RGB-D-Supported Feed-Forward Gaussian Splatting for Novel View Synthesis

A Camera Capture and Training Pipeline for Sparse-View Scene Reconstruction and Rendering with Measured Depth

Graduate

Nicolas Reichmuth

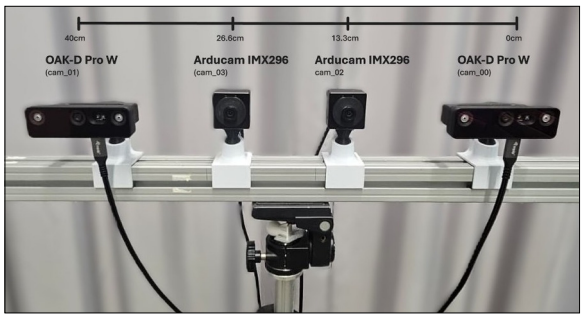
Luca van Brussel

Introduction: Novel view synthesis aims to render a scene from viewpoints that were not captured by a camera. Among recent approaches, 3D Gaussian Splatting represents scenes as sets of anisotropic Gaussian primitives and enables high-quality real-time rendering. However, its original formulation relies on per-scene optimization, which limits its applicability to rapid or live reconstructions. This work investigates whether replacing learned depth estimation with measured RGB-D depth can make feed-forward Gaussian Splatting more suitable for interactive novel view synthesis from sparse views.

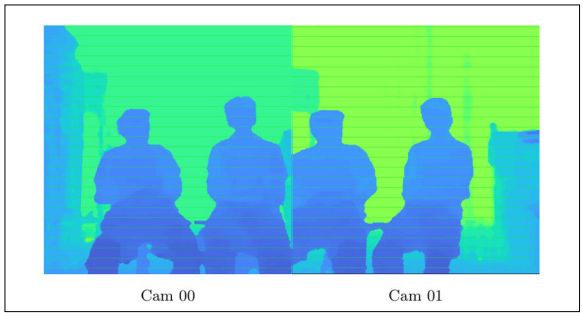
Approach: The pipeline is based on a modified version of GPS-Gaussian+, a feed-forward method that predicts Gaussian scene representations from two rectified source views without per-scene optimization. Initial experiments with the original GPS-Gaussian+ pipeline showed unstable depth estimates on the custom camera setup, due to a domain gap caused by different focal lengths. Therefore, the depth estimation module is bypassed and replaced by metric depth maps from RGB-D cameras. This decouples the pipeline from the internal stereo-depth estimation and provides direct geometric input. A four-camera rig provides two RGB-D source views and two additional target views for training and evaluation. The modified model is fine-tuned with rendering supervision and depth-residual regularization.

Result: On a hold out test-split the prototype achieves a PSNR of 26.9 dB and a SSIM of 0.894. These results indicate that measured RGB-D depth can support novel-view rendering from two source views in the custom indoor camera setup. At a resolution of 1024 × 1024, the proposed RGB-D- supported pipeline reduces the estimated per-frame latency from 338 ms for the original GPS-Gaussian+ pipeline to 136 ms, corresponding to an increase from 3.0 FPS to 7.4 FPS. This suggests that replacing learned stereo-depth estimation with measured RGB-D depth can reduce inference cost in this setup. The frame rate achieved in practice depends on the overhead caused by image acquisition, preprocessing, and data transfer.

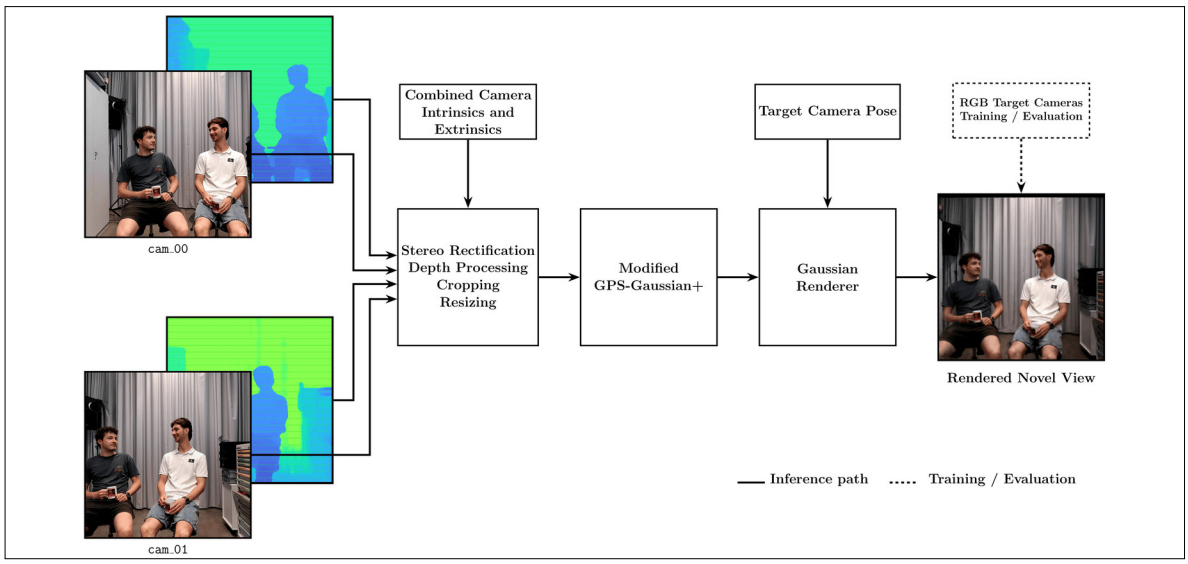
Camera Capture Setup
Own presentment



Rectified, Cropped, and Resized Depth Maps of the Source Cameras
Own presentment



Pipeline Overview
Own presentment



Advisor
Prof. Dr. Martin Weisenhorn

Co-Examiner
Jonas Schmid, Stettbacher Signal Processing

Subject Area
Image Processing and Computer Vision