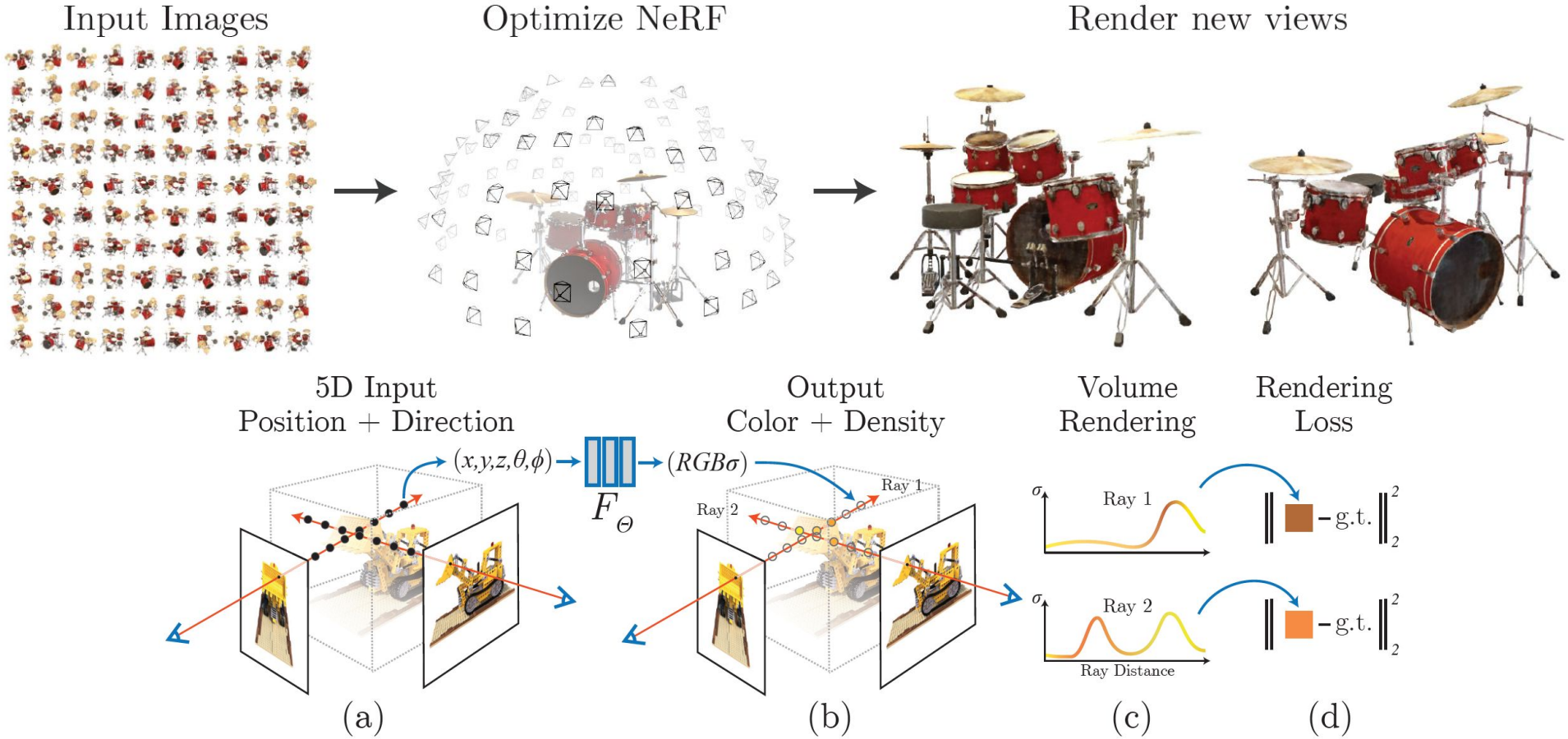
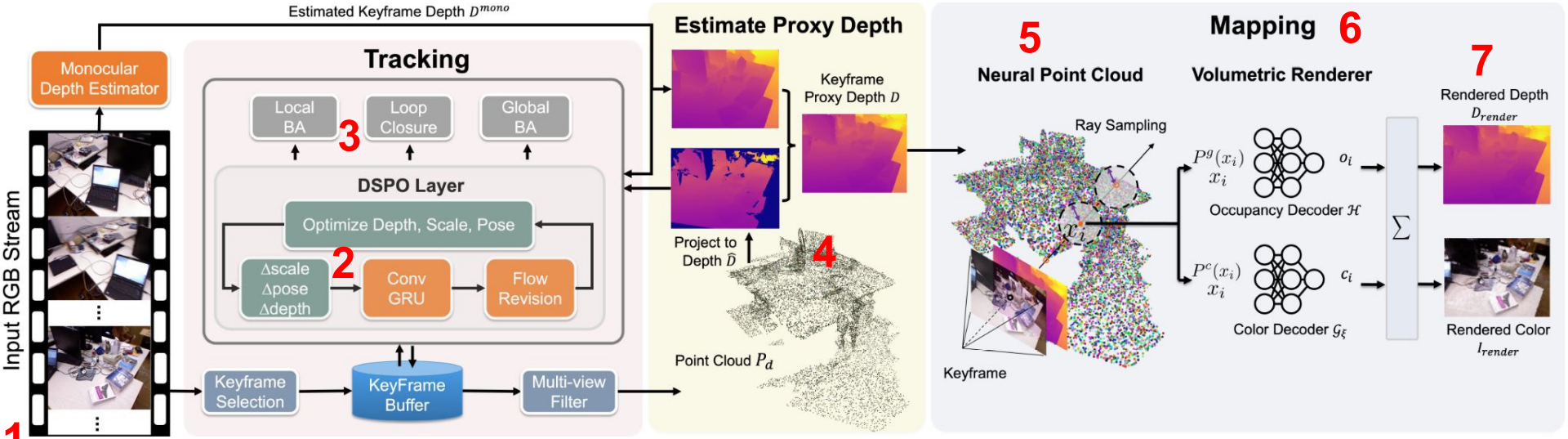
RGB-based Simultaneous Localization and Mapping (SLAM): State of the Art - Zwischenergebnisse

Johannes Decker
20.05.2025

Neural Radiance Fields (NeRFs)

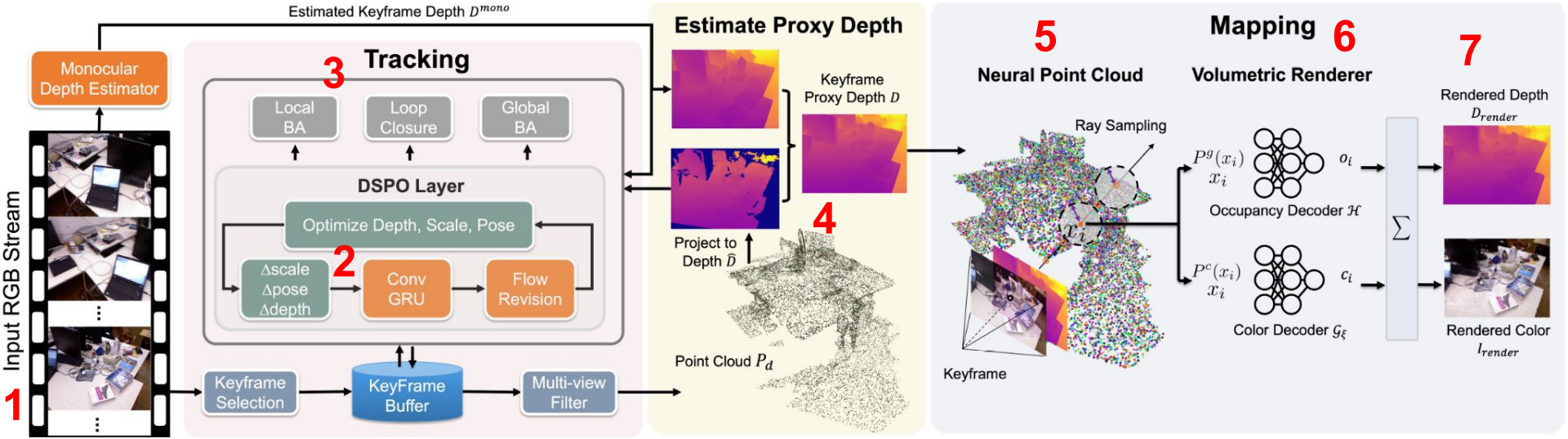


GIORIE-SLAM: Globally Optimized RGB-only Implicit Encoding Point Cloud SLAM



1. Monokulare RGB Kamera Bilder
2. Kameraposition & -tiefe wird relativ durch optischen Fluss geschätzt und optimiert
3. Schleifenschluss durch minimalen optischen Fluss & Globale Bündelanpassung über Keyframes zur Verfeinerung von Position und Tiefe
4. Proxy-Tiefenkartenschätzung durch Kombination aus Keyframe- & monokularer Tiefe

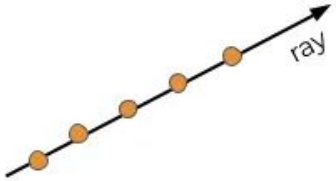
Globally Optimized RGB-only Implicit Encoding Point Cloud SLAM



5. Szenendarstellung durch MLP basierte Punktwolke (Geometrie & Farbe) - Punktpositionen werden anhand Kameraposition und/oder Tiefe aktualisiert
6. Volumen-Rendering: Erzeugen von RGB-Bildern & Tiefenkarten anhand beliebiger Kameraposen
7. Optimierung: Gerenderte Ausgaben werden mit RGB-g.t. und Tiefenkarte verglichen, und aus dem Fehler die MLP basierte Punktwolke aktualisiert

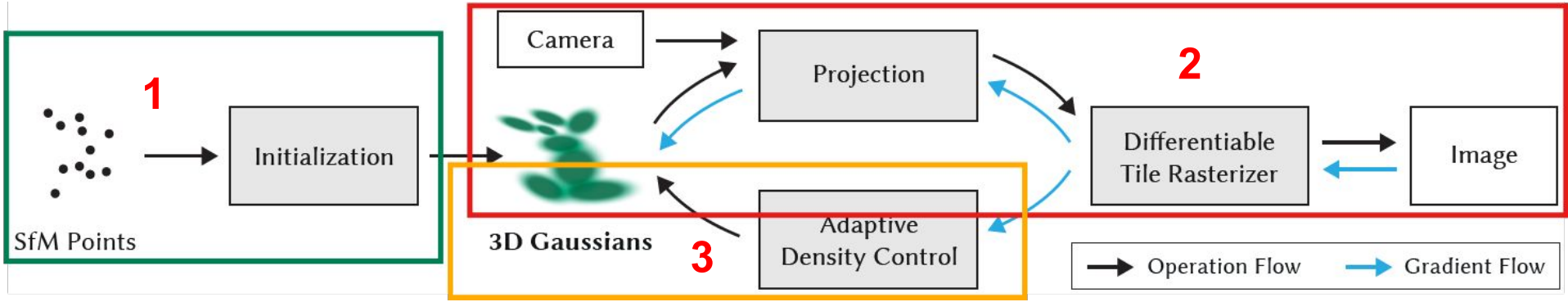
NeRFs vs 3DGS

NeRF



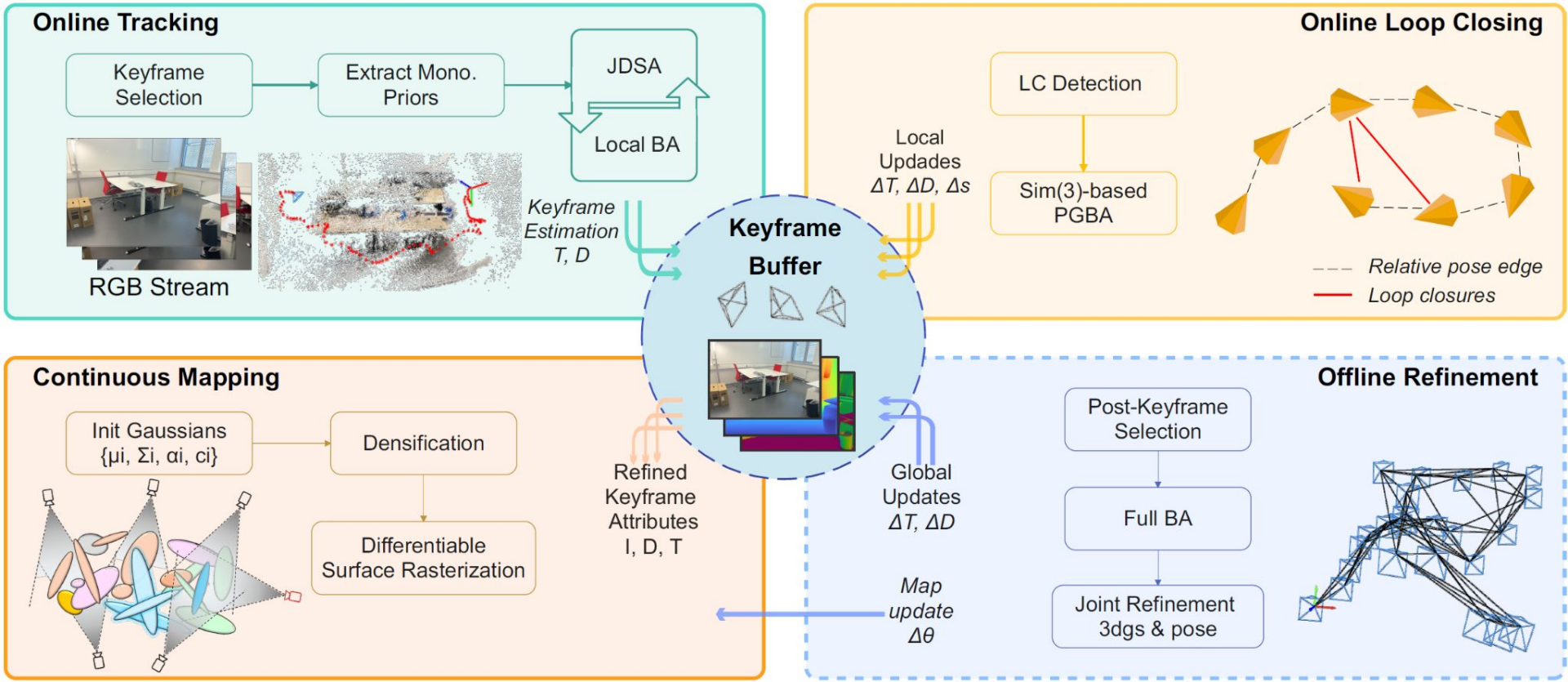
Gaussian Splatting

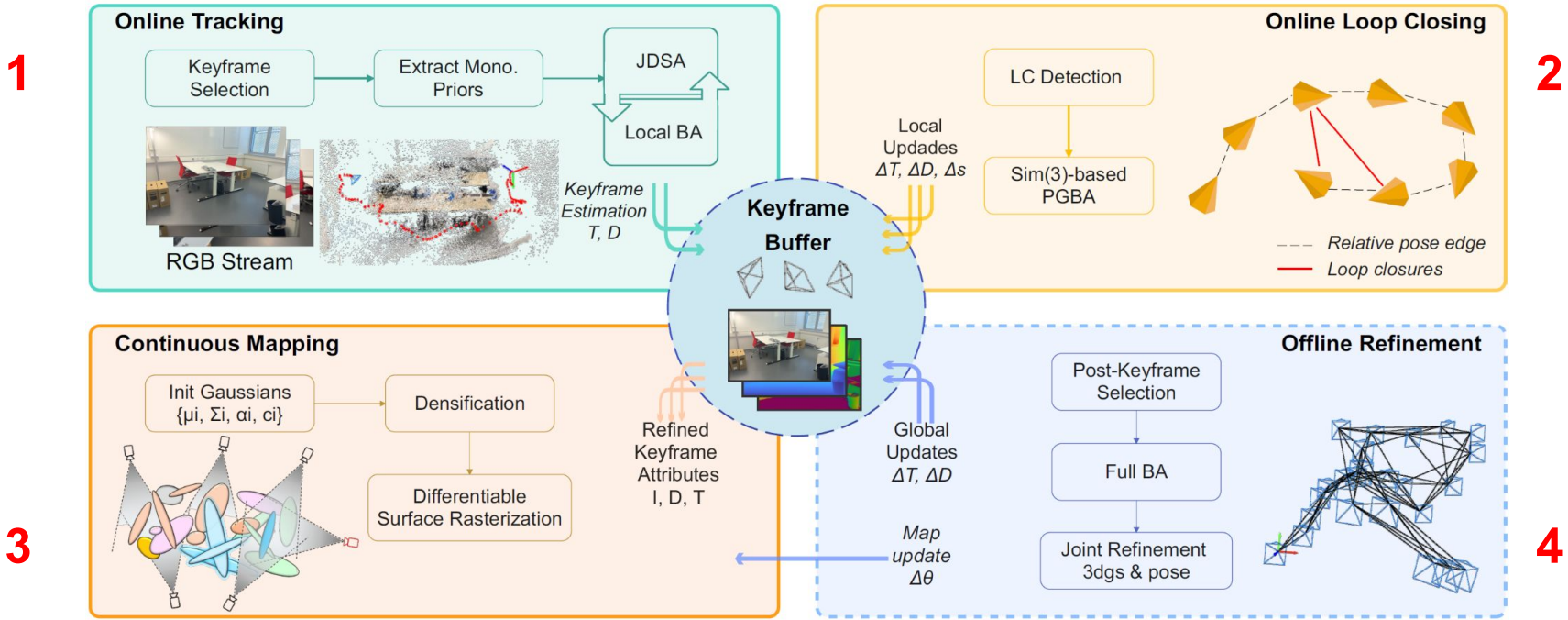
3D Gaussian Splatting (3DGS)



1. Generierung einer 3D-Punktwolke mit Structure from Motion (SfM) & Konvertierung jedes 3D-Punkts in eine Gauß-Verteilung mit Positions-, Form-, Farb- und Opazitätsparametern
2. Projizieren der 3D-Gauß-Verteilungen auf 2D-Bildebenen als elliptische “Splats” & Überblenden von “Splats” zu gerenderten Bildern für Fehlerberechnung zu RGB-g.t.
3. Optimierung der Gauß-Parameter mittels Gradientenabstieg zur Minimierung von Rendering-Fehlern & Verfeinerung des Modells durch Hinzufügen detaillierter Gauß-Verteilungen und das Entfernen von redundanten Verteilungen

HI-SLAM2: Geometry-Aware Gaussian SLAM for Fast Monocular Scene Reconstruction





1. Auswählen von Keyframes aus dem RGB-Stream & Extraktion der monokularen Tiefe, sowie des Normalenvektors. Schätzen der Kamerapose durch die Stärke des optischen Flusses und mithilfe von lokaler Bündelanpassung.
2. Schleifenschluss anhand des optischen Flusses, sowie die Ähnlichkeit der Posen
3. 3D-Gauß-”Splats” aus Tiefeninformation der Keyframes initialisieren und verdichten, sowie die Oberflächen für das Rendering rastern.
4. Post-Keyframes in unterbeobachteten Bereichen einfügen, um die Abdeckung zu verbessern. Vollständige Bündelanpassung und Optimierung 3D-Gauß-”Splats” Parameter

Performance Vergleich in der Kartografierung

Maße für die Rekonstruktionsqualität:

- **SSIM (Structural Similarity Index):** Misst die Ähnlichkeit der Wahrnehmung durch Vergleich von Leuchtdichte, Kontrast und Struktur in lokalen Bildausschnitten (je höher desto besser).
- **PSNR (Peak Signal-to-Noise Ratio):** Quantifiziert die Bildqualität anhand des pixelweisen Fehlers (je höher desto besser).
- **LPIPS (Learned Perceptual Image Patch Similarity):** Misst die Wahrnehmungsunterschiede anhand tiefer Merkmale durch Merkmalsextraktion mit einem DNN wie VGG, AlexNet, oder SqueezeNet (je niedriger desto besser).

Aktuelle Ergebnisse - basierend auf “room0” des Replica Datensatzes

Metrik	HI-SLAM2	GIORIE-SLAM
PSNR (je größer)	35,48	28,49
SSIM (je größer)	0,96	0,96
LPIPS (je kleiner)	0,04	0,13
Genauigkeit der Kameraverfolgung [cm]	0,23	0,31
Genauigkeit [cm]	1,35	2,84
Vollständigkeit [cm]	3,33	4,65
Vollständigkeit [%]	87,45	81,96
Rechenzeit	ca. 7min	ca. 2,5h

Quellen

https://pyimagesearch.com/2024/12/09/3d-gaussian-splatting-vs-nerf-the-end-game-of-3d-reconstruction/

https://medium.com/data-science/a-comprehensive-overview-of-gaussian-splatting-e7d570081362

https://www.themoonlight.io/ko/review/volumetrically-consistent-3d-gaussian-rasterization

Volumetrically Consistent 3D Gaussian Rasterization

How NeRFs and 3D Gaussian Splatting are Reshaping SLAM: a Survey

3D Gaussian Splatting for Real-Time Radiance Field Rendering

NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis

GlORIE-SLAM: Globally Optimized RGB-only Implicit Encoding Point Cloud SLAM

HI-SLAM2: Geometry-Aware Gaussian SLAM for Fast Monocular Scene Reconstruction

Fragen ?

Backup

Neural Radiance Fields (NeRFs) - Formeln

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})dt\,, \text{ where } T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right)$$

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_i\delta_i))\mathbf{c}_i\,, \text{ where } T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j\delta_j\right)$$

$$\delta_i = t_{i+1} - t_i$$

$$\alpha_i = 1 - \exp(-\sigma_i\delta_i)$$

$$t_i \sim \mathcal{U}\left[t_n + \frac{i-1}{N}(t_f - t_n),\ t_n + \frac{i}{N}(t_f - t_n)\right]$$

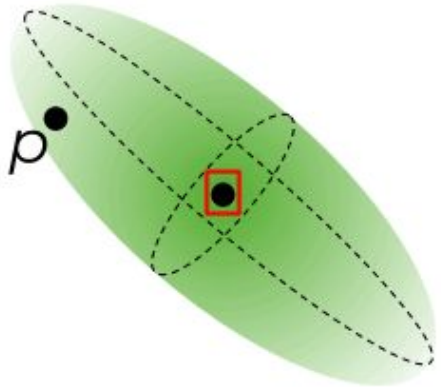
$$\hat{C}\,(\mathbf{r}) = \sum_{i=1}^{N_c} w_i c_i\,, \qquad w_i = T_i(1 - \exp(-\sigma_i\delta_i))$$

$$\mathcal{L} = \sum \|\hat{C}(r) - C(r)\|^2$$

3D Gaussian Splatting (3DGS) - Formeln

Each 3D Gaussian is parametrized by:
Mean μ interpretable as location x, y, z; Covariance Σ; Opacity σ(α); Color parameters, either 3 values for (R, G, B) or spherical harmonics (SH) coefficients.

$$f_i(p) = \sigma(\alpha_i) \exp(-\frac{1}{2}(p - \mu_i)\Sigma_i^{-1}(p - \mu_i))$$



$$C(p) = \sum_{i \in N} c_i f_i^{2D}(p) \underbrace{\prod_{j=1}^{i-1}(1 - f_j^{2D}(p))}_{transmittance}$$

$$\Sigma = RSS^T R^T$$

Structural Similarity Index (SSIM)

Formula (simplified):

$$\text{SSIM}(x, y) = [l(x, y)]^{\alpha} \cdot [c(x, y)]^{\beta} \cdot [s(x, y)]^{\gamma}$$

Where:

- $l(x, y)$ is luminance comparison
- $c(x, y)$ is contrast comparison
- $s(x, y)$ is structural comparison

Usually, $\alpha = \beta = \gamma = 1$, so it's often a straightforward product of the three components.

In summary, SSIM is a more perceptually relevant way to compare image quality than pixel-wise differences.

Peak Signal-to-Noise Ratio (PSNR)

How it works (briefly):

1. **PSNR is derived from Mean Squared Error (MSE)**, which measures the average of the squared differences between corresponding pixels in the two images.

$$\text{MSE} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} [I(i,j) - K(i,j)]^2$$

   Where I is the original image, K is the distorted image, and $m \times n$ is the image size.

2. **PSNR is then calculated using:**

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{MAX^2}{\text{MSE}} \right)$$

   - MAX is the maximum possible pixel value (e.g., 255 for 8-bit images).

Interpretation:

- **Higher PSNR = better quality** (less distortion).
- Typically, **30–50 dB** is a common PSNR range for acceptable to high-quality images.
- **PSNR is in decibels (dB)** — it's a logarithmic scale, so small changes can be perceptually significant.

📏 Accuracy (Acc. [cm] ↓)

- **What it measures:** The average **distance error** (in centimeters) between the reconstructed surface and the ground truth.
- **Goal: Lower is better** (↓) — smaller values mean the reconstruction is closer to the true surface.

🧱 Completeness (Comp. [cm] ↓)

- **What it measures:** The average **distance error from the ground truth** surface to the closest reconstructed point (opposite direction of accuracy).
- **Goal: Lower is better** (↓) — smaller values mean the reconstruction covers more of the actual scene surface.

📊 Completeness Ratio (Comp. Rat [%] ↑)

- **What it measures:** The **percentage** of the reconstructed surface that is within **5 cm** of the ground truth.
- **Goal: Higher is better** (↑) — larger percentages indicate more of the reconstruction is geometrically accurate.