N^2 mean (°)	7.6	6.5	6.4	6.3	**6.1**
Render Err.(10^{-2})	4.0	3.7	3.8	3.8	**2.7**

Table 2. Quantitative comparisons of normal estimation from 5 and 20 views. Following the notation in the main paper, vh5 and vh20 represent the initial normals reconstructed from visual hulls corresponding to 5 and 20 views, respectively. Here, wr and basic are our basic encoder-decoder network with and without rendering error map (I^{er}) and total reflection mask (M^{tr}) as inputs. Further, wr+cv represents our network with cost volume and wr+cv + opt represents the predictions after optimizing the latent vector to minimize the rendering error. Similar to the 10-view case, wr+cv + opt performs better than all other baselines for transparent shape reconstruction using both 5 and 20 views.

	CD(10^{-4})	CDN-mean(°)	CDN-med(°)	Metro(10^{-3})
RE-$\mathcal{L}_P^{\text{CD}}$	**2.00**	**6.02**	**4.38**	**5.98**
→maxPooling	2.09	6.26	4.59	6.09
– normal Diff.	2.09	6.31	4.62	6.48
– normal Skip.	2.07	6.14	4.51	6.20
standard	2.12	6.34	4.72	6.49

Table 3. Comparisons of point cloud reconstruction with different PointNet++ architectures on our synthetic dataset. Following the notation in the main paper, RE represents rendering error based view selection. $\mathcal{L}_P^{\text{CD}}$ represents the Chamfer distance loss.

Modification of standard PointNet++ [2] We examine our modifications of the standard PointNet++ architecture for point cloud reconstruction to better incorporate normal information. The quantiative numbers are summarized in Table 3. We first remove single modifications from standard

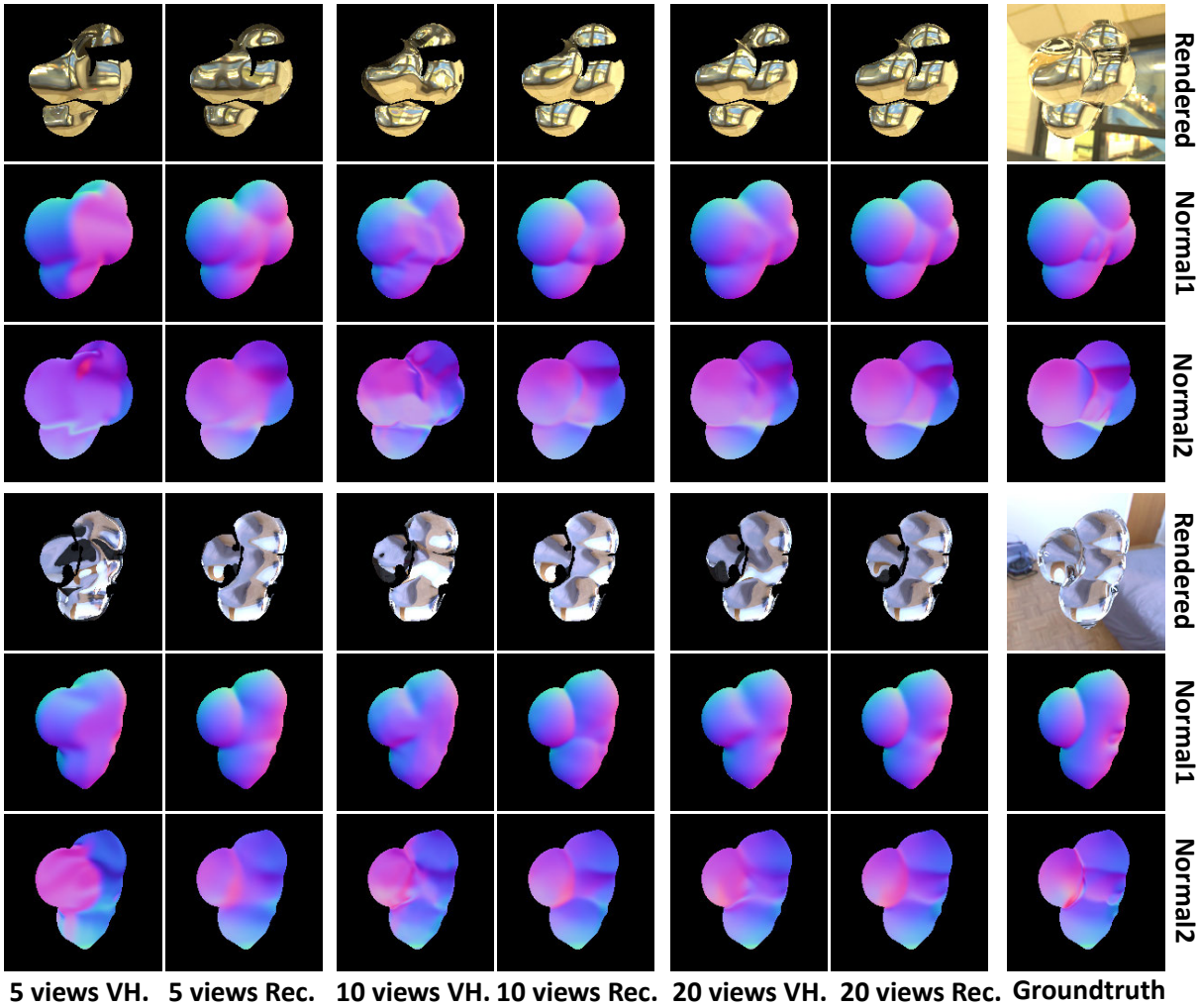

Figure 5. Normal predictions on our synthetic dataset with different number of input views. Regions with total reflection have been masked out in the rendered images. Our predicted normals are much closer to the ground truth compared to the visual hull normals.

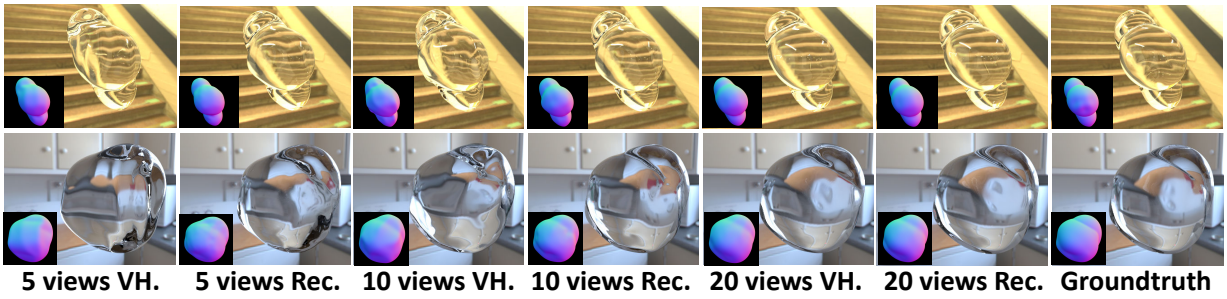

Figure 6. Transparent shape reconstruction in our synthetic dataset using 5, 10 and 20 views. Images rendered with our reconstructed shapes are much closer to the those rendered with ground truth shape, as compared to images rendered with the visual hull shapes. The inset normals are rendered from the reconstructed shapes and demonstrate the same conclusion.

PointNet++ to our novel version ($\rightarrow$ maxPooling, – normal Diff. and – normal Skip.) and then remove all the modifications to use standard PointNet++ to reconstruct the point cloud of our transparent shapes (standard). Experiments show that each of our modifications brings consistent improvements in reconstruction accuracy and removing all of them leads to a much poorer performance. This shows our modifications ease the difficulty for the network to reason about point cloud distribution based on normal predictions. Figure 7 demonstrates a real example reconstructed by our

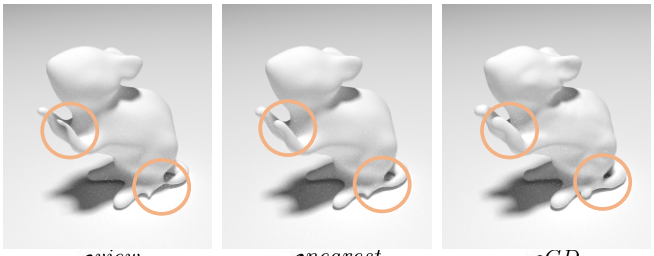

Figure 7. Comparisons of point cloud reconstruction with different loss functions on a real example. Our modified PointNet++ trained with Chamfer distance loss achieves better quality compared with the other two losses.

10 views normal reconstruction	vh10	wr+cv +op	wr+cv +opPixel
N^1 median (°)	5.5	**3.4**	3.8
N^1 mean (°)	7.5	**4.8**	4.9
N^2 median (°)	9.2	**6.6**	7.4
N^2 mean (°)	11.6	**8.4**	8.5
Render Err.(10^{-2})	6.0	2.9	**2.6**

Table 4. Quantitative comparisons of different optimization strategies for normal estimation from 10 views. `op` represents optimization the latent vector, which is the results reported in the main paper. `opPixel` represents optimization direction in the pixel space.

modified PointNet++ trained using different loss functions. It is clearly observed that our modified PointNet++ trained with Chamfer distance loss leads to a more complete and less noisy 3D reconstruction, especially for thin structures and concave regions.

Optimization of latent vector We adopt an alternating minimization strategy to optimize the latent vector. We first keep N^1 unchanged and only change N^2 by adding a large identity loss on N^1. After 500 iterations, we remove the constraint and optimize both N^1 and N^2 simultaneously. This is because the our N^1 prediction is usually more accurate and optimizing N^2 first can lead to better results. In Table 4, we compare the normal reconstruction results of optimizing the latent vector and directly optimizing the per-pixel normals. The quantitative comparison shows that while optimizing per-pixel normal can also decrease the rendering error, only by optimizing the latent vector can we observe improvements in normal reconstruction accuracy. The inherent ill-posed nature of normal prediction of transparent shapes makes it necessary to have a strong regularization to obtain meaningful outputs. In this case, the regularization is provided by the trained decoder which constrains the predicted normals to be on the natural shape manifold.

4. Building the Cost Volume

To build the cost volume (`cv`) for normal prediction, we sample ϕ uniformly from 0 to 2π and sample θ according to

	$\{\theta_k\}_{k=1}^4$	$\{\phi_k\}_{k=1}^4$
5 views	0°,25°,25°,25°	0°,0°,120°,240°
10 views	0°,15°,15°,15°	0°,0°,120°,240°
20 views	0°,10°,10°,10°	0°,0°,120°,240°

Table 5. The sampled angles for building cost volume. We set the sampled angles according to the normal error of visual hull reconstructed by different number of views.

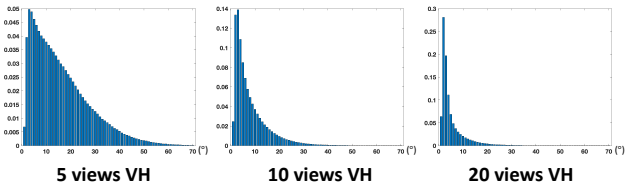

Figure 8. The error distribution of visual hull normals $\tilde{N}^1$ from different number of views.

Algorithm 1 Mapping normals to visual hull geometry

for point $\tilde{p}$ uniformly sampled from visual hull **do**
  $\tilde{p}_{N^1} \leftarrow$ the original visual hull normal
  $\tilde{p}_{M^{tr}} \leftarrow 1, \tilde{p}_{I^{er}} \leftarrow 2, \tilde{p}_v \leftarrow 0, \tilde{p}_c = 0$
for view v from 1 to V **do**
  for point $\tilde{p}$ uniformly sampled from visual hull **do**
    `isUpdate` $\leftarrow$ `False`
    if $\mathcal{V}(\tilde{p}) = 1$ **then**
      if $\mathcal{S}_v(\tilde{p}, M_v^{tr}) = 1$ **then**
        if $\tilde{p}_M^{tr} = 1$ and $\mathtt{C}_v(\tilde{p}) > \tilde{p}_c$ **then**
          `isUpdate` $\leftarrow$ `True`
      else
        if $\tilde{p}_M = 1$ **then**
          `isUpdate` = `True`
        else if $\mathcal{S}_v(\tilde{p}, I_v^{er}) < \tilde{p}_{I^{er}}$ **then**
          `isUpdate` = `True`
    if `isUpdate` = `True` **then**
      $\tilde{p}_{N^1} \leftarrow \mathcal{T}_v(\mathcal{S}_v(\tilde{p}, N_v^1)),\ \tilde{p}_{M^{tr}} \leftarrow \mathcal{S}_v(\tilde{p}, M_v^{tr})$
      $\tilde{p}_{I^{er}} \leftarrow \mathcal{S}_v(\tilde{p}, I_v^{er}),\ \tilde{p}_c = \mathcal{C}_v(\tilde{p}),\ \tilde{p}_v \leftarrow v$
$\{f\} \leftarrow$ Concatenate $\{\tilde{p}_{N^1}\}, \{\tilde{p}_{M^{tr}}\}, \{\tilde{p}_{I^{er}}\}, \{\tilde{p}_c\}$
return $\{f\}, \{\tilde{p}_v\}$

the visual hull normal error. In particular, we first randomly sample 100 scenes from our synthetic dataset and compute the angles between visual hull normals and ground truth normals. We set one θ value to be 0 and the other to larger than 85% of angles between the visual hull normal $\tilde{N}^1$ and ground truth normal $\hat{N}^1$. The distribution of visual hull normal $\tilde{N}^1$ error for 5, 10 and 20 views are presented in Figure 8. Table 5 summarizes the configurations of $\{\theta\}$ and $\{\phi\}$ angles for different number of views.

5. Details for Feature Mapping

Our feature mapping method using the rendering error based view selection is summarized in Algorithm 1. We first try to select the view with no total reflection as the best view v^*. If there is more than one view with no total reflection, we choose the view with the lowest rendering error. If for every view, the current point is in the region of total reflection, we choose the view whose optical center is closest to the point. Experiments in the main paper show

that our rendering error based view selection (RE) performs slightly better than average fusion (AV) and nearest view selection (NE) on 3D reconstruction accuracy.

References

[1] Paul J. Besl and Neil D. McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, volume 1611, pages 586–606. International Society for Optics and Photonics, 1992. 1

[2] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems*, pages 5099–5108, 2017. 3