SUPER – A Framework for Sensitivity-based Uncertainty-aware Performance and Risk Assessment in Visual Inertial Odometry

Johannes A. Gaus[1], Daniel Häufle[1], and Woo-Jeong Baek[2]

***Abstract*— Optimization-based visual odometry (VO), visual–inertial odometry (VIO) and SLAM systems usually report a single best pose estimate, but provide little information about when that estimate is becoming unreliable. This paper presents SUPER (Sensitivity-based Uncertainty-aware PErformance and Risk assessment), a lightweight framework for runtime warning signals in optimization-based visual estimation. SUPER reuses quantities already available in a Gauss–Newton or Levenberg–Marquardt backend: whitened and robust-weighted residuals, projection Jacobians, and Schur-complement landmark blocks. These terms yield per-feature projected covariances, geometric conditioning measures, and residual statistics, which are normalized over a short sliding window and fused into a frame-level risk score. The score is not treated as a calibrated physical failure probability; instead, it is an empirical proxy for the probability component of ISO-style risk and is calibrated through clean-sequence thresholds and hazard-curve analysis. We evaluate the method with a stereo VIO backend and ORB-SLAM3 on EuRoC MAV, KITTI, and TUM-VI under controlled visual corruptions. In VIO, SUPER predicts trajectory degradation 50 frames ahead with an AUC of 0.585, compared with 0.490 for residuals alone. In SLAM, a simple persistence-based caution policy reaches 89.1% recall at 8.4% false positive rate for timeout detection. The runtime overhead remains below 0.2%.**

I. INTRODUCTION

Visual odometry (VO), visual–inertial odometry (VIO), and SLAM systems achieve high accuracy, but usually return only a best pose estimate and provide little runtime information about the reliability of that estimate [1]–[5]. When blur, noise, low texture, weak parallax, or occlusion degrade the visual input, many systems detect the problem only after residuals, drift, or tracking failures have already accumulated [6]–[8]. This delay is problematic in safety-critical robotics, where an early warning signal could trigger conservative actions such as slowing down, relocalization, or increased sensing effort [9], [10]. We therefore propose SUPER, a lightweight framework for online risk assessment in VIO and SLAM.

We refer to standard ISO 12100, where the risk depends on the severity and occurrence probability of the respective event. Here, the severity is task- dependent such that SUPER estimates a backend-derived warning score that acts as an empirical measure for the occurrence probability. The calibration is obtained through clean sequence thresholds and hazard curve analyses in Sec. V-B. Technically, SUPER refers to quantities that are already computed in optimization-based backends given by the whitened residuals, projection Jacobians, and Schur-complement landmark blocks. These provide the per-feature uncertainty, geometric conditioning, and residual statistics. Particularly, these are normalized over a short temporal window and unified to one frame-level risk indicator. SUPER is backend-agnostic and does not require additional learned models, filter, or ground-truth signals at runtime. Across the data sets EuRoC MAV, KITTI, and TUM-VI, SUPER provides interpretable warning signals for visual degradation while adding less than 0.2% computational overhead.

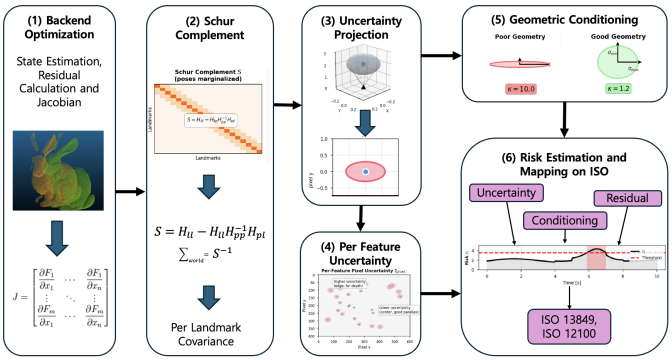

Fig. 1: **Schematic overview of SUPER. (1)** Backend optimization for state estimates. **(2)** Landmark covariance calculation. **(3)** Feature uncertainty computation. **(4)** Covariance propagation. **(5)** Geometric conditioning. **(6)** Risk Estimation.

II. RELATED WORK

Optimization-based VO/VIO/SLAM systems as OKVIS and ORB-SLAM3 rely on bundle adjustment or factor-graph optimization, but typically use Jacobians and normal-equation structure merely inside the estimator [3], [4]. Marginal covariances in backends like GTSAM and g2o are mostly treated as offline diagnostics due to their computational cost [11], [12]. Further methods estimate uncertainty through filtering or learned confidence prediction [7], [8], [13], but do not directly expose backend-derived sensitivity measures as interpretable warning signals. Prior work has linked the propagated uncertainty to hazard occurrence probability in safety-aware robotics [14]–[16]. SUPER brings this perspective into visual estimation by combining the uncertainty from the optimizer with residuals and geometric conditioning into one representative risk score.

[1] Hertie Institute for Clinical Brain Research & Center for Integrative Neuroscience, University of Tübingen, Germany.

[2] Artificial Intelligence Institute (AIIS), Seoul National University, Republic of Korea, and R&D AI Team, Division Advanced Vehicle Platform (AVP), Hyundai Motors Company, Republic of Korea.

III. The framework SUPER

A. First-Order Uncertainty Propagation

For a 3-D landmark $\mathbf{x} \in \mathbb{R}^3$ and projection $\pi : \mathbb{R}^3 \to \mathbb{R}^2$, small perturbations yield

$$\Sigma_{\text{pixel}} = J_\pi \, \Sigma_{\text{world}} \, J_\pi^\top, \tag{1}$$

where J_π is the projection Jacobian at the current linearization point. In incremental BA, residuals remain small, so Σ_{world} is approximated from the landmark marginal of the damped normal matrix $(J^\top W J + \lambda I)^{-1}$ at the current Gauss–Newton update, where W includes the backend's whitening and robust weights.

B. Schur Complement Extraction

The normal matrix from BA underlies a block structure

$$H = \begin{bmatrix} H_{pp} & H_{p\ell} \\ H_{\ell p} & H_{\ell\ell} \end{bmatrix}, \tag{2}$$

where H_{pp} contains pose-pose blocks and $H_{\ell\ell}$ landmark-landmark blocks. To marginalize out poses and obtain per-landmark covariances that account for pose-landmark coupling, we use the Schur complement

$$S = H_{\ell\ell} - H_{\ell p} H_{pp}^{-1} H_{p\ell}, \tag{3}$$

which modern solvers compute during the linear solve for efficiency [11], [12], [17], [18]. Per-landmark world covariances are approximated by $\Sigma_{\text{world},i} \approx S_{ii}^{-1}$, where S_{ii} is the 3×3 block for landmark i. Thus, $\Sigma_{\text{world},i}$ captures pose-landmark coupling.

C. Geometric Sensitivity and Conditioning

The sensitivity is obtained from the condition number of the projection Jacobian,

$$\kappa(J_\pi) = \frac{\sigma_{\max}}{\sigma_{\min}}. \tag{4}$$

Large $\kappa(J_\pi)$ indicates geometric degeneracy, typically caused by small baselines, grazing viewing angles, or distant points. In these cases, small world-space perturbations produce large image-plane changes and amplify uncertainty.

D. Frame-Level Aggregation, Risk Score and Normalization

Each frame t contains N_t tracked features. The frame-level proxies are

$$\bar{\sigma}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \sqrt{\operatorname{tr}(\Sigma_{\text{pixel},i})},$$

$$\bar{r}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \|\mathbf{r}_{\text{track},i}\|,$$

$$\kappa_t = \frac{1}{N_t} \sum_{i=1}^{N_t} \kappa_{J,i},$$

where $\mathbf{r}_{\text{track},i}$ is the reprojection residual for feature i. These aggregates reduce roughly 180–300 feature-level quantities to three frame-level scalars. To obtain a unified frame-level risk indicator, SUPER combines propagated uncertainty, residual magnitude, and geometric conditioning. Each component is z-normalized over a sliding window of 50–100 frames $\tilde{x}_t = \frac{x_t - \mu_x}{\sigma_x}$, where μ_x and σ_x denote the mean and standard deviation of the window. The normalized values are clamped to $[-3, 3]$ to suppress extreme transients. For the conditioning term, we first use $k_t = \log(1 + \kappa_t)$ and then normalize and clamp k_t in the same way. The resulting score is

$$r_t = \lambda_r \operatorname{clamp}(\tilde{r}_t) + \lambda_\sigma \operatorname{clamp}(\tilde{\sigma}_t) + \lambda_\kappa \operatorname{clamp}(\tilde{k}_t). \tag{5}$$

Unless stated otherwise, we use the default weights $(\lambda_\sigma, \lambda_r, \lambda_\kappa) = (1, 1, 1)$. This default treats the three normalized components equally, while alternative weightings can be used to tune the conservativeness of the score.

Mapping on ISO 12100. ISO 12100 formulates the risk as a combined measure of the severity and the occurrence probability P_i of the respective event. We therefore describe the risk of event i as $R_i \propto S_i P_i$. Arguing that the severity S_i is highly application-specific, we focus on P_i. SUPER estimates a warning score r_t by referring to the backend. This score can be used as an empirical estimate of P_i after calibration. Deriving the threshold and hazard-curve experiments in Sec. V-B provide this calibration by attributing r_t to observed degradation frequencies.

E. Backend Integration and Computational Cost

Our framework integrates seamlessly into any factor graph optimizer that exposes Jacobians and forms a structured Hessian (Ceres [17], g2o [12], GTSAM [11]). SUPER refers to quantities that can be directly accessed: Jacobians $\mathbf{J}_{\pi,i}$ from the linearization and the Schur complement $\mathbf{S}$ formed during the linear solve. The additional operations are limited to block extraction, small matrix products, per-feature 2×3 conditioning estimates, and frame-level aggregation. Across the evaluated datasets, this adds less than 0.2% of backend optimization time on a single CPU thread.

Comparison to alternatives. Unlike learned confidence modules or filter-based covariance propagation (D3VO [7], DeepVO [19], MSCKF [2]), SUPER does not add a separate estimator. It reuses linearization products already computed by the optimizer and adds only a lightweight monitoring layer.

IV. Experimental Setup

SUPER is evaluated on EuRoC MAV [20], TUM-VI [21], and KITTI Odometry [22].

A. Noise Injection

We corrupt camera images while keeping timestamps, calibration, and IMU data unchanged. The corruptions include Gaussian pixel noise, motion blur, JPEG compression, downsampling, partial occlusion, salt-and-pepper noise, and multiplicative intensity-dependent noise. Each corruption is applied at several severity levels, either over full sequences or short contiguous windows. For EuRoC MAV, corrupted stereo images are processed by the same VIO backend, so the experiment tests whether the visual risk signal detects

degradation while inertial measurements remain nominal. The multiplicative corruption is included only as a heteroscedastic stress test, not as the dominant camera-noise model; it exposes limits of the locally uniform first-order propagation assumption [23].

B. Pipeline, Metrics, and Baselines

Each corrupted sequence is processed with the stereo VIO pipeline of Choi *et al.* [13], comprising FAST detection, stereo matching, triangulation, and sliding-window BA with covariance extraction. The same corruption protocol is applied to ORB-SLAM3 [4]. We do not tune measurement covariances specifically for SUPER. Instead, we use the weighting, whitening, outlier rejection, and robust losses already used by the respective backend. The Hessian and Schur-complement blocks are therefore interpreted as effective local information matrices after the backend's residual weighting and robustification, not as perfectly calibrated sensor-noise covariances. This is one reason why SUPER reports a normalized warning score rather than an absolute probability.

Uncertainty is computed from pixel covariances $\Sigma_{\text{pixel}} = J_\pi \Sigma_{\text{world}} J_\pi^\top$, scalarized as $\sigma = \sqrt{\text{tr}(\Sigma_{\text{pixel}})}$ and averaged over features. Additional diagnostics include Jacobian conditioning κ_J, a triangulation-conditioning proxy κ_T, reprojection residuals, feature count, outlier ratio, and a pose-covariance proxy. The frame-level risk r_t fuses normalized uncertainty, conditioning, and residuals. The main comparison is against single-cue indicators that are commonly available in VO/VIO/SLAM systems: mean residual magnitude, feature count, Jacobian conditioning, and propagated uncertainty alone. The two evaluated backends are the Choi *et al.* VIO pipeline and ORB-SLAM3, each run under the same clean and corrupted image conditions. For the detection and policy experiments, the operational threshold r_{th} is set to the 95th percentile of r_t on clean frames for each dataset and then kept fixed across corruption conditions.

V. Results

A. Experimental Scope and Joint Trends

We evaluate SUPER on a stereo VIO pipeline [13] and ORB-SLAM3 [4] using EuRoC, KITTI, and TUM-VI under blur, noise, compression, downsampling, and occlusion corruptions. After excluding non-convergent runs, the dataset contains 314 VIO runs (1.0M frames) and 399 SLAM runs (0.75M frames); 55 severe SLAM runs reached the time budget and were retained for early-warning analysis. Across datasets, measurement corruptions increase risk moderately, whereas geometry-driven corruptions such as occlusion and the heteroscedastic stress test produce larger short-term excursions. The propagated uncertainty follows the same global trend as the risk, with strong mean correlations. Measurement corruptions increase the uncertainty, while geometry-driven corruptions produce sharper patterns especially in SLAM.

B. Detection, Prediction, and Policy Evaluation

Detection and safety metrics. We distinguish two trigger types. The margin criterion is active when the smoothed risk exceeds the clean-sequence threshold, $r_t - r_{\text{th}} > 0$. The trend criterion is active when the short-horizon derivative of the smoothed risk exceeds a fixed derivative threshold, indicating a rapid increase even before the margin threshold is crossed. For SLAM, the margin criterion detects 98.1% of timeout runs and dominates the trend criterion, indicating that persistent threshold violations are the main failure signal. For VIO, no timeouts occur, and trend-based activity mainly captures short-lived bursts rather than sustained failure. On clean KITTI and TUM-VI runs, the indicator is largely inactive with brief excursions on EuRoC.

Predictive power for near-future degradation. Instantaneous risk shows a limited correlation with cumulative trajectory error, because trajectory error accumulates over many frames. We therefore evaluate whether risk indicators predict impending degradation over a finite horizon. For each frame t, a degradation event is defined as a trajectory error exceeding $1.0\,\text{m}$ within the next $N = 50$ frames ($2.5\,\text{s}$ at $20\,\text{Hz}$). Table I reports AUC values across all 314 VIO runs. For AUC computation, the decision threshold is swept over each indicator; the fixed 95th-percentile threshold is used only for the operational policy experiments.

TABLE I: Predictive power for future degradation events. $N = 50$ frames ahead, error threshold $1.0\,\text{m}$, 314 VIO runs, 811K frames.

Indicator	AUC	vs. Chance
Mean risk (ours)	**0.585**	+17%
Mean sigma (ours)	**0.582**	+16%
Mean residual	0.490	$\approx 0\%$
Feature count	0.193	-61%
Jacobian condition	0.190	-62%

The uncertainty-based indicators achieve AUC ≈ 0.58, thereby outperforming classical heuristics. Mean residuals (AUC $= 0.49$) perform at chance level while feature count and Jacobian conditioning (AUC ≈ 0.19) are anti-correlated with future degradation. This effect occurs because some challenging scenes still contain many close-range features, yet degrade due to dynamic motion, blur, or occlusion. These are captured by the propagated uncertainty but cannot be captured solely by geometric proxies. The improvement of $\approx 20\%$ over residuals demonstrates that SUPER provides insights on impending failure that is not visible in the reprojection error.

Hazard curve analysis: For the hazard-curve analysis, frames are sorted by r_t and divided into ten equally sized risk deciles. For each decile, we compute the empirical frequency of the 50-frame degradation event defined above. The resulting curve increases from 2.3% in the lowest decile to about 45% in the highest decile, showing that the risk score is meaningfully ordered even though it is not a calibrated probability by construction.

a) Decision policy evaluation for SLAM: Additionally, we evaluate a straightforward caution policy on 399 SLAM runs.

Caution means that the system would request a conservative action, such as slowing down, increasing tracking effort, or attempting relocalization; it is not intended as a certified emergency-stop controller. The policy triggers when risk remains above the clean-sequence threshold for K consecutive frames. Table II reports detection performance across threshold values. A $K = 10$ frame ($0.5\,\mathrm{s}$) persistence requirement achieves 89% recall with 8.4% false positive rate (precision 62.8%). The false positives indicate unnecessary caution events, not necessarily system failures, and the persistence parameter can be increased or decreased depending on the desired safety-efficiency trade-off.

TABLE II: **SLAM stop/relocalization policy performance.** 399 SLAM runs, 55 timeouts. $\star$ = recommended configuration.

Threshold (frames)	Recall	FPR	Precision
5	69.1%	8.4%	56.7%
10 $\star$	**89.1%**	**8.4%**	**62.8%**
20	92.7%	25.6%	36.7%
30	92.7%	25.6%	36.7%
60	92.7%	29.9%	33.1%
120	96.4%	53.8%	22.3%

Short windows miss gradual failures, whereas long windows increase false positives during uncertainty spikes.

b) Ablation Studies: The risk combines propagated pixel uncertainty $\bar{\sigma}$, average reprojection residual $\bar{r}$, and Jacobian conditioning κ. To quantify the contribution of each component, we perform ablation studies. Since frame-level ground truth is available for VIO, the AUC values are averaged over all runs. For SLAM, frame-level ground truth is not available. Each run is treated as a single sample and timeout is used as the target. Table III summarizes the results.

TABLE III: AUC for VIO frame errors and SLAM timeouts.

Indicator	AUC (VIO)	AUC (SLAM)
Sigma only ($\bar{\sigma}$)	**0.563**	0.946
Residual only ($\bar{r}$)	0.416	0.566
Conditioning only (κ)	0.489	0.609
Risk margin	0.536	**0.989**

C. Limitations

SUPER is limited under extreme corruptions, where feature support collapses and the information matrix becomes nearly singular. Multiplicative speckle-like noise also exposes the limits of the homoscedastic first-order model. In addition, the current score focuses on visual residuals, visual geometry, and visual landmark uncertainty. IMU preintegration residuals, inertial bias uncertainty, and visual–inertial consistency checks provide complementary information and should be integrated in future versions. Extending the framework to heteroscedastic, inertial, and more strongly nonlinear regimes remains future work.

VI. Conclusion

We presented SUPER, a backend-agnostic framework for real-time risk assessment in VO, VIO, and SLAM. By reusing Schur-complement uncertainty, visual residuals, and conditioning already available in the optimizer, SUPER provides useful early warning signals with less than 0.2% overhead. It predicts VIO degradation 50 frames ahead with an AUC of 0.585 and supports a simple SLAM caution policy with 89.1% recall at 8.4% false positive rate. Future work will address heteroscedastic noise, explicit IMU residual statistics, and more severe long-horizon SLAM failures.

Acknowledgments. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Johannes A. Gaus.

References

[1] A. J. Davison *et al.*, "Monoslam: Real-time single camera slam," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 1052–1067, 2007.

[2] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint kalman filter for vision-aided inertial navigation," in *ICRA*, 2007, pp. 3565–3572.

[3] S. Leutenegger *et al.*, "Keyframe-based visual–inertial odometry using nonlinear optimization," *Int. J. Robotics Research*, vol. 34, no. 3, pp. 314–334, 2015.

[4] C. Campos *et al.*, "Orb-slam3: An accurate open-source library for visual, visual–inertial, and multi-map slam," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021.

[5] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," in *NeurIPS*, 2021.

[6] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in *ICLR*, 2019, arXiv:1903.12261.

[7] N. Yang *et al.*, "D3VO: Deep depth, deep pose and deep uncertainty for monocular visual odometry," in *CVPR*, 2020, pp. 1281–1292.

[8] M. Bloesch *et al.*, "Iterated extended kalman filter based visual-inertial odometry using direct photometric feedback," *Int. J. Robotics Research*, vol. 36, no. 10, pp. 1053–1072, 2017.

[9] C. Stachniss, J. J. Leonard, and S. Thrun, "Simultaneous localization and mapping," in *Springer Handbook of Robotics*, 2nd ed., B. Siciliano and O. Khatib, Eds. Springer, 2016, pp. 1153–1176.

[10] P. A. Lasota, T. Fong, and J. A. Shah, "A survey of methods for safe human-robot interaction," *Foundations and Trends in Robotics*, vol. 5, no. 4, pp. 261–349, 2017.

[11] F. Dellaert, "Factor graphs and gtsam: A hands-on introduction," Georgia Institute of Technology, Tech. Rep. GT-RIM-CP&R-2012-002, 2012.

[12] R. Kümmerle *et al.*, "g2o: A general framework for graph optimization," in *ICRA*, 2011, pp. 3607–3613.

[13] S. Choi *et al.*, "Statistical uncertainty learning for robust visual-inertial state estimation," *arXiv preprint arXiv:2510.01648*, 2025.

[14] W.-J. Baek and T. Kröger, "Safety evaluation of robot systems via uncertainty quantification," 02 2023.

[15] W.-J. Baek *et al.*, "Combining measurement uncertainties with the probabilistic robustness for safety evaluation of robot systems," in *IROS*, 2023, pp. 473–480.

[16] W.-J. Baek, "Uncertainty quantification and sensitivity-based optimization methods for robot systems," Ph.D. dissertation, Karlsruher Institut für Technologie (KIT), 2024.

[17] S. Agarwal, K. Mierle, and Others, "Ceres solver," *Journal of Open Source Software*, vol. 7, no. 74, p. 3735, 2022.

[18] B. Triggs *et al.*, "Bundle adjustment A modern synthesis," in *Vision Algorithms: Theory and Practice*, ser. Lecture Notes in Computer Science. Springer, 2000, vol. 1883, pp. 298–372.

[19] S. Wang *et al.*, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *ICRA*, 2017, pp. 2043–2050.

[20] M. Burri *et al.*, "The EuRoC micro aerial vehicle datasets," *The International Journal of Robotics Research*, vol. 35, no. 10, pp. 1157–1163, 2016.

[21] D. Schubert *et al.*, "The TUM-VI benchmark for evaluating visual-inertial odometry," *IEEE RAL*, vol. 3, no. 3, pp. 2637–2644, 2018.

[22] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the KITTI vision benchmark suite," in *CVPR*, 2012, pp. 3354–3361.

[23] A. Kendall and Y. Gal, "What uncertainties do we need in bayesian deep learning for computer vision?" in *NeurIPS*, 2017.