Perceptual Deep Depth Super-Resolution

Oleg Voynov[1], Alexey Artemov[1], Vage Egiazarian[1], Alexander Notchenko[1],
Gleb Bobrovskikh[1,2], Evgeny Burnaev[1], Denis Zorin[3,1]
[1]Skolkovo Institute of Science and Technology, [2]Higher School of Economics,
[3]New York University
{oleg.voinov, a.artemov, vage.egiazarian, alexandr.notchenko}@skoltech.ru,
bobrovskihg@gmail.com, e.burnaev@skoltech.ru, dzorin@cs.nyu.edu
adase.group/3ddl/projects/perceptual-depth-sr

Abstract

RGBD images, combining high-resolution color and lower-resolution depth from various types of depth sensors, are increasingly common. One can significantly improve the resolution of depth maps by taking advantage of color information; deep learning methods make combining color and depth information particularly easy.

However, fusing these two sources of data may lead to a variety of artifacts. If depth maps are used to reconstruct 3D shapes, e.g.*, for virtual reality applications, the visual quality of upsampled images is particularly important.*

The main idea of our approach is to measure the quality of depth map upsampling using renderings of resulting 3D surfaces. We demonstrate that a simple visual appearance-based loss, when used with either a trained CNN or simply a deep prior, yields significantly improved 3D shapes, as measured by a number of existing perceptual metrics. We compare this approach with a number of existing optimization and learning-based techniques.

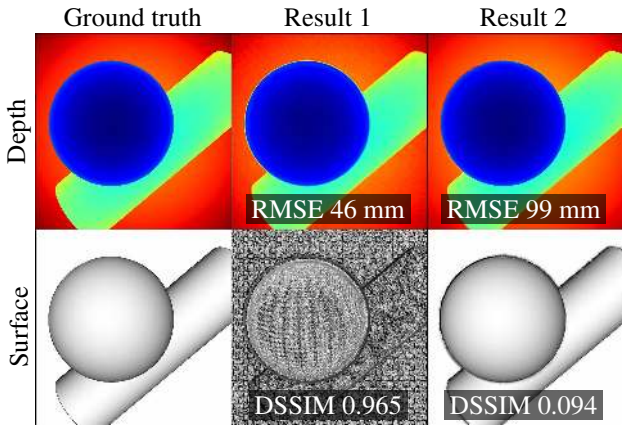

Figure 1: Visually inferior super-resolution result in the middle gets higher score according to direct depth deviation but lower score according to perceptual deviation of the rendered image of the 3D surface. While the surfaces differ significantly, the corresponding depth maps do not capture this difference and look almost identical.

1. Introduction

RGBD images are increasingly common as sensor technology becomes more widely available and affordable. They can be used for reconstruction of the 3D shapes of objects and their surface appearance. The better the quality of the depth component, the more reliable the reconstruction.

Unfortunately, for most methods of depth acquisition the resolution and quality of the depth component is insufficient for accurate surface reconstruction. As the resolution of the RGB component is usually several times higher and there is a high correlation between structural features of the color image and the depth map (*e.g.*, object edges) it is natural to use the color image for depth map super-resolution, *i.e.* upsampling of the depth map. Convolutional neural networks are a natural fit for this problem as they can easily fuse heterogeneous information.

A critical aspect of any upsampling method is the measure of quality it optimizes (*i.e.*, the loss function), whether the technique is data-driven or not. In this paper we focus on applications that require reconstruction of 3D geometry visible to the user, like acquisition of realistic 3D scenes for virtual or augmented reality and computer graphics. In these applications the *visual* appearance of the resulting 3D shape, *i.e.*, how the surface looks when observed under various lighting conditions, is of particular importance.

Most existing research on depth super-resolution is dominated by simple measures based on pointwise deviation of depth values. However, direct pointwise difference of the depth maps do not capture the visual difference between

the corresponding 3D shapes: for example, low-amplitude high-frequency variations of depth may correspond to significant difference in appearance, while conversely, relatively large smooth changes in depth may be perceptually less relevant, as illustrated in Figure 1.

Hence, we propose to compare the rendered images of the surface instead of the depth values directly. In this paper we explore depth map super-resolution using a simple loss function based on visual differences. Our loss function can be computed efficiently and is shown to be highly correlated with more elaborate perceptual metrics. We demonstrate that this simple idea used with two deep learning-based RGBD super-resolution algorithms results in a dramatic improvement of visual quality according to perceptual metrics and an informal perceptual study. We compare our results with six state-of-the-art methods of depth super-resolution that are based on distinct principles and use several types of loss functions.

In summary, our contributions are as follows: (1) we demonstrate that a simple and efficient visual difference-based metric for depth map comparison can be, on the one hand, easily combined with neural network-based whole-image upsampling techniques, and, on the other hand, is correlated with established proxies for human perception, validated with respect to experimental measurements; (2) we demonstrate with extensive comparisons that with the use of this metric two methods of depth map super-resolution, one based on a trainable CNN and the other based on the deep prior, yield high-quality results as measured by multiple perceptual metrics. To the best of our knowledge, our paper is the first to systematically study the performance of visual difference-based depth super-resolution across a variety of datasets, methods, and quality measures, including a basic human evaluation.

Throughout the paper we use the term *depth map* to refer to the depth component of an RGBD image, and the term *normal map* to refer to the map of the same resolution with the 3D surface normal direction computed from the depth map at each pixel. Finally, the *rendering of a depth map* refers to the grayscale image obtained by constructing a 3D triangulation of the height field represented by the depth map, via computing the normal map from this triangulation, and rendering it using fixed material properties and a choice of lighting. This is distinct from a commonly used depth map visualization with grayscale values obtained from the depth values by simple scaling. We describe this in more detail in Section 3.

2. Related work

2.1. Image quality measures

Quality measures play two important roles in image super-resolution: on the one hand, they are used to formulate an optimization functional or a loss function, on the other hand, they are used to evaluate the quality of the results. Ideally, the same function should serve both purposes, however, in some instances it may be optimal to choose different functions for evaluation and optimization. While in the former case the top priority is to capture the needs of the application, in the latter case the efficiency of evaluation and differentiability are significant considerations.

In most works on depth map reconstruction and upsampling a limited number of simple metrics are used, both for optimization and final evaluation. Typically these are scaled L_2 or L_1 norms of depth deviations (see *e.g.* [9]).

Another set of measures introduced in [19, 20] and primarily used for evaluation, not optimization or learning, consists of heuristic measures of various aspects of the depth map geometry: foreground flattening/thinning, fuzziness, bumpiness, etc. Most of them require a very specific segmentation of the image for detection of flat areas and depth discontinuities.

Visual similarity measures, well-established in the area of photo-processing, aim to be consistent with human judgment, in the sense of similarity ordering (which of the two images is more similar to the ground truth?). The examples include (1) the metrics based on simple vision models of *structural similarity* SSIM [51], FSIM [56], MSSIM [52], (2) based on a sophisticated model of low-level visual processing [35], or (3) on convolutional neural networks (see [57] for a detailed overview). The latter use a simple distance measure on deep features learned for an image understanding task, *e.g.* L_2 distance on the features learned for image classification, and have been demonstrated to outperform statistical measures such as SSIM.

2.2. Depth super-resolution

Depth super-resolution is closely related to a number of depth processing tasks, such as denoising, enhancement, inpainting, and densification (*e.g.*, [5, 6, 8, 21, 33, 34, 45, 46, 54]). We directly focus on the problem of super-resolution, or more specifically, estimation of high-resolution depth map from a single low-resolution depth map and a high-resolution RGB image.

Convolutional neural networks have achieved most impressive performance among learning-based methods in high-level computer vision tasks and recently have been applied to depth super-resolution [22, 30, 39, 43]. One approach [22] is to resolve ambiguity in the depth map upsampling by explicitly adding high-frequency features from high-resolution RGB data. Another, hybrid approach [39, 43] is to add a subsequent optimization stage to a CNN to produce sharper results. Different approaches to CNN-based photo-guided depth super-resolution include linear filtering with CNN-derived kernels [26], deep fusion of time-of-flight depth and stereo images [1], and generative

adversarial networks [61].

These techniques use either L_2 or L_1 norm of the depth differences as the basis of their loss functions, often combined with regularizers of different types. The recent approach of [61] is the closest to ours: it uses the difference of gradients as one of the loss terms to capture some of the visual information. For evaluation, these works report root mean square error (RMSE), mean absolute error (MAE), peak signal-to-noise ratio (PSNR), all applied directly to depth maps, and, rarely [4, 43, 44, 61], perceptual SSIM *also applied directly to depth maps*. In contrast, we propose to measure the perceptual quality of depth map *renderings*.

Dictionary learning has also been investigated for depth super-resolution [11, 13, 29], however, compared to CNNs, it is typically restricted to smaller dimensions and as a result to structurally simpler depth maps.

Variational approach aims to combine RGB and depth information explicitly by carefully designing an optimization functional, without relying on learning. Most relevant examples employ shape-from-shading problem statement for single-image [14] or multiple-image [38] depth super-resolution. These works include visual difference-related terms in the optimized functional and report normal deviation, capturing visual similarity. While showing impressive results in many cases, they typically require prior segmentation of foreground objects and depend heavily on the quality of such segmentation.

Another strategy to tackle ambiguities in super-resolution is to design sophisticated regularizers to balance the data-fidelity terms against a structural image prior [15, 24, 55]. In contrast to this approach, which requires custom hand-crafted regularized objectives and optimization procedures, we focus on the standard training strategy (*i.e.*, gradient-based optimization of a CNN) while using a loss function that captures visual similarity.

Yet another approach is to choose a carefully-designed model such as [62] featuring a sophisticated metric defined in a space of minimum spanning trees and including an explicit edge inconsistency model. In contrast to ours, such model requires manual tuning of multiple hyperparameters.

2.3. Perceptual photo super-resolution

Perceptual metrics have been considered more broadly in the context of photo processing. While convolutional neural networks for photo super-resolution trained with simple mean square or mean absolute color deviation keep demonstrating impressive results [16, 18, 58, 59], it has been widely recognized that pixelwise difference of color image data is not well correlated with perceptual image difference. For this reason, relying on a pixelwise color error may lead to suboptimal performance.

One solution is to instead use the loss function represented by the deviation of the features from a neural network trained for an image understanding task [25]. This idea can be further combined with an adversarial training procedure to push the super-resolution result to the natural image manifold [28]. Another extension to this idea is to train the neural network to generate images with natural distribution of statistical features [12, 36, 49, 50]. To balance between the perceptual quality and pixelwise color deviation, generative adversarial networks can be used [7, 31, 48].

Figure 2: Depth map renderings generated with four light directions that we use for metric calculation.

Another solution is to learn a quality measure from perceptual scores, collected from a human subject study, and use this quality measure as the loss function. Such quality measure may capture similarity of two images [57] or an absolute naturalness of the image [32].

3. Metrics

In this section, we discuss visually-based metrics and how they can be used to evaluate the quality of depth map super-resolution and as loss functions. The general principle we follow is to apply comparison metrics to *renderings* of the depth maps to obtain a measure of their difference instead of considering depth maps directly. The difficulty with this approach is that there are infinitely many possible renderings depending on lighting conditions, material properties and camera position. However, we demonstrate that even a very simple rendering procedure already yields substantially improved results. We label visually-based metrics with subscript "v" and the metrics that compare the depth values directly with subscript "d".

From depth map to visual representation. To approximate the appearance of a 3D scene depicted with a certain depth map we use a simple rendering procedure. We illuminate the corresponding 3D surface with monochromatic directional light source and observe it with the same camera that the scene was originally acquired with. We use the diffuse reflection model and do not take visibility into account. For this model, the intensity of a pixel (i, j) of the rendering I is proportional to cosine of the angle between the normal at the point of the surface corresponding to the pixel $\boldsymbol{n}_{ij}$ and direction to the light source $\boldsymbol{e}$: $I_{ij} = \boldsymbol{e} \cdot \boldsymbol{n}_{ij}$. We calculate the normals from the depth maps using first-order finite-differences. Any number of vectors $\boldsymbol{e}$ can be used to generate a collection of renderings representing the depth map, however, any rendering can be obtained as a

linear combination of three basis ones corresponding to independent light directions. Renderings for different light directions are presented in Figure 2.

Perceptual metrics. We briefly describe two representative metrics: a statistics-based DSSIM, and a neural network-based LPIPS. Either of these can be applied to three basis renderings (or a larger sample of renderings) and reduced to obtain the final value. While, in principle, they can also be used as loss functions, the choice of a loss function needs to take stability and efficiency into account, so we opt for a more conservative choice described below.

Structural similarity index measure (SSIM) [51] takes into account the changes in the local structure of an image, captured by statistical quantities computed on a small window around each pixel. For each pair of pixels of the compared images I_k, $k = 1, 2$ the luminance term ℓ, the contrast term c and the structural term s, each normalized, are computed using the means μ_k, standard deviations σ_k and cross-covariance σ_{12} of the pixels in the corresponding local windows. The value of SSIM is then computed as pixelwise mean product of these terms

$$\begin{gathered} \ell = \frac{2\mu_1\mu_2}{\mu_1^2+\mu_2^2},\ c = \frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2},\ s = \frac{\sigma_{12}}{\sigma_1\sigma_2}, \\ \mathrm{SSIM_v}(I_1, I_2) = \frac{1}{N}\sum_{ij} \ell_{ij}\cdot c_{ij}\cdot s_{ij}, \end{gathered} \tag{1}$$

where N is the number of pixels. Dissimilarity measure can be computed as $\mathrm{DSSIM_v}(I_1, I_2) = 1 - \mathrm{SSIM_v}(I_1, I_2)$.

Neural net-based metrics rely on the idea of measuring the distance between features extracted from a neural network. Specifically, feature maps $\boldsymbol{x}_{k\ell}$, $\ell = 1 \dots L$ with spatial dimensions $H_\ell \times W_\ell$ are extracted from L layers of the network for each of the compared images. In the simplest case, the metric value is then computed as pixelwise mean square difference of the feature maps, summed over the layers

$$\mathrm{NN_v}(I_1, I_2) = \sum_\ell \frac{1}{H_\ell W_\ell}\sum_{ij} \|\boldsymbol{x}_{1\ell,ij} - \boldsymbol{x}_{2\ell,ij}\|_2^2. \tag{2}$$

Learned perceptual image patch similarity (LPIPS) [57] adds a learned channel-wise weighting to the above formula and uses 5 layers from Alexnet [27] or VGG [41] or the first layer from Squezenet [23] as the CNN of choice.

Our visual difference-based metric. While the metrics described above are good proxies for human evaluation of difference between depth map renderings, they are lacking as loss functions due to their complex landscapes. Optimization with DSSIM as the loss function may produce the results actually inferior with respect to *DSSIM itself* compared to a simpler loss function we define below, as illustrated in Figure 3. LPIPS has a complex energy profile typical for neural networks, and having a neural network as the loss function for another may behave unpredictably [60].

The simplest metric capturing the difference between all

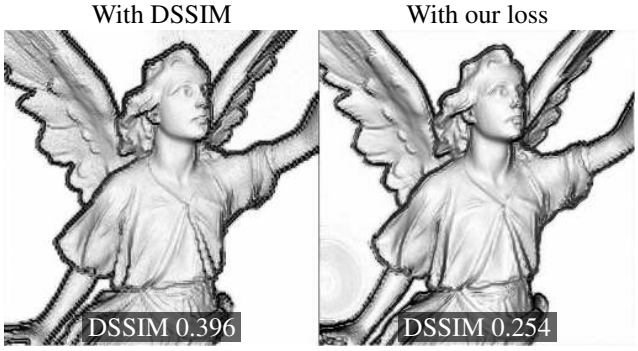

Figure 3: Optimization with DSSIM as the loss function may produce the results inferior with respect to *DSSIM itself* compared to our simpler loss function.

possible renderings of the depth maps d_k can be computed as the average root mean square deviation of three basis renderings $\boldsymbol{e}_m \cdot \boldsymbol{n}_k$ in an orthogonal basis $\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3$

$$\begin{gathered} \mathrm{RMSE_v}(d_1, d_2) = \sqrt{\mathrm{MSE_v}(d_1, d_2)}, \\ \mathrm{MSE_v}(d_1, d_2) = \frac{1}{3N}\sum_{ij,m} \|\boldsymbol{e}_m\cdot\boldsymbol{n}_{1,ij} - \boldsymbol{e}_m\cdot\boldsymbol{n}_{2,ij}\|_2^2, \end{gathered} \tag{3}$$

similarly to RMS difference of the normal maps.

We found that this simple metric for depth map comparison is efficient and stable as the loss function and at the same time, as we demonstrate in Section 5, it is well correlated with DSSIM and LPIPS, *i.e.*, situations when the value of one metric is high and the value of another is low are unlikely. Our experiments confirm that optimization of this metric also improves both perceptual metrics.

4. Methods

We selected eight representative state-of-the-art depth processing methods based on different principles: (1) a purely variational method [14], (2) a bilateral filtering method that uses a high-resolution edge map [53], (3) a dictionary learning method [13], (4) a hybrid CNN-variational method [39], (5) a pure CNN [22], (6) a zero-shot CNN [47], (7) a densification [34] and (8) an enhancement [54] CNNs. Our goals were (a) to modify the methods for using with the visual difference-based loss function, and (b) to compare the results of the modified methods with alternatives of different types. In our experiments the last two methods did not perform well compared to others, so we did not consider them further. We found that two neural network-based methods (5) and (6), that we refer to as MSG and DIP, can be easily modified for using with a visual difference-based loss function, as we explain now.

MSG [22] is a deep learning method that uses different strategies to upsample different spectral components of low-resolution depth map. In the modified version of this method, that we denote by **MSG-V**, we replaced the

original loss function with a combination of our visual difference-based metric and mean absolute deviation of Laplacian pyramid Lap_1 [2] as a regularizer

$$\mathcal{L}(d_1, d_2) = \mathrm{Lap}_1(d_1, d_2) + w \cdot \mathrm{MSE}_\mathrm{v}(d_1, d_2). \quad (4)$$

DIP [47] is a zero-shot deep learning approach, based on a remarkable observation that, even without any specialized training, the structure of CNN itself may be leveraged for solving inverse problems on images. We note that this approach naturally allows simultaneous super-resolution and inpainting. In this approach, the depth super-resolution problem would be formulated as

$$d^\mathrm{SR}_{\theta^*} = \mathrm{CNN}_{\theta^*},\ \theta^* = \arg\min_\theta \mathrm{MSE}_\mathrm{d}(\mathbf{D}d^\mathrm{SR}_\theta, d^\mathrm{LR}), \quad (5)$$

where d^LR and $d^\mathrm{SR}_{\theta^*}$ are the low-resolution and super-resolved depth maps, CNN_θ is the output of the deep neural network parametrised by θ, $\mathbf{D}$ is the downsampling operator, and MSE_d is direct mean square difference of the depth maps. To perform photo-guided super-resolution, we added a second output channel for intensity to the network

$$\begin{gathered} d^\mathrm{SR}_{\theta^*} = \mathrm{CNN}^{(1)}_{\theta^*}, \quad I_\theta = \mathrm{CNN}^{(2)}_\theta, \\ \theta^* = \arg\min_\theta \mathrm{MSE}_\mathrm{d}(\mathbf{D}d^\mathrm{SR}_\theta, d^\mathrm{LR}) + w_I \cdot \mathrm{Lap}_1(I_\theta, I^\mathrm{HR}), \end{gathered} \quad (6)$$

where I^HR is the high-resolution photo guidance, and for visually-based version **DIP-V** we further replaced the direct depth deviation MSE_d with the function from Equation 4.

We used the remaining four methods (1)-(4) for comparison as-is, as modifying them for a different loss function would require substantial changes to the algorithms.

SRfS [14] is a variational method relying on complimentarity of super-resolution and shape-from-shading problems. It already includes a visual-difference based term (the remaining methods use depth difference metrics). **EG** [53] approaches the problem via prediction of smooth high-resolution depth edges with Markov random field optimization. It does not use a loss directly, therefore cannot be easily adapted. **DG** [13] is a depth map enhancement method based on dictionary learning that uses depth difference-based fidelity term. It makes a number of modeling choices which may not be suitable for a different loss function, and typically does not perform as well as neural network-based methods. **PDN** [39] is a hybrid method featuring two stages: the first is composed of fully-convolutional layers and predicts a rough super-resolved depth map, and the second performs an unrolled variational optimization, aiming to produce a sharp and noise-free result.

5. Experiments

5.1. Data

For evaluation we selected a representative and diverse set of 34 RGBD images featuring synthetic, high-quality real and low-quality real data with different levels of geometric and textural complexity. We employed four datasets, most common in literature on depth super-resolution. *ICL-NUIM* [17] includes photo-realistic RGB images along with synthetic depth, free from any acquisition noise. *Middlebury 2014* [40], captured with a structured light system, provides high-quality ground truth for complex real-world scenes. *SUN RGBD* [42] contains images captured with four different consumer-level RGBD cameras: Intel RealSense, Asus Xtion, Microsoft Kinect v1 and v2. *ToF-Mark* [10] provides challenging real-world time-of-flight and intensity camera acquisitions together with an accurate ground truth from a structured light sensor.

In addition, we constructed a synthetic *SimGeo* dataset, that consists of 6 geometrically simple scenes with low- and high-frequency texture, and without any, using Blender. The purpose of SimGeo were to reveal artifacts that are not related to the noise or high-frequency geometry in the input data, like false geometric detail caused by color variation on a smooth surface.

We resized and cropped each RGBD image to the resolution of 512×512 and generated low-resolution input depth maps with the scaling factors of 4 and 8, that are most common among the works on depth super-resolution. We focused on two downsampling models: Box, *i.e.*, each low-resolution pixel contains the mean value over the "box" neighbouring high-resolution pixels, and Nearest neighbour, *i.e.*, each low-resolution pixel contains the value of the nearest high-resolution pixel. For additional details on our evaluation data and the results for different downsampling models please refer to supplementary material.

5.2. Evaluation details

To quantify the performance of the methods, we measured direct RMS deviation of the depth maps (denoted by RMSE_d) and deviation of their renderings with the metrics described in Section 3. For visually-based metrics we calculated their values for three orthogonal light directions, corresponding to the three left-most images in Figure 2, and the value for an additional light direction, corresponding to the right-most image. We then took the worst of the four values. With similar outcomes, we also explored different reducing strategies and a set of different metrics: BadPix and Bumpiness, applied directly to depth values, and BadPix and RMSE applied to separate depth map renderings.

Additionally, we conducted an informal perceptual study using the results on SimGeo, ICL-NUIM and Middlebury datasets, in which subjects were asked to choose the renderings of the upsampled depth maps that look most similar to the ground truth.

5.3. Implementation details

We evaluated publicly available trained models for EG, DG, and MSG and trained PDN using publicly available

code; we used the implementation of SRfS provided by the authors; we adapted publicly available implementation of DIP for depth maps, as described in Section 4; we reimplemented MSG-V in PyTorch [37] and trained it according to the original paper using the patches from Middlebury and MPI Sintel [3]. We selected the value of the weighting parameter w in Equation (4) so that both terms of the loss contribute equally with respect to their magnitudes (see supplementary material for more details).

5.4. Comparison of quality measures

To quantify how well different metrics represent the visual quality of a super-resolved depth map, we compared pairwise correlations of these metrics and calculated the corresponding values of Pearson correlation coefficient. Since LPIPS as a neural network-based perceptual metric has been experimentally shown to represent human perception well, we used its value as the reference. We found that the metrics based on direct depth deviation demonstrate weak correlation with perceptual metrics, as illustrated in Figure 4 for $\mathrm{RMSE_d}$, and hence are not suitable for measuring the depth map quality when the visual appearance plays an important role. On the other hand, we found that our $\mathrm{RMSE_v}$ correlates well with perceptual metrics, to the same extent they correlate with each other (see Figure 4).

5.5. Comparison of super-resolution methods

In Table 1 and Figure 5 we present the super-resolution results on our SimGeo dataset with the scaling factor of 4; in Table 2 and Figure 6 we present the results on ICL-NUIM and Middlebury datasets with the scaling factors of 4 and 8. We use Box downsampling model in both cases. Please find the additional results in supplementary material or online[1].

In general, we found that the methods EG, PDN and DG do not recover fine details of the surface, typically oversmoothing the result in comparison to, *e.g.*, Bicubic upsampling, the methods SRfS and original DIP suffer from false geometry artifacts in case of a smooth textured surface, and original MSG introduces severe noise around the depth edges. As illustrated in Figure 6 and Table 2, all the methods from prior works perform relatively poorly on the images with regions of missing depth measurements (rendered in black), including the ones that inpaint these regions explicitly (SRfS, DG) or implicitly (DIP). The method EG failed to converge on some images.

In contrast, we observed that integration of our visual difference-based loss into DIP and MSG significantly improved the results of both methods qualitatively and quantitatively. The visual difference-based version DIP-V do not suffer from false geometry artifacts as much as the original version. On the challenging images from Middlebury dataset, where it performed simultaneous super-resolution and inpainting, DIP-V mostly outperformed other methods as measured by the perceptual metrics and was preferred by more than 80% of subjects in the perceptual study. The visual difference-based version MSG-V produces significantly less noisy results in comparison to the original version, in some cases almost without any noticeable artifacts. On the data without missing measurements, including hole-filled "Vintage" from Middlebury, MSG-V mostly outperformed other methods as measured by the perceptual metrics and was preferred by more than 80% of subjects. On SimGeo, ICL-NUIM and Middlebury combined, one of our modified versions, DIP-V or MSG-V, was preferred over the other methods by more than 85% of subjects.

For reference, in Figure 5 we include pseudo-color visualizations of the depth maps. Notice that while the upsampled depth maps obtained with different methods are almost indistinguishable in this form of visualization, commonly used in the literature on depth processing for qualitative evaluation, the corresponding renderings and, consequently, the underlying geometry varies dramatically.

6. Conclusion

We have explored depth map super-resolution with a simple visual difference-based metric as the loss function. Via comparison of this metric with a variety of perceptual quality measures, we have demonstrated that it can be considered a reasonable proxy for human perception in the problem of depth super-resolution with the focus on visual quality of the 3D surface. Via an extensive evaluation of several depth-processing methods on a range of synthetic and real data, we have demonstrated that using this metric as the loss function yields significantly improved results in comparison to the common direct pixel-wise deviation of depth values. We have combined our metric with relatively simple and non-specific deep learning architectures and expect that this approach will be beneficial for other related problems.

We have focused on the case of single regularly sampled RGBD images, but a lot of geometric data has less regular form. The future work would be to adapt the developed methodology to a more general sampling of the depth values for the cases of multiple RGBD images or point clouds annotated with a collection of RGB images.

Acknowledgements

The work was supported by The Ministry of Education and Science of Russian Federation, grant No. 14.615.21.0004, grant code: RFMEFI61518X0004.

The authors acknowledge the usage of the Skoltech CDISE HPC cluster Zhores for obtaining the results presented in this paper.

[1] mega.nz/#F!yvRXBABI!pucRoBvtnthzHI1oqsxEvA!y6JmCajS

	Sphere and cylinder, x4				Lucy, x4				Cube, x4				SimGeo average, x4			
	$\mathrm{RMSE_d}$	$\mathrm{DSSIM_v}$	$\mathrm{LPIPS_v}$	$\mathrm{RMSE_v}$	$\mathrm{RMSE_d}$	$\mathrm{DSSIM_v}$	$\mathrm{LPIPS_v}$	$\mathrm{RMSE_v}$	$\mathrm{RMSE_d}$	$\mathrm{DSSIM_v}$	$\mathrm{LPIPS_v}$	$\mathrm{RMSE_v}$	$\mathrm{RMSE_d}$	$\mathrm{DSSIM_v}$	$\mathrm{LPIPS_v}$	$\mathrm{RMSE_v}$
SRfS [14]	70	887	1025	417	82	811	781	367	52	934	1036	361	61	711	869	311
EG [53]	55	143	326	130	69	357	426	220	43	113	214	105	53	168	306	136
PDN [39]	157	198	295	150	173	456	368	251	164	156	250	145	162	224	278	165
DG [13]	56	265	372	166	69	523	558	249	44	218	411	139	54	293	420	171
Bicubic	57	189	313	189	72	355	398	267	44	131	287	160	55	197	320	193
DIP [47]	46	965	1062	548	53	827	615	344	45	963	906	530	52	887	893	395
MSG [22]	41	626	859	229	54	444	480	259	29	445	687	176	39	374	569	194
DIP-V	**28**	560	766	142	**44**	421	446	223	**26**	352	613	146	**33**	313	524	147
MSG-V	99	**94**	**267**	**96**	74	**205**	**251**	**156**	102	**70**	**179**	**77**	96	**95**	**194**	**99**

Table 1: Quantitative evaluation on SimGeo dataset. $\mathrm{RMSE_d}$ is in millimeters, other metrics are in thousandths. Lower values correspond to better results. The best result is in bold, the second best is underlined.

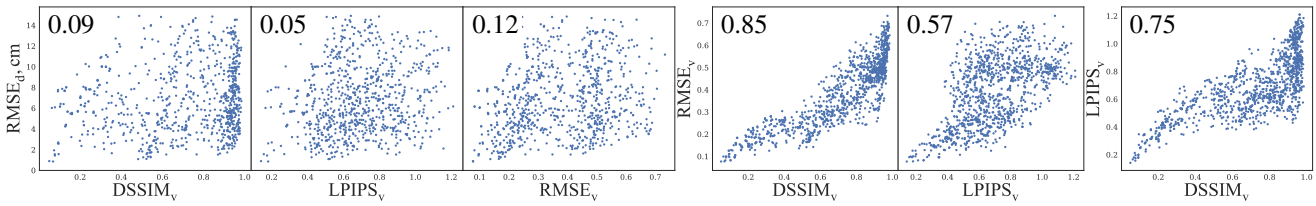

Figure 4: Scatter plots demonstrating correlation of quality measures, and the corresponding values of the Pearson correlation coefficient in the corner. Each point represents one super-resolution result.

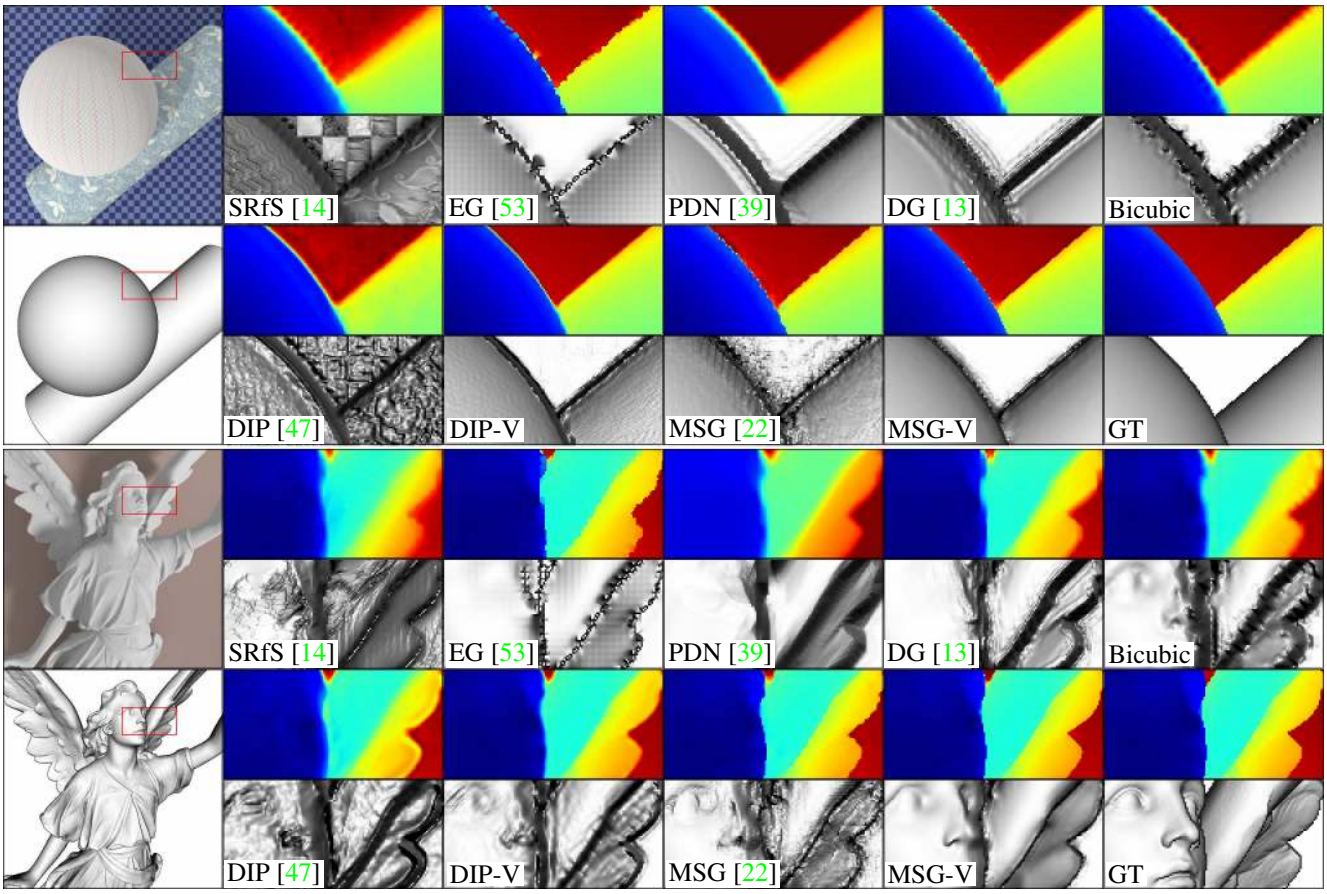

Figure 5: Super-resolution results on "Sphere and cylinder" and "Lucy" from SimGeo with the scaling factor of 4. Depth maps are in pseudo-color and depth map renderings are in grayscale. Best viewed in color.

	Plant						Vintage						Recycle						Umbrella					
	$DSSIM_v$		$LPIPS_v$		$RMSE_v$		$DSSIM_v$		$LPIPS_v$		$RMSE_v$		$DSSIM_v$		$LPIPS_v$		$RMSE_v$		$DSSIM_v$		$LPIPS_v$		$RMSE_v$	
	x4	x8	x4	x8	x4	x8	x4	x8	x4	x8	x4	x8	x4	x8	x4	x8	x4	x8	x4	x8	x4	x8	x4	x8
SRfS [14]	658	692	632	649	280	309	721	749	631	634	346	382	715	772	610	623	376	410	843	853	797	831	397	443
EG [53]	568		677		255																			
PDN [39]	574	612	659	699	269	305	663	714	706	700	319	350	635	**701**	523	589	364	457	799	**828**	847	882	367	452
DG [13]	611	622	745	785	268	291	666	669	796	840	290	300	696	719	602	617	328	383	846	878	781	856	399	457
Bicubic	562	610	688	763	249	290	558	649	602	729	258	302	**575**	721	474	576	329	398	**749**	837	747	886	323	380
DIP [47]	919	880	764	723	490	437	953	965	910	872	656	687	871	923	576	605	434	500	915	953	737	722	467	528
MSG [22]	571	645	582	**495**	234	285	708	785	**510**	**610**	292	364	741	869	624	661	485	550	834	896	678	787	442	496
DIP-V	694	707	**463**	555	262	276	804	884	579	674	343	435	**575**	735	**388**	**485**	**273**	**332**	796	854	**604**	**598**	**318**	**352**
MSG-V	**524**	**575**	639	720	**194**	**236**	**536**	**643**	670	702	**211**	**268**	603	737	520	564	368	473	778	842	800	890	348	427

Table 2: Quantitative evaluation on ICL-NUIM and Middlebury datasets. All metrics are in thousandths. Lower values correspond to better results. The best result is in bold, the second best is underlined.

Figure 6: Depth map renderings corresponding to super-resolution results on "Plant" from "ICL-NUIM" and "Vintage", "Recycle" and "Umbrella" from Middlebury datasets with the scaling factor of 4 on the left and the scaling factor of 8 on the right. Best viewed in large scale.

References

[1] Gianluca Agresti, Ludovico Minto, Giulio Marin, and Pietro Zanuttigh. Deep learning for confidence information in stereo and tof data fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 697–705, 2017. 2

[2] Piotr Bojanowski, Armand Joulin, David Lopez-Pas, and Arthur Szlam. Optimizing the latent space of generative networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 600–609, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 5

[3] Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, Oct. 2012. 6

[4] Baoliang Chen and Cheolkon Jung. Single depth image super-resolution using convolutional neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1473–1477. IEEE, 2018. 3

[5] Zhao Chen, Vijay Badrinarayanan, Gilad Drozdov, and Andrew Rabinovich. Estimating depth from rgb and sparse sensing. *CoRR*, abs/1804.02771, 2018. 2

[6] Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *European Conference on Computer Vision*, pages 108–125. Springer, Cham, 2018. 2

[7] Manri Cheon, Jun-Hyuk Kim, Jun-Ho Choi, and Jong-Seok Lee. Generative adversarial network-based image super-resolution using perceptual content losses. In Laura Leal-Taixé and Stefan Roth, editors, *ECCV Workshops*, pages 51–62, Cham, 2019. Springer International Publishing. 3

[8] Nathaniel Chodosh, Chaoyang Wang, and Simon Lucey. Deep convolutional compressed sensing for lidar depth completion. *arXiv preprint arXiv:1803.08949*, 2018. 2

[9] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in neural information processing systems*, pages 2366–2374, 2014. 2

[10] David Ferstl, Christian Reinbacher, Rene Ranftl, Matthias Rüther, and Horst Bischof. Image guided depth upsampling using anisotropic total generalized variation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 993–1000, 2013. 5

[11] David Ferstl, Matthias Ruther, and Horst Bischof. Variational depth superresolution using example-based edge representations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 513–521, 2015. 3

[12] Muhammad Waleed Gondal, Bernhard Schölkopf, and Michael Hirsch. The unreasonable effectiveness of texture transfer for single image super-resolution. In *ECCV*, pages 80–97. Springer, 2018. 3

[13] Shuhang Gu, Wangmeng Zuo, Shi Guo, Yunjin Chen, Chongyu Chen, and Lei Zhang. Learning dynamic guidance for depth image enhancement. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3769–3778, 2017. 3, 4, 5, 7, 8

[14] Bjoern Haefner, Yvain Quéau, Thomas Möllenhoff, and Daniel Cremers. Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 164–174, 2018. 3, 4, 5, 7, 8

[15] Bumsub Ham, Minsu Cho, and Jean Ponce. Robust guided image filtering using nonconvex potentials. *IEEE transactions on pattern analysis and machine intelligence*, 40(1):192–207, 2018. 3

[16] Wei Han, Shiyu Chang, Ding Liu, Mo Yu, Michael Witbrock, and Thomas S Huang. Image super-resolution via dual-state recurrent networks. In *Proc. CVPR*, 2018. 3

[17] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J. Davison. A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *IEEE Intl. Conf. on Robotics and Automation, ICRA*, Hong Kong, China, May 2014. 5

[18] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *Proc. CVPR*, 2018. 3

[19] Katrin Honauer, Ole Johannsen, Daniel Kondermann, and Bastian Goldluecke. A dataset and evaluation methodology for depth estimation on 4d light fields. In *Asian Conference on Computer Vision*, pages 19–34. Springer, 2016. 2

[20] Katrin Honauer, Lena Maier-Hein, and Daniel Kondermann. The hci stereo metrics: Geometry-aware performance analysis of stereo algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2120–2128, 2015. 2

[21] Jiashen Hua and Xiaojin Gong. A normalized convolutional neural network for guided sparse depth upsampling. In *IJCAI*, pages 2283–2290, 2018. 2

[22] Tak-Wai Hui, Chen Change Loy, and Xiaoou Tang. Depth map super-resolution by deep multi-scale guidance. In *European Conference on Computer Vision*, pages 353–369. Springer, 2016. 2, 4, 7, 8

[23] Forrest N. Iandola, Song Han, Matthew W. Moskewicz, Khalid Ashraf, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and ¡0.5mb model size. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4

[24] Zhongyu Jiang, Yonghong Hou, Huanjing Yue, Jingyu Yang, and Chunping Hou. Depth super-resolution from rgb-d pairs with transform and spatial domain regularization. *IEEE Transactions on Image Processing*, 27(5):2587–2602, 2018. 3

[25] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 3

[26] Beomjun Kim, Jean Ponce, and Bumsub Ham. Deformable Kernel Networks for Joint Image Filtering. working paper or preprint, Oct. 2018. 2

[27] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 4

[28] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. CVPR*, pages 105–114, 2017. 3

[29] Beichen Li, Yuan Zhou, Yeda Zhang, and Aihua Wang. Depth image super-resolution based on joint sparse coding. *Pattern Recognition Letters*, 2018. 3

[30] Yijun Li, Jia-Bin Huang, Narendra Ahuja, and Ming-Hsuan Yang. Deep joint image filtering. In *European Conference on Computer Vision*, pages 154–169. Springer, 2016. 2

[31] Xiaotong Luo, Rong Chen, Yuan Xie, Yanyun Qu, and Cuihua Li. Bi-gans-st for perceptual image super-resolution. In Laura Leal-Taixé and Stefan Roth, editors, *ECCV Workshops*, pages 20–34, Cham, 2019. Springer International Publishing. 3

[32] Chao Ma, Chih-Yuan Yang, Xiaokang Yang, and Ming-Hsuan Yang. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1 – 16, 2017. 3

[33] Fangchang Ma, Guilherme Venturelli Cavalheiro, and Sertac Karaman. Self-supervised sparse-to-dense: Self-supervised depth completion from lidar and monocular camera. *arXiv preprint arXiv:1807.00275*, 2018. 2

[34] Fangchang Mal and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018. 2, 4

[35] Rafat Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich. Hdr-vdp-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. In *ACM Transactions on graphics (TOG)*, volume 30, page 40. ACM, 2011. 2

[36] Roey Mechrez, Itamar Talmi, Firas Shama, and Lihi Zelnik-Manor. Learning to maintain natural image statistics. *arXiv preprint arXiv:1803.04626*, 2018. 3

[37] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017. 6

[38] Songyou Peng, Bjoern Haefner, Yvain Queau, and Daniel Cremers. Depth super-resolution meets uncalibrated photometric stereo. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2961–2968, 2017. 3

[39] David Riegler, Gernot aand Ferstl, Matthias Rüther, and Horst Bischof. A deep primal-dual network for guided depth super-resolution. In *British Machine Vision Conference*. The British Machine Vision Association, 2016. 2, 4, 5, 7, 8

[40] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German Conference on Pattern Recognition*, pages 31–42. Springer, 2014. 5

[41] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 4

[42] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 5

[43] Xibin Song, Yuchao Dai, and Xueying Qin. Deep depth super-resolution: Learning depth super-resolution using deep convolutional neural network. In *Asian Conference on Computer Vision*, pages 360–376. Springer, 2016. 2, 3

[44] Xibin Song, Yuchao Dai, and Xueying Qin. Deeply supervised depth map super-resolution as novel view synthesis. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018. 3

[45] Atsuhiko Tsuchiya, Daisuko Sugimura, and Takayuki Hamamoto. Depth upsampling by depth prediction. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1662–1666, Sept 2017. 2

[46] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *IEEE International Conference on 3D Vision (3DV)*, 2017. 2

[47] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4, 5, 7, 8

[48] Thang Vu, Tung M. Luu, and Chang D. Yoo. Perception-enhanced image super-resolution via relativistic generative adversarial networks. In Laura Leal-Taixé and Stefan Roth, editors, *ECCV 2018 Workshops*, pages 98–113, Cham, 2019. Springer International Publishing. 3

[49] Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proc. CVPR*, June 2018. 3

[50] Yifan Wang, Federico Perazzi, Brian McWilliams, Alexander Sorkine-Hornung, Olga Sorkine-Hornung, and Christopher Schroers. A fully progressive approach to single-image super-resolution. In *CVPR Workshops*, June 2018. 3

[51] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2, 4

[52] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. IEEE, 2003. 2

[53] Jun Xie, Rogerio Schmidt Feris, and Ming-Ting Sun. Edge-guided single depth image super resolution. *IEEE Transactions on Image Processing*, 25(1):428–438, 2016. 4, 5, 7, 8

[54] Shi Yan, Chenglei Wu, Lizhen Wang, Feng Xu, Liang An, Kaiwen Guo, and Yebin Liu. Ddrnet: Depth map denoising and refinement for consumer depth cameras using cascaded cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 151–167, 2018. 2, 4

[55] Jingyu Yang, Xinchen Ye, Kun Li, Chunping Hou, and Yao Wang. Color-guided depth recovery from rgb-d data using an

adaptive autoregressive model. *IEEE transactions on image processing*, 23(8):3443–3458, 2014. 3

[56] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: a feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, 20(8):2378–2386, 2011. 2

[57] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 3, 4

[58] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–301, 2018. 3

[59] Yulun Zhang, Yapeng Tian, Yu Kong, Bineng Zhong, and Yun Fu. Residual dense network for image super-resolution. In *Proc. CVPR*, June 2018. 3

[60] Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017. 4

[61] Lijun Zhao, Huihui Bai, Jie Liang, Bing Zeng, Anhong Wang, and Yao Zhao. Simultaneously color-depth super-resolution with conditional generative adversarial network. *arXiv preprint arXiv:1708.09105*, 2017. 3

[62] Yifan Zuo, Qiang Wu, Jian Zhang, and Ping An. Minimum spanning forest with embedded edge inconsistency measurement model for guided depth map enhancement. *IEEE Transactions on Image Processing*, 27(8):4145–4159, 2018. 3