Real-time Multi-material Reflectance Reconstruction for Large-scale Scenes under Uncontrolled Illumination from RGB-D Image Sequences

Lukas Bode, Sebastian Merzbach, Patrick Stotko, Michael Weinmann, Reinhard Klein
Institute of Computer Science II
University of Bonn
{lbode,merzbach,stotko,mw,rk}@cs.uni-bonn.de

Abstract

Real-time reflectance reconstruction under uncontrolled illumination conditions is well-known to be a challenging task due to the complex interplay of scene geometry, surface reflectance and illumination. Nonetheless, recent works succeed in recovering both unknown reflectance and illumination in an uncontrolled setting. However, they are either limited regarding the scene complexity (single objects / homogeneous materials) or are not suitable for real-time applications. Our proposed method enables the recovery of heterogeneous surface reflectance (multiple objects and spatially varying materials) in complex scenes at real-time frame rates. We achieve this goal in the following way: First, we perform a 3D scene reconstruction from an input RGB-D stream in real-time. We then use a deep learning based method to estimate Ward BRDF parameters from observations gathered from individual segmented scene objects. Subsequently we refine these reflectance parameters to allow for spatial variations across the object surfaces. We evaluate our method on synthetic scenes and successfully apply it to real-world data.

1. Introduction

The digitization of scenes belongs to the classical computer vision tasks with numerous applications in entertainment, advertisement, cultural heritage as well as virtual and augmented reality. However, achieving realistic models relies on the accurate capture of the underlying properties such as geometry and reflectance characteristics which is complicated by the fact that only the interplay between surface geometry, material-specific reflectance characteristics and illumination conditions can be directly measured. Additional real-time constraints further complicate this task.

Regarding the separate real-time reconstruction of 3D scene geometry, impressive results have been reported with the aid of consumer RGB-D sensors such as the Kinect [31, 5, 43, 44, 12, 13, 6]. The decoupling of reflectance and illumination characteristics, however, remains a highly ill-posed challenge due to its severely under-constrained nature. As a result, many real-time reconstruction approaches rely on strong simplifications, such as using simple color textures to represent surface appearance. However, representing a surface point using a single color value is not sufficient. One needs to take into account that color observations incrementally captured for it may strongly vary due to view- and illumination dependent shadows or high-frequency illumination characteristics. Otherwise, such effects would be stored in the surface texture, which would lead to inconsistencies for scene relighting. To improve the quality of the reflectance reconstruction by separating the aforementioned effects in real-time, existing works exploit intrinsic image decomposition for (diffuse) albedo estimation [16, 11, 29, 26, 40]. These techniques achieve real-time capabilities at a reduced reconstruction accuracy. In contrast, estimating BRDF models together with the surrounding illumination with inverse rendering frameworks yields more accurate reconstructions that also take specular reflectance into account. Inverse rendering approaches utilize alternating optimizations of reflectance and illumination based on statistical priors [39, 22, 21, 3, 47, 23, 24, 37, 2]. However, the computational burden of these approaches prevents real-time performance. Other approaches have recently demonstrated impressive real-time reconstructions by leveraging markers and mirror spheres [48] or by using the potential of deep learning, even in the absence of HDR inputs [17, 28]. However, remaining limitations include the restriction of these BRDF estimation frameworks to single objects with homogeneous reflectance characteristics.

In this paper, we address these limitations by proposing a novel multi-material reflectance reconstruction framework for large-scale scenes with spatially varying surface characteristics under uncontrolled indoor illumination. This implies taking into account near-field illumination characteristics and extending previous frameworks [17, 28] to handle inhomogeneous reflectance characteristics as well

 The final version of record is available at http://dx.doi.org/10.1109/3DV.2019.00083.

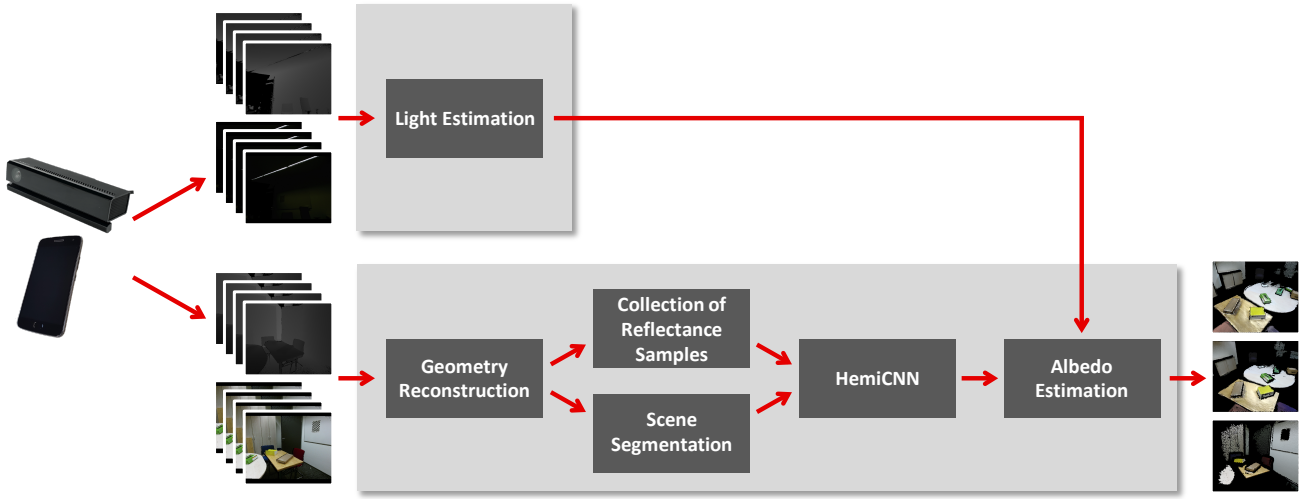

Figure 1. Overview of the proposed real-time multi-material acquisition approach.

as multiple materials in large scenarios in real-time. For this purpose, we capture near-field illumination characteristics, initially assuming that the illumination conditions in indoor scenarios remain constant during capture. In addition, the use of scene segmentation allows to associate the individual reflectance measurements to segments of homogeneous reflectance characteristics, so that within-segment observations can be exploited for the estimation of local surface reflectance behavior. In a final step, we estimate multi-material reflectance characteristics in terms of spatially varying parameters of the Ward BRDF based on the collected measurements utilizing the HemiCNN [17] with a subsequent refinement of diffuse albedo characteristics to allow handling spatially varying characteristics. Our evaluation demonstrates the potential of our approach in the scope of synthetic and real-world examples.

2. Related Work

Early work on separating reflectance and illumination includes in particular the intrinsic image decomposition [4], where an input image is decomposed into the product of a shading layer and a reflectance layer, and its numerous improvements since that time. However, the underlying representation based on two images is disadvantageous as the reflectance layer only represents the diffuse component while the specular component is stored together with the lighting in the shading layer.

Assuming known geometry, Haber et al. [10] and Diaz and Sturm [7] estimate Lambertian reflectance and illumination characteristics from images taken under uncontrolled conditions. Barron and Malik [3] estimate shape, reflectance, and illumination from a single image. Furthermore, using video frames as input, Dong et al. [8] exploit the knowledge regarding surface geometry of a rotating object to estimate spatially varying reflectance behavior and Palma et al. [33] captured SVBRFs while surrounding the object and approximating the environment with a few domination point light sources. In contrast, Wu and Zhou [48] applied the Kinect sensor as an active reflectometer in the IR spectrum and separately captured the illumination in the scene, which allows scanning the object geometry and appearance within several minutes while providing interactive visual feedback. Similarly, Knecht et al. [18] also explored the Kinect to estimate reflectance characteristics at interactive rates. In further work [47], color and depth images captured under unknown illumination serve as input to an offline joint optimization of camera poses, materials, illumination, and surface normals. On-the-fly reflectance estimation at interactive rates for objects exhibiting a homogeneous smooth surface reflectance behavior has been achieved by Kim et al. [17] based on a learned model trained on synthetic data. Solely considering flat material samples, Aittala et al. [1] exploit self-similarities in the surface reflectance behavior to fit spatially-varying BRDFs over a detailed normal map based on a flash/no-flash image pair depicting a flat material sample. Furthermore, Li et al. [19] infer BRDF characteristics for single images based on self-augmented convolutional neural networks.

Instead of assuming known surface geometry, several techniques [36, 9, 20, 25] use implicit shape priors and, hence, are tailored to objects used during the training.

Lombardi and Nishino [21, 24] employ priors for the reflectance model to extrapolate non-observed measurements in combination with illumination priors to jointly optimize for the reflectance and illumination characteristics. In subsequent work [23], this has been further improved to also handle complex scene appearance beyond single isolated objects. In all these approaches the considered objects are

assumed to exhibit a smooth, homogeneous reflectance behavior and real-time performance has not been reached. Another offline estimation approach for illumination and material properties tailored to in the wild conditions has been proposed by Richter-Trummer et al. [38].

The recent work of Meka et al. [27] has been demonstrated to allow for live reflectance estimation from single images without assuming the aforementioned priors. This has been achieved based on the coupling of various encoder-decoder architectures to derive object segmentation, as well as more detailed reflectance information. However, the approach is tailored to the capture of single objects with homogeneous reflectance characteristics.

3. Multi-material Reflectance Estimation in Large Scenes from RGB-D Sequences

As illustrated in Figure 1, our framework for real-time multi-material reflectance reconstruction takes inputs in terms of RGB-D streams from commodity depth sensors such as the Microsoft Kinect or respective RGB-D sensors in smartphones. In an initial step, we recover the illumination characteristics in the scene (see Section 3.2). Thereby, we avoid the need for special calibration targets such as chrome spheres as used by Wu and Zhou [48]. Based on the initial illumination reconstruction, we then perform a real-time reflectance reconstruction by gathering view- and illumination-dependent observations for each surface point (Section 3.4), segmenting the scene into different objects (Section 3.5), and estimating material reflectance characteristics in terms of specular (Section 3.6) and diffuse albedo (Section 3.7). In Section 3.1, we first review the underlying reflectance representation and subsequently provide more details regarding the major components of our framework.

3.1. Image Formation and Reflectance Models

Before addressing the inverse rendering problem in terms of inferring surface reflectance characteristics, we briefly focus on the underlying image formation process that describes the light exchange at surfaces as described by the rendering equation [15]:

$$L_o(x,\omega_o) = L_e(x,\omega_o) + \int_{H_i} f_*(\omega_i,x,\omega_o) L_i(x,\omega_i) \cos\theta_i \, d\omega_i. \quad (1)$$

The radiance L_o leaving some point x into direction ω_o is composed of the radiance L_e emitted from that point into direction ω_o, and the integral over the radiance L_i, incident at x from directions ω_i in the domain H_i, that gets reflected into direction ω_o according to a material-specific reflectance model f_*, weighted by the cosine of the angle between ω_i and the surface normal. Assuming that an object does not emit light on its own, we can ignore L_e.

In order to capture surface appearance, we have to recover the underlying reflectance, which is a severely ill-posed task difficult to solve in real-time. Therefore, following previous work, we assume that reflectance can be sufficiently described with parametric BRDF models [23, 47, 17]. Similar to Kim et al. [17], we use the Ward BRDF model [42]

$$f_{BRDF}(\omega_i,x,\omega_o) = \frac{\kappa_d(x)}{\pi} + \frac{\kappa_s(x)}{N} \cdot e^{\gamma}, \quad (2)$$

$$N = 4\pi\alpha^2 \sqrt{\cos\theta_i \cdot \cos\theta_o}, \quad (3)$$

$$\gamma = -\frac{\tan\theta_h{}^2}{\alpha^2}, \quad (4)$$

as it can be seen as a trade-off between simplicity and the capability to represent a wide range of materials, and has been used in the domain of material perception [34, 46]. Here, κ_d denotes the diffuse and κ_s the specular albedo. The parameter α describes the surface roughness. Another common assumption is that each scene object consists of a single homogeneous material, such that it can be sufficiently described by the 7-dimensional Ward parameters. However, since very few real-world objects follow this assumption, we relax this assumption by performing a spatially varying albedo refinement. Finally, we ignore all indirect illumination effects like self-shadowing or interreflections.

3.2. Lighting Estimation

Knowledge of the illumination conditions in the scene facilitates the estimation of surface reflectance behavior and has been addressed e.g. by using special calibration targets, such as mirroring spheres, in front of the moving camera [48]. As we focus on indoor scenarios, we have to capture near-field illumination. Since time-of-flight sensors (e.g. the Microsoft Kinect v2) are not able to measure depth for mirror-like surfaces, we instead record illumination characteristics using a separate RGB-D image sequence capturing the light sources by direct observation. During this first recording, the sensor is configured to use a low exposure in order to achieve a clear separation of light sources from the remaining scene contents in the RGB images. Note that we do not need an additional RGB-D sensor as both image sequences can be recorded sequentially. We back project pixels of the RGB images with a luminance above a given threshold according to the corresponding depth data and apply a simple spatial mean-shift clustering for each frame individually. Fusing the resulting per image point light candidates over the whole sequence yields the final illumination configuration. Alternatively, voting-based approaches could be used [45, 33].

3.3. Geometry Reconstruction

Both the estimation of near-field illumination and reflectance rely on knowledge of the surrounding scene ge-

ometry. We use the VoxelHashing 3D reconstruction framework [32, 14] that allows real-time reconstruction of large scenes. It relies on an implicit voxel-based surface representation adapted to the underlying scene geometry. Instead of allocating voxels for the entire scene volume, a sparse set of voxel blocks managed by spatial hashing is used.

3.4. Local Collection of Reflectance Observations

The inference of surface reflectance characteristics relies on collecting local observations of surface appearance at each surface point under various viewing configurations per voxel and constant illumination conditions. Therefore, an observation is given as a pair of an RGB color value and a direction from which it has been observed. For every voxel in the hash table we determine the corresponding pixel in the depth image. By comparing the depth value with the distance between voxel and camera, we check whether the voxel is corresponding to some pixel in the RGB image or not. If the two values are sufficiently close, we sample the color from the RGB image and store it together with the voxel-to-camera direction as one observation. Observations that are too far from the surface or occluded are discarded.

Similar to the VoxelHashing framework, we store all those observations in a separate large observations pool in GPU memory and access them through a hash table which maps voxel coordinates to a list of observations. Holding the observations in GPU memory allows for efficient highly parallel acquisition and processing. The GPU memory, however, is already in high demand for the geometry reconstruction itself and the machine learning framework running the CNN (Section 3.6). Due to the large number of voxels in the scene and input image sequences that usually contain hundreds of frames, the memory consumption is a very limiting factor for this step. In order to keep the memory, as well as the computational requirements tractable, we introduce some optimizations:

First, we limit the number of stored observations for a single voxel to m, while ensuring that the most important reflectance characteristic are still captured. Therefore, we approximate a uniform sampling over the hemisphere in normal direction by discarding one of the two most similar observations when exceeding the limit after storing a new one. Experimentally we determine $m = 30$ to be a reasonable number of stored observations. This solution represents a trade-off between a low chance of missing valuable specularity information and computational complexity. Since this is a real-time pipeline, we set the focus on performance.

As a second optimization, we work on a coarser voxel grid for anything regarding the reflectance observations. Instead of the usual 8^3 voxels per voxel block used for the geometry reconstruction, we only use 2^3 or 4^3 voxels for a voxel block of the same spatial dimensions in this step. This downsampling is also the reason for using a separate voxel pool and hash table instead of directly integrating the observations in the geometry reconstruction voxel data structure. Separating the reflectance from the geometric observations additionally allows decoupling the geometry reconstruction from the material estimation framework.

3.5. Segmentation

Estimating multi-material reflectance is complicated by the fact that different materials may seem similar under certain viewing and illumination configurations. Instead of performing a color-based segmentation that may not distinguish material clusters correctly and connect distant dissimilar regions, we assume that the scene contains multiple objects with locally homogeneous materials. We therefore apply the depth-based segmentation by Tateno et al. [41]. It is based on the assumption that most objects have convex shapes, and thus tend to be separated by concave boundary regions in the depth maps. The concave regions are computed using the relative normal orientations from the depth maps and are segmented using connected component analysis. In addition, we exploit the temporal coherence of such regions over image sequences to make the segmentation consistent over time.

For further processing we need to be able to randomly sample voxels of a specific segment. In order to do this, we allocate a ring buffer of fixed size per material class, which is filled with voxel references utilizing the GPU.

3.6. Specular Material Parameter Estimation

For the material estimation, we assume every extracted segment to correspond to a region with homogeneous material characteristics. We thus have to predict one set of material parameters for the voxels assigned to a specific segment. For this purpose, we use the HemiCNN [17] to estimate specular albedo κ_s and the Ward roughness parameter α. While we use κ_s and α as provided by the HemiCNN, we use a novel albedo refinement technique to compute the diffuse albedo κ_d to increase robustness against violations of our homogeneity assumption, see Section 3.7.

In a first step of the estimation process, for every segment, we loop over its ring buffer containing the segment's voxels and randomly sample 25 of them. Per segment, we use those sampled voxels to create so called HemiImages from their reflectance observations. The observations' directions are rotated such that the z-axis is aligned with the surface normal, which is stored together with the reflectance observations. This results in the observations all being contained by the hemisphere in positive z direction. All directions are now projected onto the x-y-plane such that they are contained in the unit disk around the origin. To better preserve information under flat angles, we use a parabolic mapping instead of the orthogonal projection suggested by

Kim et al. [17]. The disk containing the projected observation directions is transformed to the range $[0; 14]^2$. We subsequently use nearest neighbor interpolation on the observed colors to fill the pixel grid of 15×15 images. The created HemiImages are used as the input for HemiCNN.

We use a variation of the RMSE2 [17] as loss, i.e.

$$E(w, \hat{w}) = \lambda_d \left\| \begin{pmatrix} \lambda_l L - \lambda_l \hat{L} \\ a - \hat{a} \\ b - \hat{b} \end{pmatrix} \right\|_2^2 + \left\| \begin{pmatrix} c_r - \hat{c_r} \\ c_g - \hat{c_g} \\ c_b - \hat{c_b} \\ d - \hat{d} \end{pmatrix} \right\|_2^2 \quad (5)$$

with $w = (L, a, b, c_r, c_g, c_b, d)$ being the ground truth Ward parameters in a perceptually linear representation [35] and $\hat{w}$ analogously being the estimated parameters. For lower values for λ_d and λ_l the network focuses more on the specular estimation. Therefore, different than Kim et al. ($\lambda_d = \lambda_l = 1$), we use $\lambda_d = 0.1$ and $\lambda_l = 0.3$.

Since we estimate the scene's materials in real time, we have to run the material estimation step each frame. In order to reduce the susceptibility to noise in the individual material estimates, we fuse the material parameters over time.

Inspired by the truncated signed distance function (TSDF) update formula used in KinectFusion [30], we use an average over the materials for the single frames to temporally fuse local material parameter estimates, with a higher weight for current observations. However, instead of dividing by the number of material predictions after summing them up, we clamp the divisor (in our pipeline to 60).

3.7. Albedo Refinement

Applying the HemiCNN in a per-segment manner yields homogeneous diffuse and specular characteristics per segment. In order to relax this and address inhomogeneous reflectance characteristics, we refine the diffuse albedo while keeping the other modalities fixed, thereby allowing spatially varying surface appearance according to the observations in the voxel grid resolution.

Based on Equations 1 and 2, the k-th reflectance observation for one voxel can be expressed as

$$B_k = \sum_l \left(\frac{\kappa_d}{\pi} + \frac{\kappa_s}{N_l} \cdot e^{\gamma_l} \right) \cdot L_l \cdot \cos\theta_{i,l}, \quad (6)$$

where N_l, γ_l and $\theta_{i,l}$ are respectively the variables N, γ and θ_i for light source l. Solving for κ_d yields

$$\kappa_d = \pi \cdot \frac{B_k - \kappa_s \cdot \sum_l \frac{1}{N_l} \cdot e^{\gamma_l} \cdot L_l \cdot \cos\theta_{i,l}}{\sum_l L_l \cdot \cos\theta_{i,l}}. \quad (7)$$

For every frame we use the observations per voxel to recalculate the respective per-voxel diffuse albedo κ_d. Due to the approximately uniform sampling of the observations' directions, we achieve a high degree of temporal coherence by simply averaging the single estimates. The artifacts introduced by the rather low resolution of the observation voxel grid are reduced by applying trilinear interpolation.

4. Evaluation

After providing implementation details, we evaluate our technique for both synthetic and real-world scenarios.

4.1. Implementation Details

We performed all experiments using an Intel Core i7-4930K with 32 GB RAM and an Nvidia GeForce GTX 1080 with 8 GB VRAM. Following standard indoor 3D reconstruction approaches, we use a 3D space discretization with a resolution of 5 mm for the reconstructed model. Furthermore, we use grid resolutions of 2 cm, and 1 cm for the reflectance observations.

The data we use for training the HemiCNN is based on the SynBRDF [17] dataset. It contains 4432 RGB-D image sequences with 100 synthetic images per sequence, which show a single Ward-shaded object from different perspectives, illuminated using various environment maps. Due to its synthetic character, we know the ground truth Ward material parameters. The scenes are divided into 3574 training, 424 validation, and 434 test scenes. We use those images together with our previously described pipeline to create 500 HemiImages per sequence. Afterwards we sample 200 different random sets of 25 HemiImages to create 886400 labeled examples, on which we train our network. We train the HemiCNN in TensorFlow with the Adam optimizer, using a learning rate of 0.0001 and 150k batches, containing 32 examples each.

4.2. Synthetic Data

For synthetic data, we have direct access to the ground truth camera trajectory, segmentation, and material parameters. Our test scenarios consist of objects in a virtual scene and a camera moving around them in an oscillating manner. To generate such scenes, we utilize an OpenGL rasterization engine.

The benefits of our improved HemiCNN are shown in Figure 2. Using our modified HemiCNN allows to reconstruct the specular material characteristics more precisely (e.g. on the yellow bunny). Furthermore, our albedo refinement integrated into the material reconstruction pipeline also allows to reconstruct spatially varying diffuse albedos. A qualitative evaluation in Figure 3 shows that shading effects are mostly avoided in the refined diffuse albedo maps. Furthermore, the re-renderings with and without albedo refinement match the input RGB images closely for the cube sequence, where the individual objects are homogeneous. Figure 4 compares our reconstructed parameters to the ground truth on a synthetic scene.

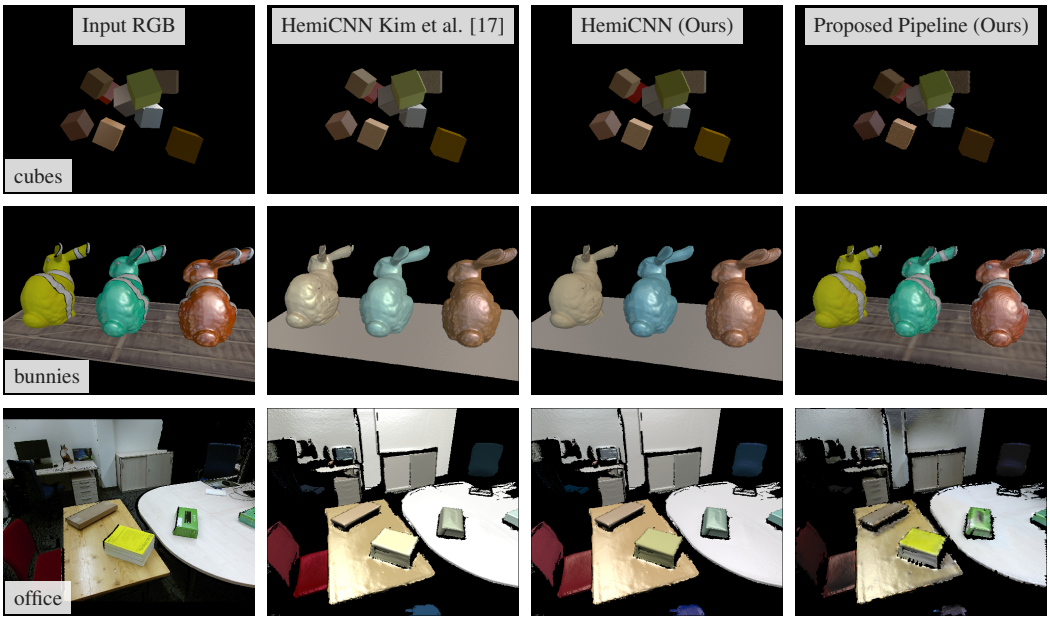

Figure 2. Comparison of our approach with the unmodified HemiCNN [17] on the cubes, bunnies, and office scenes. The first column shows the input RGB images, while the other columns show re-rendered RGB images reconstructed by the unmodified HemiCNN, HemiCNN with our proposed modifications, and our complete pipeline respectively. The reconstructions on the synthetic scenes use ground truth segmentation in order to focus the comparison on the material estimation aspect.

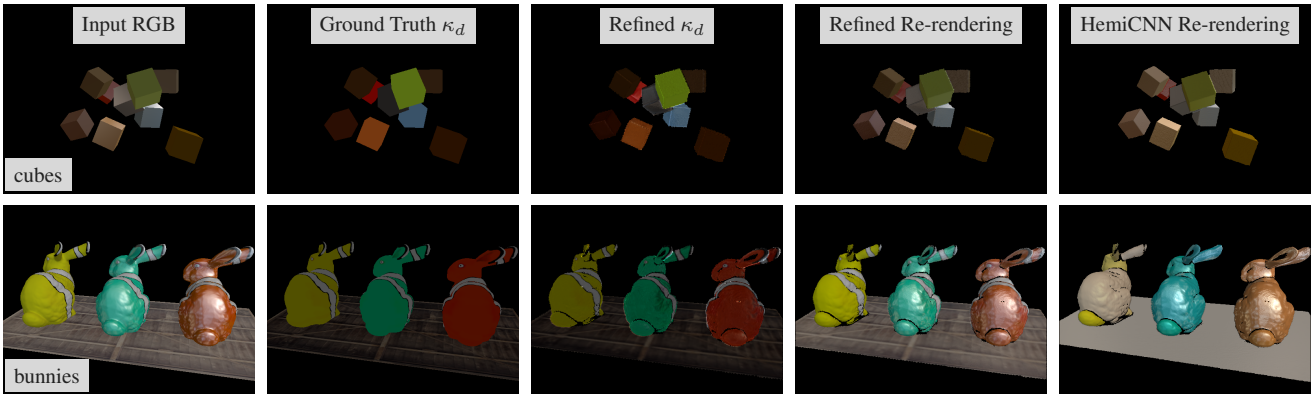

Figure 3. Results for two synthetic datasets: Image of the input sequence, ground truth diffuse albedo, refined diffuse albedo, scene re-rendering using all of the estimated Ward parameters and scene re-rendering using the diffuse albedo output of the HemiCNN directly (from left to right). In both cases, the distance between the scene's center and the camera is 4 m. The cubes have an edge length of 0.4 m and the bunnies have a height of 1.5 m.

The albedo refinement is particularly favorable for scenarios where the assumption of homogeneous materials is violated. Oscillations are induced by different viewpoints in the images. Additionally, the results in Figure 5 illustrate the influence of the albedo refinement for objects with spatially varying reflectance behavior. Further results regarding various error metrics are shown in Table 1.

4.3. Real-world Data

To demonstrate the performance of our technique on real-world data, we captured an indoor scene that contains a multitude of objects with different reflectance characteristics. For the RGB-D capturing of real-world scenes, we use the Microsoft Kinect v2 that delivers images with a resolution of 512×424 pixels at 30Hz.

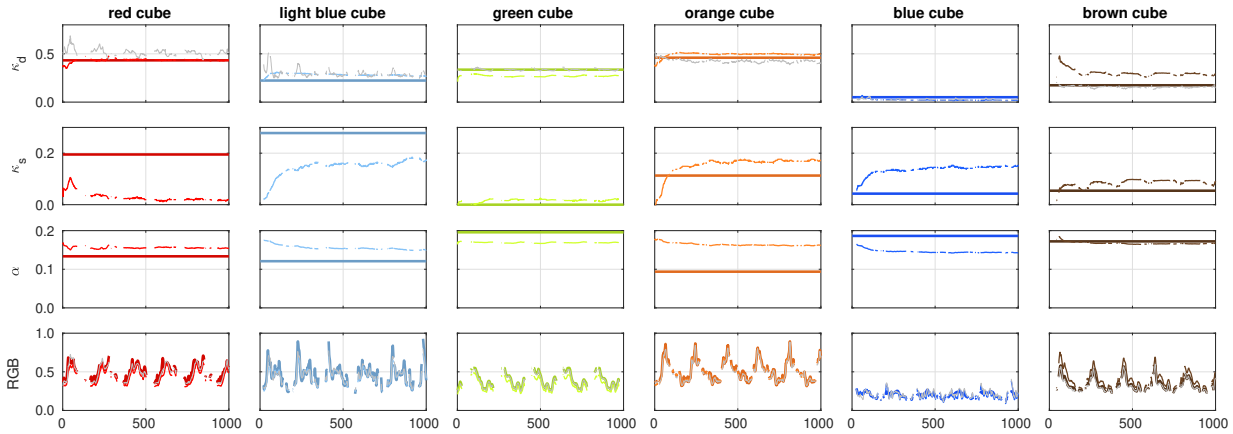

Figure 4. Quantitative evaluation over time (1000 frames) by comparison with ground truth BRDF parameters on a synthetic test scene over multiple objects. The first three rows show diffuse κ_d and specular albedo κ_s, as well as roughness parameter α. Plotted in bold are the ground truth BRDF parameters that are constant for each object (to avoid clutter, we plot only the red channel for κ_d and κ_s), the bottom row shows the red component of the rendered RGB color, where the bold line again represents ground truth. The thinner and lighter plots show our reconstructions. The gray plots show refined parameters. Gaps in the plots are caused by occlusions.

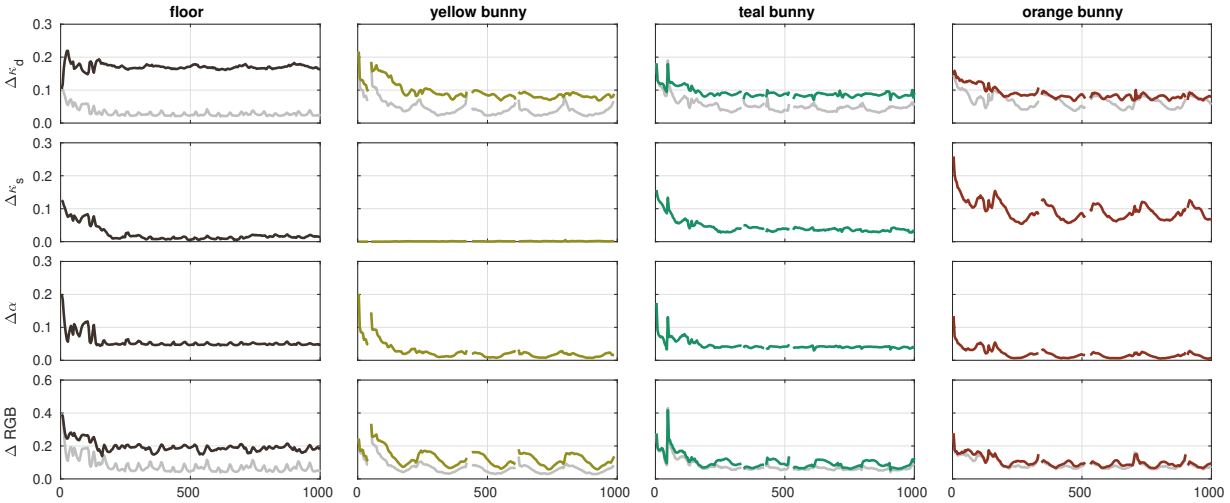

Figure 5. Quantitative evaluation over time (1000 frames) for the bunny dataset: L_1 errors (normalized over object region) of diffuse albedo (first row), specular albedo (second row), Ward roughness α, and final reconstruction error ΔRGB for the individual objects. The gray plots show the errors resulting from the refined model. Gaps in the plots are caused by occlusions.

	Δ_{L_1}					Δ_{L_2}					PSNR					SSIM				
	S1	S2	S3	S4	avg	S1	S2	S3	S4	avg	S1	S2	S3	S4	avg	S1	S2	S3	S4	avg
κ_d	0.17	0.09	0.09	0.09	0.11	0.18	0.13	0.12	0.11	0.13	13.36	19.23	20.80	19.94	18.33	0.89	0.95	0.96	0.93	0.93
$\kappa_{d,\text{ref}}$	0.03	0.05	0.05	0.07	0.05	0.06	0.10	0.09	0.10	0.09	23.57	21.65	22.82	20.75	22.20	0.94	0.97	0.96	0.93	0.95
κ_s	0.02	0.00	0.04	0.09	0.04	0.04	0.01	0.06	0.11	0.05	26.02	45.31	25.41	18.55	28.82	0.97	1.00	0.98	0.97	0.98
α	0.06	0.03	0.05	0.02	0.04	0.07	0.06	0.06	0.04	0.06	17.91	22.45	22.84	25.81	22.25	0.96	0.98	0.98	0.98	0.98
RGB	0.20	0.13	0.10	0.10	0.13	0.21	0.19	0.16	0.16	0.18	21.68	26.15	28.63	27.32	25.95	0.90	0.94	0.96	0.94	0.94
RGB_{ref}	0.08	0.07	0.08	0.09	0.08	0.15	0.14	0.15	0.15	0.15	25.16	28.83	28.96	27.48	27.61	0.92	0.94	0.97	0.95	0.95

Table 1. Different metric results averaged over the 1000 frames of the bunny dataset: mean absolute deviation (MAD, Δ_{L_1}), root mean square error (RMSE, Δ_{L_2}), peak signal to noise ratio (PSNR), structural similarity index (SSIM). The individual metrics are shown separately for the different scene objects (S1: floor, S2: yellow bunny, S3: teal bunny, S4: orange bunny), as well as averaged over the entire scene. The errors are computed on the individual model parameters (κ_d, κ_s and α), as well as the re-renderings (RGB). The second and the last row show the metrics based on the refined diffuse albedo ($\kappa_{d,\text{ref}}$) and corresponding re-renderings (RGB_{ref}). Highlighted in **bold** are the respective better results under each metric, showing that our refinements produce consistent improvements.

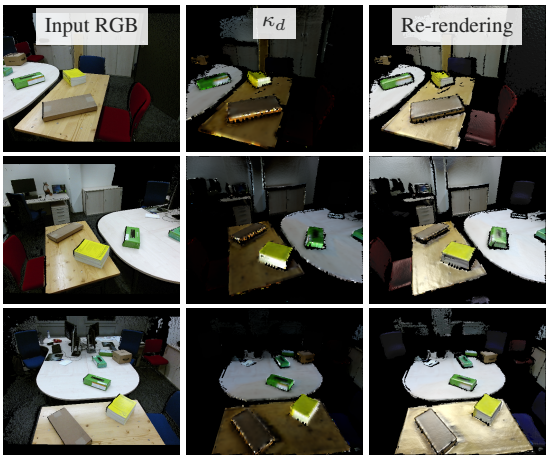

Figure 6. Results for a real-world office scene captured with the Microsoft Kinect v2 sensor: RGB input, estimated diffuse albedo, and Ward shaded re-rendering on three different frames.

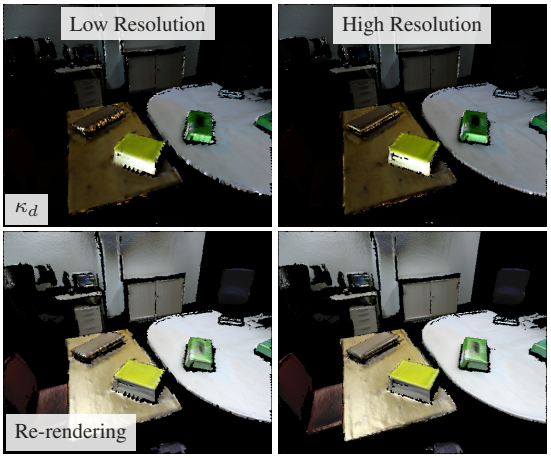

Figure 7. Comparison of results using different resolutions for the reflectance observations' voxel grid: 2 cm (left) and 1 cm (right).

As demonstrated in Figure 6, highlights are separated from the diffuse albedo, and the specular component is preserved in the reconstructed Ward parameters as illustrated by the re-rendering. Furthermore, Figure 7 shows that the reflectance observations' voxel grid resolution has only a minor effect on the re-renderings, albeit the difference being visible in the diffuse albedo maps.

4.4. Performance Evaluation

The timings needed by the individual components of our framework are shown in Table 2 and Table 3. As can be seen, our approach allows real-time material recovery on all low resolution test scenarios, as well as on the simple high resolution ones. Investigations about the α distribution in the data, as well as additional results on synthetic and real-world scenes are shown in the supplementary material.

Scene	cubes	bunnies	office
Geometry Rec.	3.480 ms	5.865 ms	8.906 ms
Refl. Obs. Collection	0.701 ms	1.474 ms	1.624 ms
Segmentation	1.154 ms	1.944 ms	1.755 ms
Specular Mat. Est.	5.997 ms	6.582 ms	6.039 ms
Albedo Refinement	5.962 ms	9.158 ms	11.438 ms
Total	17.294 ms	25.023 ms	29.763 ms

Table 2. Performance of the whole pipeline on various scenes with low (2 cm) voxel grid resolution for reflectance observations.

Scene	cubes	bunnies	office
Geometry Rec.	3.397 ms	5.867 ms	8.997 ms
Refl. Obs. Collection	2.965 ms	8.535 ms	9.967 ms
Segmentation	1.136 ms	1.980 ms	1.756 ms
Specular Mat. Est.	5.978 ms	6.507 ms	5.820 ms
Albedo Refinement	10.442 ms	21.579 ms	28.254 ms
Total	23.919 ms	44.467 ms	54.794 ms

Table 3. Performance of the whole pipeline on various scenes with high (1 cm) voxel grid resolution for reflectance observations.

4.5. Limitations and Future Work

Reconstructing scenes with high dynamic range (HDR) leads to problems with overexposure since consumer-grade RGB-D sensors like the Kinect typically only capture low dynamic range (LDR) images. This limitation could be tackled by augmenting the LDR inputs or reconstructing HDR from LDR images captured under varying exposures.

The sampling of the stored observations could be adapted to better match the object's specularity for sparsly observerd objects such as the chair and the wall in Figure 6. Further improvements include the optimization of our current implementation to allow refining the resolution of the reflectance observations and improving the segmentation by additionally considering albedo information.

5. Conclusion

In this paper, we presented a novel real-time multi-material reflectance reconstruction framework for large-scale scenes with spatially varying surface characteristics under uncontrolled static near-field indoor illumination. After an initial reconstruction of the near-field scene lighting, the framework uses the combination of real-time 3D reconstruction, scene segmentation and per-segment reflectance estimation. As demonstrated, our technique preserves specular characteristics in the estimated material parameters and additionally is capable of handling also spatially varying reflectance characteristics.

Acknowledgements

This work was supported by the DFG projects KL 1142/11-1 (DFG Research Unit FOR 2535 Anticipating Human Behavior) and KL 1142/9-2 (DFG Research Unit FOR 1505 Mapping on Demand).

References

[1] M. Aittala, T. Weyrich, and J. Lehtinen. Two-shot svbrdf capture for stationary materials. *ACM Trans. Graph.*, 34(4):110:1–110:13, July 2015.

[2] R. A. Albert, D. Y. Chan, D. B. Goldman, and J. F. O'Brien. Approximate svbrdf estimation from mobile phone video. In *Proceedings of the Eurographics Symposium on Rendering: Experimental Ideas & Implementations*, SR '18, pages 11–22, Goslar Germany, Germany, 2018. Eurographics Association.

[3] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(8):1670–1687, Aug 2015.

[4] H. G. Barrow and J. M. Tenenbaum. *Recovering Intrinsic Scene Characteristics from Images*. Academic Press, 1978.

[5] J. Chen, D. Bautembach, and S. Izadi. Scalable Real-time Volumetric Surface Reconstruction. *ACM Trans. Graph.*, 32:113:1–113:16, 2013.

[6] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Trans. Graph.*, 36(4), May 2017.

[7] M. Díaz and P. Sturm. Estimating Photometric Properties from Image Collections. *Journal of Mathematical Imaging and Vision*, 47(1-2):93–107, Sept. 2013.

[8] Y. Dong, G. Chen, P. Peers, J. Zhang, and X. Tong. Appearance-from-motion: Recovering spatially varying surface reflectance under unknown lighting. *ACM Trans. Graph.*, 33(6):193:1–193:12, Nov. 2014.

[9] S. Georgoulis, K. Rematas, T. Ritschel, M. Fritz, L. J. V. Gool, and T. Tuytelaars. Delight-net: Decomposing reflectance maps into specular materials and natural illumination. *CoRR*, abs/1603.08240, 2016.

[10] T. Haber, C. Fuchs, P. Bekaert, H. Seidel, M. Goesele, and H. P. A. Lensch. Relighting objects from image collections. In *CVPR*, pages 627–634, 2009.

[11] M. Hachama, B. Ghanem, and P. Wonka. Intrinsic Scene Decomposition from RGB-D Images. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 810–818. IEEE Computer Society, 2015.

[12] O. Kähler, V. Prisacariu, J. Valentin, and D. Murray. Hierarchical Voxel Block Hashing for Efficient Integration of Depth Images. *IEEE Robotics and Automation Letters*, 1(1):192–197, 2016.

[13] O. Kähler, V. A. Prisacariu, and D. W. Murray. Real-Time Large-Scale Dense 3D Reconstruction with Loop Closure. In *European Conference on Computer Vision*, pages 500–516, 2016.

[14] O. Kähler, V. A. Prisacariu, C. Y. Ren, X. Sun, P. Torr, and D. Murray. Very High Frame Rate Volumetric Integration of Depth Images on Mobile Devices. *Visualization and Computer Graphics, IEEE Transactions on*, 21(11):1241–1250, 2015.

[15] J. T. Kajiya. The rendering equation. In *ACM Siggraph Computer Graphics*, volume 20, pages 143–150. ACM, 1986.

[16] C. Kerl, M. Souiai, J. Sturm, and D. Cremers. Towards Illumination-invariant 3D Reconstruction using ToF RGB-D Cameras. In *International Conference on 3D Vision (3DV)*, volume 1, pages 39–46. IEEE, 2014.

[17] K. Kim, J. Gu, S. Tyree, P. Molchanov, M. Nießner, and J. Kautz. A lightweight approach for on-the-fly reflectance estimation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 20–28, 2017.

[18] M. Knecht, G. Tanzmeister, C. Traxler, and M. Wimmer. Interactive brdf estimation for mixed-reality applications. *Journal of WSCG*, 20(1):47–56, June 2012.

[19] X. Li, Y. Dong, P. Peers, and X. Tong. Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Trans. Graph.*, 36(4):45:1–45:11, July 2017.

[20] G. Liu, D. Ceylan, E. Yumer, J. Yang, and J.-M. Lien. Material editing using a physically based rendering network. *IEEE International Conference on Computer Vision (ICCV)*, pages 2280–2288, 2017.

[21] S. Lombardi and K. Nishino. Reflectance and natural illumination from a single image. In *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI*, pages 582–595, 2012.

[22] S. Lombardi and K. Nishino. Single image multimaterial estimation. In *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, CVPR '12, pages 238–245, Washington, DC, USA, 2012. IEEE Computer Society.

[23] S. Lombardi and K. Nishino. Radiometric scene decomposition: Scene reflectance, illumination, and geometry from RGB-D images. In *Fourth International Conference on 3D Vision, 3DV 2016, Stanford, CA, USA, October 25-28, 2016*, pages 305–313, 2016.

[24] S. Lombardi and K. Nishino. Reflectance and illumination recovery in the wild. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(1):129–141, 2016.

[25] M. Maximov, T. Ritschel, and M. Fritz. Deep appearance maps. *CoRR*, abs/1804.00863, 2018.

[26] A. Meka, G. Fox, M. Zollhöfer, C. Richardt, and C. Theobalt. Live User-Guided Intrinsic Video For Static Scene. *IEEE Transactions on Visualization and Computer Graphics*, 23(11):2447–2454, 2017.

[27] A. Meka, M. Maximov, M. Zollhoefer, A. Chatterjee, H.-P. Seidel, C. Richardt, and C. Theobalt. Lime: Live intrinsic material estimation. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[28] A. Meka, M. Maximov, M. Zollhöfer, A. Chatterjee, H.-P. Seidel, C. Richardt, and C. Theobalt. LIME: Live Intrinsic Material Estimation. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 6315–6324, June 2018.

[29] A. Meka, M. Zollhöfer, C. Richardt, and C. Theobalt. Live Intrinsic Video. *ACM Transactions on Graphics (TOG)*, 35(4):109:1–109:14, 2016.

[30] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohi, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Mixed and augmented reality (ISMAR), 2011 10th IEEE international symposium on*, pages 127–136. IEEE, 2011.

[31] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3D Reconstruction at Scale Using Voxel Hashing. *ACM Transactions on Graphics (TOG)*, 32(6):169:1–169:11, 2013.

[32] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3d reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (ToG)*, 32(6):169, 2013.

[33] G. Palma, M. Callieri, M. Dellepiane, and R. Scopigno. A statistical method for svbrdf approximation from video sequences in general lighting conditions. *Computer Graphics Forum*, 31(4):1491–1500, 2012.

[34] F. Pellacini, J. A. Ferwerda, and D. P. Greenberg. Toward a psychophysically-based light reflection model for image synthesis. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 55–64, New York, NY, USA, 2000. ACM Press/Addison-Wesley Publishing Co.

[35] F. Pellacini, J. A. Ferwerda, and D. P. Greenberg. Toward a psychophysically-based light reflection model for image synthesis. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 55–64. ACM Press/Addison-Wesley Publishing Co., 2000.

[36] K. Rematas, T. Ritschel, M. Fritz, E. Gavves, and T. Tuytelaars. Deep reflectance maps. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4508–4516, 2016.

[37] T. Richter-Trummer, D. Kalkofen, J. Park, and D. Schmalstieg. Instant Mixed Reality Lighting from Casual Scanning. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 27–36. IEEE, 2016.

[38] T. Richter-Trummer, D. Kalkofen, J. Park, and D. Schmalstieg. Instant mixed reality lighting from casual scanning. In *2016 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 27–36, Sept 2016.

[39] F. Romeiro and T. Zickler. Blind reflectometry. In K. Daniilidis, P. Maragos, and N. Paragios, editors, *Computer Vision – ECCV 2010*, pages 45–58, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[40] P. Stotko, M. Weinmann, and R. Klein. Albedo estimation for real-time 3d reconstruction using rgb-d and ir data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 150:213–225, 2019.

[41] K. Tateno, F. Tombari, and N. Navab. When 2.5 d is not enough: Simultaneous reconstruction, segmentation and recognition on dense slam. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 2295–2302. IEEE, 2016.

[42] G. J. Ward. Measuring and modeling anisotropic reflection. In *Proceedings of the 19th Annual Conference on Computer Graphics and Interactive Techniques*, pages 265–272, 1992.

[43] T. Whelan, M. Kaess, M. Fallon, H. Johannsson, J. Leonard, and J. McDonald. Kintinuous: Spatially Extended KinectFusion. In *RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras*, 2012.

[44] T. Whelan, M. Kaess, H. Johannsson, M. Fallon, J. J. Leonard, and J. McDonald. Real-time large-scale dense RGB-D SLAM with volumetric fusion. *The International Journal of Robotics Research*, 34(4-5):598–626, 2015.

[45] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger. ElasticFusion: Real-Time Dense SLAM and Light Source Estimation. *The International Journal of Robotics Research*, 35(14):1697–1716, 2016.

[46] J. Wills, S. Agarwal, D. Kriegman, and S. Belongie. Toward a perceptual space for gloss. *ACM Trans. Graph.*, 28(4):103:1–103:15, Sept. 2009.

[47] H. Wu, Z. Wang, and K. Zhou. Simultaneous Localization and Appearance Estimation with a Consumer RGB-D Camera. *IEEE Transactions on Visualization and Computer Graphics*, 22:2012–2023, 2016.

[48] H. Wu and K. Zhou. Appfusion: Interactive appearance acquisition using a kinect sensor. *Comput. Graph. Forum*, 34(6):289–298, Sept. 2015.