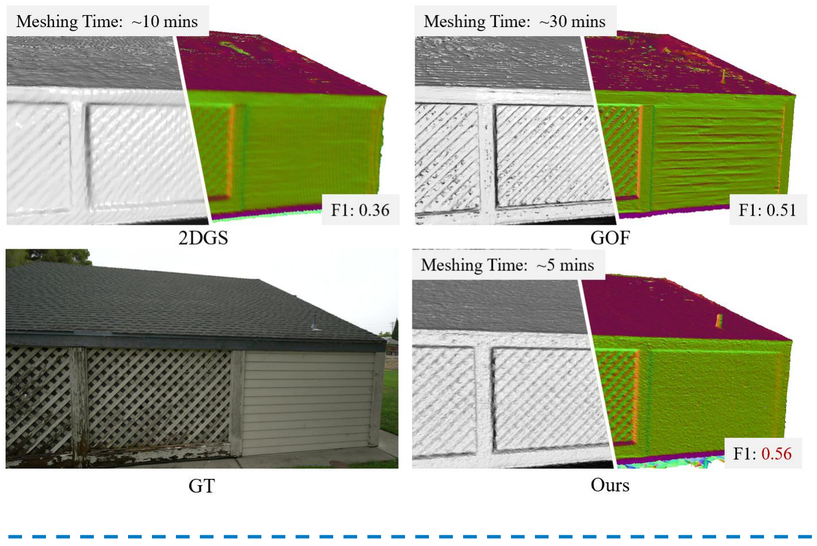
MonoGSDF: Exploring Monocular Geometric Cues for Gaussian Splatting-Guided Implicit Surface Reconstruction

Kunyi Li[1,4] Michael Niemeyer[2] Zeyu Chen[3] Nassir Navab[1,4] Federico Tombari[1,2]
[1]Technical University of Munich [2]Google Zurich [3]Tsinghua University [4]Munich Center for Machine Learning

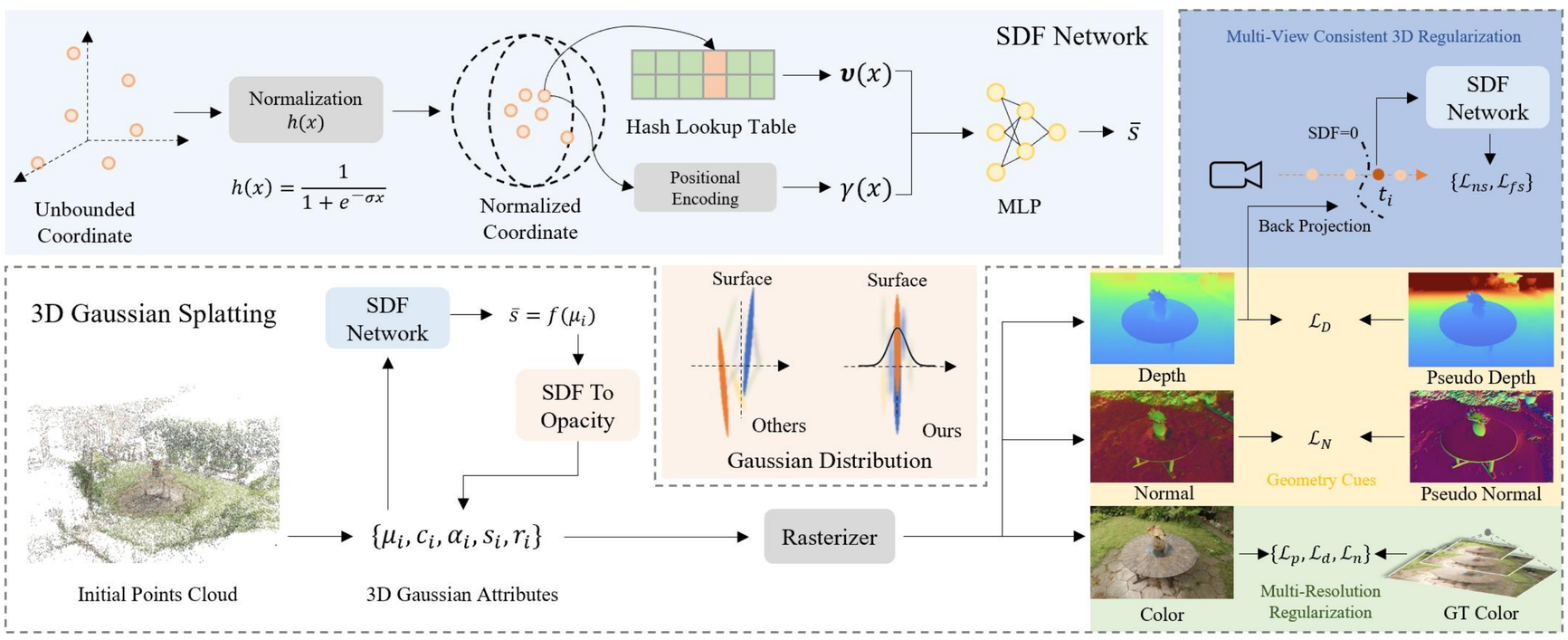
Introduction



To bridge the gap between explicit Gaussian representations and implicit surface reconstruction while fully leveraging their complementary advantages, we present MonoGSDF, a unified framework that establishes a robust connection between 3D Gaussians and SDFbased surface modeling for high-fidelity reconstruction with only monocular RGB images as input. Our contributions are:

- a Gaussian-based implicit surface reconstruction pipeline for monocular images, where the SDF guides Gaussians to better align near the surface during training and Gaussians act as priors for fast mesh extraction;
- a simple yet efficient SDF-to-Opacity function that bridges implicit SDF with explicit Gaussians, coupled with a normalization strategy that maps unbounded 3D coordinates to a (0,1) range with minimal distortion;
- a multi-resolution strategy based on wavelet transforms progressively refines details, while off-the-shelf geometric cues further enhance reconstruction quality. Experiments on real-world datasets demonstrate that our method delivers high-quality reconstructions and outperforms baselines by 13% in terms of Chamfer distance while preserving the view synthesis quality.

Method



3D Gaussian Splatting (3DGS) employs a set of 3D points to effectively render images from given viewpoints, each characterized by a Gaussian function with 3D mean covariance matrix $\boldsymbol{\mu_i} \in \mathbb{R}^3, \Sigma_i \in \mathbb{R}^{3\times3}$, opacity value $\boldsymbol{\alpha_i} \in \mathbb{R}$, RGB color values $\boldsymbol{c_i} \in \mathbb{R}^3$:

$$o_i(\mathbf{x}) = \alpha_i \times \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^T\Sigma_i^{-1}(\mathbf{x}-\mu_i)\right)$$

$$\Sigma_i = R_i S_i S_i^T R_i^T$$

Given a 3D position $\boldsymbol{x}$, $o_i(x$ represents current opacity value contributed by the i-th Gaussian. To facilitate optimization, Σ_i is factorized into the product of a scaling matrix S_i, represented by scale factors $\boldsymbol{s_i} \in \mathbb{R}^3$, and a rotation matrix $\boldsymbol{R_i}$ encoded by a quaternion $\boldsymbol{r_i} \in \mathbb{R}^4$. 3D Gaussians are then projected onto a 2D image plane according to elliptical weighted average to render images for given views. Color $\bar{\boldsymbol{C}}(\boldsymbol{u})$, depth $\bar{D}(\boldsymbol{u})$, and normal $\bar{\boldsymbol{N}}(\boldsymbol{u})$ at pixel $\boldsymbol{u}$ is rendered :

$$\{\bar{\mathbf{C}}, \bar{D}, \bar{\mathbf{N}}\}(\mathbf{u}) = \sum_{i\in N} T_i o_i \{\mathbf{c}_i, d_i, \mathbf{n}_i\}$$

Neural Signed Distance Field:

While 3DGS excels in novel view synthesis, its discrete nature limits surface reconstruction. In contrast, SDF provide a continuous, watertight surface as the zero-level set. We use a hash-based neural representation and jointly train the SDF with 3DGS, enabling more precise and efficient surface extraction.To better encode the scene geometry, we choose One-blob $\gamma(\boldsymbol{x})$ as positional embedding and a multi-resolution hash-based feature grid $\boldsymbol{V} = \{\boldsymbol{V}^l\}_{l=1}^L$. Feature $\boldsymbol{V}(h(\boldsymbol{x}))$ at any 3D point $\boldsymbol{x}$ are queried via trilinear interpolation. The geometry decoder f is an MLP which maps the 3D coordinate to an SDF value $\bar{\boldsymbol{s}}$:

$$f(\mathbf{x}) = f(\gamma(h(\mathbf{x})), \mathcal{V}(h(\mathbf{x}))) \mapsto \bar{s}$$

where $\boldsymbol{h}$ is a coordinate normalization. Therefore, for each 3D Gaussian located at $\boldsymbol{\mu_i}$, its corresponding SDF value can be written as $\bar{\boldsymbol{s}}_i = f(\boldsymbol{\mu_i})$.

Gaussian-Guided Signed Distance Field:

We propose a Gaussian-shaped transformation $g(\bar{\boldsymbol{s}})$ and a Gaussian-guided normalization $h(\boldsymbol{x})$ that maps 3D point $\boldsymbol{x} \in \mathbb{R}^3$ to a fixed $(0,1)$ range :

$$g(\bar{s}_i) = \exp(-(\beta \times \bar{s}_i)^2) \mapsto \alpha_i \qquad h(\mathbf{x}) = 1/(1+\exp(-\sigma\mathbf{x}))$$

Geometry Cues:

We incorporate dense pseudo disparities $\bar{Z}$ from a depth estimator (DepthAnything V2). However, these suffer from scale ambiguity and view inconsistency. We address this by refining pseudo depth as $\hat{D} = a\,/\,(s\bar{Z} + t) + b$ using rendered depth, then derive pseudo normals $\bar{N} = \nabla\bar{D}/|\nabla\hat{D}|\$$ for stable Gaussian supervision:

$$\mathcal{L}_D = \sum |(a/(s\times\hat{Z}+t)+b-\bar{D}| \qquad \mathcal{L}_N = \sum(1-\hat{\mathbf{N}}\cdot\bar{\mathbf{N}})$$

Multi-View Consistent 3D Regularization:

Our approach leverages monocular cues and the 3D Gaussians' advantages, where the Gaussian-rendered depth $\bar{D}$ ensures multi-view consistency and provides reliable 3D SDF supervision via back-projection:

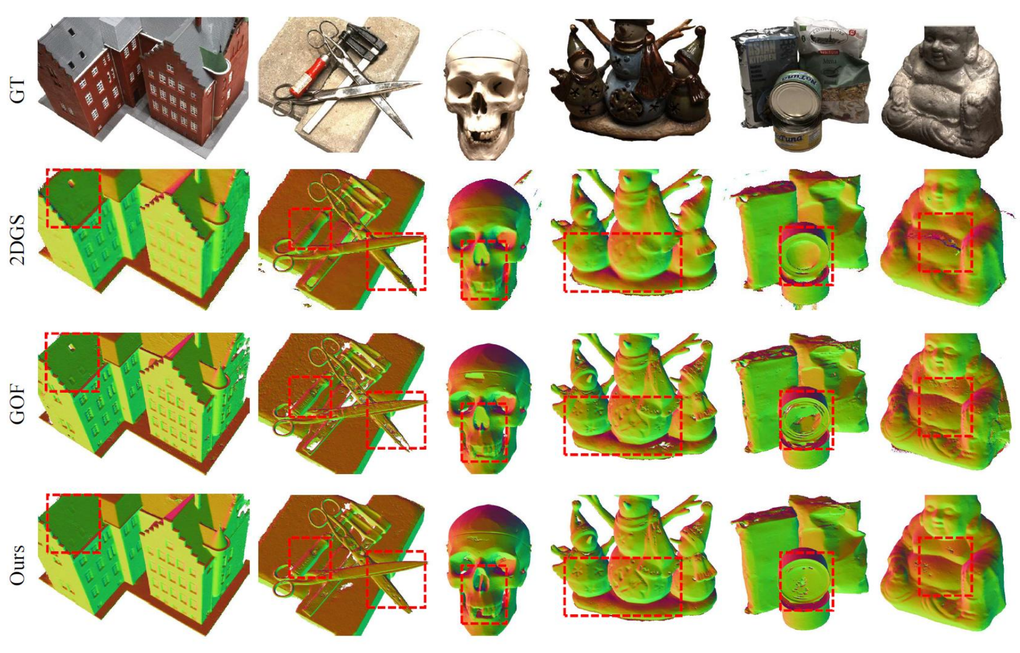
$$\mathcal{L}_{ns} = \sum_{\mathbf{u}_j\in\mathbf{u}}\sum_{t_k\in S_{tr}} ||\bar{s}_k - (\bar{D}(\mathbf{u}_j) - t_k)||_2 \qquad \mathcal{L}_{fs} = \sum_{\mathbf{u}_j\in\mathbf{u}}\sum_{t_k\in S_{fs}} ||\bar{s}_k - 1||_2$$

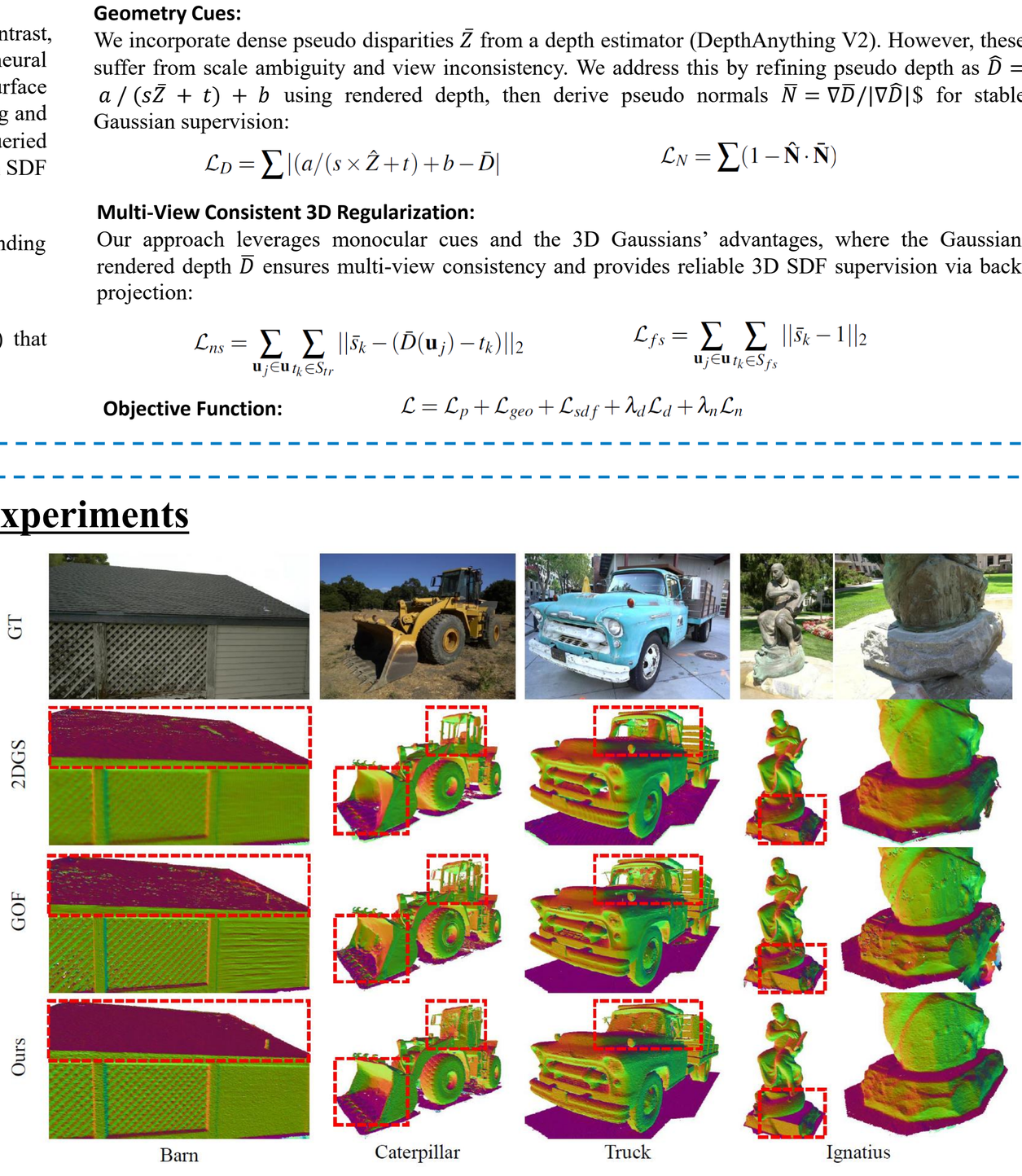
Objective Function: $\mathcal{L} = \mathcal{L}_p + \mathcal{L}_{geo} + \mathcal{L}_{sdf} + \lambda_d\mathcal{L}_d + \lambda_n\mathcal{L}_n$

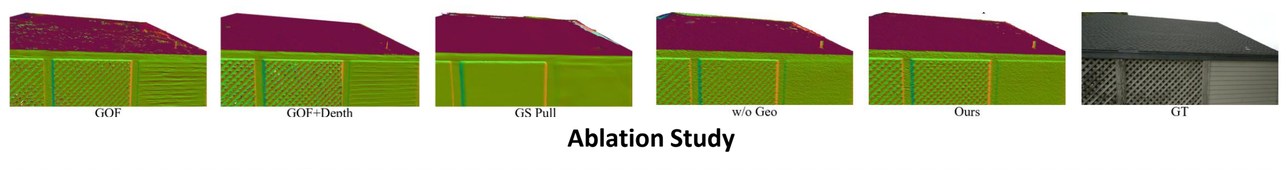
Experiments

	24	37	40	55	63	65	69	83	97	105	106	110	114	118	122	Mean
NeRF [24]	1.90	1.60	1.85	0.58	2.28	1.27	1.47	1.67	2.05	1.07	0.88	2.53	1.06	1.15	0.96	1.49
VolSDF [43]	1.14	1.26	0.81	0.49	1.25	0.70	0.72	1.29	1.18	0.70	0.66	1.08	0.42	0.61	0.55	0.86
NeuS [36]	1.00	1.37	0.93	0.43	1.10	0.65	0.57	1.48	1.09	0.83	0.52	1.20	0.35	0.49	0.54	0.84
N-angelo [20]	0.49	1.05	0.95	0.38	1.22	1.10	2.16	1.68	1.78	0.93	0.44	1.46	0.41	1.13	0.97	1.07
SuGaR [10]	1.47	1.33	1.13	0.61	2.25	1.71	1.15	1.63	1.62	1.07	0.79	2.45	0.98	0.88	0.79	1.33
2DGS [11]	0.48	0.91	0.39	0.39	1.01	0.83	0.81	1.36	1.27	0.76	0.70	1.40	0.40	0.76	0.52	0.80
GSurfel [7]	0.66	0.93	0.52	0.41	1.06	1.14	0.85	1.29	1.53	0.79	0.82	1.58	0.45	0.66	0.53	0.88
3DGSR [23]	0.68	0.84	0.70	0.39	1.16	0.87	0.77	1.48	1.26	0.87	0.69	0.80	0.42	0.64	0.60	0.81
GS-Pull [54]	0.51	0.56	0.46	0.39	0.82	0.67	0.85	1.37	1.25	0.73	0.54	1.39	0.35	0.88	0.42	0.75
GOF [47]	0.50	0.82	0.37	0.37	1.12	0.74	0.73	1.18	1.29	0.68	0.77	0.90	0.42	0.66	0.49	0.74
Ours	0.45	0.65	0.36	0.36	0.94	0.70	0.67	1.27	0.99	0.63	0.49	0.84	0.39	0.53	0.47	0.65

Quantitative Comparison on the DTU. We show the Chamfer distance







Ablation Study