Supplemental material:
Single-shot dense active stereo with pixel-wise phase estimation based on grid-structure using CNN and correspondence estimation using GCN

1. CG Images to training U-Nets for phase and code detections and GCN

1.1. CG dataset for U-Nets training

Fig. 1 shows examples of the CG-synthesized training images of U-Nets for phase and code detections. The U-Nets are trained with 200x200 patches, mini-batch sizes of 16, 5400 sets of image patches. They are sampled from 360 sets of rendered images shown in Fig. 1. The optimizer was adam, with l^2-loss functions for sinusoidal signals and cross entropy for code signals. They are trained for 300 epochs.

1.2. CG dataset for GCN training

Fig. 2 shows examples of source data for generating training data for GCN for correspondence prediction. Using CG, source pattern and corresponding IDs are projected into images. From the images, grid graphs with corresponding ID data as teacher signals are generated.

2. Phase-estimation results for various patterns

Fig. 3 shows phase detection results for various patterns with explicit/implicit grid structures. For comparison, ground truth phase values calculated from phase-shift method are also shown. The projected patterns were [2, 3], respectively. The proposed method could evaluate grid phases stably. Note that these pattern does not necessarily have explicit grid lines.

3. Network detail and parameters for training GCN for correspondence prediction

3.1. GCN-II network detail

Fig. 4 shows the architecture of the GCN for correspondence prediction, which is depicted as a box with "GCNII layer" in the Fig. 6 in the original paper. Since four different directions of up, down, left, right and adjacency are taken into account, five edge matrices are used. Thus, the number of GCN layers is five. Then, the GCN is trained with mini-batch sizes of 8, from 864 sets of graphs with ground-truth IDs that are generated from CG. The graphs from the training data are augmented by extracting their subgraphs. The optimizer was adam, with cross entropy loss between the predicted and ground-truth corresponding node IDs. The number of iterations of training was 100000 (in the number of mini-batch optimization).

3.2. Intermediate images for 3D reconstruction procedure

Fig. 5 shows intermediate images for 3D reconstruction procedure of an input image.

3.3. Images for the auto-calibration experiment described in Section 6.3 of the paper

Fig. 6 shows input images and feature points used for the auto-calibration experiment for Table 2, described in Section 6.3 of the original paper. It is confirmed that much larger number of feature points can be used for auto-calibration with the proposed method than Furukawa's *et al*. [1], resulting in better calibration.

4. Various object reconstruction results

4.1. Static objects

Fig. 7 shows examples of reconstruction results of our method. As can be seen, curved surfaces with textures are robustly estimated by our method.

4.2. Moving objects

Fig. 8 shows another example of moving object, where throwing ball is reconstructed by our method. Since ordinary camera can capture only blurry images as shown in Fig. 8(a), we set shutter speed 1ms and shape is successfully reconstructed. Kinect Azure also fails to capture correct depth, *i.e*., shapes are distorted and split into three blocks.

References

[1] Ryo Furukawa, Ryunosuke Masutani, Daisuke Miyazaki, Masashi Baba, Shinsaku Hiura, Marco Visentini-Scarzanella,

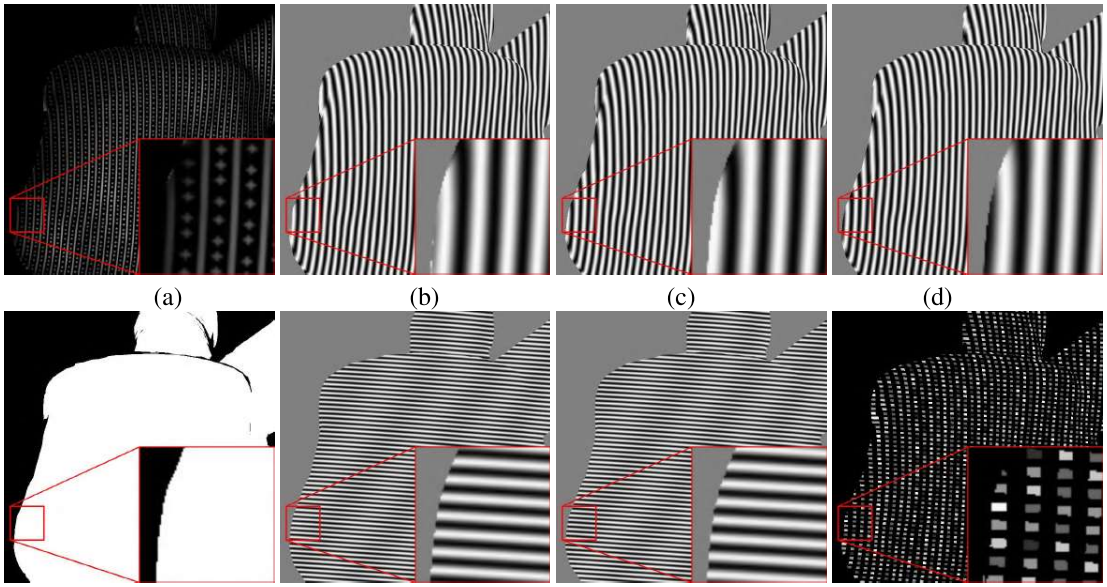

Figure 1. CG-Synthesized training images for U-Nets predicting pattern phases and codes: (a) Input image; (b) image where horizontal sinusoidal pattern with $\frac{0*2\pi}{5}$ phase-shift is projected; (c) horizontal sinusoidal pattern with $\frac{1*2\pi}{5}$ phase-shift; (d) horizontal sinusoidal pattern with $\frac{2*2\pi}{5}$ phase-shift; (e) mask image; (f) image where vertical sinusoidal pattern with 0 phase-shift is projected; (g) horizontal sinusoidal pattern with $\frac{\pi}{2}$ phase-shift; and (h) codes at grid points of the dot-line pattern. Only the white region of mask images (e) are used for training loss for sinusoidal patterns (b,c,d,f,g), and only the non-black regions are used for training loss for code images (h).

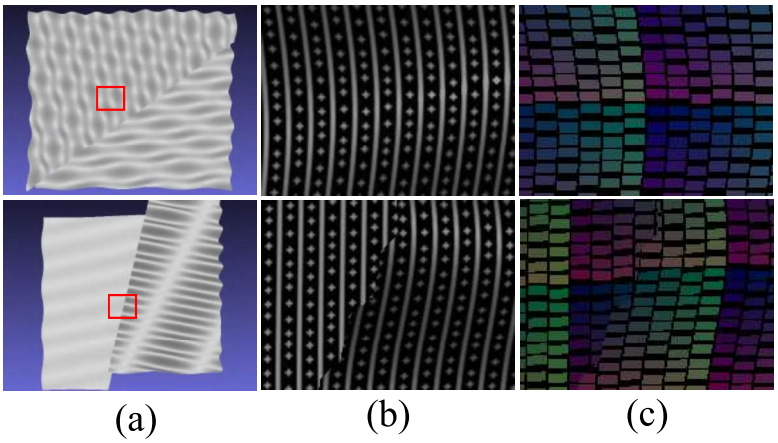

Figure 2. Source images for generating training graph data for GCN-based correspondence prediction: (a) synthesized shapes; (b) images where dot-line patterns are projected; and (c) images where ground-truth correspondence ID data are projected.

Hiroki Morinaga, Hiroshi Kawasaki, and Ryusuke Sagawa. 2-dof auto-calibration for a 3d endoscope system based on active stereo. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7937–7941. IEEE, 2015. 1, 5

[2] Ryo Furukawa, Shiro Oka, Takahiro Kotachi, Yuki Okamoto, Shinji Tanaka, Ryusuke Sagawa, and Hiroshi Kawasaki. Fully auto-calibrated active-stereo-based 3d endoscopic system using correspondence estimation with graph convolutional network. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4357–4360. IEEE, 2020. 1, 3

[3] Xiaojun Jia and Zihao Liu. One-shot m-array pattern based on coded structured light for three-dimensional object reconstruction. *Journal of Control Science and Engineering*, 2021, 2021. 1, 3

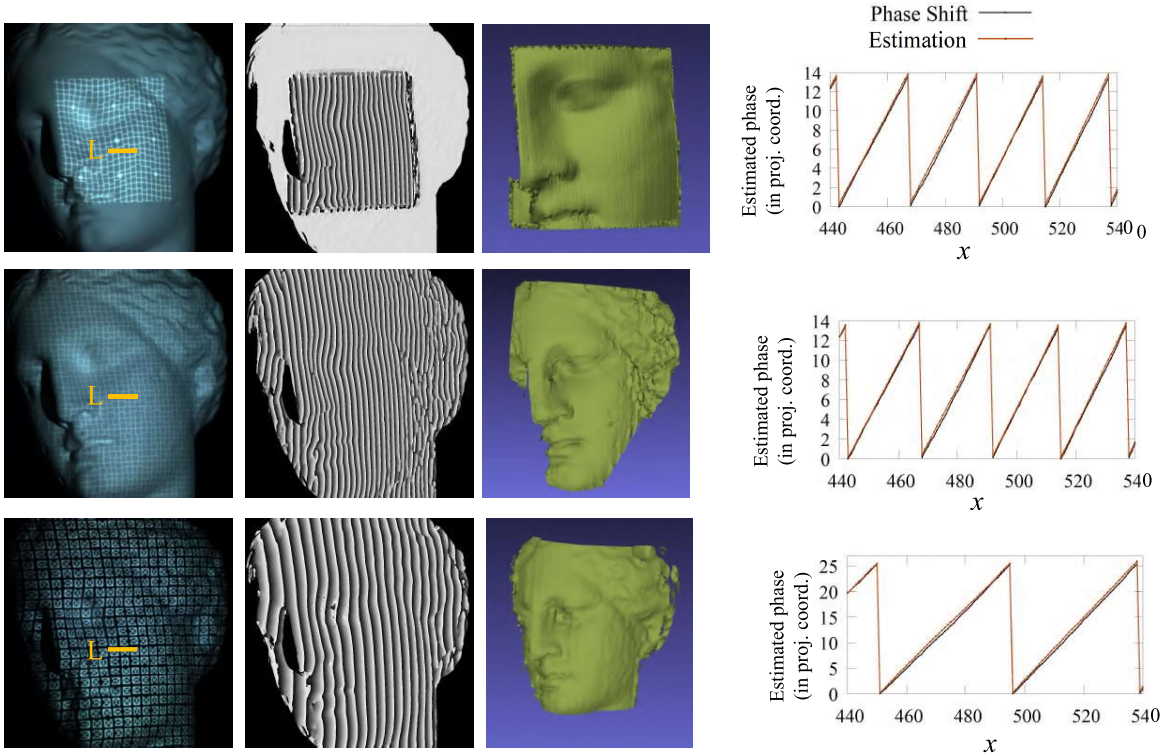

Figure 3. Phase detection results for various grid-like patterns. (a) Source images, (b)estimated phase values, (c) 3D reconstruction results, and (d) profile values compared to ground truth value (phase shift method). The pattern used are from [2] (the 1st row), an original pattern (the 2nd row), and [3] (the 3rd row), respectively.

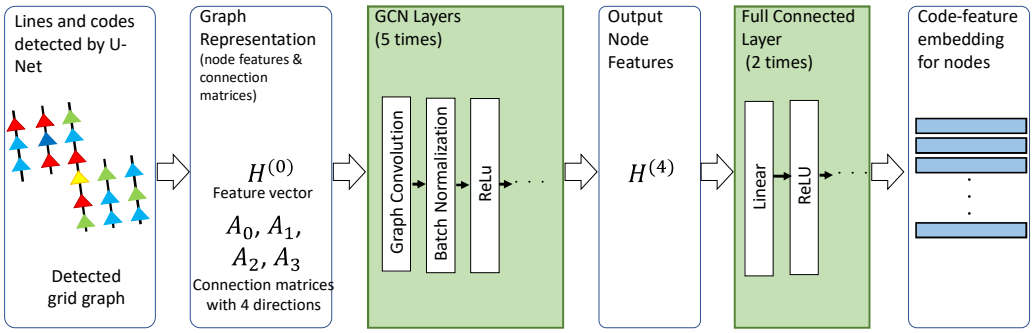

Figure 4. Network architecture of GCN for calculating node-wise code-feature embedding.

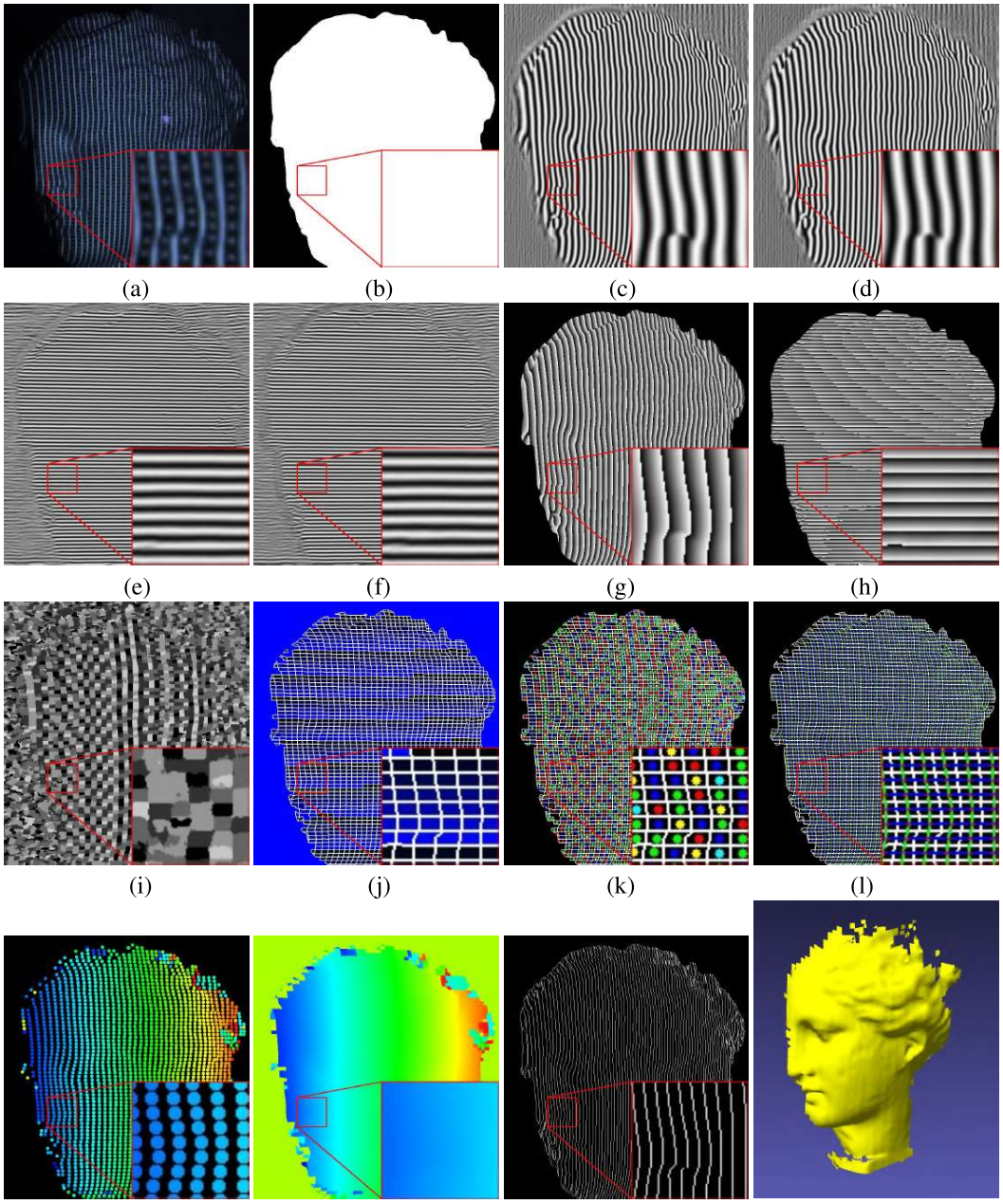

Figure 5. Intermediate images for the proposed 3D reconstruction process: (a) Input image; (b) predicted mask image; (c) predicted horizontal sinusoidal signal (phase-shift $\frac{0*2\pi}{5}$); (d) predicted horizontal sinusoidal signal (phase-shift $\frac{1*2\pi}{5}$); (e) predicted vertical sinusoidal signal (phase-shift $\frac{1*2\pi}{5}$); (f) predicted vertical sinusoidal signal (phase-shift $\frac{1*2\pi}{5}$); (g) horizontal phase signal calculated from 5 images including (c) and (d); (h) vertical phase signal calculated from 2 images of (e) and (f); (i) predicted code image; (j) phase-grid superpixels segmented from (g) and (h); (k) code information sampled at the centers of phase-graph superpixels; (l) phase-grid graph extracted as adjacency graph of phase-grid superpixels (*i.e.*, (j)); (m) GCN-based correspondence prediction results visualized by column IDs of the pattern; (n) dense correspondence estimation obtained from (m) and (g); (o) another visualization of (n) with iso-value curves; and (p) a shaded 3D reconstruction result.

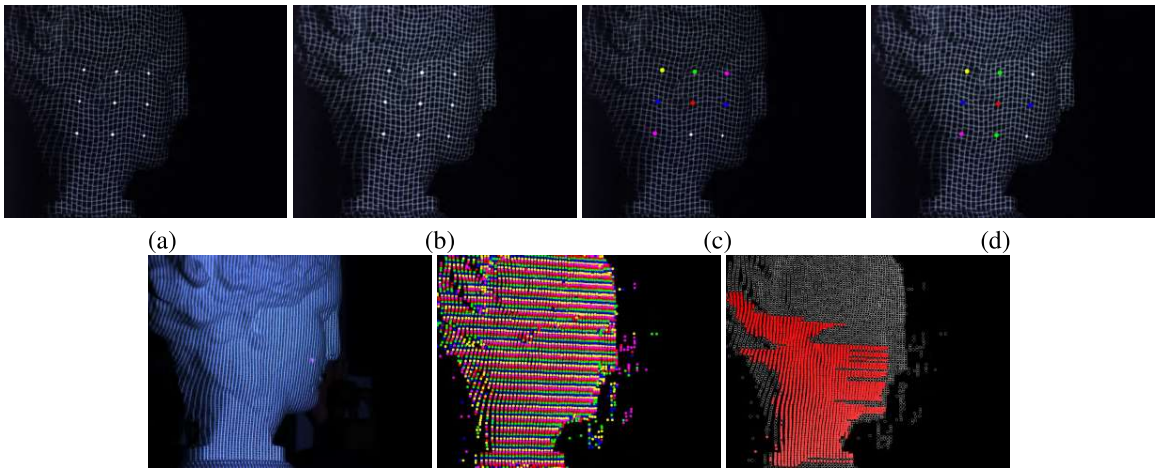

Figure 6. Images for the auto-calibration experiment (Sec. 5.2 of the paper): (a,b)Input images (2 of 9 images) for Furukawa *et al*. [1] captured with different distances to the camera; (c,d) detected points used for auto-calibration for Furukawa *et al*. [1]; (e) an input images (1 of 2 images) for proposed method; (f) GCN-predicted IDs for the grid points; (g) points used for auto-calibration (unused points are removed because of uncertain vertical IDs).

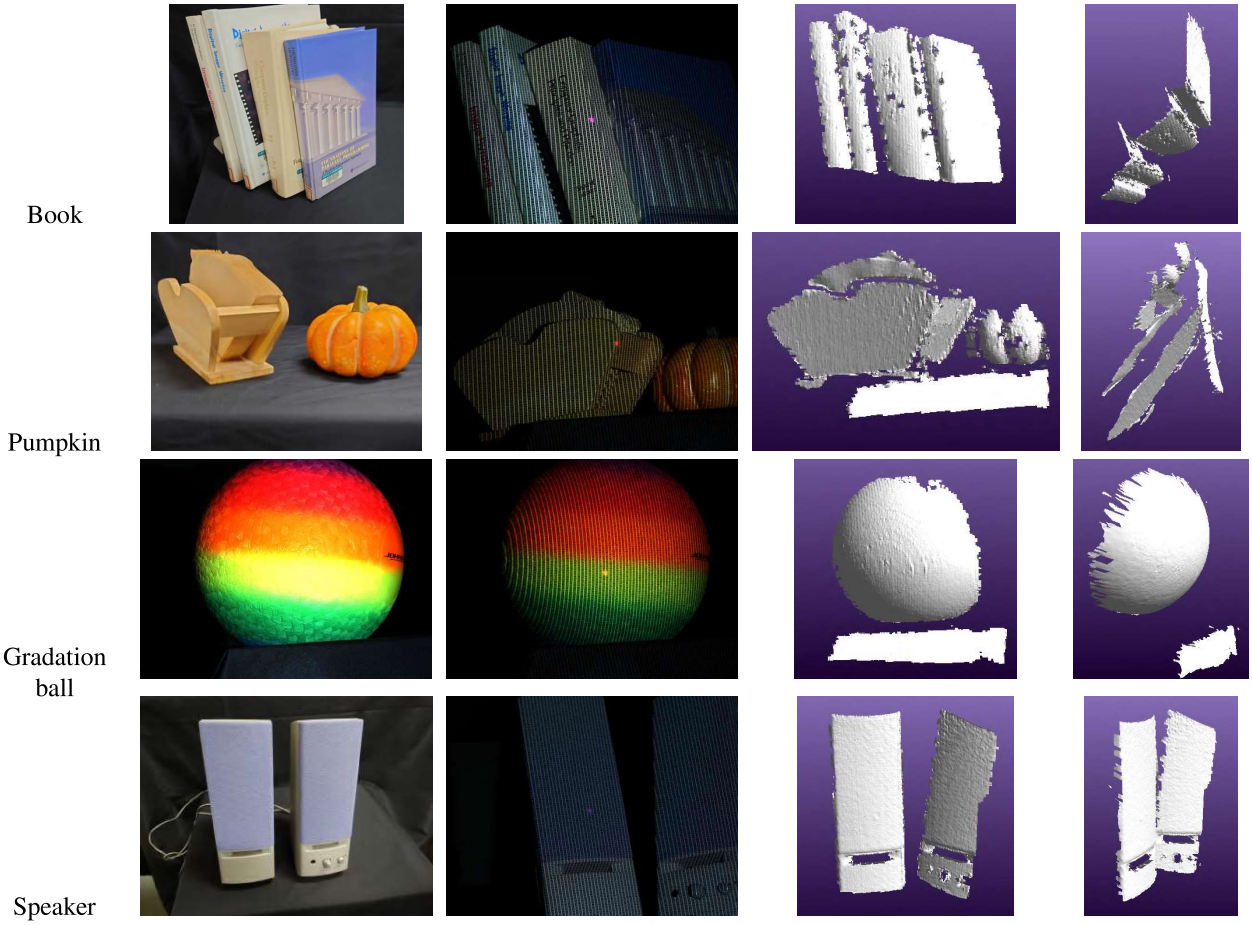

Figure 7. Reconstruction results of various static objects with textures.

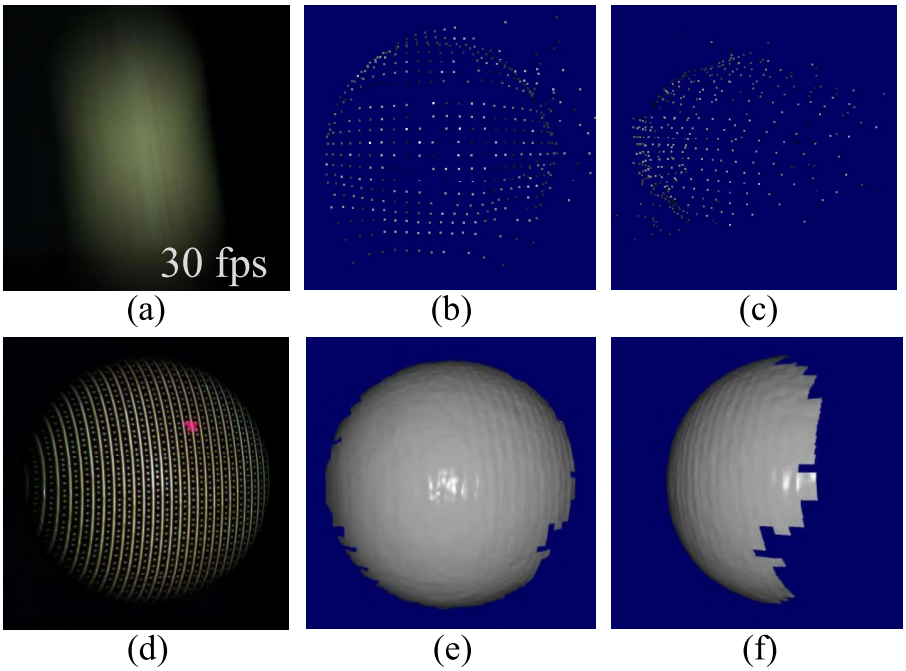

Figure 8. Measurement of fast moving object (a throwing ball). (a): Image captured by normal shutter speed. (b)(c): Measurement results by Kinect v2. (d): Image captured by fast shutter speed. (e)(f): Measurement results by the proposed method.