Supplemental: Event Camera Depth Estimation from Epipolar Plane Images

Joshua D. Rego[1], Sanjeev J. Koppal[2], and Suren Jayasuriya[1]

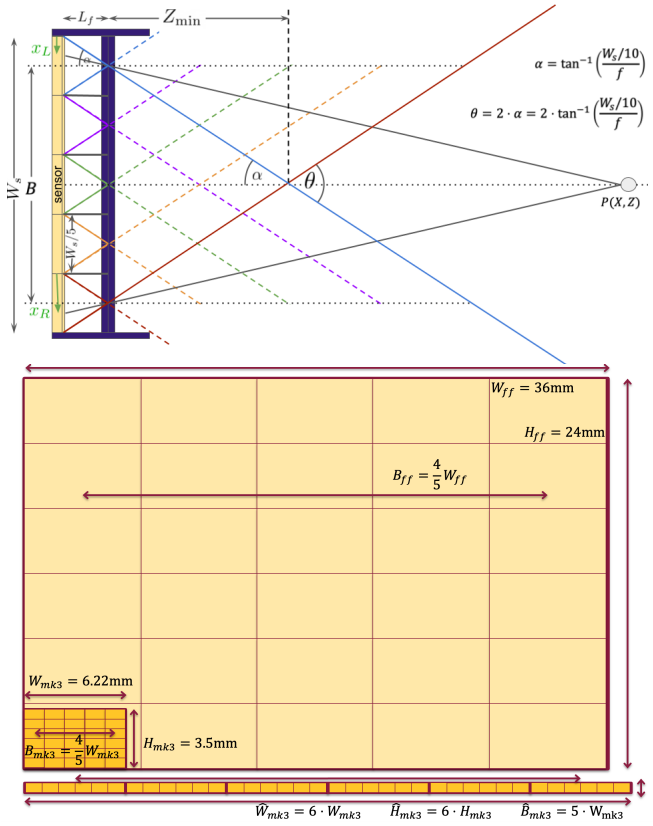

Fig. 1. (top) Theoretical 5-view camera, (bottom) sensor comparison between Prophesee Mk3 and a full-frame sensor. We can see the large difference in sensor size and baseline between the two sensors. We show results for the 5-view camera for both these sensor sizes in the main paper that shows that this difference in baseline greatly influences the possible disparity, detectable depth range, and quality of depth estimation for our system.

I. Single-shot Analysis.

In the pursuit of advancing event camera depth estimation using epipolar plane images, we encounter inherent hardware limitations that necessitate the use of simulations to explore the feasibility of single-shot depth estimation with a single event camera. For this analysis, we simulate a camera setup where a single sensor is segmented by five equally spaced lenses. This configuration allows us to capture multiple views from a single event camera snapshot, thereby enabling depth estimation using EPIs from a single capture instance. The simulations are tailored to test the effectiveness of both small sensor event cameras, representative of our current Prophesee Mk3 event camera setup, and a hypothetical full-frame event sensor configuration for a more idealized comparison. This enables us to assess the potential scalability of our method to both existing and possible future event camera technologies.

a) Parameter Selection.: Choosing the appropriate sensor size, resolution, and lens spacing is critical to achieve a suitable depth estimation range. To this end, we derive several key equations intended to guide these choices and inform us of limitations of the camera parameters, such as the achievable depth range. Fig. 1 illustrates a simplified diagram of the horizontal view of our 5-view camera with a single sensor and 5 lenses that section the sensor for each planar view. From this diagram, we can use like-triangles to derive most of the necessary equations. For a sensor width W_s, each sensor section will have a width of $W_s/5$.

The field-of-view, θ, for our 5-view camera would be defined as the angle within which a point in the scene would be captured by all five views. As seen in the diagram, this can be simply defined as the intersection of the field-of-views for the left-most and right-most views. Since the views are all equally spaced over the sensor, the half-angle, $\alpha = \theta/2$, can determined from the right angle triangle between the lens and sensor as:

$$\alpha = \tan^{-1}\left(\frac{W_s/10}{f}\right), \tag{1}$$

and the FoV, θ, as:

$$\theta = 2 \cdot \alpha = 2 \cdot \tan^{-1}\left(\frac{W_s/10}{f}\right) \tag{2}$$

For most cases of our simulation, we keep the field-of-view constant, so that the focal length, is adjusted for different sensor sizes to maintain the same field-of-view. For this it is useful to define focal length, f in terms of the sensor width, W_s, and the FoV, θ, as:

$$f = \frac{W_s/10}{\tan(\theta/2)} \tag{3}$$

To determine the minimum depth a point in the scene must be, so that it is captured by all five views of the camera, we derive the equation as:

$$Z_{min} = (2 \cdot W_s/5) \cdot \frac{f}{W_s/10} = 4 \cdot f \tag{4}$$

This work was supported by NSF IIS-1909192 and a gift from Qualcomm.

[1]J. Rego is with the School of Electrical, Computer and Energy Engineering at Arizona State University. S. Jayasuriya is with the School of Arts, Media and Engineering and the School of Electrical, Computer and Energy Engineering at Arizona State University. Contact: jdrego@asu.edu

[2]S. Koppal is with the Department of Electrical and Computer Engineering at the University of Florida.

Substituting Eq. (3) into Eq. (4), we get:

$$Z_{min} = 4 \cdot \frac{W_s/10}{\tan(\theta/2)} = \frac{2 \cdot W_s/5}{\tan(\theta/2)} \tag{5}$$

To determine the maximum range of detectable depth, we can use the stereo camera equation for disparity of two cameras on the same plane in a stereo setup. For our case of the 5-view camera, we would use the left-most and right-most views as the stereo cameras to determine the baseline, B. This is shown in Fig. 1, for a point $P(X, Z)$, where the disparity is determined as:

$$d = x_R - x_L = \frac{B \cdot f}{Z} \tag{6}$$

When determining the maximum depth, we want the disparity to be greater than the pixel pitch of the senor. We can solve for the max depth, Z_{max} from Eq. (6), by setting disparity, d, as the pixel pitch, Δx, and substituting f from Eq. (3) to give us:

$$Z_{max} = \frac{B \cdot f}{\Delta x} = \frac{B \cdot W_s/10}{\Delta x \cdot \tan(\theta/2)} \tag{7}$$

These equations help us understanding the trade-offs between sensor resolution, size, and the spacing of the lenses. For our experiments, when comparing between the sensor specifications of the Prophesee Mk3 camera and a full-frame sensor, we use these equations to determine an appropriate focal length for a fixed FoV of 60deg. and find the depth range possible for either sensor.

In the case of the Prophesee sensor, with a sensor size of 6.22x3.5mm, pixel-pitch, $\Delta x = 4.859\mu$m, baseline, $B = 4.976$mm, we determine the minimum depth, $Z_{min} = 4.309$mm and the maximum depth, $Z_{max} = 1.103$m. The full-frame sensor, with a sensor size of 36x24mm, pixel-pitch, $\Delta x = 28.125\mu$m, baseline, $B = 28.8$mm, we determine the minimum depth, $Z_{min} = 24.94$mm and the maximum depth, $Z_{max} = 6.385$m.

A. Hough Transforms

In Fig. 2, we show the Hough transform space for EPIs in 3 different cases along with the detected lines. The first Hough space (top), is for a clean EPI with four simulated random lines. The middle Hough space is for an EPI generated from a real captured scene that is noisy, while the bottom Hough space is for the EPI from the real scene after the EPI has been denoised. The Hough space from the clean simulated EPI clearly shows the separation of the sinusoids that signify all possible lines for a particular (x,y) pixel. The intersection at (r, θ) of these curves are where a line belong to multiple possible pixels. From this we can determine where the lines are most likely from the number of sinusoidal intersections.

For the Hough space for the noisy EPI, this becomes much more difficult to determine as noise can contribute to these added intersections in the Hough space where possible lines pass through the noise points. We can see after denoising that the Hough space has much better separation where the intersections are easier to distinguish.

B. Depth on simulation examples

In Fig. 3, we show the depth estimation results from our method for a sample scene that was simulated using the event simulator from a rendered scene in Blender. This scene only has three objects, a Stanford bunny at 0.5m away from the camera, a teapot at 2.5m away, and a brick wall at 3m away. As can be seen from the figure, the disparity map and 3D event points produced are fairly accurate to the true depth of the scene object. The EPI images for three rows also show closely matched detected lines to the event lines. Additionally we show comparisons of simulated scenes against our baselines in Fig. 4 of the same scene (top), as well as a more complicated classroom scene (bottom).

C. Depth on real examples

We also capture real scenes with the Prophesee Mk3 event camera on a linear motorized rail system. Results for two of these scenes are shown in Fig. 5 with the initial event frame (top-left), the EPIs with detected lines (bottom-left), the resulting disparity map (top-right), and a 3-D representation of the events with depth (bottom-right). A comparison with baseline methods for an additional scene is show in the main paper.

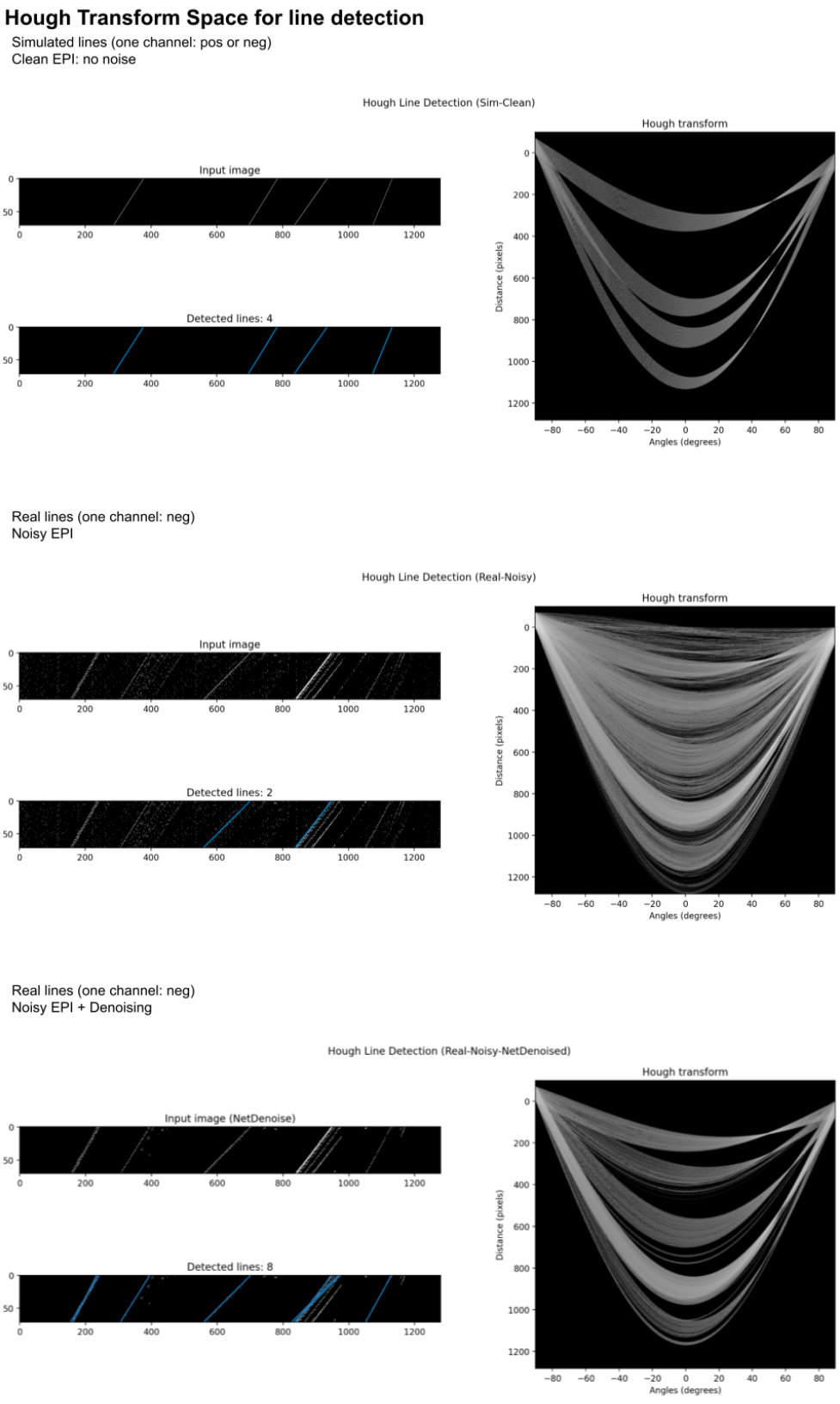

Fig. 2. EPI lines with corresponding Hough transform space for (top) clean lines, (middle) noisy scene, (bottom) denoised EPI

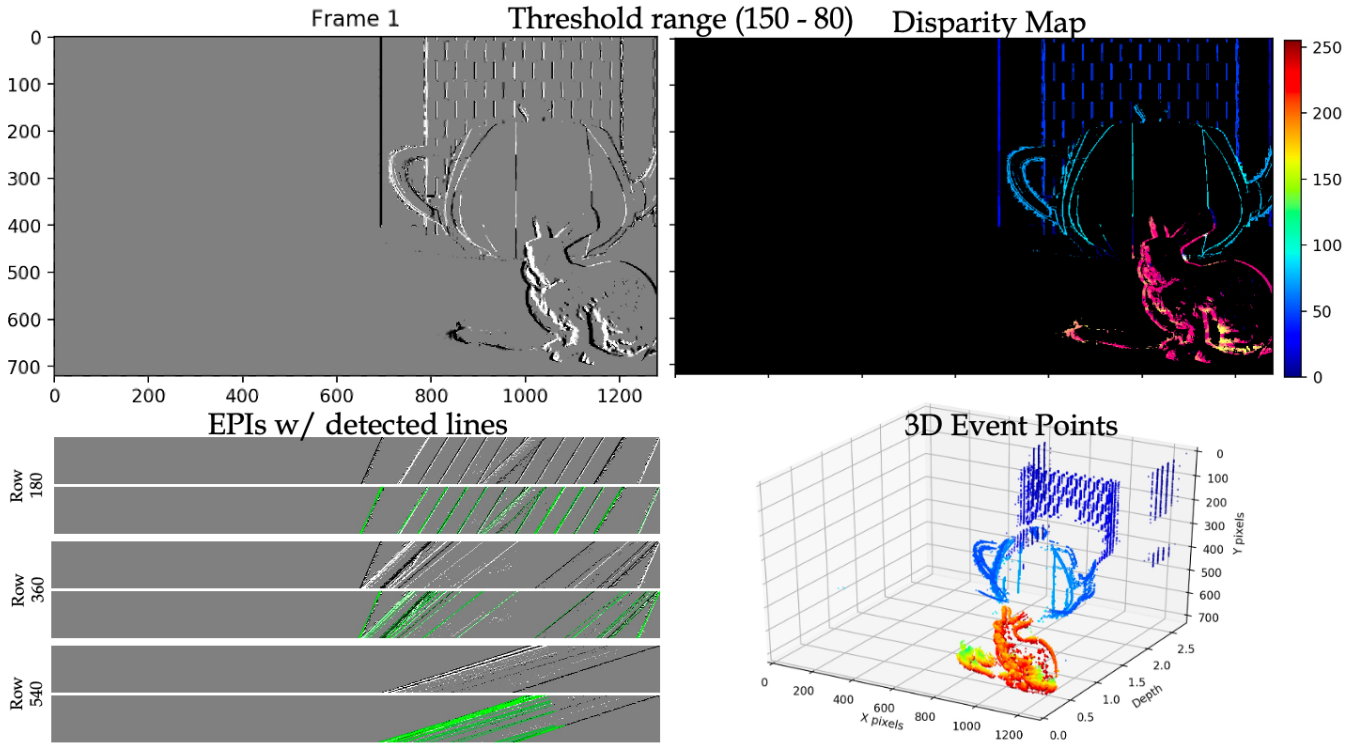

Fig. 3. EPI depth on a simulation scene.

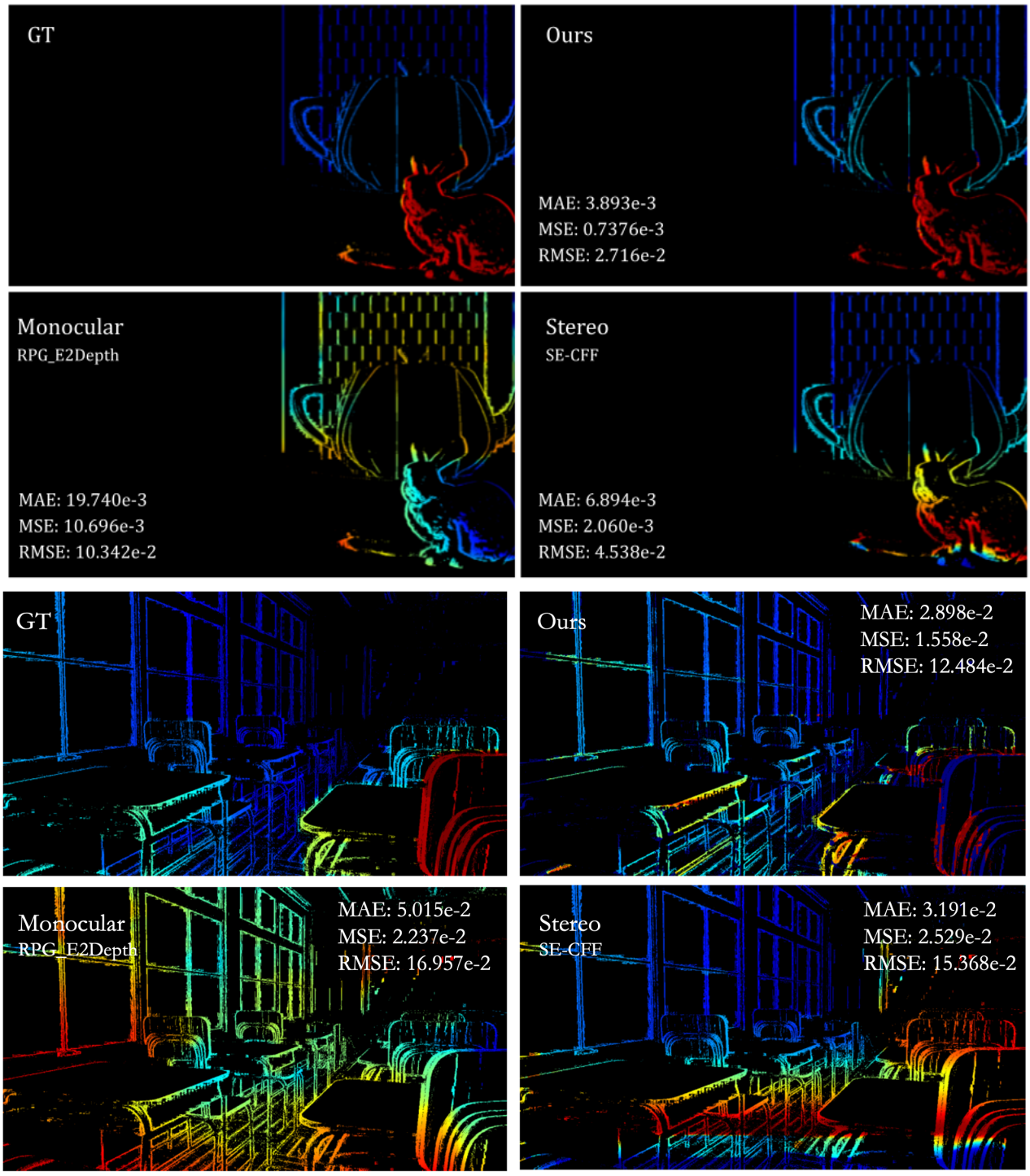

Fig. 4. Comparison with monocular and stereo depth estimation on simulated scenes

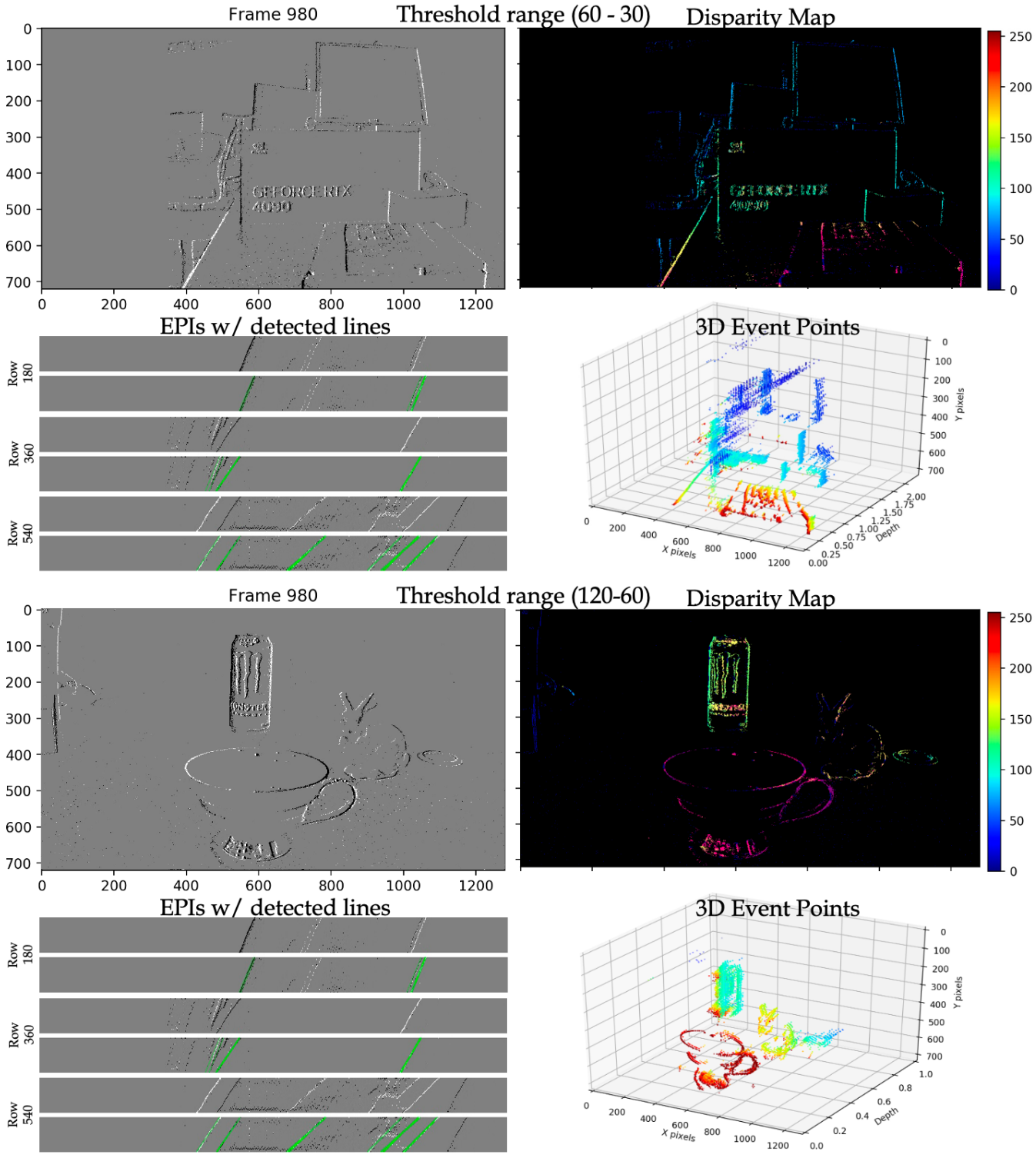

Fig. 5. EPI depth on real scenes