EUROGRAPHICS 2018
K. Hildebrandt and C. Theobalt
(Guest Editors)

Volume 37 (2018), Number 2
STAR – State of The Art Report

State of the Art on 3D Reconstruction with RGB-D Cameras

Michael Zollhöfer[1,2], Patrick Stotko[3], Andreas Görlitz[4], Christian Theobalt[1], Matthias Nießner[5], Reinhard Klein[3], Andreas Kolb[4]

[1] Max-Planck Institute for Informatics, Saarbrücken, Germany
[2] Stanford University, CA, USA
[3] Computer Graphics Group, University of Bonn, Germany
[4] Computer Graphics Group, University of Siegen, Germany
[5] Technical University of Munich, Germany

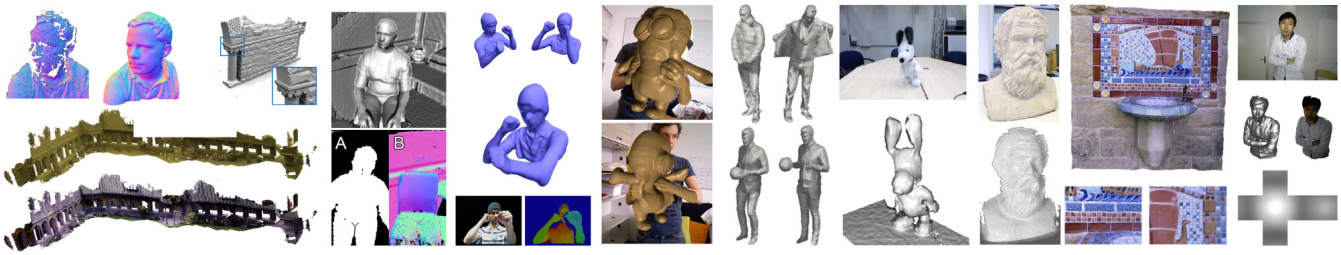


Figure 1: *This state-of-the-art report provides an overview of RGB-D scene reconstruction approaches. We discuss recent trends in the geometric reconstruction of static (left) and dynamic scenes (middle) as well as the acquisition of corresponding color and reflectance (right).*

Abstract

The advent of affordable consumer grade RGB-D cameras has brought about a profound advancement of visual scene reconstruction methods. Both computer graphics and computer vision researchers spend significant effort to develop entirely new algorithms to capture comprehensive shape models of static and dynamic scenes with RGB-D cameras. This led to significant advances of the state of the art along several dimensions. Some methods achieve very high reconstruction detail, despite limited sensor resolution. Others even achieve real-time performance, yet possibly at lower quality. New concepts were developed to capture scenes at larger spatial and temporal extent. Other recent algorithms flank shape reconstruction with concurrent material and lighting estimation, even in general scenes and unconstrained conditions. In this state-of-the-art report, we analyze these recent developments in RGB-D scene reconstruction in detail and review essential related work. We explain, compare, and critically analyze the common underlying algorithmic concepts that enabled these recent advancements. Furthermore, we show how algorithms are designed to best exploit the benefits of RGB-D data while suppressing their often non-trivial data distortions. In addition, this report identifies and discusses important open research questions and suggests relevant directions for future work.

CCS Concepts

•***Computing methodologies*** *,...,* ***Reconstruction; Appearance and texture representations;*** *Motion capture;*

1. Introduction

The core technology behind today's structured light or time-of-flight-based depth cameras already dates back several decades. However, the recent introduction of consumer grade sensors that package this technology into mass-manufactured devices of small form factor made RGB-D cameras a commodity available to a larger user base. Microsoft started this development in 2010 with the Kinect, and several other devices followed, e.g., RGB-D cameras such as the Intel RealSense, Primesense Carmine, Google Tango, or Occiptial's Structure Sensor. These cameras are not only available at low price points, but these lightweight sensors also capture per-pixel color and depth images at adequate resolution and at real-time rates. In conjunction, these features put them ahead of even some far more expensive 3D scanning systems, in particular when developing solutions for consumer grade applications. The potential of these new sensors has been quickly recognized in visual computing. For instance, the seminal KinectFusion work [NDI*11, IKH*11] had remarkable impact in the computer graphics and vision communities, and triggered an incredible response. Since then, the state of the art has been greatly advanced in computer graphics and computer

vision research developing new methods to reconstruct models of the static and dynamic world around us.

To enable this progress, many foundational research problems in this area were revisited and rethought to make best use of the new capabilities and to compensate the weaknesses of RGB-D cameras. First, highly innovative new algorithms for RGB-D-based dense 3D geometry reconstruction of static environments were developed. They have pushed the state of the art along several dimensions: new concepts to enable real-time scanning and scan integration were introduced, new ideas for drift reduction and live scanning of larger scenes were presented, and new ideas to obtain high geometry quality despite notable sensor noise were researched. Second, entirely new methods for capturing dense 3D geometry models of dynamic scenes and scene elements were proposed, such as models of moving humans and rigid objects, or of general deformable surfaces. Also in this area, the state of the art has been advanced in several ways. For example, new template-based methods have reached previously unseen runtime performance and accuracy levels, even when capturing with a single RGB-D camera. Others introduced new concepts to achieve very high reconstruction detail, yet at higher computational cost. Innovative concepts to capture space-time coherent geometry and learn shape templates on the fly have opened up further new possibilities. Third, entirely new methods were developed that capture additional scene properties from RGB-D data alongside with geometry. In particular, new methods were proposed that capture material and reflectance models of static and dynamic scenes, often times in parallel with illumination estimates, mainly with a focus on uncalibrated environments. Simultaneous shape and appearance capture is much harder, but does not only lead to more expressive scene models, but it also adds robustness to reconstruction under difficult scene conditions.

In this report, we will review and compare in detail the state-of-the-art methods from all three of these areas. We will explain the common new algorithmic concepts underlying recent innovations. In particular, we will explain and compare newly proposed concepts for RGB-D geometry processing and shape representation. We will review fundamental data structures and concepts enabling the scanning of shape, material, and illumination even on large spatial and temporal extent. Our focus will be on methods achieving interactive or real-time frame rates. However, we will also explore the fundamental basics enabling this runtime performance, and show how they arose from ideas originally developed for offline reconstruction. The report will also critically analyze recent progress and discuss open questions and avenues for future research.

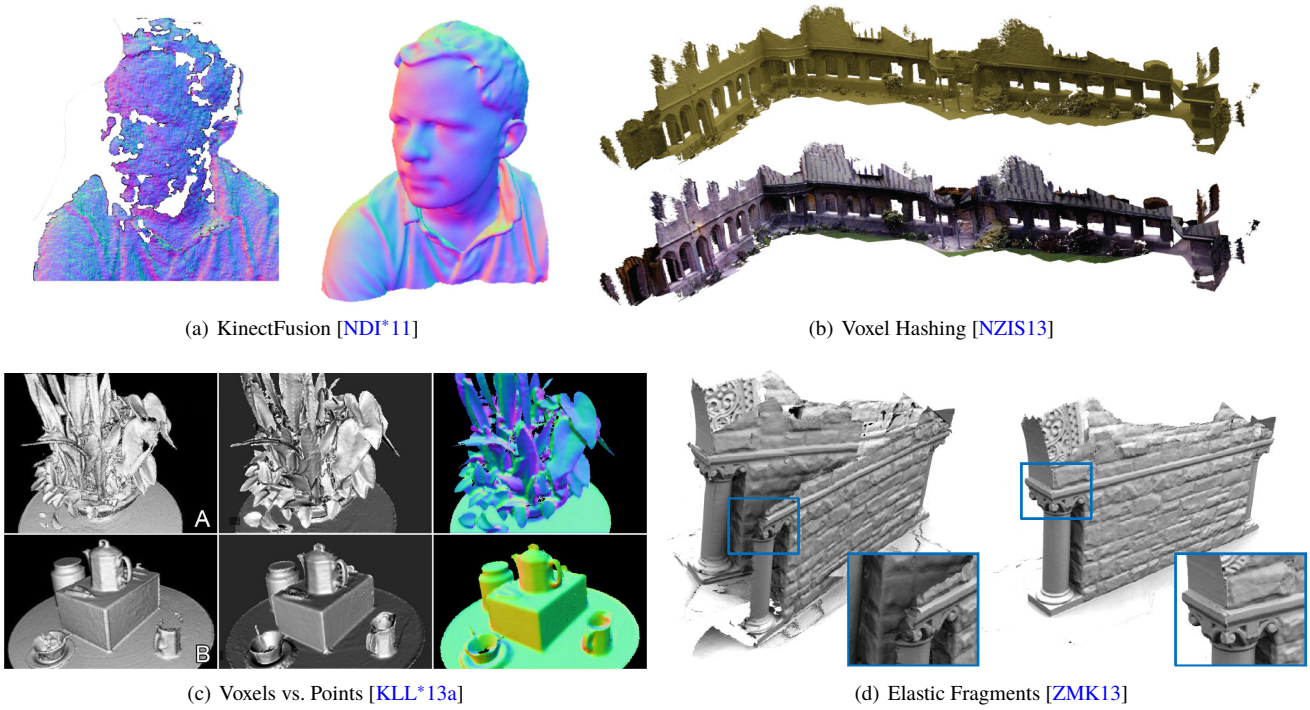
1.1. RGB-D Cameras and their Characteristics

Traditionally, there are two main approaches in range sensing, i.e., triangulation and Time-of-Flight (ToF). Triangulation can be realized as passive approach, i.e., stereo vision, or as active systems, such as structured light. While stereo vision computes the disparity between two images taken at different positions, structured light cameras project an infrared light pattern onto the scene and estimate the disparity given by the perspective distortion of the pattern due to the varying object's depth. Light detection and ranging (LIDAR) scanners and ToF cameras, on the other hand, measure the time that light emitted by an illumination unit requires to travel to an object and back to a detector. While LIDAR comprise mechanical components in order to realize the scanning approach, ToF cameras perform the time-of-flight computation on integrated circuits using standard CMOS or CCD technologies.

With early range cameras being (somewhat) accessible in the early 2000's, RGB-D camera prototypes have been set up in various research labs [LKH07, HJS08]. Up to now, mass-market RGB-D cameras rely on structured light or ToF camera approaches [KP15]. These RGB-D cameras often suffer from very specific noise characteristics and sometimes very challenging data distortions, which, in most cases, have to be taken into account in algorithm development. Functionally, there are several differences between structured light based RGB-D cameras, such as the first Kinect, and cameras on the basis of ToF, e.g., the Kinect version V2. They are related to the camera's resilience against background light, e.g. for outdoor applications, the quality of depth data, and the robustness in dealing with semi-transparent media and other, so-called multi-path effect, resulting from indirect paths taken by the active light [SLK15]. Another main difference between structured light and ToF camera approaches is that structured light requires a baseline between the illumination unit and the area sensor, which is not required for ToF.

1.2. Related STARs and Surveys

This state-of-the-art report addresses recent developments in RGB-D scene reconstruction in terms of algorithmic concepts and with respect to different application scenarios, e.g., the reconstruction of static scenes (Sec. 2), dynamic scenes (Sec. 3), and color and appearance capture (Sec. 4). There are a few surveys that are related to this STAR, they, however, focus on modeling techniques for static scenes and related public datasets [CLH15, HSR15]. Berger et al. [BTS*14] presented the "State of the Art in Surface Reconstruction from Point Clouds" in Eurographics 2014. They focus on 3D surface reconstruction from point cloud data and characterize methods with respect to imperfections of the input point cloud, the type of input data (geometry, color, and normal information), the classes of supported shapes, and the output surface format. Bonneel et al. [BKPB17a] presented a survey on "Intrinsic Decompositions for Image Editing" in Eurographics 2017. They focus on methods that decompose a color image into its reflectance and illumination layers. They classify the approaches based on the used priors that are imposed on the intrinsic decomposition problem. Weinmann et al. [WLGK16] presented a tutorial on "Advances in Geometry and Reflectance Acquisition" in Eurographics 2016. They focus on techniques that require sophisticated hardware setups to reconstruct high quality shape and reflectance information such as (spatially-varying) BRDFs and BSSRDFs from image data. For more information on the general topic of template and model-based non-rigid registration we refer to the SIGGRAPH Asia 2016 and SGP 2015 courses of Bouaziz et al. [BTP15, BTLP16]. Even though these surveys and courses are related, this state-of-the-art report has a different focus: We focus on methods that extract scene information in an online fashion, e.g., processing and accumulating data directly from a raw RGB-D input data stream. Furthermore, this STAR also describes the acquisition of dynamic scenes and of more sophisticated appearance models, such as spatially-varying BRDFs.



(a) KinectFusion [NDI*11] (b) Voxel Hashing [NZIS13]

(c) Voxels vs. Points [KLL*13a] (d) Elastic Fragments [ZMK13]

Figure 2: *Static Scene Reconstruction: single object reconstruction using KinectFusion 2(a), capturing large scenes using Voxel Hashing 2(b), comparison between voxel-based (left column) and point-based (middle and right column) representations 2(c), and the impact of drift and the corresponding loop closure solution 2(d). Images taken from [NDI*11], [NZIS13], [KLL*13a], and [ZMK13].*

2. Static Scene Reconstruction

3D reconstruction of static environments has its roots in several areas within computer vision and graphics. Online reconstruction is directly related to *Simultaneous Localization and Mapping (SLAM)*, which focuses on the problem of robot navigation in unknown environments; e.g., [MAMT15, ESC14] and many others. Here, the position and orientation of a mobile robot or an unmanned vehicle is tracked (localization) and the observed scene data, mainly the scene geometry, is fused into one common digital representation (mapping). While there is a strong focus on trajectory and pose optimization in SLAM, the reconstruction is typically limited to a sparse point cloud. In computer graphics, on the other hand, dense RGB-D reconstructions with high geometric quality are of primary interest. Most modern approaches are based on the fundamental research by Curless and Levoy [CL96] who introduced the seminal work of volumetric fusion, thus providing the foundation for the first real-time RGB-D reconstruction methods [RHHL02].

The introduction of low-cost RGB-D cameras such as the Microsoft Kinect as part of the XBox game console in combination with the ubiquitous accessibility of GPU processing power opened the door for the online reconstruction of static scenes at consumer level, using the RGB-D camera as a hand-held device. The described theoretical foundation and the availability of commodity hardware, enabled the development of modern online reconstruction methods such as Kinect Fusion [NDI*11, IKH*11], which are the main focus of this section. Poisson surface reconstruction [KBH06, KH13], based on optimizing for an indicator function, is another popular direction, which is often used in the offline context for point cloud data. An overview of RGB-D reconstruction frameworks is provided in Tab. 1.

In the following, we first give a brief overview of a reference system (Sec. 2.1) for online static scene reconstruction that leverages the depth and color information captured by a commodity RGB-D sensor. Further on, we describe the technical details and different choices for each of its constituting components, namely the data preprocessing (Sec. 2.2), the camera pose estimation (Sec. 2.3), and the underlying scene representation (Sec. 2.4).

2.1. Fundamentals of Static Scene Reconstruction

Although there are many different algorithms for RGB-D reconstruction of static scenes, most if not all of these approaches have a very similar processing pipeline, which we describe here for reference (c.f. Fig. 3).

In the first stage, the *Depth Map Preprocessing*, noise reduction and outlier removal is applied to the incoming RGB-D data. Depending on the following stages, additional information is derived from the input range map $\mathcal{V}$ and stored in additional input maps (c.f. Sec. 2.2). In the subsequent stage, the *Camera Pose Estimation*, the best aligning transformation $\boldsymbol{T}$ for the current frame (c.f. Sec. 2.3) is computed. This can be achieved in a frame-to-frame, frame-to-model, or global fashion. Finally, all points $\boldsymbol{p} \in \mathcal{V}$

Kinect fusion用NŽž eö^"
u(ÿ Poisson'Í ^úu(NŽ»~¿
'Í ^ú0

Method	Scene		Data Structure				Tracker			Data Association				Properties			Speed	
	TSDF	Surfel	Dense Grid	Sparse Octree	Sparse Hash	Sparse Points	Frame-to-Keyframe	Frame-to-Model	Global Opt.	Feature Points	Point-to-Point	Point-to-Plane	Dense Color	Loop Closure	Streaming	Rob. to Dynamics	Online	Offline
Newcombe et al. [NDI*11]	✓	-	✓	-	-	-	-	✓	-	-	-	✓	-	-	-	-	✓	-
Izadi et al. [IKH*11]	✓	-	✓	-	-	-	-	✓	-	-	-	✓	-	-	-	✓	✓	-
Whelan et al. [WKF*12]	✓	-	✓	-	-	-	-	✓	-	✓	-	✓	-	-	✓	-	✓	-
Nießner et al. [NZIS13]	✓	-	-	-	✓	-	-	✓	-	-	-	✓	-	-	✓	-	✓	-
Chen et al. [CBI13]	✓	-	-	✓	-	-	-	✓	-	-	-	✓	-	-	✓	-	✓	-
Steinbrüker et al. [SKC13]	✓	-	-	✓	-	-	✓	-	-	-	-	✓	✓	✓	-	-	✓	-
Zhou et al. [ZK13]	✓	-	✓	-	-	-	-	-	✓	✓	-	✓	-	✓	✓	-	-	✓
Zhou et al. [ZMK13]	✓	-	✓	-	-	-	-	-	✓	✓	-	✓	-	✓	✓	-	-	✓
Keller et al. [KLL*13a]	-	✓	-	-	-	✓	-	✓	-	-	-	✓	-	-	-	✓	✓	-
Choi et al. [CZK15a]	✓	-	✓	-	-	-	-	-	✓	-	✓	-	-	✓	✓	-	-	✓
Whelan et al. [WSMG*16]	-	✓	-	-	-	✓	-	✓	-	-	-	✓	✓	✓	-	-	✓	-
Dai et al. [DNZ*17]	✓	-	-	-	✓	-	-	-	✓	✓	-	✓	✓	✓	✓	-	✓	-

Table 1: *Overview of the state of the art in RGB-D based 3D reconstruction: current approaches can be differentiated in terms of the used scene representation, the employed camera tracker, the used data association strategies, support for loop closure, streaming, robustness to dynamic foreground objects, and their runtime performance. All of these dimensions are discussed in detail in this report.*

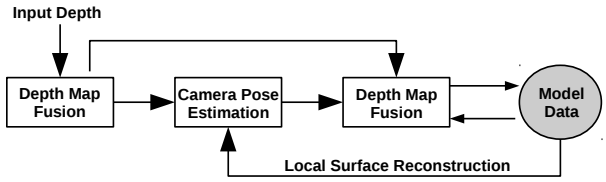


Figure 3: *Overview of the typical RGB-D reconstruction pipeline: first, pre-processed input data is aligned with respect to the current surface reconstruction; second, given the estimated camera pose, the input data is integrated/fused into the current 3D model of the reconstruction.*

from the current input frame are transformed with the estimated transformation $\boldsymbol{T}$ and are merged into the common model $\mathcal{M}$ in the *Depth Map Fusion* stage (c.f. Sec 2.4).

2.2. Depth Map Preprocessing

It has been shown that the noise of depth images of low-cost cameras depends on a variety of parameters, such as the distance to the acquired object, or the pixel position in the depth image [SLK15]. Most commonly, a bilateral filter [TM98] is applied for noise reduction and per-point normals are computed using finite differences (forward or central). Depending on the model representation, data association, and pose optimization approach, further geometric information is estimated. This includes noise or reliability information of the individual range measurements [MHFdS*12, LWK15], the radius of the corresponding 3D point [KLL*13b] or principal curvatures [LKS*17].

2.3. Camera Pose Estimation

Pose estimation computes a 6-DoF pose $\boldsymbol{T}$ for every incoming RGB-D frame with respect to the previous frame, to the so-far reconstructed model, or to all previous frames.

2.3.1. Tracking Objectives

Early works on off-line 3D shape registration heavily inspire current approaches for real-time camera tracking based on depth streams. The first proposed techniques employed simple frame-to-frame variants of the Iterative Closest Point Algorithm (ICP) [BM92, YM92] and were based on a point-to-point [BM92] or point-to-plane [YM92] error metric. Frame-to-frame tracking estimates the delta transformation $\Delta\boldsymbol{T}_{t-1}$ to the previous input frame and concatenates the estimate to the previous pose estimation result $\boldsymbol{T}_t = \Delta\boldsymbol{T}_{t-1} \cdot \boldsymbol{T}_{t-1}$. With the invention of fast and efficient variants of ICP [RL01], online in-hand scanning with live feedback became a reality [RHHL02]. This was a big step forward, since the tight feedback loop enabled the user of the system to fill reconstruction holes and decide if the object has already been completely digitized. One severe problem of the employed frame-to-frame strategies is the accumulation of tracking drift over long scanning sequences.

To reduce this problem, frame-to-model tracking has been extensively used in recent online RGB-D reconstruction frameworks [NDI*11, IKH*11, WKF*12, CBI13, NZIS13]. Tracking is based on the point-to-plane ICP described by [Low04a]. Frame-to-model tracking has two significant advantages over a simple frame-to-frame alignment strategy. First, instead of the last frame, a synthetically rendered depth map of the current reconstruction state is employed to anchor the reconstruction, thus drastically reducing temporal tracking drift. Second, if a point-to-plane distance metric

is used, the stabilized model normals can be utilized to define the tangent plane instead of the noisy input normals, which leads to higher accuracy tracking and increased robustness.

While frame-to-model tracking significantly reduces temporal tracking drift, it does not completely solve the problem of local error accumulation, since local tracking errors can still add up over time. This can lead to loop closure problems if already scanned parts of the scene are re-encountered over a different path in the same scanning session. In this case, the previously obtained reconstruction will not match the current depth observations leading to tracking failure and/or double integration of surfaces. To alleviate such problems, approaches for global pose optimization were introduced and first applied to off-line 3D scene reconstruction [ZK13, ZMK13]. The approach of Zhou et al. [ZK13] produces high-quality reconstructions using a hand-held consumer-grade sensor. They use points of interest to preserve local detail in combination with global pose optimization to distribute alignment errors equally over the scene. In a follow-up work [ZMK13], to further increase the fidelity of the reconstruction, non-rigid bundle adjustment is performed on elastic fragments to deal with low-frequency camera distortions. The off-line approach of Choi et al. [CZK15a] employs robust global pose optimization based on line processes to eliminate wrong matches and improve reconstruction quality. Very recently, the BundleFusion approach of Dai et al. [DNZ*17] enabled globally consistent reconstruction at real-time frame rates based on online bundle adjustment and surface re-integration. For computational efficiency and real-time performance, this approach organizes the input stream in a hierarchy of chunks and leverages the computational power of modern graphics hardware for data-parallel bundle adjustment. In the first step, all frames in a new chunk are globally aligned to each other (intra-chunk alignment). Afterwards each new chunk is globally aligned with respect to all previous chunks (inter-chunk alignment).

2.3.2. Data Association

Most camera tracking approaches that rely on frame-to-frame, frame-to-model, or global pose optimization, require the identification of *corresponding points* between individual frames and/or the current model. The set of corresponding point pairs is fed into the optimization (c.f. Sec. 2.3.1) in order to find the transformation that results in the best overall alignment. Essentially, there are sparse approaches, which identify specific feature points, and dense techniques that try to find correspondences to (almost) all point of the incoming frame.

Sparse Correspondences In general, a set of sparse correspondences is computed by matching feature points of the current color and depth input to detected corresponding features in the previous frames or in the model. Due to the computational complexity of dense correspondence finding, early approaches only use a subset of the incoming RGB-D observations. Statistical tests and sampling on the data can be used to increase the number of good correspondences [RL01, GIRL03]. Detected and matched sparse color features over a temporal sequence of input frames can provide a sparse set of valuable initial correspondences [GRB94]. A popular choice for the feature extraction and matching is SIFT [Low99, Low04b, LS09], which has been applied in several 3D scene reconstruction approaches [ZK15, HKH*12, WJK*13, DNZ*17]. However, there are many alternative feature sparse feature descriptors such as SURF [BTVG06], ORB [RRKB11], or more recently even learned descriptors [HLJ*15, YTLF16, ZSN*17] . Another approach is to search for correspondences across multiple frames [WG17].

Dense Correspondences All recent approaches use dense correspondence finding in conjunction with projective data association [BL95] and in combination with specific compatibility criteria in order to select the "best" model point related to a given input point by checking its projective neighborhood in image space. Most approaches [NDI*11, IKH*11, WKF*12, NZIS13, CBI13, SKC13, ZMK13, KLL*13a, WSMG*16, DNZ*17] measure spatial proximity based on a point-to-plane [YM92] error metric. The point-to-plane metric can be considered a first order approximation of the distance to the target surface. Beyond spatial proximity, recent approaches potentially consider distance related sensor uncertainty [NIL12], compatibility of surface color [GRB94, NLB*11, SKC13, RLL14], of normals [Pul99, KLL*13b], of gradients [SG14], and of local curvature [LKS*17].

2.3.3. Relocalization

The recovery from tracking failure is a crucial step in any robust camera tracking system. One of the first approaches to solve this problem was proposed by Shotton et al. [SGZ*13] by using regression forests to predict a probability density function of a pixel's location and was later extended to predict multi-modal Gaussians [VNS*15]. In their follow-up work, they propose a new general framework for minimizing the reconstruction error in an analysis-by-synthesis approach, which includes camera pose estimation using a retrieval forest and a navigation graph as search structure [VDN*16], thus also enabling RGB-to-3D-model localization. To overcome the problem of having to pre-train a regression forest for each new scene, which takes several minutes, Cavallari et al. [CGL*17] introduced an approach that can adapt a generic pre-trained forest to a new scene, thus making regression forests real-time capable.

A keyframe-based technique for relocalization was proposed by Glocker et al. [GSCI15] where, in case of tracking failure, poses of keyframes that are similar to the current frame are retrieved and used for reinitialization of the camera tracking. This approach has been used by Whelan et al. [WLSM*15] to detect previously scanned areas and, in combination with a non-rigid deformation algorithm (Sec. 3), to solve for loop closures. The global alignment strategy of Dai et al. [DNZ*17] solves tracking failures implicitly as new chunks are globally compared to all previous chunks. If the pose optimization for the current chunk fails, the chunk is ignored, otherwise its geometry is integrated into the 3D model and its pose is stored.

2.4. Geometry Representations and Fusion

The representation of the model $\mathcal{M}$ needs to be very efficient in integrating the large amount of incoming range maps. Beyond this, frame-to-model tracking requires an efficient way to generate virtual views of the model from arbitrary viewpoints in order to align

the incoming range maps with the model, mostly using projective data association. There exist mainly two different representations to accumulate the observed RGB-D data in one common 3D model. The most commonly used way is to store the information in a regular or hierarchical 3D voxel grid. Alternatively, the model can be stored as accumulated 3D point set.

2.4.1. Voxel-Based Representation

The original research by Curless and Levoy [CL96] introduced volumetric fusion by using a regular grid to store a discretized version of a signed distance function (SDF) that represents the model. This was adopted by the first real-time approach by Rusinkiewicz et al. [RHHL02], and later by modern KinectFusion methods [NDI*11, IKH*11]. Voxel-based representations are used to implicitly store the model surface using an SDF, i.e., interior and exterior voxels store the negative and positive distance to the closest surface point, respectively. The surface itself is defined as the zero-crossing of the SDF. The voxel size and the spatial extent of the grid has to be defined prior to the execution of the algorithm. Additional data, such as color, is commonly stored in per-voxel attributes.

As voxels close to the surface are of particular interest, truncated signed distance functions (TSDFs) are commonly used. The basic approach for accumulation of incoming range maps in voxel based representations is to project each voxel onto the range map using the estimated camera pose (c.f. Sec.2.3) and to evaluate its projective distance. Commonly, the merging step is accomplished by weighted averaging of the prior TSDF values in the grid and the incoming one related to the range map. This merging step, which is normally implemented using a weighted averaging scheme is very efficient in removing sensor noise based on the temporal integration of multiple distance samples.

To visualize voxel-based representations, e.g., for generating virtual views, ray-casting is applied on the voxel-grid, e.g., using digital differential analysis (DDA) [AW*87] in combination with analytic iso-surface intersection [PSL*98]; alternatively a 3D mesh can extracted with the Marching Cubes level set method [LC87].

Regular voxel grids are very inefficient in terms of memory consumption and bound to a predefined volume and resolution. In the context of real-time scene reconstruction, where most approaches heavily rely on the processing capability of modern GPUs, the spatial extent and resolution of the voxel grid is typically limited by GPU memory. In order to support larger spatial extents, various approaches have been proposed to improve the memory efficiency of voxel-based representations. In order to prevent data loss due to range images acquired outside of the current reconstruction volume, Whelan et al. [WKF*12] propose a simple dynamic shift of the voxel grid, such that it follows the motion of the camera. The approach converts the part of the volume that is shifted outside the current reconstruction volume to a surface mesh and stores it separately. While this enables larger scanning volumes, it requires heavy out-of-core memory usage and already scanned, and streamed-out surfaces, cannot be arbitrarily re-visited.

Voxel Hierarchies One way to store the surface effectively is to use a hierarchy of voxels, such as an octree, where the (T)SDF can be encoded in a sparse fashion around the actual surface. Though not real-time capable, Fuhrmann and Goesele [FG11] introduced a hierarchical SDF (hSDF) structure using an adaptive, octree-like data structure providing different spatial resolutions. Zeng et al. [ZZZL12, ZZZL13] use a fixed 4-level hierarchy and store the TSDF on the finest level only. Chen et al. [CBI13] proposed a similar 3-level hierarchy also with fixed resolution. Steinbrücker et al. [SSC14] represents the scene in a multi-resolution data structure for real-time accumulation on the CPU including an incremental procedure for mesh outputs. Henry et al. [HFBM13] subdivide the scene into many patch volumes, each consisting of a regular voxel grid of arbitrary size and resolution. The patch volumes are spatially organized in a pose graph, which is optimized in order to achieve a globally consistent model. This approach is interactive, but not fully real-time. An extremely compact hierarchy is proposed by Reichl et al. [RWW16], who only store a binary grid which is updated in a running window fashion.

Voxel Hashing Voxel hashing was introduced by Nießner et al. [NZIS13]. The approach represents a virtually infinite scene by a regular grid of smaller voxel blocks of predefined size and resolution (c.f. Fig.2(b)), whose spatial locations are addressed with a spatial *hash function*. Only the voxel blocks that actually contain geometry information are instantiated, and the corresponding indices are stored in a linearized spatial hash. This strategy significantly reduces memory consumption and allows (theoretically) infinite scene sizes. Compared to hierarchical approaches, the main benefit lies in the very efficient data insertion and access, which both is in $O(1)$. The minimal memory and the high computational efficiency makes hashing-based 3D reconstruction methods even applicable for mobile phones, such as used in Google Tango [DKSX17]. In addition, it allows to easily stream parts of the reconstruction out of core to support high resolutions and extremely fast runtime performance. Kahler et al. [KPR*15] adopt the concept of voxel block hashing, but uses a different hashing method to reduce the number of hash collisions. Further on, they provide a hashing approach for irregular grids with different resolution levels in order to capture different parts of the scene at different levels of detail [KPVM16].

2.4.2. Point-Based Representation

Alternative to voxel-based representations, the acquired range images can be directly stored and accumulated in a point- or surfel-based model [PZvBG00]. sThis sparse, point-based strategy is used by several reconstruction approaches [KLL*13b, WLSM*15] (c.f. Fig.2(c)). Additional information, such as point size/radius, color, or other information, are stored as per-point attributes. The point size, computed in the preprocessing step (Sec. 2.2), originates from the projective acquisition of the range data and intrinsically supports an adaptive resolution of the model. In order to prevent the accumulation of outliers in point based models, at least two point states are distinguished, i.e., stable and unstable points. Initially, as points get into the model, they are unstable. Points get stable after they have been merged with further incoming points at least a certain number of times (see below). This stabilization procedure may also include further confidence scores related, e.g., to the reliability of the incoming point.

Merging a new set of incoming points to the model first needs explicit point correspondences between incoming and model points.

Therefore, the neighborhood of model points $\mathcal{N}(\boldsymbol{p}) \subset \mathcal{M}$ of each incoming point $\boldsymbol{p}$ is determined by rendering an *index map*. Here, the indices of the model points are rendered in the image plane. In order to get access to neighboring points, the index map is rendered at a higher resolution than the input range map. Afterwards, the best matching model point is determined using the same or very similar rules as in dense correspondence finding; c.f. Sec. 2.3.2.

If used for visual feedback, this naive rendering of individual points, however, yields incomplete images with holes. In order to render dense images, point splatting is commonly applied [RHHL02, WWLVG09, LKS*17]. Here, each 3D point is projected as a circle with a certain radius onto the 2D image, resulting in dense images.

2.4.3. Hybrid Approaches

Salas et al. [SMGKD14] distinguish between (nearly) planar and non-planar regions with high curvature. The planar regions are clustered and labeled in the model. Points of the same cluster are forced to share the same normal, which intrinsically denoises the accumulated 3D data. The algorithm is capable of merging different clusters, which belong to the same plane, thus refining the data in the case of loop closures. Non-planar data is accumulated in a point-based model; c.f. Sec. 2.4.2.

2.5. Incorporation of Sensor Uncertainty

Sofka et al. [SYS07] improve the correspondences search by using the covariance matrix of the estimated transformation and individual point correspondences. The observed uncertainty is then used to improve correspondences and the estimated transformation by an EM-like approach.

Maier-Hein et al. [MHFdS*12] propose an ICP variant that accounts for anisotropic uncertainty in the point positions using the Mahalanobis distance. Lefloch et al. [LWK15] extended this approach to online scene reconstruction in order to tackle anisotropic sensor noise. They apply a two-sided Mahalanobis distance in the dense correspondence search and accumulation stage; c.f. Sec. 2.3.2 and Sec. 2.4.2.

2.6. Autonomous 3D Reconstruction

Even given a robust state-of-the-art online 3D reconstruction approach, completely digitizing an object or even an entire scene at high-quality is a tedious and time consuming process. Obtaining a complete reconstruction of an object requires that it is viewed from a large number of different viewpoints, and for completely digitizing an entire scene, e.g., a large living room, the user has to traverse the scene to gather depth samples of the entire surface geometry. This process can easily amount to several minutes per scan, and the chosen path through the scene impacts scanning time and reconstruction quality. Finding the optimal sensor path is a challenging problem and can be phrased as an optimization problem. Auto-scanning approaches automate the scanning process by solving the underlying optimization problem to produce the control signals for a robotics system. First approaches [KCF11, KRB*12, WSL*14, KRBS15] digitized single objects based on a controllable robot arm that holds and moves an object in front of a depth camera. Auto-scanning boils down to a view planning problem, where the next best view for scanning has to be found based on the current partial reconstruction of the object. Some approaches [KCF11, KRB*12, KRBS15] aim at minimizing the number of views to cover the complete object, while others [WSL*14] are focused on maximizing the reconstruction quality. Another class of approaches aims at digitizing complete scenes, e.g., an apartment, based on driving robots [CLKM15, XZY*17] or flying drones [HBH*11, BPH*12, SBK*13]. In these approaches the speed of scene exploration has to be balanced with respect to the ability of the system to perform simultaneous localization and mapping of the environment, and the underlying reconstruction approaches have to scale to much large environments. Driving robots are bound to the ground plane, which simplifies the underlying path planning problem to a 2D problem. In general, such systems can not reach all parts of the scene which leads to incomplete reconstructions. To tackle this problem, auto-scanning approaches based on flying drones have been proposed [HBH*11, BPH*12, SBK*13]. The underlying path planning is now a full 3D problem and thus more challenging to solve. First approaches [XLC*18] even enable to reconstruct dynamic scenes based on cooperating drones.

2.7. Datasets

There are several datasets for evaluating static 3D scene reconstruction approaches that mainly differ in the type of sensor data and the provided ground truth. Here, we focus on datasets for evaluating RGB-D scanning approaches. The datasets by Zhou et al. [ZK13] and Glocker et al. [GISC13] contain RGB-D sequences together with camera pose estimates by the respective approach. In contrast, Sturm et al. [SEE*12], as well as Pomerleau et al. [PMC*11] provide ground truth trajectories from an external, high-accuracy motion-capture system. In addition, some datasets include segmentation masks and object labels [KAJS11, SF11, NSF12, SLX15] or contain the ground truth geometry of the acquired objects [WMS16]. The SUN3D dataset [SLX15] provides a database of big indoor scenes that have been reconstructed using structure from motion from RGB-D scanning sequences. Handa et al. [HWMD14] created the ICL-NUIM dataset based on two synthetically rendered 3D scenes (Living Room, Office) using the POVRay raytracer [BC04]. They provide the ground truth camera trajectory for both scenes as well as the synthetic ground truth 3D models for the Living Room scene. In addition to the evaluation of the camera tracking accuracy, this enables an evaluation of the dense surface reconstruction error with respect to the ground truth 3D model. The augmented ICL-NUIM dataset [CZK15b] extends this dataset by adding four additional camera trajectories with multiple loop closures that emulate a realistic hand-held RGB-D scanning sequence. The synthetic RGB-D streams are generated based on a realistic noise model that emulates the deficiencies of commodity RGB-D sensors in terms of noise characteristics, quantization errors, and lens distortion. They also provide a dense point-based surface model for the Office scene, which enables the evaluation of surface reconstruction accuracy. Very recently, Bulczak et al. [BLK18] present a ToF camera simulator that incorporates sensor noise, multi-path effects, and other real-world sensor errors. For semantic classification we have seen extensive work on synthetic data, such as SceneNet [HPB*15] or SUNCG [SYZ*16], as well as annotated real-world data, including

ScanNet [DCS*17] and Matterport3D [CDF*17]. A good and recent overview and classification of the vast amount of RGB-D datasets is given by Firman [Fir16].

3. Capturing Dynamic Scenes

In addition to static components, many natural environments contain dynamic objects, such as closely interacting people. Obtaining temporally coherent reconstructions that capture the non-rigid surface motion at high quality is a highly challenging and ill-posed problem, especially if real-time performance is the target. However, fast and robust solutions have a high impact in multiple important research fields and provide key technological insights. Applications of dynamic scene reconstruction can be found in content creation for visual effects, computer animation, man-machine interaction, biomechanics, and medicine.

More recently, dynamic reconstruction techniques also find their application in the context of virtual (VR) and augmented reality (AR). A prominent recent example is the impressive Holoportation [DKD*16a, OERF*16] system, which performs online reconstruction of a dynamic scene and enables full body telepresence in augmented reality.

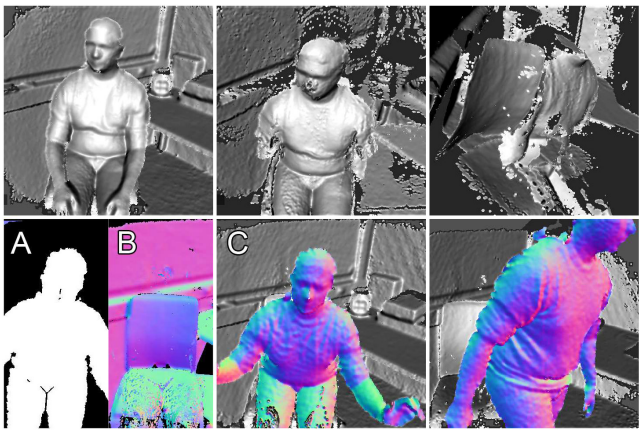


Figure 4: *A sitting person is first reconstructed and then starts to move. This non-rigid scene motion leads to a failure of camera tracking (top). The approach of Keller et al. [KLL*13a] computes a foreground segmentation based on scene dynamics (A) and excludes them from camera pose estimation (B). This enables robust camera tracking even if large parts of the scene are moving (bottom). Image taken from [KLL*13a].*

3.1. Robustness to Dynamic Objects

The first step in dealing with dynamic scenes is to make the reconstruction approach robust with respect to moving objects. In volumetric fusion approaches that are based on implicit signed distance fields [CL96] – such as KinectFusion [NDI*11, IKH*11] – the surface reconstruction is restricted to static environments. One way to address dynamic scene elements is to treat them as outliers during projective data association in the ICP tracking to avoid breaking the reconstruction [IKH*11]. If there are multiple rigid objects, one could in principal use several separate volumes to track and reconstruct each object independently. Handling dynamic foreground objects (see Fig. 4) becomes significantly easier with a surfel-based representation, as shown by Keller et al. [KLL*13a]. Jaimez et al. [JKGJC17] also classify the scene into static and dynamic parts. Other approaches [DF14] assume a pre-scanned version of the static environment as a prior. Afterwards, the camera motion is tracked online, while the dynamic parts of the scene are segmented and reconstructed separately. The *Co-Fusion* approach [RA17] enables the independent reconstruction of multiple rigidly moving objects in addition to a static background model.

3.2. Challenges of Dynamic Reconstruction

The reconstruction of dynamic scenes is computationally and algorithmically significantly more challenging than its static reconstruction counterpart. Modeling the non-rigid motion of general deforming scenes requires orders of magnitude more parameters than the static reconstruction problem [PRR03]. In general, finding the optimal deformation is a high-dimensional and highly non-convex optimization problem that is challenging to solve, especially if real-time performance is the target.

In addition to the many challenges of the static reconstruction problem, there exist infinitely many solutions [FNT*11] that non-rigidly deform one shape to another, which makes the dynamic reconstruction problem inherently ill-posed. Even if a static 3D template of the deforming scene is available, e.g., obtained by rigid fusion, estimating the non-rigid motion based on a single sensor is still a very challenging problem, since more than half of the scene is occluded at each time instance. Fast motion leads to large frame-to-frame differences, which makes tracking challenging, especially for highly deformable objects [LAGP09].

If no template can be acquired beforehand, object motion and shape have to be recovered simultaneously. This is an inherently ambiguous joint problem, since changes in the observations can be explained by both dimensions. Although more complicated, template-free approaches [BHLW12, TBW*12, LLV*12, CCS*15, NFS15, IZN*16, DKD*16b, SBCI17, DKD*16b, SBCI17] have obtained impressive results over the past years.

The key component of all state-of-the-art template-based and template-free dynamic reconstruction techniques is a robust and fast non-rigid registration framework.

3.3. Fundamentals of Non-Rigid Registration

Given a source shape $\mathbf{S} \subset \mathbb{R}^3$ and a target shape $\mathbf{T} \subset \mathbb{R}^3$, the objective of non-rigid registration is to find a warp field $W : \mathbb{R}^3 \rightarrow \mathbb{R}^3$, such that the warped source best explains the target $W(\mathbf{S}) = \mathbf{T}$. In practice, the source and target are often presented as depth maps, triangle meshes, or signed distance functions. Mixed representations are common in the literature [LSP08, LAGP09, ZNI*14, NFS15, IZN*16, DKD*16a, GXY*17, YGX*17]. Current approaches favor warp fields based on coarse deformation proxies, since this reduces complexity and enables real-time performance.

The non-rigid registration problem can be viewed from two perspectives: The problem of finding the optimal warp field W or the

problem of finding the dense shape-to-shape correspondence C. If the correspondence is available, the warp field can be easily computed. The same holds vice versa. Finding the best warp field W^* and correspondence C^* between $\mathbf{S}$ and $\mathbf{T}$ can be formalized as a highly non-linear joint optimization problem:

$$C^*, W^* = \arg\min_{C,W} E_{\text{total}}(C, W) \quad . \tag{1}$$

Since there are in general infinitely many valid mappings [FNT*11] that deform one shape to another, the energy function E_{total} normally consists of several data E_{fit} and regularization E_{reg} terms:

$$E_{\text{total}}(C, W) = E_{\text{fit}}(C, W) + \lambda \cdot E_{\text{reg}}(C, W) \quad . \tag{2}$$

While the data fitting terms measure the closeness of the model to the input, the regularization terms encode prior assumptions about the deformation behavior of the source model. λ is a trade-off factor that balances the relative importance of the two components.

Formulating the non-rigid registration problem as the joint problem of recovering both W^* and C^* simultaneously suggests a class of iterative EM-style solution strategies [DLR77] that are well known as *Non-Rigid Iterative Closest Point* (N-ICP) algorithms [IGL03, BR04, BR07]. The key insight of these approaches is to split the complicated joint optimization problem into two easier sub-steps: An initial correspondence search based on the last warp field estimate (E-step) and the update of the warp field based on the found correspondences (M-step). The N-ICP algorithm iterates between these two steps until convergence.

Many recent approaches rely on N-ICP [LSP08, LAGP09, NFS15, GXW*15, IZN*16, DKD*16a], others directly target the underlying joint optimization problem [ZNI*14, SBCI17]. In the following, we provide more details on the employed deformation proxies and the employed energy formulation.

3.3.1. Deformation Representation

The choice of a suitable deformation representation is of paramount importance, since it heavily influences the algorithmic design of every aspect of the non-rigid registration approach. Dense mesh-based representations are currently not employed in the state-of-the-art, especially not in current online approaches, since such a representation leads to a high-dimensional optimization problem. Many approaches rely on coarse deformation proxies [LSP08, CZ09, LAGP09] that decouple the optimization problem from the resolution of the underlying 3D template mesh. Current approaches vary in the choice of the employed deformation proxy.

Coarse Tetrahedralization A common choice for the deformation proxy is a coarse-scale version of the template model [BHZN10] or a coarse volumetric tetrahedralization. This representation has been extensively used in the context of off-line non-rigid tracking of bodies [AFB15, dAST*08]. More recently, Zollhöfer et al. [ZNI*14] proposed to use a coarse volumetric tetrahedralization for real-time deformable template tracking based on a single RGB-D camera. The coarse tetrahedralization has two main advantages over the initial detailed template mesh: It drastically reduces the number of free variables of the underlying optimization problem, thus enabling real-time performance. An additional advantage is the fact that the additional *Steiner* points inside the model stabilize the deformation by preserving local volume and lead to a faster propagation of the residual energy to occluded parts of the model.

Regular Volumetric Grids Another common choice, which has been widely used in recent state-of-the-art approaches are coarse regular volumetric grids [SP86, CZ09, ZSGS12, IZN*16, SBCI17]. This proxy is based on the idea of free-form deformation [SP86] and has been used before for online handle-based modeling [ZSGS12]. It shares all the advantages of the proxies that are based on a coarse tetrahedralization. In addition, its high regularity results in good access patterns and allows for easy parallelization of the optimization strategy on commodity graphics hardware [IZN*16]. While most approaches store a full 6 DoF transformation per grid cell, some recent approaches directly work on vector fields [SBCI17].

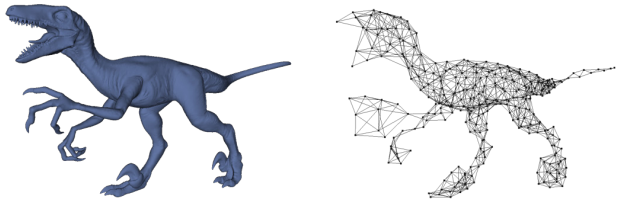


Figure 5: *The deformation graph (right) of Sumner et al. [SSP07] decouples the computational complexity from the underlying mesh (left). The deformation graph is the most commonly used deformation proxy in off-line as well as online non-rigid reconstruction techniques.*

Coarse Deformation Graph The most commonly used representation in current state-of-the-art approaches is the *Deformation Graph* introduced by Sumner et al. [SSP07], see Fig. 5. It was heavily used over the past years in many off-line [LSP08, LAGP09, DFF13, DTF*15, GXW*15, GXW*17] non-rigid registration techniques due to its simplicity and generality. More recently, it is also used frequently in online [NFS15, DKD*16a, GXY*17] non-rigid reconstruction approaches. Similar to the other deformation proxies, it decouples the optimization problem from the underlying fine template mesh. One significant advantage of the Deformation Graph is its high adaptability to the shape of the underlying mesh.

3.3.2. Data Fitting Terms

In the literature, many different data fitting terms have been proposed and used over the past years. The used terms are very similar to the ones used for camera tracking in the static reconstruction problem, see Sec. 2.3.2. Recent approaches employ a combination of terms. In the following, we discuss the terms that are used most frequently in practice.

Sparse Features Sparse constraints, such as detected and matched color features, are used in many approaches [DFF13, IZN*16]. These known correspondences are complementary to the approximate correspondences discovered during the N-ICP iteration steps and lead to faster convergence, since they guide the optimizer through the complex energy landscape. In addition, a sparse set of correspondence matches can help significantly to better enforce loop closure and stabilizes the alignment in the tangent plane of the model.

Dense Geometry Constraints Besides sparse constraints, recent state-of-the-art approaches heavily rely on dense constraints. This includes dense geometric point-to-point [LSP08, LAGP09, ZNI*14, DTF*15, GXW*15, NFS15, IZN*16, GXW*17] and point-to-plane [LAGP09, ZNI*14, NFS15, IZN*16, DKD*16a, GXY*17, GXW*15, GXW*17, GXY*17] alignment. These two terms are often used in combination to achieve a higher tracking accuracy. Besides these surface matching constraints, additional normal matching [DTF*15] and convex hull constraints [DKD*16a] can be employed. Other approaches employ point-to-tsdf alignment [DFF13]. The advantage of this constraint is that it does not require explicit correspondence search, since the optimizer can directly follow the local TSDF gradients of the target surface. Even tsdf-to-tsdf alignment [SBCI17] has been demonstrated. The key idea is to represent the source and target surface as distance fields, and align them directly.

Dense Photometric Constraints The color constancy assumption is often used to define a dense photometric term [DFF13, ZNI*14] for better alignment in the tangent plane. The direct use of color for photometric alignment is problematic if the illumination is temporally varying or if the object undergoes large deformations, since this leads to drastic appearance changes due to shading effects. To tackle this problem, Guo et al. [GXY*17] propose to use reflectance constancy. To this end, their approach jointly solves for geometry, motion, surface reflectance, and incident illumination. Instead of directly using the color information, recent state-of-the-art approaches also incorporate dense regressed correspondences [DKD*16a, WFR*16].

3.3.3. Regularization Strategies

The dynamic reconstruction problem is severely under-constrained, since there exist infinitely many solutions [FNT*11] that align two objects non-rigidly. Thus, the data fitting terms alone are in general not sufficient to uniquely constrain the solution. To resolve ambiguities, different regularization terms have been proposed that encode prior assumptions about the deformation behavior of the scene.

Linear Deformation Linear mesh deformation techniques [BS08], e.g., thin-plate splines [Boo89], are nowadays mainly employed for fine-scale alignment tasks after the model has been coarsely matched to the input [LAGP09, ZNI*14]. Such linear techniques are currently not used in the state-of-the-art for coarse model alignment, since they do not handle rotations well.

Non-Linear Deformation Non-linear regularization energies are the de facto standard in current off-line and online dynamic reconstruction techniques, due to their ability to handle large rotations. One popular regularization energy is the as-rigid-as-possible paradigm [SA07]. It enforces the deformation field to be locally as rigid as possible to prevent unnecessary stretching and shearing of the template geometry. This deformation paradigm has been applied to both real-time template-based [ZNI*14] as well as template-free [IZN*16] non-rigid reconstruction approaches. The most commonly used deformation framework is *Embedded Deformation* [SSP07]. It is used in a large percentage of recent state-of-the-art online [NFS15, DKD*16a, GXY*17] and off-line [DFF13, DTF*15, GXW*15, GXW*17, LSP08, LAGP09] approaches. It has two distinct components, a soft-constraint that enforces local rigidity and a second soft-constraint that enforces spatial smoothness of the warp field. This is in contrast to the as-rigid-as-possible paradigm [SA07] that only enforces local rigidity. Recently, a damped version of the *Killing Vector Fields* [BBSG10, SBCBG11, TSB16] regularization term has been applied to the problem of template-free online surface reconstruction [SBCI17]. Killing Vector Fields enforce the deformation to be locally isometric. Local isometry is also employed as regularization constraint by [DKD*16a].

3.4. Static Reconstruction of Quasi-Rigid Objects

The fundamental assumption of static scene reconstruction systems is that the scene remains entirely static throughout the complete scanning process. If this assumption is violated, the reconstructed 3D model will contain artifacts or the approach will completely fail. Even for static scenes, sensor calibration errors can lead to a significant non-rigid spatial distortion, such that the captured depth maps cannot be aligned rigidly [IGL03]. Allowing for a small amount of residual non-rigid deformation can alleviate this problem [IGL03]. Non-rigid registration has also been applied to online loop closure for in-hand scanning [WWLG11]. A completely static scene is difficult to guarantee in many real world scenarios, e.g., for the 3D reconstruction of an animal or a baby that will not hold still during scanning. Many approaches have been developed that enable the static reconstruction of a high-quality 3D model even if the scene undergoes slight non-rigid deformations [BR07, WHB11, TZL*12, ZZCL13, DFF13, DTF*15]. This is of particular interest for the digitization of humans with commodity depth sensors [BR07, TZL*12, LVG*13, ZZCL13, DTF*15].

3.5. Non-rigid Reconstruction of Dynamic Scenes

In the following, we discuss approaches that reconstruct the motion of dynamic scenes. We start with approaches that exploit strong scene specific priors, and highlight recent progress in the less constraint template-based and template-free reconstruction setting. In the last years, many algorithmic improvements and the steady growth of data parallel compute power led to the first online approaches that were able to handle general scenes, see Tab. 2.

3.5.1. Strong Scene Priors

Special purpose solutions exist that enable high-quality reconstruction and tracking of certain object classes based on commodity RGB-D cameras. These special purpose solutions exploit class specific knowledge and strong priors to simplify the reconstruction and tracking problem. Significant progress has recently been made in the reconstruction and tracking of faces [WBLP11, LYYB13, BWP13, TZN*15, GZC*16], hands [IOA11, TST*15, TBC*16, TPT16] and entire bodies [YLH*12, HBB*13, WSVT13, ZFYY14, BBLR15, YGX*17]. Bogo et al. [BBLR15] obtain textured detailed full-body reconstructions of moving people from RGB-D sequences using an extension of the *BlendSCAPE* model. Other approaches work for general articulated shapes [YY14, SNF15]. Given such strong priors, it is nowadays even possible to solve many of these problems at real-time frame rates [WBLP11, LYYB13, TZN*15, TST*15, TBC*16, TPT16, HBB*13, YGX*17]. For example, Thies

Method	Input				Deformation Proxy				Data Fitting Terms							Regularizers			Properties		
	Single-View	Multi-View	Commodity	Custom Rig	Tetrahedrons	Vector Field	Proxy Lattice	Proxy Graph	Point-to-Point	Point-to-Plane	TSDF	Features	Regression	Color Const.	Refl. Const.	ARAP	Emb. Def.	Killing Vectors	Template-free	Fast Motion	Top. Change
[ZNI*14]	✓	-	-	✓	✓	-	-	-	✓	✓	-	-	-	✓	-	✓	-	-	-	-	-
[NFS15]	✓	-	✓	-	-	-	-	✓	✓	✓	-	-	-	-	-	-	✓	-	✓	-	-
[IZN*16]	✓	-	✓	-	-	-	✓	-	✓	✓	-	✓	-	-	-	✓	-	-	✓	-	-
[DKD*16a]	-	✓	-	✓	-	-	-	✓	-	✓	-	-	✓	-	-	-	✓	-	✓	✓	✓
[GXY*17]	✓	-	✓	-	-	-	-	✓	-	✓	-	-	-	-	✓	-	✓	-	✓	-	-
[SBCI17]	✓	-	✓	-	-	✓	-	-	-	-	✓	-	-	-	-	-	-	✓	✓	✓	✓

Table 2: *Overview of state-of-the-art online dynamic reconstruction approaches that do not require a strong prior, such as a skeleton.*

et al. [TZN*15] reconstruct facial identity, expression and incident illumination at real-time rates [TZN*15]. They employ a parametric face and a blendshape expression model to reduce the number of unknown parameters significantly. Taylor et al. [TBC*16] track the full articulated motion of a hand based on a kinematic skeleton in real-time using only a single depth camera. These approaches achieve impressive results, but due to the employed strong prior, they only reconstruct a limited subspace of all possible deformations and do not generalize to general non-rigid scenes.

Figure 6: *The first real-time template-based tracking approach that works for general objects was proposed by Zollhöfer et al. [ZNI*14]. Real-time performance is made possible by a novel hierarchical coarse-to-fine GPU optimization strategy. Image taken from [ZNI*14].*

3.5.2. General Deformable Object Tracking

Non-rigid ICP was first proposed in the context of non-rigid 2D shape registration [PRR03] and later also extended to non-rigid registration in 3D [FNT*11]. The first approaches where employed to align multiple range scans [BR04, BR07, ZMK13] to counteract non-rigid distortions caused by imperfect camera calibration. The first non-rigid registration approaches that were able to track complex deformations [CZ09, LZW*09, LSP08, LAGP09, GXW*15, GXW*17, XLC*18] used deformation proxies to decouple the dimensionality of the optimization problem from the model complexity, but still had slow off-line runtimes. Many recent approaches for robust off-line template tracking use key frames and robust optimization [LAGP09, GXW*15, GXW*17, XLC*18]. Other approaches employ the ℓ_0-norm [GXW*15] or a robust norm [DKD*16a] to define the regularization objective. This allows for discontinuities in the deformation field, which is especially advantageous for tracking articulated motion.

Online deformable tracking of arbitrary general deforming objects, without the use of strong priors, has only been achieved quite recently. The first approach to deliver real-time performance in this setting was the template-based non-rigid tracking approach proposed by Zollhöfer et al. [ZNI*14], see Fig. 6. Input to this approach is a high-quality color and depth stream, which is captured by a custom-built RGB-D sensor. After a template acquisition step, the non-rigid object motion is tracked live at real-time frame rates using robust optimization [LSP08, Zac14a, ZNI*14]. This is made possible by a hierarchical coarse-to-fine GPU registration approach that exploits the data-parallel compute power of modern graphics hardware. In contrast to N-ICP approaches, Zollhöfer et al. [ZNI*14] jointly optimize for the best correspondence. While this approach enabled real-time tracking of general objects, a pre-acquired template mesh has to be available. Acquiring such a template for each scene is a tedious and time-consuming process that might be infeasible, e.g., for animals or small children that will not hold still.

Figure 7: *Similar to DynamicFusion by Newcombe et al. [NFS15], VolumeDeform by Innmann et al. [IZN*16] enables template-free non-rigid reconstruction of general dynamic scenes. The warp field is parameterized based on a fine-scale deformation lattice instead of a coarse-scale deformation graph, and sparse feature matches are integrated into the alignment objective. Image taken from [IZN*16].*

3.5.3. Template-free Deformable Reconstruction

If a template model of the object cannot be obtained beforehand, the more challenging joint geometry and motion reconstruction problem has to be solved. Separating object shape and motion robustly is an inherently ambiguous problem, since either of the two can explain changes in the observations. Many off-line approaches phrase temporal 3D reconstruction as a 4D space-time optimization problem [MFO*07, WJH*07, SAL*08, SWG08]. These approaches assume small deformations and small frame-to-frame motion [WAO*09, ZZCL13] to make the problem tractable. Recently, many template-free approaches have been introduced that exploit the data captured by commodity RGB-D sensors. Dou et al. [DFF13] reconstruct temporally coherent non-rigid motion based on a setup with multiple commodity sensors. Recently, this approach was extended to work with a single RGB-D camera [DTF*15].

The first approach that tackled the template-free reconstruction problem at real-time frame rates was the *DynamicFusion* approach by Newcombe et al. [NFS15]. This approach enables the joint reconstruction of object geometry and motion based on a single commodity depth camera, e.g. the Microsoft Kinect. For each new input frame, a mesh-based representation of the canonical model is first extracted from the underlying volumetric TSDF. Afterwards, a model-to-frame N-ICP approach is used to estimate a coarse warp field based on a deformation graph [SSP07]. Based on the estimated warp field, the voxels of the volumetric TSDF can be non-rigidly transformed to the space of the input depth map, which enables the update of the TSDF based on volumetric fusion [CL96, IKH*11, NDI*11]. The reconstructions are of higher quality than each single depth frame alone due to the integration of multiple surface samples and the canonical model can be completed gradually if previously unobserved parts of the object become visible for the first time. An extension of DynamicFusion [NFS15], called VolumeDeform [IZN*16] (see Fig. 7), parameterizes the warp field based on a fine-scale deformation lattice instead of a coarse-scale deformation graph allowing for higher reconstruction quality. To achieve real-time frame rates, this approach employs a hierarchical coarse-to-fine data-parallel GPU optimization strategy. In addition, more robust tracking is achieved by the integration of sparse feature matches into the alignment objective. The approach of Zhang et al. [ZX18] combines the reconstruction of static and dynamic scene components based on a Sigmoid-based Iterative Closest Point method that decouples camera from scene motion. The input sequence is segmented into static and dynamic parts which are separately reconstructed at real-time frame rates. In the following, we describe specific extensions of these baseline approaches that improve robustness and reconstruction quality.

Robustness to Tracking Failure The recently proposed *Fusion4D* approach of Dou et al. [DKD*16a], which is also the basis of the impressive Holoportation [OERF*16] system, obtains complete, temporally coherent models of a deforming scene at real-time frame rates. It is based on a complex multi-view setup that consists of 8 pots, each of which consists of 2 infrared (IR) and 1 color camera. In addition, a diffractive optical element and a laser is used to produce a pseudo-random pattern in IR. Depth is computed based on stereo matching in the IR domain, while the projected pseudo-random pattern guarantees the availability of texture. The acquired depth maps

Figure 8: *The Fusion4D approach of Dou et al. [DKD*16a], which is the basis of the impressive Holoportation [OERF*16] system, allows for topological changes in the scene by periodic resets of the reference volume. Image taken from [DKD*16a].*

are fused in a reconstruction volume. What makes this approach special is the employed key volume strategy, which makes the approach robust to tracking failures. Instead of fixing the reference volume to the first input frame, the approach periodically resets the reference to a fused local data volume, called a key volume. In addition, the approach detects tracking failures and then automatically refreshes all misaligned voxels based on the input data. This allows the approach to maintain a high quality reconstruction, even in challenging situations that bring the non-rigid tracker to its limits. The periodic resets of the reference volume also enable the reconstruction of topological changes in the scene, see Fig. 8. The downside of this approach is that global tracking information is lost, which might be required for certain types of applications, such as a temporally coherent re-texturing of the scene. The approach of Dou et al. [DDF*17] enables high speed reconstruction of arbitrary non-rigid scenes. One key ingredient is a dense 3D correspondence field between the input and the reconstruction that is estimated using a learned approximation of spectral embeddings. This enables to robustly handle fast scene motion. In addition, backward and forward alignment is employed for better handling topology changes and a detail layer is used to recover fine scale details, which would otherwise be lost.

Reflectance Constancy The approach of Guo et al. [GXY*17] employs a data term that is based on dense reflectance constancy instead of the color constancy assumption that is used in the competing approaches, see Fig. 9. Dense reflectance constancy better handles illumination changes and leads to more robust tracking under large rigid and non-rigid motion. The implementation of reflectance constancy requires material and lighting estimation, see Sec. 4 for a detailed coverage of this topic.

Fast Motion and Topological Changes The recently proposed *KillingFusion* approach of Slavcheva et al. [SBCI17] tackles the problem of very fast motion and topological changes via level set evolution, see Fig. 10. While most other approaches use variants of the N-ICP algorithm for tracking, which requires the extraction of a

Figure 9: *The approach of Guo et al. [GXY*17] employs dense reflectance constancy. This better handles changing illumination and leads to robust tracking even under large deformations. Image taken from [GXY*17].*

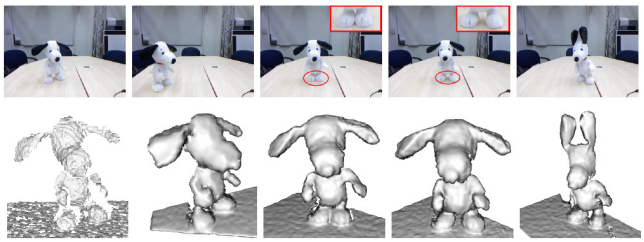


Figure 10: *The KillingFusion approach of Slavcheva et al. [SBCI17] tackles the problem of very fast motion and topological changes via level set evolution. The used regularization energy is based on a damped version of approximately Killing Vector Fields. Image taken from [SBCI17].*

mesh-based representation in each iteration step, KillingFusion directly aligns two TSDFs that encode the input and the current model. This supersedes the surface extraction step used by previous approaches and alleviates the need for explicit correspondence search. TSDF alignment is based on a damped version of approximately Killing Vector Fields [BBSG10, SBCBG11, TSB16] to warp the input volume to the current reconstruction. After each iteration step, the computed delta transformation is applied to the volume before it is volumetrically resampled. This step allows the robust handling of topological changes. The approach of Dou et al. [DKD*16a] handles topological changes based on a key frame strategy. The recently proposed *BodyFusion* approach [YGX*17] uses an articulated skeleton to define the warp field instead of a coarse general deformation graph. Since the skeleton parameterization is low dimensional, the tracking problem is drastically simplified and the approach produces more stable reconstructions for the special case of humans.

3.6. Dynamic Scene Datasets

Whereas there are many datasets for the evaluation of static 3D reconstruction based on commodity RGB-D sensors (see Sec. 2.7), only a few exist for commodity non-rigid surface tracking and reconstruction. Currently, most publicly available real world datasets captured using an RGB-D sensor do not provide geometric ground truth. For template-based tracking Guo et al. [GXW*15, GXW*17] provide several sequences used in their publication. For template-free reconstruction Innmann et al. [IZN*16] and Dou et al. [DTF*15] provide several sequences used in their publications. Quantitative comparisons to other approaches are often performed on synthetic RGB-D data streams. For this purpose the MIT dataset [VBMP08] is often used. It contains several complex and large human motion sequences that have been reconstructed using a multi-view capture system. The multi-view image input, camera calibration, and the 3D reconstructions are provided. This enables the creation of hybrid real/synthetic RGB-D streams that provide real color and synthetically rendered depth (optionally with simulated sensor noise) based on the multi-view reconstructions. The multi-view 3D reconstructions can then serve as ground truth for an evaluation of the dense surface reconstruction error. Another high-quality multi-view dataset that is often employed for quantitative evaluation is from Collet et al. [CCS*15]. Considering the recent advances regarding range sensor technology and the success of dynamic real-time 3D reconstruction systems, the lack of a large dynamic real-world RGB-D benchmark with available ground truth leaves sufficient room for further developments.

4. Color and Appearance

Besides the object and scene geometry that is of interest for many applications, surface colors and general appearance information play an important role for various virtual (VR) and augmented reality (AR) applications, and enable users to interact with virtual models in a similar way as with the real world. Obtaining such intrinsic scene properties from captured image data is highly challenging and, although the problem seems fairly related to static (see Sec. 2) and dynamic (see Sec. 3) 3D reconstruction, it is a rather orthogonal field of research. Whereas most color texture estimation approaches are tightly coupled to static reconstruction, material acquisition techniques based on RGB or RGB-D data mostly focus on objects and small scenes and solve the problem in image space. However, recent works in static [MFZ*17, WWZ16] and dynamic [GXY*17] reconstruction start to connect the fields by jointly solving for a full virtual model with the desired shape and appearance information. An overview of state-of-the-art appearance reconstruction approaches that make use of RGB-D information is provided in Tab. 3. In the following, we discuss the challenging problems related to color and material acquisition and the impressive solutions researchers have developed.

4.1. Color Textures

Reconstructing pure color information from a sequence of RGB images has been a challenging task for several years. After the recent success of state-of-the-art volumetric fusion approaches, a wealth of subsequent work tried to overcome the limitations of this system and extended it in several ways. One particular line of research has been the reconstruction of consistent color textures.

Figure 11: *Color Accumulation and Offline Texture Generation: comparing volumetric blending [NZIS13] (left) with color map optimization [ZK14] (middle), and texture generation [JJKL16] (right). Image taken from [JJKL16].*

4.1.1. Online Texture Reconstruction

In some work, extensions to the fusion process of KinectFusion have been proposed to also cover colors. Whelan et al. [WKJ*15] reject samples at object boundaries and grazing angles that would potentially lead to artifacts and inconsistent results. In subsequent work [WSMG*16], they estimate the positions and directions of light sources in the scene to further reject samples from fusion that only contain specular highlights. While this already improves the visual quality of the textures significantly, large artifacts may still arise. Most RGB-D cameras acquire color images by adjusting the exposure time dynamically depending on the currently visible illumination. This ensures that the dynamic range of the current view is faithfully mapped and quantized to the camera's brightness range. Simply fixing the exposure time is not only sometimes prohibited by the camera firmware or unsupported by the drivers, it can also quickly lead to over- or under-saturated pixels and regions in the images, since the whole possible dynamic range is far too large to represent it accurately with only 8-bit per color channel. Thus, researchers developed techniques to capture high dynamic range (HDR) color textures from low dynamic range (LDR) images with varying exposure times. In a pre-processing step, the camera-specific response curve is usually estimated to linearize the observed intensity values. Afterwards, the relative exposure time is estimated between subsequent frames to obtain HDR colors [MBC13, LHZC16, APZV17]. Finally, those values are fused together into either key frames [MBC13] or a global virtual model [LHZC16, APZV17] in real-time.

4.1.2. Offline Texture Reconstruction

In the offline context, similar ideas have been used to obtain globally consistent textures. One of the first approaches was proposed by Zhou et al. [ZK14] in which camera poses and colors of selected key frames are jointly optimized by maximizing photometric consistency (c.f. Fig. 11, middle). Narayan et al. [NA15] use a similar technique, but only consider a pixel-dependent subset of the key frames for the optimization and add an additional smoothness term along edges. This improves the visual quality of the results and especially reduces color bleeding. Substantial improvements have been achieved by Huang et al. [HDGN17] who also built upon the work of Zhou et al. [ZK14]. After a primitive abstraction of the scene has been computed, the color values are first corrected by compensating the varying exposure and white-balance and then aligned by optimizing an energy based on dense photometric, sparse feature, and primitive relationship constraints. Finally, a consistent texture is obtained using a temporally coherent sharpening operation. Zhang et al. [ZCC16] also apply exposure compensation techniques to reconstruct consistent HDR color textures and demonstrate various editing applications. Compelling results have been recently achieved by Maier et al. [MKC*17] who jointly refine camera poses, the geometry and reflectance of the model stored in a truncated signed distance field (TSDF) and optimize for the intrinsics camera and distortion parameters. Some researchers obtain a coarse 3D model of the scene via current reconstruction techniques (see Sec. 2) in a first step and then use it as a global reference to select and fuse key-frames into a refined model of much higher quality by imposing photometric consistency [MSC15, JJKL16, RTKPS16] (c.f. Fig. 11, right). Very recently, Bi et al. [BKR17] proposed a patch-based optimization approach to produce high-quality texture maps for scanned objects.

4.2. Material Acquisition

While reconstructing consistent color textures is itself a challenging task, it can still not fully explain the visual appearance of objects. In general, the observed colors in an image $\boldsymbol{I} \subset \mathbb{R}^3$ not only depend on the specific material properties but also on the surrounding scene-specific illumination. The process of reconstructing the material reflectances out of the final rendered image is called *Inverse Rendering* and is a highly ill-posed problem.

The most popular approach to tackle this problem is called *Intrinsic Image Decomposition* and was first formulated by Barrow and Tenenbaum [BT78]. Assuming that all materials predominantly are Lambertian, i.e., their appearance is independent of the view direction and only depends on the incoming light direction, the resulting image can be approximated and decomposed into two parts:

$$\boldsymbol{I}(x) = \boldsymbol{R}(x) \cdot \boldsymbol{S}(x) \quad . \tag{3}$$

Here, $\boldsymbol{I}$ is the observed image, $\boldsymbol{R}$ is the diffuse part of the surface reflectance properties, and $\boldsymbol{S}$ is the shading that depends on the surface geometry and the illumination. In addition to its simplicity, this decomposition is modeled in images space meaning that the reflectance $\boldsymbol{R}$ and the shading $\boldsymbol{S}$ are also images and the multiplication is performed component-wise per pixel x. The great advantage of this technique is that this effectively removes the burden of knowing the exact 3D geometry of the scene.

However, the problem is still highly ill-posed since for any optimal solution $(\boldsymbol{R}^*, \boldsymbol{S}^*)$ there exist infinitely many other equivalent solutions $\left(c \cdot \boldsymbol{R}^*, \frac{1}{c} \cdot \boldsymbol{S}^*\right)$ where c is a positive number. In order to avoid this scale ambiguity, further constraints are introduced and finding $\boldsymbol{R}$ and $\boldsymbol{S}$ is formulated as an energy optimization problem:

$$E_{\text{total}}(\boldsymbol{R}, \boldsymbol{S}) = E_{\text{fit}}(\boldsymbol{R}, \boldsymbol{S}) + \lambda \cdot E_{\text{reg}}(\boldsymbol{R}, \boldsymbol{S}) \quad . \tag{4}$$

The total energy E_{total} typically consists of a data fitting term E_{fit} and a regularization term E_{reg} where the parameter λ controls the relative influence between both terms. Researchers have tried many different variations and combinations of the terms to obtain plausible results. In this report, we mainly focus on recent techniques that also

Method	Shading Representation					Data Fitting			Regularizers			Properties			Speed	
	Direct Part	Indirect Part	Quadratic Function	Spherical Harmonics	Wavelets	Intr. Image Dec.	Log Intr. Image Dec.	Rendering Eq.	Sparse Reflectance	Smooth Shading	Training Data	Segmentation	Specular	IR Data	Online	Offline
Chen and Koltun [CK13]	✓	✓	-	-	-	-	✓	-	✓	✓	-	-	-	-	✓	-
Barron and Malik [BM13]	-	-	-	✓	-	-	✓	-	-	-	✓	✓	-	-	✓	-
Kerl et al. [KSSC14]	✓	-	-	-	-	✓	-	-	✓	✓	-	-	-	✓	✓	-
Hachama et al. [HGW15]	-	-	-	✓	-	✓	-	-	✓	✓	-	-	-	-	✓	-
Wu et al. [WZ15]	-	-	-	-	✓	-	-	✓	-	-	-	✓	✓	✓	-	✓
Wu et al. [WWZ16]	-	-	-	-	✓	-	-	✓	-	-	-	-	✓	-	-	✓
Richter-Trummer et al. [RTKPS16]	-	-	-	✓	-	✓	-	-	✓	-	-	✓	✓	-	-	✓
Zuo et al. [ZWZY17]	-	-	✓	-	-	✓	-	-	-	-	-	-	-	-	-	✓

Table 3: *Overview of state-of-the-art appearance reconstruction approaches that take advantage of RGB-D and IR data.*

incorporate the additionally provided depth information from RGB-D cameras. For a more detailed overview of this highly challenging field, we refer to the excellent work of Bonneel et al. [BKPB17b].

4.2.1. Data Fitting Terms

Many different data fitting terms have been proposed. All of them have in common that they guide the solution towards the defined model which is the Intrinsic Image Decomposition in our scenario. In the following, we will discuss common choices.

Intrinsic Image Decomposition The most popular fitting terms are the ones that are directly derived from the Intrinsic Image Decomposition equation. The simplest variant is a least-squares error term for the vector-valued reflectance and shading images [LBP*12, GMLMG12]. Under the assumption of white illumination, the shading image can be further constrained and simplified to be only scalar-valued which reduces the number of unknown variables [SYLJ13, KSSC14]. Both strategies require additional hard constraints to ensure that the reflectance and shading term are both non-negative. An elegant way to directly incorporate them in the energy formulation is to solve the problem in log-space which has the additional advantage of transforming the component-wise product into a less involving sum and allowing for more efficient optimization techniques. This has been applied for either the scalar [MZRT16, MFZ*17] and the vector-valued versions of the least-squares error terms [SYJL11, LZT*12, ZTD*12, BM13]. Recently, Bonneel et al. [BST*14] propose to optimize for the gradient of the log-reflectance and shading which is another convenient reformulation of the original problem. In addition, weighting the contribution of each sample adaptively based on the brightness of the observed color values further improves the robustness and reduces the influence of dark regions with less reliable color values [KGB14, MZRT16, MFZ*17, CK13].

Patch-based Optimization Some approaches use patches to reduce the number of unknown variables and directly incorporate a smoothness constraint on the shading layer. Shen et al. [STL08] group pixels with similar textures locally and non-locally and solve for the in-group reflectance intensities. Garces et al. [GMLMG12] cluster pixels based on their chromaticity and then solve for the shading image. In the context of video streams from a static scene and camera, Laffont et al. [LB15] directly incorporate time coherence in the fitting term by only allowing patches of the shading image to change over time.

Statistic-based Approaches Similar to the patch-based techniques, researchers tried to create statistics over the shape and illumination on the input image itself [TFA05, TAF06], to pre-capture training data [BM15], or use mixtures of shapes and illumination [BM13]. This additional information can be leveraged to further constrain the solution to obey the structure of the observed statistics.

Shading Decomposition Much work has been spent on further decomposing the shading layer. Chen et al. [CK13] model the vector-valued shading image by a scalar-valued direct and indirect irradiance layer, and a vector-valued light color layer. This allows to define smoothness priors per layer rather than for the whole shading image and to control the influence of each term to the whole energy independently. For outdoor scenarios, Laffont et al. [LBD13] apply a similar idea by using two layers for the sun and the sky and one for the indirect irradiance. The most prominent approaches represent shading in the spherical harmonics basis and typically consider them up to the second order to enforce global smoothness and allow for efficient optimization. While the dimensionality of the unknowns can be reduced to a constant number of lighting coefficients, the shading variation is now encoded in the basis functions that require the knowledge of surface normals. For RGB-D cameras, normals can be estimated conveniently from the depth image to provide the additionally needed information. In the context of dynamic scene reconstruction, Guo et al. [GXY*17] jointly optimize for the lighting coefficients and the observed motion between subsequent frames to improve the robustness of the motion estimation (c.f. Fig.12). In other work, per-vertex coefficients have been considered to account

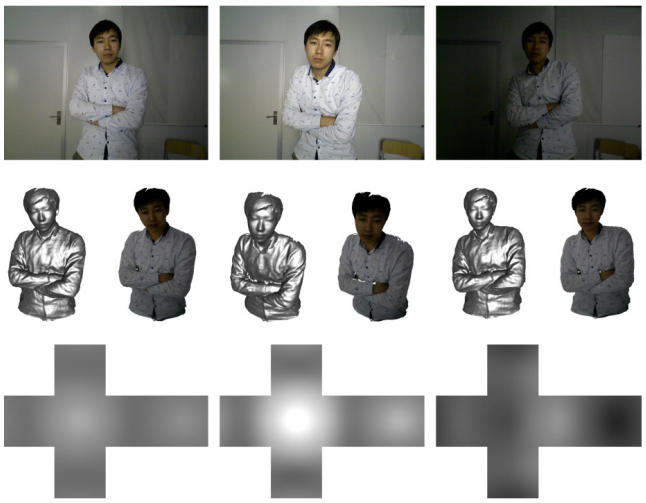


Figure 12: *The approach of Guo et al. [GXY*17] performs real-time estimation of the geometry, motion, albedo and illumination [GXY*17] of a general non-rigidly deforming scene. Image taken from [GXY*17].*

for potentially spatially varying illumination [HGW15]. Recently, shading has been also modeled by quadratic functions of the normals, which are similar to second spherical harmonics [ZWZY17].

4.2.2. Regularization Terms

Considering regularization priors, the most popular choices are based on the *Retinex theory* [LM71] stating that reflectance is in general sparse and shading is smooth. Therefore, researchers mostly rely on at least one of these observations; however, many other sources of information have been considered to further advance this field.

Reflectance Sparsity As typically only a rather small number of different materials are observed, sparsity can be enforced by penalizing reflectance variations across neighboring pixels. In case the material changes between two pixels – i.e., the image gradient is large –, the contribution of this term would be very large; hence, reflectance edges would be smoothed out to reach the minimum of the energy. Thus, each term is weighted based on a thresholded gradient image [LM71, LZT*12, ZTD*12] to allow sharp edges in the reflectance image. Instead of considering the gradient magnitude, chromaticity-based weights have been used to robustly detect reflectance edges [SYH13, ZDI*15, MZRT16, MFZ*17]. Shen et al. [SYLJ13, SYJL11] extend the original Retinex weights to also consider the observed intensity values and relax the constraint at dark regions where reflectance can be less reliably estimated and differences might not be detected accurately. A combination of both weighting strategies has been used by Chen et al. [CK13] to lower the regularization contribution in those problematic regions and obtain sharp edges. Recently, Kerl et al. [KSSC14] use the additionally provided infrared information of time-of-flight sensors to apply weights based on the estimated infrared reflectance. Compared to the typical choice of using the squared Euclidean norm of reflectance differences, other norms have been used as well. Several researchers [SYH13, KSSC14, HGW15] propose to use the more robust ℓ_1 or total variation norm which has been successfully applied in other fields such as image denoising. The more general ℓ_p-norm has also been used where $p < 1$ is a common choice to further increase the flattening effect [BST*14, MZRT16, MFZ*17]. Recently, efficient implementations of iteratively reweighted least squares (IRLS) solvers for finding optima of ℓ_p-norm constraints enabled interactive applications as the intrinsic decomposition problem could be solved in real-time [TZS*16, MZRT16, MFZ*17]. In some work, the effect of the sparsity has been extended by adding a non-local term or applying multi-resolution techniques to smooth across a wider range of the image [LZT*12, SYH13]. Li et al. [LB14] even consider probabilistic approaches to enforce sparsity in the reflectance and smoothness in the shading layer.

Shading Smoothness While reflectance changes mostly across materials, shading depends on the surface geometry and the illumination. Both of them are usually smooth leading to the observation that shading also changes slowly and smoothly. Thus, differences between neighboring values are penalized in the least-squares sense [ZTD*12, CK13, BST*14, HGW15, LB15]. Similar to reflectance sparsity, many approaches add an additional weighting to this term to allow for a more fine-grained control of the prior. Meka et al. [MZRT16, MFZ*17] reuse their reflectance weights but invert their contribution to further strengthen the idea of reflectance edges. Lee et al. [LZT*12] add a non-local term and used normal information to threshold the weights in a similar manner as for the reflectance layer. Bonneel et al. [BST*14] consider chromaticity as an approximation of reflectance and applied thresholding to the shading gradients based on them. Other norms have been also applied such as the weighted ℓ_1-norm considering the estimated shading in the infrared channel [KSSC14], and the robust Tukey function in combination with weights based on the estimated optical flow between images [KGB14]. Shen et al. [SYH13] obtain second order shading smoothness by imposing a prior using the Laplace operator.

Chromaticity Prior Although enforcing sparsity and smoothness in the respective layers already leads to plausible results, effects like indirect illumination can still lead to large deviations from the desired results, especially in darker regions. Therefore, Meka et al. [MZRT16, MFZ*17] add an additional prior that forces the chromaticity of the reflectance values in bright regions to be close to the ones of the observed image. In combination with smoothness constraints, this further improves the accuracy.

Reflectance Clustering Prior Another popular strategy to enforce reflectance sparsity is to perform clustering either via a soft or a hard constraint. Bi et al. [BHY15] approximated reflectance via image flattening techniques and clustered similar values together to obtain a sparse set of material labels. Afterwards, the intrinsic image decomposition is solved based on the labeling serving as a hard constraint. Other approaches also perform clustering but employ a soft constraint to enforce non-local sparsity either solely for an image [ZTD*12, MZRT16] or a whole video sequence [BST*14].

Reflectance Ratio Prior Laffont et al. [LBP*12] propose the idea of reflectance ratios. The key observation is that intensity variations can only be explained by different reflectance values in case their shading values are identical. Thus, the ratio between reflectance values can be approximated by the ratio of intensities. This has been applied to estimate the reflectance of a collection of images. After pixel correspondences have been computed to obtain robust estimate of the ratios, pairwise priors are defined.

IR Reflectance Coupling Whereas most approaches solely focus on RGB-D images or videos, Kerl et al. [KSSC14] experimented with time-of-flight cameras that are capable of providing an additional smooth infrared image. The illumination conditions in this channel are much more controlled as the sensor itself is the only light source and ambient radiation is negligible. They proposed to first estimate the infrared reflectance and then directly couple it with its color version. This indirectly enforces temporal consistency as the exposure time for infrared images is kept fixed by the camera.

Temporal Consistency Prior Recently, researchers not only tried to acquire material properties from a single image but also from a whole sequence of images – i.e., a video. Different strategies have been developed starting from enforcing consistency between subsequent images [MZRT16, BST*14, LBP*12, KGB14], propagating parts of the previous solution to the current frame [YGL*14], or even adding constraints between frames and a global model [MFZ*17].

User constraints Another valuable source of information are priors directly provided by the user. Compared to other priors that are typically inspired by hysterics and common use cases, these inputs can be considered as ground truth information since humans are very good at predicting material properties. In case of video sequences, they are also typically propagated between frames over time assuming that the optical flow is known. Popular choices of user inputs are strokes that enforce constant shading or reflectance [BST*14, MFZ*17] locally in some regions. Furthermore, the user can directly resolve the scale ambiguity by a fixed-illumination brush that determines the absolute quantity of a set of shading values [BPD09].

4.2.3. SV-BRDF Acquisition

The fundamental assumption in all material acquisition approaches discussed so far is that the diffuse component of the material properties is the most relevant one and dominates the specular component in all common scenarios. While for most real-world materials this approximation holds, there are still many others for which this immediately breaks, e.g., metal and all kinds of polished materials. For this group of objects, strong highlights and reflections are observed in the images leading to completely wrong estimates in those regions. Therefore, more expressive models, in particular *spatially-varying Bidirectional Reflectance Distribution Functions (SV-BRDFs)*, have been used to handle these cases.

One of the techniques for estimating SV-BRDFs from RGB-D data has been proposed by Knecht et al. [KTTW12]. They consider RGB-D and additional environment maps captured from a fish-eye lens camera to remove highlights from color observations. Using this cleaned data, a per-pixel diffuse SV-BRDF and per-cluster specular SV-BRDF is estimated interactively. The first work trying to combine SV-BRDF estimation with 3D reconstruction from RGB-D data has been done by Wu and Zhou [WZ15]. After reconstructing the object shape using KinectFusion in a first pass, its appearance is estimated afterwards. Using a mirror ball to acquire the illumination via an environment map and gray markers to photometrically calibrate the captured RGB image, they are able to interactively reconstruct an accurate SV-BRDF of an object with a Microsoft Kinect sensor. They group pixels with the same materials together and used the Ward model [War92] to estimate the specular part of the SV-BRDF per cluster. Using the infrared channel of the Kinect, the specular parameters of the Ward model are estimated by fusing all clustered sampled together similar to the volumetric fusion technique proposed by Curless and Levoy [CL96]. In subsequent work, Wu et al. [WWZ16] change the philosophy of their design from an interactive system to a more accurate offline approach. By jointly optimizing for the camera poses, material clusters, environment lighting, and spatially-varying BRDFs, they are able to obtain results of much higher quality than before. However, this approach is not able to run close to real-time anymore and requires the whole video sequence to be known beforehand. Another offline technique that jointly solves for lighting, shape, and reflectance has been developed by Lombardi and Nishino [LN16]. Richter-Trummer et al. [RTKPS16] propose an off-line system that first estimates a consistent color texture of the reconstructed model and then segments it into patches of similar material. Based on these segments, low-frequency environment lighting and per-vertex diffuse and specular reflectance components are estimated. For a detailed overview of the field of material acquisition from image data, we refer to the excellent work of Weinmann et al. [WLGK16].

4.3. Geometry Refinement and Normals

The main observation when capturing RGB-D images is that for most cameras the depth image is noisy and of rather low quality whereas the RGB image shows many fine details and exhibits only few noise. Therefore, much effort has been spent to raise the quality of the depth images to the level of the RGB images to improve the accuracy of reconstruction algorithms. Seminal work in this field was presented by Horn et al. [HB86] who introduced the concept of Shape-from-Shading (SfS) to estimate the geometry of objects from a single image. By recovering the shading, normals and the shape can be inferred. Therefore, the discussed techniques are closely related to the field of material acquisition and several ideas can be applied in both areas.

4.3.1. Natural Illumination

In several approaches, images taken from real-world scenarios with entirely uncontrolled natural illumination have been considered as the primary source of interest. Haque et al. [HCMG*14] assume Lambertian surfaces and add a first order normal smoothness and a second order Laplacian depth smoothness prior to obtain plausible results. This is closely related to the shading smoothness prior in the field of material acquisition where shading variations are assumed to be slow and smooth similar to the object's shape. In the context of refinement, the shape is constrained directly rather than

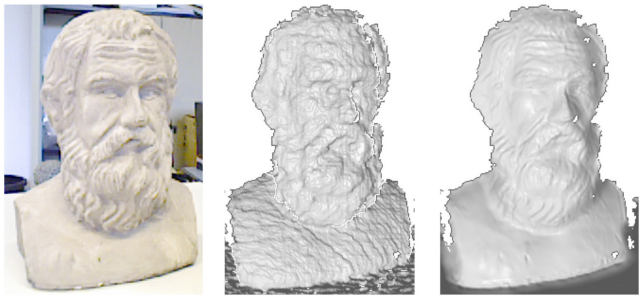


Figure 13: *The approach of Wu et al. [WZN*14] performs online shading-based surface refinement of a live depth stream captured by a commodity RGB-D sensor. Image taken from [WZN*14].*

the dependent shading. Quadratic functions have also been applied to parameterize the shading with respect to normals [HLSK13]. Zhang et al. [ZYY*12] use a multiple light source setup and optimize for refined depth and normals jointly using an adaptive visibility-based weighting scheme. Other low-rank techniques have been successfully applied to enforce shading smoothness. This includes the rank 3 brightness matrix approximation proposed by Chatterjee et al. [CMG15] and patch-based approaches that employ low-rank subspace constraints for each patch [LRL14]. Both approaches consider matrix factorization algorithms to finally obtain a refined depth image. Similar to the reflectance ratio prior in the context of material acquisition, Yu et al. [YYTL13] consider ratios computed per cluster to add further regularization priors. Recently, Zollhöfer et al. [ZDI*15] successfully refine the geometry of reconstructed global models encoded in a truncated signed distance field to obtain a refined version of much higher quality. However, their approach requires the knowledge of accurate camera poses. Maier et al. [MKC*17] address these shortcomings by jointly refining the camera poses and the surface geometry; however their method is limited to the offline setting.

Much effort has been spent in the recent years to accelerate refinement algorithms to allow real-time computations. One of the first approaches that reached this highly desirable goal was proposed by Wu et al. [WZN*14] (c.f. Fig. 13). They estimate second order spherical harmonics coefficients in a least-squares sense and then refine the depth image via a highly optimized solver on the GPU. Besides a depth fidelity term, they also enforce smoothness using a second order Laplacian constraint on the vertices and a temporal prior to the refined depth image frame the previous frame. Based on this work, Or-El et al. [OERW*15] use an extended shading model that also covers sparse specular shading and inter-reflections to increase the accuracy and robustness in problematic regions.

4.3.2. IR Illumination

Besides natural illumination, infrared information gained increasing interest due to the success of RGB-D sensors, in particular time-of-flight cameras. The acquired infrared images are comparable to their RGB variants in terms of quality and noise but the illumination conditions are typically more controlled as the camera itself emits light into the scene. Therefore, reflectance and shading information can be reconstructed more robustly to guide the refinement process. Choe et al. [CPTSK14, CPTK17] model shading with a Lambertian term and an image-wide ambient term that covers indirect light to refine the vertex positions obtained from the depth image via displacement vectors in real-time. Recently, Or-El et al. [OEHW*16] fit the Phong reflection model [Pho75] using efficient total variation techniques to also cover specularities that would otherwise negatively affect the refinement process and lead to incorrect results in those regions.

4.3.3. Normals

Also closely related to depth refinement is the field of normal estimation from images. Since many refinement algorithms directly or indirectly estimate normal information, approaches that solely estimate normals can also be considered for refinement. In addition to the already discussed algorithms, especially learning-based techniques have been used to tackle this challenging problem. In recent work, Richter et al. [RR15] apply regression forests to predict normals from a single RGB image without any assumptions about the lighting conditions that are typically exploited in Shape-from-Shading. Similar work has been done by Yoon et al. [YCK*16] who train a convolutional neural network (CNN) on uncalibrated infrared data to obtain accurate results.

4.4. Material Datasets

Whereas there are several datasets for the evaluation of static 3D reconstruction (see Sec. 2.7), only a few exist for material acquisition, especially for intrinsic image decomposition. For BRDF acquisition, widely used benchmarks are the MERL database [Mat03], which consists of over 100 measured models, the CUReT database [DVGNK99], which also contains over 60 measured Bidirectional Texture Functions (BTFs), and the KTH-TIPS database [CHM05, HCFE04], which extends the CUReT database. Weinmann et al. [WGK14] synthesized a BTF database of 84 materials measured using 22801 view-light configurations in total. The dataset also contains the corresponding surface geometry. In the field of intrinsic image decomposition, the "MIT Intrinsic Images" dataset [GJAF09] provides a database of object appearances that are composed of diffuse shading, reflectance, and specularity layers. Croudsourcing has been used in some data sets to annotate pixels with similar reflectance or shading for thousands of images [BBS14, KBSB17]. The MPI Sintel dataset [BWSB12] provides a set of computer-generated images that have similar statistics to real-world images. Ye et al. [YGL*14] created synthetic image data by rendering 3D models with a constant and diffuse shader. However, most of these datasets only contain RGB and lack depth information, making them only suitable for a subset of approaches that do not require additional range information. Some researchers tried to generate pseudo-synthetic RGB-D benchmark data by extending the "MIT Intrinsic Images" dataset [GJAF09] with depth images that were produced by the approach of Barron and Malik [BM12]. Considering the recent advances regarding range sensor technology and the success of real-time 3D reconstruction systems, the lack of RGB-D, and possibly RGB-D including IR, appearance benchmark datasets leaves sufficient room for further developments.

5. Energy Optimization

Many of the so-far discussed techniques solve the respective problem by optimizing an energy functional. Commonly, more complex optimization problems use energy functionals that consist of two parts, data fitting and regularization terms. Often the energy terms are modeled based on a non-linear least-squares objective. This part is handled rather similarly by different optimization approaches. Many other objective functions, based for example on total variation or robust kernels, heavily influence the choice of the used solver, meaning that each solver can only handle one specific category. In the following, we will discuss the most common strategies.

5.1. Non-Linear Least Squares Optimization

Many of the used data terms and regularizers are non-linear least squares optimization problems. An optimization problem in the unknowns $\boldsymbol{x} \in \mathbb{R}^N$ is a non-linear least squares problem if it has the following canonical form:

$$E(\boldsymbol{x}) = \sum_{i=1}^{M} \left[\mathbf{r}_i(\boldsymbol{x})\right]^2 = ||F(\boldsymbol{x})||_2^2 \quad . \tag{5}$$

The vector field $F : \mathbb{R}^N \rightarrow \mathbb{R}^M$ stacks the M, potenitally non-linear, residuals $\mathbf{r}_i$ in a vector. If the vector field F is linear in the unknowns $\boldsymbol{x}$, the problem reduces to a linear least squares optimization problem:

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} ||F(\boldsymbol{x})||_2^2 = \arg\min_{\boldsymbol{x}} ||\mathbf{A}\boldsymbol{x} + \mathbf{b}||_2^2 \quad . \tag{6}$$

The minimizer of such an objective function is given by the well-known normal equations $\mathbf{A}^T\mathbf{A}\boldsymbol{x} = -\mathbf{A}^T\mathbf{b}$. The resulting system of linear equations can be solved using standard iterative solvers, e.g., Gradient Descent or Preconditioned Conjugate Gradient Descent, or direct linear solvers, based on LU or Cholesky Decomposition.

Many online approaches favor iterative solution strategies, since these can be easily parallelized on modern graphics hardware [ZSGS12, ZNI*14, DKD*16a, GXY*17, MZRT16, MFZ*17].

If F is a non-linear function in the unknowns $\boldsymbol{x}$, non-linear solvers such as Gauss-Newton have to be employed. These solvers reduce the solution of the non-linear least squares optimization problem to a sequence of linear least squares problems based on a Taylor expansion of the vector field F around the last solution $\boldsymbol{x}_i$:

$$F(\boldsymbol{x}_{i+1}) \approx F(\boldsymbol{x}_i) + \mathbf{J}(\boldsymbol{x}_i)\delta, \quad \delta = \boldsymbol{x}_{i+1} - \boldsymbol{x}_i \quad . \tag{7}$$

Therefore, these approaches solve a linear least squares problem in each iteration step $\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \delta^*$ to obtain a new solution $\boldsymbol{x}_{i+1}$:

$$\delta^* = \arg\min_{\delta} ||F(\boldsymbol{x}_i) + \mathbf{J}(\boldsymbol{x}_i)\delta||_2^2 \quad . \tag{8}$$

The resulting system of equations is linear and can be solved using the normal equations. Levenberg-Marquardt is an extension of Gauss-Newton that adaptively blends between Gauss-Newton and Gradient Descent to achieve robust convergence.

5.2. Total Variation

Another increasingly popular strategy to solve and regularize highly challenging ill-posed problems is convex optimization. This particular class of energy functions has the following form:

$$E_{\mathrm{d}}(\boldsymbol{x}) + \lambda \cdot |K(\boldsymbol{x})| \quad . \tag{9}$$

Here, $\boldsymbol{x}$ is an element of a finite-dimensional vector space X, E_{d} is a convex function and represents the data term of the energy, and $|K(\boldsymbol{x})|$ is another convex function mapping from the vector space X to another one Y. The most common choice for the regularization term is the ℓ_1-norm of the gradient function, which is also called total variation norm. One of the first applications was image denoising where this concept has been successfully applied first by Rudin et al. [ROF92].

A common way to solve such a problem is to reformulate it from a minimization to a maximization task. The primal variable $\boldsymbol{x}$ is then replaced by a dual variable $\boldsymbol{y}$. This is often called the *dual problem* [ET99]. A mixture between both formulations also exists, i.e. the called *primal-dual problem*. Here, an auxiliary variable is introduced which is optimized in conjunction with the primal variable by finding a saddle point of the reformulated energy function. During the last two decades, several algorithms have been proposed to tackle the problem in one of its formulations including the solver by Chambolle and Pock [CP11], Split-Bregman methods [GO09], alternating directions of multipliers (ADMM) [Ess09, LM79], Newton-based solvers [CGM99], and many others. Recently, the concept of total variation has also been extended to enforce higher order smoothness [BKP10].

5.3. Robust Optimization

Robust optimization is often used for filtering out bad correspondences in template-based non-rigid registration approaches [LSP08, ZNI*14] and bundle adjustment [Zac14b, CZK15a]. The idea is to employ a robust kernel function instead of an ℓ_2-norm to define the objective function:

$$E(\boldsymbol{x}) = \sum_{i=1}^{M} \Psi(\mathbf{r}_i(\boldsymbol{x})) \quad . \tag{10}$$

Here, Ψ is the employed robust kernel. Due to the robust kernel, the resulting optimization problem is non-linear and not in general least-squares form. With some modifications, it is still possible to cast the optimization problem as a non-linear least-squares problem, which enables the use of the previously discussed standard solvers. There are many different possibilities how this can be achieved. A comprehensive summary and evaluation of the different approaches, which includes Iteratively Reweighted Least Squares, the Triggs correction, Square-rooting, and Lifting, has been performed by Zach et al. [Zac14b]. We also provide an explanation on how IRLS can deal with ℓ_p-norms in the next section.

5.4. Iteratively Reweighted Least Squares

Optimization problems that involve general ℓ_p-norms can be reduced to a sequence of, potentially non-linear, least squares optimization problems based on *Iteratively Reweighted Least Squares* (IRLS). This strategy has been applied to both tracking as well as intrinsic decomposition problems [TZS*16, MZRT16, MFZ*17, BST*14]. The key idea of IRLS is to split the residuals, which

involve the ℓ_p-norm, into two individual parts:

$$\|\mathbf{r}\|_2^p = \underbrace{\|\mathbf{r}\|_2^{p-2}}_{\text{const.}} \cdot \|\mathbf{r}\|_2^2 \approx c \cdot \|\mathbf{r}\|_2^2 \quad . \tag{11}$$

The first part is considered constant during each iteration step and the second part is in general least-squares form. Thus, in each iteration step, this approximation can be solved using a standard (non-linear) least-squares solver.

6. Challenges and Future Work

This report summarizes the tremendous amount of scientific progress which was spawned over the recent years by the wide accessibility of commodity RGB-D sensors. Despite this remarkable progress, there are still many difficult open challenges that future work on 3D reconstruction with RGB-D cameras needs to address, in particular for interactive and online use cases.

Geometric Aspects of 3D Reconstruction Even though the core system setups described in Sec. 2 are very mature, they are still far from being ideal. In generic scene reconstruction situations w/o prior knowledge, more efficient methods for geometry cleanup, simplification and abstraction are needed. In our report, we deliberately did not include post-processing approaches, as they are rather unattractive in the 3D reconstruction context using RGB-D cameras. Thus, we see a significant potential and need for further research, since more elaborate geometric abstraction would intrinsically support better data compression, scene completion, and removal of spurious geometry. This would help to compensate errors due to sensor noise, limited resolution, and misalignments due to drift. Resulting compression would also support scalability to larger scenes, in particular with online methods. Geometric abstraction, e.g., based on shape primitives, has been extensively researched for the purpose of static 3D model and scan simplification. Extending these concepts to the wider context of unrestricted, continuous and real-time 3D reconstruction with RGB-D cameras is not straightforward. The latter setting is much more complicated as scene dynamics can enforce to roll-back prior decisions made to specific geometric constellations, e.g., a plane needs to get back-converted into voxels/points in case of changing shape. Handling this efficiently on the level of shape abstraction is still a major challenge.

Capturing and Modeling Dynamics The reconstruction of scene dynamics has seen tremendous progress in recent past, as explained in Sec. 3. Nonetheless, existing approaches are still in their infancy and can only handle very restricted types of scenes. Independent of the specific application, robust handling of fast motions is still infeasible in general environments. In addition, approaches are not designed to handle difficult occlusions or self-occlusions, even lose interactions between multiple elements in the scene is often highly impractical. Most of them are geared towards reconstruction of individual or a low number of deforming objects. Even most offline methods break under difficult deformations and apparent topology changes. Handling such cases in real-time is a challenge of even larger scale.

Many approaches resort to some form of shape or deformation template, e.g., a skeleton, a piecewise rigid shape model, or a deformable surface, to handle dynamic shape capture. Capturing or a priori designing such a template is a challenge in its own right. It is also a limitation in practice, since template initialization for all conceivable deformable real world objects is hardly achievable. Some approaches we discussed started to look into simultaneous template building and deformable tracking. However, they only succeed on very simple and slowly moving and deforming shapes. Space-time coherent reconstruction of general deformable scenes from multiple RGB-D cameras, let alone a single RGB-B camera, is therefore still a widely open problem. One strategy that could significantly improve the capture and modeling of dynamics is therefore the active learning of more expressive adaptable deformation models. New strategies to capture dynamic scenes at larger temporal scales and finer-scale spatial detail than currently possible will also be needed in the future.

Appearance As depicted in Sec. 4, there have been groundbreaking developments in online and interactive capture and representation of 3D appearance, from simple Lambertian color and textures, via (spatially varying) BRDFs, and other scattering models to illumination estimation. Still, due to the entanglement of all these effects in real-world appearance acquisition systems, the underlying problem is highly ill-posed and it is hard to deduce more generic solutions. Even though, there are first online results, they capture rather simplistic approximations of real-world appearance and light transport complexity. The joint online estimation of high-quality material and illumination properties leaves sufficient room for improvements, involving, for instance, more appropriate reflectance, scattering, and more detailed, high-frequency illumination models.

High-Level Reconstruction From a more abstract perspective, the existing online and interactive 3D reconstruction techniques have the potential to enable a large variety of future high-level applications. This, however, requires substantial breakthroughs regarding the following three aspects. (1) More *generic scene reconstruction approaches* need to be developed that can cover a much wider scope of application scenarios in a single approach. (2) The upcoming of *mobile devices and smartphones* equipped with RGB-D sensors do not automatically induce that the existing 3D reconstruction solutions are available on these platforms. This is mainly due to the limited spatial and temporal resolution of the highly-integrated RGB-D cameras and the restricted computational power of mobile processors. (3) *Modeling of semantics* with respect to geometry, motion and appearance would strongly improve the deployability of 3D reconstruction approaches to many other fields of application, ranging from entertainment via medicine and health care to autonomous systems.

Emerging Trends and Machine Learning While not the main focus of this survey, machine learning based approaches, especially in the form of deep neural networks, are a very promising avenue for tackling the many challenges in 3D reconstruction, non-rigid tracking and material estimation. Deep learning on 3D data has seen a lot of progress recently, making it hard to exhaustively cover all the literature; we leave this for a dedicated survey paper. In the following, we shortly highlight a few approaches, which we deem

to be most relevant in the context of this STAR report. Feature matching [ZSN*17] for static reconstruction and dense correspondences [WHC*16] for non-rigidly deforming shapes can be learned using deep convolutional neural networks. Completing scans of single objects [WSK*15, DQN17], the environment around an RGB-D frame [SYZ*16], or even of complete scenes [DRB*17] is a very promising and active area of research. In addition, many approaches go beyond pure 3D reconstruction and additionally infer high-level scene semantics [SYZ*16, DCS*17, CDF*17]. Other approaches learn volumetric fusion [RUBG17] to better handle sensor noise by exploiting learned data priors. Recently, deep learning has been shown to perform well on point cloud data [QSMG16]. Besides these more geometry related approaches, machine learning has also been applied to BRDF estimation. The BRDF of an object can be inferred based on RGB-D data captured from multiple viewpoints [KGT*17]. Also very recently, learning based approaches have demonstrated BRDF estimation based on a single input image [RRF*16, GRR*17, LCY*17, LDPT17].

7. Conclusion

Online and interactive 3D reconstruction using commodity RGB-D cameras has evolved dramatically within the last years. 1000+ papers have been published in this field since the upcoming of the first Kinect generation, covering a vast variety of use cases and applications, and there is no saturation in sight. The current development covers the full reconstruction pipeline and brings about innovations at all levels and intermediate steps, from RGB-D camera hardware to high level applications related to all possible aspects of daily life.

This state-of-the-art report aggregates, reviews, compares, and critically analyzes the major aspects of 3D reconstruction using RGB-D cameras. Starting with the rather well-posed problems of static scene reconstruction, for which we outline the basic principles, we unfold various lines of research and development in this respect and evolve them to capturing scene dynamics and appearance, which are far more ill-posed problems, requiring more complex solutions, representations, and regularization techniques. We also look into the methods approaching the challenging, practically highly-relevant, and starkly ill-posed problem of combining shape, appearance, and illumination capture.

We are convinced, that this state-of-the-art report will support the further development of this field in several ways. First, even though this report cannot dive too deep into all technical details, it serves as a starting point for researchers and application engineers new to the field of 3D reconstruction. Second, it serves as reference for researcher active in this field, making them aware of approaches, which are potentially orthogonal to the methodologies they currently apply. Last, in conjunction with the presentation of this report at the Eurographics conference 2018 in Delft, NL, it will foster the discussion with respect to future approaches and potentials of the powerful 3D reconstruction toolbox at hand.

Acknowledgements

This work was supported by the ERC Starting Grant CapReal (335545), the Max Planck Center for Visual Computing and Communications (MPC-VCC), the DFG Research Training Group 1564 Imaging New Modalities, and the DFG projects Ko 2960/13-1 (Dynamic Light Fields), 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), a TUM-IAS Rudolf Mößbauer Fellowship and a Google Faculty Award.

References

[AFB15] Allain B., Franco J.-S., Boyer E.: An efficient volumetric framework for shape tracking. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (June 2015). 9

[APZV17] Alexandrov S. V., Prankl J., Zillich M., Vincze M.: Towards dense SLAM with high dynamic range colors. 14

[AW*87] Amanatides J., Woo A., et al.: A fast voxel traversal algorithm for ray tracing. In *Proc. Eurographics* (1987), vol. 87, pp. 3–10. 6

[BBLR15] Bogo F., Black M. J., Loper M., Romero J.: Detailed full-body reconstructions of moving people from monocular RGB-D sequences. In *Proc. IEEE Int. Conf. on Computer Vision* (Dec. 2015), pp. 2300–2308. 10

[BBS14] Bell S., Bala K., Snavely N.: Intrinsic images in the wild. *ACM Trans. on Graphics 33*, 4 (2014), 159. 18

[BBSG10] Ben-Chen M., Butscher A., Solomon J., Guibas L. J.: On discrete killing vector fields and patterns on surfaces. *Computer Graphics Forum 29*, 5 (2010), 1701–1711. 10, 13

[BC04] Buck D. K., Collins A. A.: POV-Ray - The Persistence of Vision Raytracer, 2004. URL: http://www.povray.org/. 7

[BHLW12] Bojsen-Hansen M., Li H., Wojtan C.: Tracking surfaces with evolving topology. *ACM Trans. on Graphics 31*, 4 (July 2012), 53:1–53:10. 8

[BHY15] Bi S., Han X., Yu Y.: An l 1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM Trans. on Graphics 34*, 4 (2015), 78. 16

[BHZN10] BorosÃąn P., Howard R., Zhang S., Nealen A.: Hybrid Mesh Editing. In *Proc. Eurographics - Short Papers* (2010), Lensch H. P. A., Seipel S., (Eds.), The Eurographics Association. 9

[BKP10] Bredies K., Kunisch K., Pock T.: Total generalized variation. *SIAM Journal on Imaging Sciences 3*, 3 (2010), 492–526. 19

[BKPB17a] Bonneel N., Kovacs B., Paris S., Bala K.: Intrinsic decompositions for image editing. *Computer Graphics Forum (Eurographics State of the Art Reports 2017) 36*, 2 (2017). 2

[BKPB17b] Bonneel N., Kovacs B., Paris S., Bala K.: Intrinsic decompositions for image editing. In *Computer Graphics Forum* (2017), vol. 36, Wiley Online Library, pp. 593–609. 15

[BKR17] Bi S., Kalantari N. K., Ramamoorthi R.: Patch-based optimization for image-based texture mapping. *ACM Trans. on Graphics (Proc. SIGGRAPH) 36*, 4 (2017). 14

[BL95] Blais G., Levine M. D.: Registering multiview range data to create 3D computer objects. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 17*, 8 (Aug 1995), 820–824. doi:10.1109/34.400574. 5

[BLK18] Bulczak D., Lambers M., Kolb A.: Quantified, interactive simulation of amcw tof camera including multipath effects. *Sensors 18*, 1 (2018). URL: http://www.mdpi.com/1424-8220/18/1/13, doi:10.3390/s18010013. 7

[BM92] Besl P. J., McKay N. D.: A method for registration of 3-d shapes. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 14*, 2 (1992), 239–256. 4

[BM12] Barron J. T., Malik J.: Shape, albedo, and illumination from a single image of an unknown object. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2012), IEEE, pp. 334–341. 18

[BM13] Barron J. T., Malik J.: Intrinsic scene properties from a single RGB-D image. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on* (2013), IEEE, pp. 17–24. 15

[BM15] Barron J. T., Malik J.: Shape, illumination, and reflectance from shading. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 37*, 8 (2015), 1670–1687. 15

[Boo89] Bookstein F. L.: Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 11*, 6 (June 1989), 567–585. 10

[BPD09] Bousseau A., Paris S., Durand F.: User-assisted intrinsic images. In *ACM Trans. on Graphics* (2009), vol. 28, p. 130. 17

[BPH*12] Bachrach A., Prentice S., He R., Henry P., Huang A. S., Krainin M., Maturana D., Fox D., Roy N.: Estimation, planning, and mapping for autonomous flight using an rgb-d camera in gps-denied environments. *Int. J. Rob. Res. 31*, 11 (Sept. 2012), 1320–1343. 7

[BR04] Brown B., Rusinkiewicz S.: Non-rigid range-scan alignment using thin-plate splines. In *Proc. Symp. 3D Data Processing, Visualization, and Transmission* (Sept. 2004). 9, 11

[BR07] Brown B., Rusinkiewicz S.: Global non-rigid alignment of 3-D scans. *ACM Trans. on Graphics (Proc. SIGGRAPH) 26*, 3 (Aug. 2007). 9, 10, 11

[BS08] Botsch M., Sorkine O.: On Linear Variational Surface Deformation Methods. *IEEE Trans. on Visualization and Computer Graphics 14*, 1 (2008), 213–230. doi:10.1109/TVCG.2007.1054. 10

[BST*14] Bonneel N., Sunkavalli K., Tompkin J., Sun D., Paris S., Pfister H.: Interactive intrinsic video editing. *ACM Trans. on Graphics 33*, 6 (2014), 197. 15, 16, 17, 19

[BT78] Barrow H. G., Tenenbaum J. M.: *Recovering Intrinsic Scene Characteristics from Images*. Academic Press, 1978. 14

[BTLP16] Bouaziz S., Tagliasacchi A., Li H., Pauly M.: Modern techniques and applications for real-time non-rigid registration. In *SIGGRAPH ASIA 2016 Courses* (New York, NY, USA, 2016), SA '16, ACM, pp. 11:1–11:25. 2

[BTP15] Bouaziz S., Tagliasacchi A., Pauly M.: Dynamic 2d/3d registration. 2

[BTS*14] Berger M., Tagliasacchi A., Seversky L., Alliez P., Levine J., Sharf A., Silva C.: State of the art in surface reconstruction from point clouds. In *Proc. Eurographics - State-of-the-Art Reports (STARs)* (2014), vol. 1, pp. 161–185. 2

[BTVG06] Bay H., Tuytelaars T., Van Gool L.: Surf: Speeded up robust features. *Computer vision–ECCV 2006* (2006), 404–417. 5

[BWP13] Bouaziz S., Wang Y., Pauly M.: Online modeling for realtime facial animation. *ACM Trans. Graph. 32*, 4 (2013), 40:1–40:10. 10

[BWSB12] Butler D. J., Wulff J., Stanley G. B., Black M. J.: A naturalistic open source movie for optical flow evaluation. In *Proc. Europ. Conf. Computer Vision* (2012), Springer, pp. 611–625. 18

[CBI13] Chen J., Bautembach D., Izadi S.: Scalable real-time volumetric surface reconstruction. *ACM Trans. on Graphics (Proc. SIGGRAPH) 32*, 4 (July 2013), 113:1–113:16. 4, 5, 6

[CCS*15] Collet A., Chuang M., Sweeney P., Gillett D., Evseev D., Calabrese D., Hoppe H., Kirk A., Sullivan S.: High-quality streamable free-viewpoint video. *ACM Trans. Graph. 34*, 4 (July 2015), 69:1–69:13. URL: http://doi.acm.org/10.1145/2766945, doi:10.1145/2766945. 8, 13

[CDF*17] Chang A., Dai A., Funkhouser T., Halber M., Niessner M., Savva M., Song S., Zeng A., Zhang Y.: Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158* (2017). 8, 21

[CGL*17] Cavallari T., Golodetz S., Lord N. A., Valentin J., Di Stefano L., Torr P. H.: On-the-fly adaptation of regression forests for online camera relocalisation. *arXiv preprint arXiv:1702.02779* (2017). 5

[CGM99] Chan T. F., Golub G. H., Mulet P.: A nonlinear primal-dual method for total variation-based image restoration. *SIAM journal on scientific computing 20*, 6 (1999), 1964–1977. 19

[CHM05] Caputo B., Hayman E., Mallikarjuna P.: Class-specific material categorisation. In *Proc. IEEE Int. Conf. on Computer Vision* (2005), vol. 2, IEEE, pp. 1597–1604. 18

[CK13] Chen Q., Koltun V.: A simple model for intrinsic image decomposition with depth cues. In *Proc. IEEE Int. Conf. on Computer Vision* (2013), pp. 241–248. 15, 16

[CL96] Curless B., Levoy M.: A volumetric method for building complex models from range images. In *Proc. Comp. Graph. & Interact. Techn.* (1996), pp. 303–312. 3, 6, 8, 12, 17

[CLH15] Chen K., Lai Y.-K., Hu S.-M.: 3d indoor scene modeling from rgb-d data: a survey. *Computational Visual Media 1*, 4 (2015), 267–278. 2

[CLKM15] Charrow B., Liu S., Kumar V., Michael N.: Information-theoretic mapping using cauchy-schwarz quadratic mutual information. In *2015 IEEE International Conference on Robotics and Automation (ICRA)* (May 2015), pp. 4791–4798. 7

[CMG15] Chatterjee A., Madhav Govindu V.: Photometric refinement of depth maps for multi-albedo objects. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2015), pp. 933–941. 18

[CP11] Chambolle A., Pock T.: A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision 40*, 1 (2011), 120–145. 19

[CPTK17] Choe G., Park J., Tai Y.-W., Kweon I. S.: Refining geometry from depth sensors using ir shading images. *International Journal of Computer Vision 122*, 1 (2017), 1–16. 18

[CPTSK14] Choe G., Park J., Tai Y.-W., So Kweon I.: Exploiting shading cues in kinect ir images for geometry refinement. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2014), pp. 3922–3929. 18

[CZ09] Chang W., Zwicker M.: Range scan registration using reduced deformable models. *Computer Graphics Forum 28*, 2 (2009), 447–456. 9, 11

[CZK15a] Choi S., Zhou Q.-Y., Koltun V.: Robust reconstruction of indoor scenes. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2015), pp. 5556–5565. 4, 5, 19

[CZK15b] Choi S., Zhou Q.-Y., Koltun V.: Robust reconstruction of indoor scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015). 7

[dAST*08] de Aguiar E., Stoll C., Theobalt C., Ahmed N., Seidel H.-P., Thrun S.: Performance capture from sparse multi-view video. *ACM Trans. on Graphics (Proc. SIGGRAPH) 27* (2008), 1–10. 9

[DCS*17] Dai A., Chang A. X., Savva M., Halber M., Funkhouser T., Niessner M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE* (2017). 8, 21

[DDF*17] Dou M., Davidson P., Fanello S. R., Khamis S., Kowdle A., Rhemann C., Tankovich V., Izadi S.: Motion2fusion: Real-time volumetric performance capture. *ACM Trans. Graph. 36*, 6 (Nov. 2017), 246:1–246:16. URL: http://doi.acm.org/10.1145/3130800.3130801, doi:10.1145/3130800.3130801. 12

[DF14] Dou M., Fuchs H.: Temporally enhanced 3D capture of room-sized dynamic scenes with commodity depth cameras. In *Proc. IEEE Conf. Virtual Reality* (2014), pp. 39–44. URL: https://doi.org/10.1109/VR.2014.6802048, doi:10.1109/VR.2014.6802048. 8

[DFF13] Dou M., Fuchs H., Frahm J.: Scanning and tracking dynamic objects with commodity depth cameras. In *Proc. IEEE Int. Symp. Mixed and Augmented Reality (ISMAR)* (2013), pp. 99–106. URL: https://doi.org/10.1109/ISMAR.2013.6671769, doi:10.1109/ISMAR.2013.6671769. 9, 10, 12

[DKD*16a] Dou M., Khamis S., Degtyarev Y., Davidson P., Fanello S. R., Kowdle A., Escolano S. O., Rhemann C., Kim D., Taylor

J., Kohli P., Tankovich V., Izadi S.: Fusion4d: Real-time performance capture of challenging scenes. *ACM Trans. on Graphics 35*, 4 (July 2016), 114:1–114:13. URL: http://doi.acm.org/10.1145/2897824.2925969, doi:10.1145/2897824.2925969. 8, 9, 10, 11, 12, 13, 19

[DKD*16b] Dou M., Khamis S., Degtyarev Y., Davidson P., Fanello S. R., Kowdle A., Escolano S. O., Rhemann C., Kim D., Taylor J., Kohli P., Tankovich V., Izadi S.: Fusion4d: Real-time performance capture of challenging scenes. *ACM Trans. on Graphics 35*, 4 (July 2016), 114:1–114:13. URL: http://doi.acm.org/10.1145/2897824.2925969, doi:10.1145/2897824.2925969. 8

[DKSX17] Dryanovski I., Klingensmith M., Srinivasa S. S., Xiao J.: Large-scale, real-time 3d scene reconstruction on a mobile device. *Autonomous Robots* (2017), 1–23. 6

[DLR77] Dempster A. P., Laird N. M., Rubin D. B.: Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B 39*, 1 (1977), 1–38. 9

[DNZ*17] Dai A., Niessner M., Zollöfer M., Izadi S., Theobalt C.: BundleFusion: real-time globally consistent 3D reconstruction using on-the-fly surface re-integration. *ACM Trans. on Graphics 36*, 3 (2017). 4, 5

[DQN17] Dai A., Qi C. R., Niessner M.: Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE* (2017). 21

[DRB*17] Dai A., Ritchie D., Bokeloh M., Reed S., Sturm J., Niessner M.: Scancomplete: Large-scale scene completion and semantic segmentation for 3d scans. *arXiv preprint arXiv:1712.10215* (2017). 21

[DTF*15] Dou M., Taylor J., Fuchs H., Fitzgibbon A. W., Izadi S.: 3D scanning deformable objects with a single RGBD sensor. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2015), pp. 493–501. URL: https://doi.org/10.1109/CVPR.2015.7298647, doi:10.1109/CVPR.2015.7298647. 9, 10, 12, 13

[DVGNK99] Dana K. J., Van Ginneken B., Nayar S. K., Koenderink J. J.: Reflectance and texture of real-world surfaces. *ACM Transactions On Graphics (TOG) 18*, 1 (1999), 1–34. 18

[ESC14] Engel J., Schöps T., Cremers D.: Lsd-slam: Large-scale direct monocular slam. In *Proc. Europ. Conf. Computer Vision* (2014), Springer, pp. 834–849. 3

[Ess09] Esser E.: Applications of lagrangian-based alternating direction methods and connections to split bregman. *CAM report 9* (2009), 31. 19

[ET99] Ekeland I., Temam R.: *Convex analysis and variational problems*. SIAM, 1999. 19

[FG11] Fuhrmann S., Goesele M.: Fusion of depth maps with multiple scales. In *ACM Trans. on Graphics* (2011), vol. 30, ACM, p. 148. 6

[Fir16] Firman M.: RGBD Datasets: Past, Present and Future. In *CVPR Workshop on Large Scale 3D Data: Acquisition, Modelling and Analysis* (2016). 8

[FNT*11] Fujiwara K., Nishino K., Takamatsu J., Zheng B., Ikeuchi K.: Locally rigid globally non-rigid surface registration. In *Proc. IEEE Int. Conf. on Computer Vision* (2011). 8, 9, 10, 11

[GIRL03] Gelfand N., Ikemoto L., Rusinkiewicz S., Levoy M.: Geometrically stable sampling for the ICP algorithm. In *Int. Conf. 3D Digital Imaging and Modeling (3DIM)* (2003), pp. 260–267. 5

[GISC13] Glocker B., Izadi S., Shotton J., Criminisi A.: Real-time rgb-d camera relocalization. IEEE. URL: https://www.microsoft.com/en-us/research/publication/real-time-rgb-d-camera-relocalization/. 7

[GJAF09] Grosse R., Johnson M. K., Adelson E. H., Freeman W. T.: Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *Proc. IEEE Int. Conf. on Computer Vision* (2009), IEEE, pp. 2335–2342. 18

[GMLMG12] Garces E., Munoz A., Lopez-Moreno J., Gutierrez D.: Intrinsic images by clustering. In *Computer Graphics Forum* (2012), vol. 31, Wiley Online Library, pp. 1415–1424. 15

[GO09] Goldstein T., Osher S.: The split bregman method for l1-regularized problems. *SIAM journal on imaging sciences 2*, 2 (2009), 323–343. 19

[GRB94] Godin G., Rioux M., Baribeau R.: Three-dimensional registration using range and intensity information. *Proc. SPIE (Videometrics III) 2350* (1994), 279–290. doi:10.1117/12.189139. 5

[GRR*17] Georgoulis S., Rematas K., Ritschel T., Gavves E., Fritz M., Gool L. V., Tuytelaars T.: Reflectance and natural illumination from single-material specular objects using deep learning. *PAMI* (2017). 21

[GSCI15] Glocker B., Shotton J., Criminisi A., Izadi S.: Real-time rgb-d camera relocalization via randomized ferns for keyframe encoding. *IEEE Trans. on Visualization and Computer Graphics 21*, 5 (2015), 571–583. 5

[GXW*15] Guo K., Xu F., Wang Y., Liu Y., Dai Q.: Robust non-rigid motion tracking and surface reconstruction using l0 regularization. In *Proc. IEEE Int. Conf. on Computer Vision* (December 2015). 9, 10, 11, 13

[GXW*17] Guo K., Xu F., Wang Y., Liu Y., Dai Q.: Robust non-rigid motion tracking and surface reconstruction using l0 regularization. *IEEE Trans. on Visualization and Computer Graphics* (2017). 9, 10, 11, 13

[GXY*17] Guo K., Xu F., Yu T., Liu X., Dai Q., Liu Y.: Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Trans. on Graphics 36*, 3 (June 2017), 32:1–32:13. URL: http://doi.acm.org/10.1145/3083722, doi:10.1145/3083722. 8, 9, 10, 11, 12, 13, 15, 16, 19

[GZC*16] Garrido P., Zollhöfer M., Casas D., Valgaerts L., Varanasi K., Perez P., Theobalt C.: Reconstruction of personalized 3D face rigs from monocular video. *ACM Trans. on Graphics (Proc. SIGGRAPH) 35*, 3 (2016), 28:1–28:15. 10

[HB86] Horn B. K., Brooks M. J.: The variational approach to shape from shading. *Computer Vision, Graphics, and Image Processing 33*, 2 (1986), 174–208. 17

[HBB*13] Helten T., Baak A., Bharaj G., Mueller M., Seidel H.-P., Theobalt C.: Personalization and evaluation of a real-time depth-based full body tracker. In *Proc. Int. Conf. 3D Vision (3DV)* (2013). 10

[HBH*11] Huang A. S., Bachrach A., Henry P., Krainin M., Maturana D., Fox D., Roy N.: Visual odometry and mapping for autonomous flight using an rgb-d camera. In *Proceedings of the International Symposium of Robotics Research (ISRR)* (2011). 7

[HCFE04] Hayman E., Caputo B., Fritz M., Eklundh J.-O.: On the significance of real-world conditions for material classification. In *Proc. Europ. Conf. Computer Vision* (2004), Springer, pp. 253–266. 18

[HCMG*14] Haque M., Chatterjee A., Madhav Govindu V., et al.: High quality photometric reconstruction using a depth camera. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2014), pp. 2275–2282. 17

[HDGN17] Huang J., Dai A., Guibas L., Niessner M.: 3DLite: Towards commodity 3D scanning for content creation. *ACM Trans. on Graphics* (2017). 14

[HFBM13] Henry P., Fox D., Bhowmik A., Mongia R.: Patch volumes: Segmentation-based consistent mapping with rgb-d cameras. In *Proc. Int. Conf 3DTV* (2013), IEEE, pp. 398–405. 6

[HGW15] Hachama M., Ghanem B., Wonka P.: Intrinsic scene decomposition from RGB-D images. In *Proc. IEEE Int. Conf. on Computer Vision* (2015), pp. 810–818. 15, 16

[HJS08] Huhle B., Jenke P., Strasser W.: On-the-fly scene acquisition with a handy multi-sensor system. *Int. Journal of Intelligent Systems Technologies and Applications 5*, 3-4 (2008), 255–263. 2

[HKH*12] Henry P., Krainin M., Herbst E., Ren X., Fox D.: Rgb-d mapping: Using kinect-style depth cameras for dense 3D modeling of indoor environments. *Int. Journal of Robotics Research 31*, 5 (2012), 647–663. 5

[HLJ*15] Han X., Leung T., Jia Y., Sukthankar R., Berg A. C.: Matchnet: Unifying feature and metric learning for patch-based matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 3279–3286. 5

[HLSK13] Han Y., Lee J.-Y., So Kweon I.: High quality shape from a single RGB-D image under uncalibrated natural illumination. In *Proc. IEEE Int. Conf. on Computer Vision* (2013), pp. 1617–1624. 18

[HPB*15] Handa A., Patraucean V., Badrinarayanan V., Stent S., Cipolla R.: Scenenet: Understanding real world indoor scenes with synthetic data. *arXiv preprint arXiv:1511.07041* (2015). 7

[HSR15] Hitomi E. E., Silva J. V., Ruppert G. C.: 3d scanning using rgbd imaging devices: A survey. In *Developments in Medical Image Processing and Computational Vision*. Springer, 2015, pp. 379–395. 2

[HWMD14] Handa A., Whelan T., McDonald J., Davison A.: A benchmark for RGB-D visual odometry, 3D reconstruction and SLAM. In *IEEE Intl. Conf. on Robotics and Automation, ICRA* (Hong Kong, China, May 2014). 7

[IGL03] Ikemoto L., Gelfand N., Levoy M.: A hierarchical method for aligning warped meshes. In *Int. Conf. 3D Digital Imaging and Modeling (3DIM)* (2003), IEEE Computer Society, pp. 434–441. 9, 10

[IKH*11] Izadi S., Kim D., Hilliges O., Molyneaux D., Newcombe R., Kohli P., Shotton J., Hodges S., Freeman D., Davison A., Fitzgibbon A.: KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera. In *Proc. ACM Symp. User Interface Softw. & Tech.* (2011), pp. 559–568. 1, 3, 4, 5, 6, 8, 12

[IOA11] Iason Oikonomidis N. K., Argyros A.: Efficient model-based 3D tracking of hand articulations using kinect. In *Proc. British Machine Vision Conference* (2011), BMVA Press, pp. 101.1–101.11. 10

[IZN*16] Innmann M., Zollhöfer M., Niessner M., Theobalt C., Stamminger M.: VolumeDeform: real-time volumetric non-rigid reconstruction. 8, 9, 10, 11, 12, 13

[JJKL16] Jeon J., Jung Y., Kim H., Lee S.: Texture map generation for 3D reconstructed scenes. *The Visual Computer 32*, 6-8 (2016), 955–965. 14

[JKGJC17] Jaimez M., Kerl C., Gonzalez-Jimenez J., Cremers D.: Fast odometry and scene flow from rgb-d cameras based on geometric clustering. In *Proc. IEEE Int. Conf. Robotics and Automation* (2017). 8

[KAJS11] Koppula H. S., Anand A., Joachims T., Saxena A.: Semantic labeling of 3d point clouds for indoor scenes. In *Advances in Neural Information Processing Systems 24*, Shawe-Taylor J., Zemel R. S., Bartlett P. L., Pereira F., Weinberger K. Q., (Eds.). Curran Associates, Inc., 2011, pp. 244–252. 7

[KBH06] Kazhdan M., Bolitho M., Hoppe H.: Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing* (2006), vol. 7. 3

[KBSB17] Kovacs B., Bell S., Snavely N., Bala K.: Shading annotations in the wild. *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2017). 18

[KCF11] Krainin M., Curless B., Fox D.: Autonomous generation of complete 3d object models using next best view manipulation planning. In *2011 IEEE International Conference on Robotics and Automation* (May 2011), pp. 5031–5037. 7

[KGB14] Kong N., Gehler P. V., Black M. J.: Intrinsic video. In *Proc. Europ. Conf. Computer Vision* (2014), Springer, pp. 360–375. 15, 16, 17

[KGT*17] Kim K., Gu J., Tyree S., Molchanov P., Niessner M., Kautz J.: A lightweight approach for on-the-fly reflectance estimation. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017* (2017), pp. 20–28. URL: https://doi.org/10.1109/ICCV.2017.12, doi:10.1109/ICCV.2017.12. 21

[KH13] Kazhdan M., Hoppe H.: Screened poisson surface reconstruction. *ACM Transactions on Graphics (TOG) 32*, 3 (2013), 29. 3

[KLL*13a] Keller M., Lefloch D., Lambers M., Izadi S., Weyrich T., Kolb A.: Real-time 3D reconstruction in dynamic scenes using point-based fusion. In *Proc. Int. Conf. 3D Vision (3DV)* (Washington, DC, USA, 2013), IEEE Computer Society, pp. 1–8. URL: http://dx.doi.org/10.1109/3DV.2013.9, doi:10.1109/3DV.2013.9. 3, 4, 5, 8

[KLL*13b] Keller M., Lefloch D., Lambers M., Izadi S., Weyrich T., Kolb A.: Real-time 3D reconstruction in dynamic scenes using point-based fusion. In *Proc. Int. Conf. 3D Vision (3DV)* (2013), p. 8. 4, 5, 6

[KP15] Kolb A., Pece F.: *Digital Representations of the Real World: How to Capture, Model, and Render Visual Reality*. CRC Press, 2015, ch. Range Imaging. 2

[KPR*15] Kähler O., Prisacariu V. A., Ren C. Y., Sun X., Torr P., Murray D.: Very high frame rate volumetric integration of depth images on mobile devices. *IEEE Trans. on Visualization and Computer Graphics 21*, 11 (2015), 1241–1250. 6

[KPVM16] Kähler O., Prisacariu V., Valentin J., Murray D.: Hierarchical voxel block hashing for efficient integration of depth images. *IEEE Robotics and Automation Letters (RA-L) 1*, 1 (2016), 192–197. 6

[KRB*12] Kriegel S., Rink C., BodenmÃijller T., Narr A., Suppa M., Hirzinger G.: Next-best-scan planning for autonomous 3d modeling. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Oct 2012), pp. 2850–2856. 7

[KRBS15] Kriegel S., Rink C., Bodenmüller T., Suppa M.: Efficient next-best-scan planning for autonomous 3d surface reconstruction of unknown objects. *J. Real-Time Image Process. 10*, 4 (Dec. 2015), 611–631. 7

[KSSC14] Kerl C., Souiai M., Sturm J., Cremers D.: Towards Illumination-invariant 3D Reconstruction using ToF RGB-D Cameras. In *Proc. Int. Conf. 3D Vision (3DV)* (2014), vol. 1, IEEE, pp. 39–46. 15, 16, 17

[KTTW12] Knecht M., Tanzmeister G., Traxler C., Wimmer M.: Interactive brdf estimation for mixed-reality applications. 17

[LAGP09] Li H., Adams B., Guibas L. J., Pauly M.: Robust single-view geometry and motion reconstruction. *ACM Trans. on Graphics (Proc. SIGGRAPH ASIA) 28*, 5 (December 2009). 8, 9, 10, 11

[LB14] Li Y., Brown M. S.: Single image layer separation using relative smoothness. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2014), pp. 2752–2759. 16

[LB15] Laffont P.-Y., Bazin J.-C.: Intrinsic decomposition of image sequences from local temporal variations. In *Proc. IEEE Int. Conf. on Computer Vision* (2015), pp. 433–441. 15, 16

[LBD13] Laffont P.-Y., Bousseau A., Drettakis G.: Rich intrinsic image decomposition of outdoor scenes from multiple views. *IEEE Trans. on Visualization and Computer Graphics 19*, 2 (2013), 210–224. 15

[LBP*12] Laffont P.-Y., Bousseau A., Paris S., Durand F., Drettakis G.: Coherent intrinsic images from photo collections. *ACM Trans. on Graphics 31*, 6 (2012). 15, 17

[LC87] Lorensen W. E., Cline H. E.: Marching cubes: A high resolution 3d surface construction algorithm. In *Proc. SIGGRAPH* (1987), vol. 21, ACM, pp. 163–169. 6

[LCY*17] Liu G., Ceylan D., Yumer E., Yang J., Lien J.-M.: Material editing using a physically based rendering network. In *Proceedings of International Conference on Computer Vision (ICCV) (spotlight presentation)* (2017), pp. 2261–2269. 21

[LDPT17] Li X., Dong Y., Peers P., Tong X.: Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Trans. Graph. 36*, 4 (July 2017), 45:1–45:11. URL: http://doi.acm.org/10.1145/3072959.3073641, doi:10.1145/3072959.3073641. 21

[LHZC16] Li S., Handa A., Zhang Y., Calway A.: HDRFusion: HDR SLAM using a low-cost auto-exposure RGB-D sensor. In *Proc. Int. Conf. 3D Vision (3DV)* (2016), IEEE, pp. 314–322. 14

[LKH07] Lindner M., Kolb A., Hartmann K.: Data-fusion of PMD-based distance-information and high-resolution RGB-images. In *Int. Sym. on Signals Circuits & Systems (ISSCS), session on Algorithms for 3D TOF-cameras* (2007), IEEE, pp. 121–124. 2

[LKS*17] Lefloch D., Kluge M., Sarbolandi H., Weyrich T., Kolb A.: Comprehensive use of curvature for robust and accurate online surface reconstruction. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)* (2017), 10.1109/TPAMI.2017.2648803. 4, 5, 7

[LLV*12] Li H., Luo L., Vlasic D., Peers P., Popović J., Pauly M., Rusinkiewicz S.: Temporally coherent completion of dynamic shapes. *ACM Trans. on Graphics 31*, 1 (January 2012). 8

[LM71] Land E. H., McCann J. J.: Lightness and retinex theory. *Josa 61*, 1 (1971), 1–11. 16

[LM79] Lions P.-L., Mercier B.: Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis 16*, 6 (1979), 964–979. 19

[LN16] Lombardi S., Nishino K.: Radiometric scene decomposition: Scene reflectance, illumination, and geometry from RGB-D images. In *Proc. Int. Conf. 3D Vision (3DV)* (2016), IEEE, pp. 305–313. 17

[Low99] Lowe D. G.: Object recognition from local scale-invariant features, 1999. 5

[Low04a] Low K.-L.: Linear least-squares optimization for point-to-plane icp surface registration. *Chapel Hill, University of North Carolina 4* (2004). 4

[Low04b] Lowe D. G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision 60* (2004), 91–110. 5

[LRL14] Lu S., Ren X., Liu F.: Depth enhancement via low-rank matrix completion. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2014), pp. 3390–3397. 18

[LS09] Lo T.-W. R., Siebert J. P.: Local feature extraction and matching on range images: 2.5D SIFT. *Computer Vision and Image Understanding 113*, 2009 (2009), 1235–1250. 5

[LSP08] Li H., Sumner R. W., Pauly M.: Global correspondence optimization for non-rigid registration of depth scans. In *Computer Graphics Forum (Proc. Geometry Processing)* (Aire-la-Ville, Switzerland, Switzerland, 2008), Eurographics Association, pp. 1421–1430. URL: http://dl.acm.org/citation.cfm?id=1731309.1731326. 8, 9, 10, 11, 19

[LVG*13] Li H., Vouga E., Gudym A., Luo L., Barron J. T., Gusev G.: 3d self-portraits. *ACM Trans. on Graphics (Proc. SIGGRAPH ASIA) 32*, 6 (November 2013). 10

[LWK15] Lefloch D., Weyrich T., Kolb A.: Anisotropic point-based fusion. In *Proc. Int. Conf. Information Fusion (FUSION)* (July 2015), pp. 1–9. 4, 7

[LYYB13] Li H., Yu J., Ye Y., Bregler C.: Realtime facial animation with on-the-fly correctives. *ACM Trans. on Graphics (Proc. SIGGRAPH) 32*, 4 (July 2013). 10

[LZT*12] Lee K. J., Zhao Q., Tong X., Gong M., Izadi S., Lee S. U., Tan P., Lin S.: Estimation of intrinsic image sequences from image+ depth video. In *Proc. Europ. Conf. Computer Vision* (2012), Springer, pp. 327–340. 15, 16

[LZW*09] Liao M., Zhang Q., Wang H., Yang R., Gong M.: Modeling deformable objects from a single depth camera. In *Proc. IEEE Int. Conf. on Computer Vision* (2009), pp. 167–174. 11

[MAMT15] Mur-Artal R., Montiel J. M. M., Tardos J. D.: Orb-slam: a versatile and accurate monocular SLAM system. *IEEE Trans. Robotics and Automation 31*, 5 (2015), 1147–1163. 3

[Mat03] Matusik W.: *A data-driven reflectance model*. PhD thesis, Massachusetts Institute of Technology, 2003. 18

[MBC13] Meilland M., Barat C., Comport A.: 3D high dynamic range dense visual slam and its application to real-time object re-lighting. In *Proc. IEEE Int. Symp. Mixed and Augmented Reality (ISMAR)* (2013), IEEE, pp. 143–152. 14

[MFO*07] Mitra N. J., Flöry S., Ovsjanikov M., Gelfand N., Guibas L., Pottmann H.: Dynamic geometry registration. In *Computer Graphics Forum (Proc. Geometry Processing)* (Aire-la-Ville, Switzerland, Switzerland, 2007), Eurographics Association, pp. 173–182. 12

[MFZ*17] Meka A., Fox G., Zollhöfer M., Richardt C., Theobalt C.: Live User-Guided Intrinsic Video For Static Scene. *IEEE Trans. on Visualization and Computer Graphics* (2017). URL: http://gvv.mpi-inf.mpg.de/projects/InteractiveIntrinsicAR/. 13, 15, 16, 17, 19

[MHFdS*12] Maier-Hein L., Franz A., dos Santos T., Schmidt M., Fangerau M., Meinzer H., Fitzpatrick J.: Convergent iterative closest-point algorithm to accomodate anisotropic and inhomogenous localization error. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 34*, 8 (Aug 2012), 1520–1532. doi:10.1109/TPAMI.2011.248. 4, 7

[MKC*17] Maier R., Kim K., Cremers D., Kautz J., Niessner M.: Intrinsic3D: High-quality 3D reconstruction by joint appearance and geometry optimization with spatially-varying lighting. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), vol. 4. 14, 18

[MSC15] Maier R., Stückler J., Cremers D.: Super-resolution keyframe fusion for 3D modeling with high-quality textures. In *Proc. Int. Conf. 3D Vision (3DV)* (2015), IEEE, pp. 536–544. 14

[MZRT16] Meka A., Zollhöfer M., Richardt C., Theobalt C.: Live Intrinsic Video. *ACM Trans. on Graphics 35*, 4 (2016), 109. 15, 16, 17, 19

[NA15] Narayan K. S., Abbeel P.: Optimized color models for high-quality 3D scanning. In *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)* (2015), IEEE, pp. 2503–2510. 14

[NDI*11] Newcombe R. A., Davison A. J., Izadi S., Kohli P., Hilliges O., Shotton J., Molyneaux D., Hodges S., Kim D., Fitzgibbon A.: KinectFusion: real-time dense surface mapping and tracking. In *Proc. IEEE Int. Symp. Mixed and Augmented Reality (ISMAR)* (2011), pp. 127–136. 1, 3, 4, 5, 6, 8, 12

[NFS15] Newcombe R. A., Fox D., Seitz S. M.: DynamicFusion: Reconstruction and Tracking of Non-Rigid Scenes in Real-Time. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (June 2015). 8, 9, 10, 11, 12

[NIL12] Nguyen C., Izadi S., Lovell D.: Modeling Kinect sensor noise for improved 3D reconstruction and tracking. In *Proc. Int. Conf. 3D Imaging, Modeling, Processing, Vis. & Transmission* (2012), pp. 524–530. 5

[NLB*11] Neumann D., Lugauer F., Bauer S., Wasza J., Hornegger J.: Real-time RGB-D mapping and 3-D modeling on the GPU using the random ball cover data structure. In *Proc. IEEE Int. Conf. Computer Vision (ICCV), Workshops* (2011), IEEE, pp. 1161–1167. 5

[NSF12] Nathan Silberman Derek Hoiem P. K., Fergus R.: Indoor segmentation and support inference from rgbd images. In *ECCV* (2012). 7

[NZIS13] Niessner M., Zollhöfer M., Izadi S., Stamminger M.: Real-time 3D reconstruction at scale using voxel hashing. *ACM Trans. on Graphics 32*, 6 (2013), 169. 3, 4, 5, 6, 14

[OEHW*16] Or-El R., Hershkovitz R., Wetzler A., Rosman G., Bruckstein A. M., Kimmel R.: Real-time Depth Refinement for Specular Objects. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2016), pp. 4378–4386. 18

[OERF*16] Orts-Escolano S., Rhemann C., Fanello S., Chang W., Kowdle A., Degtyarev Y., Kim D., Davidson P. L., Khamis S., Dou M., Tankovich V., Loop C., Cai Q., Chou P. A., Mennicken S., Valentin J., Pradeep V., Wang S., Kang S. B., Kohli P., Lutchyn Y., Keskin C., Izadi S.: Holoportation: Virtual 3D teleportation in real-time. In *Proc. Symp. User Interface Software and Technology* (New York, NY, USA, 2016), ACM, pp. 741–754. URL: http://doi.acm.org/10.1145/2984511.2984517, doi:10.1145/2984511.2984517. 8, 12

[OERW*15] Or-El R., Rosman G., Wetzler A., Kimmel R., Bruckstein A. M.: RGBD-Fusion: real-time high precision depth recovery. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2015). 18

[Pho75] Phong B. T.: Illumination for computer generated pictures. *Communications of the ACM 18*, 6 (1975), 311–317. 18

[PMC*11] Pomerleau F., Magnenat S., Colas F., Liu M., Siegwart R.: Tracking a depth camera: Parameter exploration for fast icp. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems* (Sept 2011), pp. 3824–3829. doi:10.1109/IROS.2011.6094861. 7

[PRR03] Paragios N., Rousson M., Ramesh V.: Non-rigid registration using distance functions. *Computer Vision and Image Understanding 89*, 2-3 (Feb. 2003), 142–165. 8, 11

[PSL*98] Parker S., Shirley P., Livnat Y., Hansen C., Sloan P.-P.: Interactive ray tracing for isosurface rendering. In *Proc. IEEE Visualization* (1998), IEEE, pp. 233–238. 6

[Pul99] Pulli K.: Multiview registration for large data sets. In *Int. Conf. 3D Digital Imaging and Modeling (3DIM)* (1999), pp. 160–168. 5

[PZvBG00] Pfister H., Zwicker M., van Baar J., Gross M.: Surfels: Surface elements as rendering primitives. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 2000), SIGGRAPH '00, ACM Press/Addison-Wesley Publishing Co., pp. 335–342. 6

[QSMG16] Qi C. R., Su H., Mo K., Guibas L. J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. *arXiv preprint arXiv:1612.00593* (2016). 21

[RA17] Rünz M., Agapito L.: Co-fusion: Real-time segmentation, tracking and fusion of multiple objects. In *Proc. IEEE Int. Conf. Robotics and Automation* (May 2017), pp. 4471–4478. 8

[RHHL02] Rusinkiewicz S., Hall-Holt O., Levoy M.: Real-time 3D model acquisition. *ACM Trans. on Graphics (Proc. SIGGRAPH) 21*, 3 (2002), 438–446. 3, 4, 6, 7

[RL01] Rusinkiewicz S., Levoy M.: Efficient variants of the ICP algorithm. In *Int. Conf. 3D Digital Imaging and Modeling (3DIM)* (2001), IEEE, pp. 145–152. 4, 5

[RLL14] Rhee S.-M., Lee Y. B., Lee H.-E.: Two-pass ICP with color constraint for noisy rgb-d point cloud registration. In *Proc. IEEE Int. Conf. Consumer Electronics (ICCE)* (2014), IEEE, pp. 89–90. 5

[ROF92] Rudin L. I., Osher S., Fatemi E.: Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena 60*, 1-4 (1992), 259–268. 19

[RR15] Richter S. R., Roth S.: Discriminative shape from shading in uncalibrated illumination. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2015), pp. 1128–1136. 18

[RRF*16] Rematas K., Ritschel T., Fritz M., Gavves E., Tuytelaars T.: Deep reflectance maps. In *IEEE Conference on Computer Vision and Pattern Recognition* (2016). URL: https://ivi.fnwi.uva.nl/isis/publications/2016/RematasCVPR2016. 21

[RRKB11] Rublee E., Rabaud V., Konolige K., Bradski G.: Orb: An efficient alternative to sift or surf. In *Computer Vision (ICCV), 2011 IEEE international conference on* (2011), IEEE, pp. 2564–2571. 5

[RTKPS16] Richter-Trummer T., Kalkofen D., Park J., Schmalstieg D.: Instant mixed reality lighting from casual scanning. In *Mixed and Augmented Reality (ISMAR), 2016 IEEE International Symposium on* (2016), IEEE, pp. 27–36. 14, 15, 17

[RUBG17] Riegler G., Ulusoy A. O., Bischof H., Geiger A.: Octnetfusion: Learning depth fusion from data. In *International Conference on 3D Vision (3DV) 2017* (Oct. 2017). 21

[RWW16] Reichl F., Weiss J., Westermann R.: Memory-efficient interactive online reconstruction from depth image streams. *Computer Graphics Forum* (2016), to appear. URL: http://dx.doi.org/10.1111/cgf.12779, doi:10.1111/cgf.12779. 6

[SA07] Sorkine O., Alexa M.: As-rigid-as-possible surface modeling. In *Computer Graphics Forum (Proc. Geometry Processing)* (Aire-la-Ville, Switzerland, Switzerland, 2007), Eurographics Association, pp. 109–116. URL: http://dl.acm.org/citation.cfm?id=1281991.1282006. 10

[SAL*08] Sharf A., Alcantara D. A., Lewiner T., Greif C., Sheffer A., Amenta N., Cohen-Or D.: Space-time surface reconstruction using incompressible flow. *ACM Trans. on Graphics 27*, 5 (Dec. 2008), 110:1–110:10. 12

[SBCBG11] Solomon J., Ben-Chen M., Butscher A., Guibas L. J.: As-killing-as-possible vector fields for planar deformation. *Computer Graphics Forum 30*, 5 (2011), 1543–1552. 10, 13

[SBCI17] Slavcheva M., Baust M., Cremers D., Ilic S.: Killingfusion: Non-rigid 3D reconstruction without correspondences. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (July 2017). 8, 9, 10, 11, 12, 13

[SBK*13] Sturm J., Bylow E., Kerl C., Kahl F., Cremers D.: Dense tracking and mapping with a quadrocopter. 395–400. 7

[SEE*12] Sturm J., Engelhard N., Endres F., Burgard W., Cremers D.: A benchmark for the evaluation of rgb-d slam systems. In *Proc. of the International Conference on Intelligent Robot Systems (IROS)* (2012). 7

[SF11] Silberman N., Fergus R.: Indoor scene segmentation using a structured light sensor. In *Proceedings of the International Conference on Computer Vision - Workshop on 3D Representation and Recognition* (2011). 7

[SG14] Serafin J., Grisetti G.: Using augmented measurements to improve the convergence of icp. In *Simulation, Modeling, and Programming for Autonomous Robots*. Springer, 2014, pp. 566–577. 5

[SGZ*13] Shotton J., Glocker B., Zach C., Izadi S., Criminisi A., Fitzgibbon A.: Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2013), pp. 2930–2937. 5

[SKC13] Steinbrucker F., Kerl C., Cremers D.: Large-scale multi-resolution surface reconstruction from rgb-d sequences. In *Proceedings of the IEEE International Conference on Computer Vision* (2013), pp. 3264–3271. 4, 5

[SLK15] Sarbolandi H., Lefloch D., Kolb A.: Kinect range sensing: Structured-light versus time-of-flight kinect. *Journal of Computer Vision and Image Understanding 13* (2015), 1–20. 2, 4

[SLX15] Song S., Lichtenberg S. P., Xiao J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2015). 7

[SMGKD14] Salas-Moreno R. F., Glocken B., Kelly P. H., Davison A. J.: Dense planar SLAM. In *Proc. IEEE Int. Symp. Mixed and Augmented Reality (ISMAR)* (2014), pp. 157–164. 7

[SNF15] Schmidt T., Newcombe R., Fox D.: Dart: Dense articulated real-time tracking with consumer depth cameras. *Auton. Robots 39*, 3 (Oct. 2015), 239–258. 10

[SP86] Sederberg T. W., Parry S. R.: Free-form deformation of solid geometric models. In *Proc. SIGGRAPH* (New York, NY, USA, 1986), ACM, pp. 151–160. 9

[SSC14] Steinbrücker F., Sturm J., Cremers D.: Volumetric 3D mapping in real-time on a cpu. In *Proc. IEEE Int. Conf. Robotics and Automation* (2014), IEEE, pp. 2021–2028. 6

[SSP07] Sumner R. W., Schmid J., Pauly M.: Embedded deformation for shape manipulation. In *ACM Trans. on Graphics (Proc. SIGGRAPH)* (New York, NY, USA, 2007), ACM. URL: http://doi.acm.org/10.1145/1275808.1276478, doi:10.1145/1275808.1276478. 9, 10, 12

[STL08] Shen L., Tan P., Lin S.: Intrinsic image decomposition with non-local texture cues. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2008), IEEE, pp. 1–7. 15

[SWG08] Süssmuth J., Winter M., Greiner G.: Reconstructing animated meshes from time-varying point clouds. *Computer Graphics Forum (Proc. Geometry Processing) 27*, 5 (2008), 1469–1476. 12

[SYH13] Shen L., Yeo C., Hua B.-S.: Intrinsic image decomposition using a sparse representation of reflectance. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 35*, 12 (2013), 2904–2915. 16

[SYJL11] Shen J., Yang X., Jia Y., Li X.: Intrinsic images using optimization. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2011), IEEE, pp. 3481–3487. 15, 16

[SYLJ13] Shen J., Yang X., Li X., Jia Y.: Intrinsic image decomposition using optimization and user scribbles. *IEEE Transactions on Cybernetics 43*, 2 (2013), 425–436. 15, 16

[SYS07] Sofka M., Yang G., Stewart C. V.: Simultaneous covariance driven correspondence (cdc) and transformation estimation in the expectation maximization framework. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2007), pp. 1–8. 7

[SYZ*16] Song S., Yu F., Zeng A., Chang A. X., Savva M., Funkhouser T.: Semantic scene completion from a single depth image. *arXiv preprint arXiv:1611.08974* (2016). 7, 21

[TAF06] Tappen M. F., Adelson E. H., Freeman W. T.: Estimating intrinsic component images using non-linear regression. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2006), vol. 2, IEEE, pp. 1992–1999. 15

[TBC*16] Taylor J., Bordeaux L., Cashman T., Corish B., Keskin C., Soto E., Sweeney D., Valentin J., Luff B., Topalian A., Wood E., Khamis S., Kohli P., Sharp T., Izadi S., Banks R., Fitzgibbon A., Shotton J.: Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. In *ACM Trans. on Graphics (Proc. SIGGRAPH)* (2016). 10, 11

[TBW*12] Tevs A., Berner A., Wand M., Ihrke I., Bokeloh M., Kerber J., Seidel H.-P.: Animation cartography—intrinsic reconstruction of shape and motion. *ACM Trans. on Graphics 31*, 2 (Apr. 2012), 12:1–12:15. 8

[TFA05] Tappen M., Freeman W., Adelson E.: Recovering intrinsic images from a single image. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 27*, 9 (2005), 1459–1472. 15

[TM98] Tomasi C., Manduchi R.: Bilateral filtering for gray and color images. In *Proc. IEEE Int. Conf. on Computer Vision* (1998), pp. 839–846. 4

[TPT16] Tkach A., Pauly M., Tagliasacchi A.: Sphere-meshes for real-time hand modeling and tracking. *ACM Trans. on Graphics (Proc. SIGGRAPH ASIA)* (2016). 10

[TSB16] Tao M., Solomon J., Butscher A.: Near-isometric level set tracking. *Computer Graphics Forum* (2016). 10, 13

[TST*15] Tagliasacchi A., Schroeder M., Tkach A., Bouaziz S., Botsch M., Pauly M.: Robust articulated-ICP for real-time hand tracking. *Computer Graphics Forum (Proc. Geometry Processing)* (2015). 10

[TZL*12] Tong J., Zhou J., Liu L., Pan Z., Yan H.: Scanning 3D full human bodies using kinects. *IEEE Trans. on Visualization and Computer Graphics 18*, 4 (Apr. 2012), 643–650. 10

[TZN*15] Thies J., Zollhöfer M., Nießner M., Valgaerts L., Stamminger M., Theobalt C.: Real-time expression transfer for facial reenactment. *ACM Trans. on Graphics 34*, 6 (2015). 10, 11

[TZS*16] Thies J., Zollhöfer M., Stamminger M., Theobalt C., Nießner M.: Face2Face: real-time face capture and reenactment of rgb videos. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2016). 16, 19

[VBMP08] Vlasic D., Baran I., Matusik W., Popović J.: Articulated mesh animation from multi-view silhouettes. In *ACM SIGGRAPH 2008 Papers* (New York, NY, USA, 2008), SIGGRAPH '08, ACM, pp. 97:1–97:9. URL: http://doi.acm.org/10.1145/1399504.1360696, doi:10.1145/1399504.1360696. 13

[VDN*16] Valentin J., Dai A., Nießner M., Kohli P., Torr P., Izadi S., Keskin C.: Learning to navigate the energy landscape. In *Proc. Int. Conf. 3D Vision (3DV)* (2016), IEEE, pp. 323–332. 5

[VNS*15] Valentin J., Nießner M., Shotton J., Fitzgibbon A., Izadi S., Torr P. H.: Exploiting uncertainty in regression forests for accurate camera relocalization. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2015), pp. 4400–4408. 5

[WAO*09] Wand M., Adams B., Ovsjanikov M., Berner A., Bokeloh M., Jenke P., Guibas L., Seidel H.-P., Schilling A.: Efficient reconstruction of nonrigid shape and motion from real-time 3D scanner data. *ACM Trans. on Graphics 28*, 2 (May 2009), 15:1–15:15. 12

[War92] Ward G. J.: Measuring and modeling anisotropic reflection. *Proc. SIGGRAPH 26*, 2 (1992), 265–272. 17

[WBLP11] Weise T., Bouaziz S., Li H., Pauly M.: Realtime performance-based facial animation. *ACM Trans. on Graphics (Proc. SIGGRAPH) 30*, 4 (July 2011). 10

[WFR*16] Wang S., Fanello S. R., Rhemann C., Izadi S., Kohli P.: The global patch collider. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2016), pp. 127–135. 10

[WG17] Wang C., Guo X.: Feature-based rgb-d camera pose optimization for real-time 3D reconstruction. *Computational Visual Media 3*, 2 (Jun 2017), 95–106. URL: https://doi.org/10.1007/s41095-016-0072-2, doi:10.1007/s41095-016-0072-2. 5

[WGK14] Weinmann M., Gall J., Klein R.: Material classification based on training data synthesized using a btf database. In *Proc. Europ. Conf. Computer Vision* (2014), pp. 156–171. 18

[WHB11] Weiss A., Hirshberg D., Black M.: Home 3D body scans from noisy image and range data. In *Proc. IEEE Int. Conf. on Computer Vision* (Barcelona, Nov. 2011), IEEE, pp. 1951–1958. 10

[WHC*16] Wei L., Huang Q., Ceylan D., Vouga E., Li H.: Dense human body correspondences using convolutional networks. In *Computer Vision and Pattern Recognition (CVPR)* (2016). 21

[WJH*07] Wand M., Jenke P., Huang Q., Bokeloh M., Guibas L., Schilling A.: Reconstruction of deforming geometry from time-varying point clouds. In *Computer Graphics Forum (Proc. Geometry Processing)* (Aire-la-Ville, Switzerland, Switzerland, 2007), Eurographics Association, pp. 49–58. URL: http://dl.acm.org/citation.cfm?id=1281991.1281998. 12

[WJK*13] Whelan T., Johannsson H., Kaess M., Leonard J. J., McDonald J.: Robust real-time visual odometry for dense rgb-d mapping. In *Proc. IEEE Int. Conf. Robotics and Automation* (2013), IEEE, pp. 5724–5731. 5

[WKF*12] Whelan T., Kaess M., Fallon M., Johannsson H., Leonard J., McDonald J.: Kintinuous: Spatially extended kinectfusion. In *Proc. RSS Workshop on RGB-D: Advanced Reasoning with Depth Cameras* (2012). 4, 5, 6

[WKJ*15] Whelan T., Kaess M., Johannsson H., Fallon M., Leonard J. J., McDonald J.: Real-time large-scale dense RGB-D SLAM with volumetric fusion. *Int. Journal of Robotics Research 34*, 4-5 (2015), 598–626. 14

[WLGK16] Weinmann M., Langguth F., Goesele M., Klein R.: Advances in geometry and reflectance acquisition. In *Eurographics 2016 Tutorial* (2016). 2, 17

[WLSM*15] Whelan T., Leutenegger S., Salas-Moreno R. F., Glocker B., Davison A. J.: Elasticfusion: Dense SLAM without a pose graph. In *Proceedings of Robotics: Science and Systems* (2015). 5, 6

[WMS16] Wasenmüller O., Meyer M., Stricker D.: CoRBS: Comprehensive rgb-d benchmark for slam using kinect v2. In *IEEE Winter Conference on Applications of Computer Vision (WACV)* (March 2016). URL: http://corbs.dfki.uni-kl.de/. 7

[WSK*15] Wu Z., Song S., Khosla A., Yu F., Zhang L., Tang X., Xiao J.: 3d shapenets: A deep representation for volumetric shapes. In *CVPR* (2015), IEEE Computer Society, pp. 1912–1920. 21

[WSL*14] Wu S., Sun W., Long P., Huang H., Cohen-Or D., Gong M., Deussen O., Chen B.: Quality-driven poisson-guided autoscanning. *ACM Trans. Graph. 33*, 6 (Nov. 2014), 203:1–203:12. 7

[WSMG*16] Whelan T., Salas-Moreno R. F., Glocker B., Davison A. J., Leutenegger S.: ElasticFusion: real-time dense SLAM and light source estimation. *Int. Journal of Robotics Research* (2016). 4, 5, 14

[WSVT13] Wu C., Stoll C., Valgaerts L., Theobalt C.: On-set performance capture of multiple actors with a stereo camera. In *ACM Trans. on Graphics (Proc. SIGGRAPH ASIA)* (November 2013), vol. 32, pp. 161:1–161:11. URL: http://doi.acm.org/10.1145/2508363.2508418, doi:10.1145/2508363.2508418. 10

[WWLG11] Weise T., Wismer T., Leibe B., Gool L. V.: Online loop closure for real-time interactive 3D scanning. *Computer Vision and Image Understanding 115*, 5 (May 2011), 635–648. 10

[WWLVG09] Weise T., Wismer T., Leibe B., Van Gool L.: In-hand scanning with online loop closure. In *Proc. IEEE Int. Conf. on Computer Vision Workshops (ICCV Workshops)* (2009), IEEE, pp. 1630–1637. 7

[WWZ16] Wu H., Wang Z., Zhou K.: Simultaneous localization and appearance estimation with a consumer RGB-D camera. *IEEE Trans. on Visualization and Computer Graphics 22*, 8 (2016), 2012–2023. 13, 15, 17

[WZ15] Wu H., Zhou K.: Appfusion: Interactive appearance acquisition using a kinect sensor. In *Computer Graphics Forum* (2015), vol. 34, Wiley Online Library, pp. 289–298. 15, 17

[WZN*14] Wu C., Zollhöfer M., Niessner M., Stamminger M., Izadi S., Theobalt C.: Real-time shading-based refinement for consumer depth cameras. *ACM Trans. on Graphics 33*, 6 (2014), 200. 18

[XLC*18] Xu L., Liu Y., Cheng W., Guo K., Zhou G., Dai Q., Fang L.: FlyCap: markerless motion capture using multiple autonomous flying cameras. *IEEE Trans. on Visualization and Computer Graphics* (2018). 7, 11

[XZY*17] Xu K., Zheng L., Yan Z., Yan G., Zhang E., Niessner M., Deussen O., Cohen-Or D., Huang H.: Autonomous reconstruction of unknown indoor scenes guided by time-varying tensor fields. *ACM Trans. Graph. 36*, 6 (Nov. 2017), 202:1–202:15. URL: http://doi.acm.org/10.1145/3130800.3130812, doi:10.1145/3130800.3130812. 7

[YCK*16] Yoon Y., Choe G., Kim N., Lee J.-Y., Kweon I. S.: Fine-scale surface normal estimation using a single nir image. In *Proc. Europ. Conf. Computer Vision* (2016), Springer, pp. 486–500. 18

[YGL*14] Ye G., Garces E., Liu Y., Dai Q., Gutierrez D.: Intrinsic video and applications. *ACM Trans. on Graphics 33*, 4 (2014), 80. 17, 18

[YGX*17] Yu T., Guo K., Xu F., Dong Y., Su Z., Zhao J., Li J., Dai Q., Liu Y.: Bodyfusion: Real-time capture of human motion and surface geometry using a single depth camera. In *Proc. IEEE Int. Conf. on Computer Vision* (October 2017), ACM. 8, 10, 13

[YLH*12] Ye G., Liu Y., Hasler N., Ji X., Dai Q., Theobalt C.: Performance capture of interacting characters with handheld kinects. In *Proc. ECCV* (2012), Springer, pp. 828–841. 10

[YM92] Yang C., Medioni G.: Object modelling by registration of multiple range images. *Image and Vision Computing 10*, 3 (1992), 145–155. 4, 5

[YTLF16] Yi K. M., Trulls E., Lepetit V., Fua P.: Lift: Learned invariant feature transform. In *European Conference on Computer Vision* (2016), Springer, pp. 467–483. 5

[YY14] Ye M., Yang R.: Real-time simultaneous pose and shape estimation for articulated objects using a single depth camera. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (Washington, DC, USA, 2014), IEEE Computer Society, pp. 2353–2360. 10

[YYTL13] Yu L.-F., Yeung S.-K., Tai Y.-W., Lin S.: Shading-based shape refinement of RGB-D images. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2013), pp. 1415–1422. 18

[Zac14a] Zach C.: Robust bundle adjustment revisited. In *Proc. Europ. Conf. Computer Vision* (2014), pp. 772–787. 11

[Zac14b] Zach C.: Robust bundle adjustment revisited. In *European Conference on Computer Vision* (2014), Springer, pp. 772–787. 19

[ZCC16] Zhang E., Cohen M. F., Curless B.: Emptying, refurnishing, and relighting indoor spaces. *ACM Trans. on Graphics (Proc. SIGGRAPH ASIA) 35*, 6 (2016). 14

[ZDI*15] Zollhöfer M., Dai A., Innmann M., Wu C., Stamminger M., Theobalt C., Niessner M.: Shading-based refinement on volumetric signed distance functions. *ACM Trans. on Graphics 34*, 4 (2015). 16, 18

[ZFYY14] Zhang Q., Fu B., Ye M., Yang R.: Quality dynamic human body modeling using a single low-cost depth camera. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (Washington, DC, USA, 2014), Proc. IEEE Conf. Computer Vision and Pattern Recognition, IEEE Computer Society, pp. 676–683. 10

[ZK13] Zhou Q.-Y., Koltun V.: Dense scene reconstruction with points of interest. *ACM Trans. on Graphics 32*, 4 (2013), 112. 4, 5

[ZK14] Zhou Q.-Y., Koltun V.: Color map optimization for 3D reconstruction with consumer depth cameras. *ACM Trans. on Graphics 33*, 4 (2014), 155. 14

[ZK15] Zhou Q.-Y., Koltun V.: Depth camera tracking with contour cues. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2015), pp. 632–638. 5

[ZMK13] Zhou Q.-Y., Miller S., Koltun V.: Elastic fragments for dense scene reconstruction. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2013), pp. 473–480. 3, 4, 5, 11

[ZNI*14] Zollhöfer M., Niessner M., Izadi S., Rehmann C., Zach C., Fisher M., Wu C., Fitzgibbon A., Loop C., Theobalt C., Stamminger M.: Real-time non-rigid reconstruction using an rgb-d camera. *ACM Trans. on Graphics 33*, 4 (July 2014), 156:1–156:12. doi:10.1145/2601097.2601165. 8, 9, 10, 11, 19

[ZSGS12] Zollhöfer M., Sert E., Greiner G., Süssmuth J.: Gpu based arap deformation using volumetric lattices. In *Proc. Eurographics (Short Papers)* (2012), Andújar C., Puppo E., (Eds.), Eurographics Association, pp. 85–88. 9, 19

[ZSN*17] Zeng A., Song S., Niessner M., Fisher M., Xiao J., Funkhouser T.: 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *CVPR* (2017). 5, 21

[ZTD*12] Zhao Q., Tan P., Dai Q., Shen L., Wu E., Lin S.: A closed-form solution to retinex with nonlocal texture constraints. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI) 34*, 7 (2012), 1437–1444. 15, 16

[ZWZY17] Zuo X., Wang S., Zheng J., Yang R.: Detailed surface geometry and albedo recovery from RGB-D video under natural illumination. *arXiv preprint arXiv:1702.01486* (2017). 15, 16

[ZX18] Zhang H., Xu F.: Mixedfusion: Real-time reconstruction of an indoor scene with dynamic objects. *IEEE Transactions on Visualization and Computer Graphics PP*, 99 (2018), 1–1. 12

[ZYY*12] Zhang Q., Ye M., Yang R., Matsushita Y., Wilburn B., Yu H.: Edge-preserving photometric stereo via depth fusion. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (2012), IEEE, pp. 2472–2479. 18

[ZZCL13] Zeng M., Zheng J., Cheng X., Liu X.: Templateless quasi-rigid shape modeling with implicit loop-closure. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (Washington, DC, USA, 2013), IEEE Computer Society, pp. 145–152. 10, 12

[ZZZL12] Zeng M., Zhao F., Zheng J., Liu X.: A memory-efficient kinectfusion using octree. In *Computational Visual Media*. Springer, 2012, pp. 234–241. 6

[ZZZL13] Zeng M., Zhao F., Zheng J., Liu X.: Octree-based fusion for realtime 3D reconstruction. *Graphical Models 75*, 3 (2013), 126–136. 6