ARTDECO: Towards Efficient and High-Fidelity On-the-Fly 3D Reconstruction with Structured Scene Representation

Guanghao Li[1*2*3], Kerui Ren[1*4], Linning Xu[1*5], Zhewen Zheng[1*6], Changjian Jiang[1*7], Xin Gao[1*2], Bo Dai[8], Jian Pu[2], Mulin Yu[1], Jiangmiao Pang[1]

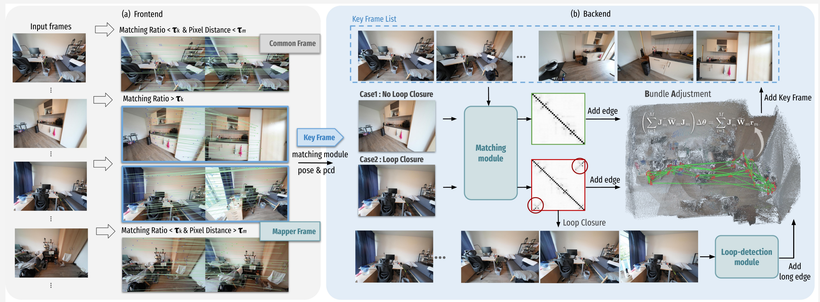
[1]Shanghai Artificial Intelligence Laboratory [2]Fudan University [3]Shanghai Innovation Institute [4]Shanghai Jiao Tong University [5]The Chinese University of Hong Kong
[6]Carnegie Mellon University [7]Zhejiang University [8]The University of Hong Kong

ICLR

Motivation & Overview

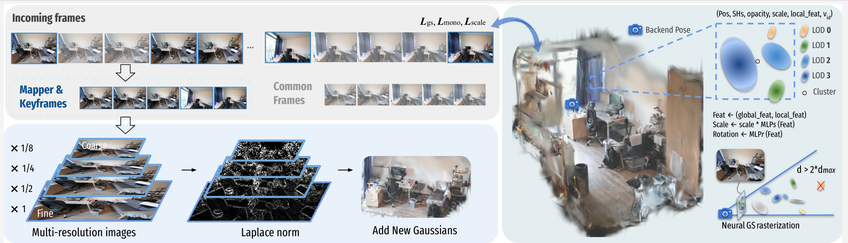
- **Challenge:** Monocular 3D recovers $T_t \in SE(3)$ and geometry online from image sequences, despite ambiguous scale and limited parallax cues.
- **Tradeoff:** 3DGS optimization is accurate but computationally expensive; feed-forward inference is faster in real time, yet often less robust and consistent.
- **Pipeline:** Foundation models predict poses and pointmaps $\{p_i\}$, then a Gaussian decoder maps multi-scale features into structured 3D Gaussians.
- **Evidence:** Across eight indoor/outdoor benchmarks, ARTDECO achieves interactive SLAM-like performance while approaching per-scene optimization quality with higher visual fidelity.

ARTDECO Pipeline



- **Overview:** the figure summarizes a streaming monocular SLAM pipeline mapping RGB frames to poses and Gaussians.
- **Frontend:** MASt3R predicts pointmaps, confidences, and matches to the latest keyframe; pose is recovered by minimizing weighted reprojection residuals.
- **Uncertainty filtering:** each current-frame point estimates local covariance from neighbors within radius δ, then projects it to keyframe space to reject unstable matches.
- **Frame routing:** a frame becomes a keyframe when valid correspondences drop below τ_k; it becomes a mapper frame when 70th-percentile displacement exceeds τ_m.
- **Backend:** each new keyframe links to the latest frame or, after loop detection, to three geometrically consistent predecessors for global BA and drift reduction.

Structured Gaussian Scene Representation



- **Input state:** RGB frames $\{I_i\}$ and poses $\{R_i|t_i\}$ initialize Gaussians $\{\mathcal{G}_i\}$. This yields a compact static 3D scene.
- **Pose-aligned seeding:** MASt3R pointmaps are warped by $SIM(3)$. Reliable pixels seed geometry in a shared world frame.

Rendering Quality Comparison

- **Best visual fidelity:** On Waymo, KITTI, MatrixCity, and Fast-LIVO2, ARTDECO achieves higher PSNR/SSIM and lower LPIPS than MonoGS, LongSplat, and S3PO-GS, with sharper edges and fewer floaters in novel views.
- **Evaluation protocol:** Every 8th frame is held out for testing and excluded from mapping, while poses are still optimized. Quality is reported with PSNR, SSIM, and LPIPS under the same split.
- **Scale robustness:** Across large-scale free-motion Fast-LIVO2 and forward-facing driving datasets, ARTDECO consistently outperforms baselines, showing robustness to scene-scale variation.

Table 1: **Rendering comparisons against baselines** across indoor and outdoor datasets. We report visual quality metrics, average running time.

Indoor-dataset	ScanNet++			ScanNet			TUM			VR-NeRF			Training
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	Time↓
MonoGS	16.71	0.682	0.600	18.87*	0.780*	0.629*	17.78	0.602	0.573	13.88	0.560	0.420	14.08 min
S3PO-GS	22.94	0.820	0.355	20.14	0.797	0.558	19.62	0.656	0.466	12.43	0.642	0.497	41.25 min
SEGS-SLAM	-	-	-	19.73*	0.839*	0.365*	19.69*	0.743*	0.307*	31.62*	0.896*	0.232*	10.84 min
OnTheFly-NVS	18.01	0.761	0.386	15.36	0.708	0.494	19.72	0.719	0.380	27.30	0.872	0.310	**2.29 min**
LongSplat	24.94*	0.827*	0.260*	19.27*	0.754*	0.404*	25.09	0.804	0.272	25.74*	0.832*	0.321*	442.96 min
Ours	**29.12**	**0.918**	**0.167**	**24.10**	**0.865**	**0.271**	**26.18**	**0.850**	**0.224**	**28.57**	**0.895**	**0.242**	5.33 min

Outdoor-dataset	KITTI			Waymo			Fast-LIVO2			MatrixCity			Training
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	Time↓
MonoGS	14.56	0.489	0.767	19.34	0.752	0.627	18.87	0.598	0.699	19.36	0.593	0.736	16.52 min
S3PO-GS	19.97	0.645	0.410	27.28	0.865	0.352	21.51	0.684	0.445	21.76	0.661	0.584	34.89 min
SEGS-SLAM	14.03	0.463	0.488	19.01*	0.698*	0.502*	24.58*	0.773*	0.307*	25.57	0.784	0.366	8.75 min
OnTheFly-NVS	16.89	0.579	0.471	25.53	0.820	0.360	18.76	0.618	0.497	21.36	0.687	0.451	**0.74 min**
LongSplat	16.86	0.532	0.447	25.61	0.795	0.326	26.37	0.792	0.276	-	-	-	313.60 min
Ours	**23.17**	**0.765**	**0.299**	**28.75**	**0.880**	**0.276**	**29.54**	**0.894**	**0.158**	**25.62**	**0.790**	**0.327**	6.58 min

*: majority of scenes successful; –: majority failed; Only compare fully successful methods.

Tracking, Efficiency & Memory

- **Tracking:** ATE RMSE uses every 8th held-out frame.
- **Efficiency:** i9-14900K + RTX 4090; second-fastest after OnTheFly-NVS.
- **Memory:** L=4, k=(0.333W,30), N_a=(23,N_c) bound state.

Table 2: **Tracking comparisons.** For tracking evaluation, we compare against SLAM- and SFM-based 3D reconstruction methods on indoor and outdoor datasets, as well as state-of-the-art SLAM systems on the TUM dataset (Following MASt3R-SLAM, 9 scenes from TUM fr1). Our method consistently achieves lower ATE RMSE.

Dataset	MonoGS	S3PO-GS	SEGS-SLAM	MASt3R-SLAM	OnTheFly-NVS	LongSplat	Ours
ScanNet++	1.217	0.632	0.245	0.025	0.891	0.602	**0.018**
TUM	0.244	0.117	0.073*	0.031	-	-	**0.025**
Waymo	7.370	1.236	-	-	3.118	4.956	**1.213**

Metric	ORB-SLAM3	DPV-SLAM++	DROID-SLAM	Go-SLAM	MASt3R-SLAM	Ours
ATE RMSE	-	0.054	0.038	0.035	0.030	**0.028**

*: majority of scenes successful; –: majority failed; Only compare fully successful methods.

Qualitative & Benchmark Results

- **Cross-dataset gains:** ARTDECO leads on TUM, ScanNet++, KITTI, and Waymo.
- **Evaluation protocol:** Every 8th frame is held out. We report PSNR, SSIM, LPIPS, and ATE RMSE.

Ablations and Design Validation

Table 3: **Quantitative results on ablation studies.** We separately listed the rendering metrics and ATE RMSE on ScanNet++ dataset for each ablation described in Sec. 4.3

Front&Backend	Full	w/ SLAM (MASt3R → π^3)	w/ Loop (π^3 → vggt)	w/o loop	w/ dense key frame
ATE RMSE	**0.018**	0.374	0.096	0.057	0.094

Mapper	Full	w/o level-of-detail	w/o implicit structure	w/o global feat	w/o mapper frame	w/o common frame
PSNR	**29.12**	28.13	28.54	28.89	26.38	27.20
SSIM	**0.918**	0.912	0.914	0.916	0.898	0.904
LPIPS	**0.167**	0.180	0.175	0.170	0.229	0.211

- **Fair evaluation:** every 8th frame is held out. These frames optimize pose only, never rendering supervision.
- **Uncertainty gating:** accept residuals if $(\Sigma c_k) < \tau$.
- **Supplementary validation:** low-texture, extreme-motion, LOD, and multi-resolution studies confirm robustness and antialiasing.

Takeaways, Limitations & Impact

- **Takeaway:** ARTDECO couples feed-forward priors with SLAM updates for monocular 3D, balancing interactive speed and geometry quality across 8 diverse benchmarks.
- **Limitations:** blur, noise, lighting shift, and OOD inputs degrade geometry; low-texture scenes, repetitive structures, and weak parallax can also induce drift artifacts.
- **Impact:** Results on 8 indoor/outdoor benchmarks support online AR/VR, robotics, and real-to-sim digitization with accurate geometry and high visual fidelity.