Real-time Acquisition and Reconstruction of Dynamic Volumes with Neural Structured Illumination

Supplementary Material

9. Supplementary Video

The video mainly consists of two parts: reconstruction results of 4 captured sequences, and comparisons with different methods on the reconstruction of a synthetic sequence.

For the first part of the video, the layout is as follows. We show the captured input images from each camera on the left. The captured image at a non-input view is in the center, with a time code displayed in the lower-left corner. And the 3D reconstruction at a view close to the center non-input view is visualized on the right.

For the second part, we compare with PINF [10] and GlobalTrans [15] on a synthetic smoke sequence with 190 frames. The same 3 input cameras are used for all methods. In the video, we compare the reconstruction results, rendered at one input view and a novel view. Quantitative errors in SSIM/PSNR/RMSE are reported at the bottom-right corner of each rendered volume. Note that SSIM and PSNR measure the 2D error of the rendered volume at a view, while RMSE measures the error over the entire 3D volume. In addition, the errors averaged over all frames are reported in Tab. 1 and 2. In all cases, our approach outperforms competing approaches in terms of result quality.

We also compare the computation time of different methods on reconstructing the synthetic sequence. For a fair comparison, we conduct all profiling experiments on a single GeForce RTX 3090 for back-compatibility with GlobalTrans, whose code cannot be executed on RTX 4090 as in our main paper. The results are 13 seconds, 13 hours and 84 hours for our approach, PINF [10] and GlobalTrans [15], respectively.

View	Ours	PINF[10]	GlobalTrans[15]
Input(1/3)	0.98/34.36	0.96/29.04	0.96/28.66
Novel	0.97/33.15	0.95/29.83	0.94/27.14

Table 1. Comparison with different methods on reconstruction quality (SSIM/PSNR) of a synthetic sequence. We list the reconstruction errors averaged over all frames shown in the final part of the supplementary video. The second row shows the reconstruction errors for one of the three input views (i.e., cam_0). The situation with other input views is similar. The third row is the reconstruction errors for a novel non-input view.

Ours	PINF[10]	GlobalTrans[15]
1.20×10^{-2}	2.72×10^{-2}	2.50×10^{-2}

Table 2. Comparison with different methods on reconstruction quality (RMSE) of a synthetic sequence. The RMSE is computed as the error averaged over each reconstructed 3D volume.

10. Calibrations

10.1. Geometric Calibration

We calibrate the intrinsic and extrinsic parameters of the projector and the cameras in the following 4 steps.

(1) We pre-calibrate the intrinsic parameters of all cameras with a chessboard pattern.

(2) We pre-calibrate the intrinsic parameters of the projector using a calibration board with printed ARTags and one of the cameras. Please refer to Fig. 13-a for an illustration. We cast vertical and horizontal lines from the projector to the board (Fig. 13-c), and take pictures with the camera. In each captured image, the screen-space coordinates of each intersection can be estimated with sub-pixel accuracy, and the extrinsic parameters of the board can be computed from the ARTags. With the additional help of the camera intrinsic parameters from the previous step, we calculate the camera-space 3D positions of each intersection. We repeat this process for different combinations of rotated board/camera. The 3D positions of the intersections along with their 2D counterparts on the projector plane are used to compute the intrinsic parameters of the projector.

(3) We pre-calibrate the extrinsic parameters of the projector and all cameras with the calibration board. The board is rotated to different angles, one at a time (Fig. 13-b). Just like in step (2), we cast vertical and horizontal lines to the board. With the intrinsic parameters of each camera, we calculate the camera-space 3D positions of each intersection. The 3D positions of all intersections at each camera view are then used to compute the extrinsic parameters of the corresponding camera with respect to the projector.

(4) Similar to existing work [40], all pre-calibrated parameters are jointly fine-tuned in an end-to-end fashion with differential optimization, by minimizing the reprojection error of each intersection at each camera view. We fine-tune the intrinsic and extrinsic parameters for 20,000 epochs with a learning rate of 10^{-3}.

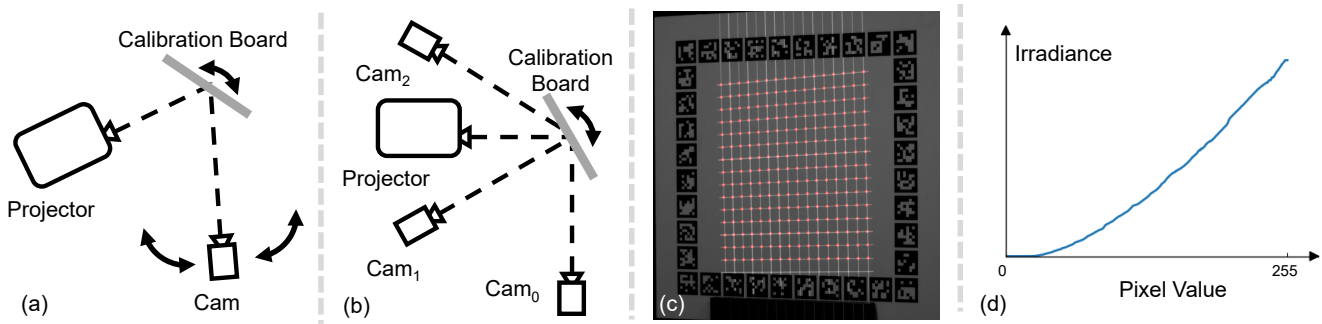

Figure 13. Geometric calibration (a-c) and the projector response curve (d). (a) Pre-calibration of the intrinsic parameters of the projector. (b) Pre-calibration of the extrinsic parameters of the projector and all cameras. (c) A photograph of the calibration board with projected horizontal and vertical lines. The reprojected intersections points are marked in red. (d) The projector response curve.

10.2. Radiometric Calibration

Our machine vision cameras can be set up to employ a linear response curve. For the projector, we directly capture its response curve as follows. We cast uniform patterns onto the calibration board, with the projector pixel intensity changes from 0 to 255. For each such pattern, we record the pixel intensity averaged over a square region observed by one calibrated camera. The collection of all pairs of projector/camera pixel intensity is the response curve, as plotted in Fig. 13-d. To linearize the projector, we apply the standard approach of inverting a 1D cumulative distribution function computed from the response curve.

10.3. Synchronization

All cameras are synchronized via a hardware trigger. In addition, we project 3 special tags along with each light pattern to facilitate projector-camera synchronization, as our projector does not support external triggers. Please refer to the inset of Fig. 2-a for tag examples.

Specifically, each tag is a white box. The center tag only appears with the first light pattern, to mark the start of our group of patterns. The left tag is projected with each odd-numbered pattern, while the right with each even-numbered pattern. An ideal synchronization will result in either the left or right tag in a captured image. If this is the case, the synchronization is finished. Otherwise, both boxes of different intensities can be observed. We then estimate the offset to the starting time of one exposure, by dividing the observed intensities by pre-calibrated intensities of the white boxes. Finally, we add this offset as a feedback to a proportional–integral–derivative (PID) algorithm, to adjust the start time of the exposure. Once the algorithm converges, the synchronization is done and we can start to capture the physical phenomenon.