

Digging into Learned Camera Self-Calibration: What Matters in Challenging Motion Sequences

Takayuki Kanai¹, Igor Vasiljevic², Vitor Guizilini², Kota Shinjo¹, and Yuto Mori¹

Abstract—Camera self-calibration is a key enabler for geometrically accurate visual perception. Traditionally, sparse descriptor-based methods excel in static scenes by leveraging long-term photometric consistency, however they still often fail in dynamic or textureless environments. In contrast, approaches that combine dense bundle adjustment (BA) with pretrained priors have recently emerged to address this limitation. Yet, they still struggle with specific camera motions, such as steep motions or nearly pure translations. These challenging conditions hinder the computation of meaningful gradients, leading to degradation or even failure. In this work, we propose a novel camera self-calibration method, ZeCNOBA, that employs two complementary BA stages to tackle such limitations. The first dense BA provides coarse calibration and per-frame confidence maps, ensuring short-term consistency. A novel zero-shot geometry-guided initialization further robustifies this step even for steep motions. Subsequently, a sparse BA refines its calibration focusing on long-term consistency. As a result, ZeCNOBA achieves robust and accurate self-calibration on casual videos. Extensive experiments across zero-shot domains exhibit the efficacy of our approach and insights for future self-calibration system design.

Index Terms—Self-Calibration, Monocular Depth Estimation, Optical Flow, Bundle Adjustment

I. INTRODUCTION

Estimation of camera parameters from casual video plays a key role in *dense* and *metric* 3D perception [1]–[3] for autonomous systems, as well as a knowledge source for learning physically plausible robotic tasks [4]–[6]. To achieve this, recent methods combine dense bundle adjustment (BA) with large-scale pretrained predictors. Such pipelines, i.e., *learning-and-BA hybrid approaches*, have shown promising results across various fields, such as structure from motion (SfM) and Simultaneous Localization and Mapping (SLAM) [7]–[10].

However, state-of-the-art (SoTA) strategies still struggle in certain challenging conditions. Even when employing depth estimation guidance [3], [9], [11], improvements remain limited (Fig. 1). Interestingly, our preliminary analysis reveals a pivotal weakness: camera parameter estimation predominantly fails in sequences with little rotational motion. We interpret this deterioration as a consequence of camera ego-motion being nearly *degenerate* (low rotation, insufficient baseline, etc.) [12], which inherently impedes reliable self-calibration and, in turn, degrades trajectory estimation. In

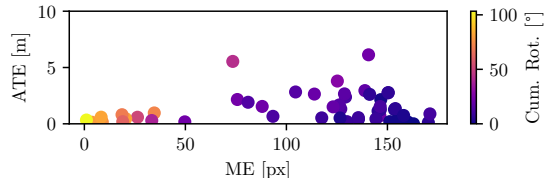


Fig. 1. Camera estimation errors (ME and ATE) labeled by cumulative egorotation on DDAD *val* split [14]. The estimates are obtained using dense-depth-guided Droidcalib [2], [7]. Accurate intrinsic estimation (low ME [15]) is achieved only for highly rotational video sequences, leading to accurate camera trajectory-estimation (low ATE).

other words, tackling the optimization difficulty related to ego-motion can be expected to have a significant impact on camera self-calibration for in-the-wild settings.

In this paper, we investigate the learning-and-BA hybrid approach with a particular focus on ego-motion-induced difficulties. We first analyze baseline methods based on Droidcalib [7], and reveal their key aspects that hinder gradient-based optimization. We show that the primary reason for this deterioration is Jacobian degeneracy. To directly tackle this issue, we introduce a complementary use of BAs that combines standard SfM [13] with dense, learned prior-guided BA [7]. In addition, we revisit the convergence issues of DROID-based BA initialization reported in [11] and propose geometry-guided zero-shot BA initialization to mitigate its unstable convergence. As a direct result of our contributions, a dramatic improvement can be achieved without relying on in-domain learning [11] or a hundred-times larger-scale backbone pretraining [9].

In short, we propose **Zero-shot Camera Self-Calibration via Neighbor-to-Omni Bundle Adjustment (ZeCNOBA)**, a **novel camera self-calibration** method, accompanied by:

- A **sensitivity analysis** of the learning-and-BA hybrid self-calibration, revealing the key difficulty: **Jacobian degeneracy**.
- **Two-step optimization**, titled **hierarchical, neighbor-to-omni keyframe pairing**, that efficiently alleviates the Jacobian issue even in challenging situations.
- A **novel geometry-guided initialization** that rectifies the failures of DROID-based BA initialization [11].

II. METHODOLOGY

A. Preliminary: DROID-based Self-Calibration

In the DROID-based approach [7], [16], dense depth and camera estimations are yielded through iterative BA guided by a neural predictor (ConvGRU). At ConvGRU’s step k (omitted hereafter unless otherwise needed), given a pair

¹T. Kanai, K. Shinjo, and Y. Mori are with Frontier Research Center, Toyota Motor Corporation (TMC), in Toyota, Aichi, Japan. {first.lastname, kota.shinjiyo}@mail.toyota.co.jp

²I. Vasiljevic and V. Guizilini are with Toyota Research Institute (TRI), in Los Altos, California, United States. {first.lastname}@tri.global

of images (I_i, I_j) and corresponding parameters such as inverted depth \mathbf{d}_i , poses $(\mathbf{G}_i, \mathbf{G}_j)$, and camera intrinsics θ , the residuals with respect to the target induced flow \mathbf{r}_{ij} and their confidence \mathbf{w}_{ij} are predicted as:

$$\mathbf{r}_{ij}(k+1), \mathbf{w}_{ij}(k+1) = \text{ConvGRU}(I_i, I_j, \mathbf{u}_{ij}(k) \mid k), \quad (1)$$

where the estimated induced flow \mathbf{u}_{ij} is obtained by projection/unprojection operators (π, π^{-1}) , a relative transformation $\mathbf{G}_{ij} := \mathbf{G}_j \circ \mathbf{G}_i^{-1}$, and grid field \mathbf{p}_i as:

$$\mathbf{u}_{ij} = \pi(\mathbf{P}'_i, \theta) \quad \text{where} \quad \mathbf{P}'_i = \mathbf{G}_{ij} \circ \pi^{-1}(\mathbf{p}_i, \mathbf{d}_i, \theta). \quad (2)$$

Subsequently, based on the neural predictions, the residual between the tentatively fixed targets (defined as $\mathbf{u}_{ij}^* = \mathbf{u}_{ij}^0 + \mathbf{r}_{ij}$) and the updating flows \mathbf{u}_{ij}^s (where $s \in [1, N_s]$ is the iteration number) is minimized using a weighted L2 cost as:

$$\mathcal{L}(\mathbf{G}_{ij}^s, \mathbf{d}^s, \theta^s) = \sum_{(i,j) \in \mathcal{E}} \|\mathbf{e}_{ij}^s\|_{\mathcal{W}_{ij}}^2, \quad \text{where} \quad \mathbf{e}_{ij}^s = \mathbf{u}_{ij}^* - \mathbf{u}_{ij}^s. \quad (3)$$

Here, \mathcal{E} is the keyframe subsets, and $\|\cdot\|_{\mathcal{W}_{ij}}$ ($\mathcal{W}_{ij} := \text{diag}[\mathbf{w}_{ij}]$) denotes the Mahalanobis distance. After the N_s updates, the neural prediction is advanced to step $k \rightarrow k+1$, and the cost (Eqn. 3) is minimized again. By leveraging standard SLAM techniques [17], this repeated optimization is solved efficiently, resulting in highly accurate estimation.

B. Key Difficulties Stemming from Egomotion Patterns

Intrinsics-related Jacobian Degeneracy. To see the reason for the failures of the DROID-based calibration (Fig. 1), we analyse the *observability* of the system with the *measurement* \mathbf{e}_{ij} (Eqn. 3) and the *state* $\xi = (\mathbf{G}, \mathbf{d}, \theta)$. In this condition, the observability is known to be analyzed with its Jacobian [18]. Here, the intrinsics-related Jacobian block $J_\theta(i, j) := \nabla_\theta \mathbf{e}_{ij}$, is obtained as¹:

$$J_\theta(i, j) = -\nabla_{\mathbf{P}'_i} \pi(\mathbf{P}'_i, \theta) \mathbf{R}_{ij} \nabla_\theta \pi^{-1}(\mathbf{p}_i, \mathbf{d}_i, \theta) - \nabla_\theta \pi(\mathbf{P}'_i, \theta), \quad (4)$$

where the rotation \mathbf{R}_{ij} satisfies $\mathbf{G}_{ij} := [\mathbf{R}_{ij} | \mathbf{t}_{ij}]$. The issue of this block can be easily anticipated under the following case: assuming a pinhole model $\theta = [f_x \ f_y \ c_x \ c_y]$ and an input video dominated by *driving-like* scenario², the J_θ drops the observability, i.e., $\nabla_{[f_x \ f_y]} \mathbf{e}_{ij} \approx \mathbf{0}$. As the DROID-based approach is originally designed to prioritize relatively short-term consistency [16], such an observability drop can easily occur. Indeed, our analysis based on the information matrix ($F_\theta(i, j) := J_\theta^\top \mathcal{W}_{ij} J_\theta$) [19] shows that the drop of observability is strongly associated with higher calibration error (Fig. 2). We thus argue that the Jacobian’s degeneracy is a key issue, which leads to self-calibration failure³.

Although the result may not be surprising, the following question still remains: *What level of degeneracy can we address?* To challenge this question, we propose a two-step optimization which addresses the degeneracy while keeping

¹Please refer to [7] and its supplementary for a detailed derivation.

² \mathbf{R}_{ij} is an almost identity matrix and its translation is $\mathbf{t}_{ij} \approx [0, 0, t_z]^T$.

³Note that the discussed degeneracy is also a matter of Fang *et al.* [20], where a gradient of the cost function provides essentially the same as $J_\theta(i, j)$. Furthermore, the cost is usually composed between the extremely short-term consistency, i.e., $j = i \pm 1$. Thus, SG-Init [11], which shares a learning principle with the way of Fang *et al.*, inevitably faces the same issue when naively extended to a self-calibrating variant.

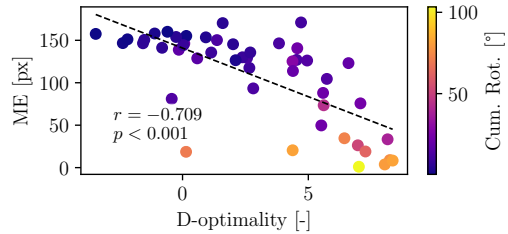


Fig. 2. **Correlation analysis between the intrinsics estimation error and Jacobian degeneracy for the baseline method.** The model and dataset follow our preliminary analysis (Fig. 1), with *seq.000196* removed as its camera is *exactly* stationary, and thus the degeneracy discussion is not meaningful. To represent the per-sequence scores, we used the state ξ_{end} after completed the calibration and computed a log determinant of the information matrix, i.e., $\log\text{-det}(\Sigma_{j \in \mathcal{J}^*} F_\theta(i^*, j) / \text{len}(\mathcal{J}^*))$, where i^* is the center index of keyframes, and \mathcal{J}^* is a subset of j which is topologically connected to i^* . This log-determinant is called *D-optimality* [19], and it exhibits greater information degeneration when it is low. By contrast, *informative* sequences (which also can be anticipated with a high Cum. Rot. score) tend to exhibit small calibration errors.

the strength of the DROID-based approach (Sec. II-C). Then, we show that our proposal successfully self-calibrates even for various *seemingly* degenerate motions.

Vulnerability from Inaccurate Initialization. Alongside the issue of camera intrinsics, the DROID-based optimization is also vulnerable to aggressive egomotions [11]. One key mitigator of these problems is providing both depth and pose initialization [11]. We demonstrate that estimating the pose via optical flow effectively satisfies the scale-consistency requirement while maintaining accuracy (Sec. II-D).

C. Hierarchical, Neighbor-to-Omni Keyframe Pairing

We introduce a novel strategy to alleviate the problem of Jacobian degeneracy while further improving accuracy (Fig. 3). In particular, we integrate a traditional SfM framework as a post-process for DROID-based BA (titled as *post-BA*, hereafter). Our key hypotheses are: (1) the neural confidence \mathbf{w}_{ij} remains *reliable* even when rotation/intrinsics estimations are erroneous, (2) filtering out low-reliability regions by \mathbf{w}_{ij} is *enough* to robustify a sparse descriptor-based SfM scheme, and (3) sparse descriptor matching on *full-resolution* images is helpful for extracting camera rotations, even in nearly *critical-type* [12] of ego-motions.

To achieve this, we first interpolate camera poses [16] and aggregate the confidence maps to cover the entire trajectory after the *backend* BA [7]. We then perform confidence-guided feature extraction, exhaustive matching, triangulation, and BA, relying on the off-the-shelf COLMAP [13].

D. Zero-shot Geometry-Guided Initialization

To provide a reliable initialization of the camera pose from the depth maps (D_i, D_j) and the optical flow \mathbf{f}_{ij} , we adopt a weighted procrustes alignment [21]. Here, the relative pose estimation \mathbf{G}_{ij}^0 is obtained by N_l 3D points, flow-guided grid sampling $\text{GS}(\cdot)$ [22], and weight map W_{ij} , as:

$$\mathbf{G}_{ij}^0 = \arg \min_{\mathbf{G}_{ij}} \sum_{l=1}^{N_l} W_{ij} \|\mathbf{G}_{ij} \circ {}^l \mathbf{P}_i - {}^l \text{GS}(\mathbf{P}_j \mid \mathbf{f}_{ij})\|^2, \quad (5)$$

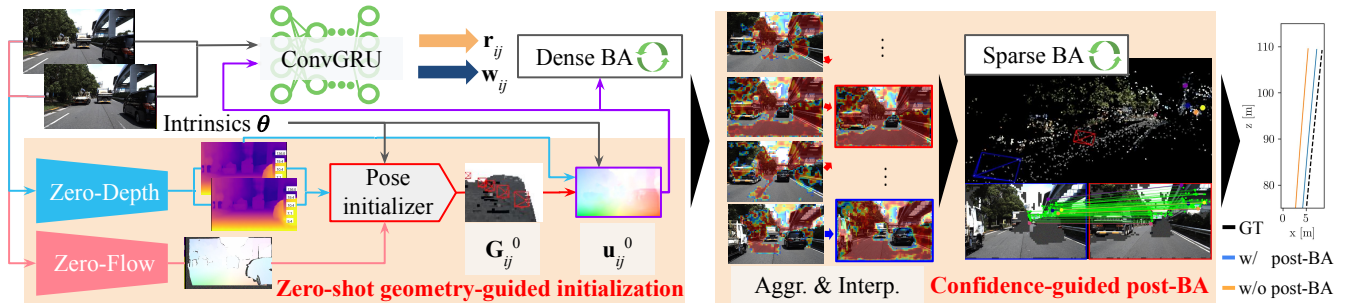


Fig. 3. **Diagram of our proposed self-calibration pipeline.** This pipeline is composed of two complementary BA stages: the first dense BA stage, which provides coarse calibration and confidence predictions, and the second sparse, descriptor-based BA stage (post-BA), which further refines the camera parameters on top of the first stage. To tackle real-world difficulties in self-calibration (Sec. II-B), we show that simple yet effective confidence aggregation and interpolation (Aggr. & Interp.; Sec. II-C), together with a zero-shot, geometry-guided initialization (Sec. II-D).

where the unprojection is defined as $\mathbf{P}_i = \pi^{-1}(\mathbf{p}_i, D_i^{-1}, \theta)$. In addition, we construct W_{ij} by multiplying the consistency mask [23] with the following to robustify the estimation:

- **A depth-driven weighting mask.** We first use $\mu_{cd} := N[1/\hat{D}_i^2]$ to emphasize the depth prediction of close points, as typically the scale of the depth estimation field is much larger than that of camera translation. Thus, this masking is expected to robustify the camera pose estimation against depth mispredictions in distant areas.
- **Photometrically invariant region mask.** To alleviate the impact of depth misprediction, we also introduce an *automask*-inspired mask, m_{ph} [24], to reject the depth prediction in photometrically less variant regions.

III. EXPERIMENTS

Datasets. We chose DDAD (Dense Depth for Autonomous Driving) [14] to benchmark the study. Notably, the evaluation on DDAD is well-suited to demonstrate self-calibration capabilities across various real-world challenges and degenerating motions. We chose 50 sequences for evaluation from *val* split. **Implementation Details.** We used a single NVIDIA RTX A6000 for all experiments except for VGGT [10], which consumes large memory footprint for inference and thus a single A100 (80GB) was employed. For the parameters of the DROID-based module, we followed previous work [11], [16]. We build our system on top of the off-the-shelf Droidcalib [7] *without fine-tuning*, which employed the relatively small-scale pretraining [25]. For our geometric priors, we adopted (1) UniMatch [26] for optical flow, and (2) ZeroDepth [2] for depth estimation. For fair comparison, we follow the typical protocol to initialize the camera intrinsics θ_0 unless otherwise described [7], [20], and also applied the same initialization to the other baselines [9], [13].

Evaluation. Table I summarizes the camera estimation results on DDAD *val* split. Ours (**ZcNOBA**) exhibits a magnitude level of improvement compared with the key elements on which ours relies [13]⁴, [27], [7]. Notably, ours outperforms SoTA methods on ATE, RTE, and ME metrics [9], [10], despite the small dataset scale for the SLAM backbone pretraining [25]. In contrast, only models

⁴Only the non-failure reconstructions (i.e., 17 out of 50) are counted.

TABLE I
UNCALIBRATED CAMERA ESTIMATION ON THE DDAD [14] VAL SPLIT.

| Method | ATE [m] | RTE [m] | RRE [°] | ME [px] |
|---------------------|--------------|--------------|--------------|--------------|
| COLMAP (17/50) [13] | 16.17 | 1.950 | 1.410 | 572.9 |
| FlowMap [27] | 4.885 | 0.250 | 0.105 | 165.3 |
| Droidcalib [7] | 4.016 | 0.348 | 0.119 | 1426 |
| VGGT+BA [10] | 1.467 | 0.043 | <u>0.043</u> | 58.80 |
| MeGaSaM [9] | 0.944 | <u>0.042</u> | 0.028 | 50.40 |
| ZcNOBA | 0.431 | 0.026 | 0.058 | 30.24 |

with large-scale pretraining surpassed ours in RRE. This implies that extending the pretraining scale is a practical way to improve further.

IV. DISCUSSION AND CONCLUSION

We introduce a novel zero-shot camera self-calibration method that hybridizes learning-based prior prediction and BA. First, we exposed key failure modes and their underlying causes, i.e., Jacobian degeneracy. Second, we proposed a novel strategy to address such an issue by (1) the complementary use of two types of BA, and (2) a zero-shot geometry-guided initialization, which addresses the BA initialization. In this strategy, (1) a sparse BA module accurately estimates camera parameters even on nearly *degenerated* ego-motions based on the first dense BA, and (2) the first of which is further robustified, particularly on aggressive ego-motion, by the use of estimated geometric priors. Our comprehensive experiments showed that zero-shot camera self-calibration can be significantly improved in various outdoor/indoor domains, without relying on large-scale SLAM backbone pretraining nor specific camera parameter initialization.

While our proposal addresses various real-world challenges with several *geometric anchors*, it still relies on the SfM principle. Therefore, camera ego-motion is still critical; footage captured by a stationary camera or in pure translation will cause issues. In addition, our method is vulnerable to mispredictions of geometric priors [16]. Integrating with a strong single-shot geometric predictor—which also alleviates the strong reliance on SfM-based estimation—while computationally efficient post-BA [28], [29] is one promising approach to further improve capability.

REFERENCES

- [1] T. Kanai, I. Vasiljevic, V. Guizilini, A. Gaidon, and R. Ambrus, "Robust self-supervised extrinsic self-calibration," in *IROS*, 2023, pp. 1932–1939.
- [2] V. Guizilini, I. Vasiljevic, D. Chen, R. Ambrus, and A. Gaidon, "Towards zero-shot scale-aware monocular depth estimation," in *ICCV*, 2023, pp. 9199–9209.
- [3] W. Yin, C. Zhang, H. Chen, Z. Cai, G. Yu, K. Wang, X. Chen, and C. Shen, "Metric3d: Towards zero-shot metric 3d prediction from a single image," *ICCV*, pp. 9043–9053, 2023.
- [4] C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, and S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots," in *RSS*, 2024.
- [5] A. Allshire, H. Choi, J. Zhang, D. McAllister, A. Zhang, C. M. Kim, T. Darrell, P. Abbeel, J. Malik, and A. Kanazawa, "Visual imitation enables contextual humanoid control," *arXiv*, 2025.
- [6] H. Chen, B. Sun, A. Zhang, M. Pollefeys, and S. Leutenegger, "Vidbot: Learning generalizable 3d actions from in-the-wild 2d human videos for zero-shot robotic manipulation," in *CVPR*, June 2025, pp. 27 661–27 672.
- [7] A. Hagemann, M. Knorr, and C. Stiller, "Deep geometry-aware camera self-calibration from video," in *ICCV*, 2023, pp. 3415–3425.
- [8] R. Murai, E. Dexheimer, and A. J. Davison, "Mast3r-slam: Real-time dense slam with 3d reconstruction priors," in *CVPR*, 2025, pp. 16 695–16 705.
- [9] Z. Li, R. Tucker, F. Cole, Q. Wang, L. Jin, V. Ye, A. Kanazawa, A. Holynski, and N. Snavely, "Megasam: Accurate, fast and robust structure and motion from casual dynamic videos," in *CVPR*, 2025, pp. 10 486–10 496.
- [10] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vvgt: Visual geometry grounded transformer," in *CVPR*, June 2025, pp. 5294–5306.
- [11] T. Kanai, I. Vasiljevic, V. Guizilini, and K. Shintani, "Self-supervised geometry-guided initialization for robust monocular visual odometry," in *IROS*, 2025, pp. 20 654–20 661.
- [12] P. Sturm, "Critical motion sequences for monocular self-calibration and uncalibrated euclidean reconstruction," in *CVPR*, 1997, pp. 1100–1105.
- [13] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *CVPR*, 2016, pp. 4104–4113.
- [14] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *CVPR*, 2020, pp. 2482–2491.
- [15] A. Hagemann, M. Knorr, H. Janssen, and C. Stiller, "Inferring bias and uncertainty in camera calibration," *IJCV*, vol. 130, no. 1, pp. 17–32, 2022.
- [16] Z. Teed and J. Deng, "DROID-SLAM: Deep Visual SLAM for Monocular, Stereo, and RGB-D Cameras," in *NeurIPS*, vol. 34, 2021, pp. 16 558–16 569.
- [17] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment - a modern synthesis," in *ICCVW*, 1999, p. 298–372.
- [18] K. Hausman, J. Preiss, G. S. Sukhatme, and S. Weiss, "Observability-aware trajectory optimization for self-calibration with application to uavs," *RA-L*, vol. 2, no. 3, pp. 1770–1777, 2017.
- [19] H. Carrillo, I. Reid, and J. A. Castellanos, "On the comparison of uncertainty criteria for active slam," in *ICRA*, 2012, pp. 2080–2087.
- [20] J. Fang, I. Vasiljevic, V. Guizilini, R. Ambrus, G. Shakhnarovich, A. Gaidon, and M. R. Walter, "Self-supervised camera self-calibration from video," in *ICRA*, 2022, pp. 8468–8475.
- [21] C. Choy, W. Dong, and V. Koltun, "Deep global registration," in *CVPR*, 2020.
- [22] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu, "Spatial transformer networks," in *NeurIPS*, vol. 28, 2015.
- [23] S. Meister, J. Hur, and S. Roth, "Unflow: Unsupervised learning of optical flow with a bidirectional census loss," *AAAI*, vol. 32, no. 1, 2018.
- [24] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *ICCV*, 2019, pp. 3827–3837.
- [25] W. Wang, D. Zhu, X. Wang, Y. Hu, Y. Qiu, C. Wang, Y. Hu, A. Kapoor, and S. Scherer, "Tartanair: A dataset to push the limits of visual slam," in *IROS*, 2020, pp. 4909–4916.
- [26] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, "Unifying flow, stereo and depth estimation," *IEEE TPAMI*, vol. 45, no. 11, pp. 13 941–13 958, 2023.
- [27] C. Smith, D. Charatan, A. Tewari, and V. Sitzmann, "Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent," in *3DV*, 2025.
- [28] L. Pan, D. Baráth, M. Pollefeys, and J. L. Schönberger, "Global structure-from-motion revisited," in *ECCV*, 2024, p. 58–77.
- [29] J. Li, H. Wang, M. Z. Irshad, I. Vasiljevic, M. R. Walter, V. C. Guizilini, and G. Shakhnarovich, "Fastmap: Revisiting dense and scalable structure from motion," in *3DV*, 2026.