# Single-view TSDF Mesh Noise Reduction under Virtual Light Using Differentiable Rendering

PILJOONG JEONG, Gwangju Institute of Science and Technology, Republic of Korea

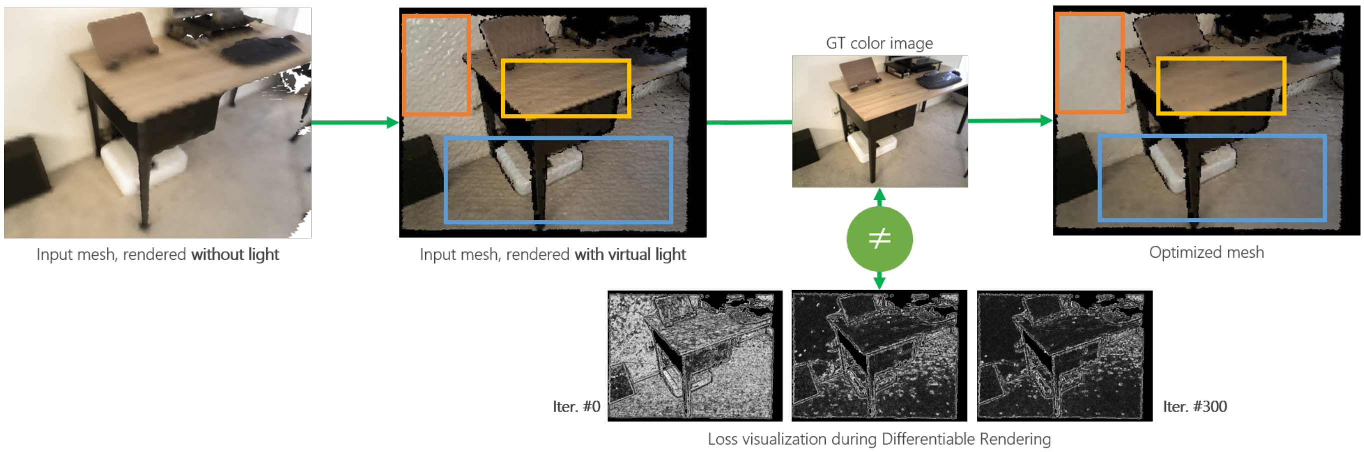


Fig. 1. Overview of our method. (From top left) We saturate noisy vertices by rendering input TSDF mesh with virtually placed light source. We extract common-shared geometric clue between rendered input mesh and target color image. Differentiable renderer iteratively minimizes loss, which is the difference between clues from rendered input mesh and target Ground Truth color image. Orange, Yellow, and Blue inset shows difference between input mesh and optimized mesh. Our method successfully reduces noise in mesh vertices. Furthermore, our provided video shows input mesh that is being optimized (right-side of the video) as iteration continues, as well as the visualization of loss (left-side of the video) that is being minimized: https://drive.google.com/file/d/10F_I89m5O-RWOIxocYoxG2QOkJc7YWlF/view?usp=sharing

Thanks to consistent evolution of SLAM (Simultaneous Localization and Mapping) and its related technologies, we can reconstruct geometric properties of where we are currently observing in real-time. Due to the limitation of current depth sensing hardware, however, we are generally able to obtain geometric features corrupted by noise. Color image is perceived as geometrically noise-free in terms of human vision-perception system, but to our best knowledge, encoding the information from Ground Truth for differentiable rendering under single view constraint is not discussed yet. In this report, we propose a bridge between the geometric information generated from color image and rendered mesh, so that differentiable renderer can optimize input i.e., noisy vertex position without any depth supervision. The key insight is that we can highlight noisy vertices by rendering mesh with virtually placed light sources. We compare our result with one of the state-of-the-art differentiable rendering method[1], and show our method outperforms previous method which requires prior depth information (silhouette image).

CCS Concepts: • **Computing methodologies** → **Shape modeling**; **Mixed / augmented reality**.

Additional Key Words and Phrases: Differentiable rendering, Mesh denoising

**ACM Reference Format:**
PilJoong Jeong. 2021. Single-view TSDF Mesh Noise Reduction under Virtual Light Using Differentiable Rendering. *ACM Trans. Graph.* 1, 1 (June 2021), 2 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Author's address: PilJoong Jeong, Gwangju Institute of Science and Technology, 123 Cheomdangwagi-ro, Buk-gu, Gwangju, 61005, Republic of Korea, piljoong.jeong@gm.gist.ac.kr.


0730-0301/2021/6-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 INTRODUCTION

To satisfy increasing demand of Augmented Reality (AR) technology, we required to obtain geometric information of observed real scene as precise as possible. As precision of geometric detail is increased, we can augment virtual objects with more consistency from computer graphics field (rendering) perspective, and we are able to estimate camera poses from perfect correspondences from computer vision field (SLAM) perspective.

Capturing detailed geometry is still remain as technical challenge due to hardware/computational limitation of SLAM. It is obvious that consumer-level depth sensing technology still has noisy observation. We can cope this by divide a real scene with detailed (fine) voxels as precise as possible when generating TSDF mesh. In this way, however, unnecessarily large number of triangles are generated even in the case of reconstructing simple geometry (e.g., plane). This directly affects to rendering performance, as rendering requires primitive traversal in order to appropriately propagates lighting information for every single frame. Due to those limitations, we are compromised to use TSDF meshes from coarse voxels, which have geometric incorrectness including noise. Nevertheless, there is strong demand to perfect geometry captured from SLAM sequences.

Recent advances of Differentiable Rendering make it possible to optimize current input parameters by observing a set of given Ground Truth images. This seems promising to SLAM, as they naturally capture Ground Truth color images whereas generating input TSDF mesh corrupted by noisy measurements. However, it is unclear that how to interpret (perceptually encoded) geometric clues within color images, reflect those information into actual geometry to minimize its imperfection. Moreover, there is some limitations hinder directly applying previous differentiable rendering techniques into SLAM dataset. We explained such difficulties in Figure. 2.

Our main contribution is bridging the gap between perceptional noise-free geometric features from C and noisy geometric feature in M. To exploit this, we borrow a novel concept of image denoising using flashlight. Images taken with flashlight can hold additional features which are hard to detect from general geometric feature (e.g., depth, normal etc). For example, in [2][3] flashy photography is used to enhance images taken from scene which has insufficient lighting condition. [4] is pioneering work that adopt image enhancement using flashlight on photorealistic rendering domain, by casting virtual flashlights to capture a scene's reflective / refractive features, which are not stored in traditional G-buffers. Based on these approaches, we saturate noisy vertices by casting virtual light, which are never detected when rendered with mesh's albedo only. Please refer the leftmost image (rendered without light) and its connected image (rendered with virtual light) in Figure 1. for details. We demonstrate our results, compare with result from previous method, and show that our method outperforms previous result.

## 2 PREVIOUS WORKS

*TSDF Mesh Noise Reduction in SLAM.* Starting from pioneering work [5], estimating true depth from noisy measurements has been one of the major challenges in SLAM. Basically, [5] first introduced the definition of 'fusion', which is accumulating noisy world positions into a voxel. This acts as similar with spatial running average filter, therefore it is known that accumulating frames that captures same region can reduce noise incrementally. However, this takes a lot of time to converge depth values to a true mean, which interferes practical AR application experience. Therefore, various depth noise reduction techniques are applied to SLAM systems, including simple bilateral filter [6], merging depth information of neighboring frames either offline [7][8], or online [9][10]. All mentioned previous works require multiple depth frames in order to generate reliable mesh which takes a long time, and they did not consider color images, which holds perceptive geometric information. Our method, in contrast, is able to infer geometric clue in color image, thus able to de-noise input mesh without multiple frames.

*Differentiable Rendering.* Differentiable rendering is the technique that optimizing input parameters by observing a set of target images via iteratively render a scene with current state of input parameters (i.e., forward pass), as well as propagate gradients along with computation graph (i.e., backward pass) built at forward pass. Due to the fact that it does not require any prior knowledge (e.g., pre-trained model trained from external dataset) other than target images, this seems an off-the-shelf optimizer for SLAM since we naturally capture target image during scanning process, whereas fused mesh is contaminated with noise due to its nature limitations. However, there are ambiguities that have to be considered in order to adapt differentiable rendering to optimize SLAM problems. Figure 2. describes the different setup between simple mesh optimization and indoor TSDF mesh optimization. Our method bypasses those ambiguities via inferring geometric clue without necessity of silhouette information, as well as multiple target observations.

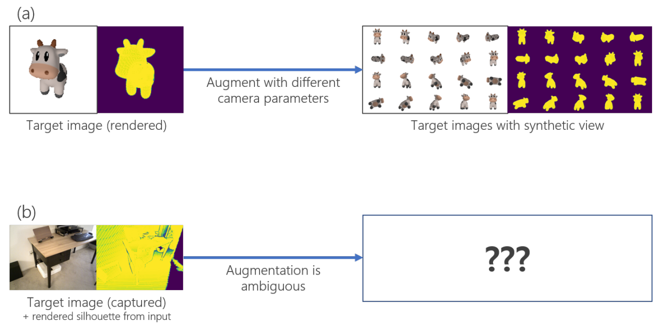


Fig. 2. Different setup between optimizing simple mesh and TSDF mesh from SLAM. Previous works aimed to optimize input geometry to set of target images captured with different camera view either synthetically[11][12][13][14] or by taking calibrated real photographs[14]. However, in the case of TSDF mesh from SLAM it is ambiguous since (1) input is not separated with its backgrounds, hence silhouette image is hard to generate (2) we cannot generate synthetic target views (3) it is hard to sample real images from SLAM sequences which captures a region that input mesh represents, since the labeled camera pose paired with target image is estimated value. It is well-known that camera poses from SLAM is inaccurate. (a) It is straightforward to generate synthetic target images if the target image can be rendered. (b) Unlike (a), it is much hard to augment target images in order to optimize indoor TSDF mesh.

## 3 METHODS

*Assumptions.* Our goal is to minimize TSDF mesh noise fused from a single, Lambertian-dominant indoor RGB-D frame. Given the scenario, we assume that we know intrinsic, extrinsic parameters of camera used to capture the frame. Specifically, we pre-computed input TSDF mesh $\mathcal{M}$ from pair of GT color image $C$ and depth image $\mathcal{D}$ with known intrinsic $K$, such that $\mathcal{M} = TSDF(K^-1(C \oplus \mathcal{D}))$. Here, $\oplus$ denotes image registration operator. We used popular SLAM framework [8] to generate $\mathcal{M}$ with voxel size as 2cm. Since $\mathcal{D}$ is noisy measurement of real scene geometry, $\mathcal{M}$, and its rendering $C$ have slightly different appearance compare with $C$. We additionally constraint that there are no strong texture changes under geometrically close, flat surfaces (e.g., chessboard pattern on flat wall). This assumption is used to ensure that strong radiance change is only made from interaction between light and hit-point geometry within C. We will elaborate this assumption later.