High-Quality Scanning Using Time-Of-Flight Depth Superresolution

Sebastian Schuon
Stanford University
schuon@cs.stanford.edu

Christian Theobalt
Stanford University
theobalt@cs.stanford.edu

James Davis
UC Santa Cruz
davis@cs.ucsc.edu

Sebastian Thrun
Stanford University
thrun@stanford.edu

Abstract

Time-of-flight (TOF) cameras robustly provide depth data of real world scenes at video frame rates. Unfortunately, currently available camera models provide rather low X-Y resolution. Also, their depth measurements are starkly influenced by random and systematic errors which renders them inappropriate for high-quality 3D scanning. In this paper we show that ideas from traditional color image superresolution can be applied to TOF cameras in order to obtain 3D data of higher X-Y resolution and less noise. We will also show that our approach, which works using depth images only, bears many advantages over alternative depth upsampling methods that combine information from separate high-resolution color and low-resolution depth data.

1. Introduction

Depth sensing is a core component of many machine vision systems. Among the technologies available, time-of-flight (TOF) based systems are attractive since they are real-time, robust, and rapidly becoming inexpensive. However their resolution is still limited. In this work, we address one of the main limitations of TOF sensors by showing that superresolution methods can be used to increase their effective resolution.

Time of flight cameras sense depth by emitting a pulse or modulated light signal and then measuring the time differential in the returning wavefront. This process is largely independent of the scene texture and full frame real-time depth estimates are possible. Unfortunately, the data is noticeably contaminated with random and systematic measurement errors. In addition the X-Y resolution of the sensors is often limited to 320x240 pixels or fewer, far below the resolution of modern cameras.

Prior researchers using TOF cameras have combined a high resolution RGB camera with a low resolution depth camera [2, 16]. Resolution is increased by assuming alignment of depth and intensity discontinuities in both views while smoothing elsewhere. These techniques work well when image features such as edges are collocated, but break down when this assumption of common scene statistics is violated. In this work we show that superresolution methods which rely *only* on the depth data perform better for these scenes.

Superresolution for traditional cameras has been well explored. Rather than reinvent these methods, we draw from the existing literature and show that it is applicable to depth cameras as well. Low resolution depth images are understood as degraded samples of a single high-resolution scene. A sequence of low resolution depth images is aligned and then merged to produce a single high quality result.

The primary contribution of this work is showing that high quality depth maps can be obtained from TOF cameras using multi-frame superresolution methods. In addition, we provide a comparison with color-fusion based superresolution, showing that multi-frame methods are superior when edge discontinuities are not collocated.

2. Related Work

Depth and color fusion: Depth image superresolution has primarily been accomplished by using a high resolution color image taken from the same location. The low resolution depth images are upsampled and regularized subject to an edge consistency term with respect to the color image. Regularization has taken the form of a MRF [2], bilateral filtering of the cost volume [16], and bilateral filtering in the image plane [8]. These methods can reproduce high frequency detail, however they incorrectly assume that color is correlated with depth. This causes difficulties with colored textures and when a true depth discontinuity is not visible in the color channel. Another approach was taken by Lindner et al. [10], who applied noise and edge aware upsampling. Using a pure upsampling method, they do not to recover details which are beyond the depth sensor's resolution limit.

Color superresolution: Image based superresolution targeted at standard color or intensity images has been well studied for many years [3][5][14]. Multiple low resolution images are aligned and then a high resolution image is estimated which explains the image stack. Interested readers will find a survey informative [1].

Some researchers have formulated a joint optimization

of superresolution together with shape-from-X. Shape from photometric cues [6] as well as defocus [12] have both been explored.

The noise and data statistics of depth data exhibit effects which may not be found in normal color images, so it is not obvious that color based methods are applicable. Indeed, earlier work targeted at depth superresolution pursued an alternate strategy. In this paper we show that color methods *are* applicable in the depth domain, and that they can perform better than the specialized depth superresolution methods previously introduced.

Improving TOF sensors: The depth accuracy of time-of-flight sensors can be increased by a variety of methods, e.g. by accounting for ambient light [4], simulating the shape of the reflected signal [7], and performing time gated superresolution [9]. While these methods improve resolution in the depth direction, they all operate at the level of peak detection in the sensor itself and are not directly related to improving resolution in the X-Y plane as discussed in this work.

3. The Depth Camera and its Characteristics

The Z-cam [15] used in our experiments exploits the time-of-flight principle of light to measure the distance of each pixel from the scene. The camera features a lens and a CMOS sensor, thus is based on video camera technology. However, it houses additional components, like a ring of infrared LEDs and a rapid controlled shutter, to enable depth rather than intensity measurement only. When capturing a single frame, the camera emits a single-pulse light wavefront from the LEDs into the scene. The returned pulse is "shaped" by the scene structure and this shape information can be extracted by gating the returned signal with the rapid shutter. After normalization, the measured intensity values can be interpreted as depth values. The Z-cam can measure full frame depth at video rate and at a resolution of 320×240 pixels. The control of the shutters also enables the definition of a 3D frustum in space in which depth measurements are taken.

In contrast to competing TOF cameras, the Z-cam features a normal video camera of 640×480 pixels in the same device which enables recording of texture-mapped geometry. Unfortunately, video and depth are not recorded through the same optics and the homographic registration of both data provided by the manufacturer can easily be several pixels off. In our comparison experiments we therefore resort to our own external color camera (Sect. 5).

Although the Z-cam delivers scene geometry at unprecedented speed and largely independently of scene texture, the quality of recovered 3D data in a single frame is not sufficient for high-quality 3D scanning, as shown in Fig. 2b. In this image three wall plugs should be visible, but are mostly masked by noise. The depth measurements are starkly contaminated by random noise which can, at $1\ m$ average scene distance, vary by up to $5\ cm$. Depth measurements also become more unreliable towards the boundary of the field of view, since there, optical aberrations like vignetting play a stronger role, and the PSNR of the returned signal naturally decreases. Fig. 1 shows the strongly increasing variance in random noise towards the field-of-view boundary. Noise variance is also much higher at mixed pixels that integrate over depth discontinuities in the scene. Fortunately, pixels with high measurement uncertainty typically exhibit low measurement intensity and therefore the camera's raw intensity data can be interpreted as a confidence map. Experimentally, we could verify that the depth readings at a single pixel location over time follow a slightly heavy-tailed distribution.

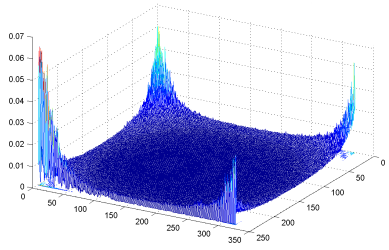


Figure 1: Variance distribution in a depth image taken at approx. $1.5\ m$ average distance from a scene. Depth images contain heavy noise near the corners.

In addition to random noise, the camera is likely to exhibit a systematic measurement bias that may depend on reflectance, angle of incidence, and environment factors like temperature and lighting. A detailed analysis of these consistent inaccuracies is beyond the scope of this paper. In our controlled lab setting, systematic errors played no significant role.

The Z-cam delivers ray-space depth maps, i.e. grayscale images that store at each pixel the distance along the ray from the center of projection to the point in the scene, Fig. 3a. For reconstructing metric 3D data, one has to unproject ray space measurements according to:

$$(X, Y, Z) = D \cdot \bar{V}\,. \tag{1}$$

Here, $V = \frac{(x,y,f)}{\sqrt{(x^2+y^2+f^2)}}$ is the measurement ray direction (viewing vector) from the camera's center of projection through the sensor pixel at location (x, y) relative to the sensor center, and f is the camera's focal length. For metric reconstruction, x and y have to be specified in terms of metric pixel size μ, i.e. $x = i_x \cdot \mu$ with i_x being the pixel index in x-direction relative to the pixel center. Further on, $D = P_d + P_w \frac{255-g}{255}$ is the depth along the measurement ray which is computed from the distance to the frontal clipping plane P_d, the depth of the 3D view frustum P_w, and the gray value g in the depth image which is quantized to eight bit.

4. Depth Superresolution

It is our goal to obtain high-quality 3D measurements of a static scene despite the significant noise in the raw data. By performing superresolution, we increase X-Y measurement resolution and, at the same time, reduce the overall random noise level. To this end, several depth maps captured from minimally displaced viewpoints are aligned, and subsequently combined into a higher resolution depth image. From this superresolved depth image, we can eventually reconstruct superresolved 3D geometry.

4.1. Setup

In our measurement setup, the depth camera is located between $50\ cm$ and $150\ cm$ away from the scene. Typically, we capture $N = 15$ images by slightly translating the camera orthogonally to the viewing direction. Please note that the alignment of images captured by the above procedure effectively leads to the creation of a multi-perspective image in which parallax effects may play a role. One way to overcome these effects would be to slightly rotate the camera around the center of projection rather than translate it. However, with as small displacements as we apply them we could experimentally not verify an increase in reconstruction quality if the camera is rotated. Therefore, we always record with translational offsets.

From the first to the last frame of a superresolution sequence, the camera is, in total, displaced by around $1\ cm$ to $1.5\ cm$. In order to cancel out random noise, we average 30 depth measurements at each camera position.

4.2. Extracting High Resolution 3D Data

By appropriately combining the low resolution depth images $\mathbf{Y_k}$, $k = 1, \ldots, N$ taken from slightly displaced viewpoints, we can create new depth maps at significantly higher resolution. Using Eq. (1), the upsampled depth maps can then be converted to high resolution 3D geometry. Our depth superresolution method is based on the approach by Farsiu et al. [3] who investigated superresolution for normal photographs.

We cast superresolution as the problem of inverting the formation process of low resolution depth images of a high resolution 3D scene. To formulate the problem, we make the simplifying assumption that the formation process of a depth image can be described in analogy to the image formation process of a normal optical camera. However, the quality of our final results shows that this simplification is valid. For a single depth image $\mathbf{Y_k}$, the formation process therefore looks as follows:

$$\mathbf{Y_k} = D_k H_k F_k \mathbf{X} + \mathbf{V_k}\,,$$

where $\mathbf{X}$ is the original scene or, in other words, the superresolved image of the 3D scene from which we sample. Henceforth, we will refer to the upsampling factor between low and high resolution images in x- and y-direction as β. F_k is a translation operator representing the motion between the superresolution image and the current low resolution image. In our setting, we assume pure translational motion. H_k is a blur operator accounting for the blur introduced during the capture process (i.e. due to the optic system or motion). In our experiments we assumed no blur, hence H_k was equivalent to the unity matrix. D_k is a decimation operator modeling the downsampling from the superresolution image to the size of the low resolution image. Finally $\mathbf{V_k}$ represents additive noise inherited during the capture process. To extract the high resolution image from the set of low resolution depth maps, we need to solve the following minimization problem:

$$\widehat{\mathbf{X}} = \underset{\mathbf{X}}{argmin} \left[\sum_{k=1}^{N} \|D_k H_k F_k \mathbf{X} - \mathbf{Y_k}\|_p^p \right], \qquad (2)$$

where [3] readily argues that $p = 1$ gives optimal results in terms of robust statistics. Since with a typical set of images this estimation problem is ill-posed, one is to add a regularization term $\Upsilon(\mathbf{X})$ with weight λ yielding

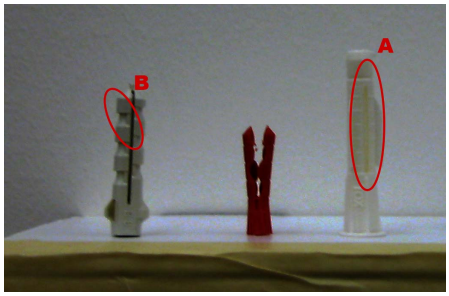
$$\widehat{\mathbf{X}} = \underset{\mathbf{X}}{argmin} \left[\sum_{k=1}^{N} \|D_k H_k F_k \mathbf{X} - \mathbf{Y_k}\|_p^p + \lambda \Upsilon(\mathbf{X}) \right] \quad (3)$$

Different regularization terms such as Tikhonov regularization or Total Variation could be imagined. For this paper, we used bilateral regularization. This robust technique, also referred to as bilateral filtering, has the advantage of preserving edges and removing random noise in areas of slowly varying depth. Also, the computation of the regularizer is relatively cheap. The bilateral regularization is given by

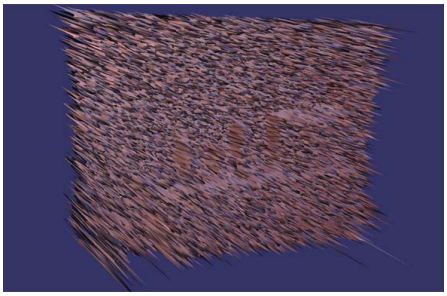
$$\Upsilon(\mathbf{X})_B = \underbrace{\sum_{l=-P}^{P} \sum_{m=0}^{P}}_{l+m>=0} \alpha^{|m|+|l|} \left\| \mathbf{X} - S_x^l S_y^m \mathbf{X} \right\|_1$$

here S_x^l and S_y^m are shift operators that perform a shift in x or y direction by l or respectively m pixels. The scalar weight α, with $0 < \alpha < 1$, controls the spatial influence area of the bilateral constraint, $P \geq 1$ specifies the size of the neighborhood used for bilateral filtering. Please refer to [3] to learn about the equivalence of the above formulation to the original bilateral filter proposed in [13]. The robust bilateral formulation in Eq. (3) is preferable over quadratic penalization since the latter would perform worse in the presence of the heavy-tailed random noise in the raw depth data, Sect. 3.

Solving the optimization problem in Eq. (3) yields a superresolved depth image of the scene. In practice, we employ the solver implementation provided by Milanfar [11] to



(a) Color recording in high resolution



(b) 3D model of single raw frame

Figure 2: Wall plug scene - details on the plugs, clearly visible in the color image, are entirely masked by random noise in an unsmoothed 3D rendering of a single depth image.

compute the solution. From the superresolution depth image, we reconstruct 3D geometry by means of Eq. (1). Prior to 3D reconstruction, we median filter the superresolution depth image with a kernel size of 3×3. Please remember that the effective metric pixel size in the high-resolution image is μ/β.

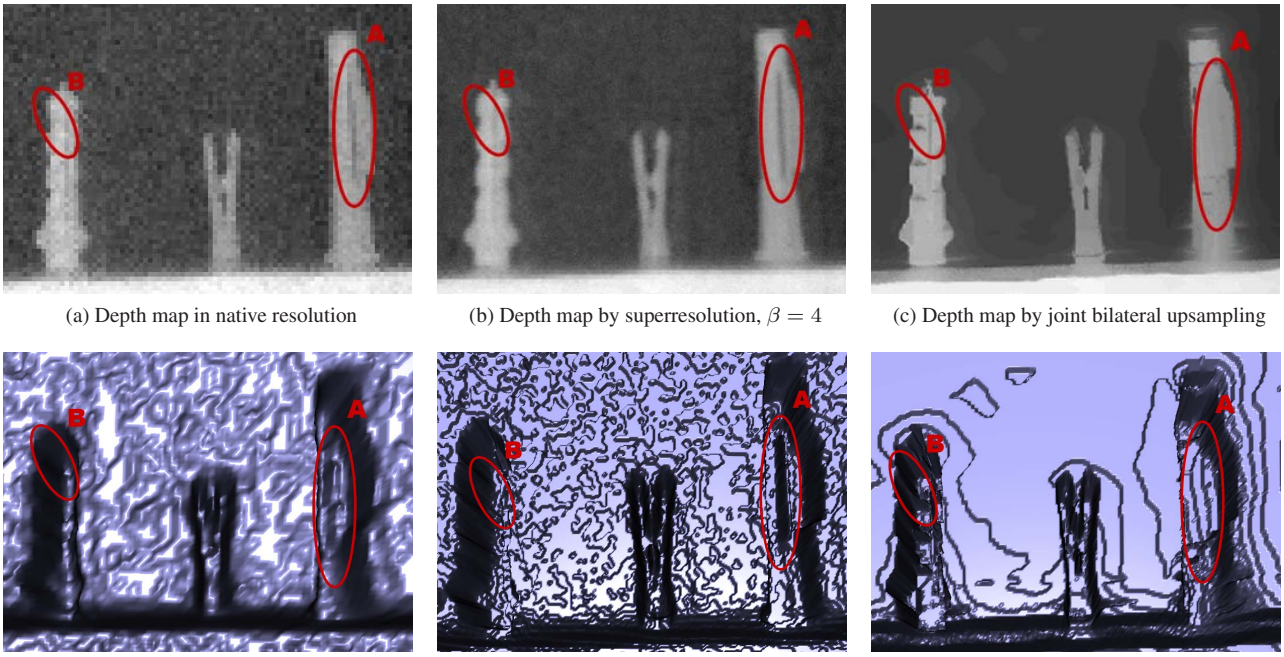
5. Results and Discussion

We have tested our approach on three different scenes, all of which show geometric detail that is close to the X-Y resolution limit of the depth camera in one frame. The test scenes also feature areas that contradict the assumption color and depth discontinuities are well-aligned, which allows us to show that methods relying on this simple prior statistics will perform worse.

Resolving thin structure: We wanted to verify that our superresolution method can resolve thin structures. Therefore our first setup shows three wall plugs in front of a white wall, Fig. 2a. The scene is approx. $50\ cm$ away from the camera, and was recorded from 15 displaced positions to perform superresolution. For this scene, the camera was configured to record objects from $0\ cm$ up to $100\ cm$ away. To illustrate the performance of our method, we focus on a dent and a long thin gap in the wall plugs which are marked as A and B, respectively, in Fig. 2a. Since these features are close to the resolution limit of the Z-cam, they do not appear well in a single depth image, Fig. 3a, and consequently also not in the corresponding low resolution 3D reconstruction, Fig. 3d. In contrast, our 4-times superresolved result accurately captures these details, as visible in the depth image Fig. 3b, and in geometry Fig. 3e where they appear as true 3D structure with correct depth. To display the 3D geometry we convert the depth maps into triangulated height fields and render them using basic Phong shading. Please note that for fair comparison we always perform superresolution at 8-bit depth precision in all tested methods, as this is the limit of the software by Milanfar et al. [11]. Therefore, discretization artifacts in the form of depth steps are visible in the renderings. To verify that our 3D reconstructions do not suffer from incorrect scaling or distortion we compared the size of several landmarks in our results to their real-world size. In all cases, this comparison showed an exact match which proves the reliability of our algorithm.

For comparison, we implemented a joint bilateral upsampling (JBU) approach [8], which uses a high-resolution color and a low resolution depth image to raise the depth resolution to the one of the color image. The color image was recorded using a standard digital camera and has been manually aligned using a homographic warp. By inspection the alignment error was determined to zero pixels for most pixels, while three pixels being the maximum error. The method's implicit assumption that color and depth edges are collocated is frequently violated in our wall plug scene causing erroneous reconstructions. Although the depth map, Fig. 3c, shows crisp edges which is visually pleasing if only the gray scale image is looked at, the actual reconstruction exhibits several errors. For instance, the method wrongly reconstructs the shadowed area *B* on the ripple of the left wall plug, Fig. 2a, as a depth discontinuity that protrudes all the way through the scene Fig. 3f. Also, joint bilateral upsampling performs excessive smoothing in areas with low image gradient. Therefore, the dent in area *A* on the right wall plug, whose edges are not clear in the color image, is entirely smoothed out. Also, shadows on the back of the table appear as geometry merged to the lower part of the plugs, and the top of the right plug is cut off due color similarity to the background. We thus conclude that a slightly higher remaining level of noise, as in in our results, is preferable over such excessive smoothing since in the latter case actual shape detail is lost or incorrectly estimated.

Preserving sharp edges: Another important characteristic of superresolution is to preserve sharp edges. Hence, a second scene, with a planar checkerboard spaced approx. $50\ cm$ from a white background, was recorded to prove that our method correctly captures both sharp edges and smooth regions, Fig. 5a. In contrast, the joint bilateral upsampling method runs into difficulties in the presence of strong texture on actually planar geometry. Here the camera was con-



(a) Depth map in native resolution (b) Depth map by superresolution, $\beta = 4$ (c) Depth map by joint bilateral upsampling

(d) 3D model from native resolution (e) 3D model from superresolution, $\beta = 4$ (f) 3D model from joint bilateral upsampling

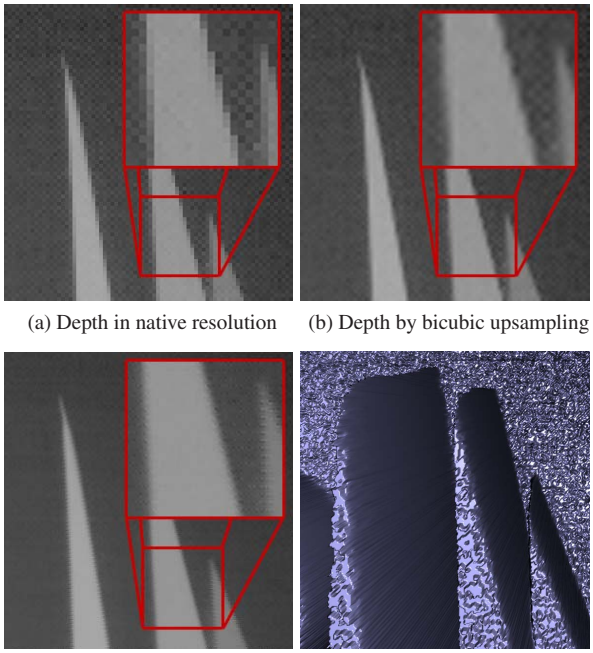
Figure 3: Wall plug scene - superresolution (b),(e) unveils fine details, previously not visible in native resolution (a),(d). Joint bilateral upsampling (c),(f) sharpens the image, but introduces false geometry. For better visibility the contrast of depth maps was enhanced.

figured for recording between 70 cm and 200 cm. The board features a color pattern with strong intensity gradients. The pattern is slightly smaller than the actual size of the board, which has a 1 cm white boundary that is visually indistinguishable from the white background. In Fig. 5a, we marked the location of the actual depth edge with lines. The low resolution depth image (Fig. 5d) has an apparent staircase effect on the edge, while the edge appears sharp and crisp in the depth map created by the proposed superresolution method (Fig. 5e). The joint bilateral upsampling method is tricked by the non-collocation of the intensity gradient (black pattern boundary) and true depth discontinuity. Consequently, the true depth edge is smoothed with the background leading to a blurred edge in the JBU depth image, Fig. 5f. This effect can be studied best in 3D. While our superresolved geometry, Fig. 5h, shows a sharp edge with sharped depth discontinuity, the edge of the joint bilateral upsampling result is incorrectly shaped like a curved ramp, Fig. 5i. The rendering of the depth edges in a cross-sectional views, Fig. 5j-5l, makes this effect even more apparent. Or result shows a sharp corner and a straight depth edge, Fig. 5j, whereas the JBU result is erroneously curved, Fig. 5l. Another problematic region for joint bilateral upsampling is the surface of the checker board itself. Whereas it appears up to noise as a plain, the color gradients in the checker board provoke the bilateral filter to emboss this structure into the geometry (Fig. 5c). In contrast, our upsampling result shows a planar board, Fig. 5b.

Gain in resolution: To further demonstrate the true gain in resolution, we recorded three planar triangular wedges 30 cm in front of a flat wall. They exhibit clear sharp depth edges and, close to the tips, fall below the resolution limit of the camera. The recording settings were $P_d = 50$ cm and $P_w = 100$ cm. While the depth map at original camera resolution exhibits strong staircase aliasing at the boundaries, Fig. 4a, our 4-times upsampled result faithfully captures crisp depth edges, Fig. 4c. Consequently, the upsampled 3D geometry also shows sharp edges, Fig. 4d. Simple bicubic upsampling of the low resolution data cannot produce the same superresolution effect. It mainly upsamples the staircase pattern and boosts the random noise, Fig. 4b.

Our method is subject to a few limitations. Since several depth images have to be combined it is, in contrast to joint bilateral upsampling, only suitable for static scenes. Also, given a runtime of approximately one minute to compute a superresolved depth map, our approach is not suitable for real-time applications. Furthermore our approach relies on faithful image registration which may be difficult in scenes with few distinct depth discontinuities. In the future, we plan to capitalize on noise characteristics and known measurement uncertainty, from which we expect improved superresolution quality.

We will also perform a more detailed analysis of the range of achievable upsampling factors in dependence on



(a) Depth in native resolution (b) Depth by bicubic upsampling

(c) Depth by superresolution (d) 3D model from superresolution

Figure 4: Wedge scene - superresolution ($\beta = 4$) achieves true resolution enhancement and shows straight alias-free edges at depth boundaries (c),(d). In contrast, staircasing artifacts are clearly visible at native resolution (a) and in the bicubic upsampled result (b). Additionally noise is significantly reduced by superresolution.

scene structure and recording conditions. Currently, we did tests with β in the range of $2 - 6$. Overall, we found that, in our test scenes, $\beta = 4$ provides the best compromise between extracted shape detail and model size.

We would also like to remark that both tested superresolution methods rely on a bilateral constraint of some form. It is not the constraint itself that makes one method preferable over the other, but the particular way how it is enforced. Joint bilateral upsampling enforces the constraint in two different data domains, namely color and depth, and implicitly relies on the wrong prior. In contrast, we enforce the constraint on depth data only and do not enforce the same excessive smoothing as the former approach which renders advantageous in our setting.

In summary, we have demonstrated that the concepts of color superresolution can be used to greatly improve 3D reconstruction quality of static scenes.

6. Conclusion

In this work we have shown that superresolution methods that were originally developed for color images can be applied to capture higher resolution 3D geometry with a time-of-flight depth camera. We have also shown that a proper formulation of superresolution only in terms of depth images frequently outperforms previous algorithms from the literature that combine information from aligned color and depth. Overall, the proposed superresolution strategy reliably increases the X-Y resolution of captured 3D geometry. Since it also severely reduces the noise level in the data, it turns the TOF camera into a viable tool for 3D shape scanning.

References

[1] S. Borman and R. L. Stevenson. Super-resolution from image sequences - a review. *Proc. Midwest Symp. Circuits and Systems*, 5, 1998.

[2] J. Diebel and S. Thrun. An application of markov random fields to range sensing. In *Advances in Neural Information Processing Systems 18*, pages 291–298. MIT Press, 2006.

[3] S. Farsiu, M. Robinson, M. Elad, and P. Milanfar. Fast and robust multiframe super resolution. *IEEE Transactions on Image Processing*, 13(10):1327–1344, Oct. 2004.

[4] H. Gonzalez-Banos and J. Davis. Computing depth under ambient illumination using multi-shuttered light. *IEEE CVPR*, 2, 2004.

[5] M. Irani and S. Peleg. Improving resolution by image registration. *CVGIP: Graph. Models Image Process.*, 53(3):231–239, 1991.

[6] M. V. Joshi and S. Chaudhuri. Simultaneous estimation of super-resolved depth map and intensity field using photometric cue. *CVIU*, 101(1):31–44, 2006.

[7] B. Jutzi and U. Stilla. Precise range estimation on known surfaces by analysis of full-waveform laser. *Proceedings of Phtogrammetric Computer Vision PCV*, 2006.

[8] J. Kopf, M. Cohen, D. Lischinski, and M. Uyttendaele. Joint bilateral upsampling. *ACM TOG*, 26(3), 2007.

[9] M. Laurenzis, F. Christnacher, and D. Monnin. Long-range three-dimensional active imaging with superresolution depth mapping. *Opt. Lett.*, 32(21):3146–3148, 2007.

[10] M. Lindner, M. Lambers, and A. Kolb. Data Fusion and Edge-Enhanced Distance Refinement for 2D RGB and 3D Range Images. *IJISTA, Issue on Dynamic 3D Imaging*, 2008.

[11] P. Milanfar. MDSP resolution enhancement software. http://soe.ucsc.edu/˜milanfar/software/superresolution.html, 2004.

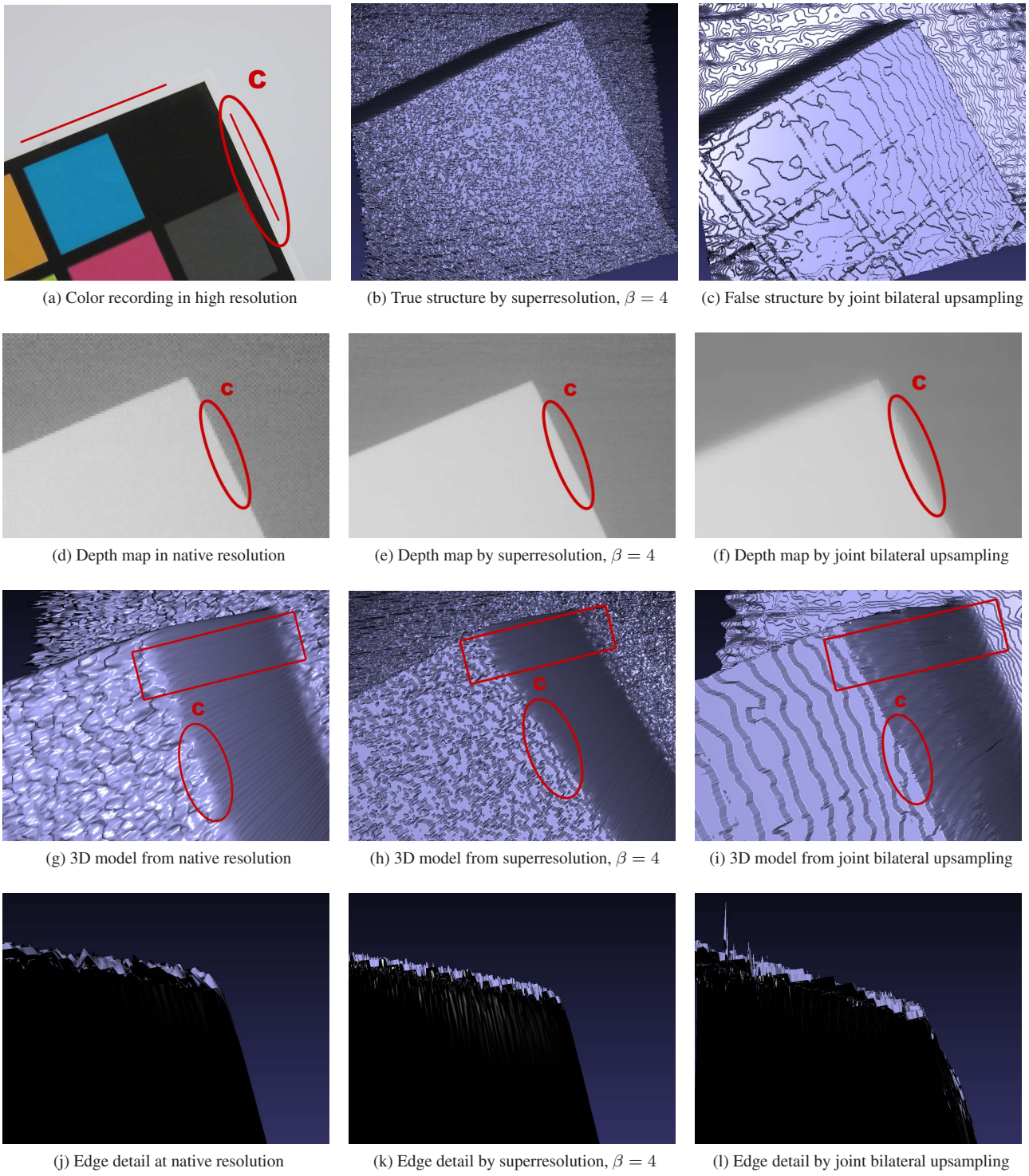
[12] D. Rajan and S. Chaudhuri. Simultaneous estimation of super-resolved scene and depth map from low resolution defocused observations. *PAMI*, 25(9):1102–1117, 2003.

[13] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *ICCV*, pages 839–846, 1998.

[14] R. Tsai and T. Huang. Multiframe image restoration and registration. *Advances in Computer Vision and Image Processing*, pages 317–339, 1984.

[15] G. Yahav, G. Iddan, and D. Mandelboum. 3d imaging camera for gaming application. In *Consumer Electronics 2007*, pages 1–2, 2007.

[16] Q. Yang, R. Yang, J. Davis, and D. Nistér. Spatial-depth super resolution for range images. In *IEEE Computer Vision and Pattern Recognition*. IEEE Computer Society, 2007.



(a) Color recording in high resolution

(b) True structure by superresolution, $\beta = 4$

(c) False structure by joint bilateral upsampling

(d) Depth map in native resolution

(e) Depth map by superresolution, $\beta = 4$

(f) Depth map by joint bilateral upsampling

(g) 3D model from native resolution

(h) 3D model from superresolution, $\beta = 4$

(i) 3D model from joint bilateral upsampling

(j) Edge detail at native resolution

(k) Edge detail by superresolution, $\beta = 4$

(l) Edge detail by joint bilateral upsampling

Figure 5: Board scene - The upper row shows that ”phantom” geometry is introduced by joint bilateral upsampling (b), whereas superresolution retains the true geometry (c). This effect is also visible in the depth maps one row below. The two lower rows show sharp edges being preserved by superresolution, while joint bilateral upsampling yields round edges.