

Appearance-Driven Automatic 3D Model Simplification

JON HASSELGREN, NVIDIA, Sweden

JACOB MUNKBERG, NVIDIA, Sweden

JAAKKO LEHTINEN, NVIDIA & Aalto University, Finland

MIIKA AITTALA, NVIDIA, Finland

SAMULI LAINE, NVIDIA, Finland



Fig. 1. We automatically approximate shape and appearance of detailed 3D scenes. Here, we approximate foliage with low-poly proxy geometry and textures, instance it 3000 times, and render the scene in a standalone path tracer. Compared to the reference, we accurately capture the appearance of the scene at just a fraction (0.4%) of the triangle count. The geometry is illustrated in the insets. The assets are from the Moana Island Scene, a publicly available data set courtesy of Walt Disney Animation Studios.

We present a suite of techniques for jointly optimizing triangle meshes and shading models to match the appearance of reference scenes. This capability has a number of uses, including appearance-preserving simplification of extremely complex assets, conversion between rendering systems, and even conversion between geometric scene representations.

We follow and extend the classic analysis-by-synthesis family of techniques: enabled by a highly efficient differentiable renderer and modern nonlinear optimization algorithms, our results are driven to minimize the image-space difference to the target scene when rendered in similar viewing and lighting conditions. As the only signals driving the optimization are differences in rendered images, the approach is highly general and versatile: it easily supports many different forward rendering models such as normal mapping, spatially-varying BRDFs, displacement mapping, etc. Supervision through images only is also key to the ability to easily convert between rendering systems and scene representations.

We output triangle meshes with textured materials to ensure that the models render efficiently on modern graphics hardware and benefit from, e.g., hardware-accelerated rasterization, ray tracing, and filtered texture lookups. Our system is integrated in a small Python code base, and can be applied at high resolutions and on large models. We describe several use cases, including mesh decimation, level of detail generation, seamless mesh filtering and approximations of aggregate geometry.

CCS Concepts: • Computing methodologies → Mesh geometry models; Reflectance modeling.

Additional Key Words and Phrases: Inverse rendering, mesh decimation, level of detail generation, seamless mesh filtering, aggregate geometry

1 INTRODUCTION

Synthesizing images of objects with complex shapes and appearances is a central goal in computer graphics. The problem can be broken down into choosing suitable representations for shape and appearance, modeling the scene according to the chosen representations, and finally, rendering it efficiently.

Creating a shape and appearance model for a particular 3D scene is inherently an inverse problem: we seek a representation that will, once fed through the renderer, result in an image that looks the way we want. Yet, most modeling tools turn the problem around: instead of providing the user with means to specify the image they want, they provide tools for editing the scene representation, leaving the modeler to iteratively proceed toward their goal.

The goal in this work is to automatically find shape and appearance representations that match, when rendered, a reference scene provided by the user. This approach is often called inverse rendering or analysis-by-synthesis. In contrast to algorithms like multi-view stereo that must make do with a small number of reference images, we focus on applications where it is possible to programmatically synthesize reference views of the target scene under arbitrary, controllable viewing and lighting conditions. Within this scope, we present inverse rendering techniques for

- **Geometric simplification (LOD).** Optimizing for the shape of a lower-resolution mesh to combat geometric aliasing and increase rendering efficiency.

- **Joint shape-appearance simplification.** Optimizing the shape and surface appearance model (mesh geometry, displacement maps, normal maps, spatially-varying BRDFs) to mimic the appearance of a more complex asset.
- **Simplification of aggregate geometry.** Dramatically simplifying complex foliage assets with little impact in visual quality. See Figure 1 for an example.
- **Animation.** Joint optimization of shape and skinning weights on reduced geometry to match a target animation.
- **Conversion between rendering systems.** Optimizing the scene representation to match images rendered by an entirely different system.
- **Conversion between shape representations.** Finding a mesh geometry and associated appearance model that captures the appearance of objects given by other shape representations, such as signed distance fields (SDF).

These kind of goals have been previously pursued with specialized algorithms (Section 2) that typically focus on a single task, and consider only specific parts of the object representation. Since our approach is based on inverse rendering and nonlinear optimization it easily generalizes over all the different regimes while allowing joint optimization of all aspects of the representation that affect the final appearance.

In all our applications, the search for the shape and appearance is driven by image-space error. This, combined with performing shape and appearance optimization together, has the significant benefit that the mechanisms of the forward rendering model can each specialize for the effects they capture best, “negotiating” how to achieve the desired outcome together. As an example, this leads to a natural division of labor between the geometry (mesh) and a normal map: geometric detail is allowed to move between the representations by, e.g., locally smoothing a mesh and baking geometric detail into the normal map or other parameters of a physically based shading model [Karis 2013].

Our source code will be publicly available at <https://github.com/NVlabs/nvdiffmodeling>.

2 PREVIOUS WORK

Mesh decimation. For a detailed overview of this mature research topic, we refer the reader to the book by Luebke et al. [2002]. Commonly used algorithms include *vertex decimation* [Schroeder 1997], *vertex clustering* [Low and Tan 1997] and *edge contraction* [Garland and Heckbert 1997]. The error metric is typically geometry based. A notable exception is view-dependent simplification [Luebke and Erikson 1997] which optimizes for silhouette quality.

Lindstrom and Turk present a framework for image-driven simplification [2000]. They use rendered images to decide which portions of a mesh to simplify. Please refer to Corsini et al. [2013] for a survey of perceptual metrics for triangle meshes. Similarly, our objective function is based on visual differences, but we leverage gradient-based optimization through differentiable rendering.

Cohen et al. [1996] introduce *simplification envelopes* for generating a hierarchy of LODs for polygonal meshes. They build envelopes around the mesh to avoid self-intersections and can guarantee a distance tolerance between the original and simplified meshes. For our

continuous level of detail application, we similarly use a sequence of meshes to represent the LODs. In contrast, we optimize for visual error in image space.

Cook et al. [2007] introduce a decimation technique that stochastically removes a subset of the geometric elements and alter the remaining elements by, e.g., scale and contrast adjustments, in order to preserve the overall appearance. This technique is particularly suited for unstructured objects such as foliage. We similarly exploit the scene graph when approximating aggregate geometry, but instead of using heuristics for preserving visual appearance, we optimize for it directly.

Appearance prefiltering. For a summary of surface appearance prefiltering techniques, please refer to Bruneton et al. [2012]. Linear texture prefiltering methods are incorrect when applied to spatially varying BRDF (*bidirectional reflectance distribution function*) and surface normal maps. Common approaches instead filter the *normal distribution function* (NDF) [Fournier 1992; Han et al. 2007; Olano and Baker 2010; Toksvig 2005], where the challenge is how to compactly represent the filtered NDF. Common representations include Gaussians [Toksvig 2005] with one or more lobes, moment statistics [Olano and Baker 2010], and spherical harmonics [Han et al. 2007].

Further work has considered, e.g., displacement mapping, and masking and shadowing effects [Dupuy et al. 2013; Wu et al. 2019]. Loubet and Neyret [2017] prefilter meshes and materials by absorbing fine details into a volumetric microflake representation. Zhao et al. [2016] downsample volumetric materials by optimizing for rendered similarity to the original.

Our work can be applied to prefiltering. In our experiments, we restrict the shading model to one diffuse lobe and one specular GGX lobe [Walter et al. 2007], and let optimization adjust the mesh shape and the material parameters so that the rendered result matches a highly supersampled reference. This model is indeed less expressive than many of those in prefiltering literature, but has the benefit that there is no change to a typical game engine or any runtime overhead. The approach is flexible, as it treats any target surface and material representation in a unified manner: only the final visual appearance is observed, and it does not matter what combination of, e.g., mesh shape, displacement, normal, and material parameters produced it.

Appearance and geometry capture. Appearance capture can be framed as seeking a digital asset (e.g., an SVBRDF map and a mesh) whose renderings visually match some real-world object. This is conceptually similar to our setup, with the exception that our targets are other digital assets.

Much of the difficulty in appearance capture originates from the desire to limit the acquisition effort for the user. Exhaustively measuring real-world appearance under all lighting and view directions is prohibitively expensive for most purposes. When only a sparse sampling is available, typical approaches employ multi-stage processing involving, e.g., clustering [Lensch et al. 2003] or multi-view stereo [Nam et al. 2018] to find a solution that reproduces the measurements and generalizes well to unseen conditions. Many approaches use special viewing configurations and lighting patterns [Gardner et al. 2003; Ghosh et al. 2009] to recover more

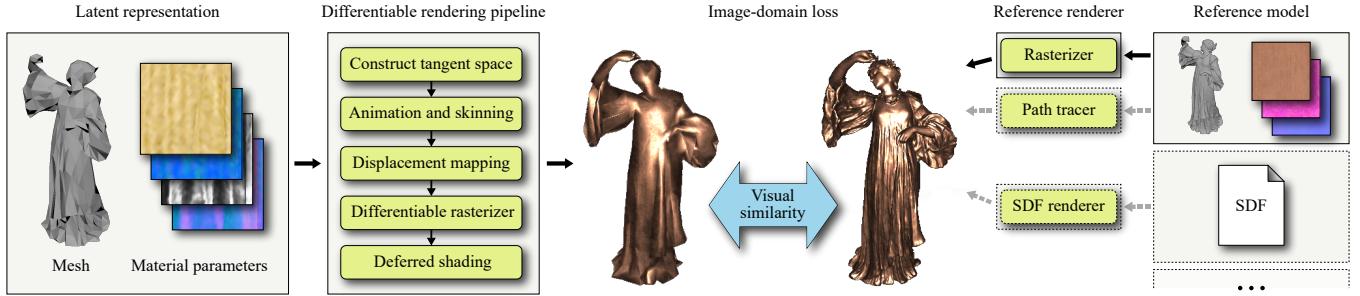


Fig. 2. An overview of our method. We render the latent representation (mesh and material parameters) in a differentiable rendering pipeline: a sequence of mesh operations followed by rasterization and deferred shading. Similarly, a target image is generated by the reference renderer under identical viewing and lighting conditions, and an image-domain loss is computed on the two images. During optimization, we iterate over a large number of image pairs with random viewing and lighting conditions. Using back-propagation and stochastic gradient descent, the latent representation is gradually morphed to produce images close to the reference.

varied reflectance information in each measurement. For recent surveys, see Dong [2019], Guarnera et al. [2016], and Weinmann and Klein [2015].

We are free to render our targets in as many viewing and lighting conditions as needed. This eliminates much of the complexity, and allows us to directly end-to-end optimize over the material parameters and vertex positions using a visual similarity loss.

Many appearance capture methods can be viewed as using simple differentiable renderers to match their predictions to the observations. For example, Gao et al. [2019] and Guo et al. [2020] optimize for SVBRDF maps that reproduce a handful of target photographs upon rendering. They assume planar geometry, and use neural network based regularization to encourage plausible generalization to unseen conditions.

Sztrajman et al. [2017] convert materials between different analytic BRDF representations by optimizing for their rendered visual similarity. Similar to us, their target images do not originate from the real world, but from renderings of the target asset. However, they do not consider geometry as part of the optimization.

Scene acquisition with neural networks. NeRF [Mildenhall et al. 2020] represents the scene as a neural radiance field. The quality of the reconstructed views are impressive, but not yet suitable for interactive applications as the radiance field needs to be densely sampled at inference time, and the generated radiance field is static. Similar to our approach, they use many observations of the scene to train the model. In contrast, we generate triangle meshes and materials, which render in real-time in a standard 3D engine.

Thies et al. [2019] achieve a similar goal by learning a latent texture map and an associated image-space decoder CNN that allows high-quality view interpolation on meshes reconstructed by multi-view stereo in fixed lighting environments.

BSP-Net [Chen et al. 2020] generates compact meshes via binary space partitioning but does not consider materials or fine detail such as normal maps.

Pixel2Mesh [Wang et al. 2018] generates a triangular mesh from a single color image by progressively deforming an ellipsoid. They use a graph convolutional neural network to represent the mesh, and a set of losses to ensure that the mesh is well-formed, including

a Laplacian regularization term. Our shape optimization is also based on deforming an existing triangular mesh, but we rely on multiple image observations to allow higher-quality reconstruction that includes materials. In addition, our approach is different in that we seek a representation directly, instead of training a neural network that would perform the conversion.

Differentiable rendering. For an overview of current approaches, please refer to Laine et al. [2020]. They also introduce a flexible set of differentiable rasterization primitives which can be used together with PyTorch or TensorFlow to build custom differentiable rendering pipelines. Several differentiable rendering packages [Jatavallabhula et al. 2019; Ravi et al. 2020; Valentin et al. 2019] provide more built-in functionality but are less customizable.

Chen et al. [2019; 2020] optimize over vertex positions, colors, normals, light directions and texture coordinates through a variety of lighting models, purely from 2D observations. We extend this approach to handle more complex shading models (GGX), more complex geometry, normal maps, displacement maps, and transparency, for higher visual quality. Additionally, we optimize shape and appearance jointly instead of training in stages, and show additional use cases including mesh filtering and animation. Chen et al. focus primarily on predicting 3D objects from single images, while we use multiple image observations to capture the shape and appearance, which leads to higher-quality reconstructions.

Li et al. [2018] present a differentiable Monte Carlo ray tracer, and Mitsuba 2 [Nimier-David et al. 2019] implements a full differentiable path tracer. In this paper, we use and extend the rasterization primitives from Laine et al. [2020] for quicker iteration times and ease of customization. Combining our approach with a differential path tracer would enable additional applications, but at a significantly higher computational cost. Given recent progress in differentiable rendering performance [Nimier-David et al. 2020], we hope to leverage a differentiable path tracer in future work.

3 OUR METHOD

Our goal is to jointly optimize shape and material parameters to match the visual appearance of images from a reference renderer. We follow the common practice of representing the shape as a triangle

mesh and using a spatially varying BRDF for materials. This ensures that our optimized representation renders directly in modern game engines and can readily exploit hardware-accelerated rasterization, ray tracing, and filtered texture-lookups.

Figure 2 outlines our method. The *latent representation* consists of a triangle mesh and a set of textures describing spatially varying material parameters from a physically based shading model, its exact form varies slightly between applications. During optimization, we render the latent representation using a differentiable rendering pipeline: a sequence of mesh operations, a rasterizer, and a deferred shading stage. An image-space loss is then computed between the resulting image and a target image produced by a reference renderer under identical viewing and lighting conditions. Because the rendering pipeline is fully differentiable, we can compute the gradient of the loss with respect to parameters of the latent representation, i.e., vertex positions and texture contents, and consequently optimize these to improve the visual similarity.

We iterate over a large number of image pairs with randomized camera and a single randomized point light, similar to a virtual photo-goniometer. Using stochastic gradient descent, the latent representation is gradually morphed to match the appearance of the reference model.

More formally, let θ denote the parameters of our latent representation (e.g., mesh vertex positions and spatially varying material parameters). The rendered image $I_\theta(c, l)$ is a function of θ , camera, c , and light, l . The reference render is another function $I_{\text{ref}}(c, l)$, parameterized by the camera and light. Given an image space loss function L , we minimize the empirical risk

$$\operatorname{argmin}_\theta \mathbb{E}_{c,l} [L(I_\theta(c, l), I_{\text{ref}}(c, l))] \quad (1)$$

using stochastic gradient descent based on gradients w.r.t. the latent parameters, $\partial L / \partial \theta$, which are obtained through differentiable rendering.

Apart from the ability to place the camera and light, we consider the reference renderer as a black box. The only information communicated between the reference renderer and our latent representation are the target images that are used in the image-domain loss. Note, in particular, that this means that the reference renderer does not need to be differentiable, or even implemented in the same framework. This allows us to convert models across renderers, and even between different geometrical representations—e.g., from signed distance fields (SDF) to triangle meshes.

Note that the choice of reference rendering algorithm depends on the intended use of the final model. Our pipeline does not render shadows or other global effects, so if the reference renderings feature no such effects either—as can usually be arranged—the optimization will converge on materials that are as close as possible to the reference model. This is because both sides of the process in Figure 2 agree upon which effects in the final image are due to the model and which are due to the rendering algorithm. How the model is ultimately used in production is an orthogonal question,

We use the word “latent” with care; though our representation and its decoder (the renderer) are fixed-function and completely interpretable, the representation is unobserved, i.e., latent, and we seek to infer it from the input images. This choice of word encompasses potential future extensions with learnable latent representations and decoders in a natural way.

and at that point, shadows or path tracing can be enabled if desired. Alternatively, if the references are rendered with, say, ambient occlusion or path tracing enabled, the optimization process will bake these into the material parameters so that rendering *without* these effects produces a reasonable approximation. We demonstrate both approaches later in Sections 4.4 and 4.5.

As often in nonlinear optimization, good initial guesses may have a dramatic effect on the speed of convergence and eventual quality of the result. When a high-resolution mesh of the target object is available, we use off-the-shelf mesh decimation tools to produce the initial guess for the latent triangle mesh. In some cases, e.g., when baking foliage as billboard clouds, we draw on prior domain knowledge and explicitly specify a suitable initial mesh. However, we find that starting from a tessellated sphere often yields surprisingly good results. Similarly, if available, we may use texture maps from the reference scene as an initial guess for material parameters. If unavailable, we start from randomly initialized texture maps. We currently do not optimize the topology or texture coordinates of the latent representation. Hence, we require the initial mesh to have a reasonable level of tessellation and non-overlapping texture parameterization.

Our rendering pipeline combines the differentiable rasterization primitives by Laine et al. [2020] with mesh operations and deferred shading in PyTorch [2017], a modern autodifferentiation framework that greatly simplifies the implementation. To encourage well-formed meshes, we include a Laplacian regularizer [Sorkine 2005] in our objective function. Each component is further detailed in Section 5. This setup is flexible and allows for differentiable rendering at high resolutions and high polygon counts at interactive rates.

4 APPLICATIONS

In this section, we present several use cases for our method: joint simplification of shape and appearance, prefiltering shape and appearance to reduce aliasing, geometric simplification of skinned character animation, approximation of aggregate geometry, and 3D mesh extraction from implicit surfaces. Unless otherwise noted, we optimize for 10k iterations at a resolution of 2048×2048 pixels, where each iteration uses a random camera and light position. We defer discussion of implementation details until Section 5. Please refer to our interactive image viewer for detailed image comparisons.

4.1 Joint Shape-Appearance Simplification

Simplifying complex assets with minimal loss in visual fidelity is our most straightforward application. We present three variants that demonstrate joint optimization over different combinations of shape and appearance: normal map baking, joint simplification that also accounts for surface reflectance, and approximating complex meshes with displacement maps applied on a coarse base domain.

Normal map baking. As an initial example, we start from a sphere with 3k triangles and optimize shape and a tangent-space normal map to approximate a highly detailed reference mesh with 735k triangles. Besides the normals, the material is otherwise fixed as diffuse uniform gray. Our result is shown in the top row of Figure 3. While some high-frequency detail is missing, the result is nonetheless encouraging, considering that the optimization process is entirely

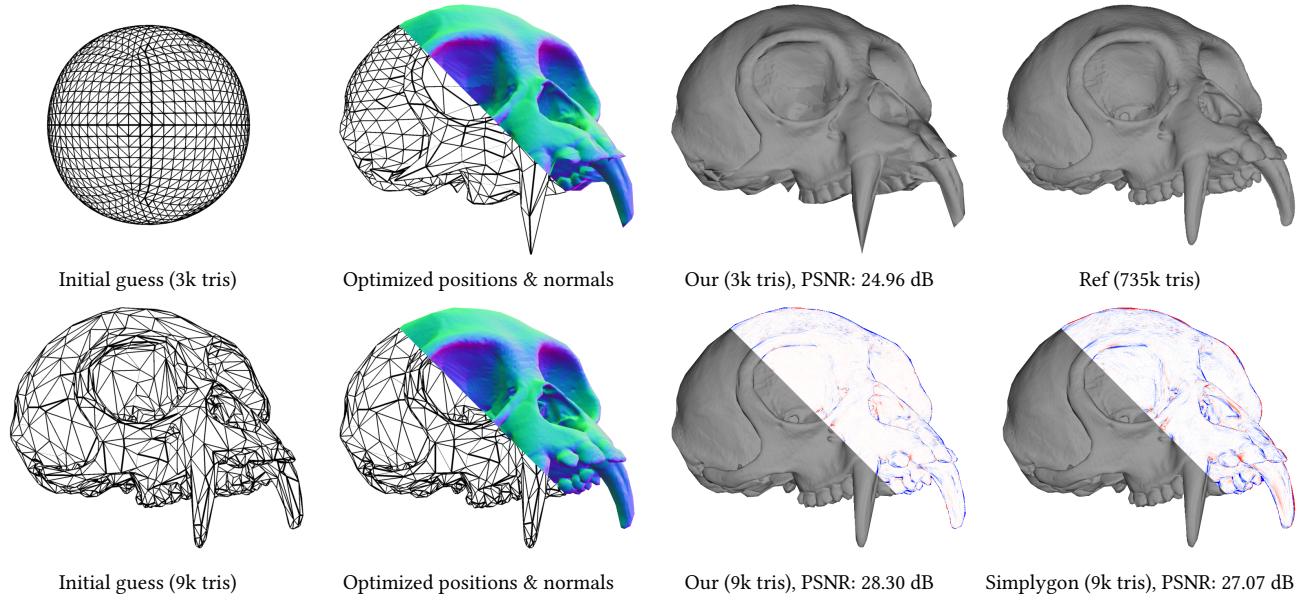


Fig. 3. **Top:** Starting from a sphere, we jointly optimize vertex positions and tangent space normal maps based on image observations of a reference mesh. **Bottom:** Starting from a reduced mesh with 9k triangles. We compare against the normal map baker in Simplygon and show split images with difference images (red/blue = too bright/dim compared to reference). The mesh is courtesy of the Smithsonian 3D Digitization project [2018].

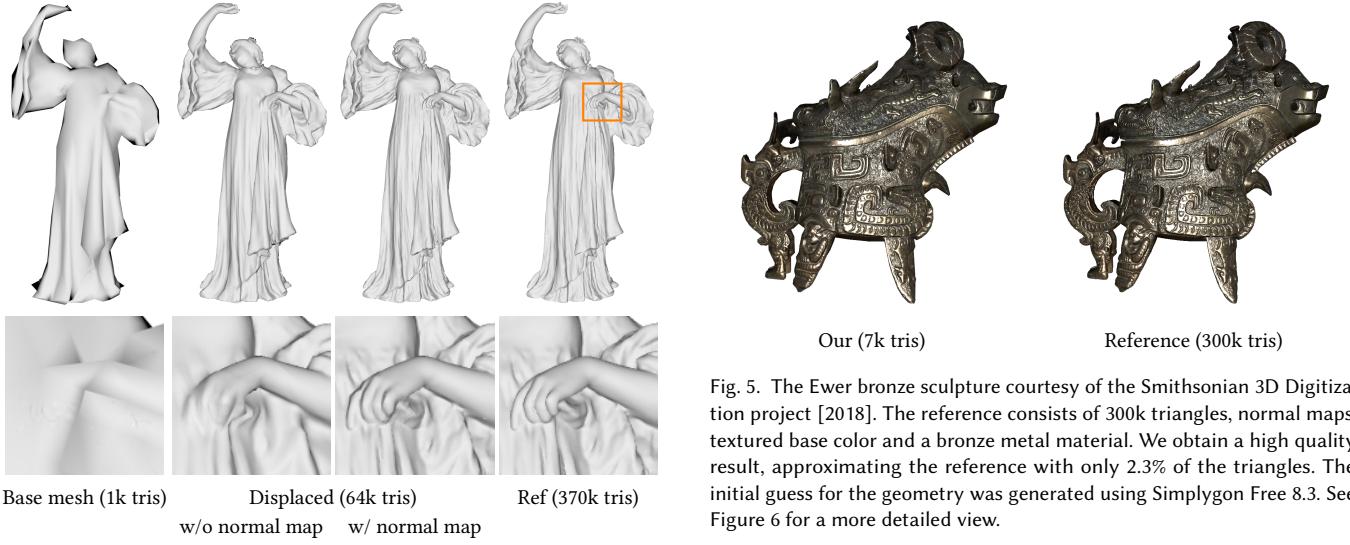


Fig. 4. We jointly optimize a base mesh, displacement map, and normal map to match the appearance of the dancer, courtesy of the Smithsonian 3D Digitization project [2018]. This is a complex optimization problem, with displacement constrained to the normal direction, and a coarsely tessellated base mesh. Still, the appearance after optimization matches the reference closely. Notably, some small details in the insets are baked into the normal map, even though they could be easily represented by displacement. We speculate that view parallax is minimal from the range of camera distances used during optimization, causing displacement and normal mapping to be equally viable. The initial decimated mesh was generated using the mesh simplifier in Autodesk Maya 2019.

automatic and uses no direct information about the reference model. In the bottom row of the figure, we repeat the experiment starting from a reduced version of the reference mesh with 9k triangles, as produced by Simplygon Free 8.3. As expected, this improves the results considerably, as it is now sufficient to fine-tune the geometry instead of discovering it from scratch, and we slightly outperform the Simplygon normal map baker. Please refer to the supplemental material for a study of how quality is impacted by the triangle count of the initial mesh.

Fig. 5. The Ewer bronze sculpture courtesy of the Smithsonian 3D Digitization project [2018]. The reference consists of 300k triangles, normal maps, textured base color and a bronze metal material. We obtain a high quality result, approximating the reference with only 2.3% of the triangles. The initial guess for the geometry was generated using Simplygon Free 8.3. See Figure 6 for a more detailed view.



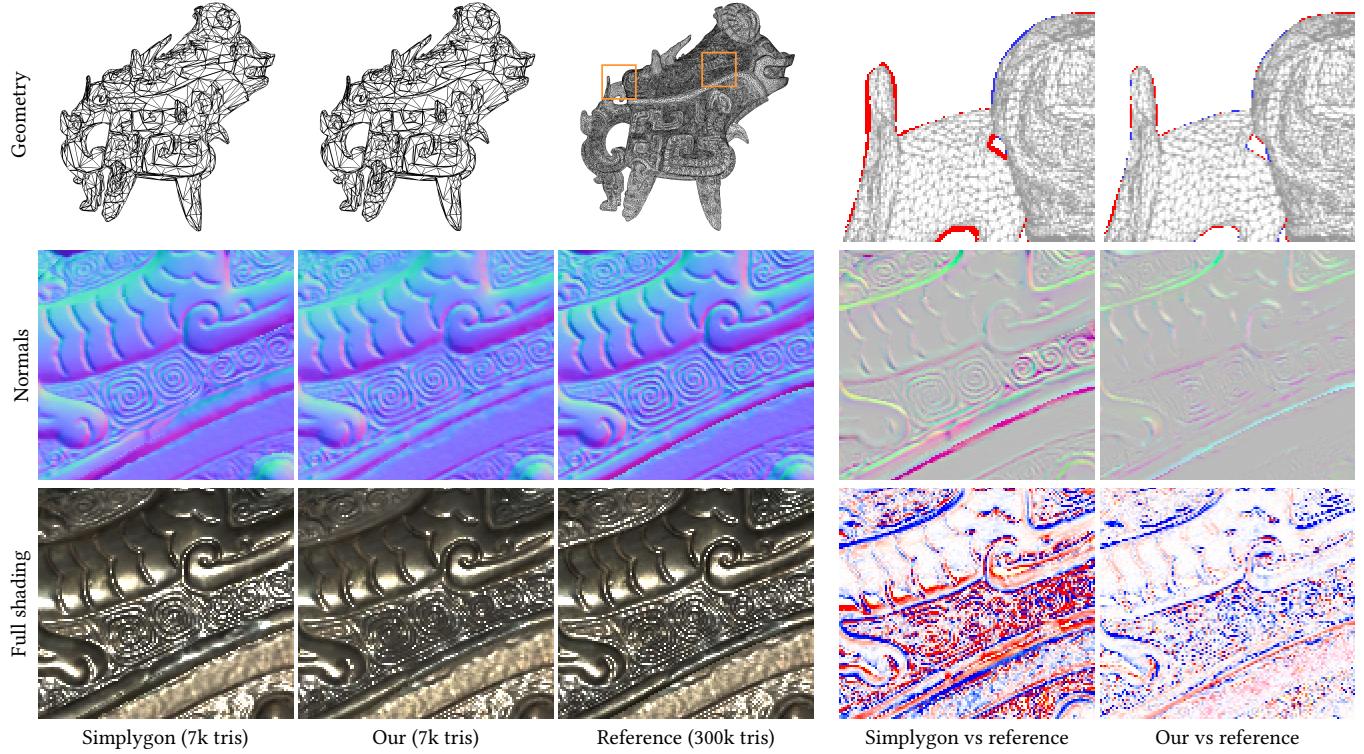


Fig. 6. Analysis of the Ewer sculpture mesh. **Top row:** Initial simplified geometry by Simplygon, our optimized geometry, and original geometry. Difference images on the right show the coverage errors of the simplified meshes overlaid on the reference mesh. Red indicates pixels covered by the simplified mesh but not the reference mesh, and blue indicates the opposite situation. As can be seen, optimizing the geometry improves the silhouettes considerably, although its effects are otherwise subtle. **Middle row:** Closeups of normals generated by Simplygon, our optimized normals, and reference normals, followed by difference images. Note that our normals are not optimized directly against reference normals but discovered via optimizing the full shading result against the reference. **Bottom row:** Full shading results rendered at 1 spp with difference images (red/blue = too bright/dim compared to reference). Our result exhibits fewer overly bright pixels because the rendering is optimized to match the reference on average, thereby reducing aliasing-induced sparkling. Some of the improvement over Simplygon’s output may be attributed to the use of more versatile shading model, but that does not explain the improvement in, e.g., normals or silhouettes.

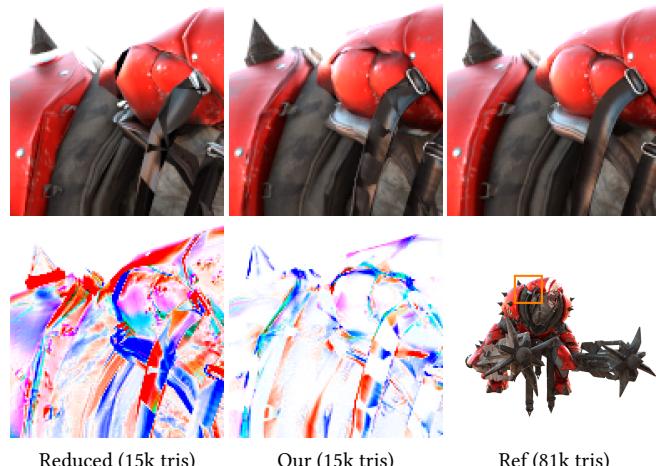


Fig. 7. A character from the Unreal Engine Paragon asset [2018]. We clear up some of the artifacts introduced by the automatically reduced mesh (generated in Autodesk Maya 2019), reattach geometry, and repair texture. The bottom row shows difference images (red/blue = too bright/dim compared to reference). Please refer to the supplemental material for full images.

Displacement map baking. In addition to normal maps, displacement mapping is an increasingly popular technique for representing complex shapes in real-time settings [Epic Games 2020]. It achieves a compact representation by tessellating a coarse *base mesh* on the fly, and displacing the resulting vertices in the direction of the interpolated surface normal by amounts read from the displacement map texture.

Our approach enables using displacement maps for approximating geometry by simply implementing the tessellation and displacement steps in our forward rendering pipeline. Figure 4 shows a displacement mapped version of the dancer mesh, rendered with diffuse shading to make the geometrical impact more apparent. Our result is obtained by jointly optimizing the pre-tessellation shape of the base mesh, the normal map, and the displacement map. As shown in the insets, our process yields a natural “division of labor” between the representations: the base mesh models the overall shape, the displacement map models mid-scale detail, and the finest detail that is not representable by the displaced surface is captured by the normal map.

While the above example is optimized for a single, fixed tessellation level, dynamic tessellation is often used for level of detail

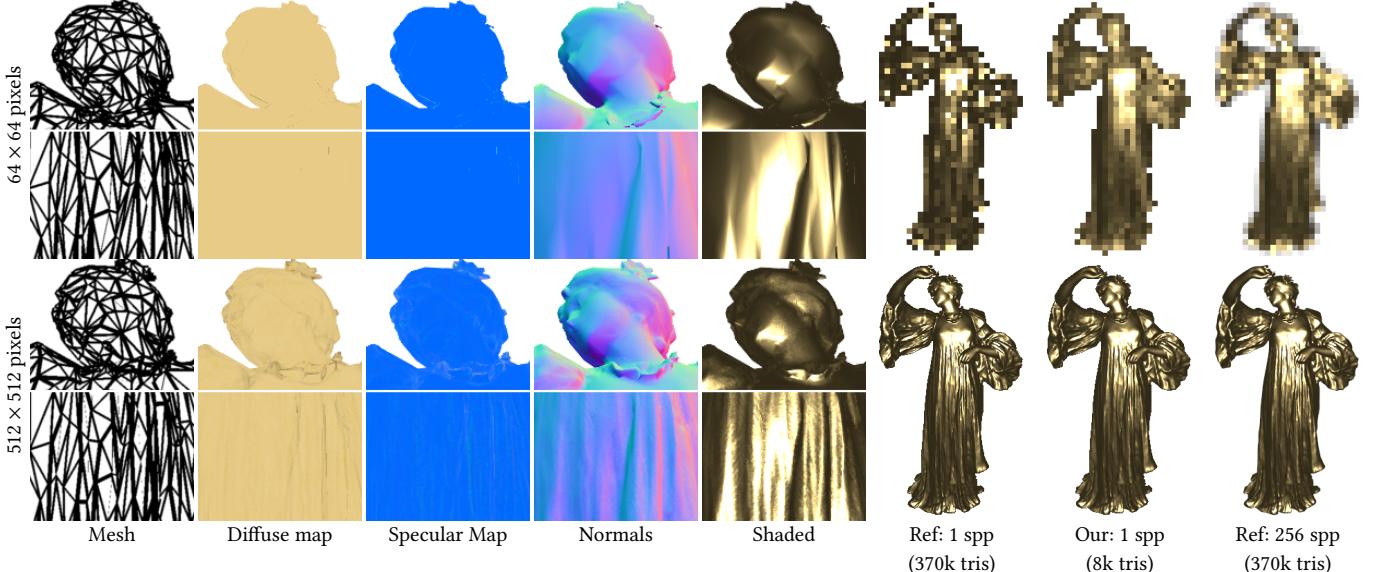


Fig. 8. We can jointly prefilter shape and appearance for a certain rendering resolution. The dancer in the **top row** is optimized for a resolution of 64×64 pixels, and for 512×512 pixels in the **bottom row**. Note in the top row how the normals are smoothed to band-limit shading. The three rightmost images show the models rendered at the intended resolution. We match the appearance of the super-sampled reference well with a single sample per pixel.

selection. We can additionally optimize a displacement mapped mesh to look good under multiple levels of tessellation. Please refer to the supplemental material for further results.

Simplification with complex materials. Next, we upgrade to a physically based shading model with one diffuse lobe and one isotropic GGX specular lobe [Walter et al. 2007] commonly used in modern game engines. Figure 5 shows a bronze sculpture of high geometric complexity, complex texture mapped materials, and normal maps. Here, we optimize jointly for shape and appearance under random views and point light directions. The specular term introduces higher frequencies and higher dynamic range, both of which make the optimization process more challenging. Hence, we use an image-domain loss robust to large floating-point values. Please refer to Section 5.3 for details. Figure 6 shows further comparisons to the reference and to the initial mesh reduced using Simplygon. Our method, optimizing based on visual differences, closely matches the true silhouette, finds accurate normals, and captures the shaded appearance of the reference. The specular highlights in our results are somewhat blurred, but it could be argued that this is preferable to the aliasing that the reference mesh shows when rendered at 1 spp.

Automatic cleanup. As a final example, Figure 7 demonstrates cleanup of the result of an unsuccessful mesh decimation operation performed in another software package. In this test, we reduced a game character from the Unreal Engine Paragon asset [2018] from 81k to 15k triangles in Autodesk Maya 2019. Note that the automatic reduction slightly decreases the volume of the mesh, detaches geometric elements, suffers from incorrect texturing, and produces some self-intersecting geometry. After optimization, we regain the

volume and automatically clean up most of the geometry and texturing issues. We do not support topology changes, so our optimized mesh is still not watertight and retains some of the artifacts, but the rendered appearance is nonetheless clearly improved. Please refer to the interactive viewer for full resolution images.

4.2 Shape and Appearance Prefiltering

In the previous section, our goal was to create faithful representations of complex assets with reduced triangle counts. A closely related variant of this problem is to find efficiently renderable approximations to original assets that are so complex that they require a substantial amount of supersampling to produce alias-free images. We call this problem joint prefiltering of shape and appearance. The resulting optimized models have the property that they reproduce, when rendered at only one sample per pixel, the appearance of assets that require potentially hundreds of samples per pixel for alias-free reproduction. The only practical differences in the optimization are specifying a typically smaller target image resolution, and rendering the reference images with enough supersampling to ensure lack of aliasing; all previously demonstrated freedoms in choosing the shape and appearance models still apply. We address some technical challenges associated with high triangle counts in Section 5.2.

Figure 8 shows joint shape and appearance prefiltering on the golden statue, targeting rendering resolutions of 64×64 and 512×512 pixels. Note that the result prefiltered for the smaller resolution of 64×64 pixels has, as one would expect, considerably smoother normals that account, together with the specular map, for the effect of averaging present in supersampling. Also note the geometric smoothing, e.g., the flower on the statue’s head. When rendered

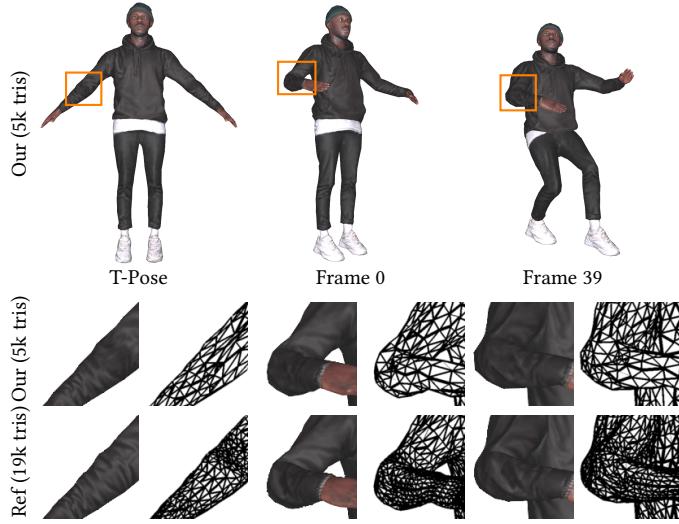


Fig. 9. Mesh decimation applied to an animated, rigged character from RenderPeople [2020]. We reduce the mesh and optimize vertex positions, skinning weights, normal maps, and material parameters over the animation.

at the intended resolution, the low-resolution mesh matches the appearance of the reference well, with no apparent aliasing.

To obtain an appropriate amount of prefiltering at different rendering resolutions (i.e., rendering distances), it is not sufficient to optimize for one resolution only. For materials, this is easy to achieve by treating each mipmap level as an independent latent variable. This yields resolution-specific material representations that a trilinear texture lookup will automatically interpolate between. Even though linear interpolation between material parameters is not generally correct [Bruneton and Neyret 2012], the optimization process will find a representation that, assuming trilinear texture fetches will be used, matches the target images as well as possible.

For geometry, we can store multiple sets of vertex positions and choose between these based on distance to mesh, average projected edge length, or a similar heuristic. As with mipmapping, linear interpolation between levels can be used to eliminate popping artifacts. To simplify the experiments, we have opted to keep the topology fixed, so no reduction in triangle count is obtained in our tests—for geometric simplification, a mesh LOD scheme would need to be incorporated into the optimization. Please see the accompanying video for an example animation of continuous, distance-dependent prefiltering.

4.3 Animation and Skinning

Having so far focused on static scenes, we now study appearance-driven simplification of animated articulated characters over entire animation sequences. More precisely, given a high-resolution reference mesh animated by skeletal subspace deformation (SSD), we optimize over the bind-pose vertex positions, normals, SVBRDF, and skinning weights (bone-vertex attachments) of a simplified model in an attempt to replicate the appearance of the reference animation. In contrast to simplifying the character in the bind pose (T-pose) only,

this holds promise for being able to strike compromises to distribute the error evenly among the frames by adjusting the geometry, skinning weights, and materials appropriately.

Implementation is straightforward: the only addition required is blending transformed vertex positions using the skinning weights, a simple linear operation. An example is shown in Figure 9, where we decimate a rigged animated mesh in a completely automated process. We include examples on rigged meshes from RenderPeople [2020] using skeletal animations from the CMU motion capture database [2020] in the accompanying video.

We assume the time-varying bone transformations are known, and treat them as constants during optimization. Joint optimization of both bone transformations and skinning weights [James and Twigg 2005] is a possible direction for future work. Furthermore, we assume that normal maps and SVBRDFs are constant in time. Modeling dynamic behavior like wrinkles opens another interesting future direction.

Above we use a reduced mesh as initial guess, but to thoroughly battle-test our ability to optimize skinning weights we instead start from a sphere and morph it into an animated figure with known skeletal animation, jointly optimizing shape, materials and skinning weights. Please refer to the video for results.

4.4 Approximating Aggregate Geometry

Stochastic aggregate geometry, such as foliage, are particularly difficult to simplify: as the overall appearance emerges from the combined effect of many small, disjoint components, techniques such as mesh decimation are ineffective.

Cook et al. [2007] introduce a stochastic decimation technique that exploits a known scene graph, and randomly remove a subset of the geometric elements and alter the remaining elements, e.g., by scaling and contrast adjustments, to preserve the overall appearance of a scene. We approach the same problem from another angle, drawing inspiration from the billboard clouds of Décoret et al. [2003]: instead of stochastically pruning the procedural scene graph, we replace the complex leaf geometries with textured quads. With the quads providing an initial guess, we then jointly optimize material parameters, shape, and *transparency* based on visual loss of rendered images. For this, we extended the differentiable rasterization primitives of Laine et. al [2020] to support order-independent transparency through depth peeling [Everitt 2001]. Please refer to Section 5.2 for details.

In Figure 10, we show two examples of simplification of aggregate geometry from the Disney Moana asset [2018]. In both cases, we create plausible approximations from only 0.8% and 0.4% of the triangles of the reference mesh, respectively. For this application, we used squared L_2 as objective function.

In Figure 1, we instance our approximation 3000× and render in a path tracer. The result closely resembles the reference with a massive reduction in geometric complexity. Our supplemental material includes a comparison with stochastic simplification [Cook et al. 2007]. Please also refer to our video for animated results.

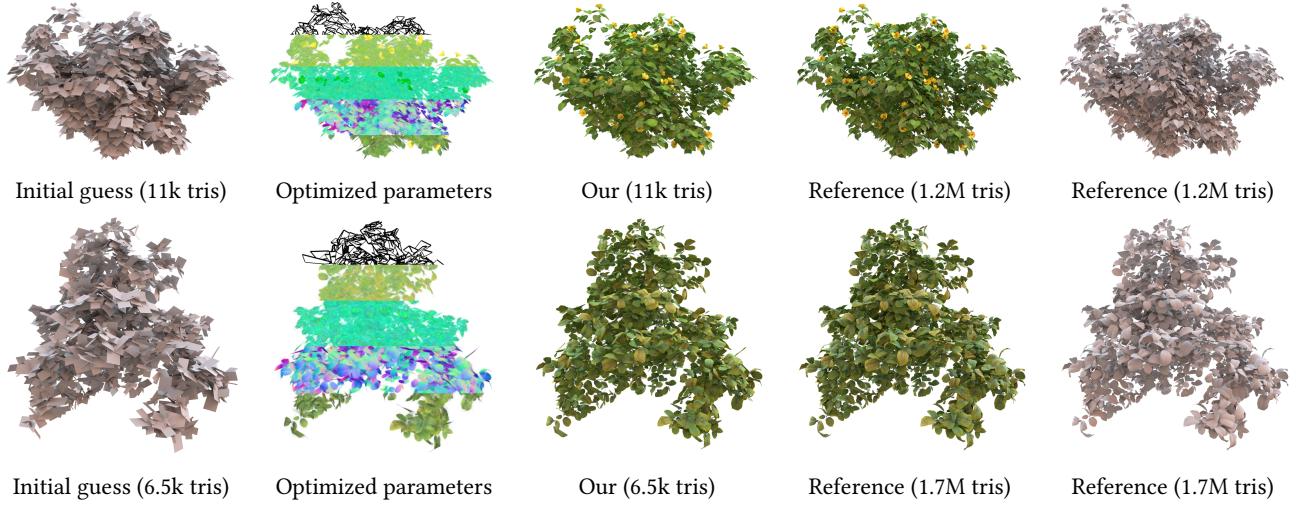


Fig. 10. Approximating aggregate geometry. We start from a low-polygon mesh and jointly optimize shape, material parameters, and transparency. The shaded results are rendered in a path tracer to illustrate that our results generalize across renderers. **Top row:** The leaves and flowers of the “isHibiscus” asset (1.2M triangles), approximated by 11k tris. **Bottom row:** The leaves from the “isGardenia” asset (1.7M triangles), approximated by 6.5k triangles. The models are taken from the Moana Island Scene [2018], a publicly available data set courtesy of Walt Disney Animation Studios.

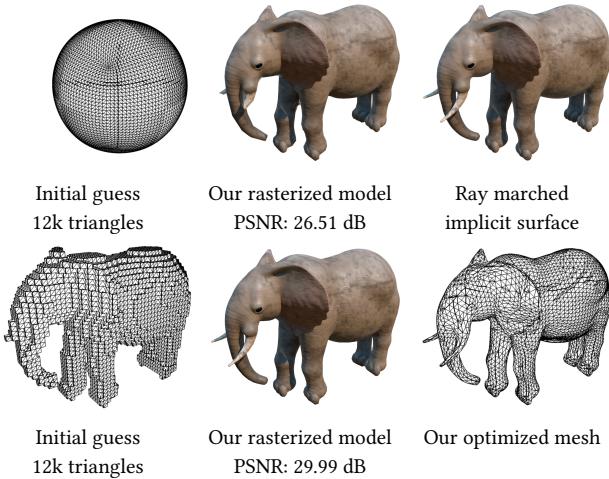


Fig. 11. We convert a ray marched implicit model, an adapted version of the ShaderToy “Elephant” ©Inigo Quilez, to a mesh with materials by optimizing for visual loss in a differentiable rasterizer. We use a tessellated sphere (12k triangles) as initial guess and jointly optimize shape and appearance. We also show the corresponding results using a better initial guess, produced by marching cubes. Note the improvements on sharp details, e.g., the tusks.

4.5 Generalizations

All examples so far have used the same geometric representation for both the latent representation and the reference model. Moreover, the reference images have been produced by the same renderer as used for optimization. Enabled by supervising the optimization strictly in image space, we now demonstrate generalization across surface representations and rendering systems. While the full scope of potential applications is vast, we illustrate this by two examples:

converting a ray marched implicit surface to a mesh, and transferring a path traced rendered model to a rasterizer.

Textured meshes from implicit surfaces. In the first example, shown in Figure 11, we adapt an implicit surface pixel shader from ShaderToy [Jeremias and Quilez 2014] to isolate the main object and match the lighting and camera model of our renderer. We automatically convert this ray marched implicit surface to a triangle mesh with materials. We use a tessellated sphere as the initial guess for the geometry. For this example, we also add an ambient material term to our latent representation to better match the lighting of the implicit surface renderer, which uses custom ambient lighting. We also use a static light position, so shadowing is captured and baked into the material parameters in the optimization process. View-dependent shading effects, e.g., specular highlight, are still captured. Please refer to the supplemental material for additional results.

Baking path traced lighting. The second example is shown in Figure 12. Here, we use the ViSII path tracer [Morrical et al. 2020] to generate reference images for the aggregate geometry decimation example of Section 4.4. We use a static light position during optimization: materials are effectively converted to our material model and shadows are baked into the textures, creating a plausible rasterized approximation of the path traced reference. Note that we control the viewing conditions in the reference images, and use matching configurations in the optimization.

5 IMPLEMENTATION

This section covers the implementation details of our method. We first describe the differentiable mesh operations that are applied to the mesh before rendering. The second part details our differentiable renderer, and the final part provides details of the optimization process.

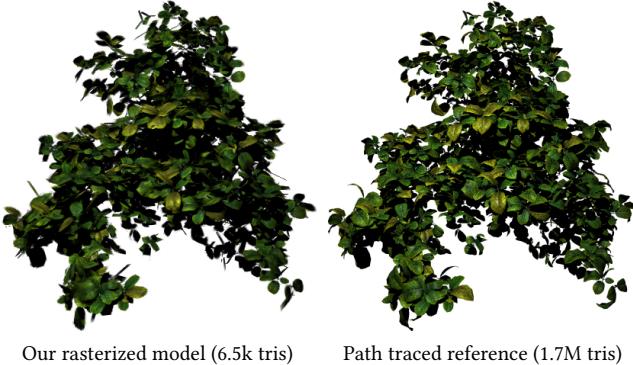


Fig. 12. We optimize a mesh and materials to have the appearance of a path traced reference, when rendered in a rasterizer. This can be used to convert between different material models, or as a simple way of baking shading into material terms. Our optimized model is rasterized with 1 spp + post-processing antialiasing. The path traced reference is rendered with 256 spp.

5.1 Mesh Operations

Referring to Figure 2, our pipeline includes differentiable mesh operations for tangent space computation, animation & skinning, and displacement mapping.

Tangent space. To optimize tangent space normal maps on deforming geometry, the tangent frame must be differentiable and dynamically updated to reflect any change in vertex position. We compute smooth vertex normals and derive tangent and bi-tangent vectors from the vertex positions and texture coordinates [Mikkelsen 2008]. Using mesh-derived smooth normals is a not a limitation because creases or other sharp features can be handled by the normal map.

Animation & skinning. We support skinning for Universal Scene Description (USD) [2016] meshes and rely on the USD API to evaluate skeleton animation. We optimize the skinning weights of animated meshes, and therefore implement a differentiable skinning operator according to:

$$\mathbf{v}_i^s = \sum_{b \in \mathcal{B}} w_{ib} M_b \mathbf{v}_i, \quad (2)$$

where \mathcal{B} is the set of bones, M_b is the bone transform matrix for the current frame, and w_{ib} is the skinning weight of bone b influencing vertex \mathbf{v}_i . Weights are typically stored using a sparse indexed representation, but we implement the full dense skinning operator to support any vertex-bone association during optimization.

Displacement mapping. Here, the latent representation consists of a coarse base mesh and a scalar displacement map. The mesh is subdivided, and the displacement map is used to displace the tessellated vertices along the interpolated normal direction. The tessellator uses edge-midpoint subdivision [Wang et al. 2018], where each triangle is split into four new triangles by inserting a new vertex along each edge. This operation changes topology, which is not a differentiable operation in our current renderer. This is simple to work around by using a tessellation criteria that does not depend on

any parameters requiring gradients. For the example in Section 4.1, we select a constant tessellation factor and precompute the topology of the tessellated mesh before optimization. The position of each vertex created by the tessellation is recomputed every iteration to ensure that gradients are propagated correctly.

For displacement lookups we deploy the differentiable texture primitive of Laine et al. [2020], and displace each vertex according to:

$$\mathbf{v}_i^d = \mathbf{v}_i + \text{tex2d}(\mathbf{t}_i) \cdot \mathbf{n}_i, \quad (3)$$

where \mathbf{v}_i is the original tessellated vertex position, \mathbf{n}_i is the interpolated normal, and \mathbf{t}_i is the texture coordinate.

5.2 Differentiable Renderer

Our renderer is based on the differentiable rasterization primitives of Laine et al. [2020]. We employ deferred shading based on a G-buffer that stores 3D position, normal, tangent, bi-tangent, and texture coordinates for each pixel.

For materials, we use a variant of the physically based model from Disney [Burley 2012] that is common in modern game engines [Karis 2013; Lagarde and de Rousiers 2014]. This allows us to easily import game assets and lets us render our optimized meshes directly in existing engines without modifications. Material models for real-time rendering commonly combine a diffuse term with an isotropic, specular GGX lobe [Walter et al. 2007]. The parameters for the diffuse lobe \mathbf{k}_d are provided in a four-component texture, where the optional fourth channel α represents transparency. The specular lobe is described by a roughness value r for the GGX normal distribution function and a metalness factor m which interpolates between plastic and metallic appearance by computing a specular highlight color according to $\mathbf{k}_s = (1 - m) \cdot 0.04 + m \cdot \mathbf{k}_d$ [Karis 2013].

For appearance filtering, we want additional flexibility in suppressing the specular lobe. To that end, we add a parameter γ that scales the specular lobe according to: $\mathbf{k}'_s = (1 - \gamma)\mathbf{k}_s$. Our specular parameters are thus represented by a three-component texture (γ, r, m) where each component may vary spatially. This representation is purposely chosen to resemble the commonly used ORM (occlusion, roughness, metalness) textures, where we have replaced the occlusion channel with γ .

Note that we have chosen to reparameterize a single, fixed BSDF model for appearance filtering to make results easily adoptable in game engines and to show that true prefiltering can be made a part of the asset pipeline. However, it would be straightforward to replace the BSDF with multi-lobe models [Bruneton and Neyret 2012], spherical harmonics representations [Han et al. 2007], or a variant of the neural representation proposed by Müller et al. [2019] for more powerful representations.

Antialiasing. For the prefiltering applications in Section 4.2, we want to match the appearance of a highly supersampled target image to a 1 spp rendering of our latent representation. In practice, we experience instabilities when optimizing for low rendering resolutions due to approximations in silhouette gradient computations [Laine et al. 2020]. We work around this by rendering our latent representation using multisampled antialiasing (MSAA), i.e., we sample visibility at a higher rate but shade only once per pixel. This setup improves visibility gradients at silhouettes, but still enforces that

our latent representation matches the shading of the supersampled reference with just a single shading sample.

This solution has the limitation that MSAA correctly integrates the visibility term, which is not expected for 1 spp rendering. This creates a small mismatch between our optimization setup and final rendering of the optimized model at 1 spp, but we have not found this to be an issue in practice. MSAA is only used during training, and all our result images and animations indicating 1 spp are rendered without it, i.e., they represent true 1 spp rendering without any antialiasing unless otherwise mentioned.

Order-independent transparency. Transparency is required for the aggregate geometry application in Section 4.4. We implement order-independent transparency through *depth peeling* [Everitt 2001], i.e., by rendering multiple passes where each pass peels off the front-most depth layer. We extend the rasterization primitive of Laine et al. [2020] by adding a two-sided depth test to their fragment shader and passing the depth output of previous rasterization pass as a parameter to the next. While we rasterize the depth layers front-to-back, we perform blending in back-to-front order (starting with the background) as this works best with their antialiasing primitive [Laine et al. 2020]. Eight passes of depth peeling were used in our experiments.

5.3 Optimization

Objective function. Our renderer uses physically based shading and produces images with high dynamic range. Therefore, the objective function must be robust to the full range of floating-point values. Following recent work in HDR image denoising [Munkberg and Hasselgren 2020], our image space loss, L_{image} , computes the L_1 norm on tone mapped colors. As tone map operator, we transform linear radiance values, x , according to $x' = \Gamma(\log(x + 1))$, where $\Gamma(x)$ is the sRGB transfer function [Stokes et al. 1996]:

$$\begin{aligned} \Gamma(x) &= \begin{cases} 12.92x & x \leq 0.0031308 \\ (1+a)x^{1/2.4} - a & x > 0.0031308 \end{cases} \\ a &= 0.055. \end{aligned} \quad (4)$$

In addition, we use a Laplacian regularizer [Sorkine 2005] on the triangle mesh in our latent representation. This is important in the beginning of optimization to keep the mesh surface intact when gradients are large. The uniformly-weighted differential δ_i of vertex v_i is given by $\delta_i = v_i - \frac{1}{|N_i|} \sum_{j \in N_i} v_j$, where N_i is the one-ring neighborhood of vertex v_i . We follow Laine et al. [2020] and use a Laplacian regularizer term given by

$$L_{\delta} = \frac{1}{n} \sum_{i=1}^n \|\delta_i - \delta'_i\|^2, \quad (5)$$

where δ'_i is the uniformly-weighted differential of the input mesh (i.e., our initial guess). When the input mesh is a poor approximation, e.g., a sphere, we use an *absolute* regularizer and set $\delta'_i = 0$.

Our combined objective function is:

$$L_{\text{opt}} = L_{\text{image}} + \lambda_t L_{\delta}, \quad (6)$$

where λ_t is the regularization weight that depends on the current optimization iteration t . We gradually reduce λ_t during optimization according to $\lambda_t = (\lambda_{t-1} - \lambda_{\min}) \cdot 10^{-kt} + \lambda_{\min}$. Here, $k = 10^{-6}$, and

λ_{\min} is chosen as 2% of the initial weight, λ_0 . The uniform Laplacian regularizer depends on tessellation, whereas image-domain loss does not. Hence, the image loss must be balanced against the Laplacian loss as our applications include meshes with greatly varying triangle counts. The initial weight, λ_0 , can either be specified by the user or by a simple heuristic: We evaluate the Laplacian error at the start of optimization, and set $\lambda_0 = 0.25 L_{\text{image}} / L_{\delta}$, which has worked well for most of our examples. Please refer to the supplemental material for an example of how the regularizer improves mesh quality.

Learning rate and runtime. We optimize the latent representation using Adam [Kingma and Ba 2015] with default parameters. Learning rate is scheduled as $\text{lr}_i = \text{lr}_0 \cdot 10^{-kt}$, where t is the iteration, lr_0 is the initial learning rate, and $k = 0.0002$. When starting from a coarse initial guess, e.g., a sphere or billboard cloud, we typically use a high initial learning rate, e.g., $\text{lr}_0 = 0.01$, to allow for large mesh deformations. When starting from an auto-decimated mesh, we want to fine-tune the result without jumping out of the already good local minima, which leads us to using lower values for $\text{lr}_0 \in [0.001, 0.003]$. We note that mini-batching is highly beneficial to reduce gradient noise, particularly during the early phases of optimization. Gradient noise is problematic for vertex positions as it can cause the mesh to fold or self-intersect, especially in highly tessellated regions. In practice, we use batch size between one and eight. For the final results, we run optimization at a resolution of 2048×2048 pixels for 10k steps, which typically takes a few hours on a single NVIDIA V100 GPU.

6 LIMITATIONS

Apart from technical limitations, e.g., large memory consumption, in particular when using many layers for depth peeling or when rendering at high resolutions, we note three main limitations in our approach.

Hyperparameter tuning. For each use case, we need to tune the constants for the Laplacian regularizer and learning rate. This is a common problem in shape and appearance optimization, but it is pronounced here as we work with a wide range of tessellation rates and initial meshes of varying quality. We use a simple heuristic to set the Laplacian regularizer weight and learning rate parameters, but robustly configuring these hyperparameters for all use cases remains an open challenge.

View parallax. The second limitation is inherent in our problem formulation. We optimize for image-domain loss, which sometimes suffers from view parallax between the latent representation and reference model, as illustrated by Figure 13. With L_1/L_2 image losses, discrepancies due to view parallax will converge to the median/mean value during optimization, which tend to blur the material parameter textures somewhat in cases where the geometry differences are large. This is particularly prominent for aggregate geometry, where transparency helps amplify the effect, and where we do the most extreme polygon reduction.

As an attempt to mitigate this problem, we experimented with perceptual VGG loss [Ledig et al. 2017] and adversarial losses [Radford et al. 2015] that are known to preserve high-frequency detail in image reconstruction tasks and should be less sensitive to small

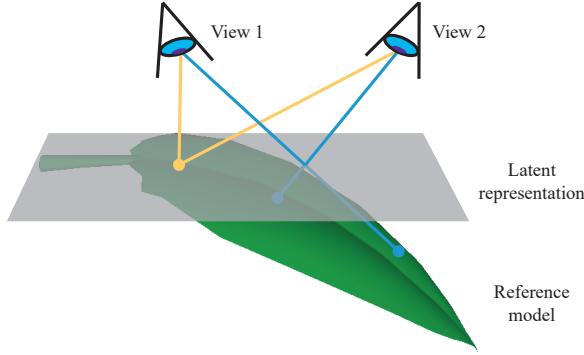


Fig. 13. View parallax can sometimes be an issue. This is most noticeable in our aggregate geometry application where a tessellated leaf is represented as a single quad. There is no problem where shape approximation is good, as illustrated by the yellow rays. However, as the blue rays show, we introduce parallax when the latent representation and reference model deviate. This leads to blurry normal maps and textures. The effect is exaggerated here for illustration purposes—in this situation, the optimization of vertex positions would tend to move the latent geometry closer to the reference model, partially because this also reduces the parallax-related discrepancies.

translations. Unfortunately we noted no or limited benefit, with both strategies adding details resembling film grain or noise patterns instead of improving the visual quality of results. These losses were also less robust in an HDR setting and did not properly preserve the high dynamic range of the results.

Rasterization-based rendering model. As our renderer is based on rasterization, it cannot handle effects such as refractive materials, specular interreflections, or subsurface scattering. If the reference renderer outputs images with these effects, our system will end up baking them into the available material parameters as well as possible, but highly view-dependent non-trivial effects such as translucency or subsurface scattering can be captured only on average.

However, differentiable rendering is a highly active field of research, and combining our approach with a differentiable path tracer [Nimier-David et al. 2020] could allow handling these kind of effects in the future.

7 CONCLUSIONS AND FUTURE WORK

We have demonstrated that a wide spectrum of modeling tasks, including simplification, conversion, and prefiltering, can be achieved in a common inverse rendering framework that supports various shape and appearance models. While we show improvements in individual examples such as mesh decimation, we believe the main strength of our approach lies in the ability to jointly optimize over shape, appearance, and animation parameters. Additionally, coupling automatic differentiation and optimization makes extensions to new applications and latent representations easy.

There are wide opportunities for future work. To keep a clear focus for this paper, we have only evaluated applications within the limitations of triangle meshes and a commonly used material model, so that the optimized representation can be used unmodified in existing renderers. However, an obvious line of extensions is to

use our framework to augment traditional fixed-function graphics models with learned representations, effectively creating hybrids between traditional graphics and the currently popular fully learned renderers [Thies et al. 2019].

Another direction is to further automate and improve our optimization process. In particular, highly tessellated meshes are difficult to optimize, as a large neighborhood of vertices must often be moved in tandem to achieve a large-scale deformation. It might be much easier to guide optimization using a hierarchical approach where more triangles are progressively generated as the result converges. This progression could further provide a natural LOD representation for the final mesh. For applications where no initial guess for the mesh can be obtained, it may be possible to improve results by using alternate geometrical representations, e.g., by using a volumetric grid and tessellating it using marching cubes.

We also envision that appearance-based optimization could be used in semi-automated modeling tools. When visualizing the optimization process, it is often evident where problems occur due to, e.g., under-tessellated regions or insufficient mesh genus, and an artist could likely clean up such areas given rudimentary interactive meshing tools.

REFERENCES

- Eric Bruneton and Fabrice Neyret. 2012. A Survey of Non-linear Pre-filtering Methods for Efficient and Accurate Surface Shading. *IEEE Transactions on Visualization and Computer Graphics* 18, 2 (2012), 242–260.
- Brent Burley. 2012. Physically Based Shading at Disney. In *SIGGRAPH Courses: Practical Physically Based Shading in Film and Game Production*.
- Wenzheng Chen, Huan Ling, Jun Gao, Edward Smith, Jaakko Lehtinen, Alec Jacobson, and Sanja Fidler. 2019. Learning to Predict 3D Objects with an Interpolation-based Differentiable Renderer. In *Advances in Neural Information Processing Systems* 32, 9609–9619.
- Zhiqin Chen, Andrea Tagliasacchi, and Hao Zhang. 2020. BSP-Net: Generating Compact Meshes via Binary Space Partitioning. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020).
- Jonathan Cohen, Amitabh Varshney, Dinesh Manocha, Greg Turk, Hans Weber, Pankaj Agarwal, Frederick Brooks, and William Wright. 1996. Simplification Envelopes. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '96)*, 119–128.
- Robert L. Cook, John Halstead, Maxwell Planck, and David Ryu. 2007. Stochastic Simplification of Aggregate Detail. *ACM Trans. Graph.* 26, 3 (2007).
- Massimiliano Corsini, Mohamed-Chaker Larabi, Guillaume Lavoué, Oldrich Petrik, Libor Váša, and Kai Wang. 2013. Perceptual metrics for static and dynamic triangle meshes. *Computer Graphics Forum* 32, 1 (2013), 101–125.
- Xavier Décoret, Frédéric Durand, François Sillion, and Julie Dorsey. 2003. Billboard clouds for extreme model simplification. *ACM Trans. Graph.* 22, 3 (2003).
- Yue Dong. 2019. Deep appearance modeling: A survey. *Visual Informatics* 3, 2 (2019), 59–68.
- Jonathan Dupuy, Eric Heitz, Jean-Claude Iehl, Pierre Poulin, Fabrice Neyret, and Victor Ostromoukhov. 2013. Linear Efficient Anti-aliased Displacement and Reflectance Mapping. *ACM Trans. Graph.* 32, 6, Article 211 (2013).
- Epic Games. 2018. Epic Games Paragon Assets. <https://www.unrealengine.com/en-US/paragon>.
- Epic Games. 2020. Unreal Engine 5: Nanite. <https://www.unrealengine.com/en-US/blog/a-first-look-at-unreal-engine-5>.
- Cass Everitt. 2001. Interactive Order-Independent Transparency.
- Alain Fournier. 1992. *Filtering Normal Maps and Creating Multiple Surfaces*. Technical Report.
- Duan Gao, Xiao Li, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. 2019. Deep Inverse Rendering for High-Resolution SVBRDF Estimation from an Arbitrary Number of Images. *ACM Trans. Graph.* 38, 4, Article 134 (2019).
- Andrew Gardner, Chris Tchou, Tim Hawkins, and Paul Debevec. 2003. Linear Light Source Reflectometry. *ACM Trans. Graph.* 22, 3 (2003), 749–758.
- Michael Garland and Paul S. Heckbert. 1997. Surface Simplification Using Quadric Error Metrics. In *SIGGRAPH '97: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, 209–216.

- Abhijeet Ghosh, Tongbo Chen, Pieter Peers, Cyrus A. Wilson, and Paul Debevec. 2009. Estimating Specular Roughness and Anisotropy from Second Order Spherical Gradient Illumination. *Computer Graphics Forum* 28, 4 (2009), 1161–1170.
- D. Guarnera, G. C. Guarnera, A. Ghosh, C. Denk, and M. Glencross. 2016. BRDF Representation and Acquisition. In *Proceedings of the 37th Annual Conference of the European Association for Computer Graphics: State of the Art Reports*. 625–650.
- Yu Guo, Cameron Smith, Miloš Hašan, Kalyan Sunkavalli, and Shuang Zhao. 2020. MaterialGAN: Reflectance Capture Using a Generative SVBRDF Model. *ACM Trans. Graph.* 39, 6, Article 254 (2020).
- Charles Han, Bo Sun, Ravi Ramamoorthi, and Eitan Grinspun. 2007. Frequency Domain Normal Map Filtering. In *ACM SIGGRAPH 2007 Papers*. 28–40.
- Doug James and Christopher Twigg. 2005. Skinning mesh animations. *ACM Trans. Graph.* 24, 3 (2005).
- Krishna Murthy Jatavallabhula, Edward Smith, Jean-Francois Lafleche, Clement Fuji Tsang, Artem Rozantsev, Wenzheng Chen, Tommy Xiang, Rev Lebaredian, and Sanja Fidler. 2019. Kaolin: A PyTorch Library for Accelerating 3D Deep Learning Research. *arXiv:1911.05063* (2019).
- Pol Jeremias and Inigo Quilez. 2014. Shadertoy: Learn to Create Everything in a Fragment Shader. In *SIGGRAPH Asia 2014 Courses*. Article 18.
- Brian Karis. 2013. Real Shading in Unreal Engine 4. *SIGGRAPH 2013 Course: Physically Based Shading in Theory and Practice* (2013).
- Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Proceedings of the 3rd International Conference for Learning Representations*.
- CMU Graphics Lab. 2020. CMU Graphics Lab Motion Capture Database. <http://mocap.cs.cmu.edu/>.
- Sébastien Lagarde and Charles de Rousiers. 2014. Moving Frostbite to Physically Based Rendering 3.0. *SIGGRAPH 2014 Course: Physically Based Shading in Theory and Practice* (2014).
- Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. 2020. Modular Primitives for High-Performance Differentiable Rendering. *ACM Transactions on Graphics* 39, 6, Article 194 (2020).
- Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. 2017. Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Hendrik P. A. Lensch, Jan Kautz, Michael Goesele, Wolfgang Heidrich, and Hans-Peter Seidel. 2003. Image-Based Reconstruction of Spatial Appearance and Geometric Detail. *ACM Trans. Graph.* 22, 2 (2003), 234–257.
- Tzu-Mao Li, Miika Aittala, Frédéric Durand, and Jaakko Lehtinen. 2018. Differentiable Monte Carlo Ray Tracing through Edge Sampling. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 37, 6 (2018), 222:1–222:11.
- Peter Lindstrom and Greg Turk. 2000. Image-Driven Simplification. *ACM Transactions on Graphics* 19, 3 (2000), 204–241.
- Guillaume Loubet and Fabrice Neyret. 2017. Hybrid mesh-volume LoDs for all-scale pre-filtering of complex 3D assets. *Computer Graphics Forum* 36, 2 (2017), 431–442.
- Kok-Lim Low and Tiow-Seng Tan. 1997. Model Simplification Using Vertex-Clustering. In *Proceedings of the 1997 Symposium on Interactive 3D Graphics*. 75–82.
- David Luebke and Carl Erikson. 1997. View-Dependent Simplification of Arbitrary Polygonal Environments. In *SIGGRAPH '97: Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. 199–208.
- David Luebke, Benjamin Watson, Jonathan D. Cohen, Martin Reddy, and Amitabh Varshney. 2002. *Level of Detail for 3D Graphics*. Elsevier Science Inc., USA.
- Morten Mikkelsen. 2008. Simulation of Wrinkled Surfaces Revisited.
- Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.
- Nathan Morrical, Jonathan Tremblay, Star Birchfield, and Ingo Wald. 2020. ViSII: Virtual Scene Imaging Interface. <https://github.com/owl-project/ViSII/>.
- Thomas Müller, Brian McWilliams, Fabrice Rousselle, Markus Gross, and Jan Novák. 2019. Neural Importance Sampling. *ACM Trans. Graph.* 38, 5, Article 145 (2019).
- Jacob Munkberg and Jon Hasselgren. 2020. Neural Denoising with Layer Embeddings. *Computer Graphics Forum* 39 (2020), 1–12.
- Giljoo Nam, Joohy Lee, Diego Gutierrez, and Min H. Kim. 2018. Practical SVBRDF Acquisition of 3D Objects with Unstructured Flash Photography. *ACM Trans. Graph.* 37, 6, Article 267 (2018).
- Merlin Nimier-David, Sébastien Speierer, Benoît Ruiz, and Wenzel Jakob. 2020. Radiative Backpropagation: An Adjoint Method for Lightning-Fast Differentiable Rendering. *ACM Trans. Graph.* 39, 4, Article 146 (2020).
- Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. 2019. Mitsuba 2: A Retargetable Forward and Inverse Renderer. *ACM Trans. Graph.* 38, 6, Article 203 (2019).
- Marc Olano and Dan Baker. 2010. LEAN Mapping. In *Proceedings of the 2010 ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*. 181–188.
- Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.
- Pixar Animation Studios. 2016. Universal Scene Description Website. <http://www.openusd.org>.
- Alec Radford, Luke Metz, and Soumith Chintala. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv:1511.06434* (2015).
- Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. 2020. Accelerating 3D Deep Learning with PyTorch3D. *arXiv:2007.08501* (2020).
- RenderPeople. 2020. RenderPeople. <https://renderpeople.com/3d-people/>.
- William Schroeder. 1997. A topology modifying progressive decimation algorithm. In *In VIS '97: 8th conference on Visualization*. 205–212.
- Smithsonian. 2018. Smithsonian 3D Digitization. <https://3d.si.edu/>.
- Olga Sorkine. 2005. Laplacian Mesh Processing. In *Eurographics 2005 - State of the Art Reports*.
- Michael Stokes, Matthew Anderson, Srinivasan Chandrasekar, and Ricardo Motta. 1996. A Standard Default Color Space for the Internet - sRGB. <https://www.w3.org/Graphics/Color/sRGB.html>
- Alejandro Sztrajman, Jaroslav Krivánek, Alexander Wilkie, and Tim Weyrich. 2017. Image-based Remapping of Material Appearance. In *Proc. 5th Workshop on Material Appearance Modeling*. 5–8.
- Justus Thies, Michael Zollhöfer, and Matthias Nießner. 2019. Deferred Neural Rendering: Image Synthesis Using Neural Textures. *ACM Trans. Graph.* 38, 4, Article 66 (2019).
- Michael Toksvig. 2005. Mipmapping normal maps. *Journal of Graphics Tools* 10, 3 (2005), 65–71.
- Julien Valentin, Cem Keskin, Pavel Pidlypenskyi, Ameesh Makadia, Avneesh Sud, and Sofien Bouaziz. 2019. TensorFlow Graphics: Computer Graphics Meets Deep Learning.
- Walt Disney Animation Studios. 2018. Moana Island Scene (v1.1). <http://technology.disneyanimation.com/islandscape/>.
- Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance. 2007. Microfacet Models for Refraction through Rough Surfaces. In *Proceedings of the 18th Eurographics Conference on Rendering Techniques*. 195–206.
- Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. 2018. Pixel2Mesh: Generating 3D Mesh Models from Single RGB Images. In *ECCV*.
- Michael Weinmann and Reinhard Klein. 2015. Advances in Geometry and Reflectance Acquisition (Course Notes). In *SIGGRAPH Asia 2015 Courses*. Article 1.
- Lifan Wu, Shuang Zhao, Ling-Qi Yan, and Ravi Ramamoorthi. 2019. Accurate Appearance Preserving Prefiltering for Rendering Displacement-Mapped Surfaces. *ACM Trans. Graph.* 38, 4, Article 137 (2019).
- Yuxuan Zhang, Wenzheng Chen, Huan Ling, Jun Gao, Yinan Zhang, Antonio Torralba, and Sanja Fidler. 2020. Image GANs meet Differentiable Rendering for Inverse Graphics and Interpretable 3D Neural Rendering. *arXiv:2010.09125* (2020).
- Shuang Zhao, Lifan Wu, Frédéric Durand, and Ravi Ramamoorthi. 2016. Downsampling Scattering Parameters for Rendering Anisotropic Media. *ACM Trans. Graph.* 35, 6, Article 166 (2016).

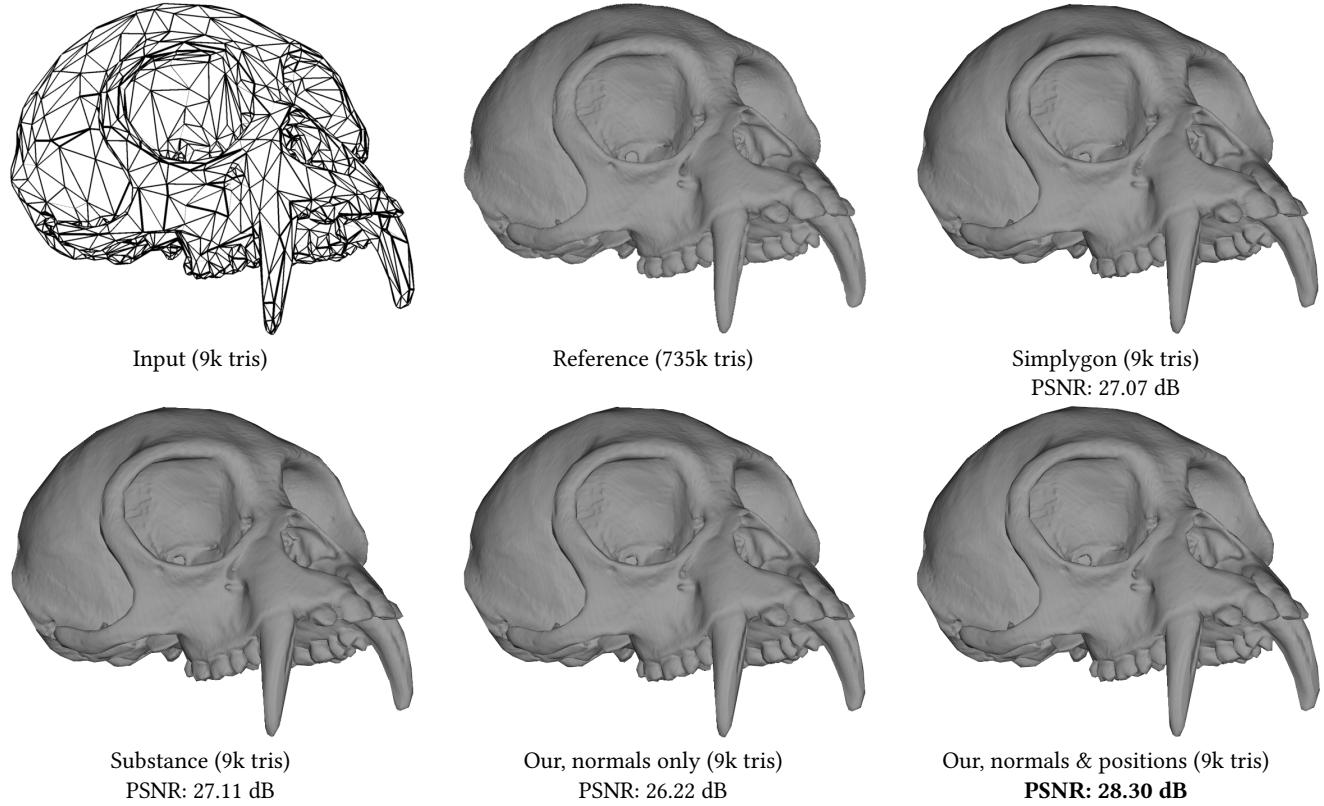


Fig. 14. A comparison with normal map baking. The starting point is a reduced base mesh with 9k triangles (reduced from the 735k triangle reference in Simplygon 8.3) and we bake a normal map from a 735k triangle reference using the normal bakers of Simplygon 8.3 and Substance Painter v2020.2.2. To our knowledge, these bakers only optimize the normal map, and leave the base mesh geometry unmodified. In our version, we jointly optimize the base geometry and normals based on rendered image observations.

A COMPARISON WITH NORMAL MAP BAKERS

In Figure 14 we compare our approach with two production quality normal map bakers (Simplygon 8.3 and Substance Painter v2020.2.2) on a scanned skull, *Hylobates* sp.: Cranium, courtesy of the Smithsonian [2018]. The normal map resolution is 2048 × 2048 texels for all versions and all versions use the same reduced input mesh with 9k tris (generated in Simplygon). To our knowledge, both Simplygon and Substance generate the normal map by ray tracing from the reduced mesh to the high resolution reference. In contrast, our version is optimized from image observations using a resolution of 2048 × 2048 pixels for 10k steps (randomized viewing conditions and lighting) in our differentiable rasterizer. As can be seen, given that we optimize both vertex positions of the base mesh and normal map texels from image observations, we obtain a slightly higher-quality result on the same triangle budget. If we restrict the optimization to only normal map texels, the quality is slightly lower (cf. “Our, normals only” versus “Ours, normals & positions”).

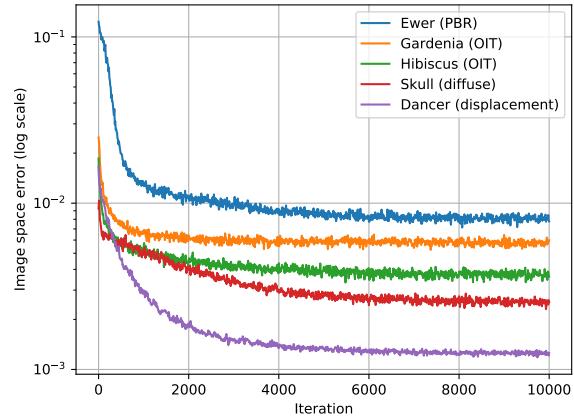


Fig. 15. Training convergence plots for five optimization examples from the paper.

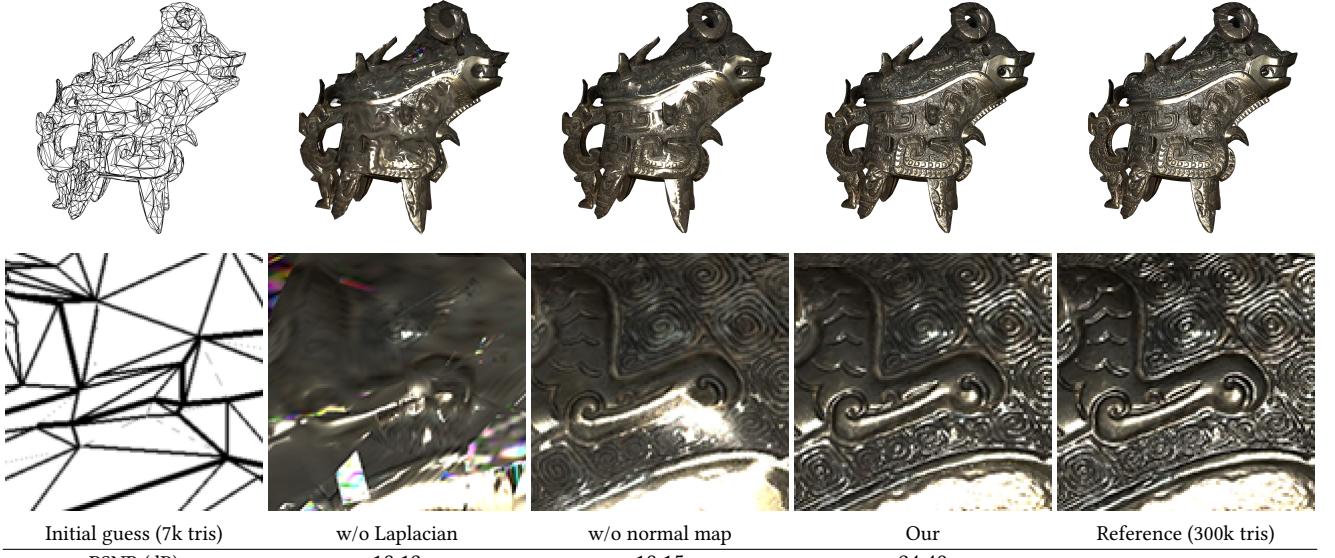


Fig. 16. The Ewer bronze sculpture, courtesy of the Smithsonian 3D Digitization project [2018]. The reference mesh consists of 300k triangles, normal maps, textured base color and a bronze metal material. We start from a reduced version of the reference mesh with 7k triangles, with randomized material parameters. In this case, we obtain a high quality result, closely approximating the reference. The insets highlight the results obtained when we disable either the normal map or the Laplacian regularizer during optimization. We note that both components are critical to obtain high quality results. Without the normal map, we lose high-frequency detail. Without the Laplacian regularizer, the mesh is malformed, which subsequently also blurs the material parameters.

B CONVERGENCE

In Figure 15 we show training convergence plots for five examples from the paper. The normal map baking example (Skull) is the easiest task, with diffuse shading, low dynamic range and joint optimization of the vertex positions and the tangent space normal map texels. The displacement map example (Dancer) performs joint optimization of shape, normal- and displacement maps and uses a lower initial learning rate. The aggregate geometry examples, Gardenia and Hibiscus, are harder, as they include order-independent transparency and PBR shading. Additionally, in those examples, we used an MSE loss function to minimize the differences against the reference when transferring the optimized assets to a path tracer, and the MSE loss is more sensitive to outliers than the tone mapped L_1 variant (see main paper) used in the other examples. Finally, we show the Ewer sculpture, which includes high-frequency specular materials and high dynamic range lighting. Note that the convergence plot show the image space *training* error, and as we randomize the light and camera position each training iteration, some remaining noise is expected throughout the training. We note that all examples converge nicely.

C INFLUENCE OF NORMAL MAP AND LAPLACIAN

In Figure 16 we show how the use of normal map and Laplacian regularizer during the optimization influences the quality of our results. We run 10k steps of optimization at a resolution of 2048 × 2048 pixels. All material textures are initialized to random values, except for the normal map, which is initialized to (0, 0, 1). Normal maps help capture micro-detail, which is clearly visible in the insets.

The Laplacian regularizer helps stabilize optimization and improves mesh quality. Without it, large initial optimization steps may cause mesh corruption or self-intersections which are hard to recover from.

D MESH DECIMATION: VARYING TRIANGLE COUNT IN THE REDUCED MESH

In Figure 17 we study quality as function of the triangle count in our initial guess for the Ewer model. All reduced versions are generated in Simplygon 8.3.

In Figure 18 we study quality as function of the triangle count in our initial guess for the reduced mesh. All versions of the dancer are optimized for 5k iterations using a resolution of 2048 × 2048 pixels. As can be seen, small details and silhouettes benefit greatly from increased triangle count. This result also shows the importance of optimizing the normal map. Details that are not part of the silhouette, e.g., folds in cloth, are captured well even at low triangle counts. Even the statue with 500 triangles does a good job estimating the overall appearance of the reference mesh and could be used as a distant level of detail.

E TESSELLATION AS LEVEL OF DETAIL

Tessellation is often used as a level of detail scheme. Here, we optimize for a single *common* base mesh, displacement map and normal map, with the objective that their renderings reproduce the reference at all levels of tessellation. Figure 19 shows increasing levels

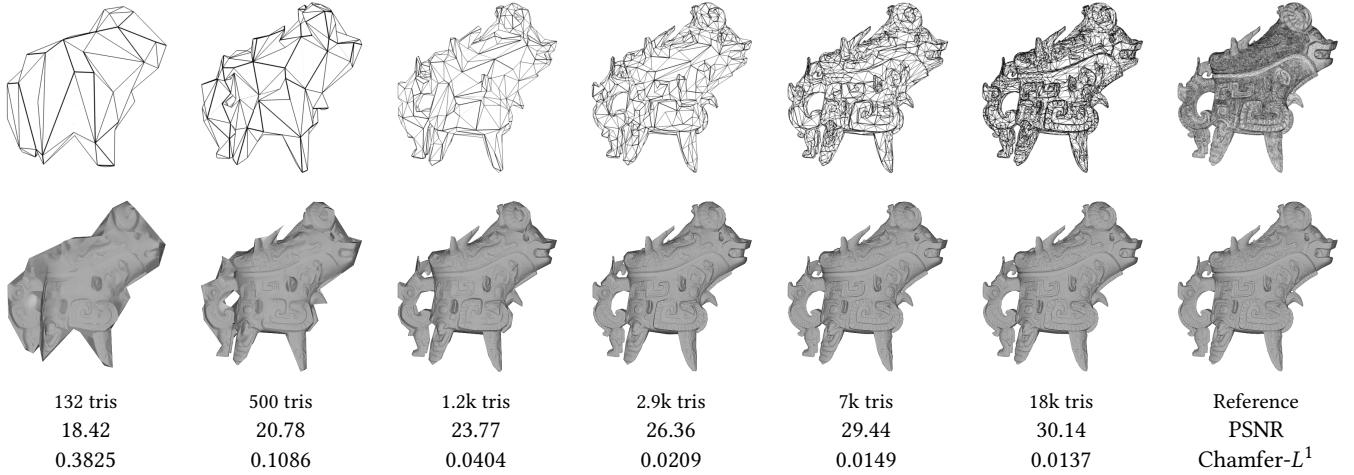


Fig. 17. Influence of initial guess. We show six different input meshes, ranging from 130 to 18k triangles, all trying to approximate a reference mesh with 300k triangles (Ewer model courtesy of the Smithsonian 3D Digitization project [2018]). The Chamfer- L^1 scores are multiplied with a factor 10^3 .

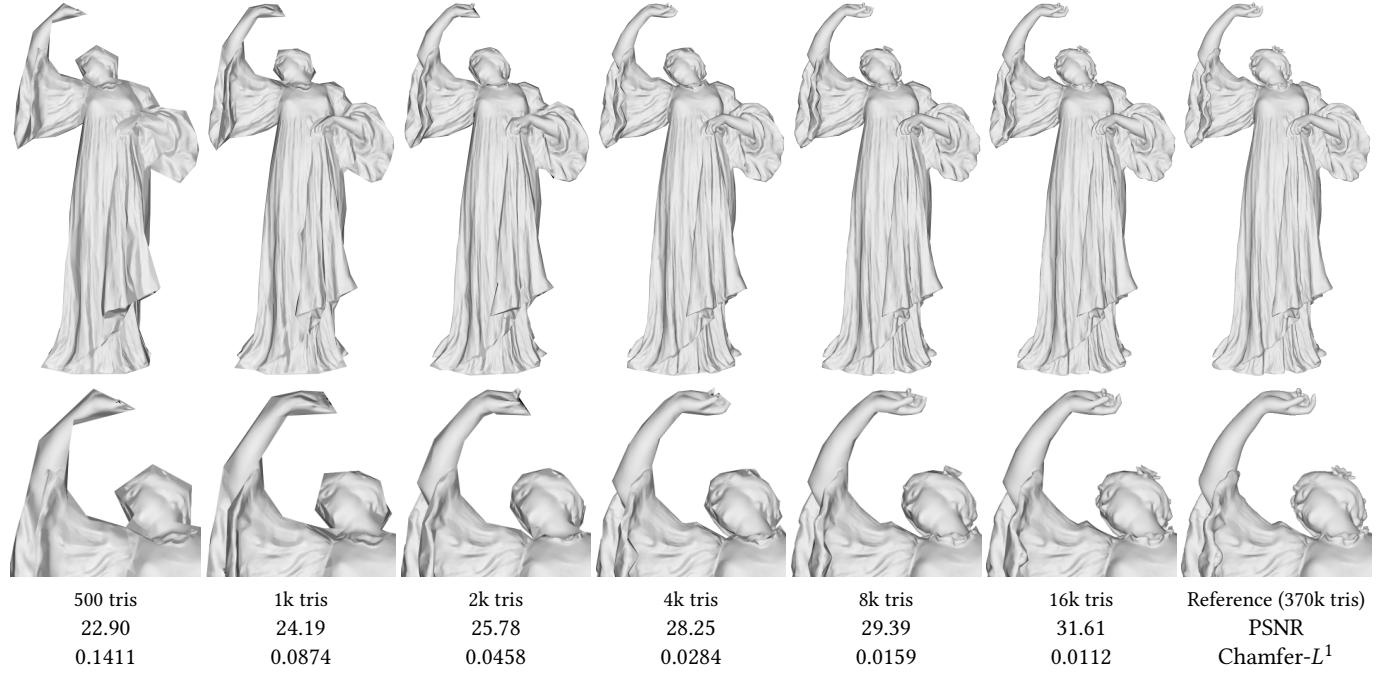


Fig. 18. Starting from an initial guess with varying triangle count, we optimize shape and normal map to match the appearance of the reference. As can be seen in the insets, small details and silhouettes benefit greatly from increased triangle count. The importance of normal mapping is evident. Details not part of the silhouette, e.g., the folds in cloth, are captured even at low triangle counts. The Chamfer- L^1 scores are computed from 1M point samples over the meshes, and are multiplied with a factor 10^3 . The dancer model is courtesy of the Smithsonian 3D Digitization project [2018]

of tessellation on the dancer model from a base mesh with 1k triangles, optimized with 2048×2048 pixels resolution and 5k iterations. We use edge-midpoint-tessellation, wherein each subdivision step quadruples the triangle count. From the insets we note that, as expected, silhouette edges and details are improved as tessellation is increased.

Note that the fingers on the right hand of the figure are never accurately captured even at the highest level of tessellation. Here, the limitation is in the base mesh. Displacement mapping cannot introduce concavities, and since the hand is not sufficiently modeled in the base mesh it cannot be recreated through displacement mapping. This would require increasing the polygon count of the

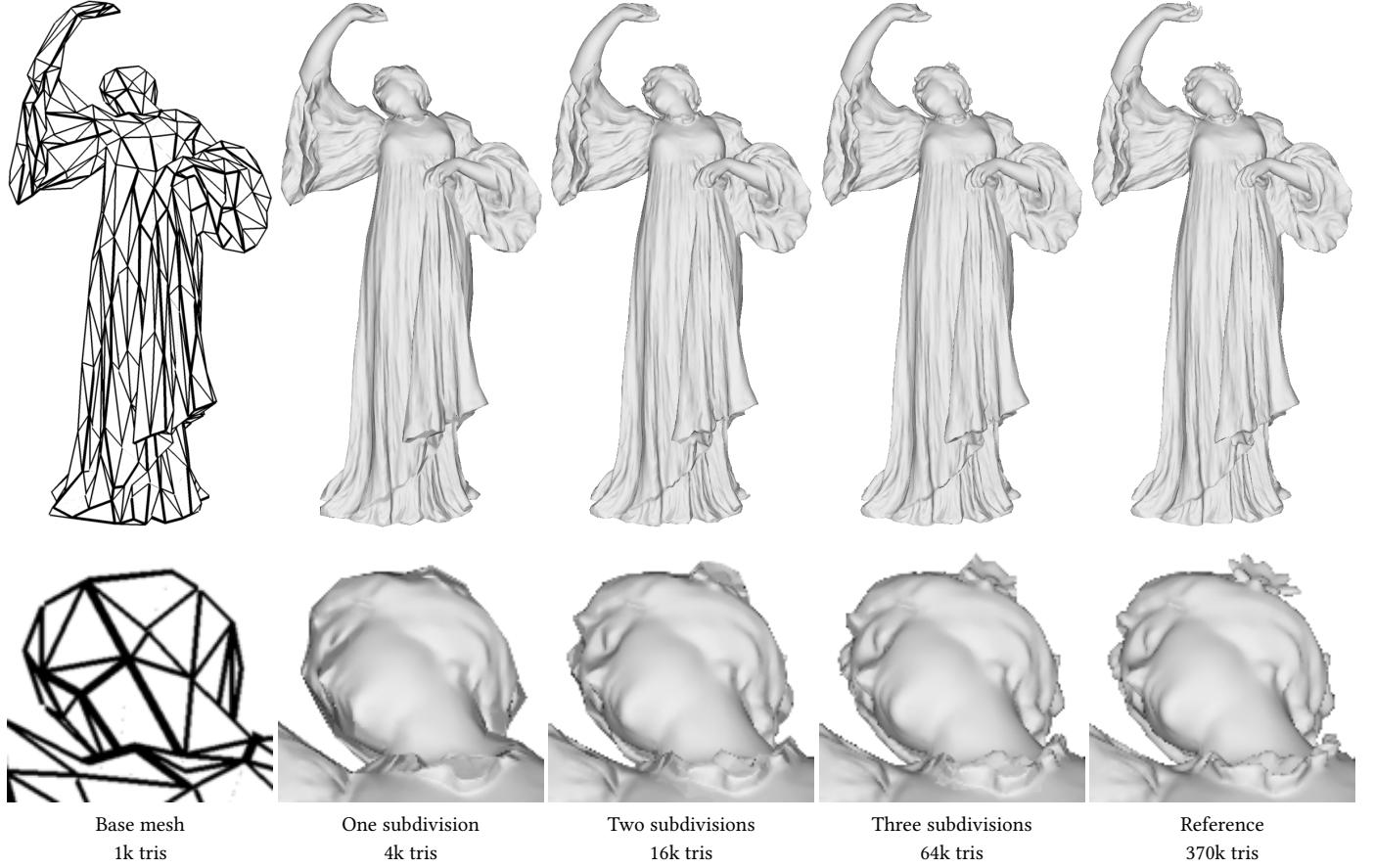


Fig. 19. A level of detail example with different levels of tessellation rendered from a single base mesh. All levels of tessellation use the same base mesh, displacement map and normal map. As can be seen in the insets, silhouette and details are improved with increased tessellation.

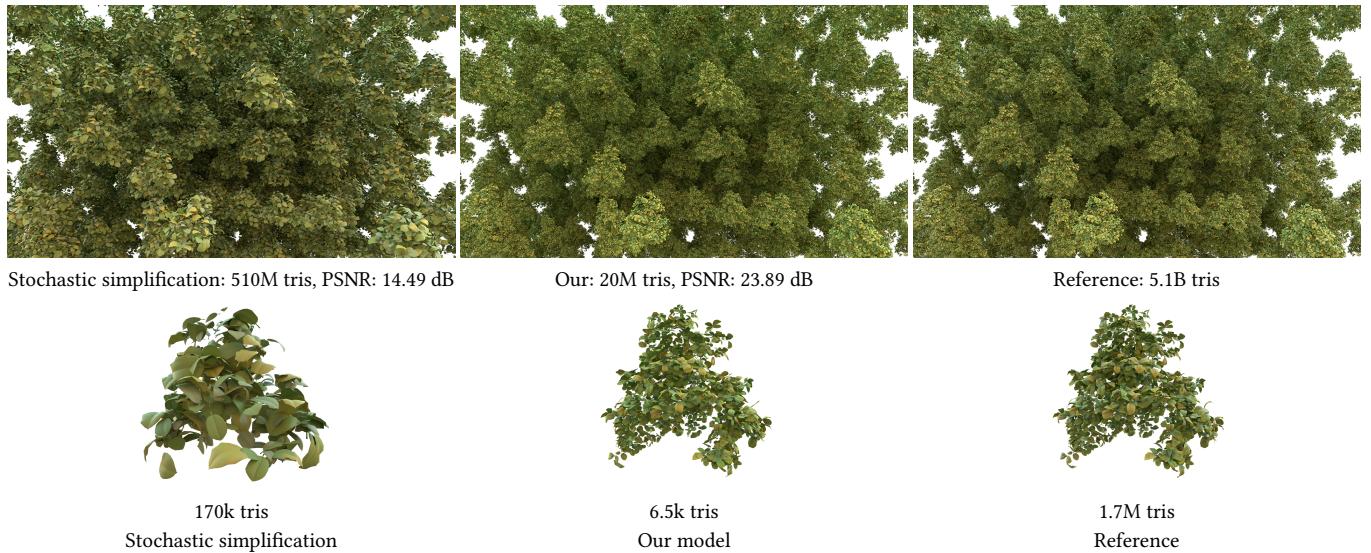


Fig. 20. Our approach compared to stochastic simplification of aggregate detail [Cook et al. 2007].

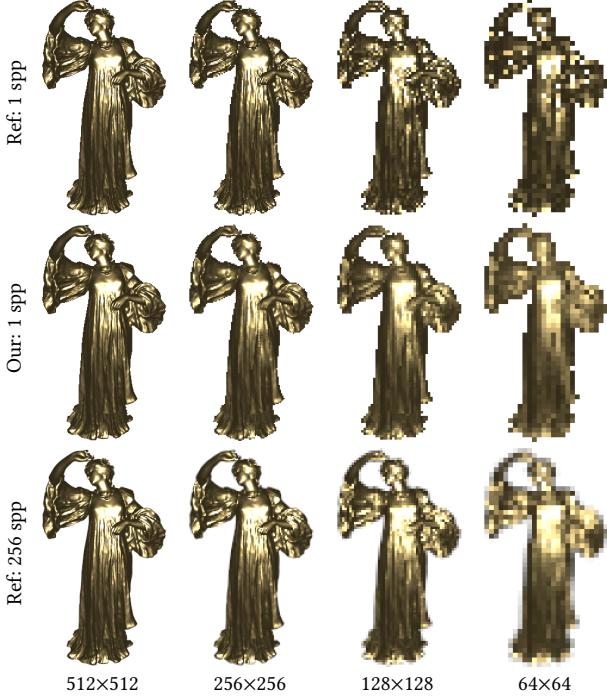


Fig. 21. We perform shape and appearance prefiltering by optimizing for a particular rendering resolution. Here, we show our results for four different resolutions. As expected, geometric details and shading are smoothed as rendering resolution decrease. **Top:** The reference dancer model rendered at 1 spp, note the aliasing. **Middle:** Our optimized model with prefiltered shape and appearance rendered at 1 spp. **Bottom:** The reference dancer model rendered, with antialiasing, at 256 spp.

base mesh, or possibly through an artist improving the initial guess to have higher tessellation in this region.

F APPROXIMATING AGGREGATE DETAIL: COMPARISON WITH STOCHASTIC SIMPLIFICATION

In Figure 20, we compare against stochastic simplification of aggregate detail. Following Cook et al. [2007], we stochastically remove 90% ($\lambda = 0.1$) of the leaves from one instance of the Disney Moana Island Gardenia asset [2018], and adjust the element area of the reduced model by scaling each leaf uniformly with a factor $\sqrt{1/\lambda}$. Finally, the contrast of each leaf texture is adjusted by modifying the color of its texels as $c'_i = \bar{c} + \sqrt{\lambda}(c_i - \bar{c})$, where \bar{c} is the average color of the texture.

Our version is optimized from image observations using a resolution of 2048×2048 pixels for 10k steps (randomized viewing conditions and lighting) in our differentiable rasterizer. We start from an initial guess with 6.5k triangles where each leaf geometry is replaced with a quad and material parameters are randomized. Please refer to the paper for a visual example of the input mesh. We visualize both a single model, and a larger scene with 3000 instances, rendered in a path tracer. Our version, with more aggressive

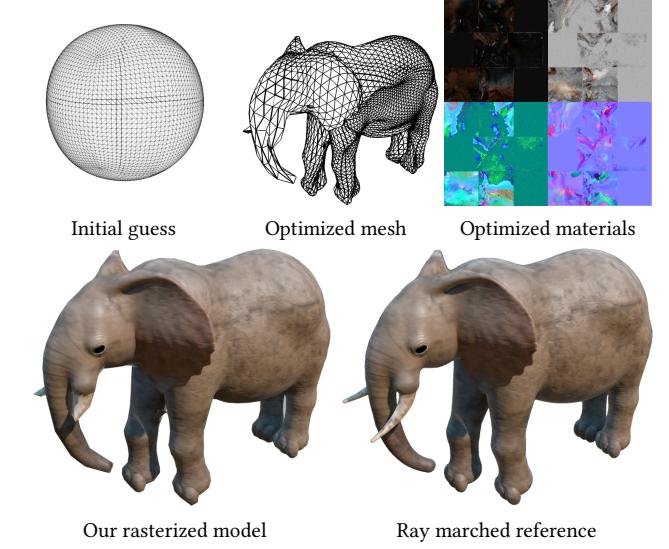


Fig. 22. We extract a mesh and materials from a ray marched distance field: the ShaderToy “Elephant,” from Inigo Quilez. We initialize the optimization process by a sphere with random material parameters and learn shape and material parameters such that our rasterized model resembles the ray ray

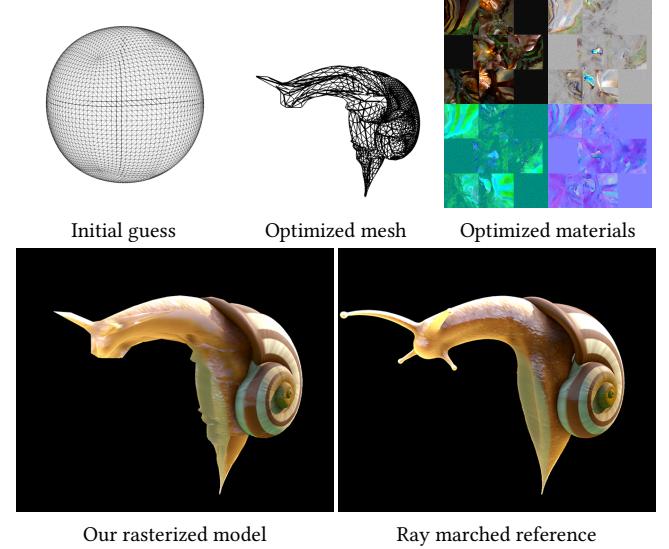


Fig. 23. Similar to Figure 22, we extract a mesh and materials from a ray marched distance field: the ShaderToy “Snail,” from Inigo Quilez.

reduction (99.6% of the triangles removed), still produces a high quality approximation by automatically moving geometry details into transparency- and normal maps. The shape and material parameters are optimized from image observations, so no heuristic for scaling or contrast adjustments is needed.

G SHAPE AND APPEARANCE PREFILTERING

Figure 21 is an extension of Figure 2 in the main paper, and shows the dancer model specifically optimized for four different rendering resolutions. Our results match the appearance of the antialiased reference well, considering the difference in sample count. As expected, geometric detail and shading are gradually smoothed as the rendering resolutions decrease, even though all results are generated from the same initial guess.

H LEARNING MESH AND MATERIALS FROM IMPLICIT SURFACES

In Figure 22 we show an example of learning shape and materials to approximate a signed distance field rendered using ray marching.

We adapt a version of the ShaderToy “Elephant” from Inigo Quilez, modified to isolate the main object. Figure 23 shows a harder example with subsurface scattering, based on the ShaderToy “Snail” from Inigo Quilez, modified to isolate the main object.

We use a sphere with 12k triangles as an initial guess for the rasterizer, and optimize at a resolution of 2048×2048 pixels for 10k steps. The appearance of the shaded result matches the reference well, and the sphere deforms to a reasonable mesh (please refer to the wireframe inset). However, we note that this example is limited by the quality of the initial guess, and further efforts would be required to generalize to more complex assets.

<https://www.shadertoy.com/view/MsXGWr>
<https://www.shadertoy.com/view/ld3Gz2>