RCTrans: Transparent Object Reconstruction in Natural Scene via Refractive Correspondence Estimation -Supplementary Material-

ACM Reference Format:
. 2025. RCTrans: Transparent Object Reconstruction in Natural Scene via Refractive Correspondence Estimation -Supplementary Material-. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3757377.3763859

1 Details of RCTrans

Here, we provide more details about RCNet and the data preparation before transparent object reconstruction.

1.1 RCNet

The two CNNs in RCNet share the same architecture, which contains three residual blocks. The transformer uses the same architecture as Xu et al. [2023], which contains six blocks, and each block consists of self-attention, cross-attention, and a small MLP.

For the transparent object image and background image at the size of $H \times W \times 3$, the CNNs would extract $8\times$ downsampled features, and the following transformer and matching layer would keep the same downsampled size. We further perform 6 additional refinements in RAFT [Teed and Deng 2020] to improve the accuracy. At each refinement, a residual correspondence is regressed with convolutions from local correlations. Then we use the upsampling method in RAFT to get the full resolution result, which computes the correspondence at each pixel as a weighted combination of a 3×3 grid of its coarse resolution neighbors. Each group of combination weights has the size of $8 \times 8 \times 3 \times 3$ and is predicted by a small CNN.

During training, the initial and refined correspondences at each step are upsampled with bilinear interpolation to the full resolution, serving as intermediate results to be also supervised with ground truth, as the Eq. 4 in the main paper.

In the main paper, all displayed correspondence estimates and warped images have been at least filtered by the object mask, both because the mask is available during reconstruction and to ensure clearer visualization. We present the original full-image estimated correspondence in Fig. 3. Note that RCNet does not automatically predict reasonable correspondence for the background areas since they have never been supervised during training, which is the same reason why we can filter regions where total reflection occurs with reconstruction error.

Fig. 1. The recovered environment lighting and background image on synthetic data

Fig. 2. The recovered background image and warped image on real data. Based on our experience, we roughly annotated some total internal reflection regions with blue masks to facilitate comparative analysis of the remaining areas.

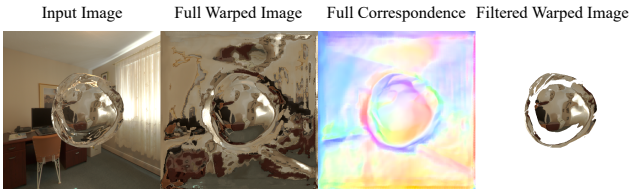

Fig. 3. The full-image estimated correspondence and warped image. The background regions do not have reasonable correspondence since they have never been trained. Note that the warped image filtered by the GT valid mask remains accurate.

Author's Contact Information:

SA Conference Papers '25, Hong Kong, Hong Kong

ACM ISBN 979-8-4007-2137-3/2025/12
https://doi.org/10.1145/3757377.3763859

Table 1. The error on the reconstruction data, measured by EPE and average averaged across all views. The correspondence coordinates are normalized to $[0, 1]$. "GT Back." represents feeding the network with ground truth background images. "Render Back." means use rendered background images. The errors of these two methods are calculated within the entire GT valid mask. "Filtered" means using the rendered background and filtering the results, whose error is only calculated on those remaining correspondences within GT valid mask.

Method	Bowl	Cat	Rabbit	Squirrel	Hand
GT Back.	0.0748	0.0796	0.0702	0.0944	0.0905
Render Back.	0.0742	0.0802	0.0702	0.0943	0.0905
Filtered	0.0674	0.0630	0.0519	0.0804	0.0757

1.2 Data Preparation

Before transparent object reconstruction, RCTrans leverages the previous method [Gao et al. 2023] to obtain the multi-view object silhouettes and the environment lighting. Specifically, given the multi-view images, we directly recover the whole scene radiance and SDF field using NeuS [Wang et al. 2021], without considering refraction. Then we project the SDF field back to the input views with the function as:

$$f_M = \int_0^{+\infty} w'(t)dt \tag{1}$$

For synthetic data, $w'(t)$ is the same as the original weight function $w(t)$ in NeuS. As for real data, $w'(t)$ filters out the support under the object, following Gao et al. [2023]. We reduced the training iteration of NeuS from 300,000 to 100,000, since it could already yield accurate silhouettes.

During this process, the environment lighting represented by nerf++ [Zhang et al. 2020] can also be recovered, which is used to render the object-free background image with a large FOV. While some refracted rays may still fall outside the background image, it has little harm to the reconstruction since data from other views would complement it. For synthetic data that has a large enough FOV (80 degrees), we directly preserve the original background part in inputs and only render to fill the area inside the object silhouette to reduce computation, as shown in Fig. 1. The quantitative results in Tab. 1 show that the rendered background image can greatly approximate the ground truth and does not interfere with the correspondence estimation.

As for real data, we render the entire background image with the FOV set as 90 degrees, which can remove the support and roughly guarantee the refractive correspondence within the background images. Although the recovered lighting on the real data is more blurred due to focusing, it is still sufficient to support the correspondence estimation, as proved by the warped image in Fig. 2.

2 More Experiment Results

We provide more visual comparisons on correspondence estimation to prove the superiority of our RCNet. As shown in Fig. 4, the regression-based baseline only produces rough results, but lacks details and precision. In contrast, our method produces highly accurate results, attributed to the explicit matching process.

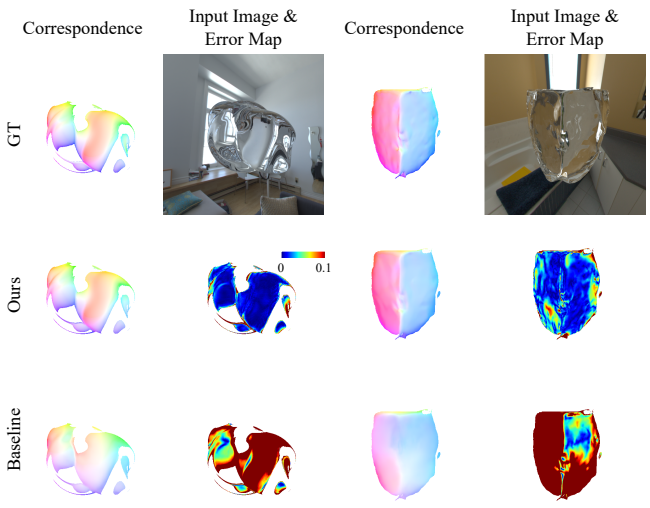

Fig. 4. The refractive correspondence estimation results on synthetic data, compared with the baseline.

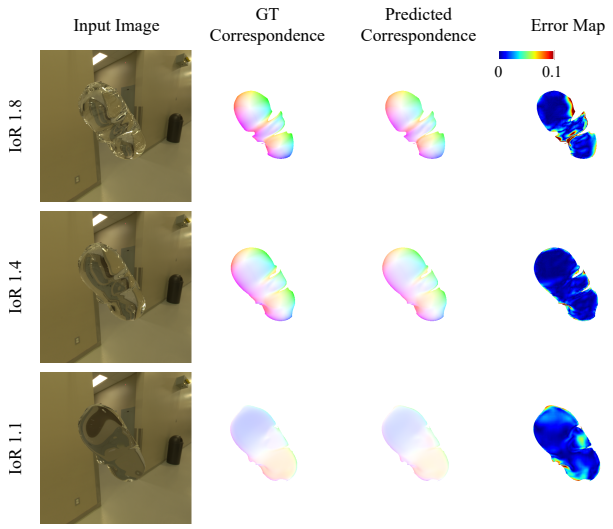

Fig. 5. The correspondence estimation results on the same sample with different IoR, including 1.8 (higher than the training data range), 1.4 (within the training data range) and 1.1 (lower than the training data range). Our method shows excellent generalization for IoRs outside the training range.

Besides, we present the quantitative error of correspondence estimation on reconstruction data in Tab. 1, including the results before and after confidence filtering. The decreased error proves that the filtering can not only reduce the influence of total reflection, but also decrease the inaccurate estimation within the valid mask. The higher error on the reconstruction set compared to the validation set primarily stems from more complex geometries and challenging viewing.

The index of refractive (IoR) range of our training and validation data is from 1.3 to 1.6, which covers the vast majority of real transparent materials, like glass, crystal, etc. But we also directly validate the generalization of RCNet to other IoRs without re-training. As shown in 5, RCNet performs well on other IoRs. Although the error in edge areas increases, the results in most regions remain accurate. Please note that we do not calculate and compare the error metrics, because these samples have different valid regions caused by different IoRs. But the visual results can illustrate the generalization. In addition, directly expanding the training data to a wider IoR range should further enhance performance on various IoRs, if such requirements arise in subsequent work.

Despite assuming the infinite-far environment lighting is convenient enough for common users, our method can be adapted to handle more complex scenes. Although trained exclusively on infinite-far backgrounds, RCNet demonstrates remarkable generalization capability to nearby background scenarios. We evaluated it on the dataset in Bemana et al. [2022] by removing the object via a bounding box to obtain background images. As illustrated in the Fig. 7, RCNet achieves highly accurate correspondence estimation. With a simple extension that accounts for background distance when converting correspondences to ray directions, our method can effectively handle such cases.

We also present a comparison of object shapes before and after our optimization, which further demonstrates the effect of our method, as shown in Fig. 6.

3 Failure Case

Benefiting from data prior, our method can effectively deal with various backgrounds, including those lacking texture, like the "BasicShape1" and "Mooncake" in Fig. 6 of the main paper. But extremely texture-less backgrounds would pose challenges to our method, as shown in 8. Despite the warped image being close to the input, the predicted correspondence does not accurately reproduce the fine changes within the object, since the large areas of pure green ground provide very little neighborhood information and cause ambiguity.

Beyond the typically intricate structures prone to total internal reflection, we observed that the combination of inclined pedestal and legs in the "Real Dog" happens to frequently exhibit total internal reflection, making these regions challenging for our method to reconstruct accurately, as shown in Fig. 9. But please note this region is also challenging for other SOTA, and our method still demonstrates superior performance, as the visual results in Fig. 9 and metrics in the main paper.

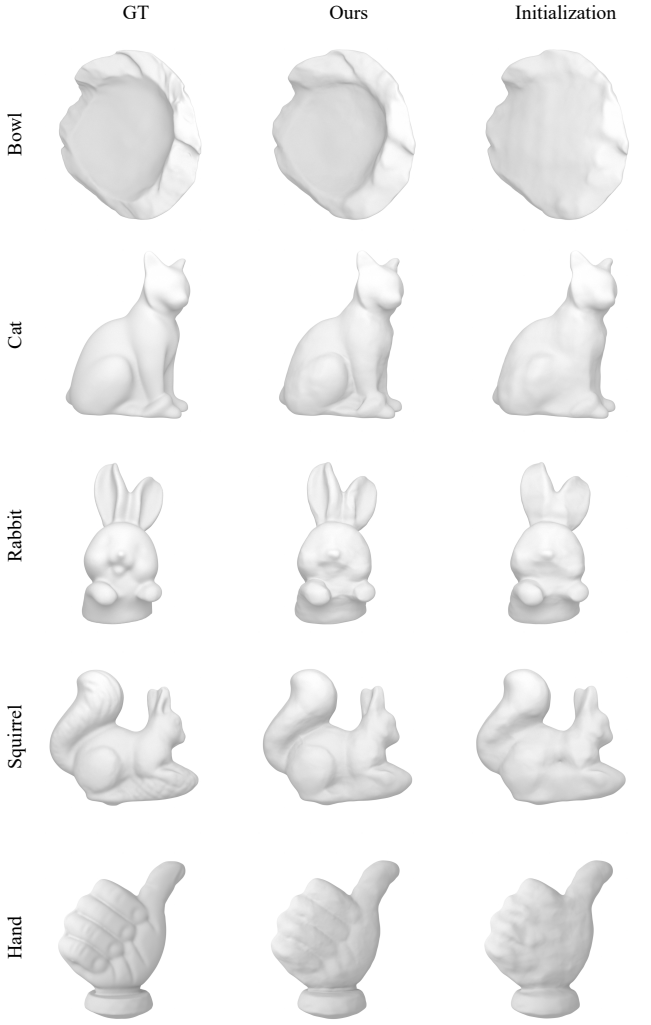

Fig. 6. Our reconstruction results on synthetic data, compared with our initial shapes and ground-truth. Our method significantly recovers the concave areas of various objects.

Fig. 7. The refractive correspondence estimation on the near background scene in Bemana et al. [2022]. Our method still produces accurate correspondence, demonstrated by the warped images. We show the object-free background image on the top left of the warped image for reference.

References

Mojtaba Bemana, Karol Myszkowski, Jeppe Revall Frisvad, Hans-Peter Seidel, and Tobias Ritschel. 2022. Eikonal fields for refractive novel-view synthesis. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–9.

Fangzhou Gao, Lianghao Zhang, Li Wang, Jiamin Cheng, and Jiawan Zhang. 2023. Transparent Object Reconstruction via Implicit Differentiable Refraction Rendering. In *SIGGRAPH Asia 2023 Conference Papers*. 1–11.

Zachary Teed and Jia Deng. 2020. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 402–419.

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In *Advances in Neural Information Processing Systems*.

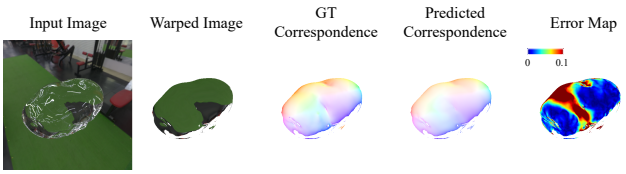

Fig. 8. The correspondence estimation on extremely texture-less backgrounds. The accuracy decrease when the background provides very little neighborhood information.

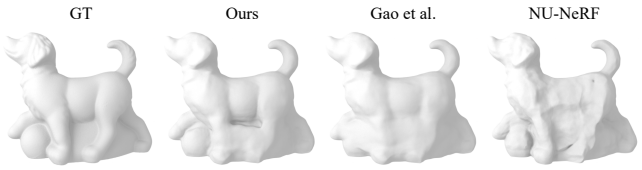

Fig. 9. Our method fails to accurately recover the pedestal and legs in the "Real Dog" due to the total reflection. But it still outperforms other methods.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. 2023. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 11 (2023), 13941–13958.

Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. 2020. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492* (2020).