Multi-Agent Pose Uncertainty: A Differentiable Rendering Cramér–Rao Bound

Arun Muthukkumar
Dept. of Computer Science
Illinois Mathematics and Science Academy
Aurora, United States
amuthukkumar@imsa.edu

***Abstract*—Pose estimation is essential for many applications within computer vision and robotics. Despite its uses, few works provide rigorous uncertainty quantification for poses under dense or learned models. We derive a closed-form lower bound on the covariance of camera pose estimates by treating a differentiable renderer as a measurement function. Linearizing image formation with respect to a small pose perturbation on the manifold yields a render-aware Cramér–Rao bound. Our approach reduces to classical bundle-adjustment uncertainty, ensuring continuity with vision theory. It also naturally extends to multi-agent settings by fusing Fisher information across cameras. Our statistical formulation has downstream applications for tasks such as cooperative perception and novel view synthesis without requiring explicit keypoint correspondences.**

***Index Terms*—Camera Pose Estimation, Uncertainty Quantification, Differentiable Rendering, Multiagent Systems, Neural Radiance Fields (NeRFs), Gaussian Splatting, Cramér–Rao Bound**

I. Introduction

Estimating 6-DoF camera pose from images is foundational in vision and robotics. Differentiable renderers (NeRF [1], Instant-NGP [2], 3D Gaussian Splatting [3]) provide *dense* photometric measurements whose pixels depend on pose, enabling gradient-based alignment; e.g., iNeRF localizes by "inverting" a pretrained field [4]. Missing, however, is a theory of *how accurately* pose can be recovered and how scene content (texture, parallax, symmetries) governs identifiability. To our knowledge, no prior work has derived closed-form *pose CRBs for dense differentiable renderers*.

Classical vision addressed accuracy via the Cramér–Rao bound (CRB), which lower-bounds estimator covariance through Fisher information; in SfM/SLAM, BA pose covariance is given by the inverse reprojection Hessian, guiding design and view planning [5], [6]. These analyses assume *feature-based* measurements. We instead treat the differentiable renderer as the observation model, linearize image formation on $\mathrm{SE}(3)$, and obtain a *render-aware* CRB: with per-pixel Jacobian J and noise Σ, $I(x) = J^\top \Sigma^{-1} J$ and $\mathrm{Cov}(\xi) \succeq I(x)^{-1}$. The eigenstructure of $I(x)$ reveals well-constrained directions (high texture/parallax) and degeneracies (low texture/symmetries), and the formulation recovers BA covariance in the pinhole/feature limit.

We further adopt an *agent* view: each camera supplies local information that is transported and summed to yield a multi-agent bound, enabling cooperative perception, fusion, and communication.

Contributions. (i) A render-aware CRB for camera pose on $\mathrm{SE}(3)$; (ii) practical autodiff recipes for per-ray Jacobians across NeRF/3DGS; (iii) connections to BA/SLAM uncertainty with degeneracy diagnostics; (iv) a compact empirical validation protocol; (v) a multi-agent extension via adjoint transport and information summation.

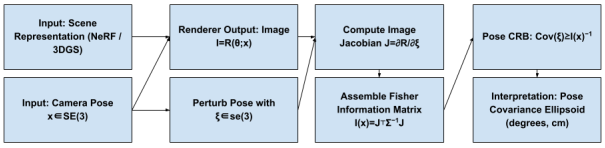

Fig. 1. Pipeline: fixed scene θ and pose $x \to$ render I; autodiff gives $J = \partial R/\partial \xi$; FIM $J^\top \Sigma^{-1} J$; pose CRB $I(x)^{-1}$; interpret as ellipsoids in rotation/translation.

II. Related Works

Differentiable neural renderers have become central to camera pose estimation, enabling analysis-by-synthesis alignment with dense photometric residuals. NeRF [1], Instant-NGP [2], and 3D Gaussian Splatting [3] provide continuous, differentiable image formation, and works such as iNeRF [4] show that a pretrained field can be directly inverted for 6-DoF localization. Extensions refine poses during training itself through bundle-adjusting neural fields [7], demonstrating that differentiable rendering can recover pose without explicit correspondences.

Uncertainty quantification for neural scene representations is a growing theme. Bayes' Rays [8] applies a Laplace approximation to NeRFs to estimate per-pixel confidence, while FisherRF [9] uses Fisher information for view selection and parameter uncertainty. These methods focus on model or scene uncertainty; in contrast, our work targets the uncertainty of *camera pose* given a fixed scene. From a complementary perspective, information-theoretic analyses such as pose-graph SLAM CRBs [5] and the Fisher Information Field for active localization [6] show how Fisher information can guide sensing, though they assume sparse feature measurements. We extend this line by formulating a dense photometric Fisher matrix via differentiable rendering, allowing us to quantify pose identifiability directly from image formation.

Finally, multi-agent and geometric estimation contexts also motivate our formulation. Distributed SLAM systems like Kimera-Multi [10] and COVINS [11] highlight the benefits of sharing information across agents, and our multi-agent CRB formalizes fusion by transporting and summing per-camera Fisher information. More broadly, our derivation follows established statistical estimation on manifolds: Barfoot's SE(3) treatment [12], Solà's micro-Lie calculus [13], and Absil et al. on Riemannian optimization [14], ensuring invariance and principled reporting of covariance in the tangent space.

III. Methodology

We define pose estimation as recovering $x \in \mathrm{SE}(3)$ from an image $I \in \mathbb{R}^M$ generated by a differentiable renderer

$$I = R(\theta; x) + \eta, \qquad \eta \sim \mathcal{N}(0, \Sigma), \tag{1}$$

with fixed scene θ and pixel-noise covariance $\Sigma \in \mathbb{R}^{M\times M}$. Let $\xi \in \mathfrak{se}(3)$ be a minimal twist so that the perturbed pose is $\exp(\xi)\, x$. Linearizing at $\xi = 0$ gives

$$\begin{aligned} R(\theta; \exp(\xi)\, x) \;&\approx\; R(\theta; x) + J\,\xi, \\ J \;&=\; \left.\frac{\partial R(\theta; \exp(\xi)\, x)}{\partial \xi}\right|_{\xi=0} \in \mathbb{R}^{M\times 6}. \end{aligned} \tag{2}$$

A. Core Derivation

Theorem 1 (Render-aware Fisher information on SE(3))**.** *Under the Gaussian model* (1) *and linearization* (2)*, the Fisher Information Matrix (FIM) for the local pose parameter ξ is*

$$\mathcal{I}(x) \;=\; J^\top \Sigma^{-1} J \;\in\; \mathbb{R}^{6\times 6}, \tag{3}$$

and the (unbiased) Cramér–Rao bound (CRB) on the local pose covariance is

$$\mathrm{Cov}(\hat{\xi}) \;\succeq\; \mathcal{I}(x)^{-1}. \tag{4}$$

If $\mathcal{I}(x)$ is singular, interpret (4) *using the Moore–Penrose pseudoinverse $\mathcal{I}(x)^+$.*

Proof sketch. For Gaussian η, $\log p(I \mid x) = -\frac{1}{2}(I - R(\theta;x))^\top \Sigma^{-1}(I - R(\theta;x)) + \text{const}$. Differentiating w.r.t. ξ through (2) gives the score $\nabla_\xi \log p = J^\top \Sigma^{-1}(I - R(\theta;x))$ with mean 0 and covariance $J^\top \Sigma^{-1} J$. The standard definition of the FIM as the covariance of the score gives $\mathcal{I}(x)$. The CRB follows. □

Reparameterization invariance.

Proposition 1. *Let $\phi : \mathbb{R}^6 \to \mathbb{R}^6$ be a local diffeomorphism relating minimal* SE(3) *coordinates ξ and $\zeta = \phi(\xi)$. Then $\mathcal{I}_\zeta = (D\phi)^{-\top} \mathcal{I}_\xi (D\phi)^{-1}$, so the CRB* (4) *is invariant up to coordinate change.*

Remark. The bound is well-defined on the manifold; we report rotation in degrees and translation in scene units.

Identifiability.

Lemma 1. *If $\mathrm{rank}(J) = 6$ on a set of nonzero measure pixels, then $\mathcal{I}(x)$ is full-rank and all pose directions are identifiable. If J loses rank (e.g., planar wall, radial symmetry), $\mathcal{I}(x)$ is singular and the CRB diverges along nullspace directions.*

Classical BA as a special case.

Corollary 1. *If R reduces to pinhole projection of 3D points $\{\mathbf{X}_k\}$ with per-point Gaussian noise $\sigma^2 I_2$, then stacking reprojection Jacobians $J_k = \partial\pi(K[R|t]\mathbf{X}_k)/\partial\xi$ gives $J = \mathrm{blkrow}(J_k)$ and $\mathcal{I}(x) = J^\top(\sigma^{-2}I)J$, the Gauss–Newton Hessian of BA, so the CRB coincides with BA covariance.*

B. Multi-Agent Extension

This extension is critical for cooperative perception, where each camera contributes partial but complementary Fisher information.

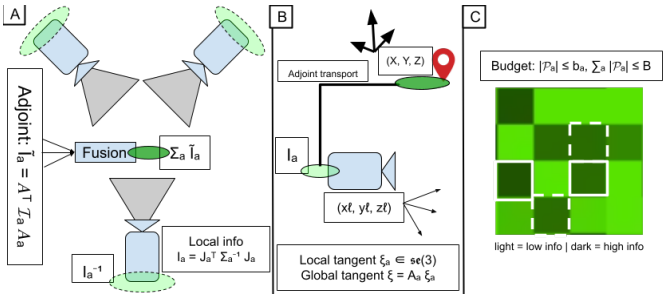

Fig. 2. A) Multi-agent fusion of Fisher information. B) Adjoint transport from local to global tangent. C) Bandwidth-aware tile selection under budget constraints.

Multi-agent FIM. For agents $a = 1{:}A$ with image Jacobians J_a and noise Σ_a, the per-agent information in the agent's local tangent is $\mathcal{I}_a = J_a^\top \Sigma_a^{-1} J_a$. To fuse in a global pose tangent (about x), we transport via the $SE(3)$ adjoint: $\tilde{\mathcal{I}}_a = A_a^\top \mathcal{I}_a A_a$, where $A_a = \mathrm{Ad}_{g_a^{-1}}$ maps the agent's local perturbations to the global frame (here g_a is the relative transform between frames, Fig. 2B). A concrete form is

$$\mathrm{Ad}_g \;=\; \begin{bmatrix} R & [t]_\times R \\ 0 & R \end{bmatrix}, \quad g = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} \in SE(3),$$

with $[t]_\times$ the skew-symmetric matrix of t. Under conditional independence of pixel noise given (θ, x), the joint information is

$$\mathcal{I}_{\text{joint}}(x) = \sum_{a=1}^{A} \tilde{\mathcal{I}}_a.$$

In an information-filter view, communicating $\tilde{\mathcal{I}}_a$ (or its Cholesky/eigen-sketch) yields consistent fusion under bandwidth limits (Fig. 2A).

Bandwidth-aware agent/tile selection. Partition each image into tiles $\{\mathcal{T}_{a,t}\}$ with tile-level Fisher blocks $\tilde{\mathcal{I}}_{a,t}$ (Fig. 2C). Given per-agent budgets b_a and a global budget B, select $\mathcal{P}_a \subseteq \{\mathcal{T}_{a,t}\}$ to maximize

$$f\Big(\mathcal{I}_0 + \sum_a \sum_{t\in\mathcal{P}_a} \tilde{\mathcal{I}}_{a,t}\Big), \quad \text{s.t.} \;\sum_a |\mathcal{P}_a| \le B, \;\; |\mathcal{P}_a| \le b_a.$$

We use $f \in \{\log\det(\cdot),\, \mathrm{tr}(\cdot),\, \lambda_{\min}(\cdot)\}$. $\log\det$ is monotone submodular (greedy gives a $(1-1/e)$ approximation under cardinality/partition constraints), tr is modular (greedy is optimal), while $\lambda_{\min}$ is not submodular (greedy is a heuristic). In practice we add a small ridge ϵI for numerical stability when computing f.

C. Computing J in practice (autodiff and VJPs)

Forming J explicitly via per-pixel gradients is memory-intensive, so we exploit vector-Jacobian products (VJPs): for any $v \in \mathbb{R}^M$, autodiff gives $J^\top v$ without materializing J. This suffices to assemble $\mathcal{I}(x) = J^\top \Sigma^{-1} J$, applying Σ^{-1} implicitly; for diagonal or block-diagonal Σ, this is cheap. Pixel subsampling and tiling further reduce cost.

Algorithm 1 CRB via implicit Jacobians (JVPs)

Require: Renderer $R(\theta; x)$; pose x; noise model Σ (apply $w \leftarrow \Sigma^{-1} v$); pixel subset $\mathcal{P} \subset \{1, \dots, M\}$

Define $f(\xi) = R(\theta; \exp(\xi)x)$ and evaluate at $\xi = 0$
for $j = 1$ to 6 **do**
  $q_j \leftarrow \mathrm{JVP}_f(e_j)$ restrict to pixels $\mathcal{P}$ ▷ column j of J
  $u_j \leftarrow \Sigma^{-1} q_j$ ▷ elementwise if Σ is (block-)diagonal
end for
$\mathcal{I}_{ij} \leftarrow \langle q_i, u_j \rangle_{\mathcal{P}}$ $(i, j = 1, \dots, 6)$ ▷ $\mathcal{I} = J^\top \Sigma^{-1} J$
return $\widehat{\mathcal{I}}(x)$ and

$$\widehat{C} = \begin{cases} \widehat{\mathcal{I}}^{-1}, & \text{if } \widehat{\mathcal{I}} \text{ is PD,} \\ \widehat{\mathcal{I}}^{+}, & \text{otherwise (Moore-Penrose, optional ridge } \epsilon I) \end{cases}$$

Complexity. With $|\mathcal{P}|$ sampled pixels, forming $\mathcal{I}(x)$ requires six columns Je_j and their weighted inner products: $O(6|\mathcal{P}|)$ renderer VJPs plus cheap reductions for diagonal Σ. For $|\mathcal{P}| = sM$ (subsampling rate $s \in (0, 1]$), cost scales linearly in sM. Tiling lowers memory, and blockwise accumulation avoids storing J, making the method practical for 512^2 images on modern GPUs.

D. Modeling Assumptions and Robustness

Noise. The derivation holds for general (possibly correlated) Σ; in practice, per-pixel variances $\hat{\Sigma} = \mathrm{diag}(\hat{\sigma}_i^2)$ can be estimated from residuals. Larger noise weakens the bound. **Photometry.** Illumination drift or tone-mapping mismatches bias J and the FIM; normalization, learned $\hat{\Sigma}$, or restricting to gradient-rich pixels help mitigate this. **Bias.** The CRB assumes unbiased estimators; at high SNR, MLEs approach the bound. Biased extensions (e.g., van Trees) are possible but not pursued.

Interpretation. Report $\sqrt{\mathrm{diag}(\mathcal{I}(x)^{-1})}$ as 1σ pose bounds (rotation in degrees, translation in scene units); eigenvalues of $\mathcal{I}(x)$ reveal ill-conditioning.

Recipe. (i) Freeze θ; (ii) treat pose as 6D input; (iii) compute Je_j via autodiff on a pixel subset; (iv) weight by Σ^{-1}; (v) assemble and invert (or pseudoinvert) $\mathcal{I}(x)$; (vi) inspect eigenstructure.

IV. Experiments

Code released at https://github.com/ArunMut/Multi-Agent-Pose-Uncertainty

We validate the render-aware CRB on Instant-NGP [2] and 3D Gaussian Splatting [3] using LLFF (texture-rich) and Tanks & Temples (low-texture). For each scene, we compute the pose FIM from per-pixel Jacobians and compare the CRB to (i) empirical errors from perturb-and-align trials (iNeRF-style [4]) and (ii) BA covariances when feature tracks are available.

From a known pose x, we render I, perturb x by random Δx, and realign by gradient descent to obtain $\hat{x}$. Across trials, RMSE in rotation/translation closely matches the CRB: high-texture scenes yield sub-degree and centimeter-level bounds, while low-texture scenes show multi-degree and decimeter-scale bounds (Table I). When keypoints exist, BA covariances (Hessian inverse) also align with our CRB within a few percent. In degenerate cases (e.g., planar white wall), near-zero eigenvalues appear in the FIM along wall-parallel translation and optical-axis rotation, so the pseudoinverse $I(x)^+$ produces large variances, consistent with BA and geometric intuition.

Scenario	Rot. error (deg)	Trans. error (cm)
High-texture (CRB)	0.4	1.3
High-texture (Empirical)	0.5	1.5
High-texture (BA Cov)	0.2	0.9
Low-texture (CRB)	5.1	21
Low-texture (Empirical)	5.5	23
Low-texture (BA Cov)	4.9	19

TABLE I
CRB VS. EMPIRICAL POSE ERROR AND BA COVARIANCE. TEXTURE-RICH VIEWS ARE TIGHTLY CONSTRAINED; LOW-TEXTURE VIEWS ARE ILL-CONDITIONED. THE CRB TRACKS BOTH EMPIRICAL AND BA UNCERTAINTIES.

We further evaluate two aspects of the bound: calibration and cooperative gains.

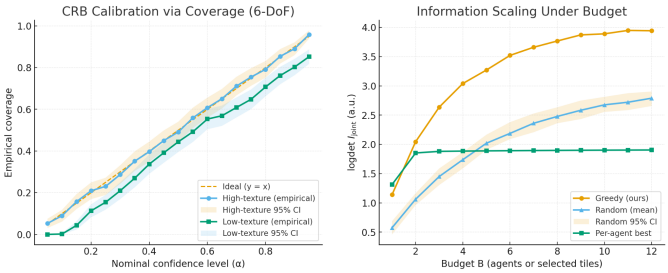

Fig. 3. CRB calibration and cooperative gains. **Left:** Coverage vs. nominal confidence shows calibration in high-texture scenes and under-coverage in low-texture ones. **Right:** log-det information grows submodularly with budget; greedy selection outperforms random and per-agent baselines.

These results suggest that the CRB can serve as both a diagnostic tool for view quality and a principled signal for multi-agent view planning.

V. Conclusion

We derived a render-aware Fisher information and Cramér–Rao bound on $\mathrm{SE}(3)$, showing how texture and geometry govern pose identifiability. The bound reduces to bundle adjustment in classical cases, matches empirical errors, and extends naturally to multi-agent settings via Fisher information fusion. Future work will address dynamic scenes and use the bound for view planning and adaptive rendering.

REFERENCES

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," in *European Conference on Computer Vision (ECCV)*, 2020.

[2] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics (SIGGRAPH)*, vol. 41, no. 4, pp. 102:1–102:15, 2022.

[3] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics (SIGGRAPH)*, vol. 42, no. 4, 2023.

[4] Y.-C. Lin, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T.-Y. Lin, "iNeRF: Inverting neural radiance fields for pose estimation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 1323–1330.

[5] Y. Chen, S. Huang, L. Zhao, and G. Dissanayake, "Cramér–rao bounds and optimal design metrics for pose-graph slam," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 627–641, 2021.

[6] Z. Zhang and D. Scaramuzza, "Beyond point clouds: Fisher information field for active visual localization," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2019, pp. 5984–5990.

[7] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "BARF: Bundle-adjusting neural radiance fields," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 5741–5751.

[8] L. Goli, C. Reading, S. Sellán, A. Jacobson, and A. Tagliasacchi, "Bayes' rays: Uncertainty quantification for neural radiance fields," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[9] W. Jiang, B. Lei, and K. Daniilidis, "Fisherrf: Active view selection and mapping with radiance fields using fisher information," 2024, extended from arXiv:2311.17874.

[10] Y. Tian, Y. Chang, F. H. Arias, C. Nieto-Granda, J. P. How, and L. Carlone, "Kimera-multi: Robust, distributed, dense metric-semantic slam for multi-robot systems," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2022–2038, 2022.

[11] P. Schmuck, T. Ziegler, M. Karrer, J. Perraudin, and M. Chli, "COVINS: Visual-inertial slam for centralized collaboration," in *IEEE International Symposium on Mixed and Augmented Reality (ISMAR) – Adjunct*, 2021.

[12] T. D. Barfoot, *State Estimation for Robotics*. Cambridge University Press, 2017.

[13] J. Solà, J. Deray, and D. Atchuthan, "A micro lie theory for state estimation in robotics," *arXiv preprint arXiv:1812.01537*, 2018.

[14] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

[15] S. Baker and I. Matthews, "Lucas–kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, no. 3, pp. 221–255, 2004.

[16] J. Engel, V. Koltun, and D. Cremers, "Direct sparse odometry," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 3, pp. 611–625, 2018.

[17] J. Engel, T. Schöps, and D. Cremers, "LSD-SLAM: Large-scale direct monocular slam," in *Proc. European Conference on Computer Vision (ECCV)*. Springer, 2014, pp. 834–849.

[18] A. Delaunoy and M. Pollefeys, "Photometric bundle adjustment for dense multi-view 3d modeling," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1486–1493.

[19] H. Alismail, B. Browning, and S. Lucey, "Photometric bundle adjustment for vision-based slam," *arXiv preprint arXiv:1608.02026*, 2016.

[20] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Transactions on Graphics (ToG)*, vol. 38, no. 4, pp. 1–14, 2019.

[21] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.