Joint Multi-view Texture Super-resolution and Intrinsic Decomposition

Vagia Tsiminaki, Wei Dong, Martin R Oswald, Marc Pollefeys

Joint Multi-view Texture Super-resolution and Intrinsic Decomposition

Vagia Tsiminaki[1]
vagia.tsiminaki@inf.ethz.ch
Wei Dong[2]
weidong@andrew.cmu.edu
Martin R. Oswald[1]
martin.oswald@inf.ethz.ch
Marc Pollefeys[1,3]
marc.pollefeys@inf.ethz.ch

[1] Computer Vision and Geometry group
ETH Zurich
Zurich, Switzerland

[2] Robotics Institute
Carnegie Mellon University
Pittsburgh, USA

[3] Microsoft
USA

Abstract

We aim to recover a high resolution texture representation of objects observed from multiple view points under varying lighting conditions. For many applications the lighting conditions need to be changed and thus require a texture decomposition into shading and albedo components. Both texture super-resolution and intrinsic texture decomposition have been separately studied in the literature. Yet, no method has investigated how these methods can be combined. We propose a framework for joint texture map super-resolution and intrinsic decomposition. To this end, we define shading and albedo maps of the 3D object as the intrinsic properties of its texture and introduce an image formation model to describe the physics of the image generation. Our approach accounts for surface geometry and camera calibation errors and is also applicable to spatio-temporal sequences. Our method achieves state-of-the-art results on a variety of datasets.

1 Introduction

Image-based 3D reconstruction has been a long time research focus in computer vision. Impressive advances have been made such that state-of-the-art methods are now able to recover fine geometric details with similar or even better accuracy than expensive laser scanners. While these methods cleverly use the information redundancy of a multi-view setup to recover high-frequency geometric details, there are few methods which do so for computing highly-detailed texture maps. With the increasing demand of 3D content for television, gaming, augmented and virtual reality applications as well as for industrial software, recovering high-resolution texture details is of equal importance. For instance, in tasks like surface or material quality inspection or in medical applications such technology in combination with commodity cameras has the potential to replace expensive task-specific sensor technology.

In this paper, we focus on recovering high resolution texture maps by solving the inverse problem of the physical generative imaging process. Since captured 3D models are often used with different lighting conditions than the ones at capturing time, it is essential to be

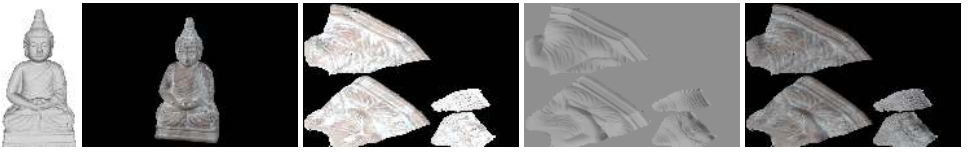

Figure 1: **Overview of our method.** We compute super-resolved texture maps while jointly decomposing the texture into albedo and shading components.

able to remove scene dependent light conditions from the high quality texture. Therefore, we propose a method that simultaneously decomposes shading and albedo while super-resolving the texture map. See Fig. 1 for an overview. In sum, we make the following **contributions**:
1) We present the first method for joint texture super-resolution and intrinsic decomposition in a 3D multi-view setting. We show that the joint estimation of both entities gives superior results than their independent estimation and demonstrate possible applications.
2) We further extend our method to the spatio-temporal case for which we show that the quality and temporal consistency of texture and albedo maps can be further improved by additionally considering further images from neighboring time steps.

2 Related Work

Since we combine super-resolution texture mapping with intrinsic decomposition in a multi-view setting, we exploit results from multiple subfields of computer vision which have been well studied in separate scenarios. This section outlines the most important related works.

2D Intrinsic Decomposition. There have been a plethora of studies performing intrinsic decomposition to retrieve shading and albedo from images. An overview and benchmark can be found in [15]. The vast majority of intrinsic decomposition methods impose priors in the log-domain [3, 5, 11, 15, 19, 21, 41, 42], emphasizing pairwise smoothness in the color space. Bell *et al*. [5] integrated multiple prevalent 2D priors and could handle most scenes, but lack the ability to deal with hard shadows. For the intrinsic decomposition of videos, temporal consistency is stressed. Weiss [41] dealt with time variant lighting assuming an albedo constancy. Kong *et al*. [19] processed videos enforcing temporal albedo consistency and shading similarity. Later, a real-time pipeline was built by Meka *et al*. [29] utilizing non-local spatial-temporal constancy. Yet these methods stick to 2D priors without exploring the underlying geometry that defines the shading, hence the problem is ill-posed to some extent.

3D Intrinsic Decomposition. Intrinsic decomposition in a 3D setting from multiple images has been studied in combination with classical image input [30, 31], but also in combination with RGB-D input [3, 18, 21]. In contrast to many 2D intrinsic decomposition methods, several 3D intrinsic decomposition define priors in the color domain rather than in the log-color domain and approximate the lighting model with spherical harmonics [26, 28, 31, 44]. In [26] ideas from shape-from-shading approaches are used for the 3D reconstruction of non-rigid monocular image sequences with human faces. Zollhöfer *et al*. [44] additionally refine the 3D model which is computed from a series of RGB-D images. A recent extension of this method [28] introduces spatially varying spherical harmonics for improved refinement results. Both [28, 44] intrinsically decompose only the chromacity channel rather than RGB.

Multi-view Texture Mapping. The simplest way for creating a texture map on an object surface from a set of photographs is to blend the weighted color values of the input [8].This,

however, leads to over-smoothed textures. Therefore, many works introduce additional registration in order to reduce the amount of ghosting artifacts [6, 9, 22, 23, 36, 37, 40]. The most generic way to correct for both geometric inaccuracies and camera calibration erros is an optical flow alignment step for registering the down-projected input images, e.g. as done in [9, 40]. Mostly these methods merge or select input appearance information with some kind of weigthed averaging scheme and thus limiting the output texture resolution to the one of the input images. In sum, they do not fully exploit the multi-view viewpoint redundancy to generate textures which exceed the resolution of the input images.

2D Image & Video Super-resolution. Although barely studied in the multi-view texture-mapping case, single image and video super-resolution has been studied in many works. Many early methods rely on a generative image formation model with blurring, warping, down-sampling and solve the corresponding inverse problem [2], follow a Bayesian approach [10, 25], or use variational approaches [32]. Tung *et al*. [39] considered a multi-view setting, yet their approach targeted on super-resolving all input videos rather than the model's texture map. Recently, machine learning-based methods have lead to significant performance improvements, e.g. with residual or generative adversarial networks [20, 35] or regression networks [1]. Impressive results with super-resolved human face images have recently been achieved by Saito *et al*. [34]. Although it is great to see how far machine learning approaches can push the state-of-art, this deep network is heavily overfit to human faces and the method is not generic to arbitrary textures. Further, these methods may hallucinate details, generating undesirable outputs. In this paper, we only use the physics of the image formation model and solve for the inverse problem.

Multi-view Texture Super-resolution. In a series of works Goldlücke *et al*. provided the first approach to compute super-resolved texture maps on arbitrary manifolds [13] which then was extended to also jointly refine the geometry [14] and camera calibration [12]. Improved super-resolution results have been achieved by Tsiminaki *et al*. [38] in which they additionally perform optical flow optimization to account for inevitable surface geometry as well as camera calibration errors. We follow the ideas of this approach and generalize it for joint intrinsic texture map decomposition. In [27] high-res textures are computed from a sequence of RGB-D images in an online setting, but without fully leveraging view redundancy. [16] compute super-resolved geometry, but no textures or intrinsic decomposition. Tab. 1 summarizes the propterties of the most related works. In sum, no existing method fully exploits multi-view redundancy to generate high-res texture maps and to decompose them into high-res albedo maps that are invariant to light conditions.

Method	Intrins. Decomp.	Super-Resolution	3D multi-view	Space-Time
Jeon *et al*. [18]	✓	✓		
Kong *et al*. [19]	✓			✓
Meka *et al*. [29]	✓			✓
Mitzel *et al*. [32]		✓		✓
Eisemann *et al*. [9]			✓	
Wächter *et al*. [40]			✓	
Melou *et al*. [31]	✓		✓	
Maier *et al*. [27]		✓	✓	
Zollhöfer *et al*. [44]	✓		✓	
Maier *et al*. [28]	✓		✓	
Goldlücke *et al*. [12]		✓	✓	
Tsiminaki *et al*. [38]		✓	✓	✓
Ours	✓	✓	✓	✓

Table 1: **Overview of related methods.**

3 Problem Formulation

Problem Setting. Our goal is to compute high-resolution, intrinsically decomposed texture maps for an arbitrary scene model from given input images. We consider an n-view multi-camera setup with given projection matrices $\{\mathrm{P}_i\}_{i=1}^n$, $\mathrm{P}_i : \mathbb{R}^3 \to \mathbb{R}^2$ and input color images $\{\mathrm{I}_i\}_{i=1}^n$, with $\mathrm{I}_i : \Omega_i \subset \mathbb{R}^2 \to \mathbb{R}^3$. For a given scene model, provided as a mesh $\mathcal{M}$, we aim to compute a super-resolved texture map T and a corresponding decomposition into

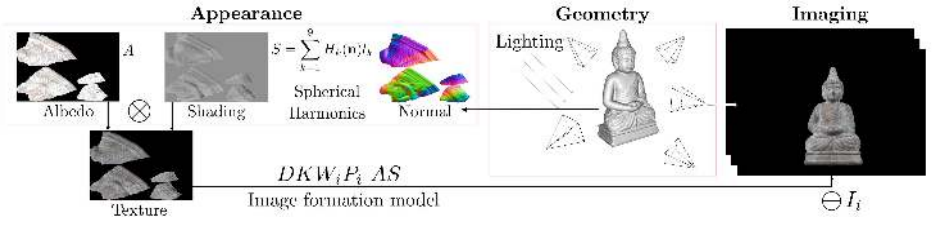

Figure 2: **Our image formation model and notations.** We compute a super-resolved texture map $\mathrm{T} = \mathrm{AS}$ decomposed into albedo A and shading S components. After projecting the high-resolution texture map into camera view i, distortion warping with W_i, blurring with kernel K and downsampling with D, the generated image should look like the corresponding low-resolution input image I_i.

an albedo map A and shading map S, such that $\mathrm{T}(x) = \mathrm{A}(x) \cdot \mathrm{S}(x)$ in every point x. In our setting, the texture, albedo and shading map will also be represented by 2D images, $\mathrm{T}, \mathrm{A} : \mathbb{T} \subset \mathbb{R}^2 \to \mathbb{R}^3$, and $\mathrm{S} : \mathbb{T} \to \mathbb{R}$ which store an unwrapped version of $\mathcal{M}$ as a texture atlas which has potentially been cut into separate texture maps. We also consider input videos and temporally changing, dynamically deforming meshes, but for simplicity of notation we first discuss the static case and extend our model for the dynamic case later on.

Image Formation Model and Super-resolution. In order to exploit the view redundancy of a multi-camera setup, we target a texture map resolution which is significantly higher than the input image resolution. Intuitively, we are observing a continuous mesh surface that is sampled with a low resolution frequency by each of the input cameras. In practice, the camera chip integrates all incoming light within the area of a pixel to a single value, which we model mathematically with a Gaussian blurring kernel K combined with a downsampling operator D. Thus, a low-res image I^{LR} can be obtained from a high-res image I^{HR} via blurring and downsampling, $\mathrm{I}^{\mathrm{LR}} = \mathrm{DKI}^{\mathrm{HR}}$. In multi-view texture mapping, we also need to model the projective mapping P_i between the texture atlas space and every input image i. Similar to [38, 40], we also consider geometric inaccuracies and camera calibration noise with an optical flow alignment step, represented with an per-image warping operator $\mathrm{W}_i : \mathbb{R}^2 \to \mathbb{R}^2$. In sum, in the ideal case a low-res input image I_i can be computed from the high-res texture atlas as a concatenation of perspective projection, optical flow warping, blurring and downsampling:

$$\mathrm{I}_i = \mathrm{DKW}_i\mathrm{P}_i \cdot \mathrm{T}. \tag{1}$$

For texture super-resolution we aim to fulfill this equation for all input views. An overview of our image formation model is depicted in Fig. 2.

Intrinsic Decomposition. As mentioned before, we express the appearance T of the object as a point-wise multiplication of the albedo map A and shading map S. The albedo map is the intrinsic color of the surface that is independent of lighting conditions while the shading map depends on the surface orientation and the local illumination conditions. Under the assumption of a Lambertian reflectance model we approximate the shading map S using spherical harmonics (SH) basis functions [33] that depend on the local surface orientation. In particular, we use a second-order spherical harmonics lighting model with nine coefficients

$$\mathrm{S} = \sum_{k=1}^{9} \mathrm{H}_k(\mathbf{n}) l_k \quad , \tag{2}$$

where $\mathrm{H}_k(\mathbf{n})$ are the spherical harmonics basis functions taken from [44] parameterized by the local surface orientation $\mathbf{n}$, and $\mathbf{l} = (l_1, l_2, \ldots, l_9)$ are the corresponding spherical harmonics coefficients. Similar to [44] we parametrize the spherical harmonics basis functions as $\{\mathrm{H}_k\}_{k=1}^{9} = \{1, n_y, n_z, n_x, n_x n_y, n_y n_z, -n_x n_x - n_y n_y + 2n_z n_z, n_z n_x, n_x n_x - n_y n_y\}$, where $\mathbf{n} = (n_x, n_y, n_z)$ is the local normalized surface orientation. This parametrization of the shading incorporates geometric information into the lighting model and simplifies the intrinsic decomposition problem.

4 Joint Intrinsic Decomposition and Super-resolution

Using the image formation model (Eq. (1)) , we aim to solve the inverse problem while accounting for noise in the input images, calibration and surface geometry. Thus, we propose an energy minimization model that effectively accounts for missing data and inaccuracies.

Energy Formulation. Since the image formation model in Fig. 2 can never be perfectly fulfilled, we minimize the residual in form of the back-projection error. To assure a well-posed energy, we assume piece-wise smooth warping functions and albedo map. The super-resolved, decomposed texture map can then be computed as the minimizer of the following energy $E(\mathrm{A}, \mathrm{S}, \mathrm{W})$ that depends on albedo, shading and optical flow warping:

$$\underset{\mathrm{A,S,W}}{\text{minimize}} \sum_{i=1}^{n} \int_{\mathbb{T}} \Big[\|\mathrm{DKW}_i \mathrm{P}_i \mathrm{AS} - \mathrm{I}_i\|_2^2 + \lambda_\mathrm{A} \|\nabla \mathrm{A}\|_2 + \lambda_\mathrm{W} \|\nabla \mathrm{W}_i\|_2 \Big] \, dx. \tag{3}$$

The weights $\lambda_\mathrm{A}, \lambda_\mathrm{W} \in \mathbb{R}_{\geq 0}$ account for the expected noise level for albedo and warping.

4.1 Optimization

To locally minimize the non-convex energy in Eq. (3) we alternate the optimization of the optical flow warp, albedo and shading independently while keeping the other entities mutually fixed. The individual energy minimizations are described in the following.

Albedo estimation. The albedo map can be estimated by computing the global minimum of Eq. (4) with the Fast Iterative Shrinkage and Thresholding Algorithm (FISTA) [4]. We denote the first quadratic term by $f_{\text{data}}(\mathrm{A})$ and the second term by $f_{TV}(\mathrm{A})$, and compute the minimizer iteratively by updating Eq. (5) until convergence:

$$\mathrm{A}^* = \arg\min_{\mathrm{A}} \sum_{i=1}^{n} \int_{\mathbb{T}} \Big[\|\mathrm{DKW}_i \mathrm{P}_i \mathrm{AS} - \mathrm{I}_i\|_2^2 + \lambda_\mathrm{A} \|\nabla \mathrm{A}\|_2 \Big] \, dx, \tag{4}$$

$$\mathrm{A}^{k+1} = \mathrm{prox}_{\gamma f_{TV}} \Big(\mathrm{A}^k - \gamma \nabla f_{\text{data}}(\mathrm{A}^k) \Big). \tag{5}$$

The gradient of the data term is $\nabla f_{\text{data}}(\mathrm{A}^k) = 2N_i^T (N_i \mathrm{A}^k - \mathrm{I}_i)$ with $N_i = \mathrm{DKW}_i \mathrm{P}_i \, \mathrm{diag}(\mathrm{S})$ and is weighted by gradient descent step size γ. The proximal operator performs a generalized projection: $\mathrm{prox}_{\gamma G}(x) = \arg\min_y \left\{ \frac{1}{2} \|x - y\|^2 + \gamma G(x) \right\}$.

Shading estimation. For the estimation of the shading parameters $\mathbf{l}$, Eq. (3) simplifies to

$$\mathbf{l}^* = \arg\min_{\mathbf{l}} \sum_{i=1}^{n} \int_{\mathbb{T}} \|\mathrm{DKW}_i \mathrm{P}_i \mathrm{AS}(\mathbf{l}) - \mathrm{I}_i\|_2^2 \, dx = \sum_{c=1}^{3} \sum_{i=1}^{n} \mathbf{M}_i^{cT} \mathbf{M}_i^c \left(\sum_{c=1}^{3} \sum_{i=1}^{n} \mathbf{M}_i^{cT} \mathrm{I}_i^c \right)^{-1}. \tag{6}$$

Finding the best SH coefficients $\mathbf{l}^*$ is straightforward. In the discretized setting, we can rewrite all symbols in Eq. (6) with matrices and vectors that cover the entire domain $\mathbb{T}$ as $\mathbf{l}^* = \arg\min_{\mathbf{l}} \sum_{c=1}^{3} \sum_{i=1}^{n} |\mathbf{M}_i^c \mathbf{l} - \mathrm{I}_i^c|^2$ with $\mathbf{M}_i^c = \mathbf{D}\mathbf{K}\mathbf{W}_i\mathbf{P}_i \operatorname{diag}(\mathrm{A}^c)\mathrm{H}$ and c being the color channel. In practice we solve this problem iteratively with a standard Matlab solver.

Optical flow warp estimation. We estimate a vector field W_i for each view $i \in \{1,\ldots,n\}$:

$$\mathrm{W}_i^* = \arg\min_{\mathrm{W}_i} \int_{\mathbb{T}} \left[\|\mathrm{D}\mathrm{K}\mathrm{W}_i\mathrm{P}_i\mathrm{A}\mathrm{S} - \mathrm{I}_i\|_2^2 + \lambda_{\mathrm{W}} \|\nabla \mathrm{W}_i\|_2 \right] dx. \tag{7}$$

We use the coarse-to-fine scheme in [25] to compute the flow field. A local minimum of Eq. (7) is obtained via iterated re-weighted least squares (IRLS). In sum, the computation of intrinsic decomposition and joint super-resolution is performed by iterating Eqs. (4)-(7).

Initialization. We initialize the albedo by utilizing the off-the-shelf intrinsic decomposition system [5] that performs well on images in the wild. The texture, treated as a regular image, can be decomposed into initial albedo and shading textures provided an active area mask.

4.2 Spatio-temporal Setting

Our approach is easily extended to process multi-view videos and an arbitrarly deforming mesh. To exploit appearance information from several time steps, we assume constant albedo within a temporal window of neighboring frames. In our experiments we found a window size of 3 to provide the best trade-off between additional accuracy and processing time. The energy for the spatio-temporal case is then defined on frames around the current time step τ.

$$E(\mathrm{A},\mathrm{S},\mathrm{W},\tau) = \sum_{t=\tau-1}^{\tau+1} \sum_{i=1}^{n} \int_{\mathbb{T}} \left[\left\|\mathrm{D}^t\mathrm{K}^t\mathrm{W}_i^t\mathrm{P}_i^t\mathrm{A}\mathrm{S}^t - \mathrm{I}_i^t\right\|_2^2 + \lambda_{\mathrm{A}} \|\nabla \mathrm{A}\|_2 + \lambda_{\mathrm{W}^t} \|\nabla \mathrm{W}_i^t\|_2 \right] dx. \tag{8}$$

The optimization is analogous to the one in Eq. (3).

5 Experiments

Setup. We carried out all experiments using a MATLAB implementation on a 2.20GHz Intel Xeon E52660 CPU with 256 GB RAM. We initialize the algorithm by first computing a weighted average texture map of visible inputs and use the code of [5] to compute the initial albedo and derive the initial shading. We threshold the relative norm of the energy to stop the optimization (usually 10-60 iterations). The execution time is in the range of 15-40 minutes per iteration depending on the dataset size, i.e. number of views and image resolution. Note that much better performance can be achieved by parallelizing the optimization on a GPU.

5.1 Joint Decomposition on Synthetic Data

We evaluate the performance of our model under varying lighting conditions on the synthetic TOAD dataset [24]. We introduce 3 scenes with different lighting scenarios: one light source on the left of the object (Left), one on the front (Front) and two light sources on the left and above the object (Left+Above). In each case, we use the ground truth geometry and albedo from [24] and the synthetic shading to render the model from 56 viewpoints (512×512).

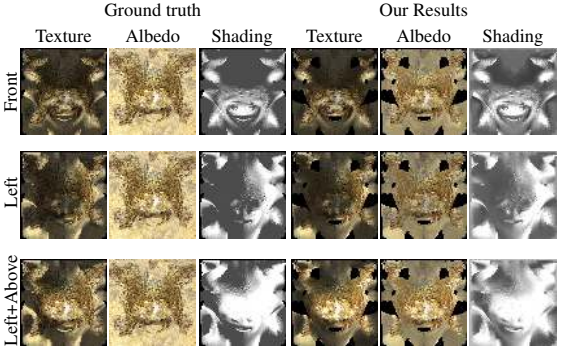

Figure 3: **Experiments with varying lighting conditions.** We have placed a single directional light in *front* of the object, *left* of the object, or two directional lights on the *left and above* the object. It can be observed in the mostly similar albedo results that our method is robust to changing lighting conditions. The recovered shading maps are similar to the ground truth indicating that the light direction is correctly estimated.

Lighting	Type	Method	MSE	SSIM
	Texture	Proposed	0.016342	0.853446
		Sequential	**0.016319**	**0.859915**
	Albedo	Proposed	**0.100509**	**0.751124**
		Sequential	0.106267	0.670265
Front	Shading	Proposed	**0.023330**	**0.848372**
		Sequential	0.029368	0.608505
	Texture	Proposed	**0.019509**	0.854575
		Sequential	0.019613	**0.863193**
	Albedo	Proposed	**0.106619**	**0.731606**
		Sequential	0.109467	0.668565
Left	Shading	Proposed	**0.032784**	**0.832347**
		Sequential	0.037840	0.570647
	Texture	Proposed	0.030629	0.829590
		Sequential	**0.029568**	**0.850900**
	Albedo	Proposed	**0.107606**	**0.722176**
		Sequential	0.109863	0.684500
Left+Above	Shading	Proposed	**0.042381**	**0.815609**
		Sequential	0.053433	0.533307

Table 2: **Comparison to sequential approach:** Super-resolution by [38] followed by 2D intrinsic decomposition [5]. The table shows MSE and SSIM scores evaluated on the ground truth texture atlases. Our method consistently yields more accurate albedo and shading maps.

We use up-sampling factor $\times 2$, i.e. we reconstruct albedo, shading and texture with an atlas resolution of 1024×1024 and compare with the ground truth, as shown in Fig. 3.

Our method yields results close to the ground truth in every case. By changing the light positions and by increasing their number the extraction of the shading becomes more challenging. Our model is able to deal with such variations of lighting conditions.

We further compare to the naive sequential approach consisting of super-resolving the

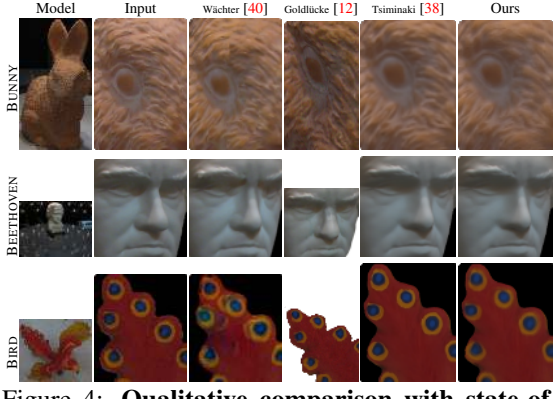

Figure 4: **Qualitative comparison with state-of-the-art texturing methods.** While our method additionally computes a texture decomposition, the combined results are comparable to [38].

	Image Domain		Texture Domain	
Accuracy	MSE	SSIM	MSE	SSIM
BUNNY	0.000056	0.997700	0.000201	0.963691
BEETHOVEN	0.000042	0.994283	0.000110	0.987816
BIRD	0.000037	0.997763	0.000141	0.979812

Table 3: **Distance to the method of Tsiminaki *et al.* [38].** Mean value of the MSE (lower is better) and SSIM (higher is better) are computed between the rendered images (image domain) and between the texture atlases (texture domain). The higher the SSIM and the lower the MSE, the closer the our output is to [38].

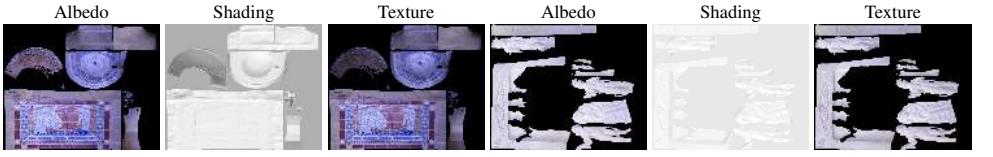

Figure 5: **Output of our method on FOUNTAIN and RELIEF datasets.** The albedo contains the color information, the shading reflects the normals of the mesh and the reconstructed texture entails high frequency details.

texture with [38] followed by 2D intrinsic decomposition [5]. In Table 2 we report the MSE and SSIM scores computed in the texture domain with respect to the reconstructed texture, albedo and shading maps. We see that our method consistently outperforms the sequential method. Note that our goal are superior results with our joint intrinsic decomposition over the sequential method rather than outperforming [38] since the texture optimization is similar.

5.2 Joint Decomposition on Real Data

We run experiments on 6 publicly available real-world datasets. The first 3 datasets BUNNY, BEETHOVEN and BIRD used in [12] are captured in a controlled capturing studio, while FOUNTAIN [43] and RELIEF [44] datasets are from less controlled environments. BUNNY, BEETHOVEN and BIRD consist of 19, 33 and 36 calibrated images with 1024×768 pixels, and FOUNTAIN and RELIEF consist of 55 and 40 key frames with 1024×1280 pixels.

We compare to the method of Wächter [40], the state-of-the-art multi-view texture super-resolution techniques by Goldlücke *et al.* [12] and Tsiminaki *et al.* [38] on BEETHOVEN, BUNNY and BIRD. We use a texture atlas resolution of $2\times$ the input image resolution and use identical 3D models as input. Our method achieves comparable results to [38], as shown

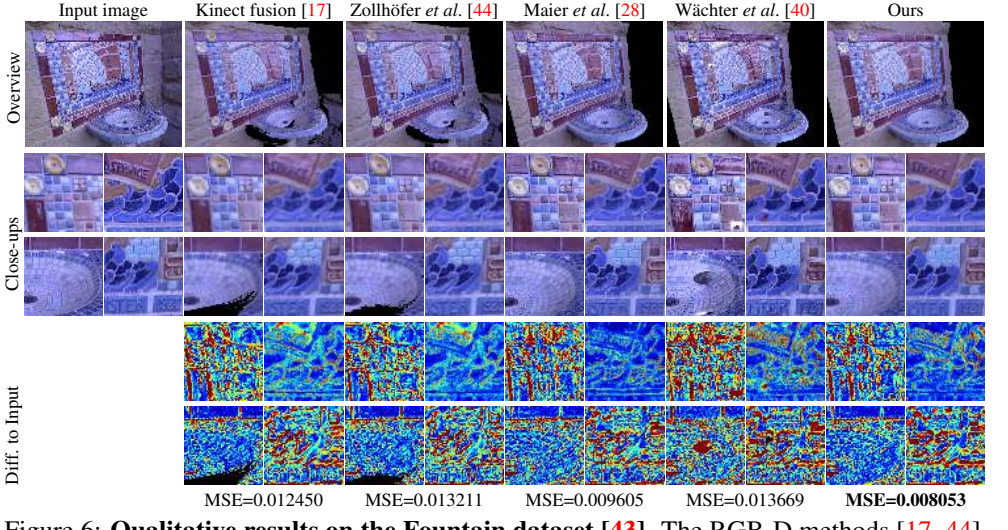

Figure 6: **Qualitative results on the Fountain dataset [43].** The RGB-D methods [17, 44] blur the texture due to low voxel resolution and camera misalignments, while [28] generates good results via camera pose and geometry optimization. [40] often introduces artifacts and seams misalignments. We recover high frequency details and remove apparent specularity.

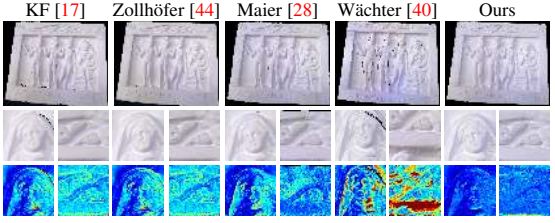

Figure 7: **Qualitative results on the Relief dataset [44].** Our method successfully denoises and recovers fine details of the texture. Similar to Fig. 6, we also show difference maps and view-averaged MSE values for each method.

Figure 8: **Relighting Example. (a)** Selected input view of TOAD in the original scene. From left to right renderings in the new scene using **(b)** super-resolved texture of [38] **(c)** output of sequential approach [38]+[5] **(d)** output of our method. Our method removes shading effects at capture time and re-rendering looks more realistic.

in Fig. 4. To quantify differences, we take the output of [38] as reference texture and compute the error between the reconstructed texture of our method as well as the error between the reprojected images. Tab. 3 shows that our method achieves comparable results to [38].

We use the same upscaling factors for the FOUNTAIN [43] and RELIEF [44] datasets. Due to the ℓ_2 data term in Eq. (3), our method averages out the non-lambertian properties and reconstructs an intrinsic albedo map that is invariant to illumination changes as well as a shading map, as shown in Fig. 5. We compare our method to Kinect fusion [17], Zollhöfer *et al.* [44], Maier *et al.* [28] and Wächter *et al.* [40]. A fair comparison of the intrinsic decomposition is not possible since the methods of Zollhöfer *et al.* [44] and Maier *et al.* [28] perform intrinsic decomposition only on the chromacity and not on the full RGB information. We thus focus on the reconstruction of the texture and compare the re-projections. Figures 6 and 7 show close ups of one selected re-projected image as well as the difference maps with the corresponding mean value of the mean square error. Our method is able to exploit the visual redundancy and recovers high-frequency details.

Extension to the Temporal Domain. We evaluate the applicability of our method on the temporal domain and demonstrate the advantage of the joint optimization. We run experiments on a selected time window of size 3 of the Running sequence of TOMAS [7] by downscaling the 64 images to 512×512. We compare our proposed joint optimization to the naive sequential approach similarly to Sec. 5.1. By introducing additional time frames, the lighting conditions change and the shading decomposition becomes more challenging. The sequential approach cannot distinguish the high-frequency details of the albedo and it incorrectly introduces them into the shading map. Our method effectively deals with these variations and correctly extracts the shading maps at each frame, as shown in Fig. 9.

5.3 Applications, Limitations and Future Work

Applications. An interesting application of our method is object relighting. We qualitatively evaluate our method on object relighting using the TOAD dataset where the light source was placed left of the object and compare it to the naive approach of using the super-resolved texture of [38] and the sequential approach presented in Sec. 5.1. To relight the object we create a new scene with new directional light sources above the object and on the

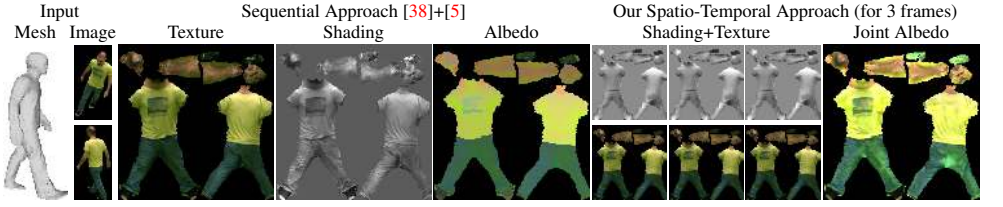

Figure 9: **Sequential vs. Spatio-temporal approach.** The sequential approach incorrectly introduces high frequency details of the albedo in the shading map like the logo on the T-Shirt. Our joint optimization successfully decomposes the shading from the albedo.

front left side. Our method successfully removes from the initial shading effects and the new renderings integrate realistically the new shading, as shown in Fig. 8.

Limitations and Future Work. The image formation model derivation contains a several common assumptions that open up directions for future work. Firstly, our data term favors Lambertian lighting and deviations like specularities are averaged out in our solution. Further, the spherical harmonic light model assumes a distant monochromatic light source and thus spatially varying lighting, cast shadows or light occlusions cannot be captured by our model. Moreover, the shading decomposition is currently governed by the surface normals of the given model and missing high-frequency model details cannot be captured by the shading model. The simultaneous optimization of the surface geometry could tackle this issue.

6 Conclusion

We presented a novel texture super-resolution approach which jointly decomposes the high-resolution texture into shading and albedo components. Our approach builds on well established state-of-the-art generative super-resolution models and generalizes them for joint intrinsic decomposition. Our method exploits knowledge about the 3D model to guide the intrinsic decomposition with surface normal information. In turn, we do not need strong priors for the decomposition and obtain superior results compared to 2D decomposition techniques. In addition to experiments on real and synthetic data of static scenes we showed the applicability of our method to spatio-temporal multi-view sequences. Future work will focus on the concurrent refinement of the surface geometry and normal information.

Acknowledgements. This research was partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/ Interior Business Center (DOI/IBC) contract number D17PC00280. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Eirikur Agustsson, Radu Timofte, and Luc Van Gool. Anchored regression networks applied to age estimation and super resolution. In *Proc. Intl. Conf. on Computer Vision (ICCV)*, Oct 2017.

[2] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 24(9):1167–1183, September 2002.

[3] Jonathan T Barron and Jitendra Malik. Intrinsic Scene Properties from a A Single RGB-D Image. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 17–24, 2013.

[4] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences (SISS)*, 2:183–202, March 2009.

[5] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic Images in the Wild. *ACM Trans. on Graphics (TOG)*, 33(4), 2014.

[6] Fausto Bernardini, Ioana M. Martin, and Holly E. Rushmeier. High-quality texture reconstruction from multiple scans. *IEEE Trans. on Visualization and Computer Graphics (TVCG)*, 7(4):318–332, 2001.

[7] Adnane Boukhayma, Vagia Tsiminaki, Jean-Sébastien Franco, and Edmond Boyer. Eigen appearance maps of dynamic shapes. In *Proc. Eur. Conf. on Computer Vision (ECCV)*, pages 230–245. Springer, 2016.

[8] Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: a hybrid geometry- and image-based approach. In *Proc. ACM SIGGRAPH*, pages 11–20. ACM, 1996.

[9] M. Eisemann, B. De Decker, M. Magnor, P. Bekaert, E. de Aguiar, N. Ahmed, C. Theobalt, and A. Sellent. Floating Textures. *Computer Graphics Forum*, 27(2): 409–418, April 2008.

[10] Rik Fransens, Christoph Strecha, and Luc Van Gool. Optical flow based super-resolution: A probabilistic approach. *Computer Vision and Image Understanding*, 106 (1):106–115, 2007.

[11] Peter V. Gehler, Carsten Rother, Martin Kiefel, Lumin Zhang, and Bernhard Schölkopf. Recovering intrinsic images with a global sparsity prior on reflectance. In *Proc. Annual Conf. on Neural Information Processing Systems (NIPS)*, pages 765–773, 2011.

[12] Bastian Goldlücke, Mathieu Aubry, Kalin Kolev, and Daniel Cremers. A super-resolution framework for high-accuracy multiview reconstruction. *Intl. J. of Computer Vision (IJCV)*, 106(2):172–191, 2014.

[13] Bastian Goldlüecke and Daniel Cremers. A superresolution framework for high-accuracy multiview reconstruction. In *Pattern Recognition (Proc. DAGM)*, 2009.

[14] Bastian Goldlüecke and Daniel Cremers. Superresolution texture maps for multi-view reconstruction. *Proc. Intl. Conf. on Computer Vision (ICCV)*, pages 1677–1684, September 2009.

[15] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *Proc. Intl. Conf. on Computer Vision (ICCV)*, pages 2335–2342, 2009.

[16] Bjoern Haefner, Yvain Quéau, Thomas Möllenhoff, and Daniel Cremers. Fight ill-posedness with ill-posedness: Single-shot variational depth super-resolution from shading. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 164–174, 2018.

[17] Shahram Izadi, Richard A. Newcombe, David Kim, Otmar Hilliges, David Molyneaux, Steve Hodges, Pushmeet Kohli, Jamie Shotton, Andrew J. Davison, and Andrew Fitzgibbon. Kinectfusion: Real-time dynamic 3d surface reconstruction and interaction. In *Proc. ACM SIGGRAPH*, pages 23:1–23:1. ACM, 2011.

[18] Junho Jeon, Sunghyun Cho, Xin Tong, and Seungyong Lee. Intrinsic image decomposition using structure-texture separation and surface normals. In *Proc. Eur. Conf. on Computer Vision (ECCV)*, pages 218–233. Springer, 2014.

[19] Naejin Kong, Peter V Gehler, and Michael J Black. Intrinsic video. In *Proc. Eur. Conf. on Computer Vision (ECCV)*, pages 360–375, 2014.

[20] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[21] Kyong Joon Lee, Qi Zhao, Xin Tong, Minmin Gong, Shahram Izadi, Sang Uk Lee, Ping Tan, and Stephen Lin. Estimation of intrinsic image sequences from image+depth video. In *Proc. Eur. Conf. on Computer Vision (ECCV)*, pages 327–340, 2012.

[22] Victor S. Lempitsky and Denis V. Ivanov. Seamless mosaicing of image-based texture maps. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, 2007.

[23] Hendrik P. A. Lensch, Wolfgang Heidrich, and Hans-Peter Seidel. A silhouette-based algorithm for texture registration and stitching. *Graphical Models*, 63(4):245–262, 2001.

[24] Andreas Ley, Ronny Hänsch, and Olaf Hellwich. Syb3r: A realistic synthetic benchmark for 3d reconstruction from images. In *Proc. Eur. Conf. on Computer Vision (ECCV)*, pages 236–251. Springer, 2016.

[25] Ce Liu and Deqing Sun. A Bayesian approach to adaptive video super resolution. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 209–216. IEEE, June 2011.

[26] Qi Liu-Yin, Rui Yu, Lourdes Agapito, Andrew Fitzgibbon, and Chris Russell. Better together: Joint reasoning for non-rigid 3d reconstruction with specularities and shading. In *Proc. British Machine Vision Conf. (BMVC)*, 2016.

[27] Robert Maier, Jörg Stückler, and Daniel Cremers. Super-resolution keyframe fusion for 3d modeling with high-quality textures. In *Proc. Intl. Conf. on 3D Vision (3DV)*, pages 536–544, 2015.

[28] Robert Maier, Kihwan Kim, Daniel Cremers, Jan Kautz, and Matthias Nießner. Intrinsic3D: High-Quality 3D Reconstruction by Joint Appearance and Geometry Optimization with Spatially-Varying Lighting. In *Proc. Intl. Conf. on Computer Vision (ICCV)*, volume 4, 2017.

[29] Abhimitra Meka, Michael Zollhöfer, Christian Richardt, and Christian Theobalt. Live intrinsic video. *ACM Trans. on Graphics (TOG)*, 35(4):109, 2016.

[30] Jean Mélou, Yvain Quéau, Jean-Denis Durou, Fabien Castan, and Daniel Cremers. Variational reflectance estimation from multi-view images. *CoRR*, abs/1709.08378, 2017.

[31] Jean Mélou, Yvain Quéau, Jean-Denis Durou, Fabien Castan, and Daniel Cremers. Beyond multi-view stereo: Shading-reflectance decomposition. In *Scale Space and Variational Methods in Computer Vision*, pages 694–705, 2017.

[32] Dennis Mitzel, Thomas Pock, Thomas Schoenemann, and Daniel Cremers. Video super resolution using duality based TV-L1 optical flow. In *Pattern Recognition (Proc. DAGM)*, pages 432–441, 2009.

[33] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proc. ACM SIGGRAPH*, pages 497–500, 2001.

[34] Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. Photorealistic facial texture inference using deep neural networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[35] Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch. Enhancenet: Single image super-resolution through automated texture synthesis. In *Proc. Intl. Conf. on Computer Vision (ICCV)*, 2017.

[36] Takeshi Takai, Adrian Hilton, and Takashi Mastuyama. Harmonised texture mapping. In *Proc. Intl. Conf. on 3D Vision (3DV)*, 2010.

[37] Christian Theobalt, Naveed Ahmed, Hendrik P. A. Lensch, Marcus A. Magnor, and Hans-Peter Seidel. Seeing people in different light-joint shape, motion, and reflectance capture. *IEEE Trans. on Visualization and Computer Graphics (TVCG)*, 13(4):663–674, 2007.

[38] Vagia Tsiminaki, Jean-Sébastien Franco, and Edmond Boyer. High resolution 3d shape texture from multiple videos. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1502–1509, 2014.

[39] Tony Tung. Simultaneous super-resolution and 3D video using graph-cuts. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, June 2008.

[40] Michael Waechter, Nils Moehrle, and Michael Goesele. Let there be color! large-scale texturing of 3d reconstructions. In *Proc. Eur. Conf. on Computer Vision (ECCV)*, pages 836–850, 2014.

[41] Yair Weiss. Deriving Intrinsic Images from Image Sequences. In *Proc. Intl. Conf. on Computer Vision (ICCV)*, volume 2, pages 68–75, 2001.

[42] Qi Zhao, Ping Tan, Qiang Dai, Li Shen, Enhua Wu, and Stephen Lin. A Closed-form Solution to Retinex with Nonlocal Texture Constraints. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 34(7):1437–1444, 2012.

[43] Qian-Yi Zhou and Vladlen Koltun. Color map optimization for 3d reconstruction with consumer depth cameras. *ACM Trans. on Graphics (TOG)*, 33(4):155:1–155:10, 2014.

[44] Michael Zollhöfer, Angela Dai, Matthias Innmann, Chenglei Wu, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Shading-based Refinement on Volumetric Signed Distance Functions. *ACM Trans. on Graphics (TOG)*, 34(4):96, 2015.