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

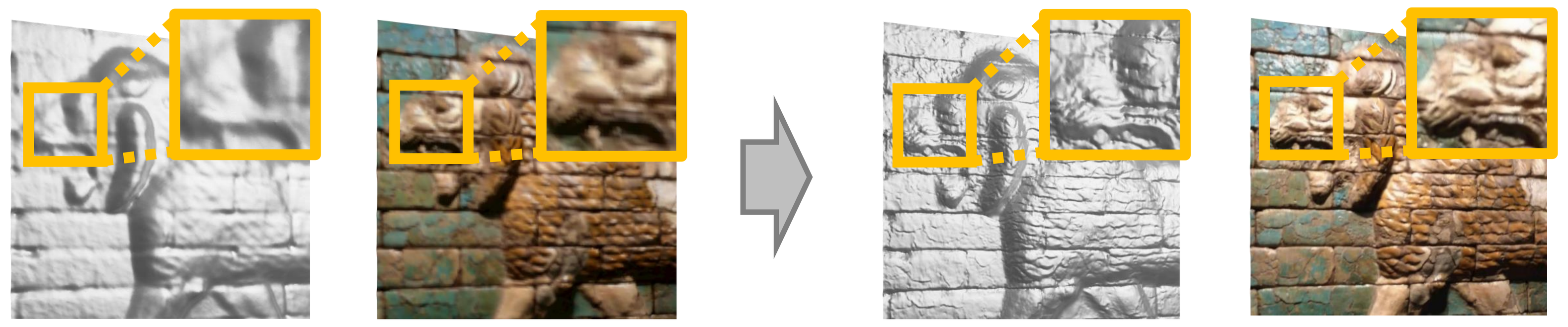
Robert Maier[1,2] Kihwan Kim[1] Daniel Cremers[2] Jan Kautz[1] Matthias Nießner[2,3]

1 NVIDIA 2 Technische Universität München 3 Stanford University

Motivation: RGB-D based 3D Reconstruction

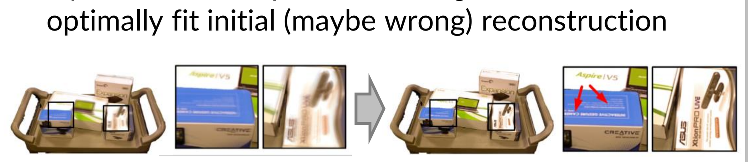
Baseline: **over-smoothed geometry bad colors**

Goal: **high-quality reconstruction of geometry and appearance**



High-Quality Colors (Zhou and Koltun [1])

Optimize camera poses and image deformations to optimally fit initial (maybe wrong) reconstruction



But: no geometry refinement involved!

High-Quality Geometry (Zollhöfer et al. [2])

Adjust camera poses in advance to improve color, use shading cues (RGB) to refine geometry

But: RGB is fixed (no color refinement based on refined geometry)

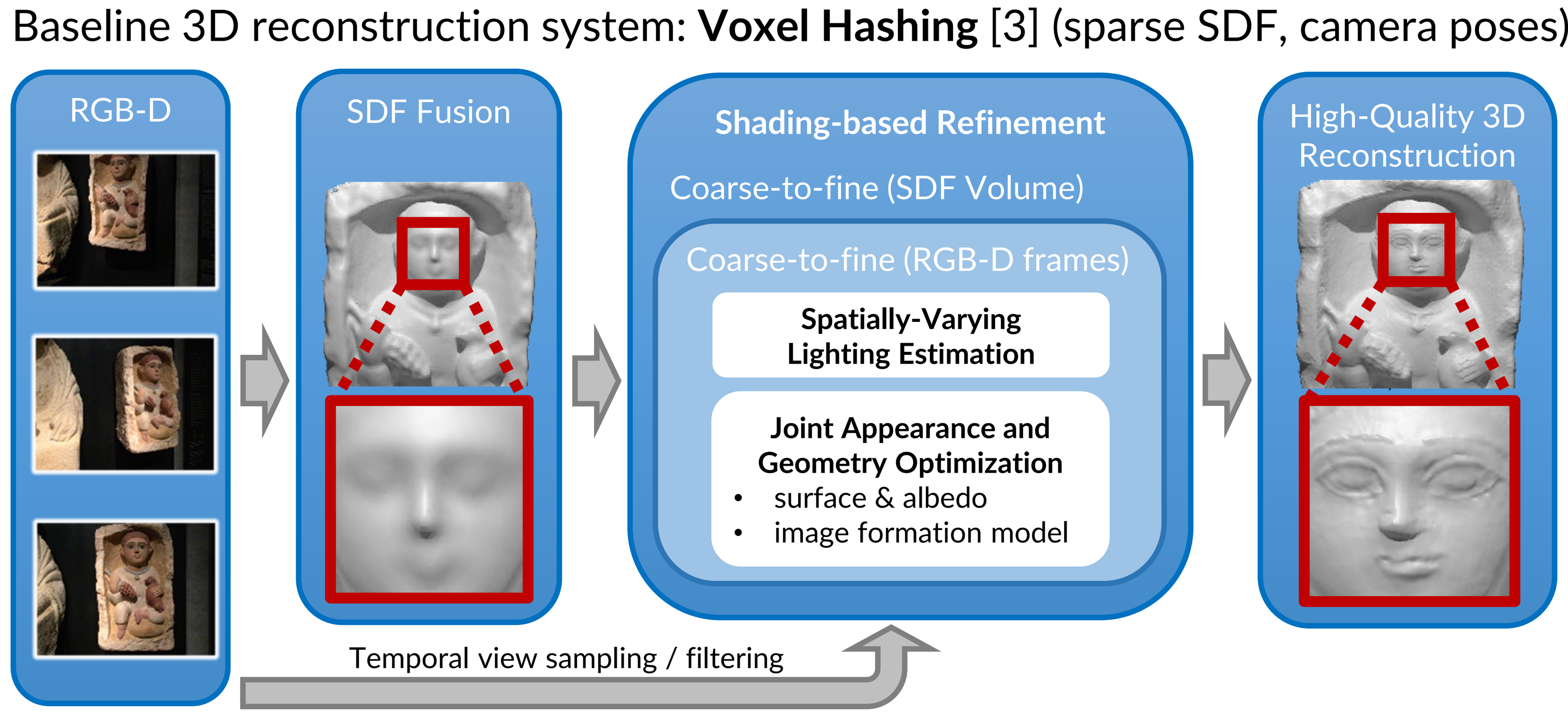
Idea: **jointly optimize for geometry, albedo and image formation model** to simultaneously obtain high-quality geometry and appearance!

Contributions

- Temporal **view sampling & filtering** techniques (input frames)
- **Joint optimization** of
  - **surface & albedo** (Signed Distance Field)
  - **image formation model** (camera poses, camera intrinsics)
- Lighting estimation using **Spatially-Varying Spherical Harmonics** (SVSH)
- **Optimized colors** (by-product)

Overview

Baseline 3D reconstruction system: **Voxel Hashing** [3] (sparse SDF, camera poses)



Sampling & Filtering

- **Keyframe selection**: frame with best blur score [4] within fixed size window
- **Sampling** of voxel observations:
  - **Collect observations** in input keyframes: $c_i^v = \mathcal{C}_i(\pi(\mathcal{T}_i^{-1} v_0))$
  - **View-dependent** observation weights (normal, depth): $w_i^v = \frac{\cos(\theta)}{d^2}$
  - Filtering: keep only **best 5 observations** by weight
- **Colorization** (weighted average): $c_v^* = \arg\min_{c_v} \sum_{(c_i^v, w_i^v) \in \mathcal{O}_v} w_i^v (c_v - c_i^v)^2$

International Conference on Computer Vision (ICCV), Venice, 2017.

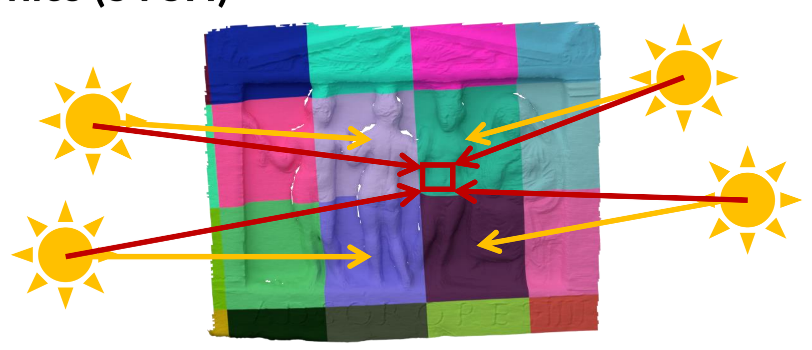
Spatially-Varying Lighting Estimation

Spherical Harmonics (SH)

- Lighting **approximation** using **only 9 SH basis functions** H_m (2nd order)
- **Shortcoming** of single global SH basis: purely directional
  → **cannot represent complex scene lighting** for all surface points simultaneously

Idea: Spatially-varying Spherical Harmonics (SVSH)

- **Partition SDF** volume into subvolumes
- Estimate **independent SH coefficients** for each **subvolume**
- **Per-voxel SH coefficients**: tri-linear interp.



Joint Optimization

Estimate SVSH coefficients for all *K* subvolumes jointly:

$$E_{\text{lighting}}(\boldsymbol{l}_1, \dots, \boldsymbol{l}_K) = E_{\text{appearance}} + \lambda_{\text{diffuse}} E_{\text{diffuse}}$$

Similarity between estimated shading and input luminance

$$\sum_{\boldsymbol{v} \in \mathbf{D}_0} (\mathbf{B}(\boldsymbol{v}) - \mathbf{I}(\boldsymbol{v}))^2$$

Smooth illumination changes (Laplacian regularizer)

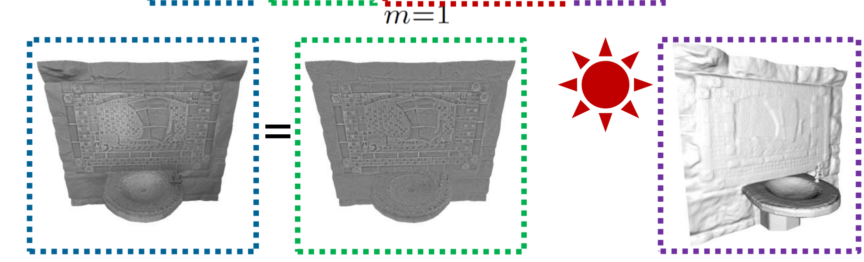
$$\sum_{s \in \mathcal{S}} \sum_{r \in \mathcal{N}_s} (\boldsymbol{l}_s - \boldsymbol{l}_r)^2.$$

Joint Appearance and Geometry Optimization

Shading-based Refinement

Shading equation: Shading albedo b^2 lighting surface normal

$$\mathbf{B}(\boldsymbol{v}) = \mathbf{a}(\boldsymbol{v}) \sum_{m=1}^{b^2} l_m H_m \mathbf{n}(\boldsymbol{v})$$



Intuition: **high-frequency changes** in surface geometry → **shading cues** in input images

1) Estimate **lighting** given surface and albedo (intrinsic material properties)
2) Estimate **surface** and **albedo** given the lighting: minimize difference between estimated shading and input luminance

Shading-based SDF Optimization

Joint optimization of geometry, albedo and image formation model (camera poses/intrinsics):

$$E_{\text{scene}}(\mathcal{X}) = \sum_{\boldsymbol{v} \in \tilde{\mathbf{D}}_0} \lambda_g E_g + \lambda_v E_v + \lambda_s E_s + \lambda_a E_a$$

with $\mathcal{X} = (\mathcal{T}, \tilde{\mathbf{D}}, \mathbf{a}, f_x, f_y, c_x, c_y, \kappa_1, \kappa_2, \rho_1)$

Gradient-based shading constraint E_g

Idea: **maximize consistency** between **estimated voxel shading** and **sampled intensities** in input luminance images (gradient for robustness)

$$E_g(\boldsymbol{v}) = \sum_{\mathcal{I}_i \in \mathcal{V}_{\text{best}}} w_i^{\boldsymbol{v}} \| \nabla \mathbf{B}(\boldsymbol{v}) - \nabla \mathcal{I}_i(\pi(v_i)) \|_2^2$$

Best views for voxel and view-dependent weights

Shading: allows for optimization of surface and albedo

Sampling: allows for optimization of camera poses/intrinsics (voxel center transformed and projected into input view) with $v_i = g(\mathcal{T}_i, \psi(\boldsymbol{v}))$, $\boldsymbol{v}_0 = \psi(\boldsymbol{v}) = \boldsymbol{v}_c - \mathbf{n}(\boldsymbol{v})\tilde{\mathbf{D}}(\boldsymbol{v})$

Volumetric regularizer E_v

Smoothness in distance values (Laplacian)

$$E_v(\boldsymbol{v}) = (\Delta \tilde{\mathbf{D}}(\boldsymbol{v}))^2$$

Surface Stabilization constraint E_s

Stay close to initial distance values

$$E_s(\boldsymbol{v}) = (\tilde{\mathbf{D}}(\boldsymbol{v}) - \mathbf{D}(\boldsymbol{v}))^2$$

Albedo regularizer E_a

Constrain albedo changes based on chromaticity (Laplacian)

$$E_a(\boldsymbol{v}) = \sum_{\boldsymbol{u} \in \mathcal{N}_{\boldsymbol{v}}} \phi(\mathbf{\Gamma}(\boldsymbol{v}) - \mathbf{\Gamma}(\boldsymbol{u})) \cdot (\mathbf{a}(\boldsymbol{v}) - \mathbf{a}(\boldsymbol{u}))^2$$

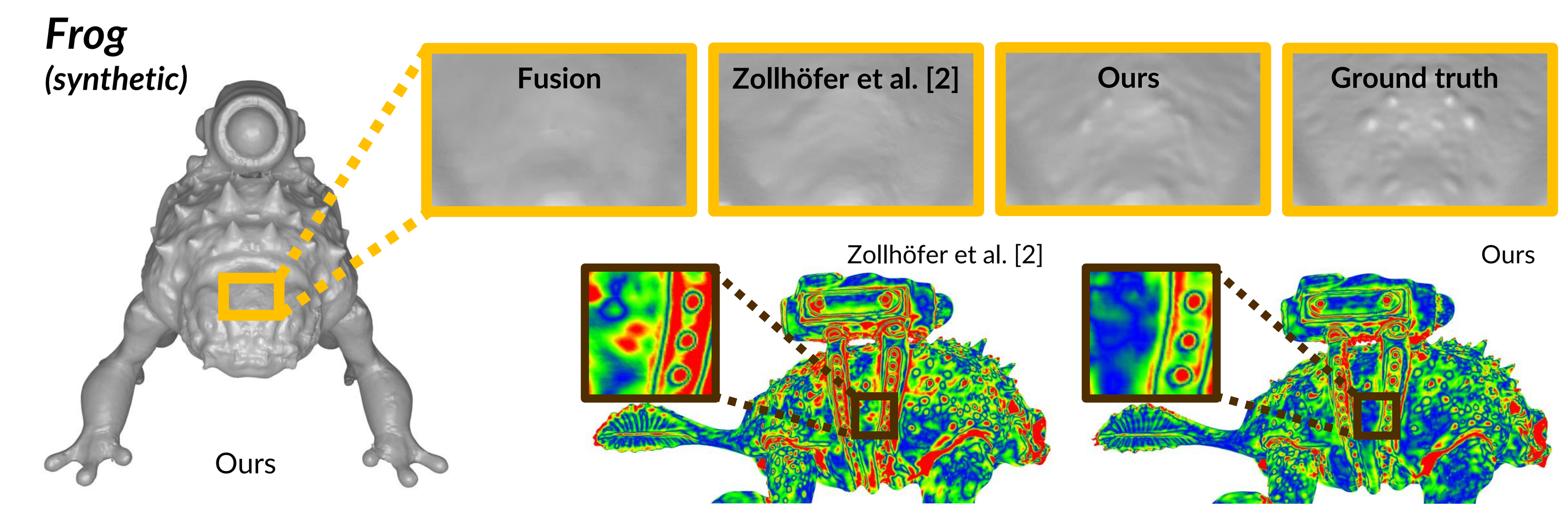
Recolorization

Recompute voxel colors after optimization at each coarse-to-fine level
→ **optimal colors** (due to optimized image formation model)

Results

Quantitative Surface Evaluation

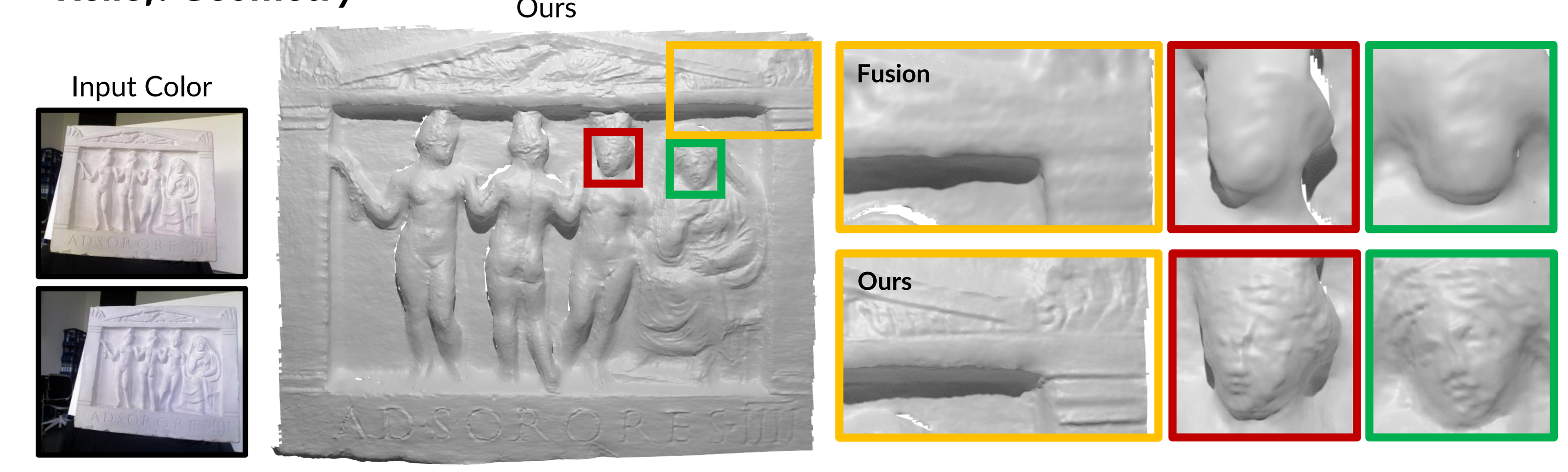
Frog (synthetic)



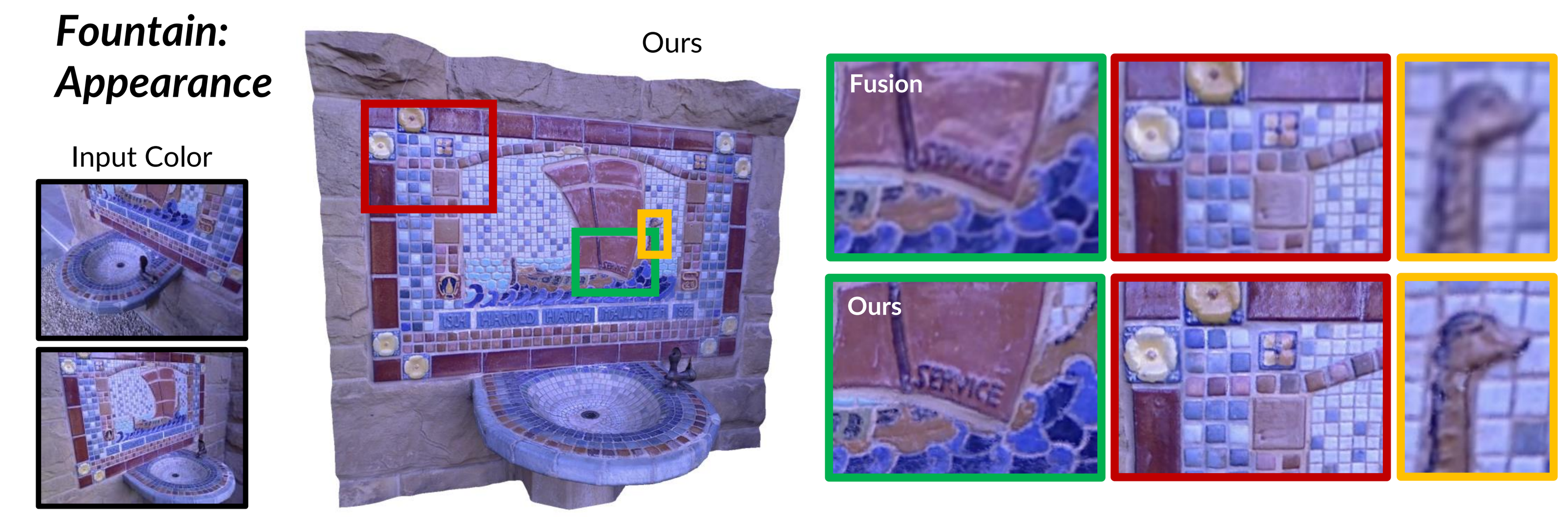
Mean abs. dev.: 0.278mm [2] vs. 0.222mm (ours) → **20.14% more accurate**

Qualitative Results

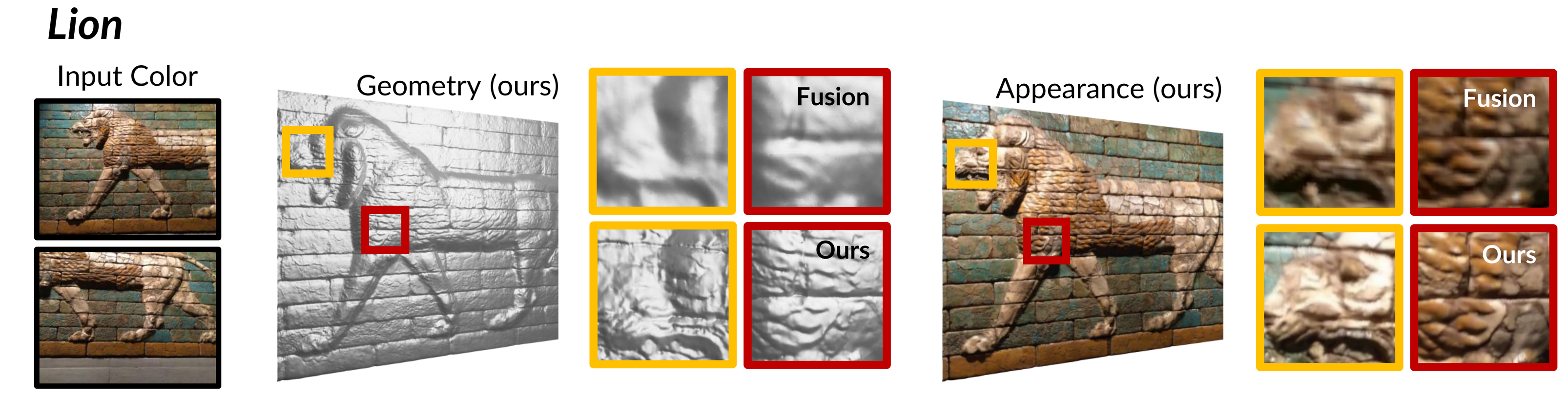
Relief: Geometry



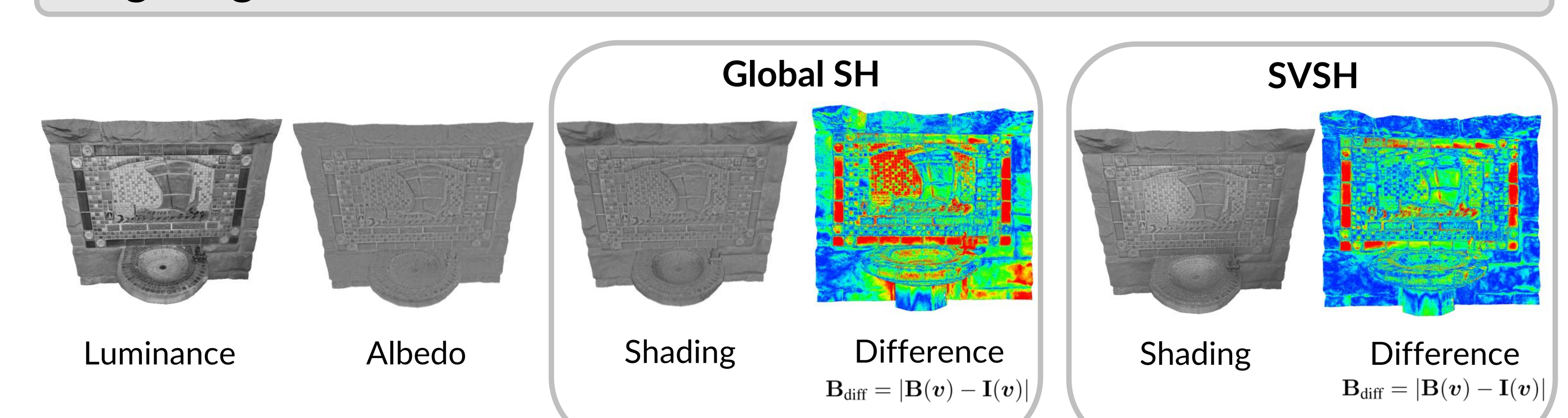
Fountain: Appearance



Lion



Lighting: Global SH vs. SVSH



References

[1] Zhou and Koltun: *Color Map Optimization for 3D Reconstruction with Consumer Depth Cameras.* ToG 2014.
[2] Zollhöfer et al.: *Shading-based Refinement on Volumetric Signed Distance Functions.* ToG 2015.
[3] Nießner et al.: *Real-time 3D Reconstruction at Scale using Voxel Hashing.* ToG 2013.
[4] Crete et al.: *The blur effect: perception and estimation with a new no-reference perceptual blur metric.* SPIE 2007.