Intrinsic3D: High-Quality 3D Reconstruction by Joint Appearance and Geometry Optimization with Spatially-Varying Lighting

Supplementary Material

Robert Maier[1,2] Kihwan Kim[1] Daniel Cremers[2] Jan Kautz[1] Matthias Nießner[2,3]
[1]NVIDIA [2]Technical University of Munich [3]Stanford University

In this document, we provide additional experiments and details. Specifically, we give an overview of the mathematical symbols in Sec. 1, and in Sec. 2 we provide a thorough quantitative evaluation regarding the geometric reconstruction quality on ground truth data (both real and synthetic). We further show qualitative results of the reconstructed models on several own and publicly-available datasets, with a focus on both reconstruction geometry and appearance; see Sec. 3. Finally, in Sec. 4, we detail additional experiments on spatially-varying lighting under both qualitative and quantitative standpoints.

1. List of Mathematical Symbols

Symbol	Description
$\boldsymbol{p}$	continuous 3D point in $\mathbb{R}^3$
$\boldsymbol{x}$	continuous 2D image point in $\mathbb{R}^2$
$\boldsymbol{v}$	position of voxel in $\mathbb{R}^3$
$\boldsymbol{v}_c$	position of voxel center of $\boldsymbol{v}$ in $\mathbb{R}^3$
$\boldsymbol{v}_0$	position of $\boldsymbol{v}$ transformed onto iso-surface in $\mathbb{R}^3$
$\mathbf{n}(\boldsymbol{v})$	surface normal at $\boldsymbol{v}$ in $\mathbb{R}^3$
$\mathbf{D}(\boldsymbol{v})$	signed distance value at $\boldsymbol{v}$
$\mathbf{C}(\boldsymbol{v}), \mathbf{I}(\boldsymbol{v})$	color (RGB) and intensity at $\boldsymbol{v}$
$\mathbf{W}(\boldsymbol{v})$	integration weight at $\boldsymbol{v}$
$\mathbf{a}(\boldsymbol{v})$	albedo at $\boldsymbol{v}$
$\tilde{\mathbf{D}}(\boldsymbol{v})$	refined signed distance value at $\boldsymbol{v}$
$\mathbf{D}_0$	iso-surface of the refined SDF
$\mathbf{B}(\boldsymbol{v})$	estimated reflected shading at $\boldsymbol{v}$
$\mathbf{\Gamma}(\boldsymbol{v})$	chromaticity at $\boldsymbol{v}$
t_{shell}	thin shell size
N	number of voxels inside the thin shell region
K, t_{sv}	number of subvolumes and subvolume size in $\mathbb{R}^3$
$\mathcal{S}$	set of subvolumes s_k
ℓ	vector of all lighting coefficients l_m
H_m	m-th spherical harmonics basis
b	number of spherical harmonics bands
M	number of input frames
$\mathcal{C}_i, \mathcal{I}_i, \mathcal{Z}_i$	color, intensity and depth image of frame i
$\mathcal{T}_i$	transformation from frame i to the base frame
t_{KF}	keyframe selection window size
$t_{\text{best}}, \mathcal{V}_{\text{best}}$	number of best views for $\boldsymbol{v}$ and corresponding set
$d_i(\boldsymbol{v})$	projective distance to voxel center in frame i
$w_i(\boldsymbol{v})$	sample integration weight of frame i
$\mathcal{O}_{\boldsymbol{v}}$	set of color observations of $\boldsymbol{v}$
$c_i^{\boldsymbol{v}}$	observed color of $\boldsymbol{v}$ in frame i
$w_i^{\boldsymbol{v}}$	observation weight of $\boldsymbol{v}$ in frame i
f_x, f_y, c_x, c_y	camera intrinsics (focal length, optical center)
$\kappa_1, \kappa_2, \rho_1$	radial and tangential lens distortion parameters
$\mathcal{X}$	stacked vector of optimization variables

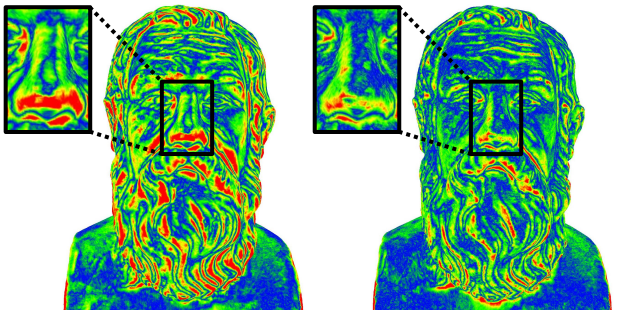


Figure 1. Surface accuracy comparison with a ground truth laser scan of the *Socrates* dataset: the approach of Zollhöfer et al. [3] (left) exhibits a higher mean absolute deviation from the ground truth compared to our method (right).

2. Quantitative Geometry Evaluation

In the following, we show a quantitative surface accuracy evaluation of our geometry refinement on the *Socrates* and *Frog* datasets.

2.1. Socrates

In order to measure the surface accuracy of our method quantitatively, we first compare our method with a ground truth laser scan of the *Socrates* Multi-View Stereo dataset from [3]. The mean absolute deviation (MAD) between our reconstruction and the laser scan is 1.09mm (with a standard deviation of 2.55mm), while the publicly-available refined 3D model of Zollhöfer et al. [3] has a significantly higher mean absolute deviation of 1.80mm (with a standard deviation of 3.35mm). This corresponds to an accuracy improvement of 39.44% of our method. Figure 1 visualizes the color-coded mean absolute deviation on the surface.

2.2. Frog

Besides a quantitative comparison with a laser scan, we also evaluate the surface accuracy of a 3D model reconstructed from synthetic RGB-D data. We therefore generated the synthetic *Frog* dataset by rendering a ground truth mesh with a high level of detail into synthetic color and depth images. We smooth the depth maps using a bilateral filter and add Gaussian noise to both the depth values and to the camera poses.

Instead of comparing the reconstructed 3D models directly with the original mesh, we instead fuse the generated noise-free RGB-D frames into a Signed Distance Field and extract a 3D mesh with Marching Cubes [1]. This extracted mesh is then used as ground truth reference and represents the best possible reconstruction given the raycasted input data in combination with an SDF volume representation.

The mean absolute deviation between our reconstruction and the ground truth mesh is 0.222mm (with a standard deviation of 0.269mm). With the reconstruction generated using our implementation of [3], we obtain a substantially higher mean absolute deviation of 0.278mm (with a standard deviation of 0.299mm). Compared to [3], our method improves the reconstruction accuracy by 20.14% and is able to reveal geometric details lost with [3]. Figure 2 visualizes the color-coded mean absolute deviation on the surface.

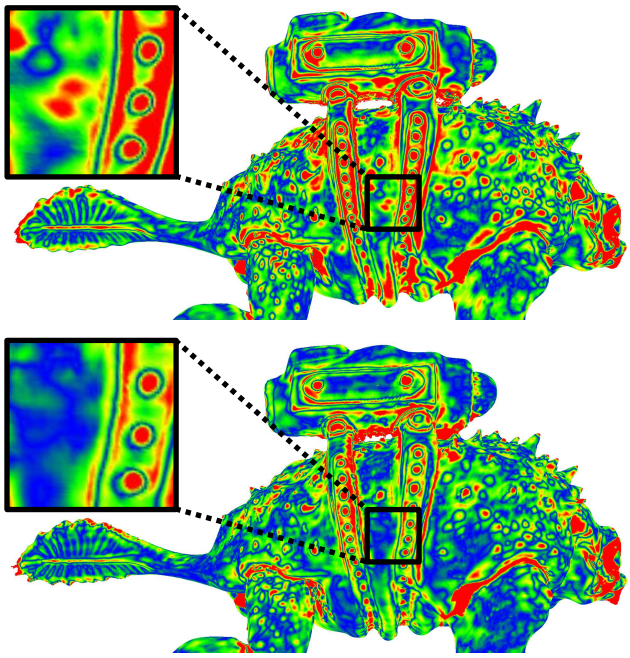


Figure 2. Surface accuracy comparison on synthetic data with a ground truth mesh of the *Frog* dataset: our method (bottom) generates more accurate results compared to Zollhöfer et al. [3] (top).

3. Examples of 3D Reconstructions

In addition to providing a thorough quantitative ground truth evaluation, we show qualitative results of 3D models reconstructed from several RGB-D datasets. In particular, we present 3D reconstructions of the publicly-available *Relief* and *Lucy* datasets from Zollhöfer et al. [3] as well as 3D models of the *Gate*, *Lion*, *Hieroglyphics*, *Tomb Statuary* and *Bricks* datasets that we acquired with a Structure Sensor.

Apart from showing the fine detailed geometry, we also demonstrate the improved appearance of the reconstructions, which we implicitly obtain by jointly optimizing for surface, albedo, and image formation model parameters within our approach.

3.1. Relief

In Figure 3, we show a comparison of the appearance generated using our method with simple volumetric fusion (e.g., Voxel Hashing [2]) and the shading-based surface refinement approach by Zollhöfer et al. [3]. The results in (a) and (b) are visualizations from the meshes that are publicly-available on the project website of [3]. The close-ups suc-

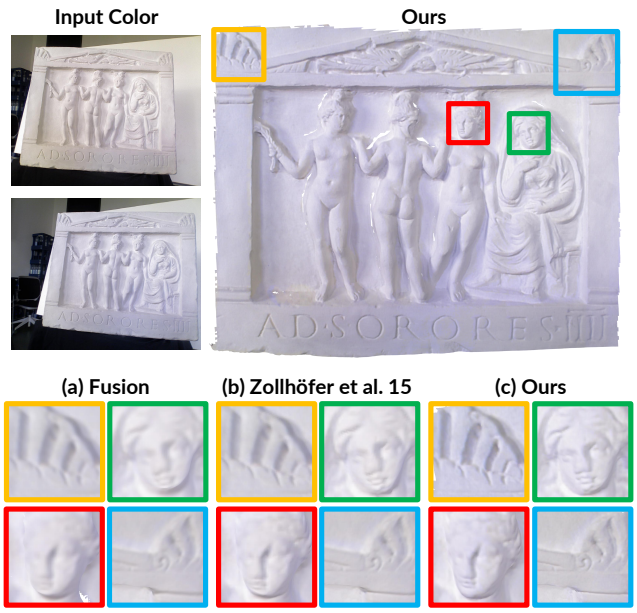


Figure 3. Refined appearance of *Relief* dataset: our method (c) reconstructs significantly sharper textures compared to (a) and (b). Close-ups of ornaments (yellow, blue) and figures (green, red) exhibit more visual details.

cessfully visualize that our method results in significantly sharper textures.

3.2. Lucy

In Figure 4, we present a visual comparison of the reconstructed surface geometry of the *Lucy* dataset. Note how volumetric fusion (a) and Zollhöfer et al. [3] (b) cannot reveal fine-scale details due to the use of averaged per-voxel colors for the refinement, while our method gives the best results and provides geometric consistency (c).

Regarding appearance, we can observe in Figure 5 that our method (c) provides a more detailed texture compared to fusion (a) and Zollhöfer et al. [3] (b).

3.3. Additional Datasets

While the *Relief* and *Lucy* datasets provided by [3] consist of rather small objects with only few input RGB-D frames and short camera trajectories, we acquired more advanced RGB-D datasets using a Structure Sensor.

Figure 6 shows the reconstruction of the *Gate* dataset, while the 3D model of the *Lion* dataset is visualized in Figure 7. The 3D reconstructions of *Hieroglyphics*, *Tomb Statuary* and *Bricks* are presented in Figure 8, Figure 9 and Figure 10 respectively. For all of these datasets, our method generates high-quality 3D reconstructions with fine-scale surface details and and compelling visual appearance with sharp texture details. In contrast, the models obtained from volumetric fusion lack fine details in both geometry and appearance.

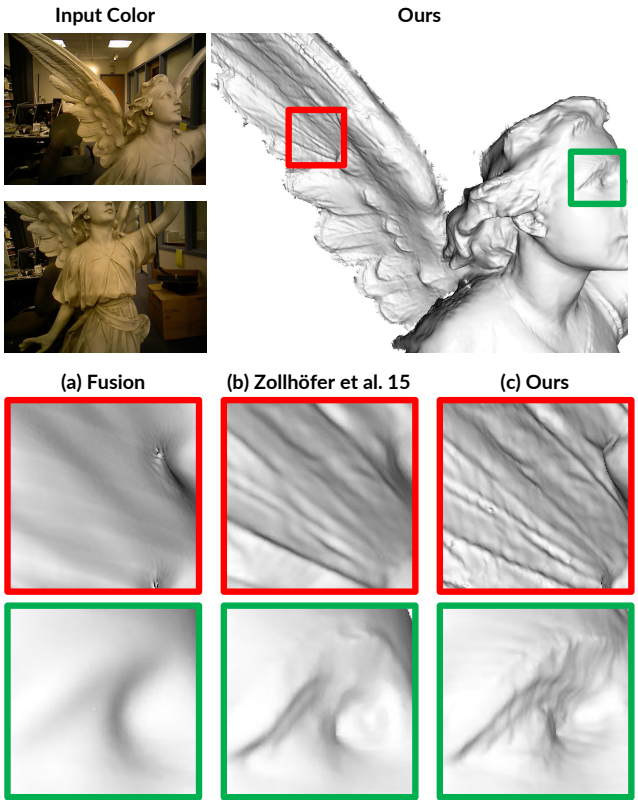


Figure 4. Refined geometry of *Lucy* dataset: volumetric fusion (a) with its strong regularization gives only coarse models. Zollhöfer et al. [3] (b) generate more details; however, limited by using averaged per-voxel colors for the refinement. Our approach that jointly optimizes for all involved parameters (c) reconstructs fine-detailed high-quality geometry.

4. Evaluation of Spatially-Varying Lighting

In this section, we present further qualitative results for lighting estimation via spatially-varying spherical harmonics (SVSH) compared to global spherical harmonics (global SH) on various datasets. We use the same underlying geometry for both variants of lighting estimation for each dataset.

Error Metric As a metric, we use the absolute difference between estimated shading and observed input luminance of a voxel $\boldsymbol{v}$; i.e.,

$$\mathbf{B}_{\text{diff}} = |\mathbf{B}(\boldsymbol{v}) - \mathbf{I}(\boldsymbol{v})|, \tag{1}$$

to determine the quality of the illumination for given geometry and albedo. Ideally, this difference should be as small as possible.

Relief For the *Relief* dataset, the differences between lighting estimation with global SH and SVSH (with a sub-volume size of 0.05m) are shown in Figure 11. It becomes obvious that even for seemingly simple scenes, a single global set of Spherical Harmonics coefficients cannot accu-

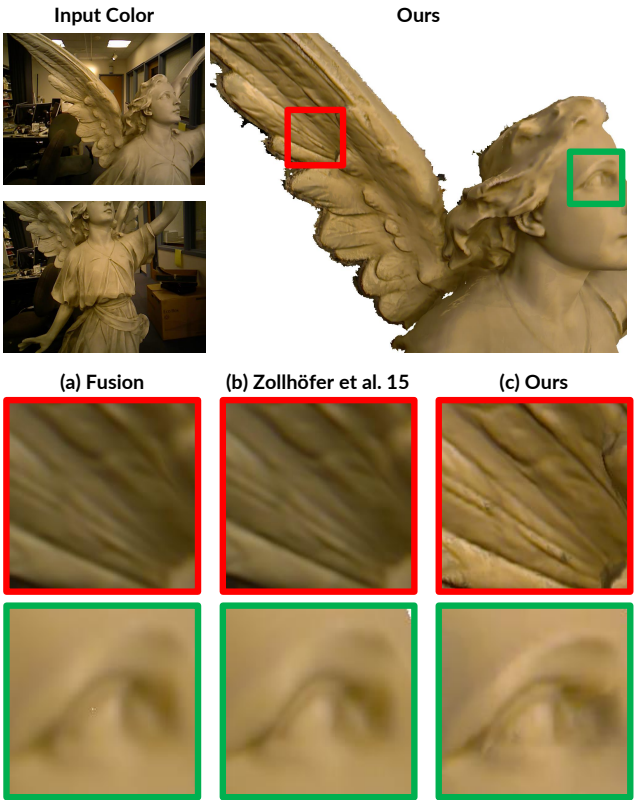


Figure 5. Refined appearance of *Lucy* dataset: in addition to precise geometry our method (c) also produces high-quality colors compared to (a) and (b).

rately reflect real-world environments with complex lighting.

Lucy Similar to the *Relief*, SVSH (with a subvolume size of 0.05m) can better approximate the complex illumination in the *Lucy* dataset than global SH. Figure 12 visualizes the differences in the estimated shadings.

References

[1] W. Lorensen and H. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM Transactions on Graphics (TOG)*, 21(4):163–169, 1987. 2

[2] M. Nießner, M. Zollhöfer, S. Izadi, and M. Stamminger. Real-time 3D reconstruction at scale using voxel hashing. *ACM Transactions on Graphics (TOG)*, 32(6):169, 2013. 2

[3] M. Zollhöfer, A. Dai, M. Innmann, C. Wu, M. Stamminger, C. Theobalt, and M. Nießner. Shading-based refinement on volumetric signed distance functions. *ACM Transactions on Graphics (TOG)*, 34(4):96, 2015. 2, 3

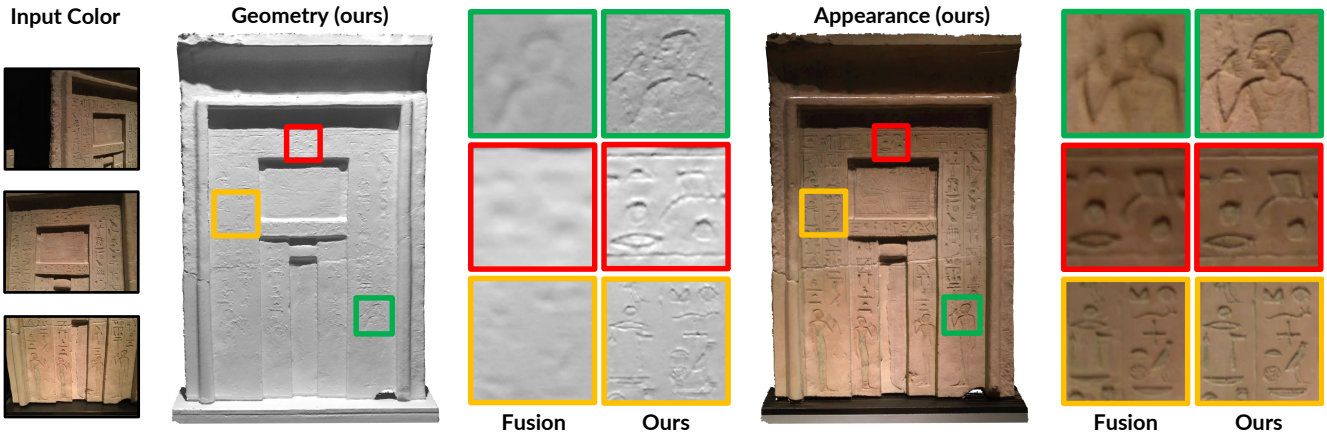


Figure 6. Reconstruction of the *Gate* dataset.

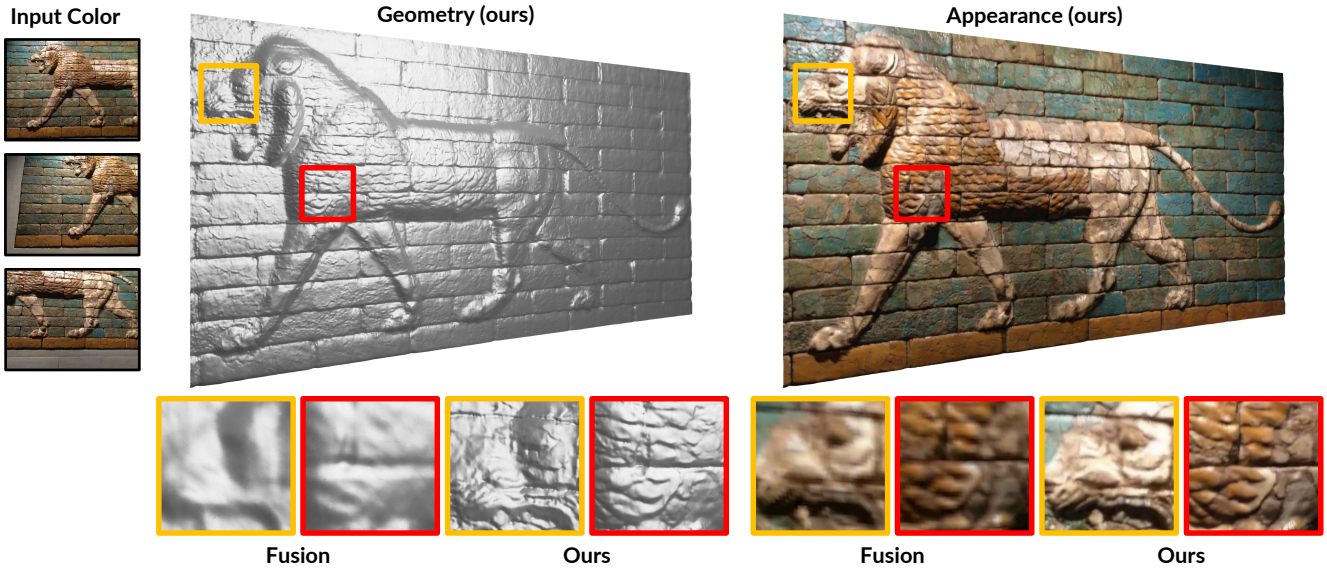


Figure 7. Reconstruction of the *Lion* dataset.

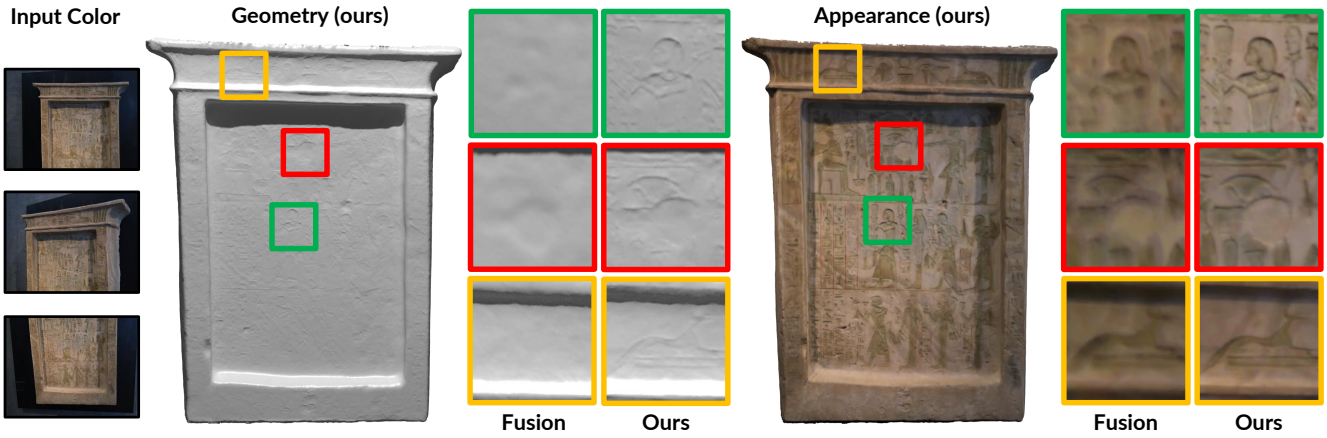


Figure 8. Reconstruction of the *Hieroglyphics* dataset.

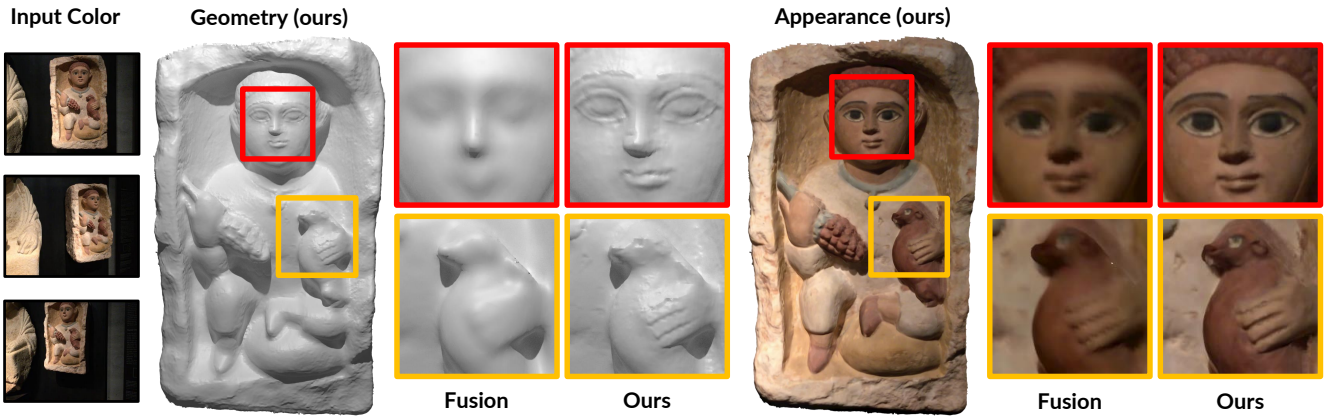


Figure 9. Reconstruction of the *Tomb Statuary* dataset.

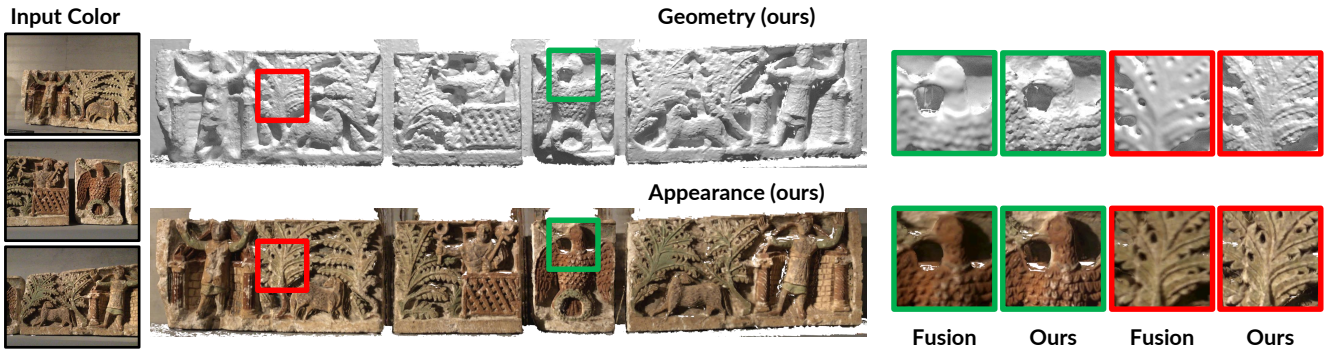
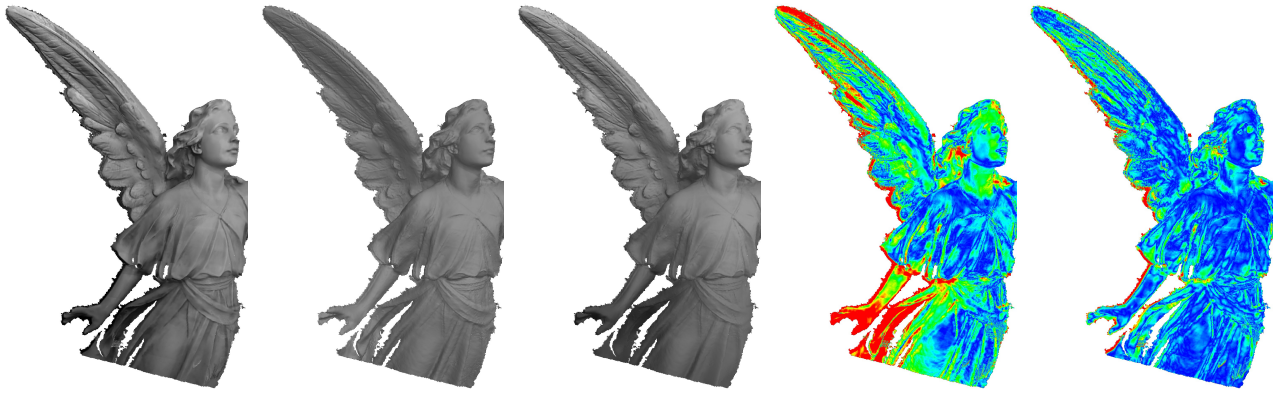


Figure 10. Reconstruction of the *Bricks* dataset.

(a) Luminance (b) Shading (global SH) (c) Shading (SVSH) (d) Difference (global SH) (e) Difference (SVSH)

Figure 11. Estimated illumination of *Relief* dataset: the differences between input luminance (a) and estimated shading (b) and (c) are less for SVSH (e) than for global SH (d), meaning a better approximation of the illumination.



(a) Luminance (b) Shading (global SH) (c) Shading (SVSH) (d) Difference (global SH) (e) Difference (SVSH)

Figure 12. Estimated illumination of *Lucy* dataset: illumination with SVSH (c) explains the illumination better than global SH only (b), resulting in less differences (e) compared to (d) between input luminance (a) and shading.