Light Field-Based Underwater 3D Reconstruction via Angular Re-Sampling

Yuqi Ding, Member, IEEE, Zhang Chen, Yu Ji, Jingyi Yu, Fellow, IEEE, and Jinwei Ye, Senior Member, IEEE

Abstract—Recovering 3D geometry of underwater scenes is challenging because of the non-linear refraction of light at the water-air interface, which is caused by the camera housing. We present a light field-based approach that leverages properties of angular samples for high-quality underwater 3D reconstruction from a single viewpoint. Specifically, we re-sample the light field image to angular patches. As underwater scenes exhibit weak view-dependent specularity, an angular patch tends to have uniform intensity when sampled at the correct depth. We thus impose this angular uniformity as a constraint for depth estimation. For efficient angular re-sampling, we design a fast approximation algorithm based on multivariate polynomial regression to approximate nonlinear refraction paths. We further develop a light field calibration algorithm that estimates the water-air interface geometry along with the camera parameters. Comprehensive experiments on synthetic and real data show our method produces state-of-the-art reconstruction of static and dynamic underwater scenes.

Index Terms—Light field imaging, underwater 3D reconstruction, underwater camera calibration.

I. INTRODUCTION

3D RECONSTRUCTION of underwater scenes is of great interest and importance to many underwater exploration tasks, for example, robotic navigation [1], [2], seafloor mapping [3], [4], and archaeological site preservation [5], [6], [7]. Affected by water as a participating medium, underwater 3D reconstruction faces challenges posed by scattering, absorption, and refraction. Nevertheless, the underwater environment also results in unique surface reflectance property that is beneficial to 3D reconstruction.

In this article, we present a light field-based approach for high-quality underwater 3D reconstruction from a single viewpoint (see Fig. 1). We use a refractive ray model to tackle the challenge posed by refraction. We consider the imaging setting that the camera is completely submerged in water by placing inside a watertight housing. The refraction thus happens on water-air interface at the lens port. We model non-linear light paths with one-time refraction for accurate depth estimation. Although there exist many solutions that adapt multi-view stereo or structure-from-motion with refractive geometry [8], [9], [10], [11], our method differs in exploiting the property of angular ray samples. Besides, by using a light field camera, our method has the advantage of being a single-view solution for 3D reconstruction.

Light field camera captures both spatial and angular information of light rays in one shot. This capacity is achieved by inserting a microlens array in between the imaging sensor and the main lens [12].
that come from a scene point. Given scene point depths, we can re-sample light field images to angular patches, which are groups of rays that come from the same scene point. Note that the angular patches that we use here are not microlens images, which in contrast collect angular samples from the same microlens. The similar angular sampling scheme is first proposed by Yu and McMillan [13] as surface light field. Because of refraction, it is non-trivial to re-sample these angular patches for underwater scenes. Even with closed-form ray tracing solutions, the computational cost is still very high because non-linear rays need to be traced for all sensor pixels at every depth candidate. We therefore design a fast approximation algorithm for efficient angular re-sampling. Specifically, we use a multivariate polynomial function to approximate the closed-form solution, which eliminates the need for per-pixel non-linear ray tracing.

The angular patches provide useful constraint for depth estimation. The key observation is that most objects exhibit much weaker view-dependent specular reflection in water than in air (see Fig. 5). Therefore, the angular patches should have uniform intensity when sampled at the correct depth. We call this property angular uniformity. With this constraint, we can estimate depth by minimizing the variance of angular patches. Although this weak-specular phenomenon of underwater reflectance is also described in several other literatures [14], [15], [16], its physical cause has not been rigorously examined. Here we provide an explanation for this phenomenon from the perspective of medium’s refractive index, by applying Fresnel equations.

As we assume known camera parameters for angular patch re-sampling, we also develop a light field camera calibration algorithm that jointly estimates the camera intrinsics, extrinsics, and parameters relevant to the water-air interface (i.e., the interface geometry and the refractive index of water). For validation, we perform extensive synthetic and real experiments. We also compare our method against off-the-self 3D scanners as well as state-of-the-art light field-based depth estimation algorithms. Results on static and dynamic scenes demonstrate that our method is highly accurate and robust for a variety of scenes. Our major contributions are summarized as follows:

- We develop a light field camera calibration algorithm that jointly estimates the camera parameters and the interface parameters.
- We propose the angular uniformity constraint for light field depth estimation based on the “non-specular” property of underwater scenes.
- We develop a fast approximation algorithm for efficient angular patch retrieval.

II. RELATED WORK

A. Underwater 3D Reconstruction

The problem of underwater 3D reconstruction has attracted much attention in the past decades. Many classical 3D reconstruction algorithms, including multi-view stereo [8], [10], [17], [18], structure-from-motion (SfM) [9], [11], [19], [20], structured light scanning [21], [22], [23], and photometric stereo [24], [25], [26] are adapted to the underwater setting either by considering the refractive geometry, or by compensating the absorption and scattering effect in the water. Notably, Chari and Sturm [18] derive the multi-view geometry in refractive medium based on curved epipolar lines. Chadbeq et al. [9], [10] use multiple refractive planes to model relative camera motions, and solve SfM with refraction. Narasimhan et al. [22] take the scattering effect into account, and refine the structured light and photometric stereo methods for underwater scenes. Tsionsios et al. [24] compensate the backscattering of point light sources in order to perform photometric stereo in murky water. Asano et al. [27], [28] leverage different absorption rate of near-infrared light for underwater depth estimation. The method is further extended as a one-shot solution for recovering dynamic scenes [29] and non-rigid scenes [30]. Several methods jointly recover the 3D of the water surface and the underwater scene with defocus cue [31], multi-view cue [32], and differentiable re-sampling [33].

As camera calibration is essential to 3D reconstruction, we also briefly review calibration methods developed for underwater settings. Lavest et al. [34] derive compensated lens models when using a camera underwater, and find that the magnifying effect of water refraction is equivalent to scaling the in-air focal length with the fluid’s refractive index. Treibitz et al. [35], [36] solve a simplified underwater calibration problem with a frontal-parallel refraction plane and known refractive index. Agrawal et al. [37] consider the problem of multiple refraction planes, and model underwater cameras with axial camera models. Haner et al. [38] study the extrinsic calibration in the presence of a single refraction. Chen and Yang [39] calibrate a stereo system in the presence of a thick flat refraction plane.

B. Light Field-Based Methods

Light field images record 4D spatial and angular samples of a scene. To capture a light field, one can either use a camera array [40], [41], or a compact light field camera with microlenses array [12], [42]. Due to the multi-view nature, light field images are widely used for image synthesis and 3D reconstruction applications, including post-defocusing [43], [44], [45], novel view synthesis [46], [47], [48], and depth reconstruction [49], [50], [51]. A comprehensive evaluation of light field-based depth estimation methods can be found in [52]. Many algorithms are developed for light field camera calibration by using different types of image features [53], [54], [55], [56].

Light field-based methods have been developed for underwater applications. Some tackle the problems of low visibility and color distortions caused by scattering and absorption via jointly recovering clear underwater image and scene depth [57], [58], [59]. Skinner and Johnson-Roberson [60] develop a fast underwater 3D reconstruction solution using a light field camera. However, this method only accounts for the absorption of light in water without compensating for refraction. Ichimaru and Kawasaki [61] synthesize refraction-free images from light field and use them for stereo-based 3D reconstruction.

Our method uses a compact light field camera as imaging device and adopts refractive ray model for underwater 3D reconstruction. Different from existing methods, we exploit the angular uniformity of underwater scenes for depth estimation.
III. PROPOSED METHOD

Here we present our light field-based method for underwater 3D reconstruction. We consider the one-time refraction on the water-air interface. We first introduce our calibration method under this setting (Section III-A). We then describe our re-sampling scheme for forming the angular patches (Section III-B). Lastly, we present our depth estimation algorithm (Section III-C).

A. Light Field Camera Calibration

Our calibration algorithm aims at jointly estimating the interface parameters, which include the interface geometry and the refractive index of water, along with the camera parameters (including both intrinsics and extrinsics).

We consider the setting as shown in Fig. 2. The intrinsic parameters are used to map a pixel on the image plane to a ray that exits the light field camera. We represent a ray with a point on the ray and its direction. Assuming the main lens plane as \( z = 0 \), a ray exiting the camera can be represented by its main lens intersection point \( P_a = (u, v, 0)^\top \), and direction \( d_a = (\sigma, \tau, 1)^\top \). A sub-aperture image pixel can be represented as a 4D coordinate \((s, t, i, j)^\top\), where \((s, t)\) is the sub-aperture image index, and \((i, j)\) is the pixel index within a sub-aperture image. Dansereau et al. [53] derive that the mapping between an image pixel and its exiting ray under the homogeneous coordinate can be expressed as a \(5 \times 5\) matrix \( H\) with 12 non-zero elements:

\[
\begin{bmatrix}
u \\ s \\ t \\ i \\ j \\ w \end{bmatrix} = 
\begin{bmatrix}
H_1 & 0 & H_2 & 0 & H_3 \\
0 & H_4 & 0 & H_5 & H_6 \\
H_7 & 0 & H_8 & 0 & H_9 \\
0 & H_{10} & 0 & H_{11} & H_{12} \\
1 & 0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
u \\ s \\ t \\ i \\ j \end{bmatrix}.
\]

We further use the sixth-order polynomial distortion model [62] to compensate for lens distortions. We refer the non-zero elements in \(H\): \(\{H_1, \ldots, H_{12}\}\) and the distortion parameters \(\{k_1, k_2, k_3, b_1, b_2\}\) as the intrinsic parameters of a light field camera.

The in-air camera ray \(r_a : P_a + \lambda d_a\) (where \(\lambda\) is the ray propagation factor) is refracted at the water-air interface. We then derive the origin and direction of the refracted ray. Assuming planar water-air interface, we have: \(Ax + By + Cz + D = 0\), where \(A, B, C, D\) are coefficients of the plane equation. The point \(P_w\) where the camera ray enters the water is determined by \(\lambda_w\):

\[
\lambda_w = -\frac{Au + Bv + D}{As + Br + C}.
\]

We use \(P_w = P_a + \lambda_w d_a\) as the origin of the refracted ray. By applying the Snell’s law, the direction of the refracted ray can be calculated as:

\[
d_w = \frac{1}{n}(d_a - (\alpha + \sqrt{n^2 - (1 - \alpha^2)\alpha})n),
\]

where \(d_a\) is the camera ray’s direction; \(n = (A, B, C)^\top/\sqrt{A^2 + B^2 + C^2}\) is the normal of water-air interface; \(\alpha = n \cdot d_a\); and \(n\) is the refractive index of water (we assume the refractive index of air is 1). The refracted ray (or underwater ray) is then written as \(r_w = P_w + \lambda d_w\).

We define the plane parameters \(\{A, B, C, D\}\) and the refractive index of water \(n\) as the interface parameters as they determine the refractive medium.

We perform calibration with in-water checkerboard patterns. We use the extrinsic parameters: \(\{R | T\}\) (where \(R \in \mathbb{R}^{3 \times 3}\) is rotation matrix and \(T \in \mathbb{R}^{3 \times 1}\) is translation vector) to transform a point \(P_m\) on the checkerboard to the camera coordinate: \(P = RP_m + T\). As \(P\) should lie on the underwater ray \(r_w\), we minimize the point-to-ray distance between \(P\) and \(r_w\) to solve the calibration parameters altogether (i.e., intrinsics, interface, and extrinsics).

We consider the point-to-ray distance in two cases that are shown in Fig. 3. We first determine \(P\) at which side of the interface plane. We compare the vector \(l = P - P_w\) with the plane normal \(n\). If \(l\) and \(n\) have opposite directions, \(P\) is on the water side; if they have the same direction, \(P\) is on the air side. Then, if \(P\) is on the water side (as it should be), we directly use its perpendicular distance from \(P\) to \(r_w\). If \(P\) is on the air side (which must be caused by incorrect parameter estimation), we use the length of \(l\) as its distance to \(r_w\). In this way, we penalize the estimations that cause \(P\) to be on the wrong side of the interface plane. The point-to-ray distance \(\varepsilon(P, r_w)\) is therefore...
computed as:

$$
\varepsilon(P, r_w) = \begin{cases} 
|1 - (1 \cdot d_w) d_w|, & 1 \cdot n \leq 0, \\
|1|, & 1 \cdot n > 0.
\end{cases}
$$

(4)

Suppose there are $M$ feature points on the checkerboard, and we use $N$ different pattern poses. We minimize the following objective function to solve for the calibration parameters:

$$
\arg\min_{I, F, \mathcal{E}} \sum_{i=1}^{N} \sum_{j=1}^{M} \varepsilon(RP_m + T, r_w),
$$

(5)

where $I = \{H_{1,\ldots,12}|k_{1,2,3}, b_{1,2}\}$ are intrinsic parameters; $F = \{A, B, C, D|a\}$ are interface parameters; and $E = \{R_{1,\ldots,N}|T_{1,\ldots,N}\}$ are extrinsic parameters. We use Levenberg-Marquardt algorithm to solve this optimization.

B. Angular Patch Re-Sampling

With the calibration parameters, we can re-sample the sub-aperture images to a bunch of angular patches, where each patch is a collection of rays that come from the same scene point. Fig. 4(a) shows an example of angular patch. We formulate angular patches with respect to the center view image. The patch is formed by tracing rays back to each sub-aperture image given a scene depth $d$. Let $\mathcal{A}(d)$ be an angular patch re-sampled at center view pixel $p$ from a scene point $P$ at depth $d$, and $\mathcal{H}^p: (s, t, d) \mapsto (i, j)$ be the mapping from a scene point at depth $d$ to its projected pixel location $(i, j)$ in sub-aperture image $(s, t)$. We can use $\mathcal{H}^p$ to locate pixels in sub-aperture images that form $\mathcal{A}(d)$.

Due to refraction, the calculation of $\mathcal{H}$ is non-trivial. We first show that a closed-form solution can be derived by applying Snell’s law. However, the closed-form solution is a complex equation that needs to be solved at all pixels for all depth labels, which is computationally expensive. We then design an approximation method based on multivariate polynomial regression.

We first derive the closed-form solution to this ray tracing problem. As shown in Fig. 4(b), given a pixel $p$ in the center view, we can trace out an underwater ray from its center of projection (CoP) $P_c$. Given depth $d$, we can sample a scene point $P$ along this ray. Let $P_a$ be the CoP of another sub-aperture image. $P_r$ and $P_a$ are the projected points of $P$ on the water-air interface. $r_a$ and $r_w$ are the refracted ray segments that project $P$ to the sub-aperture image at $P_a$. $P_a$ is the intersection point where refraction happens. To obtain the projected pixel $(i, j)$ of $P$, we need to solve for $P_a$.

To simplify notations, we denote $\gamma = ||P_a - P_w||$ and $\kappa = ||P' - P'_a||$. By applying the Snell’s law, we have

$$
\gamma^2((\kappa - \gamma)^2 + ||P' - P||^2) - n^2(\kappa - \gamma)^2(\gamma^2 + ||P'_a - P_a||^2) = 0.
$$

(6)

Given $P$ and $P_a$, (6) is a univariate fourth-order equation of $\gamma$ and has up to four real solutions. The solutions can be computed via closed-form formulas [63]. As shown in [64], only one solution is valid. With $\gamma$, we have $P_w = P'_a + \gamma(P' - P'_a)$. With $r_a = P_w - P_a$, we can locate the projected pixel $(i, j)$ with intrinsic parameters.

1) Complexity Analysis: In our approximation method, we establish a mapping between $(s, t, d)$ and $\gamma$ without solving (6) for all sub-aperture views and depth candidates. Given a pixel $p$ in the center view, we denote such mapping as $\mathcal{G}^p: (s, t, d) \mapsto \gamma$. As a microrels-based light field camera has a small baseline, $\mathcal{G}^p$ is smooth at most combinations of $(s, t, d)$ under our setup and hence we use a multivariate cubic polynomial function to approximate $\mathcal{G}^p$.

At each center view pixel $p$, we compute 64 uniform samples of $\gamma$ with respect to $(s, t, d)$ in a $4 \times 4 \times 4$ grid. With these pairs of $(s, t, d)$ and $a$, we solve the polynomial coefficients of $\mathcal{G}^p$ via least squares regression. The resulting $\mathcal{G}^p$ is then used to calculate $\gamma$ at each $(s, t)$ and depth candidate $d$. Compared with the closed-form solution, our approximation method is able to achieve $10 \times$ speedup.

We test on a light field with spatial resolution $625 \times 434$ and angular resolution $11 \times 11$, and 100 depth candidates per pixel. When implemented with Matlab without GPU acceleration, the closed-form solution takes over 20mins to compute all angular patches, as fourth-order equation needs to be solved for $625 \times 434 \times 11 \times 11 \times 100$ times. In contrast, our approximation method only takes 2mins to re-sample all the patches. The average error of $\gamma$ computed with our approximation method is 0.01 mm.

2) Uniformity of Angular Patch: We find that the angular patches tend to have uniform intensity for underwater scenes. This is because underwater surface reflectance has weak view-dependent specular reflection. We show a comparison between the underwater and in-air reflectance of the same object in Fig. 5. By applying the Fresnel equations, we show that this phenomenon is due to the low gradient of refractive index between water and common materials in real-world objects (e.g., plastic, ceramic, etc.).

The Fresnel equations give us the reflection rate of light when incident on the interface between two media:

$$
\eta = \frac{1}{2} \left( \frac{\left| n_1 k_1 - n_2 k_2 \right|^2}{n_1 k_1 + n_2 k_2} + \frac{\left| n_1 k_2 - n_2 k_1 \right|^2}{n_1 k_2 + n_2 k_1} \right),
$$

(7)

where $n_1$ and $n_2$ are the media’s refractive indices (light travels from $n_1$ to $n_2$); $\theta_i$ is the incident angle; $k_1 = \cos \theta_i$; and $k_2 = \cos \sqrt{1 - (n_2 - n_2 \sin \theta_i)/n_2}$. 

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
This equation shows that when the two media have similar refractive indices (i.e., \( n_1 \approx n_2 \)), the \( \eta \) value is likely to be small, which indicates that light will be mostly transmitted (which causes diffuse reflection), instead of being directly reflected (which causes specular reflection). As water is a dense medium, its refractive index (i.e., 1.33) is closer to common materials (e.g., plastics \( 1.3 \sim 1.7 \), porcelain \( \sim 1.5 \)) than air. These objects thus exhibit much weaker specular reflection in water than in air. This phenomenon has exception for metal surfaces, whose refractive indices are complex numbers. Direct reflection therefore still largely exists on metal surfaces.

This weak specular phenomenon results in the uniformity of angular patches that is useful for depth estimation.

C. Depth Estimation

The scene depth is estimated by solving an optimization with two constraints: 1) uniformity of angular patches, and 2) color and gradient consistency along curved epipolar lines. Note that the two constraints appear to be the same when the true depths are known, but the costs are computed in different ways. Given an angular patch, color consistency computes the summed absolute difference between the sub-aperture view pixels and the center view pixel, whereas angular uniformity computes the variance within the angular patch. The two constraints are thus complementary.

1) Angular Uniformity Constraint: In Section III-B, we show that the angular patches have uniform intensity when sampled at the correct depth because of the weak specular phenomenon. Angular ray samples traced from the wrong depth might have different colors as they in fact come from different scene points. Therefore, we use the color variance of angular patches, which is computed by summing up all pixel’s intensity difference from the mean, as a constraint for depth estimation. We plot the variances of an angular patch with respect to the depth in Fig. 6. We can see that the variance is lowest when sampled at the true depth. Wrong depth values result in the non-uniform angular patch and thus large variance. We also compare the curve computed with our approximation algorithm against the closed-form solution. We can see the two curves are close, which indicates that our approximation algorithm is valid. The cost function of the angular uniformity constraint is written as:

\[
E_{\text{angular}}(p, d) = \frac{1}{ST - 1} \sum_{s=1}^{T} \sum_{t=1}^{S} |\mu_{s,t} - \mu|^2,
\]  

where \( \mu(d) \) is the angular patch for a pixel \( p \) in the center view, sampled at depth \( d \); \( s, t \) is sub-aperture image index; \( (S, T) \) is the angular resolution; and \( \mu \) is the mean of \( \mu(d) \).

2) Color and Gradient Consistency: We also adopt the color and gradient consistency along epipolar lines as constraints for depth estimation. These two constraints measure the photo-consistency of a scene point among sub-aperture views—the same scene point should have a similar color in all sub-aperture views. They are commonly used in multi-view depth estimation. For underwater scenes, the epipolar lines are curved due to refraction [18]. The epipolar curves in sub-aperture images can be formed with the angular rays. We compare the color and color gradient of pixels along the epipolar curves with the corresponding center view pixel. Let \( f_{s,t}^p \) be the epipolar curve for center view pixel \( p \) in sub-aperture image \( (s,t) \). Our cost functions are written as:

\[
E_{\text{color}}(p, d) = \sum_{t=1}^{T} \sum_{s=1}^{S} |I(p) - I_{s,t}(f_{s,t}^p(d))|,
\]

\[
E_{\text{grad}}(p, d) = \sum_{t=1}^{T} \sum_{s=1}^{S} |G(p) - G_{s,t}(f_{s,t}^p(d))|,
\]

where \( I \) is intensity image (\( I_x \) refers to the center view); \( G \) is the combined gradient of intensity image \( I \) (i.e., \( G = \partial I/\partial x + \partial I/\partial y \)).

3) Optimization: Our final objective function combines (8), (9), and (10) and an additional smoothness term:

\[
\arg\min_d \sum_p \left( \beta_1 E_{\text{color}} + \beta_2 E_{\text{grad}} + \beta_3 E_{\text{angular}} + \beta_4 E_{\text{smooth}} \right)
\]

where \( \beta_1, \ldots, \beta_4 \) are weights for balancing the four terms (values of these parameters are given in the experiment section); \( E_{\text{smooth}} = \sum_{p_c} ||d(p) - d(p_c)|| \) is the smoothness term that enforces depth smoothness, where \( p_c \) is the set of neighboring pixels around \( p \). The optimization is solved with multi-label
We place an $\beta \times \{0, \ldots, 312.5\}$ checkerboard (in-
verse camera, we set the intrinsic parameters as follows: focal length $f = 719.92$ mm and principal point $c = 312.5$ pixel. We set the center of projection of the center view camera.

During calibration, we initialize the intrinsics and extrinsics using the air-based calibration method [53] and initialize the interface parameters as $F = \{0, 0, -1, 0.1|1.6\}$. For depth estimation, we set the hyperparameters as $\beta_1 = 0.3$, $\beta_2 = 0.3$, $\beta_3 = 0.4$ and $\beta_4 = 0.001$ in our experiments. The neighborhood size of the smoothness term is set as 3.

### A. Synthetic Experiments

We simulate light field images of underwater scenes through ray tracing. Our light field images have $11 \times 11$ sub-aperture images with resolution $625 \times 434$. The equivalent baseline is $b = 0.4$ mm. We set the center of projection of the center view image as the original of the camera coordinate. For each perspective camera, we set the intrinsic parameters as follows: focal length $f = 550$ pixel and principal point $c_x = 312.5$ pixel, $c_y = 217$ pixel. The original point $O$ is the CoP(Center of Projection) of the center view camera.

We set the refractive index of air to 1. We use three sets of interface parameters with different plane orientations and refractive indices of water: $F_1 = \{0, 0, -1, 0.1|1.33\}$, $F_2 = \{0, 0, -1, 0.1|1.45\}$, and $F_3 = \{0.1476, 0.0984, 1.33\}$. $F_1$ and $F_2$ use frontal-parallel interface planes, but different refractive indices of water. $F_3$ uses an oblique interface plane.

1) Underwater Calibration: We place a $6 \times 7$ checkerboard at the waterside. We randomly pick rotation and translation matrices to generate different pattern poses. The running time of our calibration algorithm is around 15 mins with 20 pattern poses.

We compare our intrinsic calibration with three state-of-the-art methods: 1) SV [66]: a classical calibration method for conventional camera at a single viewpoint; 2) TB [53]: the calibration algorithm from the Light Field Toolbox for Matlab; and 3) MPC [56]: a recent light field camera calibration method that uses a multi-projection-center model to reduce the parameter space. All these methods are designed for in-air calibration without considering the refraction of light. We use the center view image as input to the single view method (SV).

Table I shows the intrinsic calibration results in comparison with the ground truth under the interface settings. We can see that the estimated baselines are less affected by the refraction. The estimated focal lengths from all air-based methods are closer to the ground truth values multiplied by the refractive index (1.333). In contrast, our method separates the refraction and is able to estimate more accurate intrinsic parameters.

As our method also calibrates the interface parameters $F$ (including the interface plane parameters and the refractive index of water), we show our calibration results under the three interface settings in comparison with the ground truths in Table II. We can see that our estimations are accurate. Our method is robust under various interface orientations and refractive indices.

Finally, we evaluate our extrinsic parameter calibration. Here we also compare with two other methods that assume known intrinsics: 1) SVA: the classical calibration method [66] that uses the ground truth intrinsic parameters in air; and 2) MR [37]: the multi-refractive layer underwater calibration method that uses the ground truth intrinsic and interface parameters. For each method, we evaluate the estimated rotation and translation with respect to different numbers of pattern poses. The extrinsic parameter estimation errors under the interface settings are in Fig. 7.

In addition, we show visualization of our extrinsic calibration results in comparison the state-of-the-arts (SV [66], TB [53],...
Fig. 7. Errors of estimated extrinsic parameters under three interface settings. Each column shows results under one interface.

Fig. 8. Visualization of extrinsic parameters under different interface parameters.

Fig. 9. Analysis on the influence of interface thickness on refractive light paths. (a) An illustration of the light path when interface thickness is considered. (b) Euclidean reprojection error for not taking interface thickness into account under various incident angles, object distance and interface thickness.

MPC [56], SVA [66], and MR [37]), and the ground truth under all three interface settings ($F_1$, $F_2$, and $F_3$). Specifically, we plot all checkerboard poses (the green ones are the ground truth poses, and the red ones are the poses estimated by each method), and the water-air interface (shown in blue). Results for $F_1$, $F_2$, and $F_3$ are shown in Fig. 8. We can see that our estimated poses are highly close to ground truth ones.

2) Interface Thickness: We analyze the influence of interface thickness on refractive light paths. Specifically, we evaluate the Euclidean reprojection difference of 3D points between considering interface thickness vs. without considering. Fig. 9(a) illustrates the case when interface thickness is considered, in which there are two refraction along the light path. Fig. 9(b) shows the reprojection error for not taking interface thickness into account.
under various incident angles, object distance and interface thickness. Here we use $f_3 = \{0.1476, 0.0984, -0.9841, 0.1\}$ as interface parameters and set the refractive indices of air, water and interface as 1.0, 1.333 and 1.52 (a typical value for glass) respectively. For camera intrinsics, we use $f_x = f_y = 550$ pixel, $c_x = 312.5$ pixel and $c_y = 217$ pixel. It can be seen that the reprojection error is higher for larger incident angle/interface thickness and lower distance. However, for 2 mm and 5 mm thickness, the reprojection error is consistently smaller than 1 pixel, and the average value is even smaller (0.127 and 0.318 pixel respectively). For 10 mm thickness, although the maximum reprojection error exceeds 1 pixel, the average value is only 0.636 pixel, revealing that the reprojection error is still generally small. As a reference of interface thickness, it is usually $2 \sim 8$ mm for underwater camera housing. Based on the above analysis, interface thickness has a very small influence on calibration and reconstruction, and is therefore not explicitly modeled in our approach.

3) Depth Estimation: Given the calibration parameters, we evaluate our depth estimation algorithm. In this experiment, we use the interface parameter $F_1 = \{0, 0, -1, 0.1\}$, and test on a planar target. We compare with a state-of-the-art light field-based depth estimation method [51]. For fair comparison, we first use their algorithm to estimate a disparity map, and then use the ground truth intrinsics and interface parameters to convert disparities into depth values. Fig. 10 compares the reconstruction results. Our result is very close to the ground truth plane. In contrast, the result of [51] appears to be curved because it uses straight epipolar lines that don’t account for refraction.

B. Real Experiments

We perform real experiments to validate our algorithms. We use the Lytro Illum to capture underwater scenes submerged in a water tank of size 60cm × 30cm × 36 cm. Our experimental setup is shown in Fig. 1. The captured light field image has a resolution of 7728 × 5368. We decode it to $15 \times 15$ sub-aperture images, each with resolution 625 × 434.

1) Calibration Results: Here we show the camera calibration results of our real experiments. We use two settings in the experiments: one-camera setting and two-camera setting. We show our calibrated camera intrinsics and water-air interface parameters under the two settings in Tables III and IV, respectively. For the two-camera setting, we also show the transformation $(R_s, T_s)$ from camera 2 to camera 1. The rotation $R_s$ is represented in the form of axis-angle.

2) Ablation Study: As shown in Fig. 11, we quantitatively evaluate our method using underwater planar boards at three different depths (9 cm, 15 cm, 20 cm). The ground truth depths are measured with a ruler. Table V shows the depth estimation errors (in mm) for the three planar boards, as well as the

Fig. 10. 3D reconstruction of an underwater plane.

Fig. 11. Quantitative evaluation on three depth layers. From left to right: scene setting, center view of the captured light field, and our estimated depth map.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
ablation study on the terms in our objective function (11). We can see that our proposed angular uniformity constraint \( E_{\text{angular}} \) effectively improves the reconstruction accuracy.

3) Static Scene Reconstruction: We first use one light field camera to capture the various static underwater scenes. We compare our method with structured light (SL), time-of-flight (ToF), and a light field-based depth estimation method from [51]. Note that SL and ToF are active 3D reconstruction methods. For SL, we build a projector-camera system for 3D scanning. In the system, we use a 1280 × 720 laser projector and a 2448 × 2048 monochrome camera. Both projector and camera are calibrated with underwater patterns [67]. We use the Gray code patterns, and recover the point cloud by ray triangulation without considering refraction. For ToF, we use the Azure Kinect to scan the underwater scenes and recover its point cloud using the default SDK. We also implement a ToF variant that compensates the reconstruction results with the refractive index of water, which is denoted as ToF*. For [51], we combine it with the calibration result from our method so that the conversion from disparity to depth correctly accounts for refraction.

To obtain ground truth, we use the structured light method to scan the objects in air. The setup is the same as SL. Table VI compares the Root Mean Squared Error (RMSE) of reconstructed point clouds using different methods for five scenes. The qualitative results for different scenes are shown in Fig. 12. For underwater SL, the recovered point clouds are noisy and
distorted because the rays cannot be properly triangulated when refraction is not considered. ToF results suffer from large errors because the velocity of light is smaller in water than in air. Although ToF* greatly improves over ToF, its accuracy is still lower than our method. [51] assumes linear epipolar lines. Thus its disparity computation does not take refraction into account. As a result, its results are less accurate than our method.

We also perform experiments with two light field cameras. The two cameras are placed side by side in order to have a large overlap in the field of view. We perform 3D reconstruction on each light field and show that the point clouds reconstructed by our method can be naturally aligned without using any fitting algorithms (e.g., ICP [68]). For comparison, we combine two light field camera calibration algorithms (TB [53] and MPC [56]) with the light field depth estimation method [51]. We perform TB with in-air calibration targets, while MPC with underwater

![Fig. 12. 3D reconstruction results with one light field camera in comparison with state-of-the-arts.](image-url)
Fig. 13. Point cloud fusion results with two light field cameras. We directly fuse the point clouds without applying fitting algorithms. The second row shows zoom-in views that display the two point clouds in different colors.
ones. We also combine [51] with our calibration results. Note that TB and MPC cannot calibrate interface parameters, so TB + [51] and MPC + [51] do not take refraction into account. The point cloud fusion results are shown in Fig. 13. We can see that all the comparison methods exhibit clear misalignments in their fusion results. In contrast, our point clouds are well aligned without explicit fitting.

4) Dynamic Scene Reconstruction: Finally, we perform experiments to recover a dynamic underwater scene. Our dynamic scene consists of a goldfish swimming in front of an aquarium castle (see Fig. 14). Here we choose to use a small aquarium of size $30 \text{cm} \times 30 \text{cm} \times 30 \text{cm}$, because the field of view and focus range of the camera are small. We use one Lytro camera to capture a video at 10 frames per second. As Lytro doesn’t have the video mode, we implement a continuous triggering function with an external control board. We first calibrate the camera, and then perform 3D reconstruction on each frame of the recorded video. We show five consecutive frames of the reconstruction results in Fig. 14. Our method can well recover the fine structures of the goldfish, such as the fin and the tail. But notice that the colors of the fin and tail are dark in some frames, as they are semi-transparent and blend in the black background color. Although our frame rate is low (10 fps), the reconstruction results still illustrate the fish’s motion trajectory well. This shows that our method can be used for 3D video acquisition.

V. CONCLUSION & DISCUSSIONS

In this article, we present a light field-based method for underwater 3D reconstruction from a single viewpoint. We demonstrated using the angular uniformity for depth estimation. We designed an approximation method to address the efficiency of angular patch calculation. We also developed a calibration algorithm that jointly estimates the camera parameters with the interface attributes. Our method is validated with extensive experiments.

One limitation of our method is that the scattering and attenuation effects are not being considered. To alleviate the problem, we can preprocess underwater images with visibility enhancement algorithms (e.g., [69]) for better quality. Due to the small baseline of light field camera, our method works well for close-range, small-scale scenes. The camera array needs to be used for large-scale scenes. In the future, we plan to build more versatile imaging systems to overcome the challenges of underwater imaging.

REFERENCES

[1] M. Johnson-Roberson, O. Pizarro, S. B. Williams, and I. Mahon, “Generation and visualization of large-scale three-dimensional reconstructions from underwater robotic surveys,” *J. Field Robot.*, vol. 27, no. 1, pp. 21–51, 2010.

[2] L. Paull, S. Saeedi, M. Seto, and H. Li, “AUV navigation and localization: A review,” *IEEE J. Ocean. Eng.*, vol. 39, no. 1, pp. 131–149, Jan. 2014.

[3] S. Negahdariour and H. Madjidi, “Stereovision imaging on submersible platforms for 3-D mapping of benthic habitats and sea-floor structures,” *IEEE J. Ocean. Eng.*, vol. 28, no. 4, pp. 625–650, Oct. 2003.

[4] O. Pizarro, R. M. Eastice, and H. Singh, “Large area 3-D reconstructions from underwater optical surveys,” *IEEE J. Ocean. Eng.*, vol. 34, no. 2, pp. 150–169, Apr. 2009.

[5] J. Henderson, O. Pizarro, M. Johnson-Roberson, and I. Mahon, “Mapping submerged archaeological sites using stereo-vision photogrammetry,” *Int. J. Nautical Archaeol.*, vol. 42, no. 2, pp. 243–256, 2013.

[6] J. McCarthy, “Multi-image photogrammetry as a practical tool for cultural heritage survey and community engagement,” *J. Archaeological Sci.*, vol. 43, pp. 175–185, 2014.

[7] K. Yamaufune, R. Torres, and F. Castro, “Multi-image photogrammetry to record and reconstruct underwater shipwreck sites,” *J. Archaeological Method Theory*, vol. 24, no. 3, pp. 703–725, 2017.

[8] J. Gedge, M. Gong, and Y.-H. Yang, “Refractive epipolar geometry for underwater stereo matching,” in *Proc. IEEE Can. Conf. Comput. Robot Vis.*, 2011, pp. 146–152.

[9] F. Chadebecq et al., “Refractive structure-from-motion through a flat refractive interface,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5325–5333.

[10] F. Chadebecq et al., “Refractive two-view reconstruction for underwater 3D vision,” *Int. J. Comput. Vis.*, vol. 128, pp. 1101–1117, 2019.

[11] K. Ichimaru, Y. Taguchi, and H. Kawasaki, “Unified underwater structure-from-motion,” in *Proc. IEEE Int. Conf. 3D Vis.*, 2019, pp. 524–532.

[12] R. Ng et al., “Light field photography with a hand-held plenoptic camera,” Stanford Tech Report CTSR 2005-02, 2005.

[13] J. Yu, L. McMillan, and S. Gortler, “Surface camera (SCAM) light field rendering,” *Int. J. Image Graph.*, vol. 4, no. 04, pp. 605–625, 2004.

[14] Y. Y. Schechner and N. Karpen, “Recovery of underwater visibility and structure by polarization analysis,” *IEEE J. Ocean. Eng.*, vol. 30, no. 3, pp. 570–587, Jul. 2005.

[15] T. W. Cronin, N. Shashar, R. L. Caldwell, J. Marshall, A. G. Cheroske, and T.-H. Chiou, “Polarization vision and its role in biological signaling,” *Integrative Comp. Biol.*, vol. 43, no. 4, pp. 549–558, 2003.

[16] Y. Swirski, Y. Y. Schechner, B. Herzberg, and S. Negahdariour, “CAsse-teto: Range from light in nature,” *Appl. Opt.*, vol. 50, no. 28, pp. 89–101, Oct. 2011.
Yuqi Ding (Member, IEEE) received the B.S. and M.S. degrees in geographical information system from Wuhan University, Wuhan, China, in 2013 and 2009, respectively. He is currently working toward the Ph.D. degree with the Division of Computer Science and Engineering, Louisiana State University, Baton Rouge, LA, USA. His research interests include computational photography, computer vision, and computer graphics.

Zhang Chen received the bachelor degree in electrical and computer engineering from Shanghai Jiao Tong University, Shanghai, China, in 2014 and the Ph.D. degree in computer science from ShanghaiTech University, Shanghai, China, in 2022. His research interests include computer vision, computer graphics, and computational photography.

Yu Ji received the bachelor degree in electrical engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2009, the M.Sc. degree in digital media from Nanyang Technological University (NTU), Singapore, in 2010, and the Ph.D. degree in computer and information science from the University of Delaware (UDel), Newark, DE, USA, in 2015. He is currently a Researcher with Tencent Pixel Lab, Minneapolis, MN, USA. His research interests include computational photography, computer vision, and computer graphics.

Jingyi Yu (Fellow, IEEE) received the B.S. degree from Caltech, Pasadena, CA, USA, in 2000 and the Ph.D. degree from MIT, Cambridge, MA, USA, in 2005. He is currently a Professor and Executive Dean of the School of Information Science and Technology, ShanghaiTech University, Shanghai, China. He has authored or coauthored more than 120 papers at highly refereed conferences and journals, and holds more than 10 international patents on computational imaging. His research interests include the range of topics in computer vision and computer graphics, especially on computational photography and non-conventional optics and camera designs. He was the recipient of the NSF CAREER Award and the AFOSR YIP Award. He was an Area Chair of many international conferences including CVPR, ICCV, ECCV, ICCP and NIPS. He is a Fellow of Optica.

Jinwei Ye (Senior Member, IEEE) received the B.Eng. degree in electrical engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and the Ph.D. degree from the University of Delaware, Newark, DE, USA, in 2014. She is currently an Assistant Professor with the Department of Computer Science, George Mason University, Fairfax, VA, USA. From 2017 to 2021, she was with Louisiana State University, Baton Rouge, LA, USA. Before joining academia, she was a Researcher with the US Army Research Laboratory and Canon USA, Inc. Her research interests include computational photography, computer vision, and computer graphics. He was the recipient of NSF CAREER Award and CRII Award.