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

Arun Muthukkumar
Illinois Mathematics and Science Academy
amuthukkumar@imsa.edu

Abstract

Pose estimation is essential for many applications within computer vision and robotics. Yet few works provide rigorous uncertainty quantification for poses under dense or learned models, despite their uses. We derive a closed-form lower bound on the covariance of camera pose estimates by treating a differentiable renderer as a measurement function. We linearize image formation with respect to a small pose perturbation on the manifold and yield a render-aware *Cram'er–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.*

1. Introduction

Estimating the 6-DoF pose of a camera from images is foundational for vision and robotics. Neural rendering (NeRF [14], Instant-NGP [15], 3D Gaussian Splatting [11]) can offer a *dense*, differentiable photometric measurement model where each pixel depends on the pose. Works such as iNeRF [13] found that we may "invert" the renderer to localize cameras by photometric alignment. Despite this rapid progress on practical pose recovery, there is little theory quantifying *how accurately* pose can be estimated from these dense renderers, or how scene content (texture, depth variation, symmetries) fundamentally limits identifiability.

Classical geometric vision provides a natural lens to answer this question. The Cramér–Rao bound (CRB) lower-bounds the covariance of any unbiased estimator in terms of the Fisher information. In SfM/SLAM, the pose covariance of a bundle-adjustment (BA) solution relates to the inverse Hessian of the reprojection error. This is why CRBs have informed optimal design in pose-graph SLAM [5]; vision methods plan viewpoints by maximizing Fisher information [19]. However, these analyses typically assume *feature-based* measurements (e.g., 2D–3D correspondences). In contrast, Neural renderers give us a *dense photometric* observation governed by a complex, differentiable image formation pipeline.

We address this gap by deriving a *render-aware* CRB for pose on $\mathrm{SE}(3)$. We treat $I = R(\theta; x)$ as the observation model with fixed scene θ and pose $x \in \mathrm{SE}(3)$. Next, we can linearize image formation with respect to a tangent perturbation $\xi \in \mathfrak{se}(3)$, compute the per-pixel Jacobian $J = \partial R/\partial \xi$, and assemble a Fisher information matrix (FIM) $I(x) = J^\top \Sigma^{-1} J$. The bound $\mathrm{Cov}(\xi) \succeq I(x)^{-1}$ then quantifies the best-achievable pose accuracy.

Additionally, the eigenstructure of $I(x)$ exposes identifiability. High-texture and high-parallax regions yield large information. Low-texture or symmetric content induces degeneracies (near-zero eigenvalues). Crucially, the formulation reduces to classical BA covariance in the pinhole/feature limit, providing continuity with established theory.

While our derivation begins with a single camera, we adopt the convention of treating each camera as an *agent*. We show how this makes the formulation immediately extensible to multi-camera or cross-device settings, a useful downstream application. The method, in short, is to combine the Fisher information contributions from multiple agents, enabling efficient cooperative perception, fusion, and communication.

Contributions. (i) A general CRB for camera pose with differentiable renderers on $\mathrm{SE}(3)$; (ii) practical autodiff recipes for per-ray Jacobians across NeRF/3DGS; (iii) connections to BA/SLAM uncertainty and diagnostics for degeneracy; (iv) a compact protocol for empirical validation; (v) a multi-agent extension supporting cooperative perception and fusion

2. Related Works

Differentiable Rendering for Pose Estimation. Differentiable rendering can be used for camera pose estimation by enabling analysis-by-synthesis alignment. Neural rendering methods like NeRF provide dense and continuous scene representations that can produce photorealistic images given a camera pose. Following works (e.g. Instant-

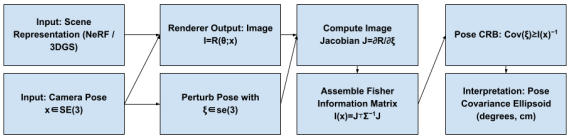

Figure 1. Pipeline: fixed scene θ and pose $x \to$ render I; autodiff gives $J = \partial R/\partial \xi$; FIM $J^\top \Sigma^{-1} J$; pose CRB $I(x)^{-1}$; interpret as ellipsoids in rotation/translation.

NGP and 3D Gaussian Splatting) now provide fast differentiable image formation. Because of these advances, gradients of the rendering process can be used for pose optimization. For example, iNeRF (Inverting NeRF) demonstrated that a pretrained radiance field can be directly "inverted" to recover 6-DoF camera pose via gradient-based photometric alignment. Such works show how differentiable neural renderers, whether used post hoc for localization [13] or in-loop during mapping [12], can reliably estimate camera pose by minimizing pixel-wise reprojection error without explicit correspondences.

Uncertainty Quantification in Neural Rendering. Quantifying uncertainty in neural scene representations is a recent goal. Bayes' Rays introduces a post-hoc Laplace approximation for NeRFs to estimate per-pixel confidence intervals [9]. FisherRF leverages Fisher information to guide view selection and quantify parameter uncertainty [10]. Current directions are focused towards scene / model uncertainty. Our work is aimed towards uncertainty for camera poses given a fixed scene. By deriving a render-aware CRB on pose covariance, we provide a complementary, pose-centric analysis that captures geometric identifiability alongside model confidence.

Information-Theoretic Analyses of Camera Pose. Information theory provides a lens to evaluate and improve pose estimation. Chen et al. derive CRBs for pose-graph SLAM and propose optimal design metrics to distribute sensing effort [5]. Zhang and Scaramuzza extend this idea by introducing the Fisher Information Field for active visual localization [19]. These approaches, however, assume feature-based measurements. In contrast, we treat a differentiable renderer as the observation model, yielding a dense photometric FIM for camera pose. By linearizing the full image formation process, our analysis bridges classical Fisher information methods with neural rendering, allowing us to quantify pose identifiability even without explicit correspondences.

Multi-Agent and Cooperative Perception. Multi-agent SLAM frameworks such as Kimera-Multi [18] and COVINS [16] demonstrate that sharing information across agents significantly improves localization accuracy and robustness. To build on top of this theme, we propose a principled method of fusing uncertainty by combining per-pixel Jacobians into a joint Fisher information matrix on a common reference frame. This yields a rigorous multi-agent pose CRB that ultimately aids cooperative view planning by communicating only the most informative observations.

Manifold and Statistical Estimation Foundations. Standard Lie-group state estimation and information theory are followed throughout our work. Barfoot's text for $\mathrm{SE}(3)$ estimation [4], Solà's micro-Lie treatment and Jacobian calculus [17], and Riemannian optimization background [1] justify local coordinates, reparameterization invariance, and reporting covariance in the tangent of $\mathrm{SE}(3)$.

3. Methodology

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

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

with fixed scene parameters θ and pixel-noise covariance $\Sigma \in \mathbb{R}^{M\times M}$ (not necessarily diagonal). Let $\xi \in \mathfrak{se}(3)$ be a minimal twist so that the perturbed pose is $\exp(\xi)\, x$. Linearizing the image formation at $\xi = 0$ gives

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

3.1. Core Derivation

Theorem 1 (Render-aware Fisher information on $\mathrm{SE}(3)$)**.** *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) yields the score $\nabla_\xi \log p = J^\top \Sigma^{-1}(I - R(\theta; x))$ with zero mean 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 (Invariance to smooth minimal pose parametrization)**.** *Let $\phi : \mathbb{R}^6 \to \mathbb{R}^6$ be a local diffeomorphism relating two minimal* $\mathrm{SE}(3)$ *coordinates ξ and $\zeta = \phi(\xi)$. Then the information transforms as $\mathcal{I}_\zeta = (D\phi)^{-\top}\mathcal{I}_\xi(D\phi)^{-1}$ and the CRB (4) is invariant (up to the coordinate change).*

Remark. The bound is thus well-defined on the manifold. We report rotation std in degrees and translation in scene units for interpretability.

Identifiability.

Lemma 1 (Local identifiability and degeneracy)**.** *If the columns of J span $\mathbb{R}^6$ on a set of nonzero measure pixels (equivalently,* $\mathrm{rank}(J) = 6$*), then $\mathcal{I}(x)$ is full-rank and all pose directions are locally identifiable. If J loses rank (e.g., constant-albedo planar wall, radial symmetry), $\mathcal{I}(x)$ becomes singular and the CRB diverges along the nullspace directions.*

Classical BA as a special case.

Corollary 1 (Bundle adjustment (BA) limit)**.** *If R reduces to pinhole projection of known* 3D *points $\{\mathbf{X}_k\}$ with per-point i.i.d. Gaussian noise $\sigma^2 I_2$, then stacking per-point reprojection Jacobians $J_k = \partial\pi(K[R|t]\mathbf{X}_k)/\partial\xi \in \mathbb{R}^{2\times 6}$ yields $J = \mathrm{blkrow}(J_k)$ and $\mathcal{I}(x) = J^\top(\sigma^{-2}I)J$, which equals the Gauss–Newton Hessian of reprojection BA; the CRB coincides with the BA covariance.*

3.2. Multi-Agent Extension

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

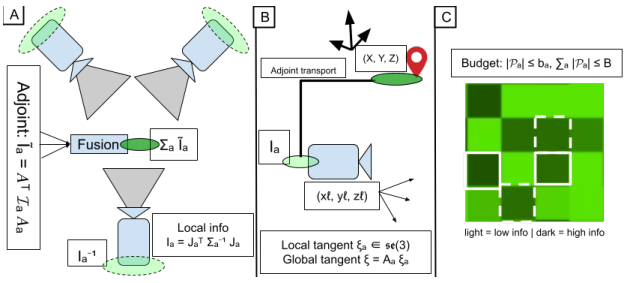

Figure 2. A) Multi-agent fusion of Fisher information. B) Adjoint transport from local to global tangent. C) Bandwidth-aware tile selection under budget constraints.

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 $(1\text{–}1/e)$ 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.

3.3. Computing J in practice (autodiff and VJPs)

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
$\mathcal{I}_{ij} \leftarrow \langle q_i, u_j\rangle_{\mathcal{P}}$ $(i, j = 1..6)$ // $\mathcal{I} = J^\top\Sigma^{-1}J$
return $\widehat{\mathcal{I}}(x)$ and

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

Directly forming J by per-pixel gradients is memory-intensive. Instead, we exploit vector-Jacobian products (VJPs): for any vector $v \in \mathbb{R}^M$, autodiff gives $J^\top v$ without materializing J. This suffices to assemble $\mathcal{I}(x) =$

$J^\top \Sigma^{-1} J$ by applying Σ^{-1} to columns of J implicitly. For diagonal (or block-diagonal) Σ, Σ^{-1} is cheap. Pixel subsampling and tiling further reduce cost.

Complexity and scalability. Let $|\mathcal{P}|$ be the number of sampled pixels. Forming $\mathcal{I}(x)$ requires 6 columns Je_j and their weighted inner products: $O(6\,|\mathcal{P}|)$ renderer VJPs plus cheap reductions for diagonal Σ. With $|\mathcal{P}| = sM$ (subsampling rate $s \in (0, 1]$), cost scales linearly in sM. Tiling amortizes memory; blockwise accumulation avoids storing J. The approach is practical for 512^2 images on modern GPUs.

3.4. Modeling assumptions and robustness

Noise. The derivation holds for general (possibly correlated) noise 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 mitigate this. **Bias.** The CRB applies to unbiased estimators; at high SNR, MLEs approach the bound. Biased extensions (e.g., van Trees) are possible but omitted here.

Interpretation and reporting. $\sqrt{\mathrm{diag}(\mathcal{I}(x)^{-1})}$ is reported as as 1σ pose bounds (rotation in degrees, translation in scene units). Eigenvalues of $\mathcal{I}(x)$ highlight ill-conditioning.

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

4. Preliminary Experiments

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

We validate the render-aware CRB on Instant-NGP [15] and 3D Gaussian Splatting [11] across LLFF (texture-rich) and Tanks & Temples (often low-texture). For each scene, we compute the pose FIM from per-pixel Jacobians and compare the resulting CRB to (i) empirical pose errors from small perturb-and-align trials (iNeRF-style [13]) and (ii) pose covariances from bundle adjustment (BA) when feature tracks are available.

Starting from a known pose x, we render I, perturb x by a small random Δx, and realign by gradient descent to obtain $\hat{x}$. Over many trials, RMSE in rotation/translation closely matches the CRB: high-texture scenes yield sub-degree and ∼centimeter-level bounds, while low-texture scenes exhibit multi-degree and decimeter-scale bounds (Table 1). When keypoints are available, BA covariances (from the Hessian inverse) also agree with our CRB in well-conditioned views, with differences only within a few percent. In degenerate cases such as a planar white wall, the FIM has near-zero eigenvalues along translation parallel to the wall and rotation about the optical axis, so the pseudoinverse $I(x)^+$ yields very large variances in those modes, 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 1. 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.

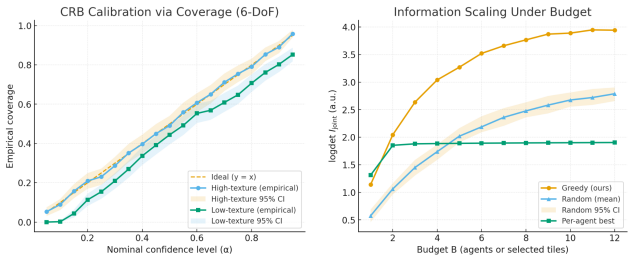

Figure 3. CRB calibration and cooperative gains. **Left:** Coverage vs. nominal confidence shows calibration in high-texture scenes and under-coverage in low-texture ones. **Right:** log-det information grows submodularly with budget; greedy selection outperforms random and per-agent baselines.

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

5. Conclusion

We derived a render-aware Fisher information and Cramér–Rao bound on $\mathrm{SE}(3)$, showing how scene texture and geometry govern pose identifiability. The bound reduces to BA in classical settings and closely tracks empirical errors, providing a principled target for pose accuracy. Future work will extend to dynamic scenes and use the bound for view planning and adaptive rendering.

References

[1] P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008. 2

[2] Hatem Alismail, Brett Browning, and Simon Lucey. Photometric bundle adjustment for vision-based slam. *arXiv preprint arXiv:1608.02026*, 2016.

[3] Simon Baker and Iain Matthews. Lucas–kanade 20 years on: A unifying framework. *International Journal of Computer Vision*, 56(3):221–255, 2004.

[4] Timothy D. Barfoot. *State Estimation for Robotics*. Cambridge University Press, 2017. 2

[5] Yongbo Chen, Shoudong Huang, Liang Zhao, and Gamini Dissanayake. Cramér–rao bounds and optimal design metrics for pose-graph slam. *IEEE Transactions on Robotics*, 37 (2):627–641, 2021. 1, 2

[6] Amaury Delaunoy and Marc Pollefeys. Photometric bundle adjustment for dense multi-view 3d modeling. In *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, pages 1486–1493, 2014.

[7] Jakob Engel, Thomas Schöps, and Daniel Cremers. LSD-SLAM: Large-scale direct monocular slam. In *Proc. European Conference on Computer Vision (ECCV)*, pages 834–849. Springer, 2014.

[8] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(3):611–625, 2018.

[9] Lily Goli, Cody Reading, Silvia Sellán, Alec Jacobson, and Andrea Tagliasacchi. Bayes' rays: Uncertainty quantification for neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[10] Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and mapping with radiance fields using fisher information. 2024. Extended from arXiv:2311.17874. 2

[11] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (SIGGRAPH)*, 42(4), 2023. 1, 4

[12] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-adjusting neural radiance fields. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5741–5751, 2021. 2

[13] Yen-Chen Lin, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. iNeRF: Inverting neural radiance fields for pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330, 2021. 1, 2, 4

[14] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 1

[15] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (SIGGRAPH)*, 41(4):102:1–102:15, 2022. 1, 4

[16] Patrik Schmuck, Thomas Ziegler, Marco Karrer, Jonathan Perraudin, and Margarita Chli. COVINS: Visual-inertial slam for centralized collaboration. In *IEEE International Symposium on Mixed and Augmented Reality (ISMAR) – Adjunct*, 2021. 2

[17] Joan Solà, Jeremie Deray, and Dinesh Atchuthan. A micro lie theory for state estimation in robotics. *arXiv preprint arXiv:1812.01537*, 2018. 2

[18] Yulun Tian, Yun Chang, Fernando Herrera Arias, Carlos Nieto-Granda, Jonathan P. How, and Luca Carlone. Kimera-multi: Robust, distributed, dense metric-semantic slam for multi-robot systems. *IEEE Transactions on Robotics*, 38(4): 2022–2038, 2022. 2

[19] Zichao Zhang and Davide Scaramuzza. Beyond point clouds: Fisher information field for active visual localization. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5984–5990, 2019. 1, 2