In-the-Wild Single Camera 3D Reconstruction Through Moving Water Surfaces Supplementary Material

Jinhui Xiong Wolfgang Heidrich
KAUST

1. Relationship to Structure-from-Motion

Inferring depth from distorted views as proposed in our work shares some similarities to structure-from-motion (SfM) [4]. SfM utilizes a series of images taken from different viewpoints to reconstruct the 3D structure of the scene. The images are usually taken with a moving camera. The reconstruction is realized with bundle adjustment, which jointly estimates camera parameters and scene geometry by solving a non-linear least square problem. By comparison, in our problem multiple viewpoints are introduced by the non-stationary water surface fluctuations, and we propose a novel differentiable framework to simultaneously estimate the structure of non-stationary water surfaces and underwater scene geometry.

The problem we studied differs from SfM mainly in the following three aspects:

- Camera parameters in SfM can be represented by a 4×4 matrix, a low-parameter model with respect to scene geometry. Since camera parameters are low-dimensional, they can be estimated by matching feature points, e.g. SIFT features [2]. On the other hand, to fully characterize a water surface, the degrees of freedom can approach the same size as the underwater scene for choppy water. To estimate water surfaces, a dense correspondence match is required, which is prone to error especially when the surface distortion is strong. Therefore, the estimation of camera parameters is more robust and less ill-conditioned.

- In SfM, the projection from 3D coordinates to the image plane is linear, and the problem could be solved via a relatively simple minimization formula. This projection becomes non-linear as the lights pass through a refractive interface, which makes the problem cannot be tackled with a scheme similar to bundle adjustment.

- In general, the baseline from moving a camera can be much larger than that caused by water surface fluctuation. Knowing that small baseline will reduce the depth estimation accuracy, which is approximately inversely proportional [1], employing water surface fluctuations for multi-view triangulation is more noise sensitive and theoretically produces scene geometry in lower depth accuracy.

Nonetheless, our proposed framework, which integrates ray casting, Snell's law and multi-time triangulation, along with the regularization terms for time-varying water surfaces and underwater scene geometry, tackles the difficulties arising from a well-studied solution to a SfM problem. We demonstrated that this novel framework is capable of recovering fully characterized time-varying water surfaces and 3D background scenes using a single camera.

2. Implementation Details

The detailed implementation of our proposed differentiable framework is illustrated in Fig. 1. Given the framework with underwater point clouds and time-varying water surfaces, the loss of the entire model is computed through forward propagation following the designed pipeline. Afterwards, the variables are simultaneously optimized from the back-propagated gradients from the model loss.

Fig. 2 shows an example of the progression of the loss function over time, as well as the corresponding geometries at the beginning, in the middle, and at the end of the optimization process.

Because of the single-view depth-normal ambiguity on the recovered water surface, the proposed global optimization problem is non-convex. Different initializations will drive the framework to different local minimums. In most initial points, the framework finds a reasonable representation of the scenes. However, we did find degenerated cases with some initializations. As discussed in the manuscript, the initialization of the underwater scene geometry is a planar surface with different axial depths. In Fig. 3, we show the reconstructions with different initial values which yields similar adequate representations, and also a failure case with improperly selected initial value. The initial value should also vary with different reconstruction data.

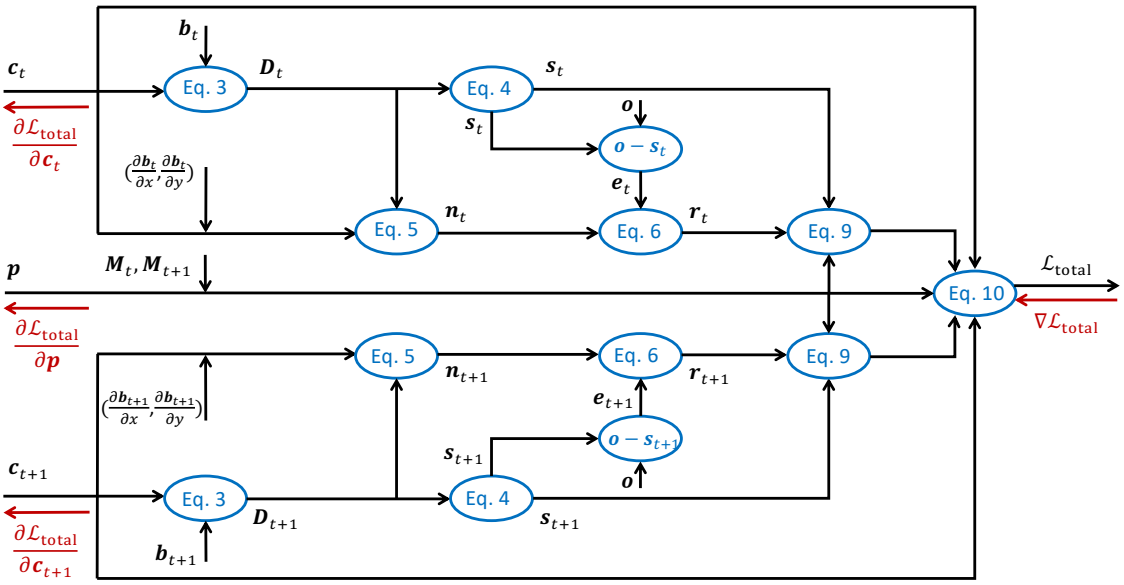


Figure 1: The flowchart of our proposed framework (each step is based on the numbered equations presented in the main manuscript). After providing the framework with the structures of water surfaces (shown here are two time steps for illustration) and the underwater point cloud, the model loss can be computed in a fully-differentiable fashion through multi-stage procedures. The gradients can be effectively back-propagated in the framework, so that all parameters can be updated in the same iteration.

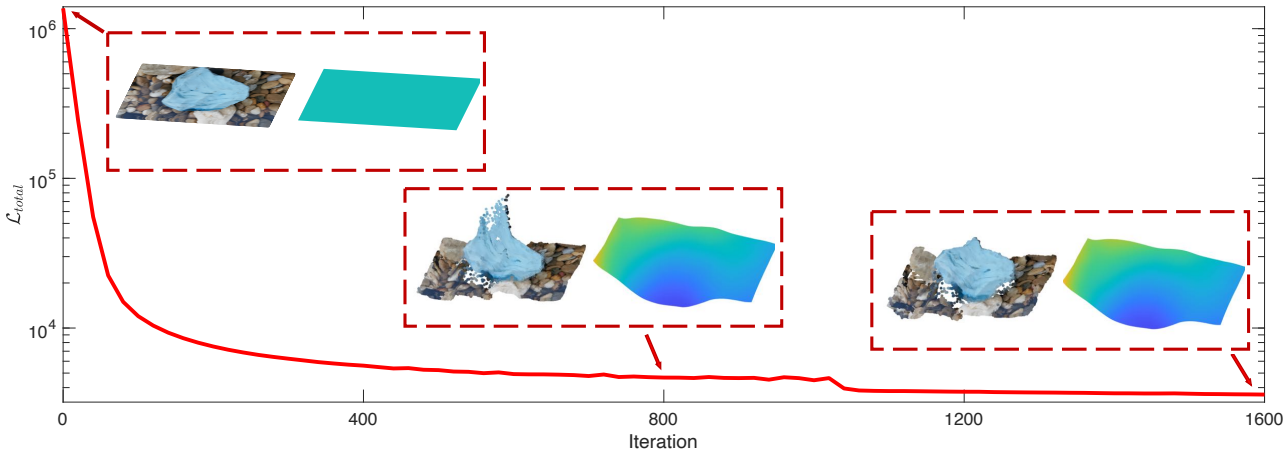


Figure 2: The evolution of the objective function versus iterations for the data shown in Fig. 1 of the main manuscript. The structures of water surfaces (one frame) and scene geometry are all initialized as planar surfaces. The objective function is effectively reduced, and accordingly, the parameters are progressively optimized, and yield a good representation of the scenes after 1600 iterations.

3. Experimental Comparisons

3.1. Water Surface

The primary focus of our work is the reconstruction of the underwater scene. However, in the process of this reconstruction, we also estimate the shape of the deforming water surface. Here, we conduct a simulated quantitative evaluation of this aspect, and compare our method to a SOTA single-camera fluid reconstruction approach [5]. Similar to our hardware setup, they use a single camera to capture refractive images of the background pattern, and the structure of fluid surfaces is estimated by a trained neural network. Their method simplifies the required equipment as compared to prior work, however, an undistorted frame is still

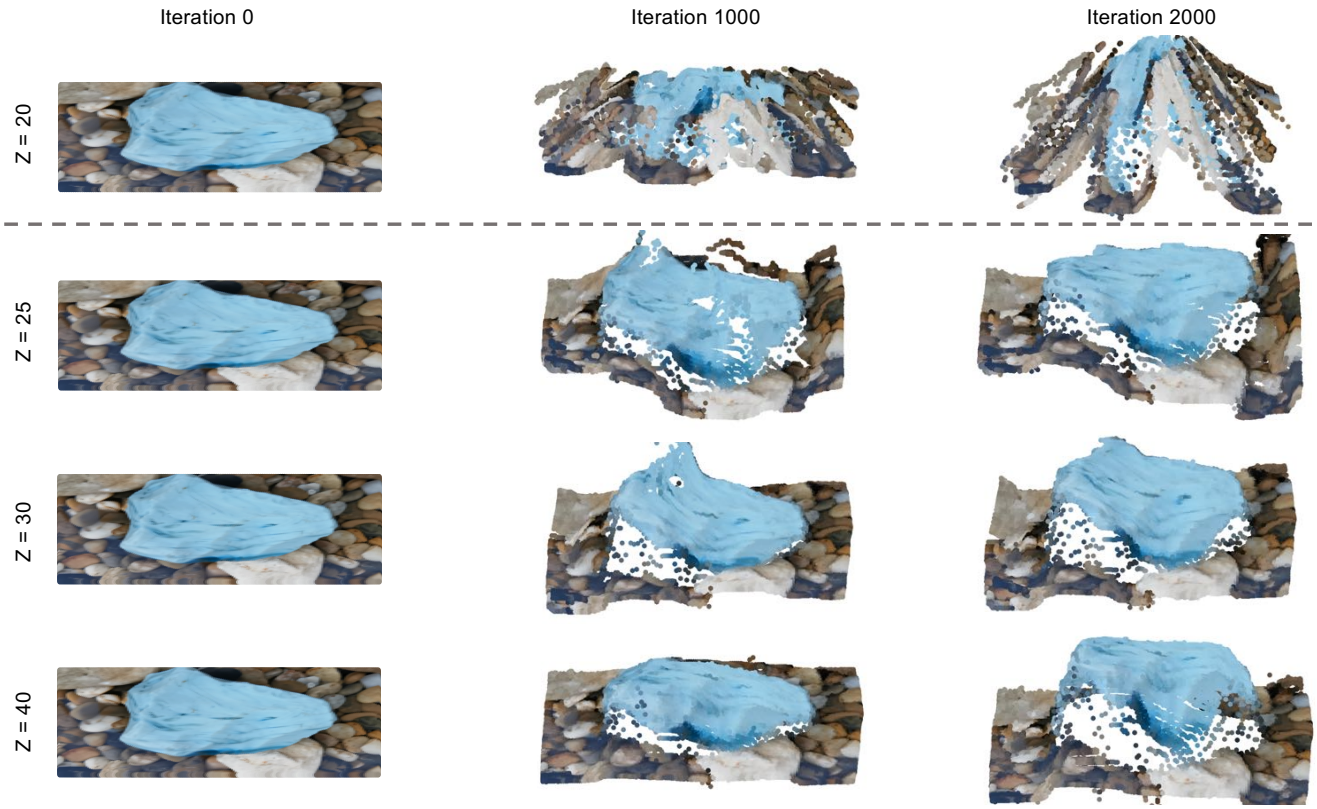


Figure 3: The reconstructions from different initial points and also the failure case at a degenerated initialization (top). The initialization of the underwater scene geometry is a planar surface with different axial depths.

required and serves as reference. For evaluation, the refractive images are rendered using the sample reference patterns as employed in [5] and synthetically generated time-varied water surfaces.

Table 1 shows the quantitative results on the recovered depth and surface normal. We use the root mean square error (RMSE) and absolute relative error (Abs Rel) as error metrics for estimated depth, and the root mean square error (RMSE) and average angular error (AAE) as error metrics for surface normal evaluation. Angular error (in degrees) measures the degrees between ground truth and the estimated surface normal. We find that our proposed model-based approach outperforms the existing single-camera fluid estimation method. It reveals that the temporal regularizer plays a significant role as it explicitly models a physical evolution of the water surface over time, and ties together all frames. The compared method employs a recurrent neural network module to encourage temporal consistency, and it may not exactly retrieve the physical process of fluid flows. We also notice that even though the water surface is modeled as a cubic b-spline surface, which implicitly enforces the spatial smoothness, we do find an explicit spatial smoothness term further improves the reconstruction.

Table 1: Quantitative results between the true and the estimated water surface.

Depth

Method	Metric	
	RMSE	Abs Rel
FSRN [5]	0.103	0.087
w/o spatial Loss	0.097	0.079
w/o temporal Loss	0.129	0.112
w/o both	0.143	0.118
Ours	**0.086**	**0.065**

Normal

Method	Metric	
	RMSE	AAE
FSRN [5]	0.064	4.33°
w/o spatial Loss	0.038	3.01°
w/o temporal Loss	0.046	3.27°
w/o both	0.049	3.69°
Ours	**0.030**	**2.21°**

We also show a qualitative comparison for the recovered

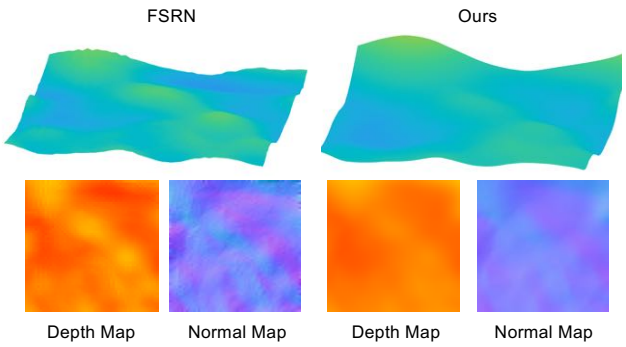


Figure 4: Qualitative comparisons with FSRN [5] for water surface estimation for the data captured in a laboratory environment. The depth and surface normal are normalized for fair comparisons. Their method requires an additional undistorted frame as input. The recovered shapes are overall consistent, while the reconstruction from ours is smoother with fewer noise-like features.

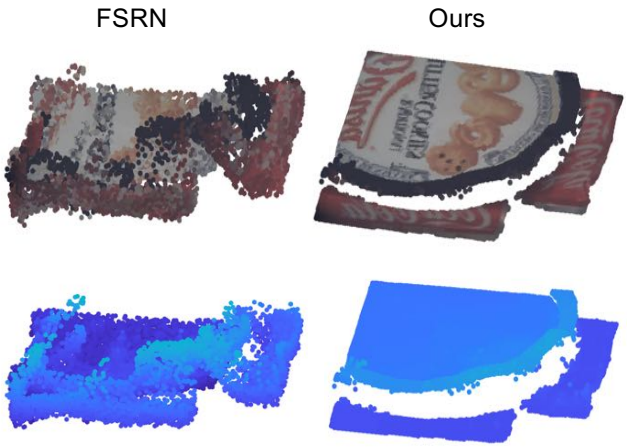
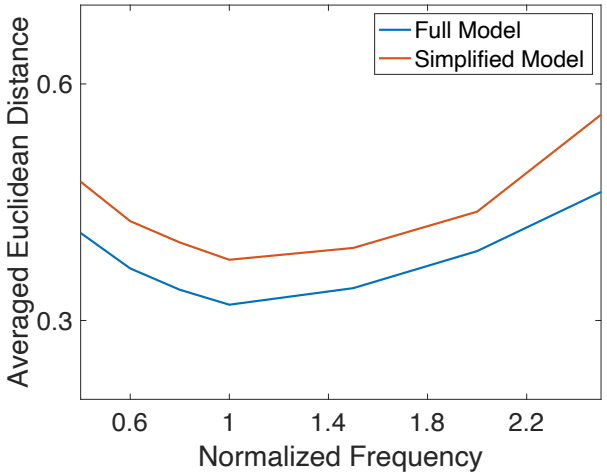


Figure 5: Qualitative comparisons with the modified FSRN [5] for underwater scene estimation.

water surface between ours and FSRN [5] in Fig. 4. The results demonstrate that the recovered structures from both methods are overall consistent. However, our method explicitly enforces smoothness, and the generated normal map exhibits fewer noise patterns. This makes our estimation to be more accord with the physical characteristics. Due to relatively lower estimation accuracy on the water surface for the compared method, inferring the 3D geometry of the underwater scene is prone to error as demonstrated in Fig. 3 of the main manuscript. We also visualize the reconstructed point clouds with color-coded depth in Fig. 5.

Table 2: Quantitative results compared with a multi-camera approach [3] on point cloud estimation.

	Ours			[3]
Number of cameras	1			9
Number of frames	30	60	120	-
AED	0.254	0.233	0.227	**0.192**



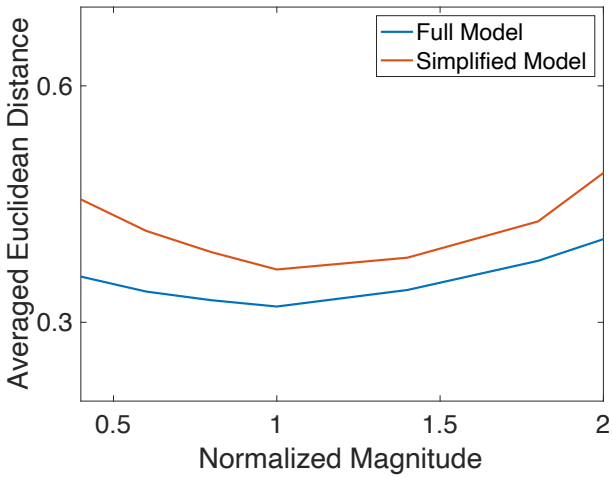


Figure 6: Average Euclidean distance between the true and the estimated point clouds on synthetic data with varying wave frequency and wave magnitude (the frequency and magnitude are normalized with respect to the frequency and magnitude which provide the best recovery, respectively).

3.2. Underwater Scene

For point cloud reconstruction evaluation, besides the ablation study we show in the main manuscript, we also conduct a numerical comparison with a multi-camera sys-

tem [3]. They use a 3×3 camera array to capture the scenes from different viewpoints. We implement their algorithm for point cloud estimation as no source code is publicly available. In their system, the parameters for the camera array are pre-calibrated and the reconstruction solves the scene geometry only. This is a standard multi-view 3D reconstruction approach which can provide a robust and accurate estimation. The baseline between adjacent cameras is set to 5 units (this could be 5-20 times larger than using water distortion for multi-view triangulation). Table 2 shows the numerical comparisons using average Euclidean distance (AED) as error metric. As expected, using the multi-camera system with a wide baseline yields a more accurate 3D geometry reconstruction, but it heavily relies on the acquisition system, which is expensive to build and calibrate such a system. We also show qualitative comparisons in Fig. 7.

We conduct another synthetic experiments with varying wave frequency and magnitude. The reconstruction error of the estimated point clouds is shown in Fig. 6. When the wave frequency or wave magnitude increases, camera observes stronger distortions, which yields a larger baseline for triangulation. Therefore, the average error of the reconstructed point clouds decreases at the early phase. However, when further increasing the frequency or magnitude, the distortion becomes severe, in which case a precise image registration cannot be achieved. The reconstruction error will increase accordingly. It also reveals that our proposed full model consistently outperforms the simplified model without using the projection loss and confidence mask. This further verifies the effectiveness of these two terms at the point cloud reconstruction.

We also show the comparisons with the underwater geometry from a 3D scan. Fig. 8 reveals the qualitative comparisons between the scanned models and our reconstructions. The scanned models were obtained without water interference. The reconstructions from a 3D scan exhibit finer recovery of its geometrical structure, however, our framework could generate an adequate representation of the scenes under such sophisticated conditions with a simple hardware setup.

4. Acknowledgements

We thank Ibrahim Alhawsawi and Hussam Altalhi from FalconViz for providing the 3D scanned models.

References

[1] Julie Delon and Bernard Rougé. Small baseline stereovision. *Journal of Mathematical Imaging and Vision*, 2007. 1

[2] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 2004. 1

[3] Yiming Qian, Yinqiang Zheng, Minglun Gong, and Yee-Hong Yang. Simultaneous 3d reconstruction for water surface and underwater scene. In *ECCV*, 2018. 4, 5

[4] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1

[5] Simron Thapa, Nianyi Li, and Jinwei Ye. Dynamic fluid surface reconstruction using deep neural network. In *CVPR*, 2020. 2, 3, 4

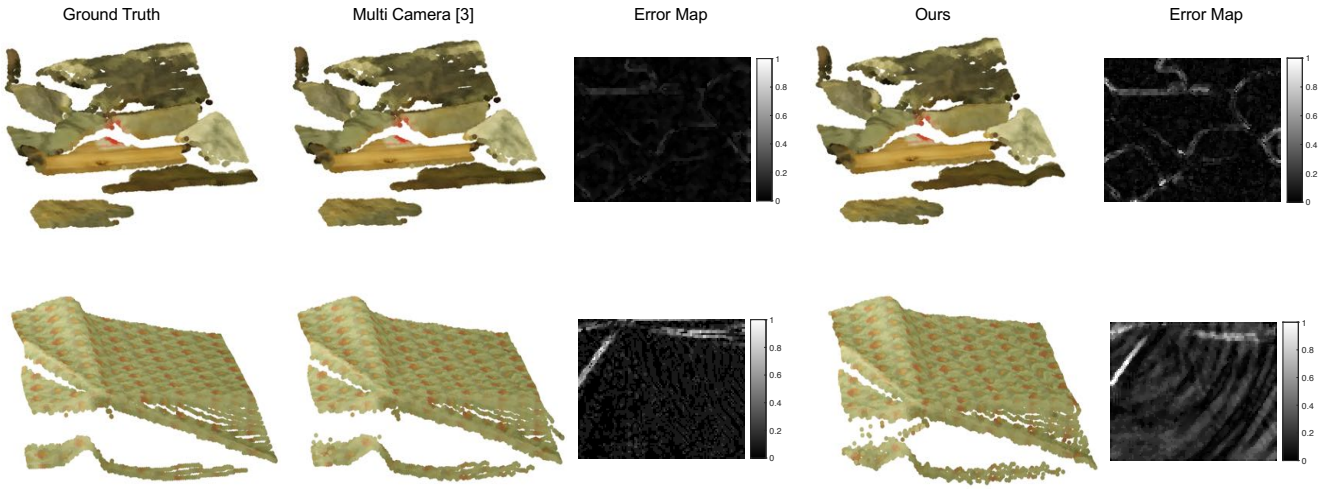


Figure 7: Qualitative comparisons between the reconstructed geometry and ground truth on synthetic data. Our recovered results exhibit a high degree of consistency with the ground truth with regard to geometric structures. The overall reconstruction error is comparably higher than using a multi-camera system (a 3×3 pre-calibrated camera array), but our acquisition system is simple and free of calibration.

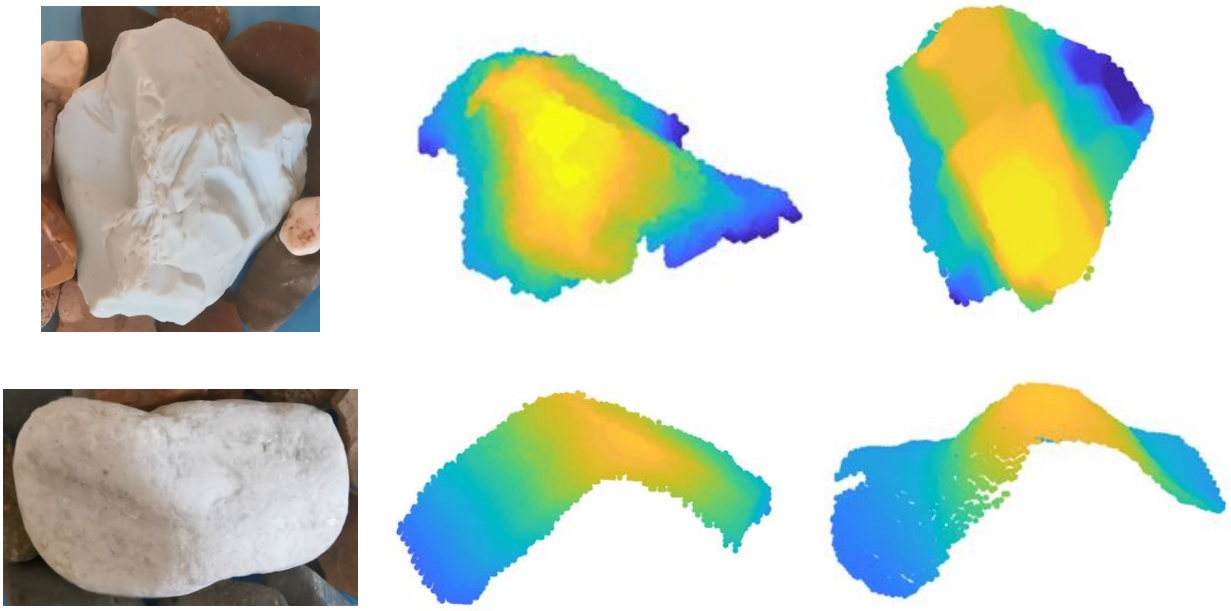


Figure 8: Visual comparisons between the 3D scanned models and our reconstructions through moving water surfaces. From left to right, the distorted frames from the video for two stone scenes, the ground truth 3D structures measured using a 3D scanner and our reconstructed geometry. The averaged absolute error on the projected depth map is 4.24 mm for the first scene and 3.01 mm for the second scene. Notice that due to the non-convexity of the problem, our solution is a locally reasonable representation of the scenes. Therefore, we scale our reconstructions with respect to the scanning results.