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

FANGZHOU GAO, Tianjin University, China
YUZHEN KANG, Tianjin University, China
LIANGHAO ZHANG, Tianjin University, China
LI WANG, Tianjin University, China
QISHEN WANG, Tianjin University, China
JIAWAN ZHANG*, Tianjin University, China

Fig. 1. We present RCTrans, capable of reconstructing precise geometry from multi-view images in uncontrolled natural environments. With the convenient acquisition setup shown on the left, our method achieves accurate geometric reconstruction as demonstrated on the right.

Transparent object reconstruction in an uncontrolled natural scene is a challenging task due to its complex appearance. Existing methods optimize the object shape with RGB color as supervision, which suffer from locality and ambiguity, and fail to recover accurate structures. In this paper, we present RCTrans, which uses ray-background intersection as a more efficient constraint to achieve high-quality reconstruction, while maintaining a convenient setup. The key technology to achieve this is a novel pre-trained correspondence estimation network, which allows us to acquire ray-background correspondence under uncontrolled scenes and camera views. In addition, a confidence evaluation is introduced to protect the reconstruction from inaccurate estimated correspondence. Extensive experiments on both synthetic and real data demonstrate that our method can produce highly accurate results, without any extra acquisition burden. The code and dataset will be publicly available.

CCS Concepts: • **Computing methodologies** → **Reconstruction**; **Mesh geometry models**.

*Corresponding authors.

Authors' Contact Information: Fangzhou Gao, Tianjin University, China, gaofangzhou@tju.edu.cn; Yuzhen Kang, Tianjin University, China, yu_zhen@tju.edu.cn; Lianghao Zhang, Tianjin University, China, lianghaozhang@tju.edu.cn; Li Wang, Tianjin University, China, li_wang@tju.edu.cn; Qishen Wang, Tianjin University, China, wangqishen@tju.edu.cn; Jiawan Zhang, Tianjin University, China, jwzhang@tju.edu.cn.

SA Conference Papers '25, Hong Kong, Hong Kong

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

Additional Key Words and Phrases: Transparent Object, Multi-view Reconstruction, Correspondence Estimation

ACM Reference Format:
Fangzhou Gao, Yuzhen Kang, Lianghao Zhang, Li Wang, Qishen Wang, and Jiawan Zhang. 2025. RCTrans: Transparent Object Reconstruction in Natural Scene via Refractive Correspondence Estimation. In *SIGGRAPH Asia 2025 Conference Papers (SA Conference Papers '25), December 15–18, 2025, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3757377.3763859

1 Introduction

Due to its unique appearance, the reconstruction of transparent objects has been a long-standing problem. The complex light paths composed of multiple reflections and refractions are beyond the scope of common approaches. Through employing a controlled environment with specialized equipment, previous methods can obtain accurate information about the real refracted rays and achieve high-precision reconstruction [Li et al. 2023; Lyu et al. 2020; Wu et al. 2018; Xu et al. 2025]. However, the complicated acquisition requirements limit their practical application.

Another line of research attempted to relax the acquisition setup and reconstruct transparent objects under uncontrolled natural lighting. Among these, ray tracing-free methods circumvent the explicit refractive ray tracing and instead predict object appearance with the ray bending network [Wang et al. 2023] or the refraction component network [Sun et al. 2024]. While the network's strong fitting capacity can directly produce plausible appearances, it causes the shape-radiance ambiguity [Zhang et al. 2020], impeding accurate geometry recovery.

In contrast, ray tracing-based methods model the appearance with physically-based rendering [Deng et al. 2024; Gao et al. 2023; Li et al.

2020]. They differentiably trace rays inside the object and render its appearance alongside natural lighting. The object geometry determines the light path and appearance through physical laws, and in turn, it can be better optimized through color supervision. However, the inherent color ambiguity in uncontrolled lighting makes it difficult for these methods to recover detailed shape, especially on concave areas. Therefore, achieving high-precision transparent object reconstruction in an uncontrolled environment remains an open challenge.

In this paper, we propose RCTrans, a novel method for high-quality transparent object reconstruction under uncontrolled natural lighting. In contrast to the color supervision that suffers from locality and ambiguity, RCTrans uses ray-background intersection to constrain the light path and object shape much more efficiently. However, acquiring the true ray-background intersection in an random environment is a highly challenging task. The key observation to solve this problem is that, despite the presence of refractive distortion, the appearance of a transparent object provides sufficient neighborhood information, which can be leveraged to match accurate correspondence with a natural background. It enables RCTrans to eliminate color ambiguity and obtain ray-background intersection under uncontrolled environments and camera views, while keeping a convenient setup.

Technically, to efficiently leverage various neighborhood information, RCTrans introduces a novel neural network to match correspondences between the transparent object and the natural background. Benefiting from the data prior and our design, the network can handle the complex appearances of transparent objects and infer precise correspondence using refraction-distorted neighborhood information. During object reconstruction, RCTrans recovers the background image from multi-view inputs and uses the pre-trained network to estimate correspondence for each view. These estimated correspondences are further used to optimize the refraction light path and object shape through physically based ray tracing. Furthermore, to handle errors in estimated results, we evaluate the result confidence at the test time, based on the error of warped images. This enables RCTrans to filter out inaccurate correspondence and incorporate only reliable constraints into the reconstruction process, achieving high-precision reconstruction.

In summary, this paper proposes a novel method for high-accuracy reconstruction of transparent objects in unknown natural environments. Its superior performance is attributed to the following technical contributions:

- A novel neural network for estimating correspondences between transparent objects and their backgrounds under natural lighting.
- A confidence evaluation for estimated correspondence to prevent the impact of erroneous correspondence on the reconstruction process.

Experiments on both synthetic and real data demonstrate the superiority of our proposed method. While maintaining a convenient acquisition process, RCTrans produces higher-quality reconstructions, particularly on complex concave areas.

2 Related Work

We first discuss research on transparent object reconstruction, which is divided into those requiring specialized equipment and controlled environments, and those in an uncontrolled natural environment. Furthermore, since a core of our method is to estimate correspondence for transparent objects, we also review previous research on environment matting, which also produces such correspondence.

2.1 Transparent Object Reconstruction

Controlled Environment. Given the complex optical properties of transparent objects, several studies have incorporated specialized equipment and techniques for reconstruction, including polarization cameras [Miyazaki and Ikeuchi 2005; Shao et al. 2024], tomography [Trifonov et al. 2006], light probes [Wetzstein et al. 2011], ToF camera [Tanaka et al. 2016] and thermal camera [Narayanan et al. 2024]. Additionally, several approaches have leveraged surface reflections under controlled lighting for transparent object reconstruction [Liu et al. 2014; Morris and Kutulakos 2007; Yeung et al. 2011b].

Another prominent line of research focuses on analyzing and reconstructing refractive light paths using conventional RGB cameras. Kutulakos and Steger [2008] analyzed the refraction light path in transparent objects and adapted the triangulation to transparent surface reconstruction. By utilizing pre-designed patterns to establish correspondence between camera rays and the background, subsequent work achieved reconstruction of single and double transparent surfaces [Qian et al. 2016; Shan et al. 2012].

Wu et al. [2018] extended the setup to multi-view and proposed the first method to reconstruct a full model of a transparent object. They employ a turntable and display to capture multi-view refractive ray directions, and then optimize a point cloud to achieve physically accurate refraction. Based on this setup, Lyu et al. [2020] introduced mesh-based differentiable refraction ray tracing, which recovers more detailed geometry while reducing the acquisition process by half. Implicit SDF field is also explored for more robust optimization and reconstruction [Li et al. 2023]. Xu et al. [2025] further extends it to natural lighting, with an iPad showing designed patterns.

Unlike these approaches, our method does not rely on extra equipment to control the capturing environment, making it more convenient for common users.

Uncontrolled Environment. Recently, there have been several works that focus on reconstructing transparent objects in natural scenes without controlled setups. Li et al. [2020] introduced a physically-based neural network that reconstructs transparent shapes under natural lighting but requires pre-captured environment maps and manually annotated silhouettes. To simplify the setup, Gao et al. [2023] proposes to project the neural field back to input views for automatically obtaining multi-view silhouettes, which is applied in our method to obtain object silhouettes and environment lighting. They further optimize the implicit object geometry through refractive rendering, along with the scene that is represented as a texture. Similarly, TNSR [Deng et al. 2024] represents the scene as a neural field and refines the object shape through physically-based rendering. Although they use volume rendering, the multiple refractive light path makes the final color still sampled from only one point, leading to local gradients and inaccurate reconstruction results.

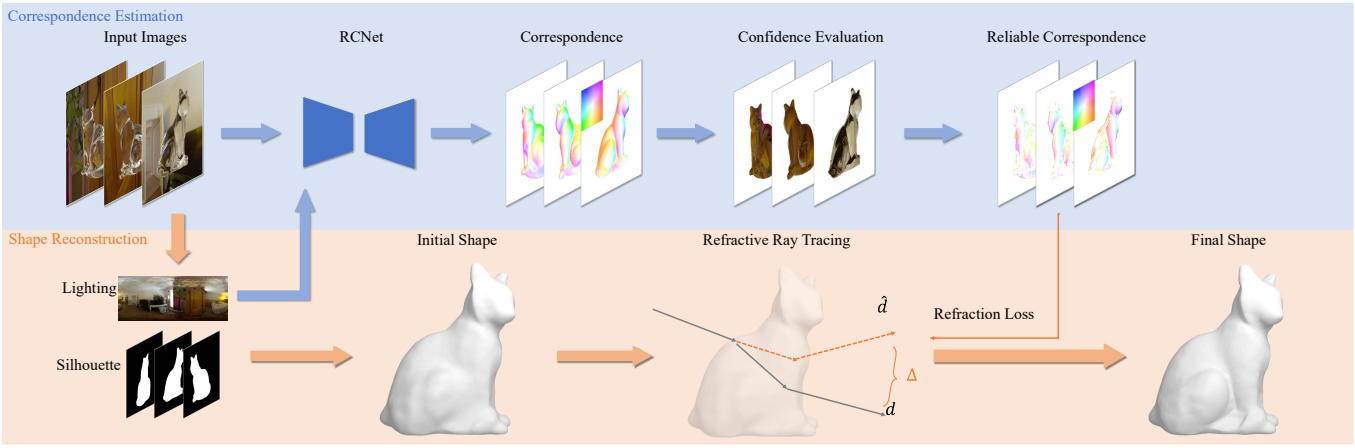

Fig. 2. The overview of RCTrans. Starting from the multi-view images, RCTrans first leverages a pre-trained network and recovered environment lighting to estimate ray-background correspondence for each view. The following confidence evaluation filters those inaccurate results, while the remaining reliable correspondences are used to optimize the initial object shape through differentiable refractive ray tracing, leading to a precise shape. The correspondence is visualized in a similar way to optical flow, with its color map shown in the top left corner.

NEMTO [Wang et al. 2023] and NU-NeRF [Sun et al. 2024] adopt a neural rendering approach to directly predict refractive colors or directions with a network, circumventing the explicit modeling of refractive light paths. While these methods can generate plausible appearance reconstructions, they struggle to recover concave geometric details due to the lack of physical refraction modeling.

In this paper, we propose a high-quality reconstruction method for thick transparent objects with obvious refraction distortion. The thin transparent object reconstruction and related research [Wu et al. 2025; Zhang et al. 2025] are beyond our scope.

2.2 Environment Matting

Environment matting aims to model how a foreground transparent object interacts with its background, enabling image synthesis through background replacement. The pioneering work [Zongker et al. 1999] assumed that each foreground pixel corresponds to a rectangular background region and employed a series of horizontal and vertical gray code patterns to establish these correspondences. Chuang et al. [2000] subsequently proposed two improvements, achieving higher precision or real-time environment matting. Subsequent research explored background patterns in frequency-domain [Qian et al. 2015] and wavelet-domain [Peers and Dutré 2003] for more efficient acquisition. To avoid the pre-designed patterns, Chen et al. [2018] proposed a convolutional neural network to regress an environment matte from a single image. But they aim to produce a visually realistic refractive effect, instead of estimating a highly accurate correspondence.

Moreover, these methods are limited to the darkroom, where phenomena such as total internal reflection are significantly simplified. Yeung et al. [2011a] proposed an approach for environment matting under natural lighting, while relying on manual annotation and only producing visually pleasing results instead of accurate correspondence.

In contrast, we propose a novel method to estimate accurate correspondence for transparent objects under natural lighting.

3 Method

3.1 Overview

Assuming a solid transparent object under unknown natural distant lighting, the target of RCTrans is to recover its geometry from multi-view RGB images. The key to high-quality reconstruction lies in recovering a shape that refracts light rays to the same background locations as in the real case.

As shown in the upper part of Fig. 2, RCTrans first recovers object-free environment lighting from multi-view images, and introduces a pre-trained network, RCNet, to estimate the correspondence between input images and backgrounds. By leveraging the neighborhood information and data prior, RCNet can estimate accurate and dense correspondences. RCTrans further introduces a confidence evaluation to filter inaccurate estimation results at test time, while the remaining reliable correspondence explicitly indicates the true intersection of refracted rays and background.

Given the object shape initialized by estimated object silhouettes, RCTrans further leverages the correspondence to optimize it through differentiable refractive ray tracing, as shown in the lower part of Fig. 2. It enforces the refracted ray to the true direction as the correspondence, therefore leading to a final high-quality shape.

In the remainder of this section, we first present the correspondence estimation (Sec 3.2), followed by the description of the complete object reconstruction process (Sec 3.3).

3.2 Refractive Correspondence Estimation

Model Formulation. Due to refractive properties, a transparent object presents a distorted version of the background. We begin by thoroughly analyzing and formulating the relationship between the transparent object's appearance and background, establishing the foundation for subsequent correspondence estimation.

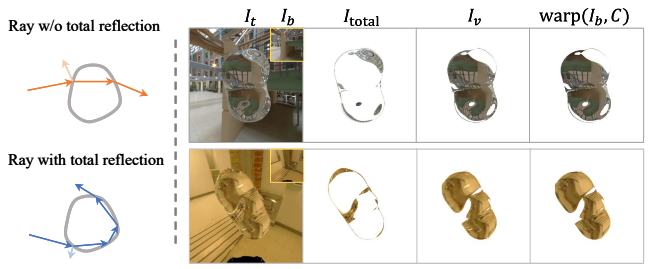

Fig. 3. The left part illustrates the light path in a transparent object when total reflection happens or not, and the right part is the different components in our appearance formulation. For reference, the background image I_b is shown at the top right corner of I_t.

Under natural illumination, the appearance of a colorless transparent object I_t can be divided into two components, as

$$I_t = I_{\text{total}} \oplus I_v, \tag{1}$$

where I_{total} represents the partial image where total reflection occurs, I_v is the remaining image without total reflection, and $\oplus$ is the pixel-wise combination to concatenate images. We separately model the appearances because of the distinct imaging model caused by total reflection. As shown in the left Fig. 3, when total internal reflection occurs, the ray continues to propagate within the object and often undergoes total reflection again, resulting in complex light paths that randomly intersect the entire illumination rather than the background. Moreover, these regions display intense specular highlights with drastic variations, providing minimal neighborhood information. Thus, I_{total} is modeled separately and excluded from correspondence estimation.

Like previous works [Chen et al. 2018; Chuang et al. 2000], we assume each surface point refractively maps a single background point. Then, according to the physically based rendering, I_v can be further formulated as:

$$I_v = (1-\rho)I_r + \rho \text{warp}(I_b, C), \tag{2}$$

,where ρ is the transmission coefficient determined by the Fresnel equation, I_r is the reflection image, I_b is the background image, C is the 2D correspondence between I_v and I_b caused by refraction, which is termed as "refractive correspondence" in this paper, and warp is the warping operation to map I_b with C.

Given the surface continuity, I_v preserves rich neighborhood information, like the distorted railing and stairs shown in the right Fig. 3, which can support inversely estimating C by matching I_v and I_b. Although I_r exists, its interference is negligible since the ρ is close to 1 in most cases. Only direct light sources would leave faint imprints that minimally affect structural information, demonstrated by the highly similar I_v and warp(I_b, C) in the right Fig. 3.

Technically, we use a neural network, RCNet, to achieve the correspondence estimation, which can leverage data prior to robustly handle various cases.

RCNet. However, estimating C is still challenging since it is difficult to separate I_v and I_{total} from I_t, and C is only well-defined within I_v. To solve it, we set RCNet to directly take I_t and I_b as input and output full-image correspondences, which are post-filtered

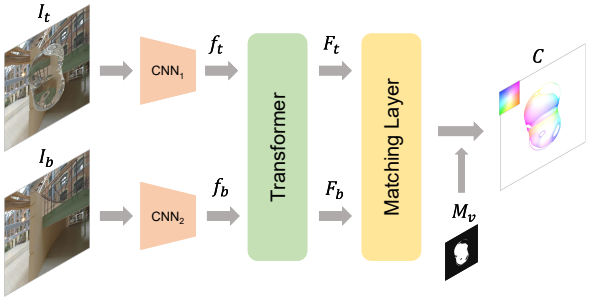

Fig. 4. The network architecture of RCNet. The features extracted from the CNN and transformer are sent to the matching layer to calculate the final correspondence, which is filtered by M_v during training. The correspondence is visualized in a similar way to optical flow, with its color map shown in the top left corner.

for selective supervision within I_v. This design not only maintains easily accessible input but also enables the network to focus on correspondence estimation, without explicitly processing I_{total}.

As shown in Fig. 4, considering the distinct appearance distribution shown in transparent objects and natural images, RCNet first employs two separate CNNs to extract features f_t, f_b from I_t and I_b, respectively. Then it uses the transformer and matching layer in the optical flow estimation GMFlow+ [Xu et al. 2023] to output 2D correspondence. f_t and f_b are fed to a transformer module to model their correlation and get the enhanced features F_t, F_b, respectively representing I_t and I_b. Then C can be globally matched by computing the feature similarity, as:

$$\begin{aligned} S &= \frac{F_t F_b^T}{\sqrt{D}} && \in \mathbb{R}^{H\times W\times H\times W} \\ C &= \text{softmax}(S)G && \in \mathbb{R}^{H\times W\times 2}, \end{aligned} \tag{3}$$

where the 4D matrix S stores the similarity between every element pair in F_t and in F_b, and D is the number of feature channels to prevent large values. The function softmax is applied to the last two dimensions so that the similarities between an element in F_t and all elements in F_b can be normalized to a probability distribution, and G is the coordinate grid. This process searches the whole image for each point to match the optimal correspondence, which can efficiently handle the large displacement caused by refraction.

As mentioned above, during training, the output C would be masked by the GT M_v, which indicates the regions of I_v and is termed as "valid mask" in this paper. Then the selected C is supervised with GT correspondence. As the common strategy in optical flow estimation [Teed and Deng 2020], we also supervise the intermediate predicted correspondence. More details about the intermediate results and the network can be found in the supplementary materials. Finally, RCNet is trained with a masked L2 loss, as

$$L_{\text{net}} = \sum_{i=1}^{N} \gamma^{N-i} \frac{\sum M_v \|C_i - C_{\text{gt}}\|_2}{\sum M_v}, \tag{4}$$

where N is the number of predictions, C_i is the ith predicted correspondence, and γ (set to 0.9) is the weight that gives higher weights

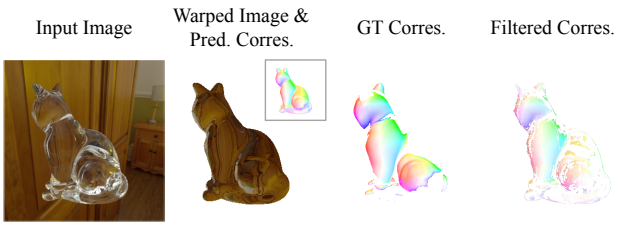

Fig. 5. The effect of confidence evaluation. We show the original predicted correspondence and its warped image (filtered by the object mask for clarity), and filter those with an average RGB error larger than 0.1. The filtered correspondence is close to the ground truth.

for later predictions. The masked supervision is essential for network training, while simple default value filling would greatly interfere with the matching process and cause the training to crash.

Test and Result Evaluation. Once trained, RCNet can estimate correspondences for transparent objects. The input I_b can be easily approximated by recovering the environment lighting from multi-view images, without requiring additional capture. However, the inaccessible M_v introduces new challenges. Without M_v filtering, the output C is a full-image 2D correspondences that erroneously include values for total reflection regions. These incorrect estimations would introduce faulty constraints in the subsequent reconstruction.

To mitigate this issue, we evaluate the confidence of estimated correspondences at test time to filter out unreliable predictions. Specifically, since total reflection regions do not have clear correspondences with the background and are never supervised during training, the predicted correspondence cannot even maintain appearance consistency. Leveraging it, we assess prediction confidence by warping the I_b with estimated C and measuring the reconstruction error against the input image I_t, as.

$$S_c = 1 - \|\text{warp}(I_b, C) - I_t\|_1, \tag{5}$$

, where three-channel color error is averaged to get a scalar confidence score. When predicting for the transparent object reconstruction, we filter the results with confidence lower than 0.9 and use the remaining as constraints.

As shown in Fig. 5, the correspondence on total reflection regions can be approximately filtered. Beyond the total reflection, this process also handles other errors, like out-of-boundary correspondences and network estimation failures, efficiently protecting the following reconstruction from incorrect constraints. Although ignoring reflections may lead to the erroneous rejection of some accurate correspondences, our analysis above demonstrates that strong reflections occupy only minimal regions, thus exerting a negligible impact on multi-view reconstruction.

3.3 Transparent Object Reconstruction

Starting with the multi-view RGB images $\{I_t\}$, RCTrans first prepares the necessary data for reconstruction. Following Gao et al. [2023], it uses neural rendering to simultaneously model the object and environment lighting from input images, and then project the neural field to get object silhouettes. It does not require any extra capture and keeps it convenient for common users. Then, per-view object-free background images $\{I_b\}$ are rendered with the recovered lighting, which is fed to the pre-trained RCNet together with $\{I_t\}$ to get reliable $\{C\}$. We use a neural SDF field N_{obj} [Wang et al. 2021] to represent the object shape and initialize it with silhouettes. We detail the preparation in the supplementary materials.

After shape initialization, RCTrans optimizes N_{obj} to recover an accurate shape by aligning its refracted rays with the estimated correspondences. As in the work of Gao et al. [2023], we use the linear interpolation by Fu et al. [2022] to locate the intersection of the ray and the implicit shape, and analytically refract the ray according to Snell's law and the refractive index of object, which is modeled as a homogeneous value and optimized along with N_{obj}. Rays that undergo exactly two refractions without total internal reflection are recorded to be supervised with correspondences.

For convenience, local per-view correspondences are transferred into the unified world coordinate system. Under the infinite-distance lighting assumption, the 2D correspondence C is transferred into the ray direction $\hat{d}$ as:

$$\hat{d} = \text{norm}(R^T K^{-1} \tilde{C}^T), \tag{6}$$

where norm is the L2 normalization, R is the rotation matrix from world coordinate system to camera coordinate system, K is the intrinsic matrix, and $\tilde{C}$ represents C in homogeneous coordinates. We adopt the infinite-distance lighting assumption for convenience. But our method can also be adapted for nearby environments, as discussed in the supplementary material.

Then $\hat{d}$ is used to supervise the traced ray directions d through a refraction loss defined as:

$$L_{\text{refraction}} = \frac{\sum \overline{M}_v \|\hat{d} - d\|_1}{\sum \overline{M}_v}, \tag{7}$$

where $\overline{M}_v$ is the mask where the ray both undergoes refraction twice and has a valid correspondence. Although a color loss can be added through rendering with environment lighting, we found that it barely changes the reconstruction results since the correspondence is accurate enough.

We also add two regularizations, including the silhouette loss to constrain the silhouette and the eikonal loss [Gropp et al. 2021] to get natural and smooth shapes. The final loss term is defined as:

$$L = \lambda_{\text{ref}} L_{\text{refraction}} + \lambda_{\text{sil}} L_{\text{silhouette}} + \lambda_{\text{eik}} L_{\text{eikonal}}, \tag{8}$$

where $\lambda_{\text{ref}}, \lambda_{\text{sil}}, \lambda_{\text{eik}}$ are set as 1.0, 1.0, 1.5, respectively.

4 Dataset Creation

We use Mitsuba3 [Nimier-David et al. 2019] to render a large synthetic dataset for training RCNet. With randomly combined transparent objects, environment map and viewpoints, I_t, I_b, C, M_t are generated under the same camera parameters. Models and environment maps are divided into the training and validation sets, ensuring that any element in the validation data is unseen during the training. In the end, there are 80,000 groups of data for training and 200 groups for validation, at the resolution of 512×512.

Transparent Object. We use the 3D basic models generated by Li et al. [2020], which are constructed by combining basic geometric shapes. To further enhance the model complexity, we also collect

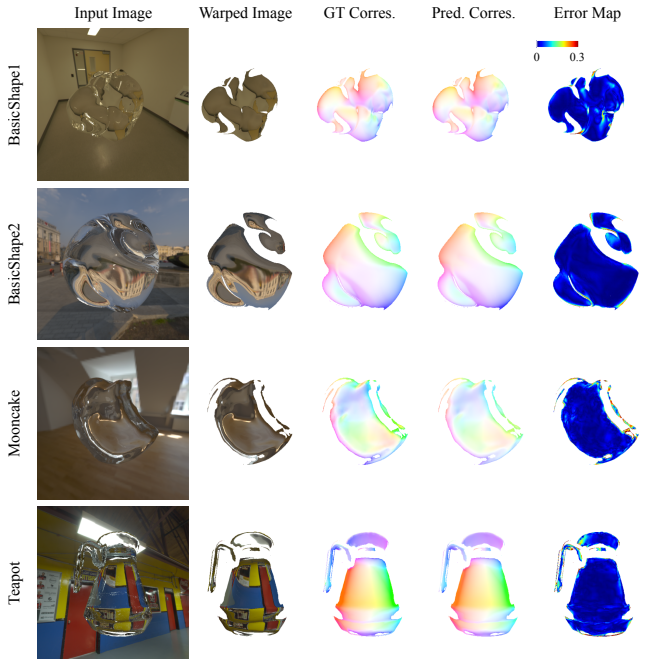

Fig. 6. The correspondence results on the validation set, filtered with the GT valid mask. The upper two samples are from the basic shape set, and the lower two are from the OmniObject3D set. The correspondence coordinates are normalized to $[0, 1]$. Besides the EPE error map of predicted correspondence, we also show the warped image to prove its accuracy. The empty areas of the images are cropped for layout.

the models in OmniObject3D [Wu et al. 2023], which are scanned from opaque real-life objects. We manually select thick models that can generate obvious refraction distortion, and divide them into training and validation sets according to the category, avoiding data leakage of models with similar structures. And we perform simplification, repairing and smoothing to remove damaged faces and noise in the scanned models. Finally, there are 4900 models for training and 200 for validation. Every model is normalized to the unit cube and randomly assigned a refractive index value between 1.3 and 1.6. We discuss the generalization to other refractive indices in the supplementary material.

Environment Map. There is a total of 3200 environment maps, with 1000 from PolyHeaven [2024] and 2200 from the Laval Indoor HDR dataset [Gardner et al. 2017]. Among these, 3000 maps are used for the training set, while 200 are reserved for the validation.

GT Correspondence. The correspondence is obtained through a simplified ray-tracing process. For each pixel, we trace and compute the refractive path of its ray until the ray intersects the background. During this process, rays that do not undergo total reflection are recorded to generate the valid mask. 2D correspondence is converted from the final exit ray directions according to the camera parameters.

Table 1. The quantitative correspondence results on the validation set. The correspondence coordinates are normalized to $[0, 1]$. EPE is the Euclidean distance between GT and predicted correspondence. P5 represents the percentage of pixels with an EPE exceeding 0.05, and P10 represents the percentage of EPE exceeding 0.1.

Method	BasicShape			OmniObject3D		
	EPE ↓	P5↓	P10↓	EPE↓	P5↓	P10 ↓
Ours	0.026	0.138	0.061	0.060	0.351	0.188
Baseline	0.092	0.668	0.346	0.122	0.713	0.467

5 Experiments

We evaluate our methods on both synthetic and real data, compared with various baselines and SOTA methods. For correspondence estimation, since there is no existing method, we construct a baseline according to the environment matting method TOM-Net+ [Chen et al. 2019], which formulates a similar problem as a regression task with U-Net [Ronneberger et al. 2015], instead of explicit matching. The baseline shares the same inputs and training iteration as RCNet, except the learning rate is 1e-3 as in the work of Chen et al. [2019].

As for the transparent object reconstruction, we compare SOTA uncontrolled transparent object reconstruction methods, including NU-NeRF [Sun et al. 2024], Gao et al. [2023], NEMTO [Wang et al. 2023] and TNSR [Deng et al. 2024]. Since there is no open-source code for Gao et al. [2023], we reproduce their method and adapt it to the environment lighting through a coarse-to-fine optimization with the recovered environment map. All methods share the same inputs for fair comparison, except for the extra required surrounding box in TNSR and the GT silhouettes and environment map in NEMTO.

5.1 Implementation Details

When training the RCNet, we follow the configuration by Xu et al. [2023], with a learning rate set as 2e-4 and a batch size set as 4. The training lasts 800,000 iterations, which takes around 55 hours on a single NVIDIA RTX 4090 GPU. As for the object reconstruction, we follow the training configuration and ray sampling strategy by Gao et al. [2023]. The shape optimization takes 30,000 iterations with a learning rate of 1e-5. The whole reconstruction process takes around 6 hours on a single NVIDIA RTX 4090 GPU, including data preparation and shape reconstruction.

5.2 Result on Synthetic Data

Besides the validation dataset for correspondence estimation, we render multi-view data for the object reconstruction evaluation. We collect models from diverse sources, including the "Rabbit" from the OminiObject3D validation set, real transparent objects "Hand" in prior work [Wu et al. 2018] and other models from web resources, to validate our method's generalization capability. These models all have complex geometrical structures with abundant concave details, which can demonstrate our reconstruction accuracy. Using different environment maps from the validation set, each model was rendered from 50-90 viewpoints at a resolution of 512×512.

Correspondence Estimation. We first separately evaluate RCNet on the correspondence estimation, presenting the quantitative results in Tab. 1 and visual results in 6. Although the more complex models

Table 2. The quantitative comparison and ablation study on synthetic data. The Chamfer distance ($\times 10^{-4}$) normalized by the bounding box diagonal is reported, and the best results are bolded. "Avg." is the average error. "Ours v1" represents our method without confidence evaluation, and "Ours v2" represents our method without refraction loss.

Methods	Bowl	Cat	Rabbit	Squirrel	Hand	Avg.
NEMTO	12.605	1.132	1.441	0.988	1.160	3.465
TNSR	12.728	1.579	1.331	3.131	1.307	4.015
NU-NeRF	7.892	1.955	1.815	2.144	0.775	2.916
Gao et al.	1.325	0.613	1.002	0.933	0.804	0.935
Ours	**0.528**	**0.243**	**0.622**	**0.453**	**0.339**	**0.437**
Ours v1	0.780	0.341	0.715	0.480	0.391	0.541
Ours v2	10.146	0.862	1.330	1.080	0.602	2.804

in OmniObject3D cause increased errors, RCNet can estimate highly accurate correspondence across diverse shapes and backgrounds, particularly avoiding large deviations in baseline. It also demonstrates strong robustness for texture-less backgrounds, such as the "BasicShape1" and "Mooncake" in Fig. 6.

We also evaluate RCNet on the reconstruction dataset. As the visual results in Fig. 5 and 7, despite the errors in some edge areas being relatively high, RCNet predicts accurate results for most areas. It greatly demonstrates the generalization of RCNet on different data distributions, confirming its effectiveness for transparent object reconstruction. We detail the quantitative results of the correspondence and confidence evaluation in the supplementary materials.

Shape Reconstruction. Then, we comprehensively evaluate RCTrans on transparent object reconstruction. As illustrated in Fig. 8, our method produces more accurate shapes with fine details. Despite other methods maintaining approximately correct silhouettes, they struggle to recover concave regions, since the ray tracing-free methods suffer from the shape-radiance ambiguity and the ray tracing-based methods are limited by the locality of color gradient. In contrast, our method accurately reconstructs these concave structures, from the large-scale depression in "Bowl" to the fine-scale finger gaps in "Hand", contributing to the efficient constraints provided by correspondence estimation.

The quantitative results in Tab. 2 further verify the superiority of our method. Our method achieves the lowest Chamfer distance across all objects, reducing the error by half compared to the second-best approach. It demonstrates the high precision and robustness of our method on diverse objects.

5.3 Result on Real Data

We collect diverse real data to verify the generalization of RCTrans. Each object is captured under 60-70 views, and the input images are resized to 512×512. The camera parameters are recovered by colmap [Schonberger and Frahm 2016], and the ground-truth shape is obtained through laser scanning the object after painting.

Since there is no ground truth correspondence for real data, we present the warped images as visual results in Fig. 9. Except for the unrecoverable regions caused by total reflection, the warped images faithfully reproduce the distorted background, which proves the excellent generalization ability of our method on real data.

Table 3. The quantitative comparison on real data. The Chamfer distance ($\times 10^{-4}$) normalized by the bounding box diagonal is reported, and the best results are bolded.

Method	Ashtray	Kitten	Cat	Squirrel	Dog	Avg.
Gao et al.	14.331	1.755	0.670	1.472	2.091	4.064
NU-NeRF	10.758	1.827	0.633	2.391	1.749	3.472
Ours	**0.367**	**0.839**	**0.504**	**0.521**	**0.914**	**0.629**

We further present the reconstruction results and compare them with Gao et al. [2023] and NU-NeRF, since TNSR and NEMTO require a different setup or additional inputs. The "Ashtray" in Fig. 9 greatly demonstrates our method's superiority. While other methods fail to recover the depression, our method accurately reconstructs it, mirroring its performance on the synthetic 'Bowl'. Other concave details, like the leg in "Kitten" and tail in "Squirrel", further confirm our advantages, along with the quantitative metrics in Tab. 3.

5.4 Ablation Study

We conduct the ablation study on synthetic data to verify the effectiveness of key components in our method. Since we use the same 3D representation and training configuration as in the work of Gao et al. [2023] but instead use correspondence supervision, it serves as a baseline to demonstrate the superiority and efficiency of correspondence supervision, compared with color supervision. We further validate other components in our method.

Besides the filtering shown in Fig. 5, we validate the effect of confidence evaluation on shape reconstruction. When removing the confidence evaluation and using all estimated correspondences for geometry optimization, we observed a degradation in reconstruction quality. Quantitative results in Tab. 2 and visual comparisons in Fig. 10 demonstrate that unreliable correspondences would lead to inaccurate object boundaries and geometry.

We further validate the effect of correspondence estimation by removing refraction loss during optimization. And our method would entirely fail to recover concave structures, as shown in Fig. 10. These experiments confirm that both correspondence estimation and confidence evaluation are essential for high-quality reconstruction.

6 Limitation and Future Work

While our method demonstrates superior accuracy compared to other SOTA, some intricate structures remain challenging to reconstruct accurately, such as decorative lines on the tail of "Squirrel" in Fig. 8 and the bowknot on back of "Real Cat" in Fig. 9. Since our approach relies on neighborhood information for correspondence inference, its precision would decrease in those tiny and isolated areas. And these areas are particularly prone to total internal reflection, presenting significant challenges for reconstruction, which is further discussed in the supplementary materials. Moreover, extremely texture-less backgrounds would also be challenging for correspondence estimation, which is detailed in the supplementary material. Addressing these limitations through network and reconstruction pipeline improvements would be a meaningful future work.

Since our method only optimizes rays that undergo refraction twice, it cannot handle nested objects like NU-NeRF [Sun et al. 2024].

Besides, our method currently assumes a colorless object. But it has the potential to be adapted, with an extended training dataset and modified confidence evaluation.

7 Conclusion

In this paper, we propose RCTrans, a novel method that leverages ray-background correspondence to reconstruct transparent objects in uncontrolled natural scenes. It contains a pre-trained neural network to estimate accurate correspondence for arbitrary transparent objects and natural backgrounds, without any extra equipment. Cooperating with a confidence evaluation, these estimated correspondences efficiently constrain the refractive light path, leading to high-quality results with precise concave areas. Extensive experiments on both synthetic and real data demonstrate that our method achieves superior accuracy compared to SOTA approaches while maintaining a convenient acquisition setup.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (62172295).

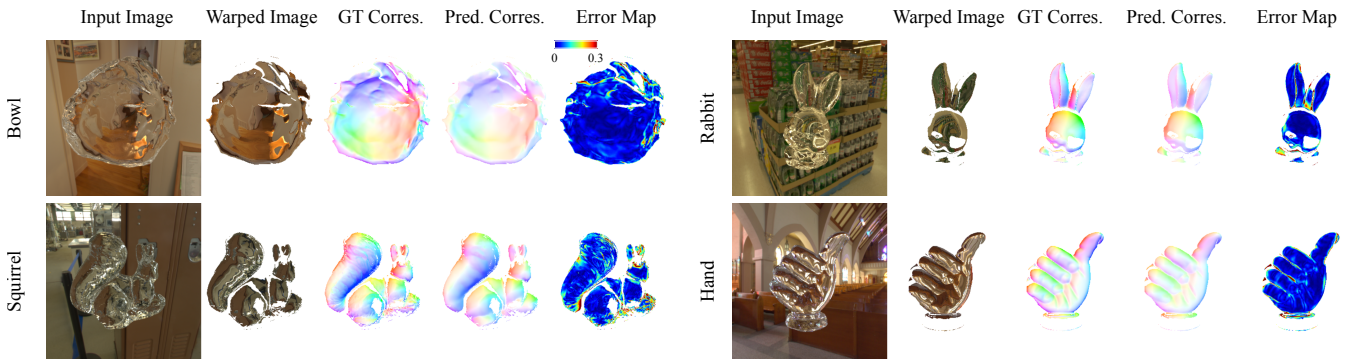

Fig. 7. The correspondence estimation results on the synthetic reconstruction dataset, with GT background as input and filtered with the GT valid mask. The error map of predicted correspondence and warped image are presented to prove the accuracy.

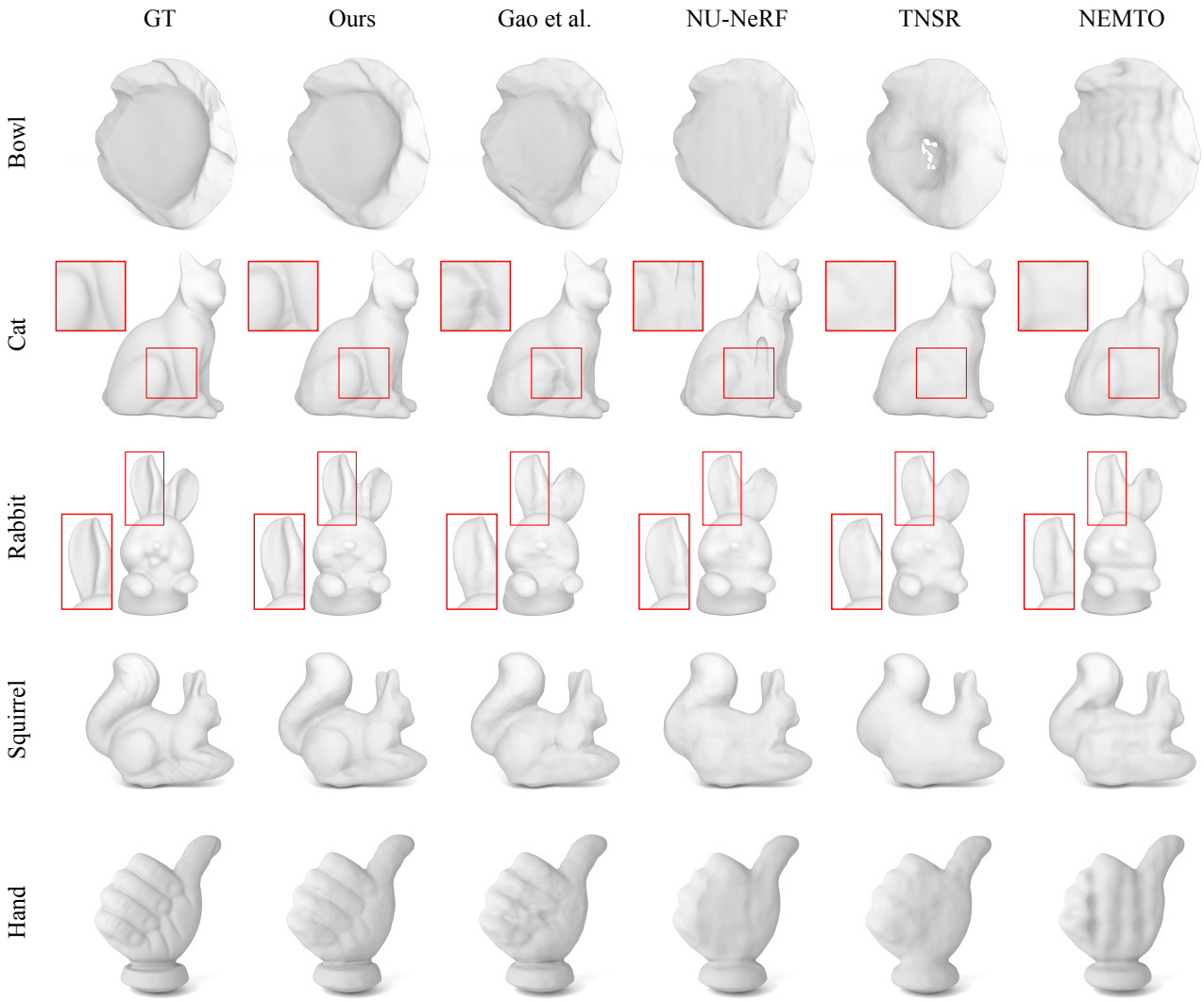

Fig. 8. The reconstruction results on synthetic data. Compared with SOTA methods, our method recovers more precise geometry, especially for those concave areas. Some concave areas are marked with the red box and zoomed in for clarity.

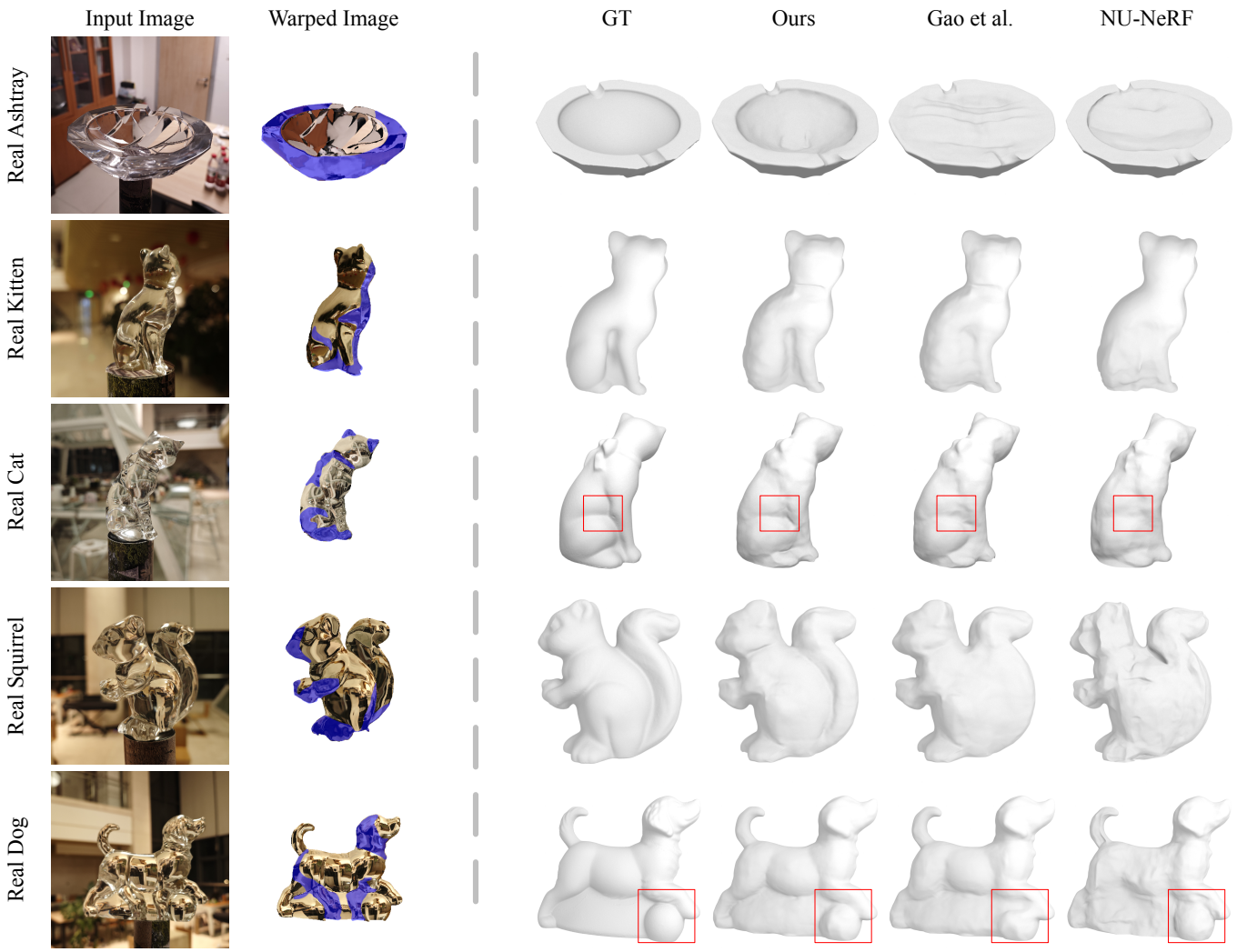

Fig. 9. The comparison of reconstruction results on real data. Our method recovers more accurate concave structures, as marked by the red box. We present the warped images to show the accuracy of the estimated correspondence. Based on our experience, we roughly annotated some total internal reflection regions with blue masks to facilitate comparative analysis of the remaining areas.

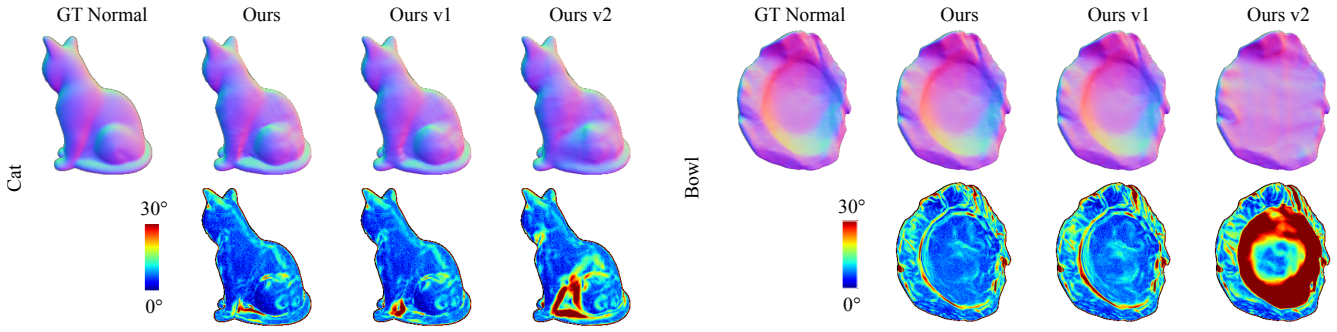

Fig. 10. The visual results of the ablation study. The normal results and their error maps are presented, with a color bar. “Ours v1” represents our method without confidence evaluation, and “Ours v2” represents our method without refraction loss.

References

Guanying Chen, Kai Han, and Kwan-Yee K Wong. 2018. Tom-net: Learning transparent object matting from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 9233–9241.

Guanying Chen, Kai Han, and Kwan-Yee K Wong. 2019. Learning transparent object matting. *International Journal of Computer Vision* 127, 10 (2019), 1527–1544.

Yung-Yu Chuang, Douglas E. Zongker, Joel Hindorff, Brian Curless, David H. Salesin, and Richard Szeliski. 2000. Environment matting extensions: towards higher accuracy and real-time capture. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques - SIGGRAPH '00*. doi:10.1145/344779.344844

Weijian Deng, Dylan Campbell, Chunyi Sun, Shubham Kanitkar, Matthew E Shaffer, and Stephen Gould. 2024. Differentiable Neural Surface Refinement for Modeling Transparent Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20268–20277.

Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. 2022. Geo-Neus: geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems* 35 (2022), 3403–3416.

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.

Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to Predict Indoor Illumination from a Single Image. *ACM Transactions on Graphics (SIGGRAPH Asia)* 9, 4 (2017).

Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. 2021. Implicit Geometric Regularization for Learning Shapes. In *37th International Conference on Machine Learning: ICML 2020, Online, 13-18 July 2020, Part 5 of 15*.

Poly Heaven. 2024. Poly Heaven. https://polyhaven.com/.

Kiriakos N Kutulakos and Eron Steger. 2008. A theory of refractive and specular 3D shape by light-path triangulation. *International Journal of Computer Vision* 76 (2008), 13–29.

Zongcheng Li, Xiaoxiao Long, Yusen Wang, Tuo Cao, Wenping Wang, Fei Luo, and Chunxia Xiao. 2023. NeTO: neural reconstruction of transparent objects with self-occlusion aware refraction-tracing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 18547–18557.

Zhengqin Li, Yu-Ying Yeh, and Manmohan Chandraker. 2020. Through the looking glass: Neural 3d reconstruction of transparent shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1262–1271.

Ding Liu, Xida Chen, and Yee-Hong Yang. 2014. Frequency-based 3d reconstruction of transparent and specular objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 660–667.

Jiahui Lyu, Bojian Wu, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. 2020. Differentiable refraction-tracing for mesh reconstruction of transparent objects. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–13.

Daisuke Miyazaki and Katsushi Ikeuchi. 2005. Inverse polarization raytracing: estimating surface shapes of transparent objects. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 2. IEEE, 910–917.

Nigel JW Morris and Kiriakos N Kutulakos. 2007. Reconstructing the surface of inhomogeneous transparent scenes by scatter-trace photography. In *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 1–8.

Sriram Narayanan, Mani Ramanagopal, Mark Sheinin, Aswin C Sankaranarayanan, and Srinivasa G Narasimhan. 2024. Shape from Heat Conduction. In *European Conference on Computer Vision*. Springer, 426–444.

Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. 2019. Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–17.

Pieter Peers and Philip Dutré. 2003. Wavelet environment matting. In *Proceedings of the 14th Eurographics workshop on Rendering*. 157–166.

Yiming Qian, Minglun Gong, and Yee-Hong Yang. 2015. Frequency-based environment matting by compressive sensing. In *Proceedings of the IEEE International Conference on Computer Vision*. 3532–3540.

Yiming Qian, Minglun Gong, and Yee-Hong Yang. 2016. 3D Reconstruction of Transparent Objects with Position-Normal Consistency. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr.2016.473

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Johannes L. Schonberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr.2016.445

Qi Shan, S. Agarwal, and B. Curless. 2012. Refractive height fields from single and multiple images. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. doi:10.1109/cvpr.2012.6247687

Mingqi Shao, Chongkun Xia, Dongxu Duan, and Xueqian Wang. 2024. Polarimetric inverse rendering for transparent shapes reconstruction. *IEEE Transactions on Multimedia* (2024).

Jia-Mu Sun, Tong Wu, Ling-Qi Yan, and Lin Gao. 2024. NU-NeRF: Neural Reconstruction of Nested Transparent Objects with Uncontrolled Capture Environment. *ACM Transactions on Graphics (TOG)* 43, 6 (2024), 1–14.

Kenichiro Tanaka, Yasuhiro Mukaigawa, Hiroyuki Kubo, Yasuyuki Matsushita, and Yasushi Yagi. 2016. Recovering Transparent Shape from Time-of-Flight Distortion. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. doi:10.1109/cvpr.2016.475

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.

Borislav Trifonov, Derek Bradley, and Wolfgang Heidrich. 2006. Tomographic reconstruction of transparent objects. In *ACM SIGGRAPH 2006 Sketches*. 55–es.

Dongqing Wang, Tong Zhang, and Sabine Süsstrunk. 2023. NEMTO: Neural environment matting for novel view and relighting synthesis of transparent objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 317–327.

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*.

Gordon Wetzstein, David Roodnick, Wolfgang Heidrich, and Ramesh Raskar. 2011. Refractive shape from light field distortion. In *2011 International Conference on Computer Vision*. IEEE, 1180–1186.

Bojian Wu, Yang Zhou, Yiming Qian, Minglun Cong, and Hui Huang. 2018. Full 3D reconstruction of transparent objects. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–11.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. 2023. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 803–814.

Tianhao Walter Wu, Fangcheng Zhong, Gernot Riegler, Shimon Vainer, Jiankang Deng, Cengiz Oztireli, et al. 2025. αsurf: Implicit surface reconstruction for semi-transparent and thin objects with decoupled geometry and opacity. In *International Conference on 3D Vision 2025*.

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.

Jiamin Xu, Zihan Zhu, Hujun Bao, and Weiwei Xu. 2025. Hybrid mesh-neural representation for 3D transparent object reconstruction. *Computational Visual Media* 11, 1 (2025), 123–140.

Sai-Kit Yeung, Chi-Keung Tang, Michael S Brown, and Sing Bing Kang. 2011a. Matting and compositing of transparent and refractive objects. *ACM Transactions on Graphics (TOG)* 30, 1 (2011), 1–13.

Sai-Kit Yeung, Tai-Pang Wu, Chi-Keung Tang, Tony F Chan, and Stanley Osher. 2011b. Adequate reconstruction of transparent objects on a shoestring budget. In *CVPR 2011*. IEEE, 2513–2520.

Haoran Zhang, Junkai Deng, Xuhui Chen, Fei Hou, Wencheng Wang, Hong Qin, Chen Qian, and Ying He. 2025. From Transparent to Opaque: Rethinking Neural Implicit Surfaces with $\boldsymbol{alpha}$-NeuS. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

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

Douglas E. Zongker, Dawn M. Werner, Brian Curless, and David H. Salesin. 1999. Environment matting and compositing. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques - SIGGRAPH '99*. doi:10.1145/311535.311558