Babylonian Journal of Artificial Intelligence
Vol.2026, **pp.** 98-122
DOI: https://doi.org/10.58496/BJAI/2026/014; ISSN: 3006-5437
https://mesopotamian.press/journals/index.php/BJAI/index

Research Article

Observability-Aware Rotation-Target Filtering for Self-Supervised Depth-Only Odometry

Farah Ibelaiden[1,*], Seif Eddine Bouziane[1], and Ouarda Lounis[1]

[1]*Intelligent Systems Engineering Department (ISE), National School of Artificial Intelligence (ENSIA), Sidi Abdellah Campus, Algiers, Algeria*

A R T I C L E I N F O

Article History
Received 08 Sep 2026
Revised 11 Nov 2026
Accepted 12 Nov 2026
Published 17 Nov 2026

Keywords
Visual Odometry
Self-Supervised Learning
Deep Learning
Depth Sensor
Observability Analysis

ABSTRACT

Plane-based pseudo-labels can provide complete pose targets even when scene geometry supports only some motion components. We introduce an observability-aware training approach that gates rotation targets using matched-plane normal separation and projects translation residuals onto separately supported directions. The student is a depth-only model at inference and receives no geometric support indicators. We compare it with unfiltered supervision using the same 620,169-parameter student, five paired random seeds, equal training budgets, and a common data protocol. The evaluation covers nine sequences from TUM RGB-D, VCU-RVI and ETH3D, including four ETH3D zero-shot transfer sequences. Across all nine sequences, the filtered model has lower mean rotation root-mean-square error (RMSE) but higher mean translation RMSE than the unfiltered model. Thus the benefit is component-specific rather than a uniform pose improvement. Finite contiguous-trajectory tests likewise show no uniform trajectory benefit and reveal unfavorable translation accumulation. TUM and ETH3D provide verified-camera physical evidence; VCU results remain conditional because the supplied reference-frame semantics are not independently verified. The method is therefore best understood as targeted supervision for rotation under geometric degeneracy, not as a guarantee of globally improved odometry.

1. INTRODUCTION

Reliable relative-pose estimation is essential for indoor navigation, augmented reality and mobile manipulation. Visual–inertial systems combine complementary measurements to estimate motion [1–3]. Filter-based ROVIO, optimization-based DM-VIO and dynamic-scene RD-VIO illustrate different approaches to this sensing problem [4–6]. RGB-D methods additionally exploit depth through point and plane constraints [7–9]. Depth reduces dependence on image texture, but does not eliminate geometric degeneracy. For a teacher based on plane normals and offsets, a single plane constrains translation along its normal and leaves rotation about that normal unresolved. Two non-parallel normals can constrain rotation locally while leaving one translation direction unsupported. Consequently, a solver can return a finite pose even when some components are not identifiable from its measurements [10, 11].

Self-supervised odometry can replace pose annotations with targets produced by geometric registration [12, 13]. However, these targets need not be reliable in every pose component. Degeneracy creates structured errors associated with the scene geometry, rather than independent zero-mean perturbations. An unfiltered student is asked to reproduce a complete target even where the teacher has insufficient constraints. Increasing network capacity alone does not remove this supervision problem: convolutional, attention-based and hybrid models can all learn unsupported target components.

* *Corresponding author. Email: farah.ibelaiden@ensia.edu.dz*

This work investigates observability-aware (OA) pseudo-label filtering for depth-only learning. A plane-alignment teacher produces relative poses, a weighted normal Gram matrix and support indicators. During training, translation residuals are projected onto supported eigendirections, while the rotation target is used only when matched normals satisfy an angular-separation rule. Translation support is determined by a normalized-spectrum threshold of 0.08. Rotation uses a separate criterion: at least one pair of matched normals must have acute separation greater than 8°. The student receives only two depth images and their validity masks at inference; neither geometric support indicators nor inertial measurements are inference inputs.

The study makes four contributions. First, it defines the geometry-filtered learning objective and derives local rotational and joint-pose ranks for the teacher's normal/offset residual model. Second, it makes rotation-target filtering explicit through a matched-normal separation gate, with translation supervision treated by a separate support transform. Third, it compares the full OA-mask with Naive training of the same 620,169-parameter CNN with attention across five paired seeds and equal optimizer budgets, including multiplicity-aware inference and exact paired sensitivity checks. Fourth, it reports completed component, mask, backbone and threshold diagnostics, zero-shot transfer, finite contiguous-trajectory accumulation and computational cost. The five-seed cohort supplies the primary inference; current seed-zero comparator rows are secondary descriptive evaluations.

The central question is whether geometry-conditioned supervision, particularly rotation-target filtering, reduces the propagation of unsupported teacher rotations. The full OA-mask remains the evaluated primary method. Five-seed component controls show that applying the rotation gate alone reduces mean rotation RMSE relative to Naive on all nine evaluation sequences, including all six TUM and ETH3D sequences, whose camera reference frames are verified. In contrast, applying the translation projector alone reduces mean rotation RMSE relative to Naive on only three of the nine sequences. Full OA has lower mean rotation RMSE than the rotation-gate-only control on seven of the nine sequences; the rotation-gate-only control has lower mean rotation RMSE on the remaining two. These exploratory contrasts indicate a strong filtering contribution but do not establish a unique causal mechanism or uniform superiority of the rotation-only control. The five-seed comparison establishes an observed rotation–translation trade-off, not universal pose accuracy or globally stable odometry.

2. RELATED WORK

2.1 Visual, RGB-D and learned odometry

Classical visual–inertial estimators range from multi-state filtering [14] to manifold preintegration and tightly coupled optimization [15, 16]. ORB-SLAM3 and OpenVINS provide established examples [1, 2]. RGB-D systems incorporate point and plane constraints, sometimes together with inertial measurements [7, 8, 17]. Dense methods such as KinectFusion and dense RGB-D SLAM use surface alignment [18, 19]; fast plane extraction and planar SLAM make the underlying structural constraints explicit [20–23]. Their effectiveness depends on the available scene geometry and the validity of data association.

Learned visual odometry includes direct motion prediction, exemplified by DeepVO, and learned geometric optimization, exemplified by DROID-SLAM and DPVO [24–26]. TartanVO studies generalization across visual domains [27], while Zhan et al. investigate what geometric quantities and objectives should be learned [28]. Photometric self-supervision provides an alternative to pose labels [29–31], with related work using three-dimensional consistency and dynamic-scene constraints [32, 33]. DVL-SLAM instead combines sparse depth with direct visual estimation [34].

Lightweight CNN–transformer models combine local processing with global interactions [35, 36]. Vision Transformer provides the patch-based architectural basis for the six-block ViT ablation evaluated here [37]; ViTVO, transformer-based monocular odometry and SWformer-VO apply attention to motion estimation [38–40]. These studies motivate the backbone comparison, but are not sensor-matched depth-only benchmark competitors. The proposed intervention concerns supervision rather than an assumption that attention is intrinsically superior to convolution.

2.2 Pseudo-labeling and knowledge distillation

Conventional knowledge distillation transfers teacher outputs to a student, often through softened classification responses [41]. Here the teacher is an offline geometric registration process that provides continuous pose targets rather than class logits. Registration-derived supervision supports LiDAR odometry methods such as DeLORA and SelfVoxeLO [12, 13]. General pseudo-labeling research considers target selection, confidence and calibration [42, 43]; a learning-based visual-odometry survey reviews domain-shift risks and reliability concerns [44]. Here, target filtering is based on geometric support. It is not treated as a calibrated prediction of target-error probability.

2.3 Degeneracy and observability

Eigenvalue and Hessian analyses diagnose poorly constrained modes in registration and state estimation. LOAM uses practical degeneracy handling, and subsequent work formalizes detection and solution remapping [10, 45]. X-ICP and Switch-SLAM incorporate localizability or sensing-condition awareness [11, 46]. Visual–inertial observability analyses provide background on unobservable modes and consistency [47, 48]; the present student uses no inertial measurements. Point-plane SLAM directly illustrates plane-normal and plane-offset constraints and degeneracy when insufficient independent planes are observed [49]. KISS-ICP supplies the point-to-point registration algorithm for the adapted sparse-depth baseline used here [50], not a sensor-matched depth-camera state-of-the-art comparison.

The present method transfers geometric support information into the training loss. Its local analysis identifies directions without direct supervision. That property is distinct from global convergence, calibrated uncertainty and long-horizon stability.

3. METHODOLOGY

Fig. 1 separates offline pseudo-label generation, loss-time filtering and depth-only student inference.

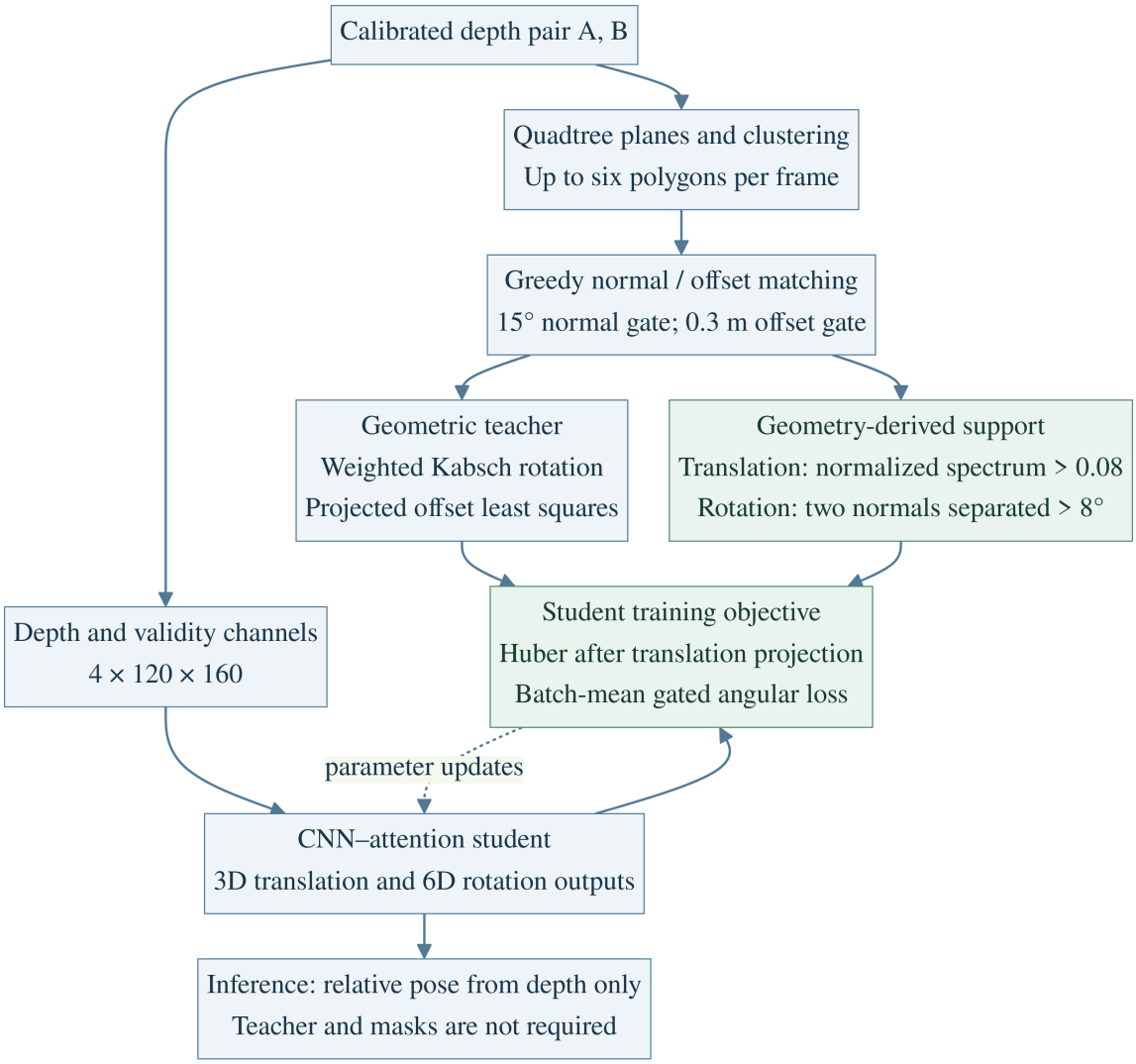


Fig. 1. Overview of the learning framework. Plane alignment and support estimation provide training targets and loss masks offline. The student receives depth and validity channels only at inference. The threshold sweep changes the loss mask while keeping teacher labels and the generation projector fixed.

3.1 Plane alignment and pose convention

A plane is represented by a unit normal n and an offset d, with $n^\top x = d$. For a matched pair k, the measurements are $(n_{A,k}, d_{A,k})$ and $(n_{B,k}, d_{B,k})$. The pose maps camera-B coordinates into camera A:

$$x_A = R_{B\to A} x_B + t_{B\to A}. \tag{1}$$

For consistent static-plane correspondences, $n_{A,k} = R n_{B,k}$ and $n_{A,k}^\top t = d_{A,k} - d_{B,k}$. Correspondences and positive area-derived weights a_k are held fixed within a solve. Let $\beta_k = a_k/(\sum_j a_j + \epsilon)$, where $\epsilon = 10^{-12}$ is a numerical guard. The rows of N_A and N_B are the matched normals.
The teacher computes $M = N_B^\top \operatorname{diag}(\beta) N_A = U_M \Sigma_M V_M^\top$ and uses the determinant-corrected weighted Kabsch rotation:

$$R_{\mathrm{pl}} = (U_M D_M V_M^\top)^\top, \qquad D_M = \operatorname{diag}\Big(1, 1, \operatorname{sign}\det(U_M V_M^\top)\Big). \tag{2}$$

Translation is estimated by row-weighted least squares and projected onto the supported subspace:

$$t_{\mathrm{LS}} = \arg\min_t \sum_k a_k^2 \left(n_{A,k}^\top t - (d_{A,k} - d_{B,k})\right)^2, \qquad t_{\mathrm{pl}} = P_{\mathrm{obs}} t_{\mathrm{LS}}. \tag{3}$$

The least-squares Gram matrix is $H_{\mathrm{off}} = \sum_k a_k^2 n_{A,k} n_{A,k}^\top$. The support proxy uses linearly normalized weights:

$$H_{\mathrm{mask}} = \sum_k \beta_k n_{A,k} n_{A,k}^\top = U \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3) U^\top, \quad \lambda_1 \ge \lambda_2 \ge \lambda_3 \ge 0. \tag{4}$$

These matrices have different relative weighting and conditioning, but the same exact nullspace on a common positive-weight support. In particular, for $G = \sum_k c_k n_k n_k^\top$ with $c_k > 0$,

$$x^\top G x = \sum_k c_k (n_k^\top x)^2, \qquad \ker G = \operatorname{span}\{n_k\}^\perp. \tag{5}$$

3.2 Local rotational and joint-pose identifiability

Consider the normal-alignment objective $F_R = \frac{1}{2}\sum_k \beta_k \|n_{A,k} - R n_{B,k}\|^2$. Under a left perturbation $R = \exp([\delta\theta]_\times) R_0$, let $p_k = R_0 n_{B,k}$ and $e_k = n_{A,k} - R n_{B,k}$. Then

$$e_k(\delta\theta) = e_k(0) + [p_k]_\times \delta\theta + O(\|\delta\theta\|^2), \qquad H_R = \sum_k \beta_k (I - p_k p_k^\top). \tag{6}$$

At a consistent alignment, $p_k = n_{A,k}$, so

$$H_R = \left(\sum_k \beta_k\right) I - H_{\mathrm{mask}}, \qquad x^\top H_R x = \sum_k \beta_k \|n_{A,k} \times x\|^2. \tag{7}$$

One plane leaves rotation about its normal unresolved and gives $\operatorname{rank}(H_R) = 2$. Two non-collinear normals give rotation rank three, although their translation Gram remains rank two. Three linearly independent normals give full translation rank. For the separable normal and offset residual model, the joint Gauss–Newton matrix is block diagonal, $\operatorname{diag}(H_R, H_{\mathrm{off}})$. The resulting local joint ranks are three, five and six for one plane, two independent normals and three spanning normals, respectively. This analysis assumes fixed correspondences and consistent plane measurements; it is not a claim about a different coupled residual or about all information present in a depth image.
Rotational support is therefore not defined by a thresholded translation eigenvalue. The operational rotation gate is

$$m_R = \mathbf{1}\Big\{\exists i < j : \ \arccos\Big(\big|n_{A,i}^\top n_{A,j}\big|\Big) > 8^\circ\Big\}. \tag{8}$$

This rule rejects parallel or nearly parallel normal configurations. It does not guarantee a well-conditioned rotation estimate when one plane has negligible weight, or correct association when moving objects violate the static-scene model. The gate is a practical support test rather than a global stability certificate.

3.3 Support masks and training objective

The normalized spectrum is $s_j = \lambda_j/(\lambda_1 + \epsilon)$. Pseudo-label generation uses

$$m_j^{\text{gen}} = \mathbf{1}\{s_j > 0.08\}, \qquad P_{\text{obs}} = U\,\text{diag}(m^{\text{gen}})U^{\mathsf{T}}, \qquad P_{\text{null}} = I - P_{\text{obs}}. \tag{9}$$

The primary OA-mask loss uses $m^{\text{loss}} = m^{\text{gen}}$. In the threshold sensitivity experiment, the loss mask is instead $m_j^{\text{loss}} = \mathbf{1}\{s_j/\max(s_1, \epsilon) \geq \sigma_{\text{th}}\}$. This branch uses an inclusive comparison. Only the loss mask changes: the teacher labels, generation projector and rotation flags remain fixed. Lowering the loss threshold can thus supervise coordinates already zeroed in the teacher translation; it does not recover additional geometric information.

The fixed 0.08 setting rejects translation directions whose normal-support eigenvalue is at most eight percent of the leading value. On the retained eigenspace, the strict rule bounds the support-proxy condition number by $\kappa < 12.5$; the inclusive loss rule gives $\kappa \leq 12.5$. This restricted bound assumes a nonzero leading eigenvalue and neglects the numerical guard. It does not bound the full singular matrix or calibrate target uncertainty. The threshold is retained as the original fixed baseline for controlled comparison, not as a validation-selected optimum. Its original selection rationale and pre-test chronology are not established by the available records. Its adverse held-out behavior is analyzed in Section 4.9.

For training example b, let U_b, s_b and m_b^{loss} denote its basis, normalized spectrum and loss mask. The translation transform is

$$W_b = \begin{cases} I, & \text{Naive or Gate-only,} \\ U_b\,\text{diag}(m_b^{\text{loss}})U_b^{\mathsf{T}}, & \text{OA-mask,} \\ U_b\,\text{diag}(\sqrt{\text{clip}(s_b, 0, 1)})U_b^{\mathsf{T}}, & \text{OA-soft.} \end{cases} \tag{10}$$

Gate-only retains a pair only if all three translation axes and the rotation flag are active. The other variants retain the common geometrically matched, sanity-filtered pool. For $r_b = \hat{t}_b - t_{\text{pl},b}$, the Huber penalty is

$$h_\delta(z) = \begin{cases} z^2/2, & |z| \leq \delta, \\ \delta(|z| - \delta/2), & |z| > \delta. \end{cases} \tag{11}$$

With batch size B, the complete objective is

$$\mathcal{L}_t = \frac{1}{B}\sum_{b=1}^{B}\sum_{j=1}^{3} h_\delta([W_b r_b]_j), \qquad \delta = 0.02\,\text{m}, \tag{12}$$

$$\mathcal{L}_R = \frac{1}{B}\sum_{b=1}^{B} w_b^R \arccos\left[\text{clip}\left(\frac{\text{tr}(\hat{R}_b^{\mathsf{T}} R_{\text{pl},b}) - 1}{2}, -0.999999, 0.999999\right)\right], \tag{13}$$

$$\mathcal{L} = \mathcal{L}_t + \lambda_R \mathcal{L}_R, \qquad \lambda_R = 1. \tag{14}$$

Here $w_b^R = m_{R,b}$ for OA-mask/OA-soft and $w_b^R = 1$ for Naive/Gate-only. The numerical weight λ_R specifies the relative scaling of the translation and angular losses; their physical units differ. The student predicts a continuous six-dimensional rotation representation, converted to a proper rotation by Gram–Schmidt [51]. Teacher rotations are encoded and reconstructed in the same representation for the loss. The angular term is averaged over the entire batch, not only its active labels. Huber is applied componentwise after the transform, rather than independently in the eigenbasis.

Let $\ell_{t,b} = \sum_j h_\delta([W_b r_b]_j)$ and $\psi_\delta(z) = \text{clip}(z, -\delta, \delta)$, applied componentwise. Then

$$\nabla_{\hat{t}_b} \ell_{t,b} = W_b^{\mathsf{T}} \psi_\delta(W_b r_b). \tag{15}$$

The transforms are symmetric, so this gradient lies in range(W_b) and is orthogonal to its nullspace. For the hard projector, changing an unsupported residual component leaves the loss unchanged. Shared network parameters may nevertheless learn priors in those directions. Furthermore, area weights are geometric proxies rather than measured inverse noise covariances; the support matrices are not claimed to be calibrated Fisher information.

3.4 Implementation details

Depth-to-plane extraction. Valid depth pixels are back-projected with dataset-specific pinhole intrinsics. A recursive quadtree fits planes by principal component analysis/singular value decomposition. Initial blocks are at most 128×128 pixels, the minimum block width is eight pixels and the minimum valid-pixel ratio is 0.9. A block is accepted when its plane mean squared residual, curvature and depth-change residual fall below t_m, t_c and f_s, respectively. VCU and ETH3D

use $t_m = 10^{-4}\,\mathrm{m}^2$, $t_c = 4 \times 10^{-3}$ and $f_s = 0.03\,\mathrm{m}$. TUM uses $t_m = 10^{-5}\,\mathrm{m}^2$, $t_c = 8 \times 10^{-4}$ and $f_s = 0.03\,\mathrm{m}$. Accepted plane regions are greedily clustered when their acute normal separation is at most 8° and their offset difference is at most 0.05 m. The generation path performs extraction and clustering without additional median denoising.
At most the six largest plane polygons are retained per image. Greedy one-to-one matching requires an acute normal angle below 15° and an offset-magnitude difference $\big||d_A| - |d_B|\big| < 0.3\,\mathrm{m}$. Candidate cost is the sum of the angle and offset difference normalized by these limits. The pair weight is the smaller polygon area. Normals and offsets are jointly sign-flipped when necessary to make matched normals consistent. The resulting measurements enter Eqs. (2) and (3). Pairs without matches are recorded with zero translation, identity rotation, zero spectrum and inactive support flags; they are excluded from self-supervised training but retained in evaluation whenever ground truth is available.

Depth calibration and sampling. VCU depth is converted from millimetres to metres and downsampled from 640×480 to 320×240. The extraction intrinsics are $(f_x, f_y, c_x, c_y) = (230.29, 230.38, 161.71, 119.98)$. Every sixth depth frame is retained, giving approximately five frames per second. TUM and ETH3D use a sampling stride of three and at most 400 frames per sequence. TUM depths are divided by 5000, downsampled to 320×240 for geometry, and use $(267.7, 269.6, 160.05, 123.8)$ at that scale. ETH3D uses its sequence calibration with focal lengths and principal point scaled by one half after spatial downsampling. Network depth images have spatial dimensions 120×160.
The input tensor is $[D_A/5, M_A, D_B/5, M_B]$, where depth is expressed in metres before normalization and M identifies finite depths greater than 0.1 m. Invalid depth values are replaced with zero. All self-supervised variants use the same input construction. Geometric labels and masks are training targets or loss controls, not network inputs.

Label filtering. Records contain the aligned pose, support basis U, normalized spectrum, translation-axis mask, rotation flag and number of matched planes. Training excludes no-match records and poses with $\|t_{\mathrm{pl}}\| > 0.35\,\mathrm{m}$. The primary strict translation threshold is 0.08. Gate-only adds full translation-support and active-rotation requirements; OA-mask and OA-soft retain partially constrained pairs and change their loss contributions as specified in Eq. (10).

Network architecture. The hybrid student accepts a $4 \times 120 \times 160$ tensor. A stride-two $4 \to 32$ stem and three residual stages with widths 32, 64 and 128 precede two pre-normalized attention blocks. Each block has feature dimension 128 and four heads over an 8×10 grid. Global average pooling feeds separate 128-unit translation and rotation heads, with three and six outputs. The hybrid contains 620,169 trainable parameters. The CNN ablation removes attention and contains 344,969 parameters; the six-block ViT uses 8×8 patches, 300 tokens and 900,617 parameters. Table XXVI reports their inference costs under a common profiling setup.

Training and augmentation. The primary Naive/OA-mask comparison uses 40 epochs, batch size 32, AdamW with learning rate 3×10^{-4} and weight decay 10^{-4}, cosine annealing with $T_{\max} = 40$, gradient clipping at norm one, two data-loader workers and incomplete final batches discarded. Implementation uses PyTorch [52] and follows the Adam/AdamW optimizer formulations [53, 54]. With probability 0.5, Gaussian depth noise of standard deviation 0.01 m is added to valid pixels. With probability 0.3, A-frame depth and validity are jointly subjected to pixel dropout with retention probability 0.95; this dropout is not applied to frame B. The same augmentation implementation and probabilities are used across variants.
Ground-truth fields in the pair files are reserved for evaluation and the separately identified supervised reference. They are not read as self-supervised learning targets. No inertial measurements enter pseudo-label generation or student inference. VCU gyroscope data are used only to estimate the evaluation-side MoCap–depth clock offset by angular-speed correlation with the reference orientation profile.

Area-weight interpretation. The polygon-area helper uses the fixed focal lengths $(366.193, 365.456)$ in the area expression $\sum_r s_{u,r} s_{v,r} z_r^2 / [f_x f_y \max(|n_{z,r}|, 0.15)]$, where $s_{u,r}, s_{v,r}$ are region widths in pixels and z_r is depth. Back-projection and fitted normals still use the dataset-specific intrinsics above. For fixed extracted regions, changing only the focal-length product in the area helper scales all areas by the same positive factor, preserving their ranking and normalized relative weights. Absolute polygon areas are therefore not treated as physically calibrated measurements; the retained implementation uses them as relative geometric weights.

4. EXPERIMENTS

4.1 Datasets and evaluation pairs

The primary training pool contains TUM Freiburg3 cabinet and structure-notexture-near, and VCU lab-simple1 and lab-dynamic1. Main held-out evaluation uses TUM structure-notexture-far and walking-static, and VCU lab-light1, corridor1

and lab-motion1. ETH3D Einstein-dark, kidnap-dark, ceiling-1 and camera-shake-3 provide zero-shot transfer tests without ETH3D training or fine-tuning. TUM and VCU recording sources are documented by their benchmark publications [9, 55]. The ETH3D recordings come from the provider's SLAM dataset release; the benchmark publication [56] documents that benchmark and its recording design. BAD SLAM is not implemented or used as an experimental comparator here.

Table I distinguishes stored pair counts, reference availability and geometric training support. Raw pairs are stored consecutive-depth image pairs before reference-validity and training filtering. The four training sequences contain 1,622 raw pairs; common filtering retains 1,547, whereas Gate-only retains 164. Main evaluation contains 2,508 raw pairs, of which 1,784 have valid reference increments. ETH3D contains 1,067 raw pairs, of which 863 have valid references. The 13 current manifests sum to 5,197 raw pairs and 4,170 reference-valid pairs. This verified current total replaces the earlier aggregate of 5,167, a difference of 30. The original sequence-level ledger is unavailable, so its historical allocation cannot be assigned to a sequence or preprocessing step. Reference validity and geometric training retention define different subsets; ground-truth availability is not required for self-supervised training.

TABLE I. DATA ACCOUNTING AFTER THE SPECIFIED PREPROCESSING. PLANES COUNTS PAIRS WITH AT LEAST ONE MATCH. RETAINED AND GATE ARE TRAINING COUNTS AFTER COMMON FILTERING AND THE ADDITIONAL GATE-ONLY CONDITION, RESPECTIVELY. GROUND TRUTH IS NOT REQUIRED FOR SELF-SUPERVISED TRAINING. TOTAL: 5,197 PAIRS; 4,170 REFERENCE-VALID PAIRS.

Sequence	Role	Pairs	GT	Planes	Rot. flag	Retained	Gate
TUM cabinet	Train	373	372	373	303	373	60
TUM near	Train	352	352	352	352	352	76
VCU dynamic1	Train	672	633	609	464	607	9
VCU simple1	Train	225	166	215	140	215	19
TUM far	Test	265	264	150	95	–	–
TUM walking	Test	240	239	145	0	–	–
VCU corridor1	Test	998	278	757	282	–	–
VCU light1	Test	479	477	403	243	–	–
VCU motion1	Test	526	526	421	176	–	–
ETH shake	Zero-shot	94	94	86	52	–	–
ETH ceiling	Zero-shot	399	289	399	282	–	–
ETH Einstein	Zero-shot	399	374	380	322	–	–
ETH kidnap	Zero-shot	175	106	172	124	–	–

Evaluation includes every reference-valid pair, including pairs without matched planes. Predicted increments map frame B into frame A. For VCU, the supplied headerless eight-field trajectory is treated as an unlabeled reference: after timestamp association, $R_{AB}^{\mathrm{ref}} = R_{WA}^{\mathsf{T}} R_{WB}$ and $t_{AB}^{\mathrm{ref}} = R_{WA}^{\mathsf{T}}(p_{WB} - p_{WA})$. All teacher, student, classical-baseline and supervised-reference predictions are scored against these same increments on the same valid pairs. This computationally consistent *VCU direct-reference convention* does not certify whether the supplied poses describe the camera, body/IMU or marker. Verified-camera-frame TUM and ETH3D evaluations provide the primary physical-accuracy evidence; VCU comparisons are secondary and frame-conditional. Section 4.7 reports the conditional body-to-camera sensitivity without selecting a frame by test error.

4.2 Comparison design and training provenance

Two evaluation cohorts have different evidentiary status. The primary cohort comprises ten documented Naive/OA-mask runs at seeds 0–4 on the four sequences in Table I, with recorded input manifests and initial-state checksums. The completed rerun campaign adds 39 new training runs, including the five-seed component controls and the five-seed Cabinet threshold sweep; all new runs use the specified 40-epoch budget. The current seed-zero comparator tables use fresh completed outputs for Gate-only, OA-soft, backbone, threshold, supervised-reference and related variants, while the Naive and full-OA seed-zero rows reuse the fresh documented primary seed-zero checkpoints. These comparator rows are excluded from the primary five-seed averages. Associated generation records retain 1,622 raw training pairs and 1,547 effective Naive/OA pairs; the current 13-manifest total is 5,197 raw pairs. The preserved initial teacher old/new comparison remains a limited fidelity audit, and numerical agreement with archived evaluator outputs does not establish byte identity or historical training-input identity.

For the five-seed primary comparison, Naive and OA-mask use the identical architecture, initialization protocol, 1,547-pair pool and optimizer settings. Each has 48 full batches per epoch and 1,920 updates over 40 epochs. Pair and initial-state checksums are recorded. Gate-only has five batches per epoch and 200 updates, so its current seed-zero comparison reflects both filtering and a reduced data/update budget. The supervised reference uses 1,523 admissible ground-truth pairs, 1,880

updates and no depth augmentation in its separate training path. It is an empirical reference, not an accuracy upper bound. Backbone comparisons are current seed-zero ablations rather than replicated architecture rankings.

The Cabinet threshold diagnostic excludes Cabinet from training, leaving 1,174 retained pairs and 1,440 updates per run. It comprises 25 completed runs: five seeds at each loss-mask threshold $0.02, 0.05, 0.08, 0.15, 0.30$, with teacher labels fixed. Cabinet is used only for this diagnostic, not as a held-out test in the primary four-sequence training protocol. The documented follow-up keeps the existing 0.08 setting as the standard baseline; the completed sweep is exploratory and does not select a replacement or establish the original selection chronology. The five-seed component controls are likewise exploratory and do not replace the primary method.

Iterative Closest Point (ICP) and KISS-ICP ("Keep It Simple, Stupid" ICP) are compared under an explicitly adapted sparse-depth protocol, not their official full-sensor benchmark protocols. An identity predictor is also evaluated to contextualize the approximately static-camera walking sequence. Published RGB, visual–inertial and LiDAR-only results are background rather than interchangeable sensor-matched competitors.

4.3 Metrics and uncertainty

For each valid reference pair, translation and rotation errors are

$$e_t = \|\hat{t} - t_{\mathrm{GT}}\|_2, \qquad e_R = \arccos\left[\mathrm{clip}\left(\frac{\mathrm{tr}(\hat{R}^{\mathsf{T}} R_{\mathrm{GT}}) - 1}{2}, -1, 1\right)\right]. \tag{16}$$

Rotation is converted to degrees. Both cohorts' reported tables and CDFs use this bounded trace-angle evaluation. RMSE is $\sqrt{N^{-1}\sum_i e_i^2}$ and MAE is $N^{-1}\sum_i e_i$. Translation uses metres in the implementation and equations, and centimetres in result tables and plots. Angular training loss uses radians. The components are not combined into a mixed-unit accuracy ranking.

Uncertainty is estimated across training seeds, not correlated video frames. For each sequence, the five paired differences are $d_s = E_{\mathrm{OA},s} - E_{\mathrm{N},s}$; negative values favor OA. Descriptive pointwise 95% intervals are $\bar{d} \pm t_{0.975,4}\, s_d/\sqrt{5}$. The strengthened analysis uses two-sided paired t tests with four degrees of freedom and Holm adjustment across all nine rotation-RMSE sequences, with a separate nine-sequence translation-RMSE family. Conservative simultaneous 95% intervals use the Bonferroni critical value $t_{1-0.05/(2\cdot 9),4} = 5.4366$. These parametric analyses assume an appropriate paired seed-difference distribution. An exact sensitivity analysis enumerates all $2^5 = 32$ paired sign flips and counts tied or more extreme absolute statistics; its smallest nonzero two-sided p value is $2/32 = 0.0625$. These additional inferential checks are exploratory, do not treat video frames as independent samples, and do not estimate population-wide environment effects.

4.4 Evaluation implementation

Both evaluation cohorts use the preprocessing, pose conventions and bounded trace-angle metric specified above. The reference implementation uses PyTorch 2.15.0 and NumPy 2.5.4 on CPU. Classical baselines use cached 120×160 depth images, rescaled intrinsics, point-sampling step two and depths in the range 0.1–8 m. The KISS-ICP sparse-depth adaptation uses version 1.3.0. Baseline predictions are scored on the same reference-valid pairs as the students.

The geometric teacher was recomputed from the specified pipeline. Table II compares its outputs with recorded aggregates using the same original evaluator for both, rather than mixing metrics. Translation RMSE and MAE reproduce on all nine sequences; rotation RMSE and MAE reproduce only on TUM walking. The eight unreproduced rotation results are TUM far, VCU light1, corridor1 and motion1, and ETH3D Einstein-dark, kidnap-dark, ceiling-1 and camera-shake-3. The largest absolute rotation-RMSE discrepancy is 9.416512° on TUM far. Switching the recomputed rotation evaluation from the original logarithm-based routine to Eq. (16) changes sequence RMSE by at most 0.011382°, far below that discrepancy. It therefore does not explain the old/new differences. Rank-deficient Kabsch freedom permits different normal-fit-equivalent rotations, but it has not been established as their historical cause. The missing per-pair teacher outputs and runtime record prevent further causal localization. Main comparison tables consistently use the recomputed teacher, not the eight unreproduced recorded rotations. Retained student RMSE/MAE aggregates reproduce in all 36 checkpoint/sequence comparisons under the original evaluator; this numerical agreement does not establish their training-input identity.

TABLE II. RECORDED VERSUS RECOMPUTED TEACHER ERRORS UNDER THE SAME ORIGINAL EVALUATOR. ROTATION RMSE/MAE ARE DEGREES; THE FINAL TWO COLUMNS ARE TRANSLATION RMSE/MAE IN CM, IDENTICAL FOR BOTH EVALUATIONS ON EVERY SEQUENCE. ONLY WALKING REPRODUCES BOTH ROTATION AGGREGATES. MAIN RESULTS USE THE RECOMPUTED TEACHER WITH THE BOUNDED TRACE-ANGLE METRIC.

Sequence	Recorded R_{RMSE}	New R_{RMSE}	Recorded R_{MAE}	New R_{MAE}	T_{RMSE}	T_{MAE}
TUM far	47.299	37.882	17.197	13.414	2.010	1.684
TUM walking	0.481	0.481	0.405	0.405	4.157	1.914
VCU light1	55.935	55.301	30.697	30.069	16.963	14.834
VCU corridor1	51.052	56.379	28.844	31.091	21.974	19.828
VCU motion1	67.658	68.606	39.855	40.089	19.643	17.460
ETH3D Einstein	36.294	31.370	19.389	17.600	8.053	5.902
ETH3D kidnap	24.304	27.777	9.134	9.801	6.359	5.543
ETH3D ceiling	43.254	43.672	27.155	27.304	8.226	7.631
ETH3D shake	55.353	57.927	36.234	37.545	5.997	4.907

4.5 Primary documented five-seed comparison

Tables III and IV report mean RMSE, seed standard deviation and paired intervals. OA-mask has lower mean rotation RMSE on all nine evaluation sequences. Six pointwise rotation-RMSE intervals exclude zero in favor of OA; the intervals for TUM far, ETH3D kidnap and camera-shake cross zero. Walking gives $4.030 \pm 1.080^\circ$ for OA versus $36.459 \pm 10.134^\circ$ for Naive. In contrast, OA has higher mean translation RMSE on all nine sequences. The benefit is therefore component-specific rather than overall pose-error dominance.

TABLE III. INDEPENDENT FIVE-SEED ROTATION RMSE (°). Δ = OA − NAIVE; NEGATIVE VALUES FAVOR OA. INTERVALS ARE POINTWISE PAIRED STUDENT-T INTERVALS, NOT MULTIPLICITY-ADJUSTED. THE SEED-ZERO COMPARATOR COHORT IS NOT INCLUDED IN THESE STATISTICS.

Sequence	Naive mean ± SD	OA-mask mean ± SD	Δ [95% paired CI]
TUM far	2.896 ± 1.732	0.865 ± 0.116	−2.031 [−4.195, 0.133]
TUM walking	36.459 ± 10.134	4.030 ± 1.080	−32.429 [−45.910, −18.948]
VCU light1	22.382 ± 6.113	8.481 ± 0.324	−13.901 [−21.611, −6.191]
VCU corridor1	22.267 ± 4.852	10.873 ± 0.188	−11.394 [−17.522, −5.265]
VCU motion1	26.364 ± 5.416	9.448 ± 0.566	−16.915 [−24.131, −9.699]
ETH Einstein	13.566 ± 4.157	3.445 ± 0.188	−10.121 [−15.089, −5.154]
ETH kidnap	8.571 ± 4.431	3.928 ± 0.089	−4.643 [−10.208, 0.923]
ETH ceiling	13.873 ± 3.889	4.736 ± 0.375	−9.137 [−13.650, −4.624]
ETH shake	16.693 ± 2.439	14.367 ± 0.258	−2.326 [−5.528, 0.875]

TABLE IV. INDEPENDENT FIVE-SEED TRANSLATION RMSE (CM). Δ = OA − NAIVE; NEGATIVE VALUES FAVOR OA. INTERVALS ARE POINTWISE PAIRED STUDENT-T INTERVALS, NOT MULTIPLICITY-ADJUSTED. THE SEED-ZERO COMPARATOR COHORT IS NOT INCLUDED IN THESE STATISTICS.

Sequence	Naive mean ± SD	OA-mask mean ± SD	Δ [95% paired CI]
TUM far	1.559 ± 0.185	1.576 ± 0.407	0.017 [−0.556, 0.589]
TUM walking	1.005 ± 0.199	4.208 ± 1.014	3.202 [2.017, 4.387]
VCU light1	12.034 ± 0.179	12.980 ± 0.310	0.946 [0.634, 1.257]
VCU corridor1	15.386 ± 0.199	16.282 ± 0.315	0.896 [0.590, 1.202]
VCU motion1	14.496 ± 0.190	15.198 ± 0.619	0.702 [0.007, 1.396]
ETH Einstein	4.158 ± 0.309	5.338 ± 0.602	1.180 [0.808, 1.552]
ETH kidnap	5.561 ± 0.134	6.857 ± 0.825	1.297 [0.342, 2.251]
ETH ceiling	7.285 ± 0.348	7.769 ± 0.878	0.484 [−1.028, 1.996]
ETH shake	4.640 ± 0.161	5.244 ± 0.361	0.604 [−0.025, 1.233]

Tables V and VI provide the corresponding MAE results. Fig. 2 displays RMSE differences and their pointwise paired intervals for both components.

TABLE V. INDEPENDENT FIVE-SEED ROTATION MAE (°). Δ = OA − NAIVE; NEGATIVE VALUES FAVOR OA. INTERVALS ARE POINTWISE PAIRED STUDENT-*T* INTERVALS, NOT MULTIPLICITY-ADJUSTED. THE SEED-ZERO COMPARATOR COHORT IS NOT INCLUDED IN THESE STATISTICS.

Sequence	Naive mean ± SD	OA-mask mean ± SD	Δ [95% paired CI]
TUM far	1.533 ± 0.720	0.782 ± 0.111	−0.751 [−1.660, 0.158]
TUM walking	26.726 ± 6.642	3.615 ± 1.016	−23.112 [−32.117, −14.107]
VCU light1	15.793 ± 3.628	7.559 ± 0.317	−8.234 [−12.846, −3.623]
VCU corridor1	16.860 ± 3.243	9.734 ± 0.201	−7.126 [−11.269, −2.982]
VCU motion1	18.871 ± 3.426	8.432 ± 0.526	−10.439 [−15.095, −5.783]
ETH Einstein	7.555 ± 2.300	3.024 ± 0.177	−4.531 [−7.228, −1.834]
ETH kidnap	6.052 ± 1.752	3.494 ± 0.149	−2.558 [−4.858, −0.258]
ETH ceiling	8.119 ± 2.219	4.389 ± 0.416	−3.731 [−6.033, −1.428]
ETH shake	13.208 ± 1.063	11.931 ± 0.326	−1.276 [−2.780, 0.228]

TABLE VI. INDEPENDENT FIVE-SEED TRANSLATION MAE (CM). Δ = OA − NAIVE; NEGATIVE VALUES FAVOR OA. INTERVALS ARE POINTWISE PAIRED STUDENT-*T* INTERVALS, NOT MULTIPLICITY-ADJUSTED. THE SEED-ZERO COMPARATOR COHORT IS NOT INCLUDED IN THESE STATISTICS.

Sequence	Naive mean ± SD	OA-mask mean ± SD	Δ [95% paired CI]
TUM far	1.443 ± 0.170	1.442 ± 0.392	−0.001 [−0.544, 0.542]
TUM walking	0.925 ± 0.191	3.912 ± 0.808	2.987 [2.038, 3.935]
VCU light1	11.017 ± 0.169	11.898 ± 0.254	0.881 [0.630, 1.132]
VCU corridor1	14.332 ± 0.195	15.185 ± 0.313	0.852 [0.572, 1.133]
VCU motion1	13.393 ± 0.176	14.078 ± 0.499	0.685 [0.140, 1.229]
ETH Einstein	3.709 ± 0.248	4.705 ± 0.517	0.996 [0.660, 1.333]
ETH kidnap	5.288 ± 0.103	6.245 ± 0.639	0.957 [0.204, 1.711]
ETH ceiling	6.943 ± 0.322	7.294 ± 0.932	0.351 [−1.198, 1.901]
ETH shake	4.305 ± 0.149	4.703 ± 0.269	0.398 [−0.101, 0.897]

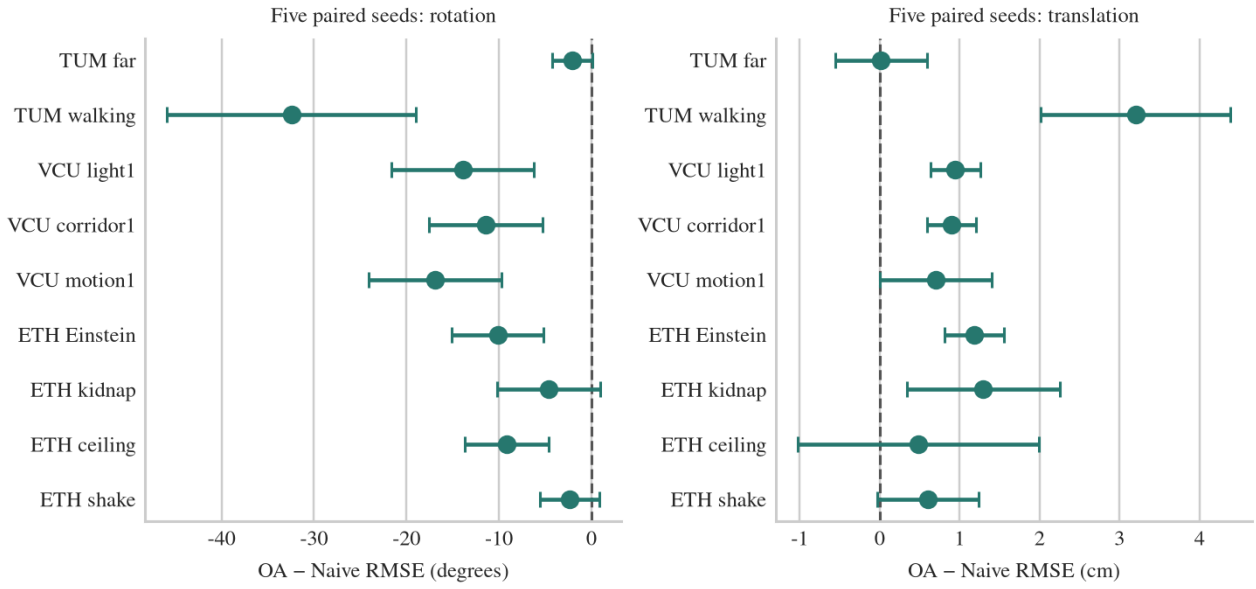


Fig. 2. OA-minus-Naive RMSE differences across five paired seeds, with pointwise 95% Student-*t* intervals. Negative differences favor OA; zero denotes no difference. Rotation and translation are reported separately.

4.5.1. Multiplicity-aware and exact paired sensitivity

Tables VII and VIII strengthen the inferential accounting without changing the ten training runs. Six rotation comparisons retain Holm-adjusted $p < 0.05$: TUM walking, VCU light1, corridor1 and motion1, and ETH3D Einstein-dark and ceiling-1. The more conservative simultaneous intervals exclude zero on four: walking, motion1, Einstein-dark and ceiling-1. The exact sign-flip values are 0.0625 for the first eight rotation sequences and 0.125 for camera-shake; none reaches 0.05 under that sensitivity test. Thus the sample-mean pattern is broader than the simultaneous inferential support.

Translation has higher OA mean RMSE on all nine sequences. Four adverse differences remain below Holm-adjusted $p = 0.05$ and have zero-excluding simultaneous intervals: walking, VCU light1 and corridor1, and ETH3D Einstein-dark. Translation exact-test values likewise do not reach 0.05. The statistical units remain the five training seeds; neither the nine sequences nor the constituent video frames are assumed to provide independent replicated training runs.

TABLE VII. STRENGTHENED PAIRED INFERENCE FOR ROTATION RMSE (DEGREES). Δ = OA − NAIVE. HOLM ADJUSTS THE NINE-SEQUENCE COMPONENT-SPECIFIC FAMILY; SIMULTANEOUS INTERVALS USE BONFERRONI CORRECTION. EXACT *P* ENUMERATES ALL 32 PAIRED SIGN FLIPS AND CANNOT BE SMALLER THAN 0.0625 WITH FIVE SEEDS. MEAN AND SD ARE REPORTED IN THE PRIMARY DESCRIPTIVE TABLES.

Sequence	Δ	Holm p	Exact p	Simultaneous 95% CI
TUM far	-2.031	0.1790	0.0625	[−6.269, 2.206]
TUM walking	-32.429	0.0235	0.0625	[−58.826, −6.032]
VCU light1	-13.901	0.0337	0.0625	[−28.998, 1.196]
VCU corridor1	-11.394	0.0337	0.0625	[−23.394, 0.606]
VCU motion1	-16.915	0.0235	0.0625	[−31.045, −2.786]
ETH Einstein	-10.121	0.0337	0.0625	[−19.848, −0.395]
ETH kidnap	-4.643	0.1790	0.0625	[−15.541, 6.256]
ETH ceiling	-9.137	0.0337	0.0625	[−17.975, −0.300]
ETH shake	-2.326	0.1790	0.1250	[−8.596, 3.943]

TABLE VIII. STRENGTHENED PAIRED INFERENCE FOR TRANSLATION RMSE (CM). Δ = OA − NAIVE. HOLM ADJUSTS THE NINE-SEQUENCE COMPONENT-SPECIFIC FAMILY; SIMULTANEOUS INTERVALS USE BONFERRONI CORRECTION. EXACT *P* ENUMERATES ALL 32 PAIRED SIGN FLIPS AND CANNOT BE SMALLER THAN 0.0625 WITH FIVE SEEDS. MEAN AND SD ARE REPORTED IN THE PRIMARY DESCRIPTIVE TABLES.

Sequence	Δ	Holm p	Exact p	Simultaneous 95% CI
TUM far	0.017	0.9394	1.0000	[−1.105, 1.138]
TUM walking	3.202	0.0101	0.0625	[0.882, 5.523]
VCU light1	0.946	0.0087	0.0625	[0.335, 1.556]
VCU corridor1	0.896	0.0087	0.0625	[0.297, 1.496]
VCU motion1	0.702	0.1941	0.0625	[−0.658, 2.062]
ETH Einstein	1.180	0.0083	0.0625	[0.451, 1.908]
ETH kidnap	1.297	0.0980	0.0625	[−0.573, 3.166]
ETH ceiling	0.484	0.8494	0.6250	[−2.477, 3.444]
ETH shake	0.604	0.1941	0.1250	[−0.628, 1.835]

4.6 Current seed-zero comparator diagnostics

Tables IX and X report current completed seed-zero comparisons alongside the recomputed teacher and adapted classical baselines. Naive and full OA-mask seed-zero rows are the fresh documented primary seed-zero checkpoints; the other learned rows are fresh completed seed-zero comparators. These rows are descriptive and are not the five-seed average. On the current walking row, OA-mask has rotation RMSE 4.28° versus 46.93° for Naive, while its translation RMSE is 4.28 cm versus 0.76 cm.

The same walking case shows that OA does not dominate the geometric teacher: teacher and identity rotation RMSE are approximately 0.481°, while identity translation RMSE is 0.2146 cm. Under the specified sparse-depth protocol and primary VCU reference interpretation, KISS-ICP/OA translation-RMSE ratios are approximately 3.39, 4.69 and 4.91 on light1, corridor1 and motion1. These ratios are restricted to those adapted-baseline comparisons. The supervised reference has lower translation RMSE than OA-mask on four of the five main sequences: TUM walking, VCU light1, VCU corridor1 and VCU motion1. On TUM far, OA-mask has lower translation RMSE (1.13 versus 1.48 cm).

TABLE IX. REEXECUTED DOCUMENTED SEED-ZERO STUDENT/COMPARATOR VALUES AND UNCHANGED DOCUMENTED TEACHER/CLASSICAL VALUES: ROTATION RMSE (°). NAIVE AND OA-MASK REUSE THE FRESH DOCUMENTED PRIMARY SEED-ZERO CHECKPOINTS; ALL OTHER LEARNED ROWS ARE FRESH SEED-ZERO COMPARATORS. VCU ROWS ARE CONDITIONAL UNDER THE SUPPLIED REFERENCE CONVENTION.

Sequence	Teacher	ICP	KISS-ICP	Naive	Gate	OA-mask	OA-soft	GT reference
TUM far	37.88	0.38	0.78	3.16	1.49	0.76	1.09	0.93
TUM walking	0.48	0.81	0.60	46.93	4.64	4.28	3.91	2.49
VCU light1	55.30	10.64	14.42	20.28	8.46	9.02	8.59	6.34
VCU corridor1	56.39	15.65	24.24	23.34	10.57	11.13	10.84	9.42
VCU motion1	68.60	13.82	19.68	24.55	8.89	10.25	10.04	8.24

TABLE X. REEXECUTED DOCUMENTED SEED-ZERO STUDENT/COMPARATOR VALUES AND UNCHANGED DOCUMENTED TEACHER/CLASSICAL VALUES: TRANSLATION RMSE (CM). NAIVE AND OA-MASK REUSE THE FRESH DOCUMENTED PRIMARY SEED-ZERO CHECKPOINTS; ALL OTHER LEARNED ROWS ARE FRESH SEED-ZERO COMPARATORS. VCU ROWS ARE CONDITIONAL UNDER THE SUPPLIED REFERENCE CONVENTION.

Sequence	Teacher	ICP	KISS-ICP	Naive	Gate	OA-mask	OA-soft	GT reference
TUM far	2.01	1.17	2.26	1.75	1.65	1.13	1.85	1.48
TUM walking	4.16	3.45	2.53	0.76	2.83	4.28	3.03	2.29
VCU light1	16.96	28.54	44.14	11.81	12.13	13.01	12.82	9.47
VCU corridor1	21.97	39.76	76.76	15.22	15.68	16.37	16.86	13.04
VCU motion1	19.64	33.76	71.99	14.21	14.75	14.67	15.41	12.83

To define the descriptive rotation-gap comparison, let E_{N}, E_{OA} and E_{GT} denote sequence-level rotation RMSE for Naive, OA-mask and the supervised reference. For a nonzero denominator,

$$G = 100\,\frac{E_{\mathrm{N}} - E_{\mathrm{OA}}}{E_{\mathrm{N}} - E_{\mathrm{GT}}}\ \%. \tag{17}$$

Table XI provides all five values and the underlying errors. The approximate 80.8–96.0% range applies to walking and the three VCU cases; TUM far exceeds 100% because OA has lower rotation error than the supervised reference there. This ratio is neither a recovery probability nor a theoretical accuracy ceiling.

TABLE XI. DESCRIPTIVE REEXECUTED ROTATION-GAP FRACTION $100(E_N - E_{OA})/(E_N - E_{GT})$. THE NAIVE, OA-MASK AND GT-REFERENCE ENTRIES ARE THE CURRENT PRINTED SEED-ZERO ROTATION RMSE VALUES FROM TABLE IX; GAP FRACTIONS ARE RECOMPUTED FROM THOSE VALUES USING EQ. (17). VALUES ARE NOT CLIPPED AND THE SUPERVISED REFERENCE IS NOT AN UPPER BOUND.

Sequence	Naive	OA-mask	GT reference	Gap fraction
TUM far	3.16	0.76	0.93	107.6%
TUM walking	46.93	4.28	2.49	96.0%
VCU light1	20.28	9.02	6.34	80.8%
VCU corridor1	23.34	11.13	9.42	87.7%
VCU motion1	24.55	10.25	8.24	87.7%

4.7 Conditional reference-frame sensitivity

Let $E = {}^{B}T_{C}$ denote the documented extrinsic mapping camera coordinates into body coordinates. Under the hypothesis that the supplied reference increment is a body increment T_B, the corresponding camera increment is $T_C = E^{-1}T_BE$. Under the direct-reference convention, the supplied increment is used unchanged without assigning it a verified physical frame. The rounded extrinsic rotation is projected to the nearest proper rotation, and its translation lever arm is retained. Explicitly, if the supplied pose is confirmed as body/IMU, $R_C = R_{BC}^{\mathsf{T}}R_gR_{BC}$ and $t_C = R_{BC}^{\mathsf{T}}[t_g + (R_g - I)t_{BC}]$, where T_{BC} maps camera coordinates into body coordinates. A marker-target interpretation would require an additional documented hand-eye transform.

Tables XII and XIII evaluate fixed predictions under these two interpretations. The comparison tests sensitivity to reference convention, not the physical validity of either hypothesis. A marker-frame reference would require its own documented transform. The five-seed OA rotation-RMSE intervals remain favorable on all three VCU sequences under

both interpretations, but the numerical errors change. Thus stronger physical accuracy claims require an independently verified reference-frame schema.

TABLE XII. CONDITIONAL FRAME SENSITIVITY OF FIXED PREDICTIONS. NAIVE AND OA-MASK USE THE CURRENT DOCUMENTED PRIMARY SEED-ZERO CHECKPOINTS REPORTED IN TABLES IX AND X; KISS-ICP USES THE SAME FIXED CLASSICAL PREDICTIONS AS THOSE TABLES. A: SUPPLIED REFERENCE INCREMENT USED UNCHANGED. B: GT ASSUMED BODY/IMU AND TRANSFORMED WITH THE DOCUMENTED BODY-TO-CAMERA EXTRINSIC. NEITHER CASE CERTIFIES THE CSV SEMANTIC FRAME.

Sequence	Method	R_A (°)	R_B (°)	T_A (cm)	T_B (cm)
VCU corridor1	Naive	23.34	21.91	15.22	14.91
VCU corridor1	OA-mask	11.13	10.25	16.37	14.47
VCU corridor1	KISS-ICP	24.24	24.88	76.76	73.55
VCU light1	Naive	20.28	19.19	11.81	11.73
VCU light1	OA-mask	9.02	7.72	13.01	12.14
VCU light1	KISS-ICP	14.42	11.27	44.14	38.17
VCU motion1	Naive	24.55	24.14	14.21	13.90
VCU motion1	OA-mask	10.25	11.49	14.67	13.63
VCU motion1	KISS-ICP	19.68	19.57	71.99	69.17

TABLE XIII. FIXED-PREDICTION FIVE-SEED VCU ROTATION RMSE (DEGREES); THIS TABLE RETAINS ITS DOCUMENTED FIVE-SEED PROVENANCE UNDER REFERENCE CASES A/B. ALL THREE ROTATION INTERVALS FAVOR OA UNDER BOTH STATED HYPOTHESES; NEITHER HYPOTHESIS CERTIFIES A CAMERA/BODY/MARKER CSV SCHEMA.

Sequence	Case	Naive mean ± SD	OA mean ± SD	Δ [95% CI]
VCU light1	A	22.38 ± 6.11	8.48 ± 0.32	−13.90 [−21.61, −6.19]
VCU light1	B	21.51 ± 6.83	7.14 ± 0.41	−14.37 [−22.98, −5.76]
VCU corridor1	A	22.27 ± 4.85	10.87 ± 0.19	−11.39 [−17.52, −5.27]
VCU corridor1	B	21.03 ± 5.73	9.71 ± 0.58	−11.32 [−18.23, −4.42]
VCU motion1	A	26.36 ± 5.42	9.45 ± 0.57	−16.92 [−24.13, −9.70]
VCU motion1	B	25.77 ± 6.27	10.22 ± 0.75	−15.55 [−23.64, −7.46]

4.8 Error distributions and zero-shot transfer

Fig. 3 illustrates common-reference-set error distributions for textureless structure, static-camera/moving-foreground, handheld motion and camera shake. Figs. 4 and 5 retain all nine sequences and their full tails. A symmetric-log horizontal scale allows small errors and large tails to remain visible. The walking distribution confirms that the teacher can outperform OA; universal distributional dominance is not supported.

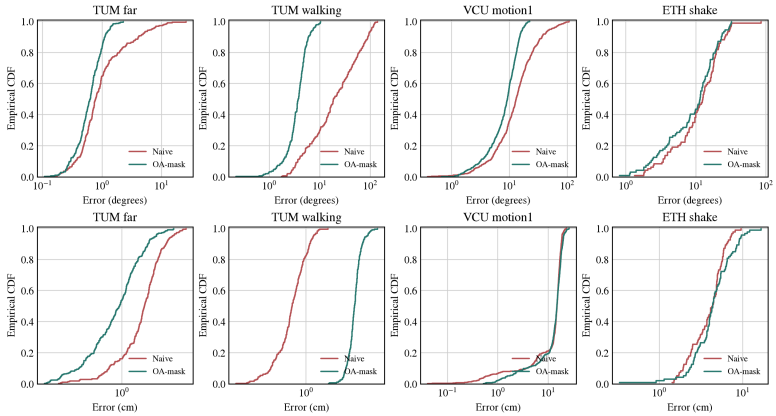


Fig. 3. Empirical error CDFs for four scene types. Rotation is in degrees and translation in centimetres. The symmetric-log horizontal scale has a linear threshold of 0.1. All valid reference pairs and full tails are retained. Student predictions use the current seed-zero comparator cohort; teacher values use the specified pipeline. VCU reference interpretation is conditional.

Table XIV reports complete rotation and translation RMSE/MAE for ETH3D transfer of the current seed-zero comparator cohort. There is no ETH3D fine-tuning. Rotation and translation need not improve together, and the four transfer sequences do not establish generalization beyond the tested sensor and conditions.

TABLE XIV. REEXECUTED SEED-ZERO ETH3D TRANSFER USING THE SAME EVALUATOR AND RETAINED INPUT ARRAYS. NO ETH3D TRAINING OR FINE-TUNING IS PERFORMED.

Sequence	Method	R_{RMSE} (°)	R_{MAE} (°)	T_{RMSE} (cm)	T_{MAE} (cm)
ETH Einstein	Teacher	31.370	17.600	8.05	5.90
ETH Einstein	Naive	11.574	5.722	3.97	3.57
ETH Einstein	Gate	3.559	3.183	4.95	4.56
ETH Einstein	OA-mask	3.401	3.078	5.10	4.48
ETH Einstein	OA-soft	3.275	2.867	4.94	4.34
ETH Einstein	GT reference	2.509	2.191	4.93	4.41
ETH kidnap	Teacher	27.777	9.801	6.36	5.54
ETH kidnap	Naive	7.165	5.373	5.53	5.27
ETH kidnap	Gate	4.282	3.910	5.39	5.05
ETH kidnap	OA-mask	3.867	3.420	5.81	5.47
ETH kidnap	OA-soft	3.953	3.553	7.63	6.76
ETH kidnap	GT reference	3.741	3.423	6.53	6.11
ETH ceiling	Teacher	43.672	27.304	8.23	7.63
ETH ceiling	Naive	9.726	6.524	7.56	7.21
ETH ceiling	Gate	3.723	3.395	6.34	6.10
ETH ceiling	OA-mask	4.616	4.288	7.18	6.77
ETH ceiling	OA-soft	4.771	4.410	8.52	7.91
ETH ceiling	GT reference	4.932	4.463	7.18	6.75
ETH shake	Teacher	57.927	37.545	6.00	4.91
ETH shake	Naive	17.553	13.864	4.54	4.20
ETH shake	Gate	14.296	12.358	5.81	5.32
ETH shake	OA-mask	14.213	11.792	5.61	4.95
ETH shake	OA-soft	14.659	12.375	5.59	5.02
ETH shake	GT reference	14.146	12.007	6.51	5.87

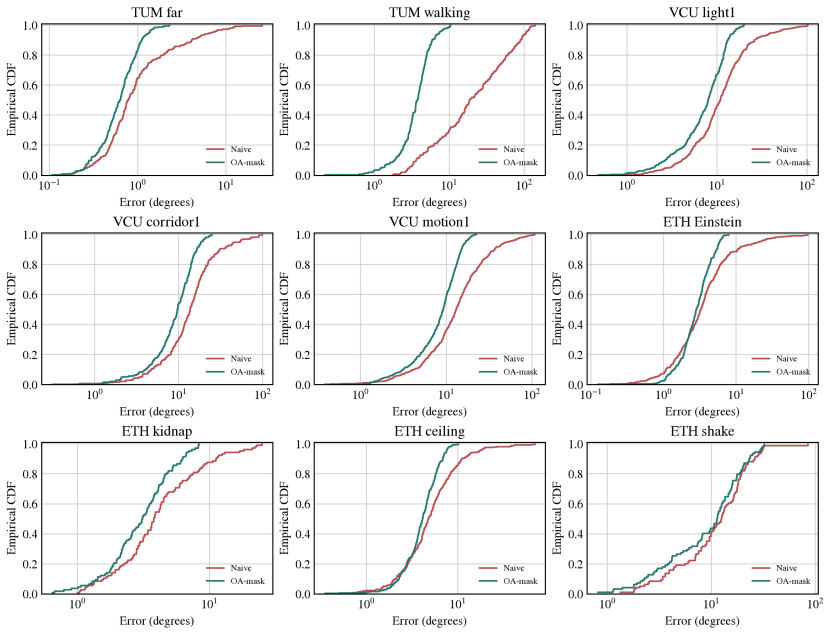


Fig. 4. Rotation CDFs for all nine evaluation sequences, using every valid reference pair and retaining the complete bounded-angle range.

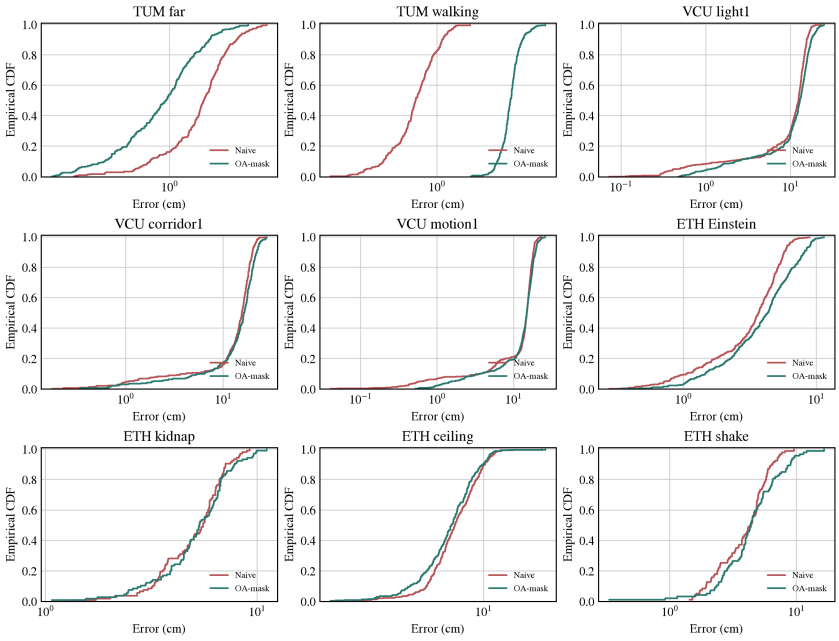


Fig. 5. Translation CDFs for all nine evaluation sequences, in centimetres. The symmetric-log scale retains the full error range.

4.9 Mask, backbone and threshold ablations

Tables XV, XVI and XVII report both error statistics for both pose components. Gate-only is not uniformly inferior to OA-mask: it has slightly lower rotation RMSE on VCU light1, corridor1 and motion1, and some lower translation errors. Its much smaller retained set and update budget prevent attributing these differences to loss filtering alone. CNN and ViT outcomes also vary by sequence and component; the current seed-zero backbone study does not establish a statistically reliable architecture ordering.

TABLE XV. REEXECUTED SEED-ZERO MASK/COMPARATOR COMPARISON; GATE RETAINS ITS DOCUMENTED DISTINCT BUDGET.

Sequence	Variant	R_{RMSE} (°)	R_{MAE} (°)	T_{RMSE} (cm)	T_{MAE} (cm)
TUM far	Naive	3.16	1.68	1.75	1.62
TUM far	Gate	1.49	1.38	1.65	1.54
TUM far	OA-mask	0.76	0.68	1.13	1.00
TUM far	OA-soft	1.09	0.96	1.85	1.74
TUM walking	Naive	46.93	33.46	0.76	0.70
TUM walking	Gate	4.64	4.51	2.83	2.75
TUM walking	OA-mask	4.28	3.89	4.28	4.19
TUM walking	OA-soft	3.91	3.54	3.03	2.80
VCU light1	Naive	20.28	14.39	11.81	10.80
VCU light1	Gate	8.46	7.90	12.13	11.40
VCU light1	OA-mask	9.02	8.06	13.01	11.87
VCU light1	OA-soft	8.59	7.60	12.82	11.79
VCU corridor1	Naive	23.34	17.37	15.22	14.17
VCU corridor1	Gate	10.57	9.74	15.68	14.91
VCU corridor1	OA-mask	11.13	10.03	16.37	15.23
VCU corridor1	OA-soft	10.84	9.59	16.86	15.70
VCU motion1	Naive	24.55	17.88	14.21	13.13
VCU motion1	Gate	8.89	8.27	14.75	13.81
VCU motion1	OA-mask	10.25	9.22	14.67	13.63
VCU motion1	OA-soft	10.04	8.87	15.41	14.26

TABLE XVI. REEXECUTED SEED-ZERO OA BACKBONE COMPARISON; NO ARCHITECTURE IS SELECTED FROM HELD-OUT SCORES.

Sequence	Variant	R_{RMSE} (°)	R_{MAE} (°)	T_{RMSE} (cm)	T_{MAE} (cm)
TUM far	OA CNN	0.86	0.74	1.79	1.66
TUM far	OA-mask	0.76	0.68	1.13	1.00
TUM far	OA ViT	0.89	0.79	1.10	0.99
TUM walking	OA CNN	6.13	5.44	6.12	5.23
TUM walking	OA-mask	4.28	3.89	4.28	4.19
TUM walking	OA ViT	0.66	0.60	3.60	3.58
VCU light1	OA CNN	8.32	7.49	12.52	11.57
VCU light1	OA-mask	9.02	8.06	13.01	11.87
VCU light1	OA ViT	7.05	6.20	12.03	11.20
VCU corridor1	OA CNN	10.81	9.55	15.95	14.92
VCU corridor1	OA-mask	11.13	10.03	16.37	15.23
VCU corridor1	OA ViT	9.35	8.35	15.41	14.50
VCU motion1	OA CNN	9.36	8.29	14.25	13.28
VCU motion1	OA-mask	10.25	9.22	14.67	13.63
VCU motion1	OA ViT	7.56	6.79	14.58	13.62

TABLE XVII. REEXECUTED SEED-ZERO FIXED-THRESHOLD SENSITIVITY ON THE FIVE MAIN SEQUENCES; NO THRESHOLD IS SELECTED FROM THESE VALUES.

Sequence	Variant	R_{RMSE} (°)	R_{MAE} (°)	T_{RMSE} (cm)	T_{MAE} (cm)
TUM far	$\sigma = 0.02$	0.93	0.86	1.82	1.69
TUM far	$\sigma = 0.05$	1.01	0.89	1.14	1.03
TUM far	default 0.08	0.76	0.68	1.13	1.00
TUM far	$\sigma = 0.15$	0.78	0.70	1.83	1.71
TUM far	$\sigma = 0.30$	0.96	0.85	1.75	1.63
TUM walking	$\sigma = 0.02$	3.60	3.27	3.12	3.03
TUM walking	$\sigma = 0.05$	3.74	3.23	2.25	2.14
TUM walking	default 0.08	4.28	3.89	4.28	4.19
TUM walking	$\sigma = 0.15$	3.41	3.09	5.55	4.97
TUM walking	$\sigma = 0.30$	3.62	3.30	2.87	2.46
VCU light1	$\sigma = 0.02$	8.67	7.77	12.63	11.59
VCU light1	$\sigma = 0.05$	8.57	7.63	12.44	11.44
VCU light1	default 0.08	9.02	8.06	13.01	11.87
VCU light1	$\sigma = 0.15$	8.56	7.65	12.89	11.79
VCU light1	$\sigma = 0.30$	8.58	7.75	12.41	11.45
VCU corridor1	$\sigma = 0.02$	11.12	9.96	16.31	15.22
VCU corridor1	$\sigma = 0.05$	10.93	9.70	15.71	14.64
VCU corridor1	default 0.08	11.13	10.03	16.37	15.23
VCU corridor1	$\sigma = 0.15$	10.89	9.69	16.30	15.17
VCU corridor1	$\sigma = 0.30$	11.02	9.85	16.46	15.41
VCU motion1	$\sigma = 0.02$	10.33	9.17	14.36	13.33
VCU motion1	$\sigma = 0.05$	9.66	8.68	13.86	12.85
VCU motion1	default 0.08	10.25	9.22	14.67	13.63
VCU motion1	$\sigma = 0.15$	9.41	8.47	15.22	14.01
VCU motion1	$\sigma = 0.30$	9.50	8.52	13.88	12.95

The independent Cabinet-held-out diagnostic is reported in Table XVIII and Fig. 6. Threshold changes affect only the loss mask, with fixed projected teacher labels and rotation flags. The completed experiment uses 25 runs, five seeds per threshold, and shows sensitivity to this choice without selecting the primary default or isolating a new source of geometric information.

TABLE XVIII. HELD-OUT CABINET SENSITIVITY ACROSS 25 COMPLETED RUNS (FIVE SEEDS PER FIXED THRESHOLD), REPORTED AS MEAN ± SD. THIS IS EXPLORATORY AND DOES NOT SELECT AN OPTIMUM.

σ_{loss}	R_{RMSE} (°)	R_{MAE} (°)	T_{RMSE} (cm)	T_{MAE} (cm)
0.02	2.161 ± 0.204	1.775 ± 0.138	3.554 ± 1.277	3.365 ± 1.271
0.05	2.083 ± 0.276	1.702 ± 0.199	3.432 ± 1.322	3.233 ± 1.282
0.08	2.384 ± 0.579	1.882 ± 0.295	3.221 ± 0.418	2.997 ± 0.370
0.15	2.327 ± 0.303	1.900 ± 0.207	4.136 ± 0.883	3.882 ± 0.892
0.30	2.127 ± 0.213	1.751 ± 0.206	3.922 ± 1.534	3.637 ± 1.504

For comparison with the completed five-seed means, the previously documented single-seed Cabinet values are retained in Table XIX as an explicitly old diagnostic. They are not substituted for the current reruns. Paired contrasts against the fixed 0.08 default are summarized in Table XX; all reported exploratory intervals span zero.

TABLE XIX. PRESERVED ORIGINAL DOCUMENTED ONE-SEED CABINET DIAGNOSTIC. IT IS RETAINED AS AN OLD DIAGNOSTIC EVEN WHEN THE NEW FIVE-SEED MEANS DIFFER.

σ_{loss}	Row provenance	R_{RMSE} (°)	R_{MAE} (°)	T_{RMSE} (cm)	T_{MAE} (cm)
0.02	archived old seed 0	2.405	1.825	2.403	2.217
0.05	archived old seed 0	2.068	1.665	2.894	2.672
0.08	archived old seed 0	3.365	2.294	3.945	3.637
0.15	archived old seed 0	2.313	1.754	3.499	3.205
0.30	archived old seed 0	2.272	1.684	2.984	2.636

TABLE XX. PAIRED SEED CONTRASTS FOR EACH ALTERNATIVE CABINET THRESHOLD AGAINST FIXED 0.08; CONTRASTS ARE DESCRIPTIVE AND NO OPTIMUM IS SELECTED. TRANSLATION CONTRASTS ARE REPORTED IN CM.

Compared τ	Metric	Other − default mean ± SD	95% paired CI
0.02	rotation rmse deg	−0.223 ± 0.487	−0.223 [−0.828, 0.382]
0.02	rotation mae deg	−0.107 ± 0.330	−0.107 [−0.516, 0.303]
0.02	translation rmse cm	0.333 ± 1.557	0.333 [−1.600, 2.266]
0.02	translation mae cm	0.368 ± 1.510	0.368 [−1.507, 2.242]
0.05	rotation rmse deg	−0.301 ± 0.663	−0.301 [−1.124, 0.522]
0.05	rotation mae deg	−0.180 ± 0.372	−0.180 [−0.642, 0.283]
0.05	translation rmse cm	0.211 ± 1.382	0.211 [−1.505, 1.927]
0.05	translation mae cm	0.236 ± 1.336	0.236 [−1.423, 1.895]
0.15	rotation rmse deg	−0.057 ± 0.736	−0.057 [−0.971, 0.857]
0.15	rotation mae deg	0.018 ± 0.478	0.018 [−0.575, 0.611]
0.15	translation rmse cm	0.915 ± 1.152	0.915 [−0.515, 2.346]
0.15	translation mae cm	0.885 ± 1.111	0.885 [−0.494, 2.265]
0.30	rotation rmse deg	−0.257 ± 0.593	−0.257 [−0.993, 0.479]
0.30	rotation mae deg	−0.131 ± 0.460	−0.131 [−0.702, 0.440]
0.30	translation rmse cm	0.701 ± 1.736	0.701 [−1.454, 2.856]
0.30	translation mae cm	0.640 ± 1.666	0.640 [−1.428, 2.709]

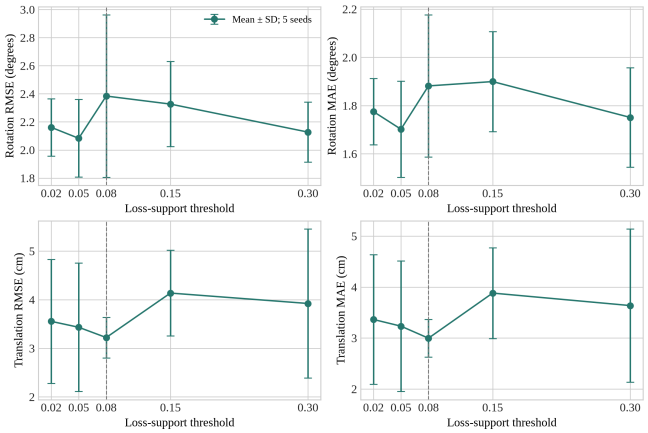


Fig. 6. Cabinet-held-out loss-mask sensitivity across five seeds per threshold. Error bars are sample standard deviations. Cabinet is excluded from the corresponding training pool; teacher labels and rotation flags remain fixed. The dashed line identifies the fixed primary threshold, not a test-selected optimum.

Interpretation of the Cabinet-held-out result. Across the completed five-seed means, the fixed 0.08 setting ranks fifth of five for rotation RMSE (2.384 ± 0.579°) and fourth of five for rotation MAE (1.882 ± 0.295°), but first of five for translation RMSE (3.221 ± 0.418 cm) and translation MAE (2.997 ± 0.370 cm). The lowest rotation means occur at 0.05 for both RMSE and MAE; the lowest translation means occur at 0.08. These descriptive means and the paired contrasts do not select an optimum or justify changing the standard threshold.

A direct audit of the 1,174 retained non-Cabinet training spectra gives mean active translation ranks of 1.6405, 1.5886 and 1.5196 at loss thresholds 0.05, 0.08 and 0.15, respectively. Cabinet's corresponding mean ranks are 1.9705, 1.9598 and 1.9437 across its 373 stored pairs. The 0.08 mask is intermediate, not a uniquely extreme rank configuration. No audited eigenvalue ratio lies on the 0.08 boundary within 10^{-12}, so the strict generation and inclusive loss masks agree there. The differing metric patterns are not explained by the rank count alone. Changing the translation loss mask changes the gradient entering the shared feature extractor and can therefore affect rotation predictions despite fixed rotation targets and flags. This coupling makes non-monotone component errors possible; it does not identify the cause of the observed 0.08 spike. Threshold-dependent optimization and the held-out distribution are not separated by this diagnostic. Keeping 0.08 as a fixed baseline permits a controlled comparison; it does not assert optimality or justify choosing another threshold from these outcomes. The sweep also cannot isolate a generation-threshold effect because projected teacher labels are fixed.

Fig. 7 summarizes the optimization histories of the five-seed comparison. Because the rotation term is gated and averaged over the entire batch, loss magnitudes are not directly comparable as unmasked pose-error measures.

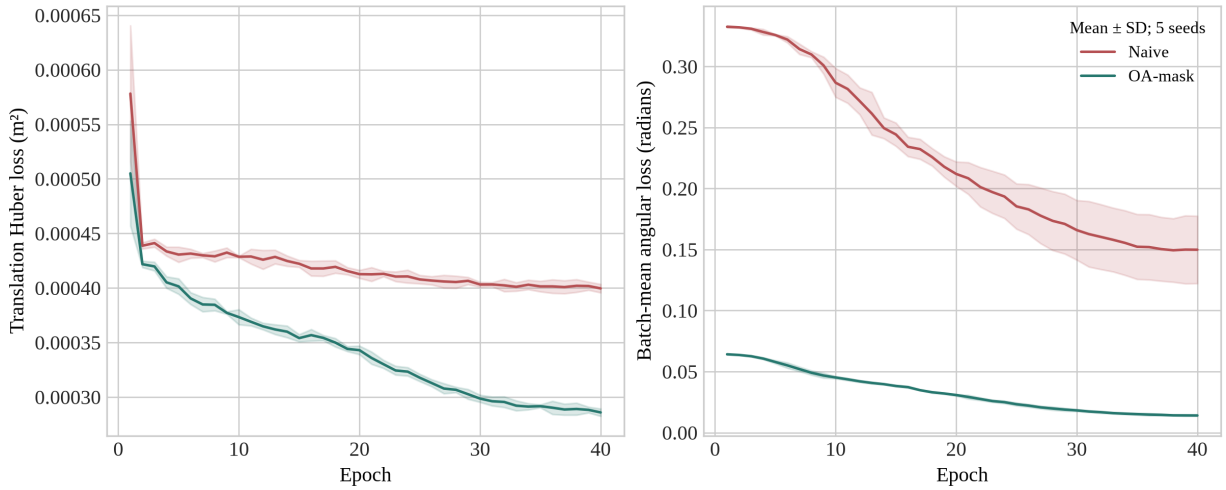


Fig. 7. Mean and standard deviation of the 40-epoch loss histories over five seeds. Translation Huber and angular loss have distinct units. Rotation gating changes the effective contribution of active targets, so a lower training objective alone does not imply greater pose accuracy.

Table XXI gives five-seed loss-component controls under the same 1,547-pair pool, 40 epochs, batch size 32, 1,920 updates and paired initialization protocol. *R gate only* retains angular target filtering but uses $W_t = I$; *T projector only* retains $W_t = P_{\text{obs}}$ but gives every rotation target unit weight. Teacher labels and the generation projector remain unchanged. These exploratory controls distinguish loss mechanisms without replacing the five-seed full OA-mask comparison.

TABLE XXI. FIVE-SEED COMPONENT CONTROLS AND THE SAME PAIRED-SEED NAIVE/FULL-OA REFERENCES. CONTROLS ARE EXPLORATORY; VCU ROWS REMAIN CONDITIONAL.

Sequence	Variant	R_{RMSE} mean ± SD	R_{MAE} mean ± SD	T_{RMSE} mean ± SD (cm)	T_{MAE} mean ± SD (cm)
TUM far	Naive	2.896 ± 1.732	1.533 ± 0.720	1.559 ± 0.185	1.443 ± 0.170
TUM far	Full OA-mask	0.865 ± 0.116	0.782 ± 0.111	1.576 ± 0.407	1.442 ± 0.392
TUM far	R gate only	1.055 ± 0.349	0.948 ± 0.289	1.937 ± 0.146	1.826 ± 0.145
TUM far	T projector only	2.585 ± 1.736	1.368 ± 0.620	1.353 ± 0.189	1.178 ± 0.164
TUM walking	Naive	36.459 ± 10.134	26.726 ± 6.642	1.005 ± 0.199	0.925 ± 0.191
TUM walking	Full OA-mask	4.030 ± 1.080	3.615 ± 1.016	4.208 ± 1.014	3.912 ± 0.808
TUM walking	R gate only	3.890 ± 1.419	3.578 ± 1.351	1.602 ± 0.486	1.508 ± 0.459
TUM walking	T projector only	42.201 ± 10.671	30.711 ± 8.466	4.992 ± 1.700	4.627 ± 1.602
VCU light1	Naive	22.382 ± 6.113	15.793 ± 3.628	12.034 ± 0.179	11.017 ± 0.169

Sequence	Variant	R_{RMSE} mean ± SD	R_{MAE} mean ± SD	T_{RMSE} mean ± SD (cm)	T_{MAE} mean ± SD (cm)
VCU light1	Full OA-mask	8.481 ± 0.324	7.559 ± 0.317	12.980 ± 0.310	11.898 ± 0.254
VCU light1	R gate only	8.568 ± 0.254	7.676 ± 0.176	12.531 ± 0.192	11.451 ± 0.167
VCU light1	T projector only	22.959 ± 5.298	15.808 ± 2.988	12.568 ± 0.352	11.561 ± 0.337
VCU corridor1	Naive	22.267 ± 4.852	16.860 ± 3.243	15.386 ± 0.199	14.332 ± 0.195
VCU corridor1	Full OA-mask	10.873 ± 0.188	9.734 ± 0.201	16.282 ± 0.315	15.185 ± 0.313
VCU corridor1	R gate only	10.870 ± 0.209	9.748 ± 0.147	15.850 ± 0.131	14.757 ± 0.118
VCU corridor1	T projector only	22.474 ± 3.892	16.717 ± 2.185	15.672 ± 0.274	14.681 ± 0.276
VCU motion1	Naive	26.364 ± 5.416	18.871 ± 3.426	14.496 ± 0.190	13.393 ± 0.176
VCU motion1	Full OA-mask	9.448 ± 0.566	8.432 ± 0.526	15.198 ± 0.619	14.078 ± 0.499
VCU motion1	R gate only	9.846 ± 0.825	8.791 ± 0.602	15.320 ± 0.185	14.116 ± 0.169
VCU motion1	T projector only	27.755 ± 7.262	19.061 ± 4.319	14.454 ± 0.506	13.432 ± 0.454
ETH Einstein	Naive	13.566 ± 4.157	7.555 ± 2.300	4.158 ± 0.309	3.709 ± 0.248
ETH Einstein	Full OA-mask	3.445 ± 0.188	3.024 ± 0.177	5.338 ± 0.602	4.705 ± 0.517
ETH Einstein	R gate only	3.513 ± 0.321	3.071 ± 0.193	3.992 ± 0.110	3.603 ± 0.120
ETH Einstein	T projector only	13.350 ± 6.573	7.748 ± 3.046	5.749 ± 0.964	5.166 ± 0.940
ETH kidnap	Naive	8.571 ± 4.431	6.052 ± 1.752	5.561 ± 0.134	5.288 ± 0.103
ETH kidnap	Full OA-mask	3.928 ± 0.089	3.494 ± 0.149	6.857 ± 0.825	6.245 ± 0.639
ETH kidnap	R gate only	4.075 ± 0.412	3.636 ± 0.356	5.427 ± 0.159	5.199 ± 0.129
ETH kidnap	T projector only	11.905 ± 5.798	7.235 ± 2.488	7.090 ± 1.236	6.530 ± 1.038
ETH ceiling	Naive	13.873 ± 3.889	8.119 ± 2.219	7.285 ± 0.348	6.943 ± 0.322
ETH ceiling	Full OA-mask	4.736 ± 0.375	4.389 ± 0.416	7.769 ± 0.878	7.294 ± 0.932
ETH ceiling	R gate only	5.046 ± 0.493	4.714 ± 0.492	7.322 ± 0.751	7.002 ± 0.737
ETH ceiling	T projector only	12.159 ± 4.358	7.536 ± 1.240	8.173 ± 1.497	7.704 ± 1.505
ETH shake	Naive	16.693 ± 2.439	13.208 ± 1.063	4.640 ± 0.161	4.305 ± 0.149
ETH shake	Full OA-mask	14.367 ± 0.258	11.931 ± 0.326	5.244 ± 0.361	4.703 ± 0.269
ETH shake	R gate only	14.479 ± 0.235	12.087 ± 0.260	4.737 ± 0.151	4.359 ± 0.120
ETH shake	T projector only	16.796 ± 1.776	13.565 ± 0.873	5.405 ± 0.334	4.830 ± 0.252

4.10 Temporal behavior

Temporal diagnostics compose nonoverlapping windows of 10, 30 and 60 consecutive pose increments within contiguous reference-valid segments. Each window starts at a common local origin. Predicted and reference endpoints are compared without trajectory alignment, and no chain crosses missing-reference intervals. Increments are not assumed uniformly timed. Table XXII reports mean seed RMSE for 30-increment endpoints; Fig. 8 compares window lengths.

TABLE XXII. MEAN ACROSS FIVE SEEDS OF UNALIGNED ENDPOINT-ERROR RMSE OVER NONOVERLAPPING 30-INCREMENT WINDOWS. NO CHAIN CROSSES AN INVALID GT ROW; MISSING GT IS NOT INTERPOLATED. INCREMENTS ARE NOT ASSUMED UNIFORMLY TIMED. THESE ARE BOUNDED-WINDOW DIAGNOSTICS, NOT A STABILITY CERTIFICATE.

Sequence	Windows	R_N (°)	R_O (°)	T_N (cm)	T_O (cm)
TUM far	8	41.16	15.88	42.29	39.16
TUM walking	7	112.05	73.25	13.49	99.48
VCU light1	15	140.91	147.16	220.50	194.86
VCU corridor1	8	142.02	132.06	234.19	215.05
VCU motion1	17	139.81	143.47	216.36	204.30
ETH Einstein	8	105.87	59.44	90.96	115.58
ETH kidnap	1	48.83	26.25	110.15	119.91
ETH ceiling	8	118.25	93.19	197.74	207.02
ETH shake	3	100.79	95.99	37.61	47.51

Accumulated errors can be large, and lower per-pair rotation error does not ensure lower endpoint error on every sequence. For example, the 30-increment OA rotation endpoint error exceeds Naive on light1 and motion1. OA translation accumulation is also unfavorable in several cases. Corridor1 has 278 reference-valid pairs among 998 stored pairs, so these diagnostics do not cover its untracked interior or establish a global trajectory bound. They describe bounded-window behavior, not control-theoretic stability.

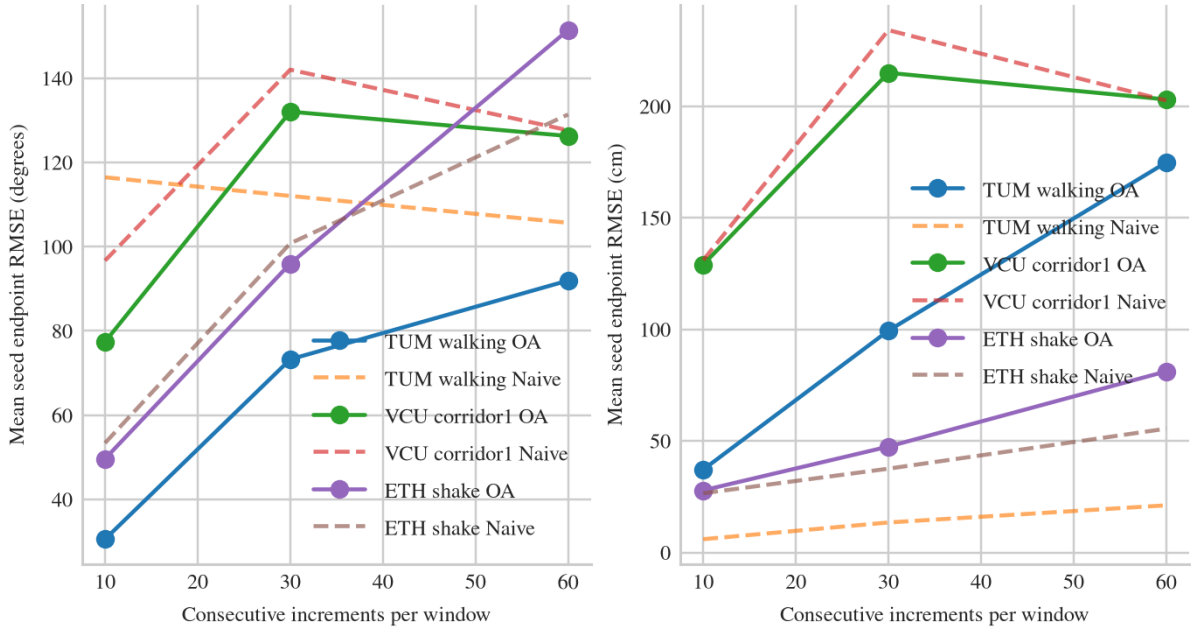


Fig. 8. Unaligned endpoint RMSE versus window length for walking, corridor and camera-shake cases. Values are means across five seeds of per-seed window RMSE, not independent-window confidence estimates. Missing-reference intervals are not bridged.

4.10.1. Full maximal contiguous trajectories

The five-seed predictions were additionally composed over every maximal contiguous reference-valid segment. A chain terminates at an invalid reference row, non-adjacent pair indices or a timestamp discontinuity; no reference gap is interpolated or bridged. Prediction and reference start from the same local identity pose, without subsequent trajectory alignment. For K increments, endpoint errors compare the final composed poses, and trajectory RMS averages position/orientation errors over the $K + 1$ composed poses. Tables XXIII and XXIV give the longest-segment extent and mean/SD across training seeds. Segment selection depends only on available reference duration, not model performance. The longest verified-camera physical horizon is TUM far: 27.280 s, 264 increments and 4.530 m. VCU motion1 reaches 105.196 s, 526 increments and 68.272 m only under the conditional direct-reference convention. OA endpoint translation error is 2.402 m versus 1.453 m for Naive, while endpoint rotation error is 134.28 versus 149.63 degrees. Across the nine longest segments, OA has higher mean endpoint translation error on all nine and higher trajectory translation RMS on eight; endpoint rotation is higher on six. On verified-frame TUM walking (24.768 s), the identity comparator has endpoint errors of 0.096 m and 9.38 degrees, versus 3.774 m and 106.06 degrees for OA. These results expose accumulation and reversals that per-pair rotation RMSE alone does not capture. They are finite-trajectory measurements, not evidence of bounded long-horizon error. The 720 reference-invalid corridor rows remain untested, and no statement extends beyond the available horizons.

TABLE XXIII. EXTENT OF THE LONGEST CONTIGUOUS REFERENCE-VALID SEGMENT PER SEQUENCE. N_{seg} COUNTS ALL AVAILABLE MAXIMAL SEGMENTS, K IS THE INCREMENT COUNT AND PATH LENGTH USES THE SUPPLIED REFERENCE. VCU HORIZONS ARE CONDITIONAL ON THE DIRECT-REFERENCE CONVENTION. NO MISSING-REFERENCE INTERVAL IS CROSSED.

Sequence	N_{seg}	Duration (s)	K	Reference path (m)
TUM far	1	27.280	264	4.530
TUM walking	1	24.768	239	0.435
VCU light1	1	95.382	477	51.431
VCU corridor1	2	27.970	140	19.024
VCU motion1	1	105.196	526	68.272
ETH Einstein	9	9.843	89	3.698
ETH kidnap	5	3.428	31	1.507
ETH ceiling	4	14.303	129	9.805
ETH shake	1	10.653	94	3.816

TABLE XXIV. UNALIGNED ACCUMULATED ERRORS ON THE LONGEST SEGMENT SPECIFIED IN TABLE XXIII. T_e, R_e ARE ENDPOINT ERRORS; $T_{\text{traj}}, R_{\text{traj}}$ ARE TRAJECTORY RMS ERRORS FROM A COMMON LOCAL ORIGIN. VALUES ARE MEAN ± SD OVER FIVE TRAINING SEEDS, NOT INDEPENDENT-FRAME INTERVALS. IDENTITY IS A DETERMINISTIC STATIC-CAMERA COMPARATOR. VCU VALUES ARE FRAME-CONDITIONAL.

Sequence	Method	T_e (m)	R_e (°)	T_{traj} (m)	R_{traj} (°)
TUM far	Naive	3.109 ± 0.551	86.51 ± 37.37	2.100 ± 0.297	63.80 ± 13.08
TUM far	OA-mask	3.265 ± 1.121	99.04 ± 41.12	2.109 ± 0.803	60.60 ± 24.30
TUM walking	Naive	0.564 ± 0.437	126.29 ± 60.95	0.405 ± 0.271	120.66 ± 9.22
TUM walking	OA-mask	3.774 ± 0.810	106.06 ± 43.46	2.630 ± 0.460	110.55 ± 4.40
TUM walking	Identity	0.096	9.38	0.073	7.04
VCU light1	Naive	0.753 ± 0.561	116.32 ± 42.72	1.853 ± 0.214	123.99 ± 5.51
VCU light1	OA-mask	3.296 ± 1.736	131.24 ± 34.77	3.158 ± 0.535	126.02 ± 7.40
VCU corridor1	Naive	4.124 ± 0.497	117.88 ± 35.14	1.971 ± 0.210	124.74 ± 3.88
VCU corridor1	OA-mask	4.142 ± 0.450	149.67 ± 20.70	1.890 ± 0.166	127.25 ± 4.47
VCU motion1	Naive	1.453 ± 0.900	149.63 ± 29.12	2.024 ± 0.288	128.49 ± 3.37
VCU motion1	OA-mask	2.402 ± 0.703	134.28 ± 28.82	2.460 ± 0.420	127.36 ± 3.26
ETH Einstein	Naive	1.825 ± 0.315	98.69 ± 24.56	0.918 ± 0.364	105.71 ± 21.68
ETH Einstein	OA-mask	3.099 ± 0.519	141.20 ± 30.67	1.767 ± 0.432	104.35 ± 19.44
ETH kidnap	Naive	1.082 ± 0.029	49.73 ± 25.33	0.758 ± 0.021	33.45 ± 15.37
ETH kidnap	OA-mask	1.170 ± 0.147	25.64 ± 23.48	0.853 ± 0.062	18.56 ± 14.03
ETH ceiling	Naive	0.674 ± 0.985	133.63 ± 53.32	2.950 ± 0.416	130.31 ± 12.14
ETH ceiling	OA-mask	2.214 ± 0.770	150.94 ± 23.20	3.306 ± 0.622	114.31 ± 33.76
ETH shake	Naive	0.454 ± 0.065	131.34 ± 46.92	0.439 ± 0.046	107.38 ± 40.55
ETH shake	OA-mask	0.570 ± 0.257	134.25 ± 26.38	0.653 ± 0.090	110.97 ± 16.85

4.10.2. Full ETH3D recording extension

A separate finite-recording extension evaluates the complete fixed input prefixes for the four ETH3D sequences, without changing the original 5,197-pair primary accounting. The prefixes contain 679 frames/678 pairs/543 reference-valid pairs for Einstein-dark, 176/175/106 for kidnap-dark, 532/531/384 for ceiling-1, and 95/94/94 for camera-shake-3. The longest physical horizons are 10.727, 3.428, 14.303 and 10.653 seconds, respectively; none exceeds the 27.280-second TUM far physical headline. Endpoint translation is adverse for OA on all four sequences: the mean OA versus Naive values are 3.744 versus 3.123 m, 1.170 versus 1.082 m, 2.214 versus 0.674 m, and 0.570 versus 0.454 m, respectively. Endpoint rotation is mixed: OA is higher on Einstein-dark (124.53 versus 108.18 degrees), ceiling-1 (150.94 versus 133.63 degrees) and camera-shake-3 (134.25 versus 131.34 degrees), but lower on kidnap-dark (25.64 versus 49.73 degrees). These finite full-recording chains are post hoc trajectory diagnostics over fixed valid segments, not evidence of global stability.

TABLE XXV. FULL-RECORDING ETH3D FINITE-HORIZON DIAGNOSTICS ON THE LONGEST CONTIGUOUS REFERENCE-VALID SEGMENT OF EACH SEQUENCE. MEAN ± SAMPLE SD ACROSS FIVE FIXED SEEDS; TRANSLATION IS IN METRES AND ROTATION IN DEGREES. THE FIXED STRIDE-3 EXTENSION PRESERVES THE PRIMARY INPUT PREFIX AND NEVER CHAINS ACROSS MISSING REFERENCES.

Sequence	Model	Duration (s)	Path (m)	End T (m)	End R (°)	RMS (m)
einstein_dark	Naive	10.727	4.392	3.123 ± 0.386	108.18 ± 55.55	2.159 ± 0.177
einstein_dark	OA-mask	10.727	4.392	3.744 ± 1.106	124.53 ± 20.17	2.528 ± 0.694
kidnap_dark	Naive	3.428	1.507	1.082 ± 0.029	49.73 ± 25.33	0.758 ± 0.021
kidnap_dark	OA-mask	3.428	1.507	1.170 ± 0.147	25.64 ± 23.48	0.853 ± 0.062
ceiling_1	Naive	14.303	9.805	0.674 ± 0.985	133.63 ± 53.32	2.950 ± 0.416
ceiling_1	OA-mask	14.303	9.805	2.214 ± 0.770	150.94 ± 23.20	3.306 ± 0.622
camera_shake_3	Naive	10.653	3.816	0.454 ± 0.065	131.34 ± 46.92	0.439 ± 0.046
camera_shake_3	OA-mask	10.653	3.816	0.570 ± 0.257	134.25 ± 26.38	0.653 ± 0.090

4.11 Computational requirements

Table XXVI reports batch-one FP32 CPU timing with three Torch threads on an Intel(R) Xeon(R) Processor at 2.50 GHz. Each timing series uses 30 warmups and 200 repeats. The hybrid forward pass averages 5.249 ± 0.888 ms, with a 95th percentile of 7.067 ms. Input assembly and rotation conversion are measured in a separate combined series. The combined-series mean is not an arithmetic increment to the forward-only mean; independent timing variability explains why

the hybrid combined mean is slightly smaller. Sensor acquisition, decoding, disk I/O and teacher processing are excluded.

TABLE XXVI. BATCH-ONE FP32 CPU PROFILING: 30 WARMUPS, 200 REPEATS AND THREE TORCH THREADS. PREP+POSE MEASURES CACHED-INPUT ASSEMBLY, NETWORK INFERENCE AND ROTATION CONVERSION IN AN INDEPENDENT TIMING SERIES. IT IS NOT THE FORWARD MEAN PLUS A SEPARATELY ADDED OVERHEAD. ACQUISITION, DECODING, DISK I/O AND TEACHER PROCESSING ARE EXCLUDED.

Backbone	Parameters	Forward ms mean ± SD	p95 ms	Prep+pose ms
hybrid	620,169	5.249±0.888	7.067	5.083
cnn	344,969	3.476±0.442	4.256	5.243
vit	900,617	8.583±1.159	10.956	9.410

Offline teacher geometry is profiled separately on twelve sampled real frames each from TUM far and VCU light1, excluding disk I/O, reference synchronization and network inference. The measured costs are given in Table XXVII. These small-sample measurements do not establish full-sequence throughput. The student forward cost alone does not establish complete sensor-to-pose real-time performance.

TABLE XXVII. SMALL-SAMPLE OFFLINE TEACHER PROFILING ON TWELVE REAL SAMPLED FRAMES PER SEQUENCE. EXTRACTION IS PER FRAME; MATCHING/SUPPORT/ALIGNMENT IS PER ADJACENT PAIR. THE LAST COLUMN COMBINES THE TWO MEANS AS A SEQUENTIAL ESTIMATE. DECODING, DISK I/O, REFERENCE SYNCHRONIZATION AND STUDENT INFERENCE ARE EXCLUDED.

Sequence	Frames	Extraction ms	Matching/solve ms	Sequential ms
TUM far	12	352.040	0.075	352.115
VCU light1	12	143.865	0.370	144.236

5. DISCUSSION AND LIMITATIONS

The documented paired training study identifies a component-specific error trade-off. OA-mask has lower sample-mean per-pair rotation RMSE and higher sample-mean translation RMSE on all nine sequences. Six rotation comparisons retain Holm-adjusted paired-t $p < 0.05$, while four have favorable simultaneous intervals. The exact five-seed sign-flip sensitivity does not resolve any comparison at 0.05. The physical-accuracy interpretation rests on TUM and ETH3D; VCU remains conditional on its supplied-reference convention. Current seed-zero comparator rows are descriptive completed evaluations, not additional five-seed replications.

This pattern is consistent with the supervision geometry, but does not by itself establish causality for every sequence. The angular gate can prevent training on arbitrary full rotations from rank-deficient normal alignment. Translation projection removes direct penalties in unsupported directions; the network may still learn a prior there through shared parameters and the image structure. Finite boundaries and nonplanar depth cues contain information omitted by the teacher's normal/offset residual model. Accordingly, its local rank analysis is not an information bound on the complete student input.

The completed five-seed component controls show a strong but nonexclusive role for rotation-target filtering. Applying the rotation gate alone reduces mean rotation RMSE relative to Naive on all nine evaluation sequences, including all six TUM and ETH3D sequences, whose camera reference frames are verified. In contrast, applying the translation projector alone reduces mean rotation RMSE relative to Naive on only three of the nine sequences, all of which are TUM or ETH3D sequences with verified camera reference frames. Full OA has lower mean rotation RMSE than the rotation-gate-only control on seven of the nine sequences; the rotation-gate-only control has lower mean rotation RMSE on the remaining two. This does not support a replicated claim that the rotation-only control is superior overall. The rotation-gate-only control has lower mean translation RMSE than Naive on only two of the nine sequences. Full OA has lower mean rotation RMSE than the translation-projector-only control on all nine sequences, and translation-only filtering is not a consistent improvement. These contrasts are exploratory and do not identify a unique causal mechanism.

Teacher quality remains a central limitation. A single plane leaves a family of rotations with identical normal-alignment cost, and moving foreground objects can invalidate static correspondence assumptions. The approximately static-camera walking sequence also favors the identity predictor and teacher over OA, directly contradicting a universal teacher-distillation advantage. Plane extraction, greedy matching, calibration and depth completeness can all produce misleading support indicators. The binary rotation gate additionally discards partially informative rotational constraints and changes the effective angular-loss contribution with the prevalence of active targets. Relative area weights and fixed thresholds are not calibrated uncertainty estimates.

Baseline interpretation must respect sensor and training differences. The approximately 3.39, 4.69 and 4.91 ratios concern adapted sparse-depth KISS-ICP translation errors on VCU light1, corridor1 and motion1 only. Gate-only has a smaller training budget; the supervised reference uses different admissible data and augmentation. Neither provides a controlled superiority test for every aspect of the method. Reference-frame ambiguity also changes VCU numerical errors: fixed-prediction sensitivity checks retain the rotation benefit under both stated hypotheses, but cannot certify a camera, body or marker schema from error values.

Full contiguous-trajectory diagnostics expose substantial accumulation and reversals of the per-pair advantage. OA has higher endpoint translation error on every longest segment and higher trajectory translation RMS on eight of nine. The longest verified-camera physical horizon is TUM far at 27.280 s; the 105.196-s VCU motion1 horizon remains conditional, and the missing corridor interior is not evaluated. The separate full-ETH3D extension in Section 4.10.2 remains finite-recording evidence and makes no global stability claim. Stronger seed-level inference improves the statistical accounting but retains five-seed distributional and resolution limits. The Cabinet diagnostic does not establish an optimum for the fixed 0.08 threshold; the mask-only design cannot identify a generation-threshold optimum.

The practical advantage is separation of expensive offline geometric supervision from a lightweight depth-only student. The measured forward latency supports low computational cost for the tested model and hardware, but complete real-time applicability requires acquisition, decoding and preprocessing to be profiled together. Stronger claims require larger replicated datasets, verified calibration/reference conventions and longer continuous trajectories.

6. CONCLUSION

Observability-aware rotation-target filtering suppresses angular supervision when the plane teacher lacks independent normal constraints, while translation residuals receive a separate support projection. The documented five-seed full OA-mask comparison shows lower mean per-pair rotation RMSE than Naive on all nine sequences, with six Holm-adjusted paired-t comparisons and four simultaneous intervals favoring OA. Translation RMSE is higher on all nine, and accumulated trajectory error is not uniformly reduced. Completed component controls show a strong filtering contribution but do not establish a unique causal mechanism or uniform superiority of the rotation-only control. The contribution is therefore targeted supervision of unsupported rotations, with an explicit translation and accumulation trade-off. Verified-frame TUM/ETH3D evidence and conditional VCU comparisons are distinguished.

Future work should retain partially informative rotational constraints, calibrate plane weights, improve dynamic-scene association and test joint generation/loss-threshold choices on disjoint validation data. Robustness studies should include failed extraction, missing depth and nonplanar scenes. Longer continuously referenced trajectories and end-to-end computational profiling are needed to assess deployment behavior. Verified VCU reference-frame semantics are necessary for stronger physical accuracy claims. The explicitly stated preprocessing, loss definitions, training budgets and evaluation conditions delimit the reported findings and provide the methodological information needed for independent implementation.

DECLARATIONS

Acknowledgment. The authors acknowledge the providers of the TUM RGB-D, VCU-RVI and ETH3D SLAM benchmarks for making their recordings available.

Funding. This research received no external funding.

CRediT Author Contributions. Farah Ibelaiden: Conceptualization, Methodology, Software, Validation, Writing – original draft, Writing – review & editing; Seif Eddine Bouziane: Investigation, Writing – review & editing; Ouarda Lounis: Investigation, Writing – review & editing. All authors have read and approved the manuscript and agree to be accountable for their respective contributions and for ensuring that questions concerning the accuracy or integrity of the work are appropriately investigated and resolved.

Conflicts of Interest. The authors declare no conflicts of interest.

Ethical Approval. Not applicable. This study uses existing public benchmark recordings and involves no newly recruited participants or interventions. Applicable provider data-use conditions remain in force.

Data Availability. The study uses the public TUM RGB-D, VCU-RVI and ETH3D SLAM benchmark recordings cited in the Experiments section. Access is governed by the respective providers' conditions. The sequences, preprocessing and evaluation subsets used in this study are specified in the manuscript.

Declaration of AI Assistance. Claude was used to support language editing. The authors retain full responsibility for the manuscript and its scientific content.

REFERENCES

[1] C. Campos, R. Elvira, J. J. G. Rodríguez, J. M. M. Montiel, and J. D. Tardós, "ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM," *IEEE Transactions on Robotics*, vol. 37, no. 6, pp. 1874–1890, 2021. [Online]. Available: https://doi.org/10.1109/TRO.2021.3075644

[2] P. Geneva, K. Eckenhoff, W. Lee, Y. Yang, and G. Huang, "OpenVINS: A research platform for visual-inertial estimation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 4666–4672. [Online]. Available: https://doi.org/10.1109/ICRA40945.2020.9196524

[3] T. Qin, P. Li, and S. Shen, "VINS-Mono: A robust and versatile monocular visual-inertial state estimator," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 1004–1020, 2018. [Online]. Available: https://doi.org/10.1109/TRO.2018.2853729

[4] M. Bloesch, S. Omari, M. Hutter, and R. Siegwart, "Robust visual inertial odometry using a direct EKF-based approach," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 298–304. [Online]. Available: https://doi.org/10.1109/IROS.2015.7353389

[5] L. von Stumberg and D. Cremers, "Dm-vio: Delayed marginalization visual-inertial odometry," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1408–1415, 2022. [Online]. Available: https://doi.org/10.1109/LRA.2021.3140129

[6] J. Li *et al.*, "Rd-vio: Robust visual-inertial odometry for mobile augmented reality in dynamic environments," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 10, pp. 6941–6955, 2024. [Online]. Available: https://doi.org/10.1109/TVCG.2024.3353263

[7] C. X. Guo and S. I. Roumeliotis, "IMU-RGBD camera navigation using point and plane features," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013, pp. 3164–3171. [Online]. Available: https://doi.org/10.1109/IROS.2013.6696806

[8] Z. Shan, R. Li, and S. Schwertfeger, "RGBD-inertial trajectory estimation and mapping for ground robots," *Sensors*, vol. 19, no. 10, p. 2251, 2019. [Online]. Available: https://doi.org/10.3390/s19102251

[9] H. Zhang, L. Jin, and C. Ye, "The VCU-RVI benchmark: Evaluating visual inertial odometry for indoor navigation applications with an RGB-D camera," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 6209–6214. [Online]. Available: https://doi.org/10.1109/IROS45743.2020.9341713

[10] J. Zhang, M. Kaess, and S. Singh, "On degeneracy of optimization-based state estimation problems," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 809–816. [Online]. Available: https://doi.org/10.1109/ICRA.2016.7487211

[11] T. Tuna, J. Nubert, Y. Nava, S. Khattak, and M. Hutter, "X-icp: Localizability-aware LiDAR registration for robust localization in extreme environments," *IEEE Transactions on Robotics*, vol. 40, pp. 452–471, 2024. [Online]. Available: https://doi.org/10.1109/TRO.2023.3335691

[12] J. Nubert, S. Khattak, and M. Hutter, "Self-supervised learning of LiDAR odometry for robotic applications," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 9601–9607. [Online]. Available: https://doi.org/10.1109/icra48506.2021.9561063

[13] Y. Xu *et al.*, "Selfvoxelo: Self-supervised LiDAR odometry with voxel-based deep neural networks," in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 155. PMLR, 2021, pp. 115–125. [Online]. Available: https://proceedings.mlr.press/v155/xu21a.html

[14] A. I. Mourikis and S. I. Roumeliotis, "A multi-state constraint Kalman filter for vision-aided inertial navigation," in *Proceedings of the 2007 IEEE International Conference on Robotics and Automation (ICRA)*, 2007, pp. 3565–3572. [Online]. Available: https://doi.org/10.1109/ROBOT.2007.364024

[15] C. Forster, L. Carlone, F. Dellaert, and D. Scaramuzza, "On-manifold preintegration for real-time visual-inertial odometry," *IEEE Transactions on Robotics*, vol. 33, no. 1, pp. 1–21, 2017. [Online]. Available: https://doi.org/10.1109/TRO.2016.2597321

[16] Y. Yang, P. Geneva, K. Eckenhoff, and G. Huang, "Tightly-coupled aided inertial navigation with point and plane features," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 6094–6100. [Online]. Available: https://doi.org/10.1109/ICRA.2019.8794078

[17] H. Zhang and C. Ye, "DUI-VIO: Depth uncertainty incorporated visual inertial odometry based on an RGB-D camera," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 5002–5008. [Online]. Available: https://doi.org/10.1109/IROS45743.2020.9341592

[18] R. A. Newcombe *et al.*, "KinectFusion: Real-time dense surface mapping and tracking," in *2011 10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2011, pp. 127–136. [Online]. Available: https://doi.org/10.1109/ISMAR.2011.6092378

[19] C. Kerl, J. Sturm, and D. Cremers, "Dense visual SLAM for RGB-D cameras," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2013, pp. 2100–2106. [Online]. Available: https://doi.org/10.1109/IROS.2013.6696650

[20] C. Feng, Y. Taguchi, and V. R. Kamat, "Fast plane extraction in organized point clouds using agglomerative hierarchical clustering," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 6218–6225. [Online]. Available: https://doi.org/10.1109/ICRA.2014.6907776

[21] P. F. Proença and Y. Gao, "Fast cylinder and plane extraction from depth cameras for visual odometry," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 6813–6820. [Online]. Available: https://doi.org/10.1109/IROS.2018.8593516

[22] M. Kaess, "Simultaneous localization and mapping with infinite planes," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 4605–4611. [Online]. Available: https://doi.org/10.1109/ICRA.2015.7139837

[23] M. Hsiao, E. Westman, G. Zhang, and M. Kaess, "Keyframe-based dense planar SLAM," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 5110–5117. [Online]. Available: https://doi.org/10.1109/ICRA.2017.7989597

[24] S. Wang, R. Clark, H. Wen, and N. Trigoni, "Deepvo: Towards end-to-end visual odometry with deep recurrent convolutional neural networks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 2043–2050. [Online]. Available: https://doi.org/10.1109/ICRA.2017.7989236

[25] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," in *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, 2021, pp. 16 558–16 569. [Online]. Available: https://proceedings.neurips.cc/paper/2021/hash/89fcd07f20b6785b92134bd6c1d0fa42-Abstract.html

[26] Z. Teed, L. Lipson, and J. Deng, "Deep patch visual odometry," in *Advances in Neural Information Processing Systems*, vol. 36, 2023. [Online]. Available: https://doi.org/10.52202/075280-1696

[27] W. Wang, Y. Hu, and S. Scherer, "Tartanvo: A generalizable learning-based vo," in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 155. PMLR, 2021, pp. 1761–1772. [Online]. Available: https://proceedings.mlr.press/v155/wang21h.html

[28] H. Zhan, C. S. Weerasekera, J.-W. Bian, and I. Reid, "Visual odometry revisited: What should be learnt?" in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4203–4210. [Online]. Available: https://doi.org/10.1109/icra40945.2020.9197374

[29] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6612–6619. [Online]. Available: https://doi.org/10.1109/CVPR.2017.700

[30] C. Godard, O. Mac Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 3827–3837. [Online]. Available: https://doi.org/10.1109/iccv.2019.00393

[31] R. Li, S. Wang, Z. Long, and D. Gu, "Undeepvo: Monocular visual odometry through unsupervised deep learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 7286–7291. [Online]. Available: https://doi.org/10.1109/ICRA.2018.8461251

[32] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5667–5675. [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2018/html/Mahjourian_Unsupervised_Learning_of_CVPR_2018_paper.html

[33] H. Li, A. Gordon, H. Zhao, V. Casser, and A. Angelova, "Unsupervised monocular depth learning in dynamic scenes," in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 155. PMLR, 2021, pp. 1908–1917. [Online]. Available: https://proceedings.mlr.press/v155/li21a.html

[34] Y.-S. Shin, Y. S. Park, and A. Kim, "DVL-SLAM: sparse depth enhanced direct visual-LiDAR SLAM," *Autonomous Robots*, vol. 44, no. 2, pp. 115–130, 2020. [Online]. Available: https://doi.org/10.1007/s10514-019-09881-0

[35] M. Maaz *et al.*, "Edgenext: Efficiently amalgamated CNN-transformer architecture for mobile vision applications," in *Computer Vision – ECCV 2022 Workshops*. Springer, 2023, pp. 3–20. [Online]. Available: https://doi.org/10.1007/978-3-031-25082-8_1

[36] J. Pan *et al.*, "Edgevits: Competing light-weight cnns on mobile devices with vision transformers," in *Computer Vision – ECCV 2022*. Springer, 2022, pp. 294–311. [Online]. Available: https://doi.org/10.1007/978-3-031-20083-0_18

[37] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

[38] C.-C. Chiu, H.-K. Yang, H.-W. Chen, Y.-W. Chen, and C.-Y. Lee, "Vitvo: Vision transformer based visual odometry with attention supervision," in *2023 18th International Conference on Machine Vision and Applications (MVA)*. IEEE, 2023, pp. 1–4. [Online]. Available: https://doi.org/10.23919/mva57639.2023.10215538

[39] A. O. Françani and M. R. Maximo, "Transformer-based model for monocular visual odometry: A video understanding approach," *IEEE Access*, vol. 13, pp. 1–15, 2025. [Online]. Available: https://doi.org/10.1109/access.2025.3531667

[40] Z. Wu and Y. Zhu, "Swformer-vo: A monocular visual odometry model based on swin transformer," *IEEE Robotics and Automation Letters*, vol. 9, no. 5, pp. 4301–4308, 2024. [Online]. Available: https://doi.org/10.1109/lra.2024.3384911

[41] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021. [Online]. Available: https://doi.org/10.1007/s11263-021-01453-z

[42] P. Kage, J. C. Rothenberger, P. Andreadis, and D. I. Diochnos, "A review of pseudo-labeling for computer vision," *Journal of Artificial Intelligence Research*, 2026. [Online]. Available: https://www.jair.org/index.php/jair/article/view/19656

[43] J. Gawlikowski *et al.*, "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 1513–1589, 2023. [Online]. Available: https://doi.org/10.1007/s10462-023-10562-9

[44] T.-D. Phan and G.-W. Kim, "Toward specialized learning-based approaches for visual odometry: A comprehensive survey," *Journal of Intelligent & Robotic Systems*, vol. 111, Art. no. 44, 2025. [Online]. Available: https://doi.org/10.1007/s10846-025-02245-0

[45] J. Zhang and S. Singh, "Loam: LiDAR odometry and mapping in real-time," in *Robotics: Science and Systems*, 2014. [Online]. Available: https://doi.org/10.15607/RSS.2014.X.007

[46] J. Lee, R. Komatsu, M. Shinozaki, T. Kitajima, H. Asama, and Q. An, "Switch-slam: Switching-based LiDAR-inertial-visual slam for degenerate environments," *IEEE Robotics and Automation Letters*, vol. 9, no. 8, pp. 7270–7277, 2024. [Online]. Available: https://doi.org/10.1109/LRA.2024.3421792

[47] J. A. Hesch, D. G. Kottas, S. L. Bowman, and S. I. Roumeliotis, "Camera-IMU-based localization: Observability analysis and consistency improvement," *The International Journal of Robotics Research*, vol. 33, no. 1, pp. 182–201, 2014. [Online]. Available: https://doi.org/10.1177/0278364913509675

[48] G. Huang, "Visual-inertial navigation: A concise review," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 9572–9582. [Online]. Available: https://doi.org/10.1109/ICRA.2019.8793604

[49] Y. Taguchi, Y.-D. Jian, S. Ramalingam, and C. Feng, "Point-plane SLAM for hand-held 3D sensors," in *2013 IEEE International Conference on Robotics and Automation (ICRA)*, 2013, pp. 5182–5189. [Online]. Available: https://doi.org/10.1109/ICRA.2013.6631318

[50] I. Vizzo, T. Guadagnino, B. Mersch, L. Wiesmann, J. Behley, and C. Stachniss, "Kiss-icp: In defense of point-to-point icp – simple, accurate, and robust registration if done the right way," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 1029–1036, 2023. [Online]. Available: https://doi.org/10.1109/lra.2023.3236571

[51] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5738–5746. [Online]. Available: https://doi.org/10.1109/CVPR.2019.00589

[52] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019, pp. 8024–8035. [Online]. Available: https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: https://arxiv.org/abs/1412.6980

[54] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

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

[56] T. Schöps, T. Sattler, and M. Pollefeys, "Bad slam: Bundle adjusted direct rgb-d slam," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 134–144. [Online]. Available: https://doi.org/10.1109/cvpr.2019.00022