Patch Based Synthesis for Single Depth Image Super-Resolution

Oisin Mac Aodha, Neill D.F. Campbell, Arun Nair, and Gabriel J. Brostow

University College London
http://visual.cs.ucl.ac.uk/pubs/depthSuperRes/

Abstract. We present an algorithm to synthetically increase the resolution of a solitary depth image using only a generic database of local patches. Modern range sensors measure depths with non-Gaussian noise and at lower starting resolutions than typical visible-light cameras. While patch based approaches for upsampling intensity images continue to improve, this is the first exploration of patching for depth images.

We match against the height field of each *low resolution* input depth patch, and search our database for a list of appropriate high resolution candidate patches. Selecting the right candidate at each location in the depth image is then posed as a Markov random field labeling problem. Our experiments also show how important further depth-specific processing, such as noise removal and correct patch normalization, dramatically improves our results. Perhaps surprisingly, even better results are achieved on a variety of real test scenes by providing our algorithm with only *synthetic* training depth data.

1 Introduction

Widespread 3D imaging hardware is advancing the capture of depth images with either better accuracy, *e.g.* Faro $Focus^{3D}$ laser scanner, or at lower prices, *e.g.* Microsoft's *Kinect*. For every such technology, there is a natural upper limit on the spatial resolution and the precision of each depth sample. It may seem that calculating useful interpolated depth values requires additional data from the scene itself, such as a high resolution intensity image [1], or additional depth images from nearby camera locations [2]. However, the seminal work of Freeman *et al.* [3] showed that it is possible to explain and super-resolve an *intensity* image, having previously learned the relationships between blurry and high resolution image patches. To our knowledge, we are the first to explore a patch based paradigm for the super-resolution (SR) of single depth images.

Depth image SR is different from image SR. While less affected by scene lighting and surface texture, noisy depth images have fewer good cues for matching patches to a database. Also, blurry edges are perceptually tolerable and expected in images, but at discontinuities in depth images they create jarring artifacts (Fig. 1). We cope with both these problems by taking the unusual step of matching inputs against a database at the *low* resolution, in contrast to using interpolated high resolution. Even creation of the database is also harder

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part III, LNCS 7574, pp. 71–84, 2012.

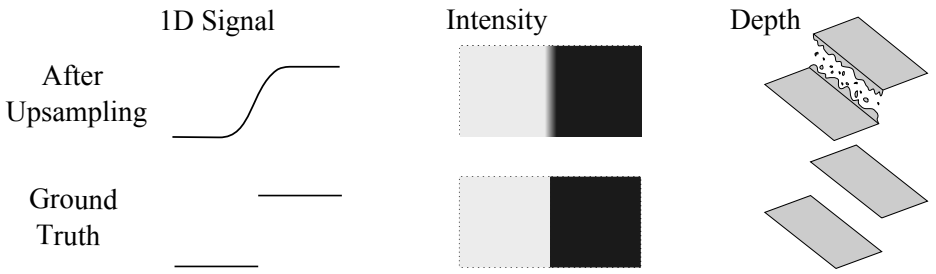

Fig. 1. On the bottom row are three different, high resolution, signals that we wish to recover. The top row illustrates typical results after upsampling its low resolution version by interpolation. Interpolation at intensity discontinuities gives results that are perceptually similar to the ground truth image. However in depth images, this blurring can result in very noticeable jagged artifacts when viewed in 3D.

for depth images, whether from Time-of-Flight (ToF) arrays or laser scanners, because these can contain abundant interpolation-like noise.

The proposed algorithm infers a high resolution depth image from a *single* low resolution depth image, given a generic database of training patches. The problem itself is novel, and we achieve results that are qualitatively superior to what was possible with previous algorithms because we:

- Perform patch matching at the low resolution, instead of interpolating first.
- Train on a synthetic dataset instead of using available laser range data.
- Perform depth specific normalization of non-overlapping patches.
- Introduce a simple noisy-depth reduction algorithm for postprocessing.

Depth cameras are increasingly used for video-based rendering [4], robot manipulation [5], and gaming [6]. These environments are dynamic, so a general purpose SR algorithm is better if it does not depend on multiple exposures. These scenes contain significant depth variations, so registration of the 3D data to a nearby camera's high resolution intensity image is approximate [7], and use of a beam splitter is not currently practicable. In the interest of creating visually plausible super-resolved outputs under these constraints, we relegate the need for the results to genuinely match the real 3D scene.

2 Related Work

Both the various problem formulations for super-resolving *depth* images and the successive solutions for super-resolving *intensity* images relate to our algorithm. Most generally, the simplest upsampling techniques use nearest-neighbor, bilinear, or bicubic interpolation to determine image values at interpolated coordinates of the input domain. Such increases in resolution occur without regard for the input's frequency content. As a result, nearest-neighbor interpolation

turns curved surfaces into jagged steps, while bilinear and bicubic interpolation smooth out sharp boundaries. Such artifacts can be hard to measure numerically, but are perceptually quite obvious both in intensity and depth images. While Fattal [8] imposed strong priors based on edge statistics to smooth "stair step" edges, this type of approach still struggles in areas of texture. Methods like [9] for producing high quality antialiased edges from jagged input are inappropriate here, for reasons illustrated in Fig. 1.

Multiple Depth Images: The SR problem traditionally centers on fusing multiple low resolution observations together, to reconstruct a higher resolution image, *e.g.* [10]. Schuon *et al.* [2] combine multiple (usually 15) low resolution depth images with different camera centers in an optimization framework that is designed to be robust to the random noise characteristics of ToF sensors. To mitigate the noise in each individual depth image, [1] composites together multiple depths from the same viewpoint to make a "single" depth image for further super-resolving, and Hahne and Alexa [11] combine depth scans in a manner similar to exposure-bracketing for High Dynamic Range photography. Rajagopalan *et al.* [12] use an MRF formulation to fuse together several low resolution depth images to create a final higher resolution image. Using GPU acceleration, Izadi *et al.* made a system which registers and merges multiple depth images of a scene in real time [13]. Fusing multiple sets of noisy scans has also been demonstrated for effective scanning of individual 3D shapes [14]. Compared to our approach, these all assume that the scene remains static. Though somewhat robust to small movements, large scene motion will cause them to fail.

Intensity Image Approaches: For intensity images, learning based methods exist for SR when multiple frames or static scenes are not available. In the most closely related work to our own, Freeman *et al.* [15,16] formulated the problem as multi-class labeling on an MRF. The label being optimized at each node represents a high resolution patch. The unary term measures how closely the high resolution patch matches the interpolated low resolution input patch. The pairwise terms encourage regions of the high resolution patches to agree. In their work, the high resolution patches came from an external database of photographs. To deal with depth images, our algorithm differs substantially from Freeman *et al.* and its image SR descendants, with details in Section 3. Briefly, our depth specific considerations mean that we i) compute matches at low resolution to limit blurring and to reduce the dimensionality of the search, ii) model the output space using non-overlapping patches so depth values are not averaged, iii) normalize height to exploit the redundancy in depth patches, and iv) we introduce a noise-removal algorithm for postprocessing, though qualitatively superior results emerge before this step.

Yang *et al.* [17] were able to reconstruct high resolution test patches as sparse linear combinations of atoms from a learned compact dictionary of paired high/low resolution training patches. Our initial attempts were also based on sparse coding, but ultimately produced blurry results in the reconstruction stage. Various work has been conducted to best take advantage of the statistics of natural image patches. Zontak and Irani [18] argue that finding suitable high

resolution matches for a patch with unique high frequency content could take a prohibitively large external database, and it is more likely to find matches for these patches within the same image. Glasner *et al.* [19] exploit patch repetition across and within scales of the low resolution input to find candidates. In contrast to depth images, their input contains little to no noise, which can not be said of external databases which are constrained to contain the same "content" as the input image. Sun *et al.* [20] oversegment their input intensity image into regions of assumed similar texture and lookup an external database using descriptors computed from the regions. HaCohen *et al.* [21] attempt to classify each region as a discrete texture type to help upsample the image. A similar approach can be used for adding detail to 3D geometry; object specific knowledge has been shown to help when synthesizing detail on models [22,23]. While some work has been carried out on the statistics of depth images [24], it is not clear if they follow those of regular images. Major differences between the intensity and depth SR problems are that depth images usually have much lower starting resolution and significant non-Gaussian noise. The lack of high quality data also means that techniques used to exploit patch redundancy are less applicable.

Depth+Intensity Hybrids: Several methods exploit the statistical relationship between a high resolution intensity image and a low resolution depth image. They rely on the co-occurrence of depth and intensity discontinuities, on depth smoothness in areas of low texture, and careful registration for the object of interest. Diebel and Thrun [25] used an MRF to fuse the two data sources after registration. Yang *et al.* [1] presented a very effective method to upsample depth based on a cross bilateral filter of the intensity. Park *et al.* [7] improved on these results with better image alignment, outlier detection, and also by allowing for user interaction to refine the depth. Incorrect depth estimates can come about if texture from the intensity image propagates into regions of smooth depth. Chan *et al.* [26] attempted to overcome this by not copying texture into depth regions which are corrupted by noise and likely to be geometrically smooth. Schuon *et al.* [27] showed that there are situations where depth and color images can not be aligned well, and that these cases are better off being super-resolved just from multiple depth images.

In our proposed algorithm, we limit ourselves to super-resolving a single low resolution depth image, without additional frames or intensity images, and therefore no major concerns about baselines, registration, and synchronization.

3 Method

We take, as an input, a low resolution depth image $\mathbf{X}$ that is generated from some unknown high resolution depth image $\mathbf{Y}^*$ by an unknown downsampling function $\downarrow_{d*}$ such that $\mathbf{X} = (\mathbf{Y}^*) \downarrow_{d*}$. Our goal is to synthesize a plausible $\mathbf{Y}$. We treat $\mathbf{X}$ as a collection of N non-overlapping patches $\mathbf{X} = \{\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}\}$, of size $M \times M$, that we scale to fit in the range [0..1], to produce normalized input patches $\hat{\mathbf{x}}_i$. We recover a plausible SR depth image $\mathbf{Y}$ by finding a minimum of a discrete energy function. Each node in our graphical model (see Fig. 3) is

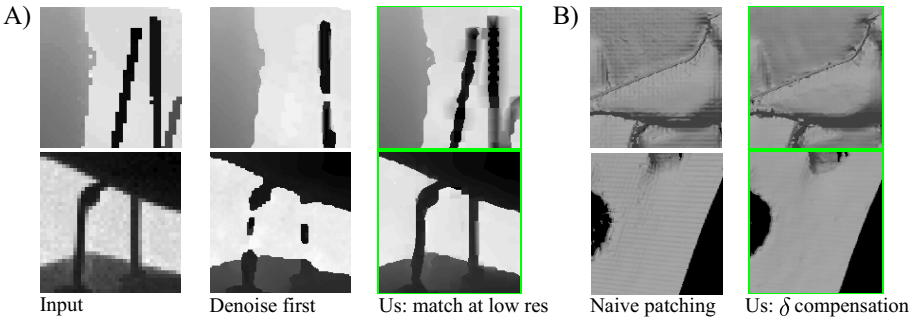

Fig. 2. A) Noise removal of the input gives a cleaner input for patching but can remove important details when compared to our method of matching at low resolution. B) Patching artifacts are obvious (left) unless we apply our patch min/max compensation (right).

associated with a low resolution image patch, $\hat{\mathbf{x}}_i$, and the discrete label for the corresponding node in a Markovian grid corresponds to a high resolution patch, $\mathbf{y}_i$. The total energy of this MRF is

$$E(\mathbf{Y}) = \sum_i E_{\mathrm{d}}(\hat{\mathbf{x}}_i) + \lambda \sum_{i,j \in \mathcal{N}} E_{\mathrm{s}}(\mathbf{y}_i, \mathbf{y}_j), \tag{1}$$

where $\mathcal{N}$ denotes the set of neighboring patches.

The data likelihood term, $E_{\mathrm{d}}(\hat{\mathbf{x}}_i)$, measures the difference between the normalized input patch and the normalized downsampled high resolution candidate:

$$E_{\mathrm{d}}(\hat{\mathbf{x}}_i) = ||\hat{\mathbf{x}}_i - (\hat{\mathbf{y}}_i) \downarrow_d ||_2. \tag{2}$$

Unlike Freeman *et al.* [3], we do not upsample the low resolution input using a deterministic interpolation method to then compute matches at the upsampled scale. We found that doing so unnecessarily accentuates the large amount of noise that can be present in depth images. Noise removal is only partially successful and runs the risk of removing details, see Fig. 2 A). A larger amount of training data is also needed to explain the high resolution patches and increases the size of the MRF. Instead, we prefilter and downsample the high resolution training patches to make them the same size and comparable to input patches.

The pairwise term, $E_{\mathrm{s}}(\mathbf{y}_i, \mathbf{y}_j)$, enforces coherence in the abutting region between the neighboring unnormalized high resolution candidates, so

$$E_{\mathrm{s}}(\mathbf{y}_i, \mathbf{y}_j) = ||O_{ij}(\mathbf{y}_i) - O_{ji}(\mathbf{y}_j)||_2, \tag{3}$$

where O_{ij} is an overlap operator that extracts the region of overlap between the extended versions of the unnormalized patches $\mathbf{y}_i$ and $\mathbf{y}_j$, as illustrated in Fig. 3 A). The overlap region consists of a single pixel border around each patch. We place the non-extended patches down side by side and compute the pairwise

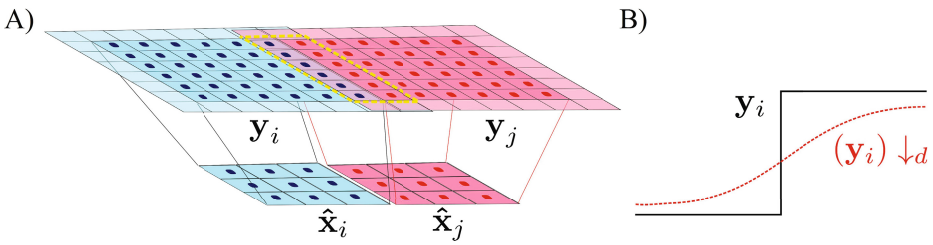

Fig. 3. A) Candidate high resolution patches $\mathbf{y}_i$ and $\mathbf{y}_j$ are placed beside each other but are not overlapping. Each has an additional one pixel border used to evaluate smoothness (see Eqn. 3), which is not placed in the final high resolution depth image. Here, the overlap is the region of 12 pixels in the rectangle enclosed by yellow dashed lines. B) When downsampling a signal its absolute min and max values will not necessarily remain the same. We compensate for this when unnormalizing a patch by accounting for the difference between the patch and its downsampled version.

term in the overlap region. In standard image SR this overlap region is typically averaged to produce the final image [3], but with depth images this can create artifacts.

The high resolution candidate, $\hat{\mathbf{y}}_i$, is unnormalized based on the min and max of the input patch:

$$\mathbf{y}_i = \hat{\mathbf{y}}_i(\max(\mathbf{x}_i)\delta_i^{\max} - \min(\mathbf{x}_i)\delta_i^{\min}) + \min(\mathbf{x}_i)\delta_i^{\min}, \tag{4}$$

where the δ_i terms account for the differences accrued during the downsampling of the training data (see Fig. 3 B)):

$$\delta_i^{\min} = \min(\mathbf{y}_i)/\min((\mathbf{y}_i)\downarrow_d),$$
$$\delta_i^{\max} = \max(\mathbf{y}_i)/\max((\mathbf{y}_i)\downarrow_d).$$

The exclusion of the δ_i terms results in noticeable patching artifacts, such as stair stepping and misalignment in the output depth image, due to the high resolution patch being placed into the output with the incorrect scaling (see Fig. 2 B)). We experimented with different normalization techniques, such as matching the mean and variance, but, due to the non-Gaussian distribution of depth errors [28], noticeable artifacts were produced in the upsampled image. To super-resolve our input, we solve the discrete energy minimization objective function of (1) using the TRW-S algorithm [29,30].

3.1 Depth Image Noise

Depending on the sensor used, depth images can contain a considerable amount of noise. Work has been undertaken to try to characterize the noise of ToF sensors [31]. They exhibit phenomena such as flying pixels at depth discontinuities due to the averaging of different surfaces, and return incorrect depth readings

from specular and dark materials [28]. Coupled with low recording resolution (compared to intensity cameras), this noise poses an additional challenge that is not present when super-resolving an image. We could attempt to model the downsampling function, $\downarrow_{d*}$, which takes a clean noiseless signal and distorts it. However, this is a non trivial task and would result in a method very specific to the sensor type. Instead, we work with the assumption that, for ToF sensors, most of the high frequency content is noise. To remove this noise, we bilateral filter [32] the input image $\mathbf{X}$ before the patches are normalized. The high resolution training patches are also filtered (where $\downarrow_d$ is a bicubic filter), so that all matching is done on similar patches.

It is still possible to have some noise in the final super-resolved image due to patches being stretched incorrectly over boundaries. Park *et al.* [7] identify outliers based on the contrast of the min and max depth in a local patch in image space. We too wish to identify these outliers, but also to replace them with plausible values and refine the depth estimates of the other points. Using the observation that most of the error is in the depth direction, we propose a new set of possible depth values $\mathbf{d}$ for each pixel, and attempt to solve for the most consistent combination across the image. A 3D coordinate $\mathbf{p^w}$, with position $\mathbf{p^{im}}$ in the image, is labelled as an outlier if the average distance to its T nearest neighbours is greater than τ_{3d}. In the case of non-outlier pixels, the label set $\mathbf{d}$ contains the depth values $p_z^{im} + n\gamma p_z^{im}$ where $n = [-N/2, ..., N/2]$ and each label's unary cost is $|n\gamma p_z^{im}|$, with $\gamma = 1\%$. For the outlier pixels, the label set contains the N nearest non-outlier depth values with uniform unary cost. The pairwise term is the truncated distance between the neighboring depth values i and j: $||p_{i_z}^{im} - p_{j_z}^{im}||_2$. Applying our outlier removal instead of the bilateral filter on the input depth image would produce overly blocky aliased edges that are difficult for the patch lookup to overcome during SR. Used instead as a postprocess, it will only act to remove errors due to incorrect patch scaling.

4 Training Data

For image SR, it is straightforward to acquire image collections online for training purposes. Methods for capturing real scene geometry, *e.g.* laser scanning, are not convenient for collecting large amounts of high quality data in varied environments. Some range datasets do exist online, such as the Brown Range Image Database [24], the USF Range Database [33] and Make3D Range Image Database [34]. The USF dataset is a collection of 400 range images of simple polyhedral objects with a very limited resolution of only 128×128 pixels and with heavily quantized depth. Similarly, the Make3D dataset contains a large variety of low resolution scenes. The Brown dataset, captured using a laser scanner, is most relevant for our SR purposes, containing 197 scenes spanning indoors and outdoors. While the spatial resolution is superior in the Brown dataset, it is still limited and features noisy data such as flying pixels that ultimately hurt depth SR; see Fig. 4 B). In our experiments we compared the results of using Brown Range Image vs. synthetic data, and found superior results using synthetic

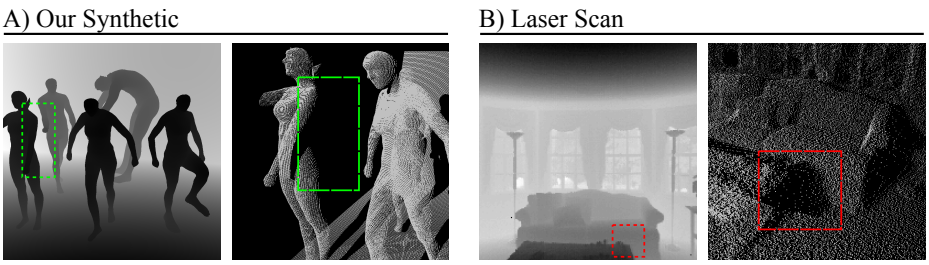

Fig. 4. A) Example from our synthetic dataset. The left image displays the depth image and the right is the 3D projection. Note the clean edges in the depth image and lack of noise in the 3D projection. B) Example scene from the Brown Range Image Database [24] which exhibits low spatial resolution and flying pixels (in red box).

data (see Fig. 7). An example of one of our synthesized 3D scenes is shown in Fig. 4 A). Some synthetic depth datasets also exist, *e.g.* [35], but they typically contain single objects.

Due to the large amount of redundancy in depth scenes (*e.g.* planar surfaces), we prune the high resolution patches before training. This is achieved by detecting depth discontinuities using an edge detector. A dilation is then performed on this edge map (with a disk of radius 0.02 $\times$ the image width) and only patches with centers in this mask are chosen. During testing, the top K closest candidates to the low resolution input patch are retrieved from the training images. Matches are computed based on the $||.||_2$ distance from the low resolution patch to a downsampled version of the high resolution patch. In practice, we use a k-d tree to speed up this lookup. Results are presented using a dataset of 30 scenes of size 800×800 pixels (with each scene also flipped left to right), which creates a dictionary of 5.3 million patches, compared to 660 thousand patches in [16].

5 Experiments

We performed experiments on single depth scans obtained by various means, including a laser scanner, structured light, and three different ToF cameras. We favor the newer ToF sensors because they do not suffer from missing regions at depth-discontinuities as much as Kinect, which has a comparatively larger camera-projector baseline. We run a sliding window filter on the input to fill in missing data with local depth information. The ToF depth images we tested come from one of three camera models: PMD CamCube 2.0 with resolution of 200×200, Mesa Imaging SwissRanger SR3000 with 176×144, or the Canesta EP DevKit with 64×64. We apply bilateral prefiltering and our postprocessing denoising algorithm only to ToF images. Unless otherwise indicated, all experiments were run with the same parameters and with the same training data. We provide comparisons against the Example based Super-Resolution (EbSR) method of [16] and the Sparse coding Super-Resolution (ScSR) method of [17].

A)

		Cones	Teddy	Tsukuba	Venus
x2	MRF RS*	**0.740**	0.527	0.401	0.170
	Cross Bilateral*	0.756	**0.510**	**0.393**	**0.167**
	Nearest Neighbor	1.220	1.006	0.612	0.294
	ScSR	2.065	1.518	0.705	1.003
	EbSR	1.447	0.969	0.617	0.332
	Our method	1.227	0.977	0.601	0.296
x4	MRF RS*	1.141	0.801	0.549	0.243
	Cross Bilateral*	**0.993**	**0.690**	**0.514**	**0.216**
	Nearest Neighbor	2.013	1.351	0.833	0.419
	ScSR	2.614	1.733	0.840	1.044
	EbSR	1.669	1.214	0.869	0.393
	Our method	1.779	1.184	0.833	0.395

B)

	Scan 21	Scan 30	Scan 42
Nearest Neighbor	0.024	**0.016**	0.051
ScSR	0.031	0.035	0.051
EbSR	**0.020**	0.017	0.074
Our method	**0.020**	0.017	**0.040**

C)

	ScSR	EbSR	Our Method
Training time (s)	600	420	750
Testing time (s)	1601	1377	292

Fig. 5. A) RMSE comparison of our method versus several others when upsampling the downsampled Middlebury stereo dataset [36] by a factor of $\times 2$ and $\times 4$. ***MRF RS [25] and Cross Bilateral [1] require an additional intensity image at the same high resolution as the upsampled depth output.* B) RMSE comparison of our method versus two other image based techniques for upsampling three different laser scans by a factor of $\times 4$. See Fig. 6 for visual results of Scan 42. C) Training and testing times in seconds.

5.1 Quantitative Evaluation against Image Based Techniques

In the single image SR community, quantitative results have been criticized for not being representative of perceptual quality [17,18], and some authors choose to ignore them completely [19,21,20]. We first evaluated our technique on the Middlebury stereo dataset [36]. We downsampled the ground truth (using nearest neighbor interpolation) by a factor of $\times 2$ and $\times 4$, and then compared our performance at reconstructing the original image. The error for several different algorithms is reported as the Root Mean Squared Error (RMSE) in Fig. 5 A). Excluding the algorithms that use additional scene information, we consistently come first or second and achieve the best average score across the two experiments. It should be noted that, due to the heavy quantization of the disparity, trivial methods such as nearest neighbor interpolation perform numerically well but perceptually they can exhibit strong artifacts such as jagged edges. As the scale factor increases, nearest neighbor's performance decreases. This is consistent with the same observation for bilinear interpolation reported in [7].

We also report results for three laser scans upsampled by a factor of $\times 4$; see Fig 5 B). Again our method performs best overall. Fig. 6 shows depth images along with 3D views of the results of each algorithm. It also highlights artifacts for both of the intensity image based techniques at depth discontinuities. EbSR [16] smooths over the discontinuities while ScSR [17] introduces high frequency errors that manifest as a ringing effect in the depth image and as spikes in the 3D view. Both competing methods also fail to reconstruct detail on the object's surface such as the ridges on the back of the statue. Our method produces sharp edges like the ones present in the ground truth and also detail on the object's surface.

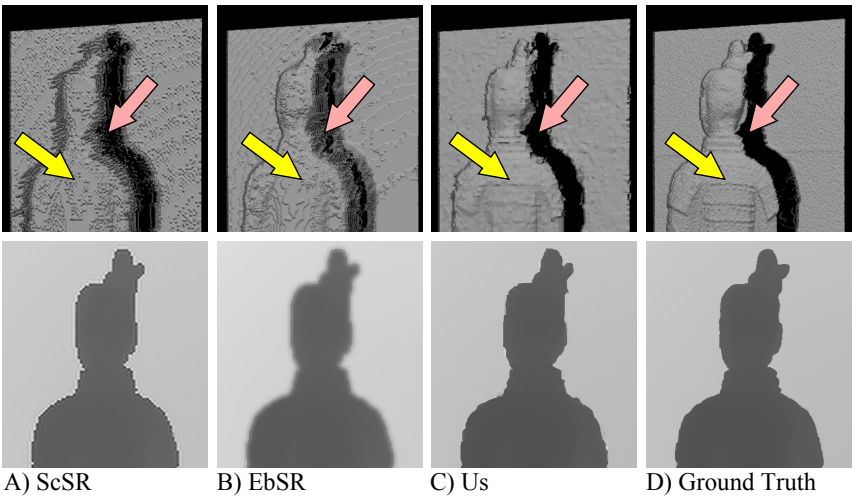

Fig. 6. 3D synthesized views and corresponding depth images (cropped versions of the original) of Scan 42 from Fig. 5 B) upsampled $\times 4$. A) ScSR [17] introduces high frequency artifacts at depth discontinuities. B) EbSR [16] over-smooths the depth due to its initial interpolated upsampling. This effect is apparent in both the 3D view and depth image (pink arrow). C) Our method inserts sharp discontinuities and detail on the object surface (yellow arrow). D) Ground truth laser scan.

5.2 Qualitative Results

As described in Section 4, noisy laser depth data is not suitable for training. Fig. 7 shows the result of SR when training on the Brown Range Image Database [24] (only using the indoor scenes) as compared with our synthetic dataset. The noise in the laser scan data introduces both blurred and jagged artifacts.

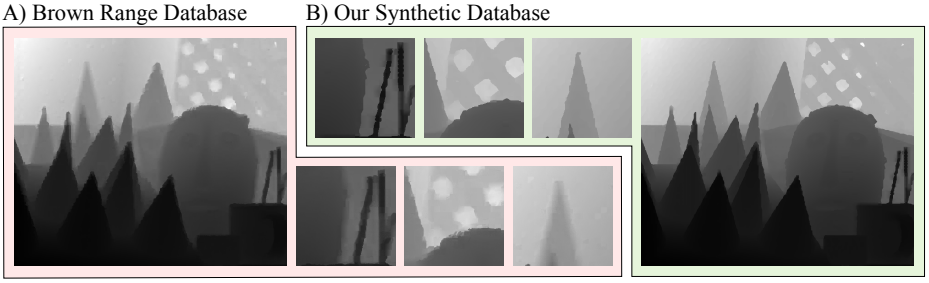

Fig. 7. Result of super-resolving a scene using our algorithm with training data from two different sources: A) Laser scan [24] B) Our synthetic. Note that our dataset produces much less noise in the final result and hallucinates detail such as the thin structures in the right of the image.

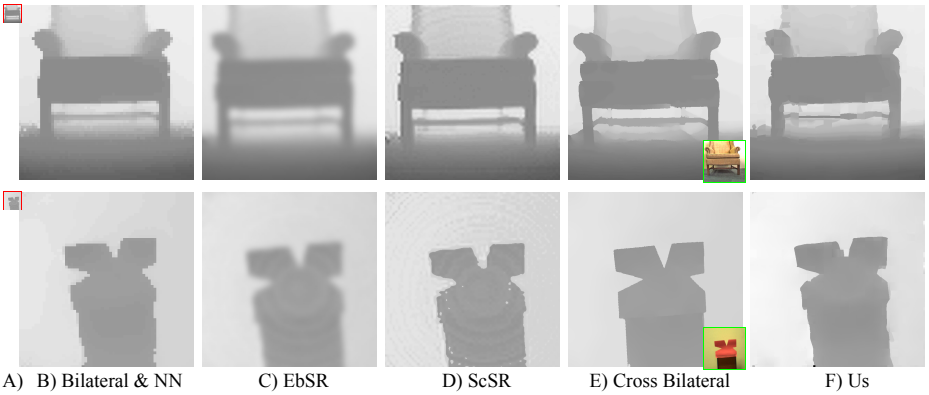

Fig. 8. Upsampling input ToF image from a Canesta EP DevKit (64×64) by a factor of $\times 10$ [1]. A) Input depth image shown to scale in red. B) Bilateral filtering of input image (to remove noise) followed by upsampling using nearest neighbor. C) EbSR [16] produces an overly smooth result at this large upsampling factor. D) ScSR [17] recovers more high frequency detail but creates a ringing artefact. E) The Cross Bilateral [1] method produces a very sharp result, however, the method requires a high resolution intensity image at the same resolution as the desired super-resolved depth image (640×640), shown inset in green. F) Our method produces sharper results than C) and D).

We also compare ourselves against the cross bilateral method of Yang *et al.* [1]. Their technique uses an additional high resolution image of the same size as the desired output depth image. Fig. 8 shows results when upsampling a Canesta EP DevKit 64×64 ToF image by a factor of $\times 10$. It is important to note that to reduce noise, [1] use an average of many successive ToF frames as input. The other methods, including ours, use only a single depth frame.

Fig. 9 shows one sample result of our algorithm for a noisy ToF image captured using a PMD CamCube 2.0. The image has a starting resolution of 200×200 and is upsampled by a factor of $\times 4$. The zoomed regions in Fig. 9 C) demonstrate that we synthesize sharp discontinuities. For more results, including moving scenes, please see our project webpage.

Implementation Details

Fig. 5 C) gives an overview of the the training and test times of our algorithm compared to other techniques. All results presented use a low resolution patch size of 3×3. For the MRF, we use 150 labels (high resolution candidates) and the weighting of the pairwise term, λ, is set to 10. The only parameters we change are for the Bilateral filtering step. For scenes of high noise we set the window size to 5, the spatial standard deviation to 3 and range deviation to 0.1; for all other scenes we use 5, 1.5 and 0.01 respectively. We use the default parameters provided in the implementations for ScSR [17] and EbSR [16]. To encourage comparison and further work, code and training data are available from our project webpage.

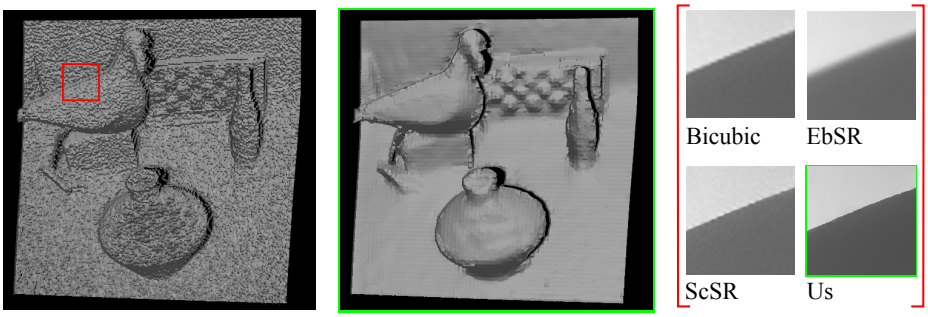

Fig. 9. CamCube ToF input. A) Noisy input. B) Our result for SR by $\times 4$; note the reduced noise and sharp discontinuities. C) Cropped region from the input depth image comparing different SR methods. The red square in A) is the location of the region.

6 Conclusions

We have extended single-image SR to the domain of depth images. In the process, we have also assessed the suitability for depth images of two leading intensity image SR techniques [17,16]. Measuring the RMSE of super-resolved depths with respect to known high resolution laser scans and online Middlebury data, our algorithm is always first or a close second best. We also show that perceptually, we reconstruct better depth discontinuities. An additional advantage of our single frame SR method is our ability to super-resolve moving depth videos. From the outset, our aim of super-resolving a lone depth frame has been about producing a qualitatively believable result, rather than a strictly accurate one. Blurring and halos may only become noticeable as artifacts when viewed in 3D.

Four factors enabled our algorithm to produce attractive depth reconstructions. Critically, we saw an improvement when we switched to low resolution searches for our unary potentials, unlike almost all other algorithms. Second, special depth-rendering of clean computer graphics models depicting generic scenes outperforms training on noisy laser scan data. It is important that these depths are filtered and downsampled for training, along with a pre-selection stage that favors areas near gradients. Third, the normalization based on min/max values in the low resolution input allows the same training patch pair to be applied at various depths. We found that the alternative of normalizing for mean and variance is not very robust with 3×3 patches, because severe noise in many depth images shifts the mean to almost one extreme or the other at depth discontinuities. Finally, we have the option for refining our high resolution depth output using a targeted noise removal algorithm. Crucially, our experiments demonstrate both the quantitative and perceptually significant advantages of our new method.

Limitations and Future Work

Currently, for a video sequence we process each frame independently. Our algorithm could be extended to exploit temporal context to both obtain the most reliable SR reconstruction, and apply it smoothly across the sequence. Similarly, the context used to query the database could be expanded to include global scene information, such as structure [37]. This may overcome the fact that we have a low local signal to noise ratio from current depth sensors. As observed by Reynolds *et al*. [28], flying pixels and other forms of severe non-Gaussian noise necessitates prefiltering. Improvements could be garnered with selective use of the input depths via a sensor specific noise model. For example, *Kinect* and the different ToF cameras we tested all exhibit very different noise characteristics.

Acknowledgements. Funding for this research was provided by the NUI Travelling Studentship in the Sciences and EPSRC grant EP/I031170/1.

References

1. Yang, Q., Yang, R., Davis, J., Nister, D.: Spatial-depth super resolution for range images. In: CVPR (2007)
2. Schuon, S., Theobalt, C., Davis, J., Thrun, S.: Lidarboost: Depth superresolution for tof 3D shape scanning. In: CVPR (2009)
3. Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning low-level vision. IJCV (2000)
4. Kuster, C., Popa, T., Zach, C., Gotsman, C., Gross, M.: Freecam: A hybrid camera system for interactive free-viewpoint video. In: Proceedings of Vision, Modeling, and Visualization, VMV (2011)
5. Holz, D., Schnabel, R., Droeschel, D., Stückler, J., Behnke, S.: Towards semantic scene analysis with time-of-flight cameras. In: RoboCup International Symposium (2010)
6. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR (2011)
7. Park, J., Kim, H., Tai, Y.W., Brown, M., Kweon, I.: High quality depth map upsampling for 3D-ToF cameras. In: ICCV (2011)
8. Fattal, R.: Upsampling via imposed edges statistics. SIGGRAPH (2007)
9. Yang, L., Sander, P.V., Lawrence, J., Hoppe, H.: Antialiasing recovery. ACM Transactions on Graphics (2011)
10. Irani, M., Peleg, S.: Improving resolution by image registration. In: CVGIP: Graph. Models Image Process. (1991)
11. Hahne, U., Alexa, M.: Exposure Fusion for Time-Of-Flight Imaging. In: Pacific Graphics (2011)
12. Rajagopalan, A.N., Bhavsar, A., Wallhoff, F., Rigoll, G.: Resolution Enhancement of PMD Range Maps. In: Rigoll, G. (ed.) DAGM 2008. LNCS, vol. 5096, pp. 304–313. Springer, Heidelberg (2008)
13. Izadi, S., Kim, D., Hilliges, O., Molyneaux, D., Newcombe, R.A., Kohli, P., Shotton, J., Hodges, S., Freeman, D., Davison, A.J., Fitzgibbon, A.: Kinectfusion: Real-time 3D reconstruction and interaction using a moving depth camera. In: UIST (2011)

14. Cui, Y., Schuon, S., Derek, C., Thrun, S., Theobalt, C.: 3D shape scanning with a time-of-flight camera. In: CVPR (2010)
15. Freeman, W.T., Jones, T.R., Pasztor, E.C.: Example-based super-resolution. Computer Graphics and Applications (2002)
16. Freeman, W.T., Liu, C.: Markov random fields for super-resolution and texture synthesis. In: Advances in Markov Random Fields for Vision and Image Processing. MIT Press (2011)
17. Yang, J., Wright, J., Huang, T.S., Ma, Y.: Image super-resolution via sparse representation. IEEE Transactions on Image Processing (2010)
18. Zontak, M., Irani, M.: Internal statistics of a single natural image. In: CVPR (2011)
19. Glasner, D., Bagon, S., Irani, M.: Super-resolution from a single image. In: ICCV (2009)
20. Sun, J., Zhu, J., Tappen, M.: Context-constrained hallucination for image super-resolution. In: CVPR (2010)
21. HaCohen, Y., Fattal, R., Lischinski, D.: Image upsampling via texture hallucination. In: ICCP (2010)
22. Gal, R., Shamir, A., Hassner, T., Pauly, M., Cohen-Or, D.: Surface reconstruction using local shape priors. In: Symposium on Geometry Processing (2007)
23. Golovinskiy, A., Matusik, W., Pfister, H., Rusinkiewicz, S., Funkhouser, T.: A statistical model for synthesis of detailed facial geometry. SIGGRAPH (2006)
24. Huang, J., Lee, A., Mumford, D.: Statistics of range images. In: CVPR (2000)
25. Diebel, J., Thrun, S.: An application of markov random fields to range sensing. In: NIPS (2005)
26. Chan, D., Buisman, H., Theobalt, C., Thrun, S.: A Noise-Aware Filter for Real-Time Depth Upsampling. In: Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (2008)
27. Schuon, S., Theobalt, C., Davis, J., Thrun, S.: High-quality scanning using time-of-flight depth superresolution. In: CVPR Workshops (2008)
28. Reynolds, M., Doboš, J., Peel, L., Weyrich, T., Brostow, G.J.: Capturing time-of-flight data with confidence. In: CVPR (2011)
29. Kolmogorov, V.: Convergent tree-reweighted message passing for energy minimization. PAMI (2006)
30. Wainwright, M., Jaakkola, T., Willsky, A.: Map estimation via agreement on (hyper)trees: Message-passing and linear programming approaches. IEEE Transactions on Information Theory (2002)
31. Frank, M., Plaue, M., Rapp, H., Köthe, U., Jähnea, B., Hamprecht, F.A.: Theoretical and experimental error analysis of continuous-wave time-of-flight range cameras. Optical Engineering (2009)
32. Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: ICCV (1998)
33. (USF Range Database), `http://marathon.csee.usf.edu/range/DataBase.html`
34. Saxena, A., Sun, M., Ng, A.: Make3d: Learning 3d scene structure from a single still image. PAMI (2009)
35. Saxena, A., Driemeyer, J., Kearns, J., Ng, A.: Robotic grasping of novel objects. In: NIPS (2006)
36. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. IJCV (2002)
37. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor Segmentation and Support Inference from RGBD Images. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 746–760. Springer, Heidelberg (2012)