Proceedings of the 8th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2026.36664

Multi-Cue Dynamic Feature Scoring and Robust Filtering for RGB-D Visual Odometry

Yu Jian

School of Artificial Intelligence and Computer Science, North China University of Technology, Beijing, China
jianshanyue@outlook.com

Abstract. Feature-based RGB-D visual odometry and simultaneous localization and mapping (SLAM) commonly rely on an approximately static scene; moving objects can introduce observations inconsistent with camera motion and reduce pose-estimation reliability. We present DynamicScore-VO, a multi-cue dynamic-risk-aware front end that retains ORB extraction while assessing individual keypoints before ORB-SLAM2 Tracking. A dense semantic prior, epipolar inconsistency, background-relative motion residual, and temporally accumulated evidence are fused into a continuous dynamic-risk score rather than a calibrated probability. A frame-adaptive threshold based on the median and median absolute deviation (MAD) removes high-risk keypoints without redesigning the downstream pose-estimation pipeline. On five Freiburg 3 dynamic sequences from the TUM RGB-D benchmark, the five-run median absolute trajectory error (ATE) root-mean-square error (RMSE) is lower on four sequences. Relative to Original ORB-SLAM2, median ATE RMSE decreases by approximately 98.44 percent and 94.93 percent on the two highly dynamic walking sequences. The results show pronounced gains under strong dynamics but scene-dependent behavior under weaker dynamics, including a mild degradation on sitting_xyz, indicating that dynamic-feature filtering is not uniformly beneficial across motion regimes.

Keywords: RGB-D visual odometry, dynamic environments, dynamic feature filtering, multi-cue fusion, ORB-SLAM2

1. Introduction

Feature-based RGB-D visual odometry/SLAM relies on mostly static correspondences [1]; moving objects can bias pose estimation. Semantics identifies movable classes [2, 3] but not instantaneous motion, geometry exposes static-model violations [4] but is mismatch-sensitive, and optical flow mixes camera, structure, depth, and object motion [5]. DynamicScore-VO therefore fuses complementary evidence at point level.

DynamicScore-VO uses semantics as a soft prior and fuses semantic, epipolar, background-relative motion, and temporal risks into a continuous DynamicScore. A frame-adaptive median/MAD threshold rejects high-risk ORB keypoints before ORB-SLAM2 Tracking (Figure 1). The contributions are: (1) a point-level four-cue dynamic-risk representation; (2) integration

DOI: 10.54254/2755-2721/2026.36664

between ORB extraction and Tracking with a nonlearned adaptive rejection rule while preserving downstream matching and pose estimation; and (3) evaluation on five TUM RGB-D Freiburg 3 dynamic sequences [6] using repeated-run ATE, translational RPE, cue visualization, ablation, and threshold sensitivity, including both strong walking-scene gains and weaker-dynamic regressions.

2. Related work

ORB-SLAM2 [1] underlies this work. Detect-SLAM [7], DS-SLAM [3], and DynaSLAM [2] use detection/segmentation for dynamic filtering; DGS-SLAM and SG-SLAM combine semantic and geometric evidence [8, 9]; PVO [10] and Panoptic-SLAM [11] further integrate flow, depth, panoptic masks, association, or epipolar reasoning. Hard semantic filtering may discard temporarily static features and miss unknown movers.

ReFusion uses registration residuals/free space [4], feature-based systems use epipolar or motion consistency [2, 3, 8, 9, 11], and RDMO-SLAM [5], PVO [10], and DynaPix [12] exploit flow-related evidence. DynamicScore-VO differs in point-level admission: dense semantics, one-sided epipolar residuals, background-relative displacement, and grid temporal evidence form a continuous risk, followed by robust normalization and a frame-wise median/MAD decision.

3. Proposed method

DynamicScore-VO is inserted between ORB extraction and RGB-D Tracking. RGB supports semantic inference, grayscale supports ORB and GFTT-LK processing, and depth remains available to the original pipeline. Each ORB keypoint receives four risks whose fused score and adaptive threshold determine admission; downstream Tracking and pose estimation are unchanged. Loop closing and global bundle adjustment are disabled for evaluation.

GFTT points are propagated by pyramidal Lucas-Kanade flow with RANSAC two-view geometry. Valid-track geometry/motion responses are rasterized by radius-4 disks (pixel-wise maximum on overlap) and sampled at rounded ORB locations; uncovered pixels are zero. Rejected ORB entries and their associated arrays/spatial index are compacted consistently.

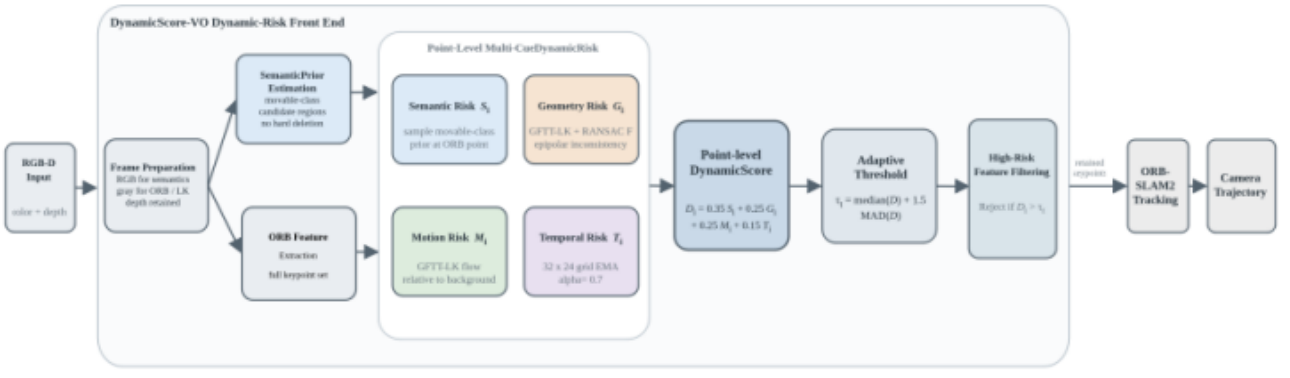

Figure 1. Overall architecture of the proposed DynamicScore-VO front end

3.1. Multi-cue dynamic risk estimation

Semantic risk uses TorchScript DeepLabV3 [13] with MobileNetV3 [14]. The pixel-wise maximum softmax probability over C_m={2,6,7,14,15,19} (bicycle, bus, car, motorbike, person, train) is

DOI: 10.54254/2755-2721/2026.36664

resized and sampled as S_i. The separate argmax movable-class mask does not gate this map, so S_i may remain nonzero for a nonmovable argmax; an unavailable score map contributes zero.

For valid GFTT-LK correspondence j, homogeneous image points and the RANSAC fundamental matrix F_t define

$$x_j = [u_j, v_j, 1]^T,\ x\prime_j = [u\prime_j, v\prime_j, 1]^T, [a_j,\ b_j, c_j]^T = F_t x_j \tag{1}$$

The one-sided residual, robust normalization, and clipped geometry response are

$$r_j^g = \frac{|a_j u\prime_j + b_j v\prime_j + c_j|}{\sqrt{a_j^2 + b_j^2}} \tag{2}$$

$$\sigma_g = median(\Omega_g) + 3 \times 1.4826 \times MAD(\Omega_g) \tag{3}$$

$$\widetilde{G}_j = clip(\frac{r_j^g}{\sigma_g}, 0, 1) \tag{4}$$

This is not the Sampson distance. Degenerate norms (<=10^-9) give zero residual; a degenerate scale falls back to the maximum residual and then one. The rasterized response is sampled as G_i; high values may reflect independent motion or tracking/model error.

For motion risk, d_j=x'_j-x_j. B uses tracks with semantic risk <=0.2, falling back to all valid tracks when fewer than eight remain. Background displacement, residual, scale, and response are

$$d_{bg} = \begin{bmatrix} median_{j \in B}(d_{j,u}) \\ median_{j \in B}(d_{j,v}) \end{bmatrix} \tag{5}$$

$$e_j^m = \|d_j - d_{bg}\|_2 \tag{6}$$

$$\sigma_m = median(\Omega_m) + 3 \times 1.4826 \times MAD(\Omega_m) \tag{7}$$

$$\widetilde{M}_j = clip\left(\frac{e_j^m}{\sigma_m}, 0, 1\right) \tag{8}$$

The same fallback/rasterization yields M_i. Because this image-plane cue does not explicitly compensate rotation, depth-dependent parallax, or articulated 3-D motion, it is fused rather than used alone.

Temporal risk averages available cue evidence in a 32x24 grid (20x20-pixel cells at 640x480):

$$C_t(c) = \frac{1}{N_t(c)} \sum_{q \in A_t(c)} q_t(c) \tag{9}$$

$$T_t(c) = \alpha T_{t-1}(c) + (1 - \alpha) C_t(c), \quad \alpha = 0.7 \tag{10}$$

The first valid frame initializes the grid with T_i=0; afterward alpha=0.7 and keypoints sample updated cells, so T_i is accumulated but correlated with current cues.

At each ORB keypoint, the bounded risk vector and complete Semantic-Geometry-Motion-Temporal (SGMT) score are

$$s_i = [S_i, G_i, M_i, T_i]^T \in [0, 1]^4 \tag{11}$$

DOI: 10.54254/2755-2721/2026.36664

$$D_i = 0.35S_i + 0.25G_i + 0.25M_i + 0.15T_i \tag{12}$$

The 0.35/0.25/0.25/0.15 weights are fixed empirical coefficients. Ablations zero inactive cues without renormalization: S, SG, SM, SGM, and SGMT use (S,G,M,T) tuples (0.35,0,0,0), (0.35,0.25,0,0), (0.35,0,0.25,0), (0.35,0.25,0.25,0), and (0.35,0.25,0.25,0.15).

For the post-extraction keypoint set K_t, the frame score set, robust center/spread, adaptive threshold, and strict decision are

$$\Omega_t = \{D_i \mid i \in K_t\} \tag{13}$$

$$m_t = median(\Omega_t), MAD_t = median_{D_i \in \Omega_t} |D_i - m_t| \tag{14}$$

$$\tau_t = clip(m_t + 1.5MAD_t, 0, 1) \tag{15}$$

$$decision(i) = \begin{cases} reject, & D_i > \tau_t \\ retain, & D_i \le \tau_t \end{cases} \tag{16}$$

The coefficient 1.5 is fixed and clipping enforces [0,1]; equality is retained. If fewer than 300 keypoints would remain, the unfiltered post-extraction set is preserved, while missing cue maps contribute zero.

4. Experiments

TUM RGB-D experiments use fr3/walking_xyz, walking_halfsphere, sitting_xyz, sitting_halfsphere, and sitting_static [6]. Original ORB-SLAM2, DS-SLAM, Detect-SLAM, DynaSLAM, and DynamicScore-VO share a visual-odometry-only protocol with loop closing/global bundle adjustment disabled; dynamic baselines were locally reproduced in Docker.

ATE RMSE after fixed-scale SE(3) alignment is primary; translational RPE RMSE uses a 1 s interval on three sequences. Each method completed 5/5 runs per sequence. The platform was Ubuntu 22.04, i7-11800H (2.30 GHz), 8 GB RAM, no GPU/CUDA, OpenCV 3.2, PyTorch 2.4.0, and 640x480 input.

Table 1. Five-run median ATE RMSE (m) on selected TUM RGB-D dynamic sequences

Method	walking_xyz	walking_halfsphere	sitting_xyz	sitting_halfsphere	sitting_static
ORB-SLAM2 [1]	1.0225	0.7055	0.0115	0.0299	0.0084
DS-SLAM [3]	0.0948	0.1284	0.0176	0.0248	0.0102
Detect-SLAM [7]	0.0613	0.0856	0.0153	0.0215	0.0093
DynaSLAM [2]	0.0286	0.0479	0.0131	0.0196	0.0081
DynamicScore-VO	0.0159	0.0358	0.0124	0.0167	0.0070

DynamicScore-VO is best on four of five sequences and among reproduced dynamic baselines on all five. Versus original ORB-SLAM2, ATE drops 98.44% on walking_xyz (1.0225 to 0.0159 m) and 94.93% on walking_halfsphere (0.7055 to 0.0358 m); sitting_halfsphere/sitting_static also improve, while sitting_xyz mildly degrades from 0.0115 to 0.0124 m, showing weak-dynamic over-filtering.

DOI: 10.54254/2755-2721/2026.36664

Table 2. Five-run median translational RPE RMSE (m) on representative sequences

Sequence	ORB-SLAM2	DynamicScore-VO	Reduction
walking_xyz	0.028296	0.013250	53.17%
sitting_static	0.016885	0.006340	62.45%
walking_halfsphere	0.030548	0.017624	42.31%
Mean RPE	0.025243	0.012405	50.86%

RPE reductions are 53.17%, 62.45%, and 42.31%, with the three-sequence mean reduced by 50.86%.

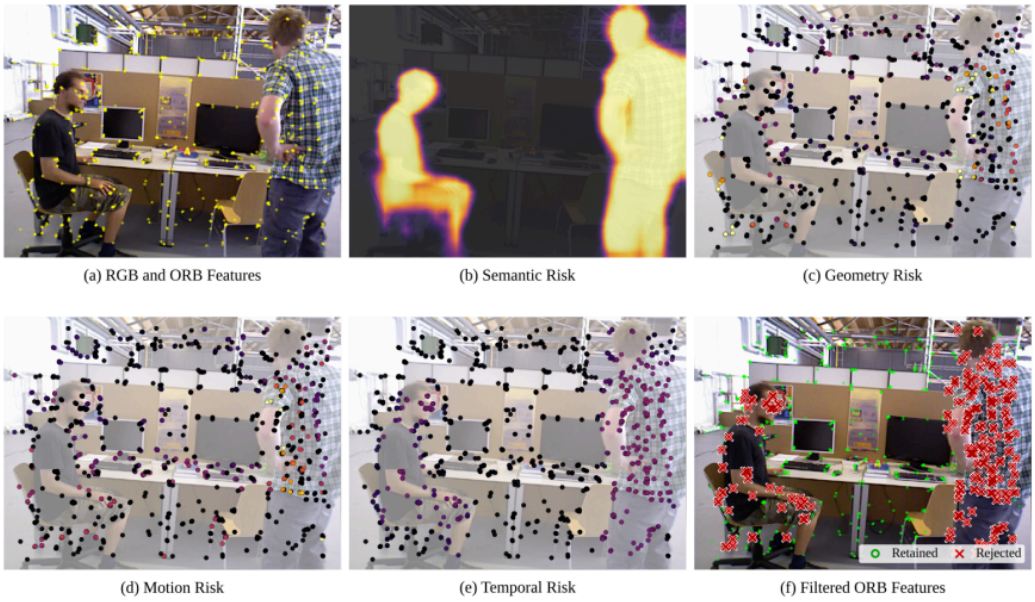

Figure 2. Multi-cue dynamic-risk responses and resulting feature filtering on the same representative frame

At timestamp 1341846313.686156, Figure 2 shows the four risks and filtering: 365/1004 ORB features (36.35%) are rejected and 639 retained, mainly removing points on the two people while preserving background support.

Table 3. Cue-ablation ATE RMSE (m) on representative TUM RGB-D sequences

Configuration	walking_xyz	walking_halfsphere	sitting_xyz
S	0.0169	0.0372	0.0152
SG	0.0168	0.0319	0.0129
SM	0.0164	0.0312	0.0134
SGM	0.0196	0.0334	0.0121
SGMT	0.0159	0.0358	0.0124

The ablation is non-monotonic across scenes, including scene-dependent temporal utility.

DOI: 10.54254/2755-2721/2026.36664

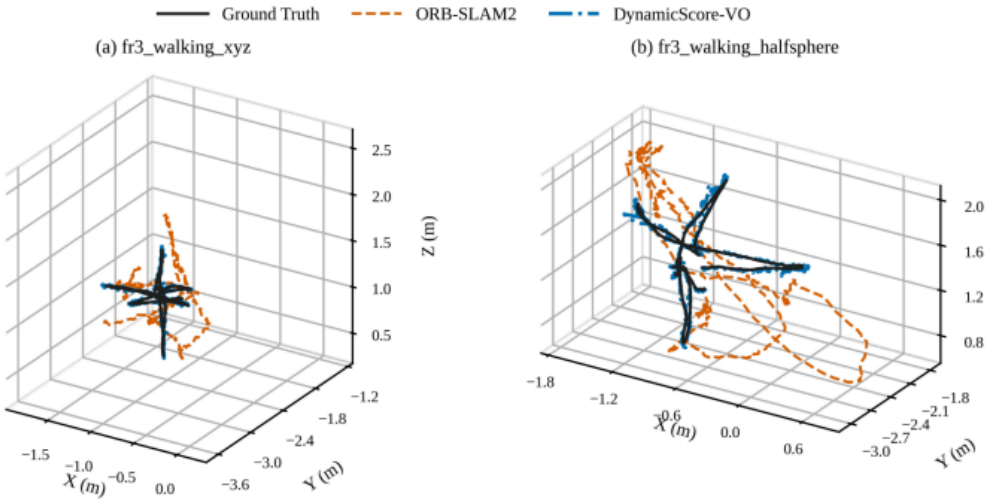

Figure 3. Camera trajectories on representative highly dynamic TUM RGB-D sequences; DynamicScore-VO remains closer to ground truth overall than original ORB-SLAM2

Table 4. Sensitivity of fixed and adaptive DynamicScore thresholds in ATE RMSE (m)

Threshold Strategy	walking_xyz	walking_halfsphere	sitting_xyz
Low (τ=0.188)	0.0174	0.0394	0.0140
Medium (τ=0.258)	0.0161	0.0363	0.0135
High (τ=0.324)	0.0167	0.0373	0.0121
Adaptive	0.0159	0.0358	0.0124

Fixed thresholds 0.188/0.258/0.324 are pooled Q25/Q50/Q75 adaptive-threshold values from the five-sequence evaluation set and are post-hoc sensitivity points, not held-out tuning. Adaptive is best on both walking sequences, while Fixed High is slightly better on sitting_xyz; the rule therefore favors a unified frame-dependent operating point over sequence-wise optimality.

5. Conclusion

DynamicScore-VO fuses semantic, geometric, motion, and temporal evidence into a point-level risk and applies a median/MAD adaptive threshold before ORB-SLAM2 Tracking. Median ATE improves on four of five sequences, most strongly on the two walking sequences, with a mild regression on sitting_xyz; RPE, ablation, and sensitivity results likewise show scene-dependent cue and threshold behavior. Future work will investigate more adaptive cue weighting and threshold generalization to reduce over-filtering under weak dynamics.

References

[1] Mur-Artal, R., & Tardos, J. D. (2017). ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics, 33*(5), 1255–1262. https://doi.org/10.1109/TRO.2017.2705103

[2] Bescos, B., Facil, J. M., Civera, J., & Neira, J. (2018). DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes. *IEEE Robotics and Automation Letters, 3*(4), 4076–4083. https://doi.org/10.1109/LRA.2018.2860039

[3] Yu, C., et al. (2018). DS-SLAM: A semantic visual SLAM towards dynamic environments. In *Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 1168–1174).

DOI: 10.54254/2755-2721/2026.36664

https://doi.org/10.1109/IROS.2018.8593691

[4] Palazzolo, E., Behley, J., Lottes, P., Giguere, P., & Stachniss, C. (2019). ReFusion: 3D reconstruction in dynamic environments for RGB-D cameras exploiting residuals. In *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 7855–7862). https://doi.org/10.1109/IROS40897.2019.8967590

[5] Liu, Y., & Miura, J. (2021). RDMO-SLAM: Real-time visual SLAM for dynamic environments using semantic label prediction with optical flow. *IEEE Access, 9*, 106981–106997. https://doi.org/10.1109/ACCESS.2021.3100426

[6] Sturm, J., Engelhard, N., Endres, F., Burgard, W., & Cremers, D. (2012). A benchmark for the evaluation of RGB-D SLAM systems. In *Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 573–580). doi: 10.1109/IROS.2012.6385773.

[7] Zhong, F., Wang, S., Zhang, Z., Chen, C., & Wang, Y. (2018). Detect-SLAM: Making object detection and SLAM mutually beneficial. In *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (pp. 1001–1010). doi: 10.1109/WACV.2018.00115.

[8] Yan, L., Hu, X., Zhao, L., Chen, Y., Wei, P., & Xie, H. (2022). DGS-SLAM: A fast and robust RGBD SLAM in dynamic environments combined by geometric and semantic information. *Remote Sensing, 14*(3), Art. no. 795. doi: 10.3390/rs14030795.

[9] Cheng, S., Sun, C., Zhang, S., & Zhang, D. (2023). SG-SLAM: A real-time RGB-D visual SLAM toward dynamic scenes with semantic and geometric information. *IEEE Transactions on Instrumentation and Measurement, 72*, Art. no. 7501012. doi: 10.1109/TIM.2022.3228006.

[10] Ye, W., et al. (2023). PVO: Panoptic visual odometry. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9579–9589). doi: 10.1109/CVPR52729.2023.00924.

[11] Abati, G. F., Soares, J. C. V., Medeiros, V. S., Meggiolaro, M. A., & Semini, C. (2024). Panoptic-SLAM: Visual SLAM in dynamic environments using panoptic segmentation. In *Proceedings of the 2024 21st International Conference on Ubiquitous Robots (UR)* (pp. 1–8). doi: 10.1109/UR61395.2024.10597506.

[12] Xu, C., Bonetto, E., & Ahmad, A. (2025). DynaPix SLAM: A pixel-based dynamic visual SLAM approach. In *Pattern Recognition: 46th DAGM German Conference, DAGM GCPR 2024, Proceedings, Part II, Lecture Notes in Computer Science, 15298* (pp. 168–184). doi: 10.1007/978-3-031-85187-2_11.

[13] Chen, L.-C., Papandreou, G., Schroff, F., & Adam, H. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587. https://arxiv.org/abs/1706.05587.

[14] Howard, A., et al. (2019). Searching for MobileNetV3. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 1314–1324). doi: 10.1109/ICCV.2019.00140.