DOI: 10.1111/cgf.14707
Pacific Graphics 2022
N. Umetani, E. Vouga, and C. Wojtan
(Guest Editors)

Volume 40 (2021), Number 7

ShadowPatch: Shadow Based Segmentation for Reliable Depth Discontinuities in Photometric Stereo

M. Heep[1] and E. Zell[1]

[1]PhenoRob, University of Bonn, Germany

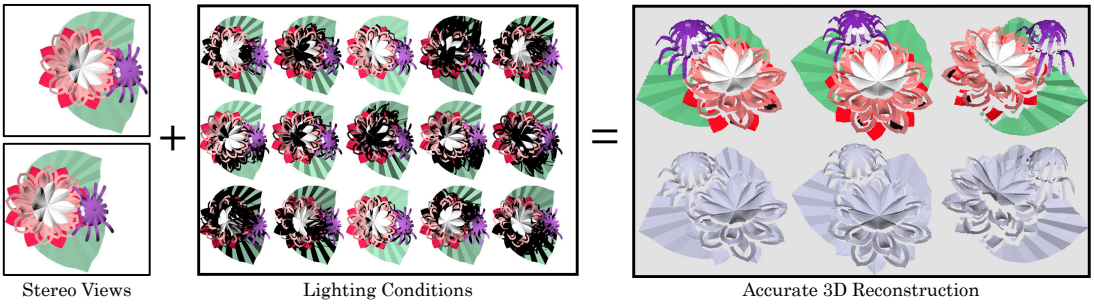


Figure 1: *By combining and deeply integrating depth-from-stereo with photometric stereo, and by exploiting the visual structure introduced by self-shadowing we create accurate 3D reconstructions with accurate handling of depth discontinuities.*

Abstract

Photometric stereo is a well-established method with outstanding traits to recover surface details and material properties, like surface albedo or even specularity. However, while the surface is locally well-defined, computing absolute depth by integrating surface normals is notoriously difficult. Integration errors can be introduced and propagated by numerical inaccuracies from inter-reflection of light or non-Lambertian surfaces. But especially ignoring depth discontinuities for overlapping or disconnected objects, will introduce strong distortion artefacts. During the acquisition process the object is lit from different positions and self-shadowing is in general considered as an unavoidable drawback, complicating the numerical estimation of normals. However, we observe that shadow boundaries correlate strongly with depth discontinuities and exploit the visual structure introduced by self-shadowing to create a consistent image segmentation of continuous surfaces. In order to make depth estimation more robust, we deeply integrate photometric stereo with depth-from-stereo. Having obtained a shadow based segmentation of continuous surfaces, allows us to reduce the computational cost for correspondence search in depth-from-stereo. To speed-up computation further, we merge segments into larger meta-segments during an iterative depth optimization. The reconstruction error of our method is equal or smaller than previous work, and reconstruction results are characterized by robust handling of depth-discontinuities, without any smearing artifacts.

CCS Concepts

• ***Computing methodologies*** → ***Reconstruction;*** *Image segmentation; Shape inference;*

1. Introduction

Accurate reconstruction of 3D shape and appearance is a fundamental research problem with many applications beyond graphics. Photometric stereo and photogrammetry are well-established and among the most popular methods, each coming with its benefits and disadvantages. While photometric stereo is the method of choice to capture stunning surface details as well as accurate albedo textures and material properties, photogrammetry estimates absolute depth more robustly. Not surprisingly, combinations of both methods have been suggested to counteract the shortcomings of each approach, where e.g. shape-from-silhouette [LMC19] serves as a 3D shape proxy. Even recent high-end solutions [GLD*19] design the 3D reconstruction as a two-stage process where a coarse 3D model is reconstructed from depth-from-stereo and surface details are added by photometric stereo.

Algorithms, where the building blocks of photometric stereo and photogrammetry are fully integrated at a deeper level, lifting the true potential of the synergies, are less common. Lightcodes [SNA*21] or photogeomtric scene flow [GSSM15] are among the few examples. While lightcodes extend the pixel-wise matching to all lighting conditions, photogeometric scene flow, replaces the smoothness term in multi-view stereo with the locally defined surface information from photometric stereo. Our work extends previous contributions, by identifying continuous surfaces and depth discontinuities at an early stage via a shadow-based segmentation. Starting from the rather trivial observation that objects located closer to the camera are fully illuminated, but cast shadows on objects further away from the camera, we identified that depth discontinuities correlate strongly with shadow boundaries. To exploit this correlation algorithmically, we divide the image into segments based on the structure introduced through self-shadowing. The derived content sensitive image segmentation is, at least in theory, independent of the image resolution. We take advantage of this compact data structure in our variation of the depth-from-stereo algorithm and match segments instead of pixels.

While in classical photometric stereo self-shadowing is considered as an unavoidable downside, our approach turns it for the better and utilizes the additional information to identify reliably depth discontinuities. Therefore, addressing an important property for more accurate 3D reconstructions from photometric stereo. Results and comparisons show that our method outperforms state of the art methods especially in challenging scenarios for photometric stereo with many depth discontinuities.

2. Related Work

Photometric stereo is best understood as inverse rendering; images taken from the same viewpoint but under different lighting conditions are used to recover parameters such as surface normals, texture albedo and sometimes even specularity. Applications of photometric stereo range from detailed acquisition of small objects [LMC19], humans [VPB*09, GSSM15] to buildings [SSB*14]. An in-depth overview on photometric stereo can be found in a survey [AG15] and we focus in this section on the most relevant work to ours. A serious limitation of photometric stereo is that depth can only be estimated up to a relative scale. A common strategy is therefore to refine the geometry of low-resolution scans, e.g. structured light [NRDR05], laser scanner [BFR14] or low-resolution depth-from-stereo [JK07] with photometric stereo. With the popularity of normal maps in rendering pipelines, it became more practical to compute the normal map and use it directly within the shader for high-quality visualizations [PSM*16, GLD*19].

If the object category is known in advance or is largely convex, a template model (e.g. a generic face) [WGTS13], or an initial 3D model based on shape from silhouette [HVC08, VPB*09, PSM*16, LMC19] is refined by photometric stereo. Hernandez et al. [HVC08] use a setup combining multiview with photometric stereo. Their approach operates directly on a mesh using the visual hull as an initial shape estimate and does not require knowledge about the light or camera positions. Similarly, Vlasic et al. [VPB*09] use the visual hull to get an initial depth estimate which is then refined first by means of photometric stereo and later by matching the surfaces obtained from different views. Different to Hernandez et al. [HVC08] their method addresses depth discontinuities using a threshold value in combination with the visual hull. Other methods use a two-stage approach of multi-view-stereo with photometric stereo [GT15] or refine depth using back-projection and a brute-force approach on the GPU [GWT*18]. Logothetis et al. [LMC19] extend multi-view stereo and photometric stereo to non-Lambertian surfaces. Again starting from a rough shape estimate of their model, they use photometric stereo to directly compute signed distance functions, effectively refining the geometry in each step. Shadow masks are estimated at each iteration to calculate the self-shadowing based on the current geometry.

More similar to our work, Du et al. [DGS11] use two cameras in conjunction with photometric stereo and incorporate a filter based approach which has the benefit of being convex but does not scale well to higher resolutions. Gotardo et al. [GSSM15] combine photometric stereo with stereo vision to reconstruct dense facial animations. Their method is based on a combination of colored lights, to effectively capture three lighting conditions in a single frame, together with optical flow, to combine lighting conditions from neighbouring frames. Quantitative comparison between our method and the work of Gotardo et al. [GSSM15] (Section 6) show that our method is more accurate at depth discontinuities and more reliable for slanted surfaces, while both methods achieve similar performance for smooth, camera-facing surfaces. Some authors focus on special topics in photometric stereo, like robust normal integration without [QDA18, CSFM21] and with additional handling of discontinuities [XWW*19]. Others suggest to increase the number of light conditions for small image sets, relying on RGB color space [GSSM15] and multi-spectral cameras [ZDJ*20]. Recently convolutional neural networks gained attraction within the photometric stereo community, mainly to predict normals for non-Lambertian surfaces, facilitate arbitrary BRDFs [Ike18, LBMC20, SSS*20, YLF*20] or jointly optimize for light positions and normals [KKO*21].

Within the depth-from-stereo field, most publications leading public benchmarks, e.g. [THZ*21] are based on convolutional neural networks (CNN) and a still very recent overview is given by Laga et al. [LJBB20]. While these methods typically deliver a great performance in terms of both, runtime and matching quality, they also come with their downsides: The maximal resolution and baseline is typically limited by available GPU-memory. The underlying 3D convolutions quickly consume a lot of memory. Images of 2000x2000 pixels are considered as big and a baseline of roughly 200px is the standard. In addition, requiring training data with very precise depth/disparity information leads in practice either to using a large data set generated from synthetically generated data (e.g. SceneFlow [MIH*16]) or a small data set obtained from scans of real world objects (e.g. KITTI [GLU12]). Both datasets struggle to represent the virtually infinite variety of image configurations (lighting, occlusions, etc.): The latter simply by number, the former due to the remaining gap between synthetic and real data (domain gap) [SYZ*21].

Image segmentation to improve stereo matching was suggested previously by a number of methods [HC04, KSK06, WL11], all sharing similar principles: After searching for the optimal dispar-

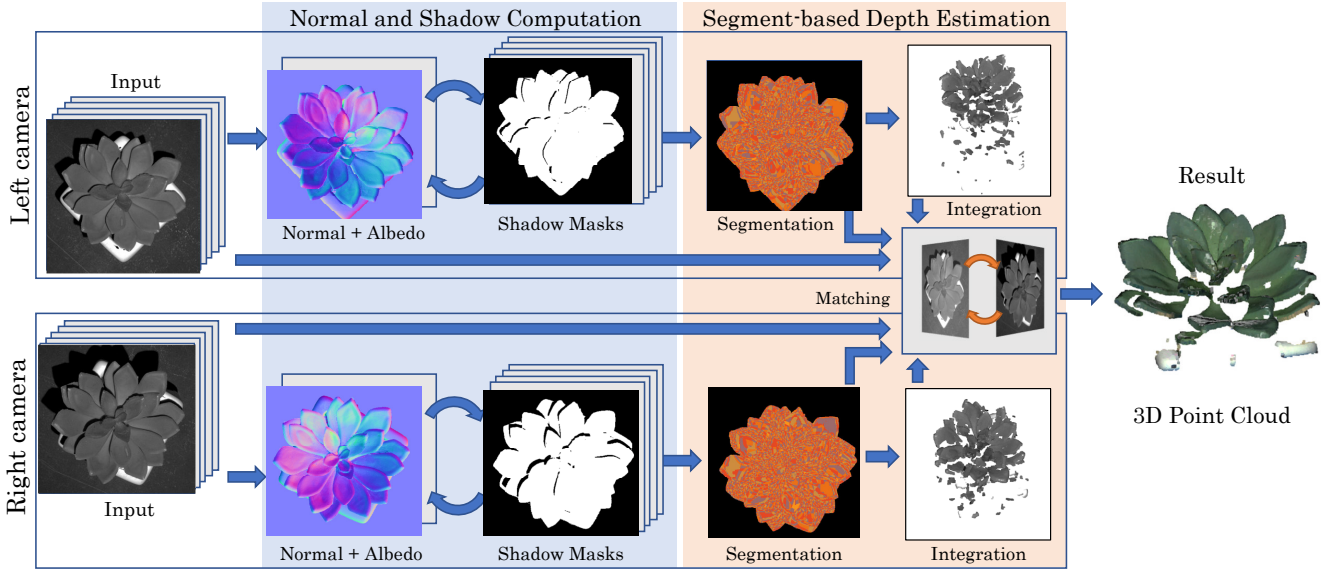


Figure 2: *Overview of our algorithmic flow: For each camera view, we first estimate normals, albedo and shadow masks and separate the image into smaller, semantically plausible segments. We then iteratively refine the depth for each segment by minimizing the integration error and matching energy between botch camera views.*

ity for each individual pixel, a segmentation is constructed based on color uniformity, e.g. using mean-shift filtering, without further taking the 3D shape structure into account. The disparity values in each segment are used to construct (several) disparity planes for each segment. These disparity planes are then propagated to neighbouring segments through graph cuts or believe propagation. A fundamental shortcoming of these algorithms is the assumption of flat segments. PatchMatch [BSFG09, BRR11], overcomes this problem by using a disparity plane per pixel instead of per segment. At each iteration, a set of candidate disparity planes is generated and the best one is adopted. Later extensions [GLS15, LZYZ16, RM19] proposed variations for candidate generation, ranging from estimating based on local neighbourhood, random perturbations or projections from different views. One downside of PatchMatch and similar methods is the missing support to encourage smooth, continuous surfaces.

The most similar depth-from-stereo method to ours, was introduced by Taniai et al. [TMSN17]. Like PatchMatch, it uses pixel-wise disparity planes but iteratively refines them in pixel-blocks via expansion moves, thus overcoming the smoothness prior limitation. By combining photometric stereo with a shadow based segmentation, and iteratively refining the depth over all segments, we reduce the dimensionality of the matching problem. Furthermore, our method outperforms the reconstruction quality even of learning based depth-from-stereo methods. ShadowCut [CAK07] introduces a robust shadow detection for photometric stereo by using graph cuts which we adopt in our approach. Different to ShadowCut, we rely on depth-from-stereo for accurate, absolute depth reconstruction instead of relying on a single underconstrained view. Detecting depth-discontinuities from shadow contours was introduced in the context of non-photorealistic rendering [RTF*04]. Compared to our method, we can report that this approach reduces the initial number of segments even further. However, it is more prone to missing depth discontinuities, causing diminishing reconstruction results. Our more defensive approach to rather over segment at the beginning and quickly merge nearby segments into larger meta-segments seems to be more advantageous in this case.

3. Overview

For our capturing setup, we assume that images are taken under different lighting conditions, by a calibrated stereo camera, implying that camera intrinsics and extrinsics are known. During capturing only one light is switched on for each frame. The intensity of all lights is equal and all lights are placed on a hemisphere around the object, such that the distance between the object and the light is equal as well. Having obtained a set of images with an object lit under different conditions, we compute normals, albedo (Section 4.1), and shadow masks (Section 4.2) within an alternating optimization framework. In a follow-up procedure we estimate depth from stereo views, where we first segment the image based on the structure introduced through self-shadowing (Section 5.1). We then refine the depth of each segment (Section 5.4) relying on the integration of normals from photometric stereo (Section 5.3) and by matching segments between the left and right camera views (Section 5.2). Dividing the image into segments has the advantage that: (a) the search space is defined by the number of segments instead of the number of image pixels, (b) depth discontinuities are detected and handled correctly from the beginning, and (c) large continuous surfaces remain connected during the optimization. Once the segment-wise optimization converged, we address small discontinuities between segments, e.g. due to discretization errors, in an additional post-processing step (Section 5.5).

4. Normal and Shadow Computation

In line with many previous publications on photometric stereo, e.g. [AG15], we assume a Lambertian surface, where the measured light intensity i of a pixel is given by $i = \alpha(\mathbf{l} \cdot \mathbf{n})$. The light vector $\mathbf{l}$, describes both the *direction and the brightness* of the incoming light and the albedo α is a material constant determining the fraction of light that is reflected. The light vector can be sufficiently estimated from the light positions, which implies that the light positions are known for the setup or were obtained in advance during a calibration procedure. In general, each color channel of a pixel has its own albedo, but for normal reconstruction, using multiple color channels has no advantage over using just a single channel. When the pixel-wise intensity is measured in RGB space, we consider only the luminance of the YUV color space for normal computations to save memory and derive the RGB albedo *after* the 3D reconstruction (Section 5.5).

We summarize pixel-wise intensities under different light conditions within a single intensity vector $\mathbf{i} = (i_1, ..., i_L)^t$ and introduce a binary shadow mask $\mathbf{s} = (s_1, ..., s_L)^t$ to exclude pixels in shadow from normal estimation, where

$$s_l = \begin{cases} 0 & \text{if the pixel is in shadow for light } l, \\ 1 & \text{else.} \end{cases}$$

Especially on discontinuous surfaces, that we aim to reconstruct, self-shadowing is very present and must be addressed explicitly to retain the assumptions of the Lambertian reflectance model.

4.1. Normal Update

To compute shadows, normals and albedo, we employ an alternating optimization method, where we exploit the observation that binary shadows spread across several pixels (Section 4.2), while normals and albedo change more from pixel to pixel. When estimating the normals and albedo, the shadow mask remains fixed. The pixel-wise energy term minimizes, the squared error between the shading model prediction and the observed pixel intensities

$$\Psi(\alpha, \mathbf{n}, \mathbf{s}) = \frac{1}{2\sigma_i^2} \sum_{l \in \mathcal{L}} \left(i_l - s_l \alpha_p (\mathbf{l}_l \cdot \mathbf{n})\right)^2 , \tag{1}$$

where σ_i can be understood as a scale for image noise. We estimate image noise through the mean squared deviation between neighbouring pixels [BJ01] with $\mathcal{N}$ being the set of all pairs of adjacent pixels.

$$\sigma_i^2 = \frac{1}{|\mathcal{N}|} \sum_{(p,q) \in \mathcal{N}} \|\mathbf{i}_p - \mathbf{i}_q\|^2 \tag{2}$$

By introducing an auxiliary variable $\mu := \alpha\mathbf{n}$, that summarizes the albedo and the normal, Eq. (1) is minimized. For a given shadow mask $\mathbf{s}$ finding the optimum

$$\mu^* = \underset{\mu \in \mathbb{R}^3}{\operatorname{argmin}} \Psi\left(\|\mu\|, \frac{\mu}{\|\mu\|}, \mathbf{s}\right)$$

is equivalent to solving a linear equation. The albedo and normal can be recovered through:

$$\alpha = \|\mu^*\| \qquad\qquad \mathbf{n} = \frac{\mu^*}{\|\mu^*\|}$$

4.2. Shadow Detection

To make shadow detection more robust than simple thresholds over intensity [GSSM15], we exploit the observation that binary shadows spread across several pixels and extend the energy term from Eq.(1) by a neighborhood constraint, where $\|\cdot\|_1$ is the L^1-norm or Hamming distance. This term vanishes if the shadow masks of adjacent pixels are identical, thus favoring smooth shadow masks.

$$E_{\text{Photo}}(\mathbf{N}, \mathbf{S}) := \sum_{p \in \mathcal{P}} \Psi_p(\alpha_p, \mathbf{n}_p, \mathbf{s}_p) + \lambda \sum_{(p,q) \in \mathcal{N}} \omega_{pq} \|\mathbf{s}_p - \mathbf{s}_q\|_1 . \tag{3}$$

The matrix $\mathbf{N} = (\mathbf{n}_1, ..., \mathbf{n}_P)$ contains all pixel-wise normals and similarly for $\mathbf{S}$ and all pixel-wise shadow masks. In our experiments, $\lambda = 5$ turned out to be a good choice. In line with previous work, e.g. [BJ01, RKB04, TMSN17], we introduce a similarity weight between two adjacent pixels p and q

$$\omega_{pq} := \max\left(\exp\left(-\frac{1}{2\sigma_i^2}\|\mathbf{i}_p - \mathbf{i}_q\|_2^2\right), \omega_{\min}\right), \tag{4}$$

encouraging smoothness especially in uniform image regions. Please note that this pairwise term does *not* include the surface normals and has therefore no influence on the normal update. In our experiments we used $\omega_{\min} = 0.05$.

To find the global minimum, we adopt the ShadowCut method [CAK07] based on graph cuts. Ψ_p acts as the so-called unary term for each pixel: If the pixel is lit, we expect agreement with the Lambertian surface model; if the pixel is in the shadow its intensity should be next to zero. For fixed albedos and normals, Eq. (3) turns into a submodular boolean optimization problem and an optimal solution can be found via graph cuts [KZ04]. Since every shadow mask is coupled to a single lighting condition the graph cuts are computed in parallel for each light condition using the implementation by [BK04].

We combine the albedo-normal update Eq. (1), and the shadow detection Eq. (3) into a joint optimization scheme. At initialization, all shadow masks are set to one, assuming there are no shadows. We then iterate between the shadow detection and the joint albedo-normal update. The optimization terminates typically after five iterations. Other initialization values for shadow masks, converged after a comparable number of iterations to similar results.

5. Segment-based Depth Estimation

Matching pixels across two cameras is a fundamental problem in depth-from-stereo estimation, where the search space complexity is increased by image resolution. In photometric stereo robust identification of discontinuities is key for accurate 3D reconstructions. We introduce a shadow based image segmentation, by leveraging the structure created by self-shadowing. Aggregating pixels into segments, reduces the matching complexity between image pairs, while we reliably identify depth discontinuities from the segment boundaries. In theory, the number of created segments will depend only on the number of lights and is independent of the actual image resolution. Due to rasterization artifacts of the pixel-wise defined image, the initial number of segments is smaller then the theoretical limit and converges to the upper limit with the image resolution.

Dataset	Pixels	Segments	Meta-Segments
Artichoke	163 k	5.5 k	598
Female	1.6 M	17 k	54
Male	2.0 M	23 k	181
Nose	79 k	1.5 k	145
Origami	187 k	14 k	511
Succulent plant	100 k	5.4 k	295

Table 1: *Comparison between the number of pixels, segments and the number of remaining meta-segments after the expansion moves. Please note the drastic reduction between segments and pixels and then again between segments and meta-segments.*

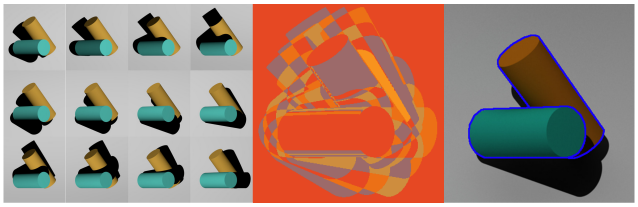


Figure 3: *A simple example illustrating the semantic image segmentation. Starting from twelve, differently illuminated images (left), an image segmentation mask is computed (center), where the segment boundaries, fully cover depth discontinuities (right).*

In our experiments we achieve an impressive reduction of complexity, where the number of segments is 10-100 times smaller than the number of pixels. By merging segments to meta-segments during the optimization, we reduce the matching complexity by an additional factor of 10 to 300 (Table 1).

5.1. Shadow Based Image Segmentation

A core contribution of our work, is the exploitation of self-shadowing to reveal continuous surface patches. We observe that depth discontinuities are located at shadow boundaries, but not every shadow boundary implies a depth discontinuity (Figure 3). Noting that a perfect segmentation is virtually impossible to achieve, we relax our segmentation goal; instead of a perfect segmentation, we aim for a segmentation that captures *at least* all discontinuities. Even with a fair amount of over-segmentation, the dimensionality of the problem is reduced significantly. In fact, the *Photometric Expansion Move* that we layout in Section 5.4 simultaneously solves for the absolute depth of each segment and fuses adjacent segments into larger *meta-segments*, reducing the dimensionality even further in the course.

The segmentation is build from shadow masks; two adjacent pixels, defined by the indices p and q belong to the same segment, if their shadow masks $\mathbf{s}_p$ and $\mathbf{s}_q$ are identical. Using this relation in a floodfill-like algorithm, we naturally end up with a number of segments $\mathcal{S}_1, ..., \mathcal{S}_S$ covering the entire image. To simplify access of the segment's neighbourhood, segments are saved within a 2D graph. Some of the resulting segments cover only a few pixels. At the same time the run-time of our segment matching algorithm is proportional to the number of segments. To speed up computation time, we recommend to merge tiny segments to the most similar adjacent segment, measured by ω_{pq}, Eq. (4), until a minimum segment size is reached. In our implementation we define a minimum segment size in relation with the image resolution of 4:10^6.

5.2. Stereo Matching and Lightcode

Having defined the image segmentation, we continue with the definition of the matching cost between the left and the right image of a stereo pair. To simplify notation and description, we will start with pixel-wise matching energies and extend this later to a segment-based matching energy. Furthermore, we cover only the matching for the left image. For the right image, matching is performed analogously. The matching terms consists of three components: Consistency of shadow masks, pixel intensity and normal consistency as well as depth consistency.

Consistent illumination across all lighting conditions, which can be simplified to consistent shadow masks across all lighting conditions, is an efficient, but still rarely used metric to reduce matching ambiguities along the scanline in depth-from-stereo methods [SNA*21]. We express the matching through a shadow mask consistency weight

$$\gamma_p = \exp\left(-\frac{1}{2\sigma_S^2}\left\|\mathbf{s}_p - \mathbf{s}_p^{(R\to L)}\right\|^2\right), \quad \text{with } 0 \le \gamma_p \le 1 \tag{5}$$

where $\mathbf{s}_p^{(R\to L)}$ is the projection of the pixel-wise defined shadow mask from the right into the left camera image. In our experiments, $\sigma_S = 2$ proved to be give a good trade-off between rejection and acceptance based on the shadow masks. However, matching shadow masks alone is not a sufficient criterion, e.g., within a segment (Section 5.1) all pixels have identical shadow masks.

Consistent pixel intensities permit a finer differentiation then shadow masks and incorporate additional shading and texture information. In order to better handle uniform image regions, we also include normals into the pixel-wise matching term

$$\Phi_p := \omega_i \left\|\mathbf{i}_p - \mathbf{i}_p^{(R\to L)})\right\|^2 + \omega_n \left\|\mathbf{n}_p - \mathbf{n}_p^{(R\to L)}\right\|^2. \tag{6}$$

$\mathbf{i}_p^{(R\to L)}$ and $\mathbf{n}_p^{(R\to L)}$ are, in line with the previous notation, the image-intensities or normals projected from the right into the left camera image.

Our method refines depth iteratively, and segment-wise for the left and right image (Section. 5.4). In the end, both views should agree on the same depth. We express depth consistency based on the point-to-plane distance, which has been shown to be more accurate for highly slanted surfaces in the iterative closest point context [RL01, CSFM21].

$$\Xi_p = \mathbf{n}_p^{(L)} \cdot \left(\mathbf{x}_p^{(L)} - \mathbf{x}_p^{(R)}\right)^2, \tag{7}$$

where $\mathbf{x}_p$ are the 3D screen coordinates based on the current depth estimates and $\mathbf{x}_p^{(R\to L)}$ was projected from the right to the left image. As in previous work [BRR11, TMSN17], we introduce ceiling cutoffs for the pixel intensity, the normal and the depth matching

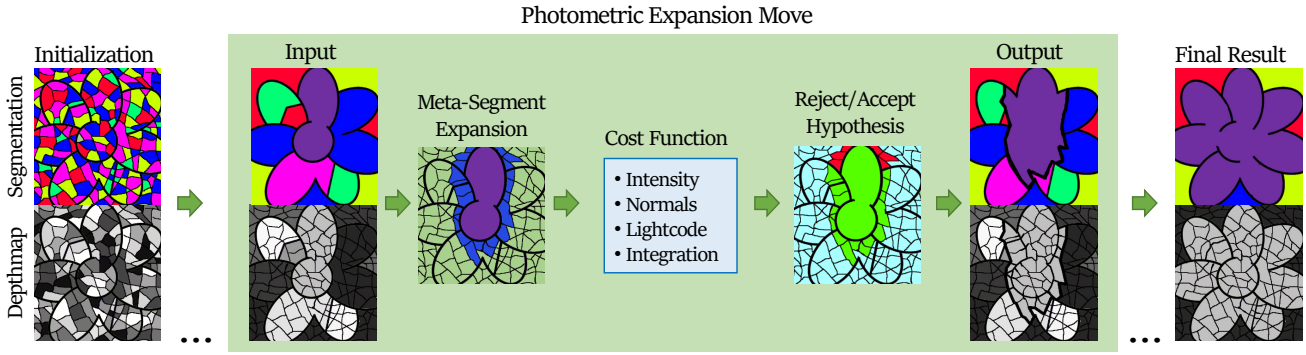


Figure 4: *Illustrative example showing the shadow based image segmentation and random depth assignment at initialization (left), one iteration of the expansion move (green box) and the final result after multiple iterations (right). At each iteration a meta-segment (purple) together with its one-ring neighbourhood is integrated and the new depth candidates are evaluated based on the cost function. For each segment, the new depth hypothesis is either accepted (green) or rejected (red).*

term. Due to these cutoffs, there is a maximal value of the matching terms $\Phi_{\max}$ and $\Xi_{\max}$ which we consider as a complete mismatch. The depth consistency is only taken into account if normals and pixel intensities match up. Therefore:

$$\Omega_p = \operatorname{lerp}\left(\Xi_p, \Xi_{\max}, \frac{\Phi_p}{\Phi_{\max}}\right).$$

$\operatorname{lerp}(x,y,\gamma)$ interpolates linearly between x and y for $\gamma \in [0,1]$, thus, depth consistency is only taken into account if $\Phi_p < \Phi_{\max}$.

When combining all, previously defined matching terms, we want to enforce matching shadow masks, and define the total matching function as:

$$E_{\text{Mat}}(\mathbf{d}) = \sum_{p \in \mathcal{S}} \operatorname{lerp}\left(\Phi_{\max} + \omega_d \Xi_{\max}, \Phi_p + \omega_d \Omega_p, \gamma_p\right). \quad (8)$$

In consequence, matches with bad shadow mask consistency are devalued. To speed-up computations, we introduce early termination criteria if $\Phi_p = \Phi_p^{(max)}$ and assign a mismatch to the total matching energy.

5.3. Normal Integration

Having ensured a semantically plausible image segmentation that reliably cuts along depth discontinuities, we estimate surfaces for each segment independently. For the normal integration we follow the idea of Durou et al. [DC07] and consider normals and tangents *between adjacent pixels*.

However, we replace the traditional gradient formulation by a more robust tangent term for highly slanted surfaces [CSFM21] The final energy function for the normal integration is:

$$E_{\text{Int}}(\mathbf{d}) = \sum_{(p,q)\in\mathcal{N}} \omega_{pq} \min\left((\mathbf{n}_{pq} \cdot \mathbf{t}_{pq}(d_p, d_q))^2, \tau_{\text{Int}}^2\right), \quad (9)$$

where the weight ω_{pq} between two adjacent pixels was defined in Eq. (4), $\mathbf{t}_{qp}$ is the surface tangent between two pixels and $\mathbf{n}_{pq}$ is the normalized mean of the normals at pixels p and q. If $(\mathbf{n}_{pq} \cdot \mathbf{t}_{pq})^2$ is above the cut-off τ_{Int}^2, it is interpreted as a depth discontinuity between those pixels.

A limitation, all normal integration methods have in common, is that the resulting surface is only unique up to a constant scaling factor, since scaling leaves the tangent *directions* untouched. This ambiguity cannot be resolved within the normal integration framework itself, and at least one single surface point is required to fix the entire surface. In Section 5.4 the ambiguity will be resolved by testing different depth candidates. Nevertheless, using the traditional definition of the surface tangent $\mathbf{t}_{pq}$ like [NRDR05, DGS11] would introduce a bias for small depth values, putting surfaces close to the camera. In this case, the tangents would be linear in depth d, i.e. scaling depth means scaling the tangents and hence the energy Eq. (9) is scaled as well.

Instead, we define the tangent vector based on logarithmic depth.

$$\mathbf{t}_{pq}(d_p, d_q) = (\mathbf{v}_p - \mathbf{v}_q) + (\mathbf{v}_p + \mathbf{v}_q)\big(\log(d_p) - \log(d_q)\big) \quad (10)$$

This alleviates the bias while remaining linear - in the new logarithmic depth. Thus, minimizing Eq. (9) is still equivalent to solving a linear system of equations. While depth is invariant regarding multiplication with a constant scale, the logarithmic depth is invariant regarding the addition of a constant value. This should come as no surprise as Eq. (10) contains only differences of logarithmic depths.

5.4. Photometric Expansion Move

Having defined the pixel-wise matching function in Section 5.2 and the segment-wise integration in Section 5.3, we combine both terms in the final energy function to estimate the (logarithmic) depth maps.

$$E(\mathbf{d}) = E_{\text{Mat}}(\mathbf{d}) + \omega_{\text{Int}} \cdot E_{\text{Int}}(\mathbf{d}) \quad (11)$$

To minimize the energy function, we apply expansion moves, which is an extended version of graph-cuts for non-binary problems. Expansion moves allow an iterative refinement of the (logarithmic) depth while simultaneously propagating depth to neighbouring segments and fusing segments into larger *meta-segments*. To access easier the local neighborhood of each segment, we save all segments $\mathcal{S}_1, \ldots, \mathcal{S}_S$ in a 2D graph structure [DJK11], with edges encoding adjacent segments. In addition, we create a second graph

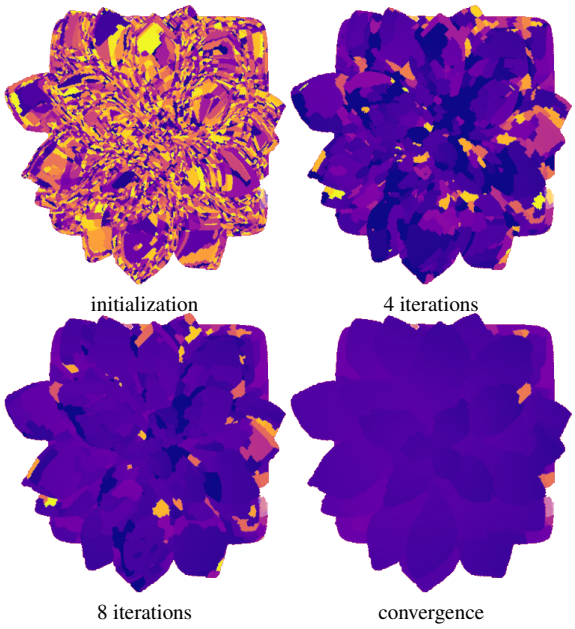


Figure 5: *From left to right: Depth maps before the expansion move, after four and eight iterations and after the expansion move. Segments grow into larger meta-segments and continuously converge to the final position.*

saving all meta-segments, $\mathcal{M}_1, ..., \mathcal{M}_M$ with $M \leq S$. At initialization time, every segment is its own meta-segment. During the optimization meta-segments are fused and the total number of meta-segments will decrease.

We initialize the photometric expansion move by integrating each meta-segment individually by Eq. (9) and scatter the segments in the capturing volume by assigning randomly uniform scaling along depth. Operating in logarithmic depth space Eq. (10), this results in adding a uniform log-depth offset for each segment. Afterwards we select meta-segment by meta-segment, and test whether each meta-segment and its its one-ring neighbourhood of segments form one continuous surface. Testing only adjacent segments, and not adjacent meta-segments, is motivated by providing a possibility to separate meta-segments again and recover from non-optimal decisions at the beginning of the optimization. We obtain the hypothetical surface as the solution of the minimization problem

$$\min_{d^{(\text{cand})}} \quad \sum_{(p,q)} \omega_{pq} \left(\mathbf{n}_{pq} \cdot \mathbf{t}_{pq}(d_p^{(\text{cand})}, d_q^{(\text{cand})}) \right)^2$$
$$\text{s.t.} \quad \sum_{p \in \mathcal{M}} \log\left(d_p^{(\text{cand})}\right) = \sum_{p \in \mathcal{M}} \log(d_p),$$

where (p,q) are all pixels of the meta-segment and the one-ring neighbourhood. Note that this is equivalent to Eq. (9) without the cut-off. By explicitly omitting the cut-off, we make the assumption that meta-segment and the one-ring neighbourhood form a single continuous surface. The additional constraint is needed to make the surface unique and ensures that current depth and candidate depth agree as much as possible within the meta-segment.

Algorithm 1 Photometric Expansion Move

```
S = {1, ..., S}                                   ▷ image segments
M = S                                             ▷ meta-segments
for left & right image do
    for i = 1, ..., M do                          ▷ all meta-segments
        select S_j ∈ N of M_i                     ▷ one-ring neighbourhood
        generate d_cand from Eq. (9) + Δ(log d)
        b* = argmin (Eq. (12))
        for each S_j do
            if b*_j == true then
                d ← d_cand in S_j                 ▷ adopt depth
                fuse(S_j, M_i)                    ▷ add to meta-segment
            end if
        end for
        update M
    end for                                       ▷ end meta-segments
end for                                           ▷ end image pairs
```

In addition to testing for continuous surfaces, we add a perturbation $\Delta(\log d)$ to the candidate depth, leading to a depth refinement over time. Remember, adding a uniform offset does not influence optimality with regards to the normal integration. In order to effectively sample the search space the perturbation is sampled from a normal distribution of width σ_{pert}. Large perturbations explore the search space while small perturbations refine matches that we already found. Therefore, we decrease σ_{pert} in the course of the algorithm. The total energy function Eq. (11) allows us to test whether the new depth is better or worse for the selected meta-segment and its one-ring neighbourhood. Based on the result of the boolean optimization problem, we decide for each segment whether to adopt or reject the depth hypothesis.

On the one hand, there is the matching error which is a sum over pixel-wise errors. The matching error for any segment $\mathcal{S}_i$ is simply the sum over all pixel-wise errors for $p \in \mathcal{S}_i$. Depending on whether we use the current depth or the candidate depth, there are up to two outcomes for the error which we denote $E_{\text{Mat}}^{(\mathcal{S}_i)}(0)$ for the current and $E_{\text{Mat}}^{(\mathcal{S}_i)}(1)$ for the candidate depth.

On the other hand, there is the integration error which is a sum over all adjacent pixels pairs p and q. Each pixel belongs to one unique segment. If $p \in \mathcal{S}_i$, then q must be in $\mathcal{S}_i$ as well, or in one $\mathcal{S}_j$ within the one-ring neighbourhood of $\mathcal{S}_i$. If p and q are in the same segment, this segment can either adopt or reject the candidate depth leading to two possible integration errors $E_{\text{Int}}^{(\mathcal{S}_i,\mathcal{S}_i)}(0,0)$ and $E_{\text{Int}}^{(\mathcal{S}_i,\mathcal{S}_i)}(1,1)$. If p and q belong to two adjacent segments $\mathcal{S}_i$ and $\mathcal{S}_j$ there are two more options: One segment can adopt the candidate depth while the other rejects it, leading to the integration errors $E_{\text{Int}}^{(\mathcal{S}_i,\mathcal{S}_j)}(0,1)$ and $E_{\text{Int}}^{(\mathcal{S}_i,\mathcal{S}_j)}(1,0)$ at the boundary of $\mathcal{S}_i$ and $\mathcal{S}_j$.

By summing over all segments and all pairs of adjacent segments we obtain a boolean optimization problem from Eq. 11:

$$E(\mathbf{b}) = \sum_i E_{\text{Mat}}^{(\mathcal{S}_i)}(b_i) + \omega_{\text{Int}} \sum_{(i,j)} E_{\text{Int}}^{(\mathcal{S}_i,\mathcal{S}_j)}(b_i, b_j) \qquad (12)$$

where $b_i = 1$ if segment $\mathcal{S}_i$ adopts the candidate depth and $b_i = 0$

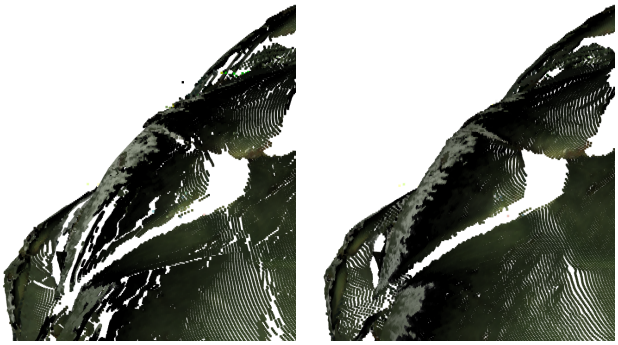


Figure 6: *Visual comparison of the reconstruction before (left) and after post-processing (right). The post-processing removes wrong matches from occluded areas and closes small gaps around (meta-) segment boundaries.*

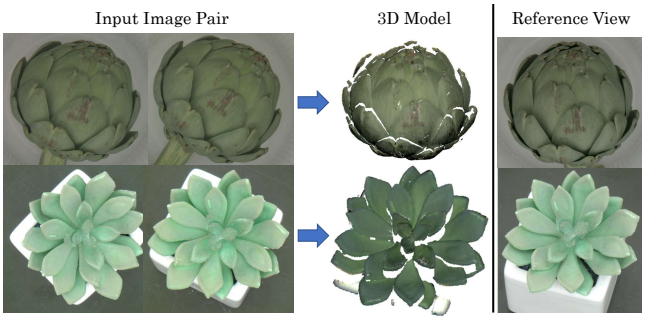


Figure 7: *Qualitative comparison for real data. From left to right: original image pair taken by a stereo camera, 3D reconstruction matching the reference view, image taken by a third (reference) camera.*

if the candidate depth is rejected. Any combination for the meta-segment and its one-ring neighbourhood represented by $\mathbf{b}$ leads to a different new depth d_{new}. The condition $E(d_{\text{new}}) = E(\mathbf{b})$ is valid for each value of $\mathbf{b}$. Hence,

$$\mathbf{b}^* = \underset{\mathbf{b}}{\operatorname{argmin}} E(\mathbf{b}) \tag{13}$$

describes the optimal adoption/rejection strategy. All segments adopting the depth hypothesis are fused into one meta-segment. If some adopting segments are disconnected, segments are fused into different meta-segments.

Different to the graph-cut based shadow detection, we cannot guarantee that all binary terms $E_{\text{Int}}^{(\mathcal{S}_i,\mathcal{S}_j)}$ are strictly submodular. They mostly are as long as $\Delta(\log d)$ is large compared to the noise in the normal data. Still, graph cuts can be used to obtain a (partial) solution that is guaranteed to not increase the energy [KR07].

It might appear that the photometric expansion moves requires an escalating amount of normal integrations on different iterations. In theory, this is true. However, the integration can be speed up significantly. We are using the conjugate gradient algorithm [GJ*10] to solve the linear system involved in the normal integration. As an iterative solver, conjugate gradient benefits from a good initial guess. Due to the structure of the normal integration problem, such a guess can be easily obtained from previous iterations, because the internal shape of a segment is typically quite close to the optimum already. The matter is more about aligning the segments to each other. For the selected meta-segment, this was full-filled in the previous iteration. For segments in the one-ring neighbourhood, we can use the additive degree of freedom to minimize the integration error along the boundary to the meta-segment; Segments are moved as a whole to line up the boundaries.

The total energy also includes the photometry term Eq. (3). However, the effect of depth on lighting and thus normals is neglible and can be omitted for the depth update. Still, we update shadow masks as well as albedo/normals every four iterations of the expansion move with E_{Int} acting as a regularization term encouraging orthogonality to the surface.

5.5. Post-Processing

The expansion moves converge closely to the global minimum. However, this is only valid under the assumption that the integration of meta-segments is free of errors. Due to numerical and discretization errors, noise within data or deviations from the Lambertian model small discontinuities might be introduced between meta-segments. To improve the visual reconstruction quality, we further minimize our energy function Eq. (11) using the Gauss-Newton algorithm, by linearizing the errors, solving for a descent direction and perform a line search. The energy function is minimized for each meta-segment while keeping the surrounding meta-segments fixed, cf. Eq. (9). Since meta-segments are a suitable prior for connected regions, this gives a good balance between performance and quality as opposed to a global optimization. We run five iterations of Gauss-Newton for each meta-segment one after the other until convergence. It is worth noting that Gauss-Newton converges to the next best minimum and is thus only meaningful in conjunction with previous expansion moves to bring us close to the desired optimum. After the Gauss-Newton iteration, we check for depth consistency between both views and reject inconsistent pixels, which is a well-established method to improve reconstruction results and to handle occlusions, cf. e.g. [BRR11]. A close-up of the artichoke reconstruction before and after post-processing can be found in Figure 6. After the expansion move, all segments are approximately in the right position but there are still visible gaps between the (meta-) segments and some amount of noise. These gaps close during post-processing leading to a smoother surfaces and boundaries.

6. Evaluation

The central element of our algorithm is the iterative fusion of segments into larger and larger meta-segments. The number of pixels, against the number of segments at initialization time and the final number of fused meta-segments can be found in Table 1. Obviously, the number of segments is at least one magnitude smaller than the number of foreground pixels. The advantage seems to be even larger for the two higher resolution datasets. However, the

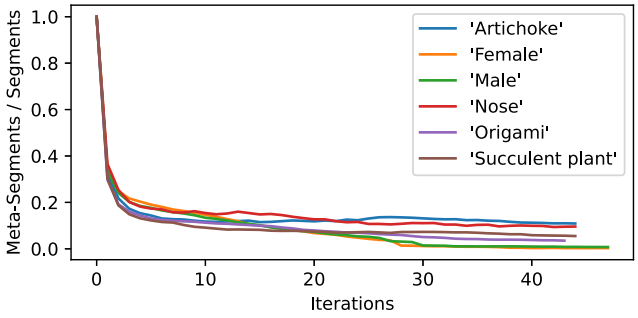


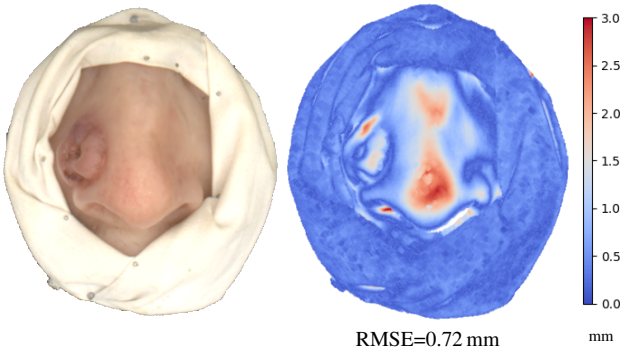
Figure 8: *The ratio between meta-segments and the original number of segments in the course of the algorithm. The number of meta-segments decreases quickly after a few iterations.*

strong reduction in complexity becomes apparent if we look at the number of remaining meta-segments which are again, several magnitudes less compared to the original number of segments. The number of meta-segments during the optimization is illustrated in more detail in Figure 8. As is easy to see, already a handful of iterations are sufficient to fuse a substantial amount of segments into meta-segments. This is consistent with the evolvement of the depth maps in Figure 5.

Real World Examples We perform qualitative as well as quantitative tests on some challenging examples. Here, the images where acquired in a lightstage similar setup [SSWK13] using 41 lights, that roughly uniformly sample a hemisphere. The first object is a so-called moulage, a medical wax model of a human nose depicting a skin disease. The model together with the error map is depicted in Figure 9. The groundtruth data stems from a structured-light scanner. The RMSE is well below 1 mm. Despite micro-structure and the wrinkles of the cloth, it is reconstructed at a high precision. The same is true for the area right around the nostrils. Both the tip and the bridge of the nose illustrate that our pipeline can handle non-Lambertian materials albeit with some effect on the reconstruction quality. Additionally, we have reconstructions for an artichoke and a succulent plant. Lacking groundtruth data, we rely on the visual similarity and compare the reconstructed 3D model against a reference view obtained with an additional camera (Figure 7).

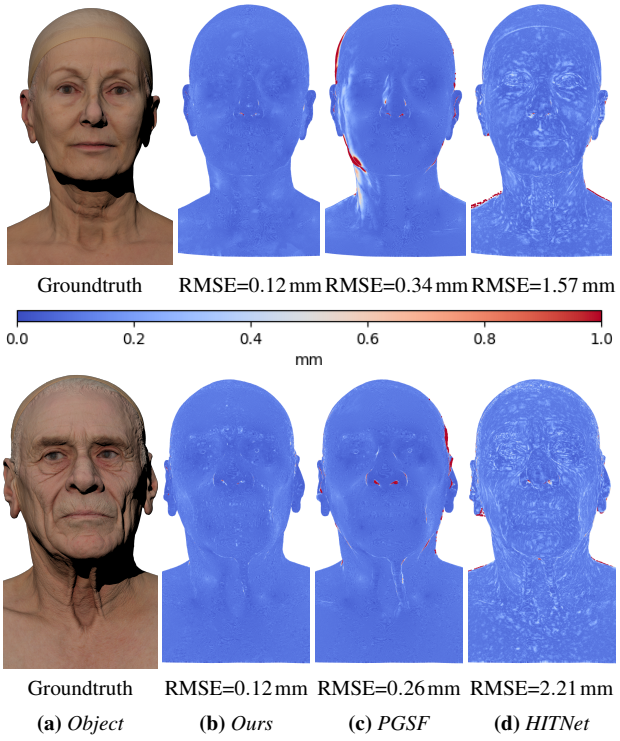
Previous work: For quantitative analysis and comparisons to previous work we rely on three synthetic datasets: A male and a female 3D face scan with pore-deep geometry details, purchased from 3D Scanstore and a highly self-occluding Origami inspired scene modelled in 3D. All scenes were rendered as HDR-images with the Arnold render engine, using 73 lights uniformly spread across a hemisphere at about 3 m distance to the object and observed through a stereo camera at 2 m distance and a baseline of 15 cm. For the faces, we use a resolution of 2048 by 2048 pixels and 1024 by 1024 pixels for the Origami scene.

We compare our method to the photogeometric scene flow (PGSF) algorithm [GSSM15] that was developed for face scanning and combines photometric stereo with depth-from-stereo and optical flow. For a fair comparison, we adapt the method to using all lights and omit the scene flow part as we are dealing with



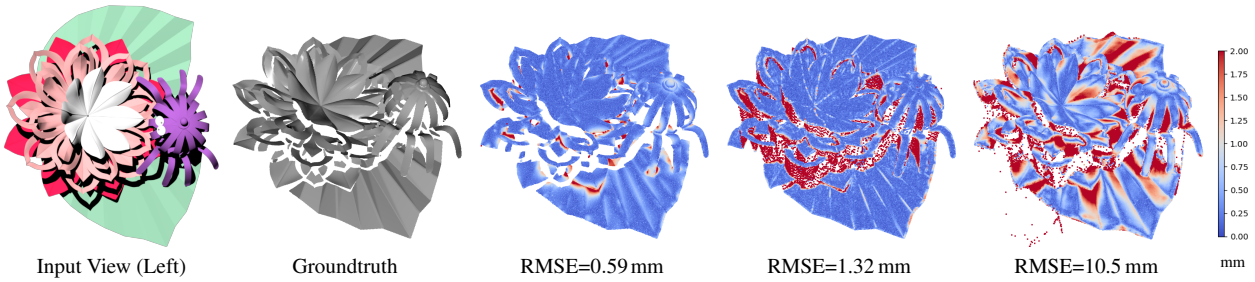
(a) *Input View* **(b)** *Reconstruction Error*

Figure 9: *Quantitative comparison to the 'Nose' dataset. The object is a medical wax model of a human nose depicting a skin disease. groundtruth data was acquired through a structured light scanner. Both images where rotated for an upright depiction.*



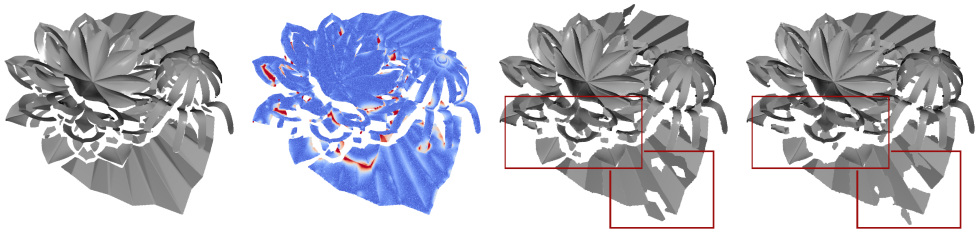
(a) *Object* **(b)** *Ours* **(c)** *PGSF* **(d)** *HITNet*

Figure 10: *Quantitative comparison of the 'Male' and 'Female' dataset. From left to right: A rendering of the object, error maps and root-mean-squared error (RMSE) for our method, the method in [GSSM15] and HITNet [THZ*21]*



Input View (Left) **(a)** *Object* | Groundtruth **(b)** *Object* | RMSE=0.59 mm **(c)** *Ours* | RMSE=1.32 mm **(d)** *PGSF* | RMSE=10.5 mm **(e)** *HITNet*

Figure 11: *Quantitative comparison of the 'Origami' dataset. From left to right: A rendering of the object, ground truth result from a tilted view, error maps and root-mean-squared error (RMSE) for our method, PGSF [GSSM15] and HITNet [THZ*21]. Our results, show substantially less smearing artifacts at depth discontinuities.*



(a) *Groundtruth* **(b)** *All* **(c)** *Normals and Intensity* **(d)** *Intensity only*

Figure 12: *Ablation Studies: (a) Ground truth result from a tilted view, (b) reconstruction result with the full matching term, (c) reconstruction result with normal and intensity matching only (without lightcodes) and (d) finally matching based on pixel intensities alone. Especially for occluded areas - like the lower flower petals - and uniform areas - like the lowest leaf layer - the reconstruction quality noticeably decreases. Some segments are misplaced, others are removed during post-processing due inconsistencies across views.*

static scenes. Furthermore, we compare against HITNet [THZ*21], a state of the art depth-from-stereo algorithm, leading at the time of comparison the synthetic SceneFlow [MIH*16] benchmark. To make the comparison as fair as possible, the image pairs were rendered without shadows with a headlight at the camera position. Materials in all scenes are fully Lambertian. The results for the faces are shown in Figure 10 and for the Origami scene in Figure 11. The Origami dataset, together with the reconstructed 3D point clouds is part of the supplemental material. In both comparisons, our method has the smallest root-mean-squared error (RMSE). While we outperform PGSF only in areas with depth discontinuities, The RMS-error of our method is an order of magnitude smaller compared to HITNet.

Ablation Studies: Finally, we evaluate the effects of the different parts of our algorithm. A key element of our algorithm is the introduction of lightcodes, i.e. shadow masks and normals our into matching term. For comparison, we performed our reconstruction without lightcodes as well as without lightcode and normals, i.e. only based on image intensities. The results are pictured in Figure 12. Clearly, omitting parts of the matching term results in misplaced or missing segments. Missing segments occur if both views do not agree on a consistent depth. The effects are especially apparent for uniform regions like the lowest leaf but also for occluded regions like the lower flower petals.

7. Conclusion and Future Work

We proposed a photometric 3D reconstruction algorithm that fully integrates photometric stereo and depth-from-stereo into a single reconstruction pipeline. We demonstrated that this integrated pipeline reconstructs both synthetic rendering as well as real footage and achieved state of the art performance. In future work, we would like to investigate more complex shading models, especially in order to handle specular reflections. Our experiments proved that the core component of our algorithm, the shadow-based segmentation, reduces the matching complexity between stereo pairs. At the same time, cuts are introduced reliably along depth discontinuities. This, together with the light-code based matching, resulted in very clean edges without any smearing artefacts. Furthermore, our method proved to be capable of rapidly fusing segments together into even larger meta-segments, thus reducing the complexity even further. Due to the high reconstruction quality, we believe that the iterative reduction of meta-segments demonstrates the full potential that lies in deeply integrating photometric stereo with depth-from-stereo methods. We would like to investigate how our approach scales with the number of views in a multi-view stereo setup.

Acknowledgments

Special thanks to: Ralf Sarlette for technical advice and support on the hardware setup and ground truth comparison with real-data, Oh-Hun Kwon for computing the HITNet results. This work has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy, EXC-2070 – 390732324 (PhenoRob). Open Access funding enabled and organized by Projekt DEAL.

References

[AG15] ACKERMANN J., GOESELE M.: A Survey of Photometric Stereo Techniques. *Foundations and Trends in Computer Graphics and Vision 9*, 3-4 (2015), 149–254. doi:10.1561/0600000065. 2, 4

[BFR14] BERKITEN S., FAN X., RUSINKIEWICZ S.: Merge2-3D: Combining Multiple Normal Maps with 3D Surfaces. In *International Conference on 3D Vision* (Dec. 2014), vol. 1, pp. 440–447. doi:10.1109/3DV.2014.22. 2

[BJ01] BOYKOV Y. Y., JOLLY M.-P.: Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D Images. In *Proceedings of the IEEE International Conference on Computer Vision. ICCV* (2001), vol. 1, IEEE, pp. 105–112. 4

[BK04] BOYKOV Y., KOLMOGOROV V.: An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*, 9 (2004), 1124–1137. 4

[BRR11] BLEYER M., RHEMANN C., ROTHER C.: PatchMatch Stereo - Stereo Matching with Slanted Support Windows. In *Procedings of the British Machine Vision Conference BMVC* (Dundee, 2011), British Machine Vision Association, pp. 14.1–14.11. doi:10.5244/C.25.14. 3, 5, 8

[BSFG09] BARNES C., SHECHTMAN E., FINKELSTEIN A., GOLDMAN D.: PatchMatch: A Randomized Correspondence Algorithm for Structural Image Editing. *ACM Transaction on Graphics 28* (Aug. 2009). doi:10.1145/1531326.1531330. 3

[CAK07] CHANDRAKER M., AGARWAL S., KRIEGMAN D.: ShadowCuts: Photometric Stereo with Shadows. In *Procedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Minneapolis, MN, USA, June 2007), IEEE, pp. 1–8. doi:10.1109/CVPR.2007.383288. 3, 4

[CSFM21] CAO X., SHI B., FUMIO O., MATSUSHITA Y.: Normal Integration via Inverse Plane Fitting With Minimum Point-to-Plane Distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2021), pp. 2382–2391. 2, 5, 6

[DC07] DUROU J.-D., COURTEILLE F.: Integration of a Normal Field without Boundary Condition. In *Proceedings of the Workshop on Photometric Analysis For Computer Vision-PACV 2007* (2007), INRIA, pp. 8–p. 6

[DGS11] DU H., GOLDMAN D. B., SEITZ S. M.: Binocular Photometric Stereo. In *Procedings of the British Machine Vision Conference BMVC* (2011), vol. 4, British Machine Vision Association, p. 8. 2, 6

[DJK11] DEZS B., JÜTTNER A., KOVÁCS P.: LEMON - an Open Source C++ Graph Template Library. *Electronic Notes in Theoretical Computer Science (ENTCS) 264*, 5 (July 2011), 23–45. doi:10.1016/j.entcs.2011.06.003. 6

[GJ*10] GUENNEBAUD G., JACOB B., ET AL.: Eigen v3, 2010. 8

[GLD*19] GUO K., LINCOLN P., DAVIDSON P., BUSCH J., YU X., WHALEN M., HARVEY G., ORTS-ESCOLANO S., PANDEY R., DOURGARIAN J.: The Relightables: Volumetric Performance Capture of Humans with Realistic Relighting. *ACM Transactions on Graphics (ToG) 38*, 6 (2019), 1–19. 1, 2

[GLS15] GALLIANI S., LASINGER K., SCHINDLER K.: Massively Parallel Multiview Stereopsis by Surface Normal Diffusion. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (Santiago, Chile, Dec. 2015), IEEE, pp. 873–881. doi:10.1109/ICCV.2015.106. 3

[GLU12] GEIGER A., LENZ P., URTASUN R.: Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2012), IEEE, pp. 3354–3361. 2

[GSSM15] GOTARDO P. F. U., SIMON T., SHEIKH Y., MATTHEWS I.: Photogeometric Scene Flow for High-Detail Dynamic 3D Reconstruction. In *2015 IEEE International Conference on Computer Vision (ICCV)* (Santiago, Chile, Dec. 2015), IEEE, pp. 846–854. doi:10.1109/ICCV.2015.103. 2, 4, 9, 10

[GT15] GROCHULLA M., THORMÄHLEN T.: Combining Photometric Normals and Multi-View Stereo for 3D Reconstruction. In *Proceedings of the European Conference on Visual Media Production* (London United Kingdom, Nov. 2015), ACM, pp. 1–8. doi:10.1145/2824840.2824846. 2

[GWT*18] GAN J., WILBERT A., THORMÄHLEN T., DRESCHER P., HAGENS R.: Multi-view photometric stereo using surface deformation. *The Visual Computer 34*, 11 (Nov. 2018), 1551–1561. doi:10.1007/s00371-017-1430-5. 2

[HC04] HONG L., CHEN G.: Segment-based Stereo Matching Using Graph Cuts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2004), vol. 1, IEEE, pp. I–I. 2

[HVC08] HERNANDEZ C., VOGIATZIS G., CIPOLLA R.: Multiview photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence 30*, 3 (2008), 548–554. 2

[Ike18] IKEHATA S.: CNN-PS: CNN-based Photometric Stereo for General Non-Convex Surfaces. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 3–18. 2

[JK07] JOSHI N., KRIEGMAN D. J.: Shape from Varying Illumination and Viewpoint. In *IEEE International Conference on Computer Vision ICCV* (2007), IEEE, pp. 1–7. 2

[KKO*21] KAYA B., KUMAR S., OLIVEIRA C., FERRARI V., VAN GOOL L.: Uncalibrated Neural Inverse Rendering for Photometric Stereo of General Surfaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 3804–3814. 2

[KR07] KOLMOGOROV V., ROTHER C.: Minimizing Nonsubmodular Functions with Graph Cuts-A Review. *IEEE Transactions on Pattern Analysis and Machine Intelligence 29*, 7 (2007), 1274–1279. 8

[KSK06] KLAUS A., SORMANN M., KARNER K.: Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure. In *International Conference on Pattern Recognition (ICPR)* (2006), vol. 3, IEEE, pp. 15–18. 2

[KZ04] KOLMOGOROV V., ZABIN R.: What energy functions can be minimized via graph cuts? *IEEE Transactions on Pattern Analysis and Machine Intelligence 26*, 2 (Feb. 2004), 147–159. doi:10.1109/TPAMI.2004.1262177. 4

[LBMC20] LOGOTHETIS F., BUDVYTIS I., MECCA R., CIPOLLA R.: A CNN Based Approach for the Near-Field Photometric Stereo Problem. In *British Machine Vision Virtual Conference (BMVC)* (Oct. 2020). 2

[LJBB20] LAGA H., JOSPIN L. V., BOUSSAID F., BENNAMOUN M.: A Survey on Deep Learning Techniques for Stereo-based Depth Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020). 2

[LMC19] LOGOTHETIS F., MECCA R., CIPOLLA R.: A Differential Volumetric Approach to Multi-View Photometric Stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 1052–1061. 1, 2

[LZYZ16] LI L., ZHANG S., YU X., ZHANG L.: PMSC: PatchMatch-Based Superpixel Cut for Accurate Stereo Matching. *IEEE Transactions on Circuits and Systems for Video Technology 28*, 3 (2016), 679–692. 3

[MIH*16] MAYER N., ILG E., HAUSSER P., FISCHER P., CREMERS D., DOSOVITSKIY A., BROX T.: A Large Dataset To Train Convolutional Networks for Disparity, Optical flow, and Scene Flow Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 4040–4048. 2, 10

[NRDR05] NEHAB D., RUSINKIEWICZ S., DAVIS J., RAMAMOORTHI R.: Efficiently Combining Positions and Normals for Precise 3D Geometry. *ACM transactions on graphics (TOG) 24*, 3 (2005), 536–543. 2, 6

[PSM*16] PARK J., SINHA S. N., MATSUSHITA Y., TAI Y.-W., KWEON I. S.: Robust Multiview Photometric Stereo Using Planar Mesh Parameterization. *IEEE transactions on pattern analysis and machine intelligence 39*, 8 (2016), 1591–1604. 2

[QDA18] QUÉAU Y., DUROU J.-D., AUJOL J.-F.: Normal Integration: A Survey. *Journal of Mathematical Imaging and Vision 60*, 4 (2018), 576–593. 2

[RKB04] ROTHER C., KOLMOGOROV V., BLAKE A.: "GrabCut" - Interactive Foreground Extraction using Iterated Graph Cuts. *ACM transactions on graphics (TOG) 23*, 3 (2004), 309–314. 4

[RL01] RUSINKIEWICZ S., LEVOY M.: Efficient Variants of the ICP Algorithm. In *Proceedings on 3-D Digital Imaging and Modeling* (2001), IEEE, pp. 145–152. 5

[RM19] ROMANONI A., MATTEUCCI M.: TAPA-MVS: Textureless-Aware PAtchMatch Multi-View Stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (Seoul, Korea (South), Oct. 2019), IEEE, pp. 10412–10421. doi:10.1109/ICCV.2019.01051. 3

[RTF*04] RASKAR R., TAN K.-H., FERIS R., YU J., TURK M.: Non-photorealistic camera: Depth edge detection and stylized rendering using multi-flash imaging. *ACM transactions on graphics (TOG) 23*, 3 (2004), 679–688. 3

[SNA*21] SUN T., NAM G., ALIAGA C., HERY C., RAMAMOORTHI R.: *Human Hair Inverse Rendering Using Multi-View Photometric Data*. The Eurographics Association, 2021. doi:10.2312/sr.20211301. 2, 5

[SSB*14] SHEN F., SUNKAVALLI K., BONNEEL N., RUSINKIEWICZ S., PFISTER H., TONG X.: Time-Lapse Photometric Stereo and Applications. *Computer Graphics Forum 33*, 7 (2014), 359–367. doi:10.1111/cgf.12504. 2

[SSS*20] SANTO H., SAMEJIMA M., SUGANO Y., SHI B., MATSUSHITA Y.: Deep Photometric Stereo Networks for Determining Surface Normal and Reflectances. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2020), 1–1. doi:10.1109/TPAMI.2020.3005219. 2

[SSWK13] SCHWARTZ C., SARLETTE R., WEINMANN M., KLEIN R.: DOME II: A Parallelized BTF Acquisition System. In *Material Appearance Modeling* (2013), Citeseer, pp. 25–31. 9

[SYZ*21] SONG X., YANG G., ZHU X., ZHOU H., WANG Z., SHI J.: AdaStereo: A Simple and Efficient Approach for Adaptive Stereo Matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 10328–10337. 2

[THZ*21] TANKOVICH V., HANE C., ZHANG Y., KOWDLE A., FANELLO S., BOUAZIZ S.: HITnet: Hierarchical Iterative Tile Refinement Network Ror real-time Stereo Matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), pp. 14362–14372. 2, 9, 10

[TMSN17] TANIAI T., MATSUSHITA Y., SATO Y., NAEMURA T.: Continuous 3D Label Stereo Matching using Local Expansion Moves. *IEEE Transactions on Pattern Analysis and Machine Intelligence 40*, 11 (2017), 2725–2739. 3, 4, 5

[VPB*09] VLASIC D., PEERS P., BARAN I., DEBEVEC P., POPOVIĆ J., RUSINKIEWICZ S., MATUSIK W.: Dynamic Shape Capture using Multi-View Photometric Stereo. In *ACM Transaction on Graphics (ToG)*. 2009, pp. 1–11. 2

[WGTS13] WANG Z., GROCHULLA M., THORMÄHLEN T., SEIDEL H.-P.: 3D Face Template Registration Using Normal Maps. In *International Conference on 3D Vision - 3DV* (June 2013), pp. 295–302. doi:10.1109/3DV.2013.46.

[WL11] WANG D., LIM K. B.: Obtaining depth map from segment-based stereo matching using graph cuts. *Journal of Visual Communication and Image Representation 22*, 4 (2011), 325–331. 2

[XWW*19] XIE W., WANG M., WEI M., JIANG J., QIN J.: Surface Reconstruction from Normals: A Robust DGP-based Discontinuity Preservation Approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), pp. 5328–5336. 2

[YLF*20] YAO Z., LI K., FU Y., HU H., SHI B.: GPS-Net: Graph-based Photometric Stereo Network. *Advances in Neural Information Processing Systems 33* (2020), 10306–10316. 2

[ZDJ*20] ZHOU M., DING Y., JI Y., YOUNG S. S., YU J., YE J.: Shape and Reflectance Reconstruction using Concentric Multi-Spectral Light Field. *IEEE Transactions on Pattern Analysis and Machine Intelligence 42*, 7 (2020), 1594–1605. 2