Photometric Depth Super-Resolution – Supplementary Material

Bjoern Haefner*, Songyou Peng*, Alok Verma*, Yvain Quéau, and Daniel Cremers

Abstract—This supplementary material explores the three proposed approaches of the paper "Photometric Depth Super-Resolution" in more detail towards experimental evaluation on self-generated and publicly available synthetic and real-world datasets. Also, a more thorough discussion on the theoretical aspects of the RGB image formation model and the problems arising in depth super-resolution and shape-from-shading are drawn. A unified comparison on a publicly available photometric stereo benchmark eventually highlights the pros and cons of each proposed method.

Index Terms—RGB-D cameras, depth super-resolution, shape-from-shading, photometric stereo, variational methods, deep learning.

1 Organization of the document

This document is structured as follows. Section 2 contains general comments on photometric 3D-reconstruction and depth super-resolution: the derivation of the RGB image formation model used through the paper, a visual description of the ambiguities arising in depth super-resolution and in shape-from-shading, and some general information regarding the reflectance learning-based approach. The rest of the document is devoted to the individual experimental evaluation of each of the proposed methods: Section 3 contains the shape-from-shading experiments, Section 4 the reflectance learning ones, and Section 5 evaluates the uncalibrated photometric stereo-based approach. Section 6 eventually concludes the document by presenting a unified comparison of the results obtained with the three proposed methods.

2 Generalities

2.1 Derivation of the RGB image formation model

This subsection is devoted to the derivation of the RGB image formation model (Eq. (3) in the main paper), which relates the irradiance measurements and the surface normals. The following derivation is adapted from [1, Sect. 2.2], with an extension of the model to RGB images and spherical harmonics lighting.

We first assume that the surface is Lambertian, i.e. its appearance is independent from the viewing angle. A consequence of this assumption is that the surface's reflectance ρ at a surface point is a simple scalar quantity called the albedo, which is independent from the incident light direction.

Next, we assume that the surface is lit by a single, infinitely distant light source represented by a direction ω on the visible hemisphere. The spectral radiance at a surface point is thus given by

$$L(\lambda,\omega) = \phi(\lambda,\omega)\,\frac{\rho(\lambda)}{\pi}\,\max\{0,\mathbf{s}(\omega)^\top \mathbf{n}_{z,\nabla z}\}, \qquad (1)$$

with λ the wavelength, $\phi(\cdot,\omega)$ the spectrum of the source associated with direction ω, $\rho(\cdot)$ the spectral reflectance of the surface point, $\mathbf{s}(\omega)$ the unit-length vector pointing towards the light source associated with direction ω, and $\mathbf{n}_{z,\nabla z}$ the outer unit-length surface normal.

Now, let us assume that the surface is observed under natural illumination, rather than lit by one single light source. Let us represent natural illumination by a collection of infinitely distant point light sources, each of them being represented by a direction ω. The total spectral radiance of a surface point is obtained by summing the individual contributions from each source, i.e. by integrating (1) over the visible hemisphere:

$$L(\lambda) = \frac{\rho(\lambda)}{\pi}\int_{\mathbb{S}^2} \phi(\lambda,\omega)\,\max\{0,\mathbf{s}(\omega)^\top \mathbf{n}_{z,\nabla z}\}\,\mathrm{d}\omega. \qquad (2)$$

We further assume that the sensor's response is linear, and that the RGB camera is focused on the surface. Then, the sensor's spectral irradiance, in the pixel conjugate to the surface point, is given by

$$E(\lambda) = \beta\cos^4\alpha\, L(\lambda), \qquad (3)$$

where β depends on the sensor's aperture and magnification, and where α is the angle between the viewing angle and the optical axis (the $\cos^4\alpha$ factor is thus responsible for darkening at the periphery of images).

The intensity recorded by the camera in channel $\star$, $\star \in \{R,G,B\}$, is proportional to the sum of all spectral sensor's irradiances, weighted by the camera's transmission spectrum. Denoting by γ this proportionality coefficient, this writes as

$$I_\star = \gamma\int_{\mathbb{R}^+} c_\star(\lambda)E(\lambda)\,\mathrm{d}\lambda, \qquad (4)$$

with $c_\star(\lambda)$ the transmission spectrum of camera's channel $\star$.

We further assume that all the light sources are achromatic, i.e. that

$$\phi(\lambda,\omega) = \phi(\omega) \qquad (5)$$

(this assumption implies that color will be interpreted in terms of surface's reflectance by our algorithms, rather than in terms of lighting).

* *Those authors contributed equally*

Plugging Equations (2), (3) and (5) into (4) yields

$$I_\star = \rho_\star \int_{\mathbb{S}^2} \phi(\omega) \max\{0, \mathbf{s}(\omega)^\top \mathbf{n}_{z,\nabla z}\} \, \mathrm{d}\omega, \tag{6}$$

with

$$\rho_\star := \frac{\gamma\beta\cos^4\alpha}{\pi} \int_{\mathbb{R}^+} c_\star(\lambda)\rho(\lambda) \, \mathrm{d}\lambda \tag{7}$$

the "albedo", relatively to channel $\star$ (note that $\rho_\star$ does not characterize the surface, since it depends upon the sensor's response, its aperture and magnification, etc.).

Next, we approximate the integral in (6) using spherical harmonics [2], [3]. In this work we consider the first-order case, which already captures more than 85% of natural illumination [4], and leave the extension to second-order spherical harmonics as future work. The spherical harmonics approximation reads

$$\int_{\mathbb{S}^2} \phi(\omega) \max\{0, \mathbf{s}(\omega)^\top \mathbf{n}_{z,\nabla z}\} \, \mathrm{d}\omega \approx \mathbf{l}^\top \mathbf{m}_{z,\nabla z} \tag{8}$$

with $\mathbf{l} \in \mathbb{R}^4$ the achromatic "light vector" (which is the same for all pixels), and

$$\mathbf{m}_{z,\nabla z} := \begin{bmatrix} \mathbf{n}_{z,\nabla z} \\ 1 \end{bmatrix} \tag{9}$$

a geometric vector depending upon the surface normals.

Plugging (8) into (6), we obtain

$$I_\star = \rho_\star \, \mathbf{l}^\top \mathbf{m}_{z,\nabla z}, \quad \star \in \{R, G, B\}. \tag{10}$$

Denoting

$$\mathbf{I} := \begin{bmatrix} I_R \\ I_G \\ I_B \end{bmatrix} \quad \text{and} \quad \boldsymbol{\rho} := \begin{bmatrix} \rho_R \\ \rho_G \\ \rho_B \end{bmatrix}, \tag{11}$$

and assuming that (10) is satisfied up to additive noise, we eventually obtain the RGB image formation model (Eq. (3) in the paper) by plugging together the three equations in (10):

$$\mathbf{I} = \mathbf{l}^\top \mathbf{m}_{z,\nabla z} \, \boldsymbol{\rho} + \boldsymbol{\eta}_\mathbf{I}, \tag{12}$$

with $\boldsymbol{\eta}_\mathbf{I}$ the realisation of a stochastic process.

2.2 Ambiguities in Depth Super-resolution and Shape-from-shading

This subsection illustrates the ambiguities arising in depth super-resolution and in photometric 3D-reconstruction, in order to visually motivate the choice of their joint solving. As can be seen in Figure 1, in super-resolution high-frequency geometric clues are missing and thus there exist infinitely many ways to interpolate between low-resolution samples. On the contrary, shape-from-shading suffers from the concave / convex ambiguity: though the surface orientation is unambiguous in critical points (arrows in Figure 2), two such singular points may be connected either by "going up" or by "going down". Therefore, it seems reasonable to rely on high-frequency photometric clues to disambiguate depth super-resolution, and on low-frequency geometric clues to disambiguate photometric 3D-reconstruction.

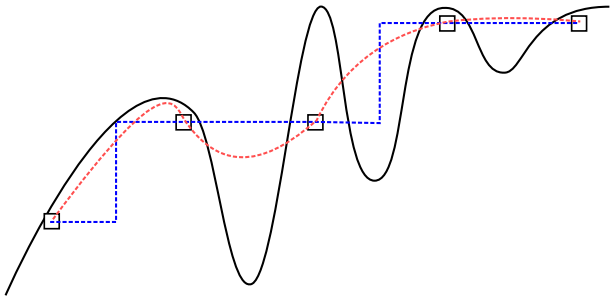

Fig. 1: There exist infinitely many ways (dashed lines) to interpolate between low-resolution depth samples (rectangles). Our disambiguation strategy builds upon shape-from-shading applied to the companion high-resolution color image (cf. Figure 2), in order to resurrect the fine-scale geometric details of the genuine surface (solid line).

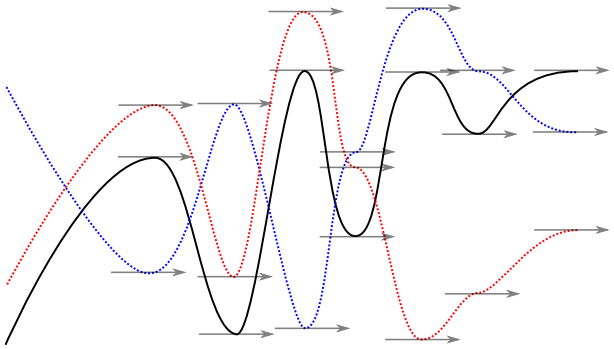

Fig. 2: Shape-from-shading suffers from the concave / convex ambiguity: the genuine surface (solid line) and both the surfaces depicted by dashed lines produce the same image, if lit and viewed from above. We put forward low-resolution depth clues (cf. Figure 1) for disambiguation.

2.3 Generalities on Reflectance Learning-based Depth Super-resolution

We now illustrate the creation of the training dataset and the network's architecture, and justify why we focused on a particular class of objects in the learning-based approach.

Figure 3 illustrates the generation of training data. We consider ground truth geometry and reflectance of various human faces from the ICT-3DRFE database [5]. A rendering software is used to generate multiple images of these faces under different viewing and lighting scenarios. Lighting variations are created by turning off and on several extended sources, emulating usual indoor lighting conditions.

Figure 4 illustrates the architecture of the neural network. It is a U-Net architecture comprising an initial convolution layer of kernel size 4, stride 2 and padding 1; after which there are repeated blocks of 8 ReLU-Conv-BatchNorm layers. This results in downsampling of a 512x512 resolution image to a 1x512 vector at the bottleneck of the "U". Then, the 1D array is upsampled to input resolution with multiple ReLU-Transpose Convolution-BatchNorm layers. Dropout is also used in a few layers to allow for randomness while learning the mapping from input images to albedo maps. Finally, the L1 loss is considered, which favors sharper output compared to the L2 loss.

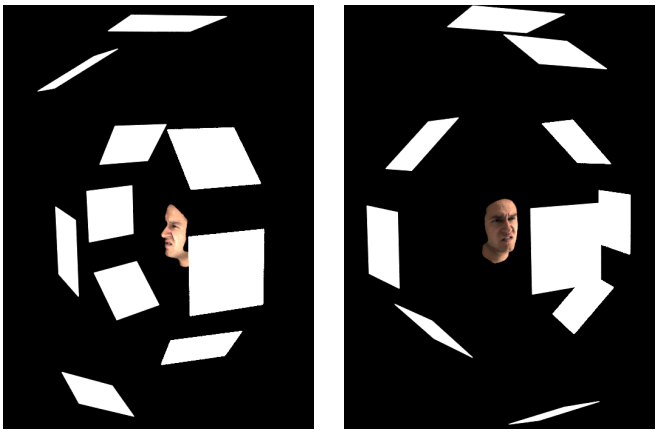

Fig. 3: Rendering of synthetic faces for generating training data. The white planes represent switchable extended light sources, which are independently controlled to create multiple illumination conditions. Multiple images can then be captured under different illumination and viewing angles.

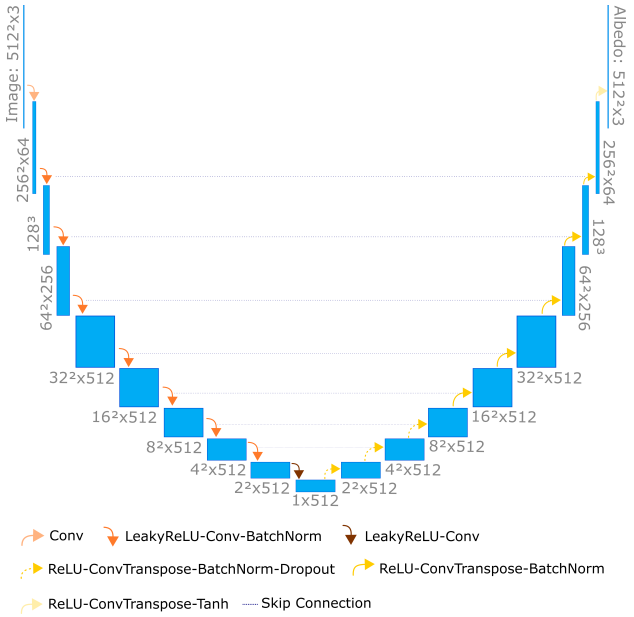

Fig. 4: The U-Net Architecture used for albedo estimation. The top two layers are the input and output, respectively. The arrows' color represent the operations of the other hidden layers. Skip connections (dotted lines) concatenate the left and right layers.

Eventually, Figure 5 illustrates the lack of inter-class generalisation which is inherent to learning-based methods. For instance, the approach of [6] (trained on Sintel [7] and MIT [8] datasets) performs well on the MIT object but poorly on the ShapeNet car image, because such an object was not present in the learning database. For the same reason, the alternative approach of [9] (trained on ShapeNet objects [10]) performs well on the ShapeNet car but fails on the MIT object, and both approaches fail on the face image since the latter resembles none of the training data. Due to this lack of inter-class generalisation, we choose to focus in our approach on the specific class of human faces.

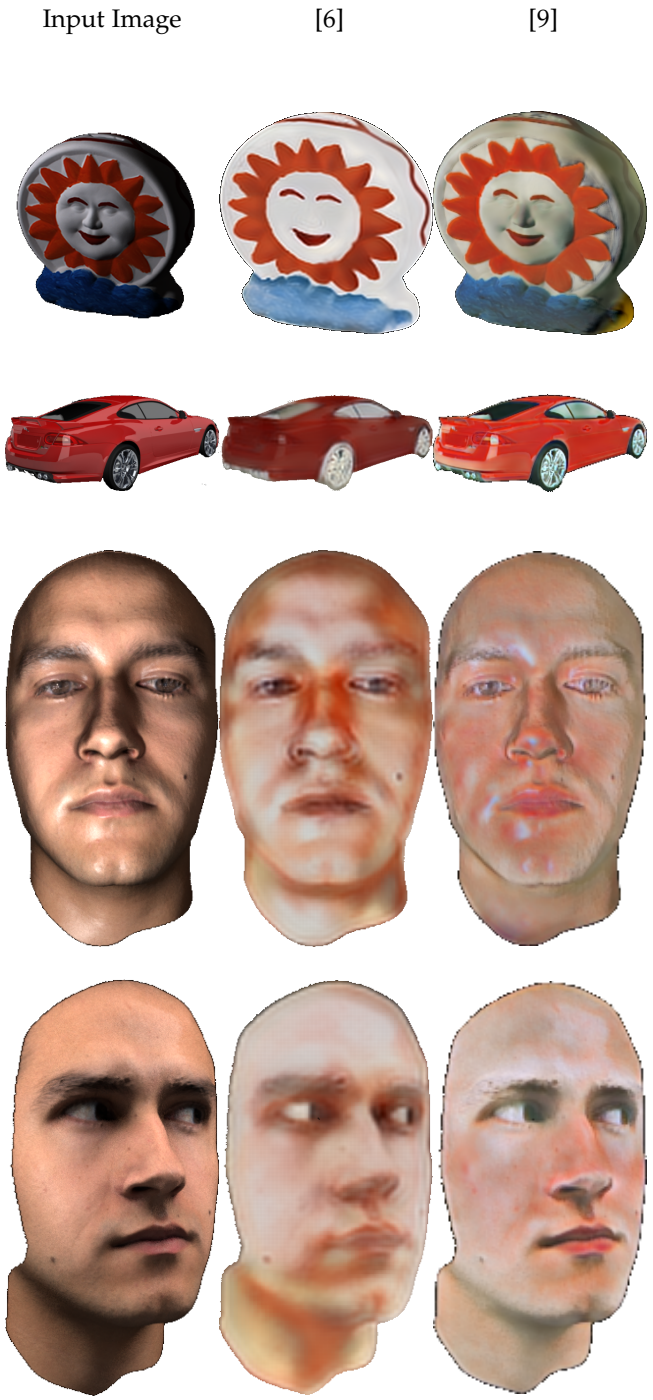

Fig. 5: Learning-based albedo estimation applied to an object from the MIT database (first row), a car from the ShapeNet dataset (second row), and two images of human faces we generated with a renderer using the ICT-3DRFE database [5]. This illustrates the lack of inter-class generalisation inherent to learning-based techniques: the approach from [6], trained on the MIT dataset, fails on the ShapeNet car and on faces, and the one from [9], trained on the ShapeNet dataset, fails on the MIT object and on faces: in both cases albedo estimation is not satisfactory since the objects do not resemble the training data.

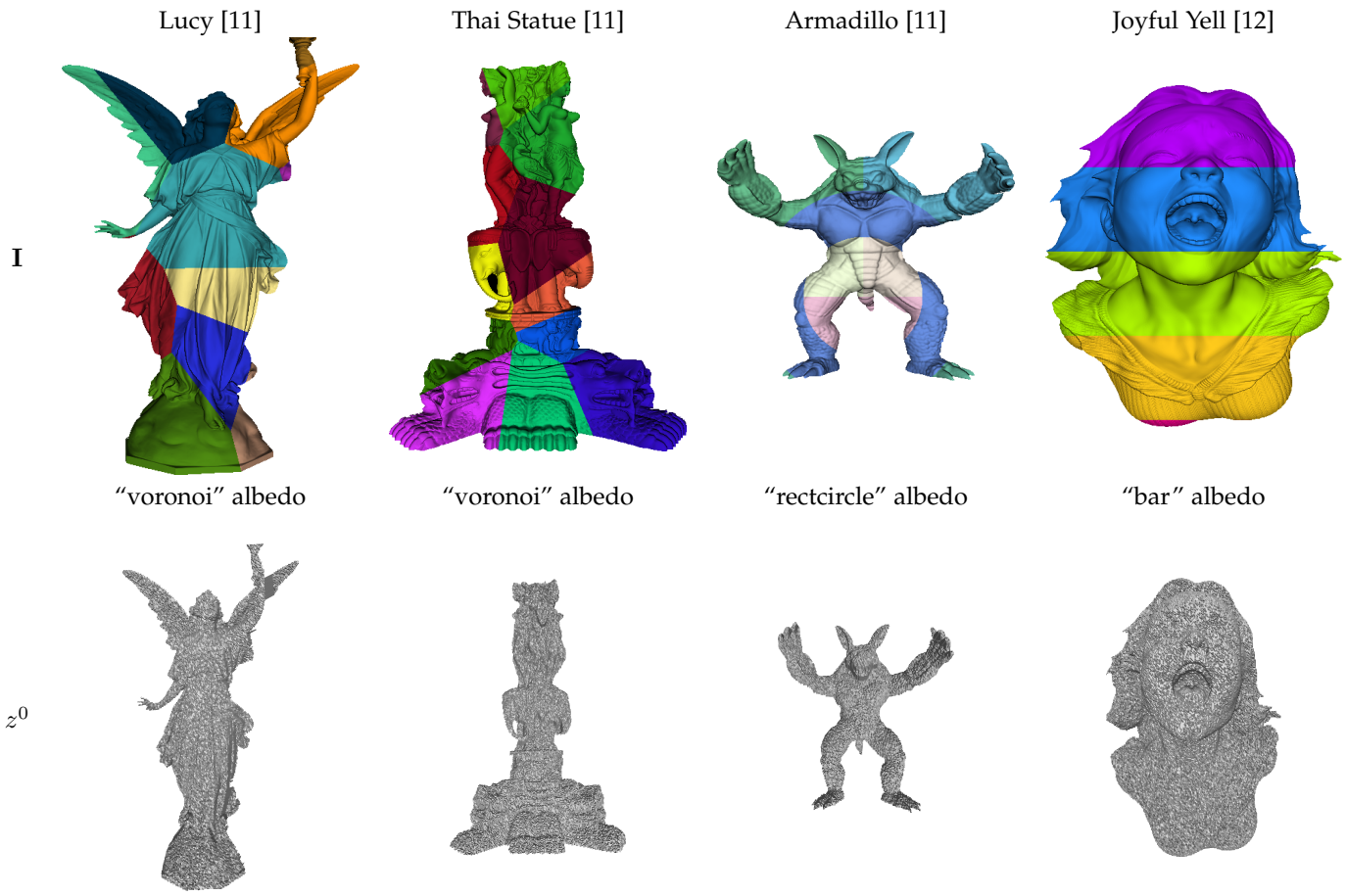

Fig. 6: Illustration of synthetic data used for evaluation of the single-shot approach based on shape-from-shading. High-resolution RGB images $\mathbf{I}$, of size 480×640, are generated using high-resolution ground truth depth and reflectance maps, and adding noise. Low-resolution depth maps z^0 are created by downsampling the ground truth depth maps with scaling factors of 8, 4 and 2 (the second row shows the low-resolution depth maps with a scaling factor of 2), and adding noise.

3 Evaluation of the Single-shot Approach based on Shape-from-shading

3.1 Creation of the Synthetic Data

Figure 6 illustrates the synthetic data used for evaluation, which is generated using four different 3D-shapes ("Lucy", "Thai Statue", "Armadillo" and "Joyful Yell"), each of them rendered using three different albedo maps ("voronoi", "rectcircle" and "bar") and three different scaling factors (2, 4 and 8) for the low-resolution depth image. To this end, 3D-meshes are rendered into high-resolution ground truth depth maps of size 480×640, which are then downsampled. Then, additive zero-mean Gaussian noise with standard deviation 10^{-4} times the squared original depth value (consistently with real-world measurements from [13]) is added to the low-resolution depth maps, which are eventually quantised. High-resolution RGB images are rendered from the ground truth depth map using the first-order spherical harmonics model with $\mathbf{l} = [0, 0, -1, 0.2]^\top$ using the three different high-resolution reflectance maps, and an additive zero-mean Gaussian noise with standard deviation 1% the maximum intensity is eventually added to the RGB images.

3.2 Tuning the Hyper-parameters

In Figure 7, we use the "Joyful Yell" dataset from Figure 6 in order to determine appropriate values for the hyper-parameters (μ, ν, λ). For quantitative evaluation, we consider the root mean squared error (RMSE) on the estimated depth and reflectance maps, and the mean angular error (MAE) on surface normals. To select an appropriate set of values for them, we initially set $\mu = 0.5$, $\nu = 0.01$ and $\lambda = 1$. We then evaluate the impact of each parameter by varying it while keeping the remaining two fixed. As could be expected, large values of μ force the depth map to keep close to the noisy input, while small values make the depth prior less important so not capable of disambiguating shape-from-shading. Inbetween, the range $\mu \in [10^{-1}, 10]$ seems to provide appropriate results. As for ν, large values produce over-smoothed results and small ones result in slightly noisier depth estimates, although the albedo estimate seems unaffected by this choice. Overall, the range $\nu \in [0.5, 10^2]$ seems appropriate. The parameter λ strongly impacts both the resulting albedo and depth: too small (resp., high) values for λ result in over (resp., under)-segmentation problems, and in both cases shading information gets propagated to the albedo. We found $\lambda \in [10^{-1}, 10]$ to be a reasonable choice. Overall, we opted for $(\mu, \nu, \lambda) = (0.1, 0.7, 1)$.

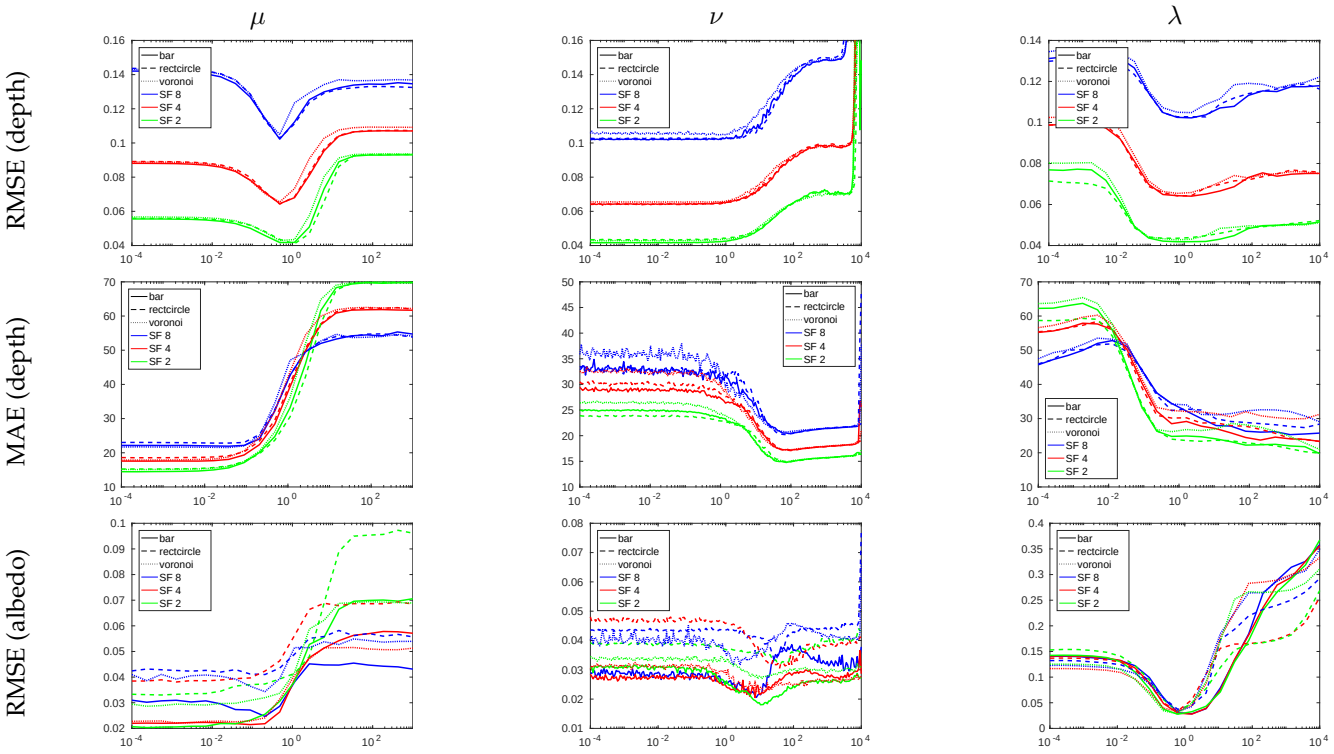

Fig. 7: Impact of the parameters (μ, ν, λ) on the accuracy of the albedo and depth estimates. The accuracy of the albedo is evaluated by the root mean square error (RMSE), and that of the depth by the RMSE and the mean angular error (MAE). Based on these experiments, the set of hyper-parameters $(\mu, \nu, \lambda) = (0.1, 0.7, 1)$ is selected.

3.3 Comparison against the State-of-the-art on the Synthetic Dataset

Next, we compare the results obtained by our single-shot approach against the state-of-the-art, on the synthetic dataset from Figure 6. We consider two alternative depth super-resolution methods: the image-based one from [14], and the learning-based one from [15] (since the authors only provide trained data for a factor of 4, this method was evaluated only for this factor). To emphasise the interest of joint shape-from-shading and depth super-resolution over shading-based depth refinement using downsampled images, we also consider [16]. Qualitative results are presented in Figure 8, and quantitative ones in Table 1. As can be seen, our method systematically overcomes the competitors in terms of MAE, which indicates that high-frequency geometric details are better recovered. The RMSE on depth rather evaluates the overall (low-frequency) fit to ground truth, and for this metric our results are comparable with [14], which achieves the best results.

Interestingly, for scaling factors of 4 and 8, our approach seems less accurate than [14] in terms of RMSE. However, Figure 8 clearly shows that our results are significantly better: we thus believe that only the order of magnitude of the RMSE is meaningful, yet comparison using this metric might not really indicate which method is the best, and MAE should be preferred for this purpose. A more thorough discussion on the relevance of RMSE for evaluation can be found in [17].

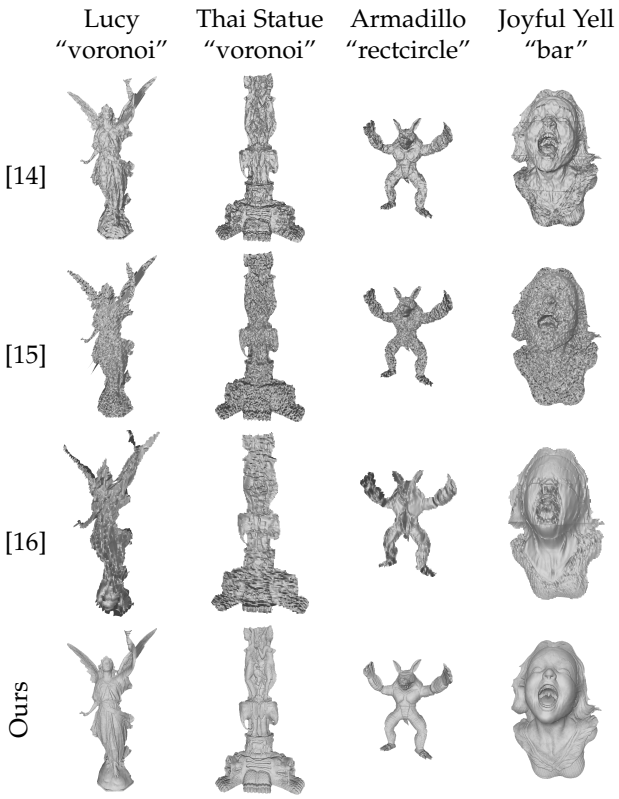

Fig. 8: Qualitative comparison between our single-shot results and state-of-the-art's ones (the scaling factor is 4).

Albedo	3D-shape	SF	[14]		[15]		[16]		Ours	
			RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
bar	Armadillo	2	0.043643	38.6274	–	–	0.41993	67.2643	**0.034655**	**16.7496**
		4	**0.051558**	42.2277	0.17865	45.6972	0.45139	66.2117	0.054679	**19.0314**
		8	**0.072466**	43.5649	–	–	0.58837	69.3262	0.091263	**20.8836**
	Joyful Yell	2	0.05089	29.1719	–	–	0.1721	47.4836	**0.050694**	**16.7414**
		4	**0.066517**	33.0843	0.084094	42.611	0.22867	32.9784	0.079271	**19.0695**
		8	**0.10212**	36.565	–	–	0.37923	31.2894	0.128	**21.9886**
	Lucy	2	0.057987	39.4714	–	–	0.21309	66.5525	**0.053989**	**25.0955**
		4	**0.068502**	42.7169	0.50472	47.605	0.34091	69.2566	0.081005	**28.3044**
		8	**0.098713**	46.4775	–	–	0.43619	59.5434	0.1195	**30.1058**
	Thai Statue	2	0.040821	42.8976	–	–	0.12948	63.06	**0.035736**	**23.9147**
		4	**0.050296**	47.1017	0.22363	49.9553	0.15489	54.6139	0.057313	**28.492**
		8	**0.066515**	49.8604	–	–	0.22835	56.4247	0.087054	**31.65**
rectcircle	Armadillo	2	0.044026	39.108	–	–	0.34323	70.8526	**0.03494**	**18.4909**
		4	**0.052115**	43.3175	0.17782	45.6324	0.2338	50.6919	0.056727	**18.8487**
		8	**0.069467**	45.4735	–	–	0.61917	70.9363	0.09155	**21.9959**
	Joyful Yell	2	**0.051296**	30.7886	–	–	0.14841	41.5424	0.05226	**17.134**
		4	**0.066911**	33.3	0.10328	42.7531	0.28311	51.0665	0.080387	**19.8717**
		8	**0.10201**	36.2961	–	–	0.39518	35.4817	0.1281	**22.8027**
	Lucy	2	0.058495	39.7374	–	–	0.19546	64.8212	**0.054383**	**24.8427**
		4	**0.069893**	43.9016	0.50464	48.1068	0.23235	53.2901	0.082547	**28.7517**
		8	**0.099402**	46.3739	–	–	0.39583	64.3269	0.12283	**29.1531**
	Thai Statue	2	0.039821	40.6144	–	–	0.11355	58.2254	**0.036845**	**23.9036**
		4	**0.04973**	46.1154	0.20894	49.4124	0.16749	52.9663	0.05866	**28.155**
		8	**0.067799**	50.6515	–	–	0.21058	50.9074	0.094688	**33.5308**
voronoi	Armadillo	2	0.043635	38.9089	–	–	0.33005	69.3157	**0.034751**	**17.6873**
		4	**0.051989**	41.57	0.17182	45.5833	0.4407	65.5811	0.056032	**20.168**
		8	**0.07077**	43.1987	–	–	0.50548	63.8618	0.090708	**22.2767**
	Joyful Yell	2	**0.052002**	28.7903	–	–	0.16893	47.72	0.052429	**17.0453**
		4	**0.066557**	32.3448	0.086394	43.1744	0.24753	39.6569	0.079888	**19.6512**
		8	**0.10238**	35.8017	–	–	0.47694	47.4707	0.12916	**21.6663**
	Lucy	2	0.058222	36.2327	–	–	0.29164	72.9002	**0.054442**	**26.1333**
		4	**0.068253**	40.8878	0.5066	48.0387	0.32955	71.1042	0.079877	**28.4506**
		8	**0.099838**	43.7671	–	–	0.37839	57.6856	0.11877	**29.6331**
	Thai Statue	2	0.039872	39.6508	–	–	0.13261	65.8352	**0.037607**	**25.6126**
		4	**0.049783**	45.7178	0.22688	49.4132	0.16533	58.3933	0.058957	**28.6314**
		8	**0.065577**	48.7962	–	–	0.21927	49.6711	0.091959	**32.0347**
Median		2	0.047458	39.0085	–	–	0.18378	65.3282	**0.044151**	**21.1973**
		4	**0.059316**	42.4723	0.19379	46.6511	0.24067	53.952	0.069114	**24.1615**
		8	**0.085589**	44.6203	–	–	0.3955	57.0551	0.10673	**25.9779**
Mean		2	0.048392	36.9999	–	–	0.22154	61.2978	**0.044394**	**21.1126**
		4	**0.059342**	41.0238	0.24812	46.4986	0.27298	55.4842	0.068779	**23.9521**
		8	**0.084754**	43.9022	–	–	0.40275	54.7438	0.1078	**26.4768**

TABLE 1: Quantitative comparison between our single-shot results and three state-of-the-art methods, on all the synthetic datasets. Our results are always superior in terms of mean angular error (MAE) and in terms of root mean square error (RMSE) when the scaling factor is 2. For larger synthetic factors our RMSE values are slightly higher than those from [14], but Figure 8 shows that our results are actually of better quality than the latter, so the RMSE values might not be as relevant as the MAE ones.

3.4 Comparison against the State-of-the-art on a Public Real-world Dataset

In Figure 9, we qualitatively compare our single-shot results against the state-of-the-art, using the real-world DiLiGenT photometric stereo dataset [18] (only one out the 96 images of each object was used). To create noisy low-resolution input depths with a scaling factor of 2, 4 and 8, the ground truth depth is downsampled and Gaussian noise is then added, as in the previous subsection.

On objects which match our assumption of a Lambertian surface with piecewise-constant albedo (e.g., "bear" and "pot1"), we obtain very satisfactory results. However, the strong dependency of our approach on the piecewise-constant albedo assumption is clearly visible in the "cat" results, which are not as satisfactory: the dark structures in the image are too thin to be appropriately interpreted as piecewise-constant albedo areas and this creates artifacts in the geometry.

Besides, the "cow", "pot2" and "reading" results demonstrate that our approach also strongly depends upon the Lambertian assumption: the specular highlights in the images get propagated into the estimated depth. A natural future extension of our method would thus be to cope with such non-Lambertian effects, either by resorting to robust estimation techniques [19], or by adapting our approach to a non-Lambertian image formation model [20].

Nevertheless, and despite these important limitations, our results remain qualitatively superior to those of the state-of-the-art in all the experiments. This can also be observed in the quantitative evaluation of Table 2, which confirms the conclusions of the synthetic quantitative evaluation from Table 1.

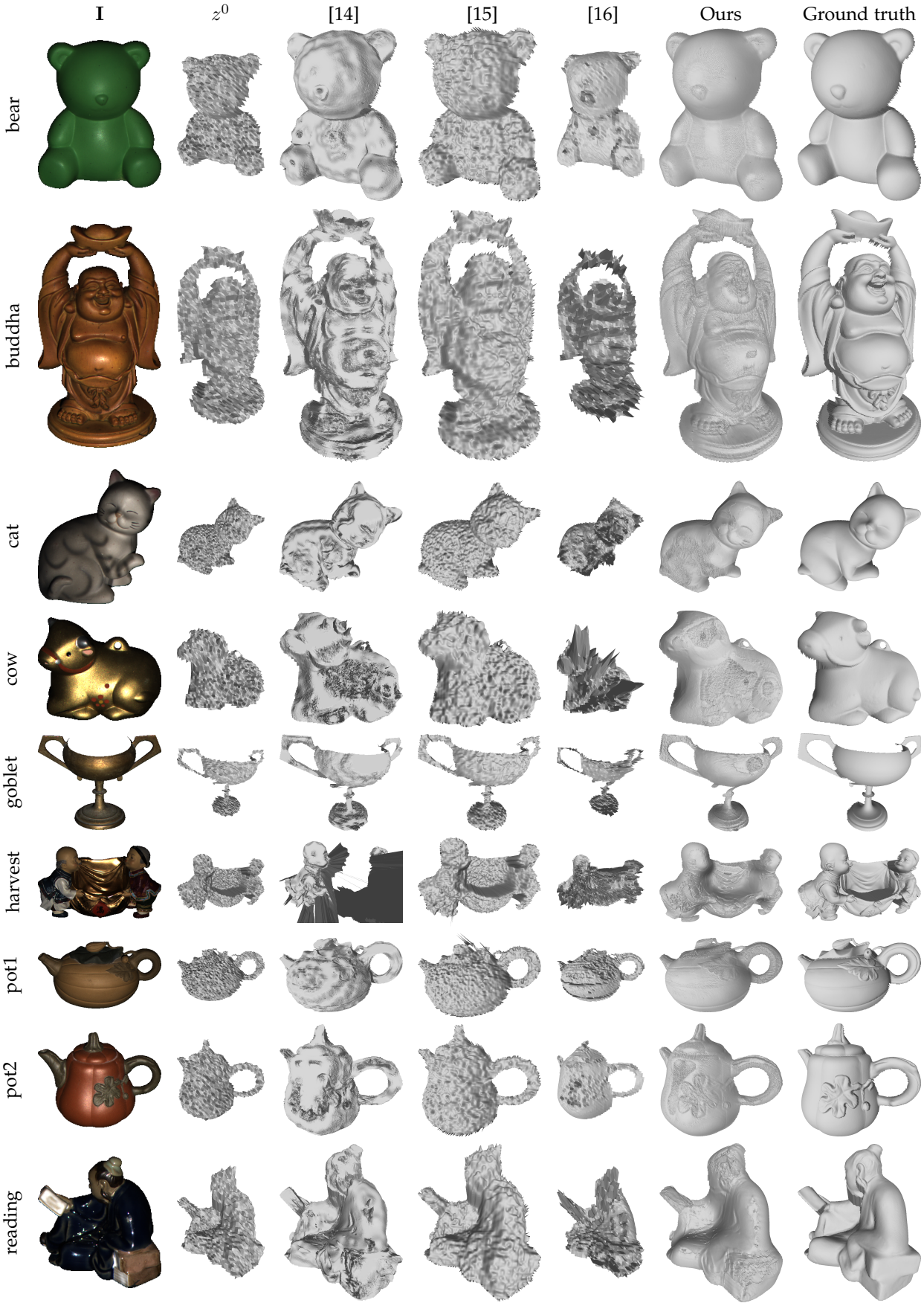

Fig. 9: Qualitative comparison between our single-shot results and those from the state-of-the-art, on the DiLiGenT dataset [18] (the scaling factor is 4). Our approach outperforms the state-of-the-art in all the experiments.

3D-shape	SF	[14]		[15]		[16]		Ours	
		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
bear	2	0.0066575	17.2655	–	–	0.014616	27.9357	**0.0047136**	**12.8781**
	4	**0.0065535**	19.5072	0.8825	31.5392	0.028849	31.8918	0.0085904	**14.8113**
	8	0.97126	76.4581	–	–	0.055159	31.0276	**0.018022**	**20.341**
buddha	2	0.0099968	37.3338	–	–	0.02972	69.0274	**0.0080152**	**26.6017**
	4	**0.0099935**	39.1319	0.86352	36.8237	0.038584	68.6713	0.012027	**31.0774**
	8	1.3683	71.2403	–	–	0.047353	57.6881	**0.019676**	**39.0075**
cat	2	0.0085294	23.3362	–	–	0.028382	44.6708	**0.0084811**	**18.8204**
	4	**0.0096136**	27.826	0.80869	30.7428	0.042872	54.1746	0.01353	**21.4786**
	8	**0.015137**	30.8242	–	–	0.065853	53.4602	0.023393	**25.3616**
cow	2	0.0086552	32.7633	–	–	0.037772	59.2638	**0.0049385**	**14.806**
	4	0.0090334	33.8093	0.84557	33.7576	0.055621	55.3108	**0.0089681**	**16.9767**
	8	**0.010392**	31.6684	–	–	0.059261	53.5979	0.017596	**21.03**
goblet	2	**0.01019**	30.2473	–	–	0.032588	59.1553	0.011007	**23.0414**
	4	**0.011121**	31.1036	1.3435	34.0517	0.048727	56.7471	0.017208	**24.2692**
	8	**0.015451**	36.2801	–	–	0.084675	51.7091	0.031125	**25.7217**
harvest	2	**0.014169**	33.9026	–	–	0.041792	66.3635	0.01594	**31.1557**
	4	2.651	63.9349	0.75973	37.0383	0.05696	66.5893	**0.023588**	**33.6957**
	8	115.5837	79.2204	–	–	0.074651	50.9501	**0.037176**	**35.9762**
pot1	2	0.0077563	22.6961	–	–	0.020767	48.3748	**0.007147**	**16.9523**
	4	**0.0086358**	26.2298	0.72979	31.8426	0.03114	39.7103	0.010863	**17.6975**
	8	**0.013278**	29.6214	–	–	0.05537	38.9525	0.019307	**19.9866**
pot2	2	0.0081729	28.8295	–	–	0.021455	50.4214	**0.0055283**	**18.0749**
	4	0.0088839	32.7579	0.90388	33.4448	0.028528	28.5455	**0.0088442**	**19.2421**
	8	**0.014079**	35.288	–	–	0.054661	47.9005	0.01623	**22.4169**
reading	2	0.011767	28.7648	–	–	0.030566	53.4663	**0.0097283**	**19.2611**
	4	**0.011428**	30.4347	0.93384	31.764	0.047677	53.7065	0.015536	**22.91**
	8	**0.01607**	32.2913	–	–	0.071794	52.5448	0.028808	**29.0107**
Median	2	0.0086552	28.8295	–	–	0.02972	53.4663	**0.0080152**	**18.8204**
	4	**0.0096136**	31.1036	0.86352	33.4448	0.042872	54.1746	0.012027	**21.4786**
	8	**0.015451**	35.288	–	–	0.059261	51.7091	0.019676	**25.3616**
Mean	2	0.0095439	28.3488	–	–	0.028629	53.1865	**0.0083887**	**20.1769**
	4	0.30292	33.8595	0.89678	33.4449	0.042106	50.5941	**0.013239**	**22.4621**
	8	13.112	46.988	–	–	0.063197	48.6479	**0.023481**	**26.5391**

TABLE 2: Quantitative comparison between our single-shot results and those from the state-of-the-art, on the DiliGenT dataset [18]. Our approach systematically outperforms the state-of-the-art, consistently with the conclusions from the synthetic experiments drawn in Table 1.

3.5 Comparison against State-of-the-art Multi-view Techniques on Publicly Available Real-world Datasets

Figure 10 shows four qualitative comparisons with state-of-the-art multi-view approaches on publicly available datasets. The "Augustus", "Lucy" and "Relief" datasets [21] were created using a PrimeSense camera, whereas "Gate" [22] was acquired using a Structure Sensor for the iPad. The scaling factor for "Augustus", "Relief" and "Gate" is 2, whereas it is 1 for "Lucy" (in this case, our approach only performs shading-based depth refinement without super-resolution). Although our approach needs significantly less data (a single RGB-D image) compared to multi-view approaches, we are still able to recover fine geometry close to the degree of detail of [21], [23]. Even under more complex lighting, as for instance in the "Gate" experiment, our approach can result in high-resolution depth maps with fine-scale details.

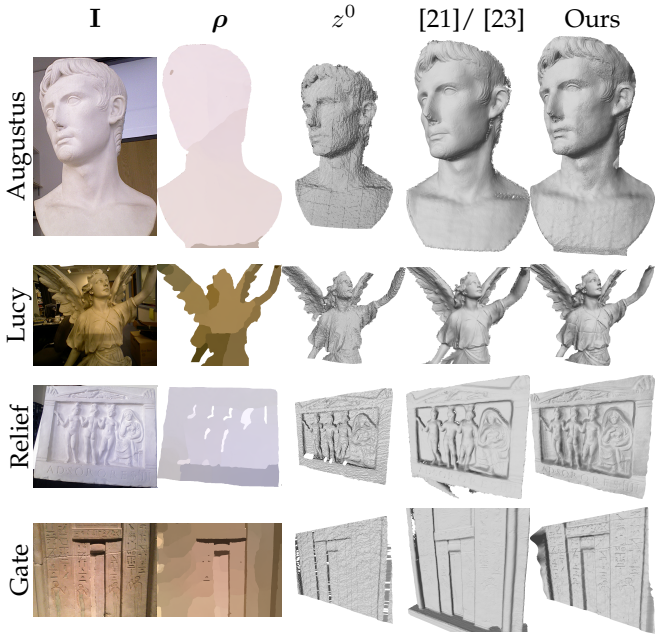

Fig. 10: Qualitative comparison against state-of-the-art multi-view approaches. Although it uses a single RGB-D frame, our approach results in depth maps whose quality is comparable with those obtained using multi-view data.

3.6 Additional Comparison against State-of-the-art Single-shot Techniques on Real-world Datasets we Captured Ourselves

Figure 11 shows additional qualitative comparison of single-shot results, on data we captured using an Asus Xtion Pro Live camera (scaling factor of 4). Once again, our approach outperforms the state-of-the-art, even though under- or over-segmentation of the reflectance may happen.

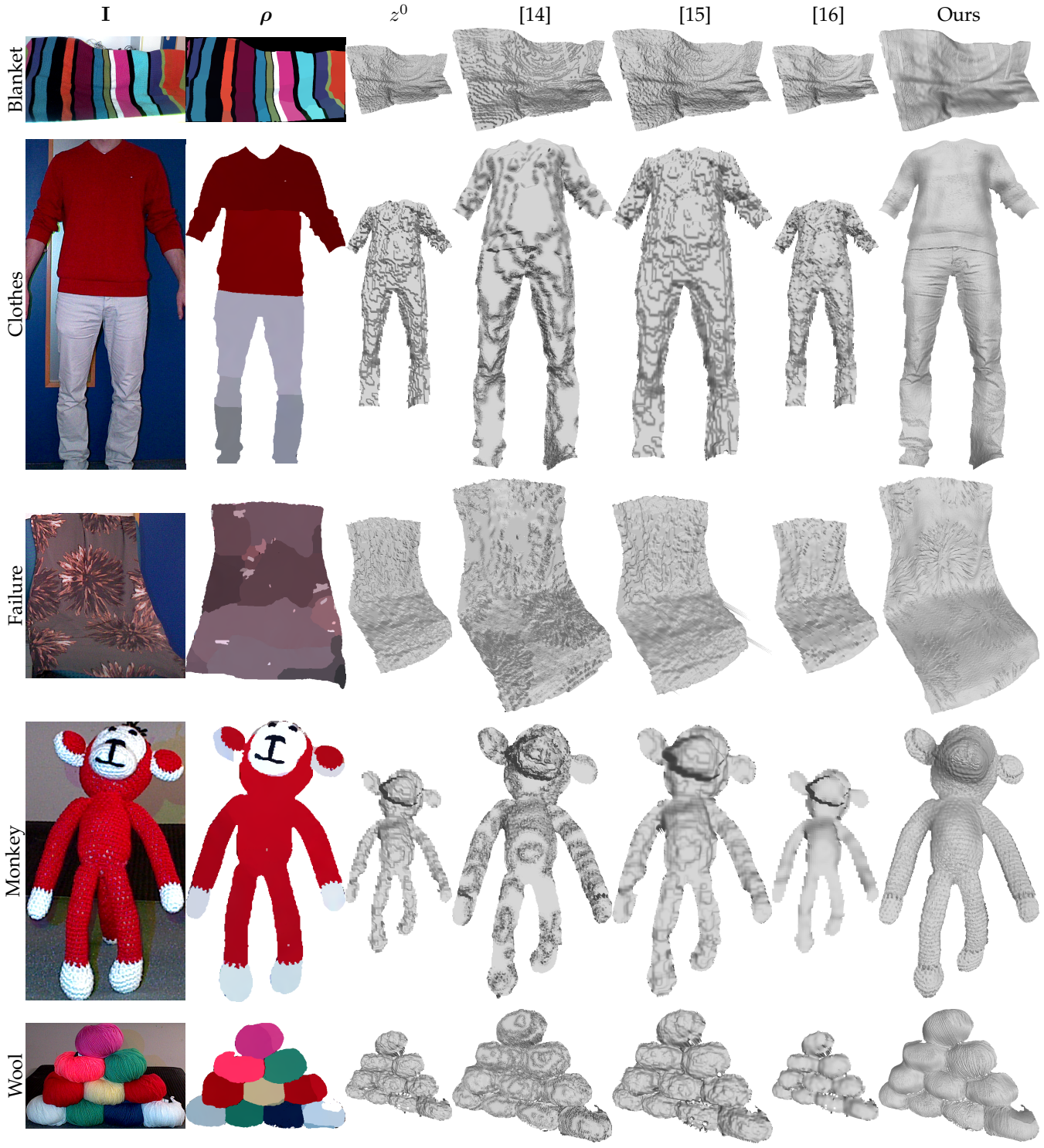

Fig. 11: Qualitative comparison of state-of-the-art single-view approaches on five real-world datasets captured with an Asus Xtion Pro Live camera at resolution 1280×960 for the RGB images and 320×240 for the low-resolution depth.

The "Clothes" experiment illustrates a case where over-segmentation of reflectance happens, but interestingly this does not seem to impact depth recovery. Whenever color gets saturated (some of the balls of "Wool") or too low (black areas in the "Blanket"), then minimal surface drives super-resolution: the areas where brightness is not informative are simply smoothed out, which adds robustness. Our method only fails when reflectance does not fit the Potts prior, as shown in the "Failure" experiment. In this case of an object with smoothly-varying reflectance, under-segmentation of reflectance happens, and all the thin brightness variations are interpreted in terms of geometry. Two alternative strategies are investigated in this work to cope with this issue: estimate reflectance without a piecewise-constant prior (learning-based strategy), or actively control lighting (photometric stereo-based strategy).

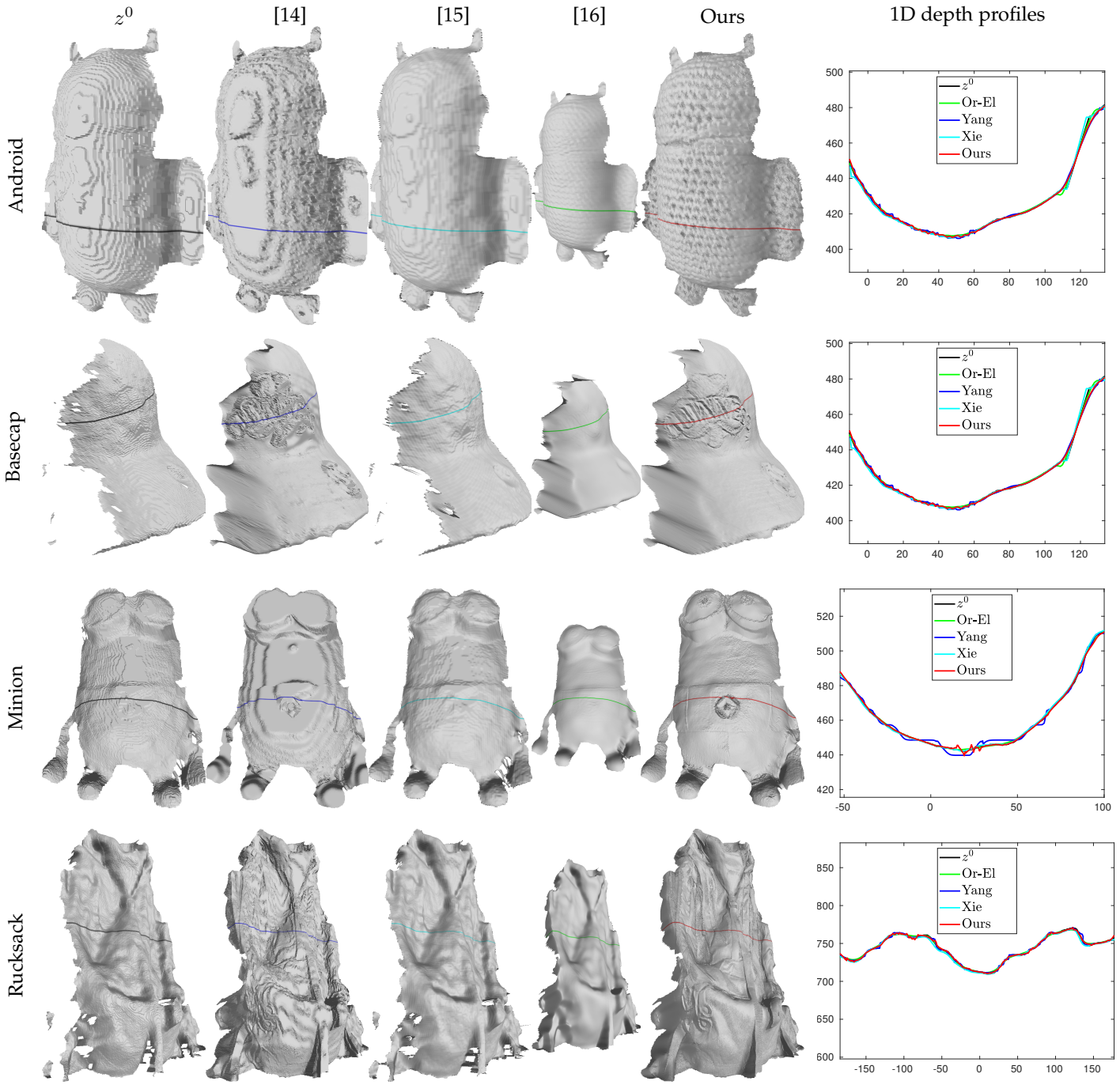

Fig. 12: Qualitative comparison between our single-shot results and those from the state-of-the-art, on the datasets from Figure 3 in the main paper. Our approach systematically outperforms the state-of-the-art. The rightest column shows 1D depth profiles corresponding to the lines drawn on the 3D-shapes: although the depth estimated using all the methods overall fit well together, ours is the only which provides reasonable fine-scale details.

Eventually, Figure 12 presents another qualitative comparison on the real-world data from Figure 3 in the main paper (captured using a RealSense D415 camera). Note that [14] seems to give good depth estimates whenever the underlying assumption (an edge in the RGB image coincides with an edge in the depth image) is met, cf. "Rucksack" experiment, but it fails to provide detail-preserving depth maps when reflectance is uniform or changes only slightly ("Android" and "Minion" experiments), since it uses only a sparse set of information from the RGB data. Unsurprisingly, the method from [15] cannot hallucinate surface details, since it does not use the color image. The shading-based depth refinement method of [16] does a much better job at improving geometry, but it is largely overcome by the proposed shading-based depth super-resolution approach, because the latter uses information from a higher-resolution RGB image.

4 Evaluation of the Reflectance Learning-based Approach

4.1 Creation of the Synthetic Data

Let us first recall that the reflectance learning-based approach was trained on data extracted from the ICT-3DRFE Database [5]. In order to evaluate this approach, we considered two subjects from this database as well, each one enacting 10 different facial expressions. Of course, in order to avoid any bias, these subjects were not used when training the neural network.

The high-resolution RGB and low-resolution depth images were created in a similar manner as in the previous section: high-resolution RGB images of the faces were rendered at 512×512 resolution from the ground truth albedo and depth under first-order spherical harmonics lighting $\mathbf{l} = [0, 0, -1, 0.2]^\top$; and the low-resolution depth maps were created by downsampling the ground truth depth by a scaling factor of 2, 4 and 8. Zero-mean Gaussian noise with standard deviation 1% the maximum RGB intensity was then added to the RGB images, and zero-mean Gaussian noise with standard deviation 10^{-4} the squared original depth value (consistently with the real-world measurements from [13]) was added to the low-resolution depth maps, before quantisation.

These synthetic faces were then used for quantitative evaluation of the proposed reflectance learning-based approach against the state-of-the-art and against the proposed fully variational solution, as discussed in the next subsection.

4.2 Comparison against the State-of-the-art on the Synthetic Dataset

We next evaluate our method, which first estimates reflectance using deep learning and then achieves variational depth super-resolution using the RGB image, in comparison with end-to-end deep learning solutions for geometry estimation.

For comparison, we first consider the method introduced in [24], which is an end-to-end depth super-resolution technique based on low-resolution depth data and high-resolution RGB image, i.e. the same inputs as our methods. It can be seen in Figure 13 that this end-to-end solution fails to recover surface details which are visible in the RGB image.

In order to evaluate the ability of deep networks to reconstruct geometry from a single RGB image, similarly to shape-from-shading techniques, we also show the results of SfSNet [25], which is a deep learning-based method estimating albedo and surface normals (which we further integrated into a depth map using the quadratic integation method discussed in [26]) out of a single RGB image. SfSNet is limited to RGB images of size 128×128, so it was evaluated only for a scaling factor of 4 and, since it does not perform depth super-resolution, the ground truth depth was downsampled for the quantitative evaluation of this method. Figure 13 shows that reasonable results can be expected using SfSNet, yet geometry is slightly oversmoothed in comparison with what can be obtained using the proposed combination of machine learning and variational approaches.

Eventually, we compare this combined approach with the fully variational one from the previous section. The latter does not completely fail at recovering a reasonable geometry, but since the estimated albedo is piecewise-constant and departs significantly from the ground truth, artifacts and noise are propagated to the estimated geometry. This is confirmed by the quantitative evaluation in Table 3, which clearly indicates that the proposed combination of machine learning and variational methods is more efficient than both end-to-end learning solutions from the state-of-the-art and the proposed fully variational approach.

4.3 Qualitative Comparison against the State-of-the-art on Real-world Datasets we Captured Ourselves

In Figure 14, we show additional qualitative comparisons of our results against those from the state-of-the-art, on the dataset from Fig. 5 in the main paper. This dataset consists of RGB-D frames of human faces which we acquired ourselves using an Intel Realsense D415 camera (the scaling factor between the high-resolution RGB image and the low-resolution depth map is 4).

This qualitative comparison validates the conclusions from the synthetic experiment in the previous subsection: combining variational and machine learning techniques yields more detailed 3D-reconstructions than end-to-end learning solutions based on neural networks for solving the shape-from-shading [25] or the depth super-resolution [24] problems.

4.4 Comparison against the State-of-the-art on a Public Real-world Dataset

Eventually, we compare qualitatively in Figure 15, and quantitatively in Table 4, the results of the proposed reflectance learning-based approach against those of the state-of-the-art, on data extracted from the DiLiGenT dataset [18]. Note that the datasets are exactly the same as the ones used for the evaluation of the fully variational solution in Figure 9 and Table 2, so that the results of the fully variational solution and those of the combined approach can also be compared.

Let us emphasize that the proposed reflectance learning-based solution was trained on a faces dataset, while none of the objects in this experiment resembles a face. Therefore, this test is rather intended as a test of robustness, and we are not expecting to overcome the results of the fully variational solution.

Indeed, the results obtained with the combined approach are both qualitatively and quantitatively less satisfactory on this dataset than those obtained with the fully variational solution. However, they remain surprisingly competitive, in comparison with the state-of-the-art.

Obviously, such a combination of machine learning and variational methods could still be improved by increasing the size of the training database using multiple classes of objects, but the present results already demonstrate its potential.

Fig. 13: Qualitative comparison of the results obtained using the deep learning-based depth super-resolution technique from [24], the deep learning-based shape-from-shading approach from [25], the proposed variational approach to shape-from-shading (denoted by SfS), and the proposed combination of deep learning and variational methods. The latter seems the most effective, and this is confirmed by the quantitative evaluation provided in Table 3.

Subject (S)	Expression (E)	SF	[24]		[25]		Ours (SfS)		Ours	
			RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
4	0	2	0.1572	48.3553	–	–	0.1613	14.1551	**0.1355**	**9.9354**
		4	0.13284	36.9774	0.6071	13.4647	0.10629	11.6301	**0.086867**	**8.5443**
		8	0.63	38.0244	–	–	0.24354	14.2278	**0.19869**	**11.0617**
	1	2	0.15451	48.4316	–	–	0.15824	15.2649	**0.13413**	**10.6932**
		4	0.13069	36.4169	0.7185	14.3254	0.10972	12.8361	**0.087058**	**9.5301**
		8	0.61295	35.7076	–	–	0.25859	15.2818	**0.20735**	**12.0648**
	2	2	0.15358	49.0454	–	–	0.14589	14.516	**0.14005**	**14.221**
		4	0.13266	37.1973	0.8821	18.5108	0.12322	13.9566	**0.10993**	**13.2265**
		8	0.97825	38.8272	–	–	0.27613	17.4243	**0.25379**	**16.0565**
	3	2	0.15657	48.4417	–	–	0.1614	14.776	**0.14468**	**11.1379**
		4	0.13335	37.0725	0.8554	16.0271	0.11759	13.165	**0.09558**	**10.604**
		8	0.95333	38.6755	–	–	0.26179	15.8035	**0.21567**	**13.0665**
	4	2	0.15155	48.3665	–	–	0.14914	14.5008	**0.13132**	**11.1096**
		4	0.13093	37.2093	0.6301	15.0882	0.1108	12.6113	**0.091216**	**10.0432**
		8	0.81628	37.4412	–	–	0.24872	15.1804	**0.18053**	**12.4699**
	5	2	**0.15404**	47.7565	–	–	0.17	15.3645	0.16346	**13.0845**
		4	0.17413	37.2071	0.9004	18.8166	0.187	15.0677	**0.16933**	**13.3297**
		8	0.81401	37.1801	–	–	0.35861	18.885	**0.31589**	**16.5856**
	6	2	0.15457	47.8468	–	–	0.16725	14.4807	**0.15373**	**12.0069**
		4	0.13863	36.681	0.8684	19.1934	0.14573	13.5136	**0.12169**	**11.3427**
		8	0.49746	36.8001	–	–	0.31234	17.0774	**0.26064**	**14.4585**
	7	2	**0.15476**	48.4094	–	–	0.17713	15.3454	0.1602	**12.6357**
		4	0.18215	36.0914	0.8460	19.8673	0.18528	14.1876	**0.15723**	**11.8129**
		8	0.82718	38.4132	–	–	0.34986	17.1481	**0.29932**	**14.3226**
	8	2	0.15437	48.3719	–	–	0.15093	15.9026	**0.13533**	**12.1738**
		4	0.13062	37.3586	0.4986	13.6524	0.107	13.3844	**0.085065**	**10.5078**
		8	0.71791	37.543	–	–	0.23366	15.5826	**0.19305**	**12.6953**
	9	2	0.15989	49.2317	–	–	0.15939	13.879	**0.13843**	**11.097**
		4	0.13373	38.022	0.6107	14.3473	0.11108	12.4866	**0.091548**	**9.9358**
		8	0.53732	36.8758	–	–	0.25022	15.1722	**0.20378**	**12.7385**
11	0	2	0.16035	48.3088	–	–	0.15248	15.1914	**0.13971**	**9.7571**
		4	0.13743	36.6588	1.0125	11.8150	0.11775	12.5609	**0.10129**	**8.6988**
		8	0.51743	32.4395	–	–	0.26081	14.6577	**0.22472**	**11.6671**
	1	2	0.15231	48.2292	–	–	0.14523	15.381	**0.13328**	**9.9593**
		4	0.12957	35.6881	0.8798	10.8757	0.11237	12.7309	**0.097136**	**8.5511**
		8	0.52279	32.5115	–	–	0.25173	14.7836	**0.20387**	**11.6099**
	2	2	0.15548	47.5781	–	–	0.15421	15.7925	**0.14821**	**12.0207**
		4	0.1393	36.4907	0.9789	19.5521	0.13943	14.875	**0.12114**	**11.8059**
		8	0.66616	36.1001	–	–	0.30543	18.2869	**0.26649**	**15.2768**
	3	2	0.16131	48.4766	–	–	0.15472	15.3901	**0.14409**	**10.0102**
		4	0.13652	36.0848	1.2922	13.2403	0.12219	13.3513	**0.10614**	**8.9464**
		8	0.94169	37.7036	–	–	0.27622	16.4698	**0.2266**	**12.5186**
	4	2	0.15879	48.3293	–	–	0.15457	15.0479	**0.14001**	**10.8615**
		4	0.13926	36.8105	0.8897	12.7579	0.12404	13.557	**0.10533**	**9.5906**
		8	0.72556	35.9876	–	–	0.27086	15.786	**0.22974**	**12.7579**
	5	2	0.16252	47.6152	–	–	0.16964	17.0446	**0.15787**	**10.6522**
		4	0.15556	36.695	1.1557	14.7778	0.17783	15.3102	**0.15392**	**10.6452**
		8	0.81958	35.1608	–	–	0.32727	18.6277	**0.28649**	**14.4657**
	6	2	0.15936	48.2603	–	–	0.15054	15.5422	**0.14255**	**10.4559**
		4	0.13906	36.2701	0.7581	13.9221	0.13145	13.7609	**0.1157**	**9.8813**
		8	0.68759	35.3423	–	–	0.29362	18.0689	**0.25192**	**14.3041**
	7	2	**0.15783**	46.3708	–	–	0.19123	16.3544	0.17274	**10.4441**
		4	**0.20118**	35.1363	1.2066	18.6458	0.23771	16.3544	0.20955	**11.5822**
		8	0.73165	35.5369	–	–	0.41273	19.1395	**0.36912**	**14.8272**
	8	2	0.1601	48.3084	–	–	0.14637	18.5782	**0.12985**	**13.3089**
		4	0.13852	37.6211	0.7112	13.2194	0.11509	15.9898	**0.095155**	**11.6081**
		8	0.78491	37.1651	–	–	0.25296	18.0263	**0.20633**	**14.3745**
	9	2	0.15292	48.2978	–	–	0.13997	14.5447	**0.12648**	**10.2274**
		4	0.13424	36.6469	0.9484	12.6980	0.11997	13.0994	**0.10137**	**9.4048**
		8	0.63803	34.7198	–	–	0.26044	15.9719	**0.21383**	**12.7267**
Median		2	0.15693	48.3609	–	–	0.15494	15.1423	**0.14043**	**11.0633**
		4	0.13568	36.769	0.8741	14.3363	0.11767	13.1322	**0.10094**	**9.9086**
		8	0.72174	36.838	–	–	0.26285	15.7971	**0.22566**	**12.7326**
Mean		2	0.15694	48.2983	–	–	0.15773	15.3515	**0.1432**	**11.1919**
		4	0.14095	36.7644	0.8625	15.2399	0.12908	13.4354	**0.10935**	**10.1732**
		8	0.72675	36.4789	–	–	0.27802	16.2705	**0.23276**	**13.117**

TABLE 3: Quantitative comparison between two state-of-the-art methods, the proposed fully variational approach based on shape-from-shading (denoted by SfS), and the proposed combination of deep learning and variational methods, on the synthetic dataset. The combined solution is the most effective.

Fig. 14: Qualitative comparison of our reflectance learning-based results against state-of-the-art methods, on six RGB-D frames of human faces which we captured using an Intel RealSense D415 camera (scaling factor of 4). Our method provides the most detailed 3D-reconstructions.

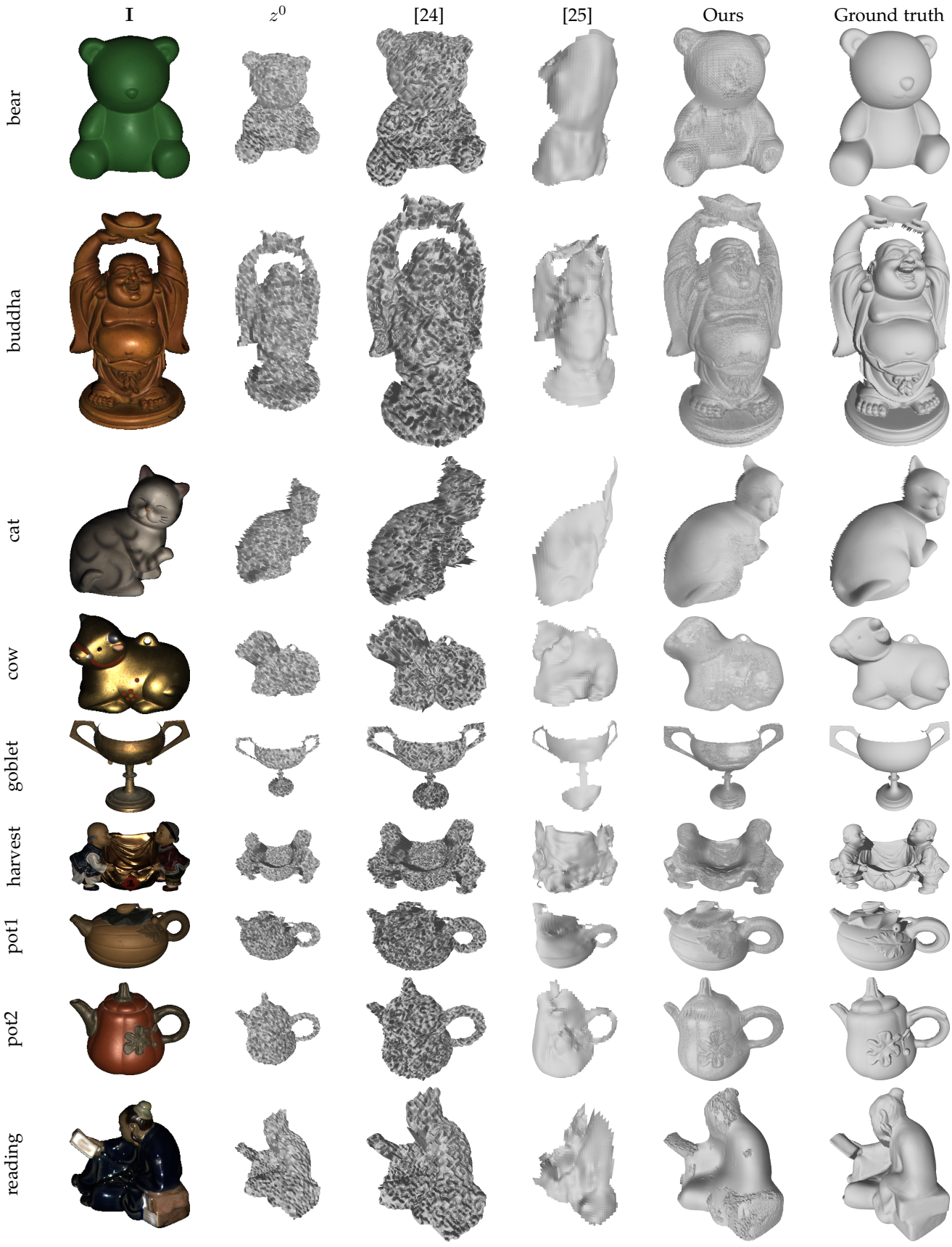

Fig. 15: Qualitative comparison between our method combining deep learning and variational methods, and state-of-the-art deep learning-based methods, on the DiLiGenT dataset [18] (the scaling factor is 4). Our approach outperforms the state-of-the-art in all the experiments.

3D-shape	SF	[24]		[25]		Ours	
		RMSE	MAE	RMSE	MAE	RMSE	MAE
bear	2	0.010946	62.708	–	–	**0.0046569**	**22.5961**
	4	0.010609	49.8753	0.1096	41.9262	**0.0086235**	**23.2417**
	8	**0.012821**	36.8812	–	–	0.018246	**30.7021**
buddha	2	0.011778	63.0557	–	–	**0.0082909**	**29.6526**
	4	0.012539	52.2028	0.0518	40.1120	**0.011945**	**33.5974**
	8	**0.015423**	45.132	–	–	0.019568	**41.0636**
cat	2	0.013194	62.903	–	–	**0.008389**	**15.1466**
	4	0.0137	50.278	0.0647	36.9720	**0.013534**	**19.1494**
	8	**0.015258**	38.3265	–	–	0.023363	**26.7149**
cow	2	0.011679	64.5302	–	–	**0.0053628**	**17.6086**
	4	0.011237	50.6826	0.0562	39.3336	**0.0092811**	**18.8318**
	8	**0.014157**	42.9122	–	–	0.017689	**21.1007**
goblet	2	0.013153	61.8508	–	–	**0.011713**	**30.1888**
	4	**0.01379**	48.7097	0.1414	36.4712	0.017615	**29.6286**
	8	**0.016659**	36.6476	–	–	0.03133	**28.7208**
harvest	2	0.0167	64.113	–	–	**0.016649**	**39.602**
	4	**0.019409**	53.9958	0.1757	54.1461	0.024208	**41.0901**
	8	**0.028625**	44.4953	–	–	0.037441	**41.1994**
pot1	2	0.011218	61.9779	–	–	**0.0070793**	**18.4819**
	4	0.011597	50.0199	0.1051	35.0139	**0.010794**	**18.6248**
	8	**0.01495**	40.4749	–	–	0.019198	**20.5408**
pot2	2	0.010693	61.9083	–	–	**0.0057831**	**20.0908**
	4	0.011123	50.5484	0.0575	32.0884	**0.0090011**	**20.7887**
	8	**0.014105**	40.3902	–	–	0.016243	**23.1403**
reading	2	0.012058	61.2583	–	–	**0.0098101**	**20.5263**
	4	**0.012927**	49.0756	0.0817	55.4988	0.015531	**24.2634**
	8	**0.017714**	41.0243	–	–	0.028793	**28.8291**
Median	2	0.011778	62.708	–	–	**0.0082909**	**20.5263**
	4	0.012539	50.278	0.0732	38.1528	**0.011945**	**23.2417**
	8	**0.015258**	40.4749	–	–	0.019568	**28.7208**
Mean	2	0.01238	62.7006	–	–	**0.0086371**	**23.766**
	4	**0.012992**	50.5987	0.0918	41.2045	0.013392	**25.4684**
	8	**0.016635**	40.6982	–	–	0.023541	**29.1124**

TABLE 4: Quantitative comparison between other state-of-the-art methods and our method combining machine learning and variational methods. Although the results are not as accurate as the fully variational solution (cf. Table 2), since none of the objects here resembles the faces from the training database, they remain superior to the state-of-the-art.

5 Evaluation of the Multi-shot Approach Based on Photometric Stereo

5.1 Creation of the Synthetic Data

In order to quantitatively evaluate the proposed photometric stereo-based solution, we consider the same four 3D-shapes as in the shape-from-shading experiments, i.e. "Lucy", "Thai Statue", "Armadillo" and "Joyful Yell". However, this time we consider much more complex albedo maps since the multi-shot approach is not limited to piecewise-constant albedos. The albedo maps we consider are "ebsd"[1], "mandala"[2] and "rectcircle". The rest of the process for creating the dataset (rendering the high-resolution RGB and low-resolution depth images, and adding noise) is exactly the same as for shape-from-shading, except that multiple RGB images are acquired under randomly varying lighting. Three RGB images of each dataset under three different illumination conditions are presented in Figure 16, and the corresponding depth maps are those from Figure 6.

1. https://mtex-toolbox.github.io/files/doc/EBSDSpatialPlots.html
2. http://www.cleverpedia.com/mandala-coloring-books-20-coloring-books-with-brilliant-kaleidoscope-designs/

5.2 Selecting the Number of Images and Tuning the Hyper-parameters

Figure 17 illustrates the effect of the hyper-parameter γ on shape and reflectance estimation. For this purpose, we consider sets of $n = 10$ images from the Joyful Yell dataset, and evaluate the RMSE and MAE on depth, as well as the RMSE on albedo, as functions of the number of input images. As can be seen, when $\gamma \to 0$ the estimated depth map sticks to the noisy input, thus results are deceiving. But as soon as γ is large enough, photometric stereo drives super-resolution and the accuracy dramatically increases. Interestingly, results remain stable even when $\lambda \to \infty$. This tends to indicate that the ambiguities of uncalibrated photometric stereo vanish as soon as a depth prior is available: it is not necessary to seek a compromise between the depth prior and the photometric 3D-reconstruction, only to plug the information from the former into the latter.

Next, we evaluate the number n of input RGB images which would result in the best compromise between accuracy of the 3D-reconstruction and runtime. For this purpose, we consider once again the Joyful Yell synthetic dataset, and evaluate the RMSE and MAE on depth, the RMSE on albedo and the total runtime required to attain convergence, as functions of n. As can be seen in Figure 18, the accuracy of the estimation very quickly increases with n, while the runtime increases linearly with n. Overall, the choice $n \in [10, 30]$ seems to represent a good compromise.

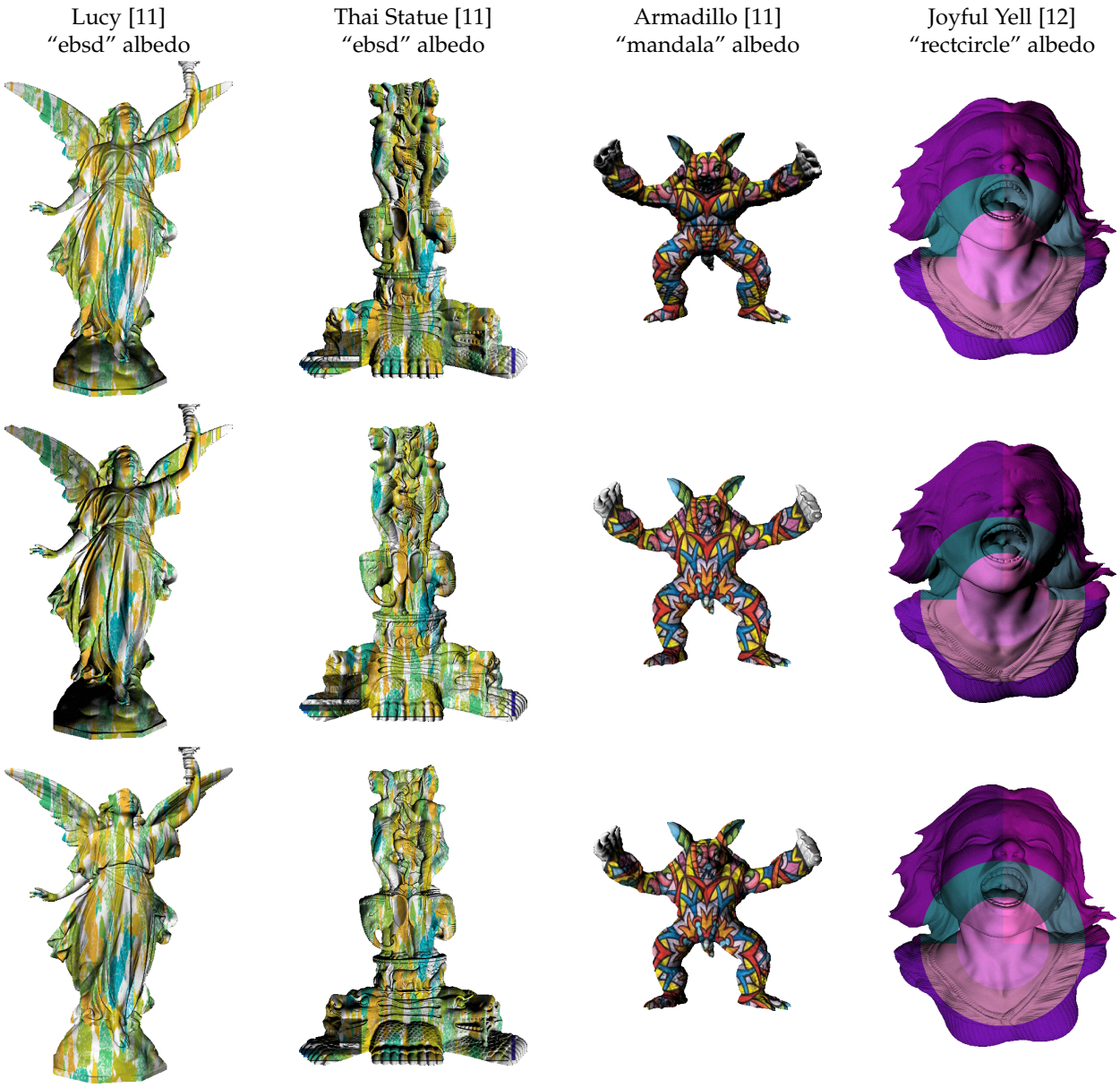

Fig. 16: Illustration of the synthetic RGB data used for quantitatively evaluating the multi-shot depth super-resolution approach based on photometric stereo. Each row represents a different illumination condition. Remark that much more complex albedo maps are considered, in comparison with the ones used in the single-shot approach, cf. Figure 6.

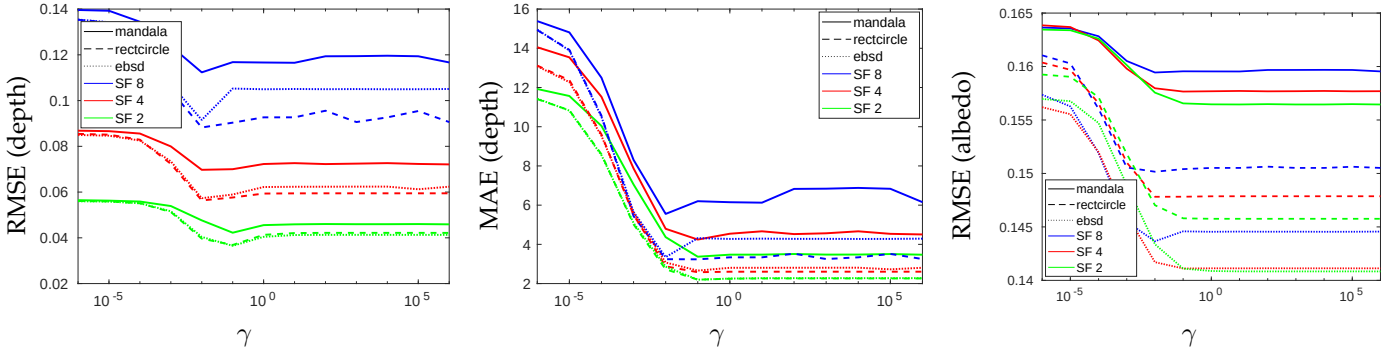

Fig. 17: Impact of the parameter γ on the accuracy of the albedo and depth estimates using our multi-shot photometric stereo approach ($n = 10$ in this experiment). Based on these results, the value $\gamma = 0.01$ was retained.

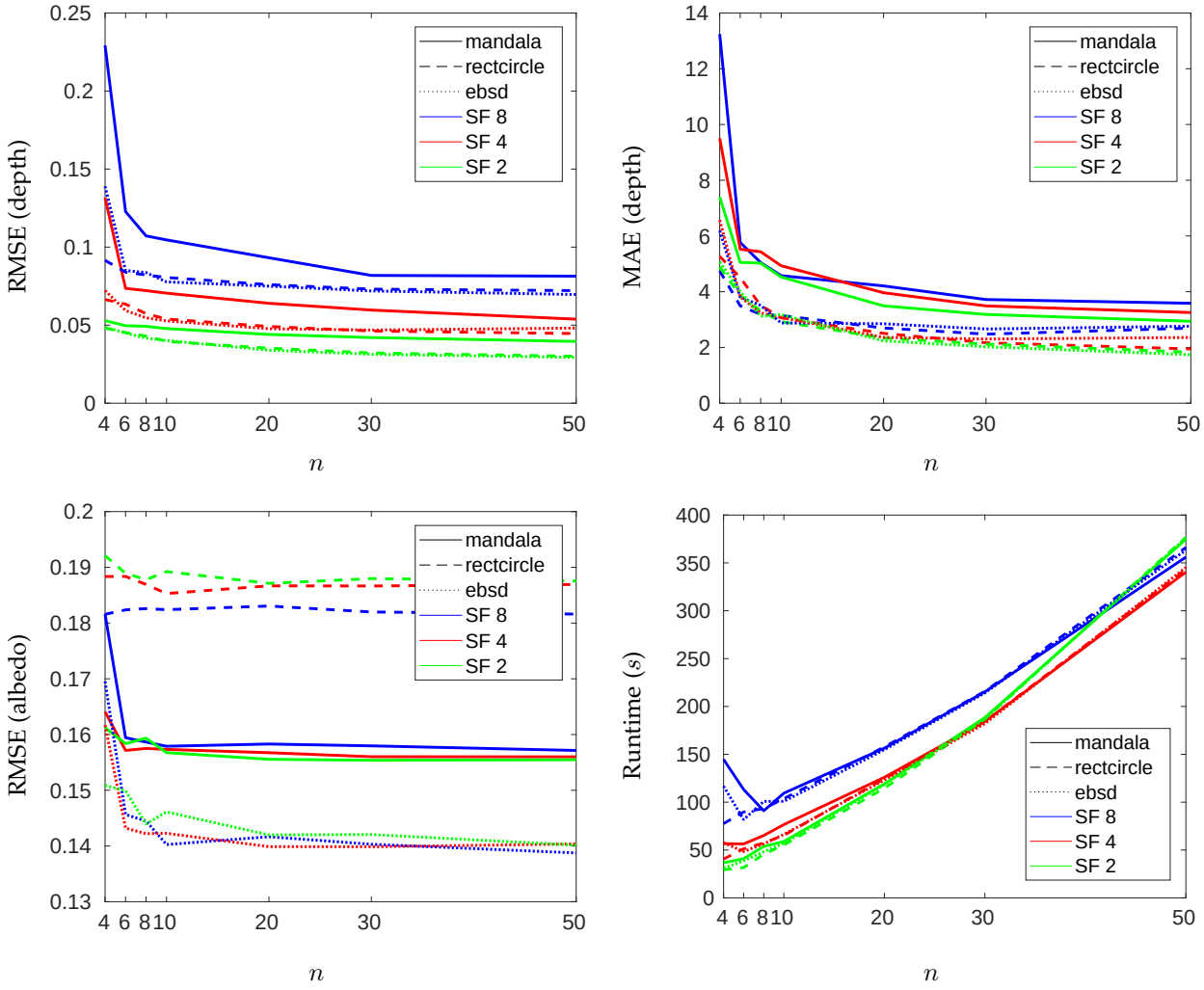

Fig. 18: Impact of the number of images n on the accuracy of the albedo and depth estimates using our multi-shot photometric stereo approach ($\gamma = 0.01$ in this experiment). The range $n \in [10, 30]$ represents a reasonable compromise between accuracy and runtime.

5.3 Comparison against the State-of-the-art on the Synthetic Dataset

Next, we compare our multi-shot approach against the state-of-the-art, on all the synthetic datasets (consistently with the results from the previous subsection, $n = 20$ images are considered for each dataset, and $\gamma = 0.01$ in all the experiments). Our results are expected to overcome both pure depth super-resolution and pure uncalibrated photometric stereo, as well as single-shot depth refinement methods acting on low-resolution data.

To highlight the interest of an explicit photometric model, we first compare our results against an image-based multi-shot depth super-resolution approach adapted from [27], [28]. It is a personal combination of these papers which achieves variational depth super-resolution by fusing the n low-resolution depth maps, while regularising the gradient of the estimated high-resolution depth map in an anisotropic manner. Here, the anisotropy coefficient is derived from the gradients of the RGB image. This approach is thus a "pure depth super-resolution" one, which uses RGB clues but without any explicit photometric model.

In contrast, we also consider the "pure" uncalibrated photometric stereo method from [29], which estimates lighting, albedo and high-resolution geometry from the n high-resolution RGB images. In this method, an explicit photometric model is used, as in ours, yet no low-resolution depth clue is considered hence the underlying bas-relief ambiguity may affect the quality of the results.

As in the evaluation of the shape-from-shading-based approach from Section 3, we also show the results of RGB-D refinement [16] applied to the low-resolution RGB-D frame, selecting one image out of n.

The qualitative comparison in Figure 19, and the quantitative ones in Table 5, show that our methods result in much more satisfactory high-resolution geometry, in comparison with these methods. This proves that using an explicit model for driving image-based depth super-resolution, and using low-resolution depth clues to disambiguate uncalibrated photometric stereo, both are worthwile.

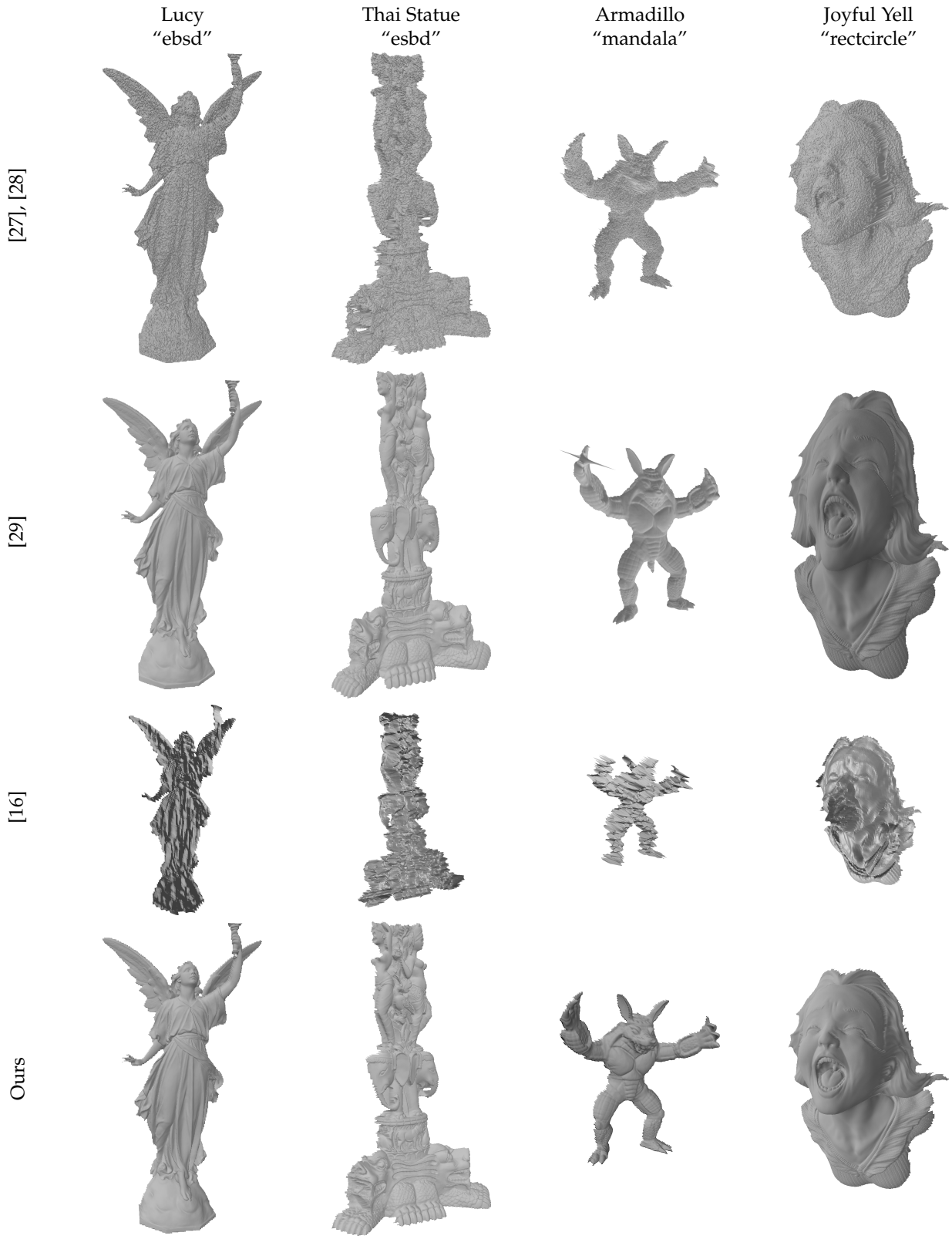

Fig. 19: Qualitative comparison of our multi-shot approach against state-of-the-art methods, on four synthetic datasets (scaling factor of 4). Image-based depth super-resolution adapted from [27], [28] results in noisy geometry, uncalibrated photometric stereo results from [29] are slightly flattened due to the underlying bas-relief ambiguity, and RGB-D fusion [16] of the low-resolution data is not really successful here. In comparison, the results of the proposed method are extremely satisfactory.

Albedo	3D-shape	SF	Image Based depth SR		[29]*		[16]		Ours	
			RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
mandala	Armadillo	2	0.031468	46.4149	0.51996	17.4225	0.4320	69.7311	**0.023266**	**2.883**
		4	0.042467	43.5403	–	–	0.3948	63.7382	**0.037789**	**2.8391**
		8	0.088849	42.6184	–	–	0.5961	83.7853	**0.073928**	**2.9196**
	Lucy	2	0.043889	46.4903	0.37197	15.7192	0.4755	84.6189	**0.036334**	**3.5842**
		4	0.065857	44.0677	–	–	0.4951	82.0516	**0.051316**	**3.6142**
		8	0.12668	42.8905	–	–	0.5231	64.7317	**0.084713**	**4.7864**
	Joyful Yell	2	0.048887	45.0552	1.0735	14.2243	0.3757	70.2724	**0.044198**	**3.3143**
		4	0.069088	42.644	–	–	0.2985	55.6927	**0.063392**	**3.6407**
		8	0.13103	40.0426	–	–	0.4240	44.2549	**0.1046**	**3.753**
	Thai Statue	2	0.032432	47.8575	0.37738	13.372	0.4615	70.3271	**0.022446**	**3.579**
		4	0.053061	45.5618	–	–	0.4211	90.2134	**0.036245**	**3.6985**
		8	0.094911	43.838	–	–	0.3371	53.2791	**0.049733**	**4.1133**
rectcircle	Armadillo	2	0.028459	41.506	0.52582	18.0902	0.2844	55.3096	**0.020885**	**2.0047**
		4	0.038966	38.7345	–	–	0.3031	48.1000	**0.035145**	**1.9458**
		8	0.11182	36.3801	–	–	0.5805	80.4625	**0.073139**	**2.1436**
	Lucy	2	0.040635	42.3051	0.32285	13.6126	0.4868	85.9076	**0.026858**	**1.8617**
		4	0.062747	39.0783	–	–	0.4685	75.9166	**0.041968**	**2.2851**
		8	0.12325	37.956	–	–	0.3767	56.5020	**0.075311**	**3.8793**
	Joyful Yell	2	0.045765	39.9946	0.84162	11.4847	0.2012	41.3053	**0.038698**	**2.7879**
		4	0.064537	37.1175	–	–	0.3189	37.2107	**0.053871**	**3.1022**
		8	0.09492	34.7218	–	–	0.4432	36.3990	**0.084381**	**3.2463**
	Thai Statue	2	0.030859	44.4276	0.38981	13.3935	0.2625	66.0562	**0.018374**	**2.1086**
		4	0.045516	41.7235	–	–	0.3151	85.4734	**0.028457**	**2.2876**
		8	0.10507	39.7697	–	–	0.2389	55.0568	**0.041552**	**3.0519**
ebsd	Armadillo	2	0.031939	46.9515	0.49466	16.3427	0.3473	65.4823	**0.021037**	**2.0398**
		4	0.04424	44.2571	–	–	0.5933	58.6932	**0.036102**	**2.0035**
		8	0.10062	42.2539	–	–	0.6453	81.5187	**0.073138**	**1.8159**
	Lucy	2	0.04299	47.5844	0.32989	13.0463	0.4141	84.9623	**0.028555**	**1.9483**
		4	0.072388	44.5851	–	–	0.4541	75.3771	**0.04325**	**2.1771**
		8	0.16385	42.4252	–	–	0.6460	74.8618	**0.079427**	**3.6839**
	Joyful Yell	2	0.049515	46.0065	1.0052	13.1767	0.2645	55.3462	**0.034162**	**2.1722**
		4	0.069491	43.4654	–	–	0.2770	42.4242	**0.04818**	**2.3335**
		8	0.11255	40.9818	–	–	0.4589	38.8507	**0.073515**	**2.5774**
	Thai Statue	2	0.03307	48.7666	0.30254	12.0112	0.2371	69.6653	**0.019305**	**2.3639**
		4	0.046843	45.6104	–	–	0.2792	77.7622	**0.029185**	**2.4529**
		8	0.089646	43.7591	–	–	0.2847	64.3520	**0.041307**	**2.9642**
Median		2	0.036853	46.2107	0.44223	13.503	0.12186	45.0229	**0.025062**	**2.2681**
		4	0.057904	43.5029	–	–	0.18929	41.3767	**0.039879**	**2.3932**
		8	0.10844	41.6178	–	–	0.31159	41.3102	**0.073722**	**3.1491**
Mean		2	0.038326	45.28	0.54626	14.3246	0.11516	42.7392	**0.027843**	**2.554**
		4	0.056267	42.5321	–	–	0.18488	40.6331	**0.042075**	**2.6984**
		8	0.11193	40.6364	–	–	0.29819	40.1205	**0.071228**	**3.2446**

TABLE 5: Quantitative comparison of the results attained with the proposed multi-shot approach and the state-of-the-art (*: to make the comparison fair, we run the algorithm of [29] on the high resolution RGB images, as it performs uncalibrated photometric stereo on the RGB images without super-resolution – the scaling factor is thus actually 1 in this case). Our approach overcomes the state-of-the-art in all the experiments.

5.4 Qualitative Comparison against the State-of-the-art on Real-world Datasets we Captured Ourselves

Figure 20 shows four qualitative comparisons against the state-of-the-art, on real-world data from Figures 1 and 6 in the main paper, which was captured with an Asus Xtion Pro Live camera (scaling factor of 4).

It can be seen that image-based depth super-resolution approach hallucinates reflectance information as geometric information, since the underlying concept allows larger depth variations where strong image gradients are present. The uncalibrated photometric stereo results from [29] contain much more relevant details, but the approach clearly suffers from a low-frequency bias due to the underlying bas-relief ambiguity, cf. "Tablet Case" and "Vase". In these experiments the RGB-D fusion results from [16] are reasonable, but not as accurate as the ones obtained with the proposed multi-shot approach.

5.5 Comparison against the State-of-the-art on a Public Real-world Dataset

Eventually, we compare our results against the state-of-the-art on the DiLiGenT dataset [18]. Qualitative results are presented in Figure 21, and quantitative ones in Table 6. Once again, our method most of the times overcomes the state-of-the art in terms of surface details recovery. It is also interesting to compare these results with the corresponding ones in the previous sections: this comparison clearly shows that resorting to a multi-shot strategy based on photometric stereo is the only way to cope with general reflectance.

Still, it can be observed that even with redundant data, some results such as the "harvest" one remain somewhat disappointing: this is because the proposed method explicitly builds upon the Lambertian assumption, which is not met in this example. Future extensions could thus include coping with non-Lambertian phenomena.

Fig. 20: Comparison between the proposed multi-shot method and the state-of-the-art, on real-world datasets captured using an Asus Xtion Pro Live camera. These results confirm the conclusion of the synthetic experiments in Figure 19.

Fig. 21: Qualitative comparison of our uncalibrated photometric stereo-based approach against state-of-the-art methods, on the DiLiGenT dataset [18] (the scaling factor is 2). Our method overcomes the state-of-the-art in all the experiments.

3D-shape	SF	[27], [28]		[29]*		Ours	
		RMSE	MAE	RMSE	MAE	RMSE	MAE
bear	2	0.0077882	23.2799	0.029124	8.65	**0.0064907**	**7.056**
	4	**0.0077919**	19.6628	–	–	0.0083983	**7.2645**
	8	**0.0079796**	23.8495	–	–	0.013453	**7.0708**
buddha	2	0.0077863	31.0075	0.041827	18.0718	**0.0066078**	**12.7816**
	4	0.0078303	28.5663	–	–	**0.0077671**	**13.0276**
	8	**0.0076309**	20.1206	–	–	0.012959	**13.6987**
cat	2	**0.0078205**	24.5162	0.039112	11.0118	0.008108	**6.1952**
	4	**0.0076492**	20.6365	–	–	0.010542	**6.5739**
	8	**0.0078364**	21.2045	–	–	0.015403	**7.3812**
cow	2	0.0078497	31.5175	0.030244	18.1343	**0.0055052**	**10.4445**
	4	**0.007844**	26.7532	–	–	0.0083455	**11.3151**
	8	**0.0085472**	17.225	–	–	0.015231	**12.7818**
goblet	2	**0.0078938**	32.2235	0.13005	71.5669	0.010771	**11.16**
	4	**0.0078725**	29.261	–	–	0.015434	**11.6484**
	8	**0.008322**	24.9651	–	–	0.030694	**13.9542**
harvest	2	**0.0078757**	32.6288	0.06847	**29.3081**	0.024211	30.4736
	4	**0.0078363**	**30.6866**	–	–	0.029344	31.9109
	8	**0.0077605**	**33.427**	–	–	0.040837	33.5636
pot1	2	0.0078648	25.4586	0.01869	10.3055	**0.0063032**	**7.3048**
	4	**0.0078397**	22.6612	–	–	0.0080599	**7.514**
	8	**0.0079306**	30.9277	–	–	0.014455	**7.9022**
pot2	2	0.0077881	29.7433	0.022896	14.5031	**0.0048177**	**9.4492**
	4	0.0080123	26.261	–	–	**0.0066391**	**9.5829**
	8	**0.0076366**	21.8009	–	–	0.012587	**10.0768**
reading	2	**0.0077277**	29.1401	0.069057	25.0014	0.0098433	**16.7382**
	4	**0.0076277**	26.4486	–	–	0.014885	**19.6366**
	8	**0.0078612**	**18.6829**	–	–	0.027963	23.2138
Median	2	0.0078205	29.7433	0.039112	18.0718	**0.0066078**	**10.4445**
	4	**0.0078363**	26.4486	–	–	0.0083983	**11.3151**
	8	**0.0078612**	21.8009	–	–	0.015231	**12.7818**
Mean	2	**0.0078216**	28.835	0.049941	22.9503	0.0091842	**12.4004**
	4	**0.0078115**	25.6597	–	–	0.012157	**13.1638**
	8	**0.007945**	23.5781	–	–	0.020398	**14.4048**

TABLE 6: Quantitative Comparison between other state-of-the-art methods and our multi-shot approach based on photometric stereo (*: to make the comparison fair, we run the algorithm of [29] on the high resolution RGB images, as it performs uncalibrated photometric stereo on the RGB images without super-resolution – the scaling factor is thus actually equal to 1 in this case). Our approach overcomes the state-of-the-art in terms of the level of geometric details which can be recovered, while being only slightly less accurate in terms of overall RMSE fit.

6 Unified Comparison of our Results on a Public Real-world Dataset

Eventually, we present in Figure 22 a unified qualitative comparison of the results obtained with the three proposed methods, on the 9 objects of the DiLiGenT dataset [18]. This dataset illustrates well the cases where the single-shot approach can be used (when reflectance is uniform, as for instance in the "bear" example) and when it completely fails because the piecewise-constant albedo assumption is not satisfied (e.g., "Cat"). This method could thus still be improved by designing a more general reflectance prior. The multi-shot approach based on uncalibrated photometric stereo estimates a much more reasonable albedo map, and thus a much more satisfactory depth map, because it does not rely on any assumption regarding piecewise-constantness. Yet, it could still be improved in order to reduce artifacts due to specularirites (e.g., "reading"). Eventually, the albedo estimated by deep learning is sometimes reasonable (e.g., "buddha"), but most of the times it is not really satisfactory. This is because the objects do not resemble the training set, which consists only of faces: to cope with a wider variety of objects, the training dataset should contain a broader range of object classes.

7 Conclusion

We evaluated in depth the applicability of photometric techniques to resolve depth super-resolution in the context of RGB-D sensing. Multiple self-captured real-world, publicly available real-world and self-generated synthetic datasets were used in order to qualitatively and quantitatively compare the three proposed strageties against state-of-the-art variational, optimization-based and deep learning methods. It appeared that each of the three proposed methods beats the corresponding state-of-the-art ones, which provides an empirical evidence for the soundness of considering photometry as a valuable clue for depth super-resolution in RGB-D sensing.

In order to have at hand a unified comparison of the three methods presented in this work, we also considered a publicly available real-world photometric stereo benchmark across all experimental sections. This permitted us to clearly highlight the respective strengths and weaknesses of each method. They could still be improved towards, respectively, a more general reflectance prior (single-shot strategy), a broader training dataset (reflectance learning), and the handling of specularities (uncalibrated photometric stereo).

Fig. 22: Comparison of the albedo and high-resolution depth maps estimated by the proposed variational approach to shape-from-shading (SfS), the combination of SfS and deep reflectance learning, and the uncalibrated photometric stereo (UPS)-based approach, on the DiLiGenT dataset [18]. For quantitative evaluation, we refer the reader to Tables 2, 4 and 6.

REFERENCES

[1] Y. Quéau, B. Durix, T. Wu, D. Cremers, F. Lauze, and J.-D. Durou, "LED-based Photometric Stereo: Modeling, Calibration and Numerical Solution," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 3, pp. 313–340, 2018.

[2] R. Basri and D. P. Jacobs, "Lambertian reflectances and linear subspaces," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 2, pp. 218–233, 2003.

[3] R. Ramamoorthi and P. Hanrahan, "An Efficient Representation for Irradiance Environment Maps," in *Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques*, 2001, pp. 497–500.

[4] D. Frolova, D. Simakov, and R. Basri, "Accuracy of spherical harmonic approximations for images of Lambertian objects under far and near lighting," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2004, pp. 574–587.

[5] G. Stratou, A. Ghosh, P. Debevec, and L. Morency, "Effect of illumination on automatic expression recognition: A novel 3d relightable facial database," in *Face and Gesture*, 2011, pp. 611–618.

[6] M. M. Takuya Narihira and S. X. Yu, "Direct intrinsics: Learning albedo-shading decomposition by convolutional regression," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[7] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012, pp. 611–625.

[8] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman, "Ground truth dataset and baseline evaluations for intrinsic image algorithms," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2009, pp. 2335–2342.

[9] J. Shi, Y. Dong, H. Su, and S. X. Yu, "Learning non-lambertian object intrinsics across shapenet categories," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5844–5853.

[10] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An Information-Rich 3D Model Repository," Stanford University — Princeton University — Toyota Technological Institute at Chicago, Tech. Rep. arXiv:1512.03012 [cs.GR], 2015.

[11] M. Levoy, J. Gerth, B. Curless, and K. Pull, "The stanford 3d scanning repository," 2005, http://www-graphics.stanford.edu/data/3dscanrep.

[12] "The joyful yell," 2015, https://www.thingiverse.com/thing:897412.

[13] K. Khoshelham and S. O. Elberink, "Accuracy and resolution of Kinect depth data for indoor mapping applications," *Sensors*, vol. 12, no. 2, pp. 1437–1454, 2012.

[14] Q. Yang, R. Yang, J. Davis, and D. Nistér, "Spatial-depth super resolution for range images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[15] J. Xie, R. S. Feris, and M.-T. Sun, "Edge-guided single depth image super resolution," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 428–438, 2016.

[16] R. Or-El, G. Rosman, A. Wetzler, R. Kimmel, and A. Bruckstein, "RGBD-Fusion: Real-Time High Precision Depth Recovery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5407–5416.

[17] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE signal processing magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[18] B. Shi, Z. Mo, Z. Wu, D. Duan, S. K. Yeung, and P. Tan, "A benchmark dataset and evaluation for non-lambertian and uncalibrated photometric stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 271–284, 2019.

[19] Y. Quéau, T. Wu, F. Lauze, J.-D. Durou, and D. Cremers, "A Non-Convex Variational Approach to Photometric Stereo under Inaccurate Lighting," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 350–359.

[20] L. Chen, Y. Zheng, B. Shi, A. Subpa-Asa, and I. Sato, "A microfacet-based reflectance model for photometric stereo with highly specular surfaces," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3162–3170.

[21] M. Zollhöfer, A. Dai, M. Innman, C. Wu, M. Stamminger, C. Theobalt, and M. Nießner, "Shading-based refinement on volumetric signed distance functions," *ACM Transactions on Graphics*, vol. 34, no. 4, pp. 96:1–96:14, 2015.

[22] R. Maier, K. Kim, D. Cremers, J. Kautz, and M. Nießner, "Intrinsic3D Dataset," 2017, http://vision.in.tum.de/data/datasets/intrinsic3d.

[23] ——, "Intrinsic3d: High-quality 3D reconstruction by joint appearance and geometry optimization with spatially-varying lighting," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 3114–3122.

[24] T.-W. Hui, C. C. Loy, and X. Tang, "Depth map super-resolution by deep multi-scale guidance," in *Proceedings of European Conference on Computer Vision (ECCV)*, 2016, pp. 353–369.

[25] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs, "SfSNet: Learning Shape, Refectance and Illuminance of Faces in the Wild," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6296–6305.

[26] Y. Quéau, J.-D. Durou, and J.-F. Aujol, "Variational methods for normal integration," *Journal of Mathematical Imaging and Vision*, vol. 60, no. 4, pp. 609–632, 2018.

[27] M. Werlberger, W. Trobin, T. Pock, A. Wedel, D. Cremers, and H. Bischof, "Anisotropic Huber-L1 Optical Flow," in *Proceedings of the British Machine Vision Conference*, 2009, pp. 108.1–108.11.

[28] M. Unger, T. Pock, M. Werlberger, and H. Bischof, "A convex approach for variational super-resolution," in *Joint Pattern Recognition Symposium*, 2010, pp. 313–322.

[29] T. Papadhimitri and P. Favaro, "A closed-form, consistent and robust solution to uncalibrated photometric stereo via local diffuse reflectance maxima," *International Journal of Computer Vision*, vol. 107, no. 2, pp. 139–154, 2014.