УДК 004.932.2:004.89
DOI: 10.26102/2310-6018/2026.59.8.013

Spatial-temporal fusion of depth data for estimating three-dimensional distances

X. Yu

Peoples' Friendship University of Russia named after Patrice Lumumba, Moscow, the Russian Federation

Abstract. Depth enhancement is commonly evaluated with pixel-wise errors, although robotic vision and three-dimensional metrology ultimately use geometric distances. This study develops and tests a protocol that evaluates enhancement methods by the error of Euclidean distances between back-projected point pairs. The primary experiments use the TUM Freiburg1 desk and xyz sequences. Additional validation includes the dynamic Freiburg3 walking_static sequence, controlled stress tests with moving depth boundaries, long range and elevated noise, three pseudo-reference variants, and a computational-cost benchmark. With temporal half-window N = 10, temporal mean fusion reduces distance RMSE by 24.07 % on desk and 23.66 % on xyz relative to raw depth. Direct metric comparison shows that temporal median has the lowest pixel MAE, whereas temporal mean has the lowest distance RMSE; Spearman rank correlations are 0.70 and 0.90. Temporal mean remains ranked first under all three pseudo-reference constructions. On the dynamic sequence, a temporal-stability gate retains 18.92 % of pixels and changes the method ranking, demonstrating that moving regions must be excluded or motion-compensated. For 640 × 480 frames at N = 10, the research CPU implementation requires 56.2 ms for temporal mean, 217.5 ms for temporal median, and 3155.6 ms for spatial-temporal median. The results delimit the practical use of the protocol and the systematic limitations of pseudo-reference evaluation.

Keywords: depth data, three-dimensional measurement, depth map, pseudo-reference, spatial-temporal fusion, geometric accuracy, depth enhancement.

For citation: Yu X. Spatial-temporal fusion of depth data for estimating three-dimensional distances. *Modeling, Optimization and Information Technology.* 2026;14(8). URL: https://moitvivt.ru/ru/journal/article?id=2414 DOI: 10.26102/2310-6018/2026.59.8.013

Пространственно-временное слияние глубинных данных для оценки трёхмерных расстояний

С. Юй

Российский университет дружбы народов имени Патриса Лумумбы, Москва, Российская Федерация

Резюме. Актуальность исследования обусловлена тем, что качество обработки глубинных данных часто оценивают по пиксельным ошибкам, хотя в робототехнике и трехмерной метрологии конечным результатом являются пространственные расстояния. Цель работы – разработать и проверить протокол оценки методов улучшения глубины по ошибке евклидовых расстояний между восстановленными точками. Основные эксперименты выполнены на последовательностях TUM Freiburg1 desk и xyz; дополнительная валидация включает динамическую последовательность Freiburg3 walking_static, контролируемые сценарии с движением границ, дальними глубинами и повышенным шумом, три варианта псевдоэталона и измерение вычислительных затрат. При полуширине окна N = 10 временное среднее уменьшило RMSE расстояний на 24,07 % для desk и на 23,66 % для xyz по сравнению с необработанной глубиной. Прямое сопоставление метрик показало, что минимальную пиксельную MAE дает временная медиана, тогда как минимальную RMSE расстояний – временное среднее;

коэффициенты ранговой корреляции Спирмена равны 0,70 и 0,90. Временное среднее сохранило первое место при трех способах построения псевдоэталона. На динамической последовательности маска временной стабильности сохранила 18,92 % пикселей и изменила ранжирование, что подтверждает необходимость исключения или компенсации движущихся областей. На CPU обработка кадра 640 × 480 при N = 10 заняла 56,2 мс для временного среднего, 217,5 мс для временной медианы и 3155,6 мс для пространственно-временной медианы в исследовательской реализации. Полученные результаты определяют область применимости протокола и ограничения псевдоэталонной оценки.

Ключевые слова: глубинные данные, трехмерные измерения, карта глубины, псевдоэталон, пространственно-временное слияние, геометрическая точность, улучшение глубины.

Для цитирования: Юй С. Пространственно-временное слияние глубинных данных для оценки трёхмерных расстояний. *Моделирование, оптимизация и информационные технологии.* 2026;14(8). (На англ.) URL: https://moitvivt.ru/ru/journal/article?id=2414 DOI: 10.26102/2310-6018/2026.59.8.013

Introduction

Depth cameras that provide range measurements are increasingly used in three-dimensional reconstruction, robotic perception, indoor mapping, navigation, and measurement-oriented scene analysis. In these applications, the depth map is rarely the final result of processing. It is usually an intermediate representation from which three-dimensional points, object dimensions, local surfaces, trajectories, and point-to-point distances are computed. For this reason, the practical value of a depth enhancement method depends not only on the visual quality of the depth map, but also on the accuracy of the geometric quantities obtained after back-projection.

Pixel-wise approaches. Depth enhancement is most often compared by per-pixel MAE, RMSE, absolute relative error, and valid-pixel coverage. Early studies quantified Kinect depth accuracy and noise [1, 2], and subsequent work compared structured-light and time-of-flight sensing [3]. Classical bilateral and guided filters reduce edge-preserving image noise [4, 5], whereas temporal-spatial denoising and RGB-D fusion exploit local and temporal redundancy [6, 7]. U-Net-type architectures and self-supervised depth denoising later broadened the set of learning-based enhancement tools [8, 9].

Structural approaches. Public RGB-D benchmarks provide reproducible sequences for evaluating geometric reconstruction and motion [10], while paired lower- and higher-quality sensors support self-supervised denoising [11]. RGB and sparse-depth fusion preserves image-guided structure in dense completion [12]. Downstream structural consistency is also central to static-object localization and mobile-robot SLAM [13, 14], as well as navigation and reference-guided depth inpainting [15, 16]. Unsupervised completion extends this line to sparse depth without direct labels [17]. Recent methods combine convolution-transformer completion with recurrent updating [18, 19], bilateral propagation with variable-density completion [20, 21], and geometry-aware decomposition with practical dToF enhancement [22, 23]. The latest work addresses sparse direct-ToF measurements and lightweight RGB-ToF fusion [24, 25]. These criteria and models are closer to scene structure than scalar pixel errors, but they do not directly quantify the final distance between reconstructed points.

Geometric approaches. Back-projection, point-cloud, distance, surface, and trajectory criteria evaluate the quantities used by mapping, localization, navigation, and metrology. Sensor characterization supplies the physical basis for back-projection and range error [1, 3], while benchmark and SLAM studies connect depth to trajectory and mapping quality [10, 14]. However, direct comparisons of pixel-level and distance-level rankings on exactly the same processed depth maps remain limited, and evaluations without an external high-precision reference require an explicit analysis of reference bias.

Position of this work. The proposed protocol complements rather than replaces pixel and structural metrics. It tests the downstream Euclidean-distance task, directly compares pixel and geometric rankings, evaluates pseudo-reference sensitivity, challenges the protocol with dynamic and controlled noisy scenes, and reports computational cost.

The research problem addressed in this paper is not that pixel and geometric errors are universally uncorrelated, but that their correlation and method ranking are task- and scene-dependent. Smoothing may reduce average depth residuals while displacing object boundaries or changing the relative geometry of two points. A measurement-oriented system therefore needs both a direct distance criterion and evidence that the chosen reference does not mechanically favor one processing operator.

The aim of the study is to develop and experimentally test a measurement-oriented evaluation framework for depth enhancement methods, where the principal quality criterion is the error of Euclidean distances between reconstructed point pairs. The tasks are: (1) to define a reproducible back-projection and pair-sampling protocol; (2) to compare pixel-wise and distance-level rankings on the same processed data; (3) to quantify sensitivity to pseudo-reference construction and slow scene motion; (4) to extend validation to dynamic, sharp-boundary, long-range, and high-noise regimes; and (5) to estimate time and memory complexity as the temporal window grows.

The scientific contribution consists of four linked elements: a reproducible distance-level protocol for depth enhancement without arbitrary physical point selection; a direct empirical test of agreement between pixel and geometric rankings; a quantitative pseudo-reference sensitivity analysis with a temporal-stability gate for dynamic regions; and a joint robustness-and-complexity characterization. The contribution is methodological; the engineering recommendations derived from it are stated separately in the conclusion.

Materials and methods

The primary experiments use two public indoor depth sequences, rgbd_dataset_freiburg1_desk and rgbd_dataset_freiburg1_xyz, from the Technical University of Munich RGB-D benchmark. The first sequence contains a cluttered desk scene with repeated viewpoint changes, while the second has more regular motion and scene structure. Robustness is additionally tested on rgbd_dataset_freiburg3_walking_static, in which two people move while the camera remains nearly stationary. The archive contains 723 depth frames; 60 uniformly spaced target frames are evaluated. Sixteen-bit depth values are converted to meters with scale factor 5000, and invalid observations are excluded from reconstruction and pair sampling.

Let $z_t(u,v)$ denote the depth value at pixel coordinates (u,v) in frame t. If $z_t(u,v) > 0$, and the camera intrinsic parameters are focal lengths f_x, f_y and principal point (c_x, c_y), the corresponding three-dimensional point is computed by the standard back-projection model:

$$P_t(u,v) = (X_t, Y_t, Z_t)^T = \left(\frac{(u-c_x)z_t(u,v)}{f_x}, \frac{(v-c_y)z_t(u,v)}{f_y}, z_t(u,v)\right)^T. \quad (1)$$

For two valid pixels a and b, the recovered geometric quantity is the Euclidean distance between the corresponding three-dimensional points:

$$D_t(a,b) = \|P_t(a) - P_t(b)\|_2. \quad (2)$$

If $a = (u_a, v_a)$ and $b = (u_b, v_b)$, the squared distance can be written as follows:

$${D_t}^2(a,b) = (X_t(a) - X_t(b))^2 + (Y_t(a) - Y_t(b))^2 + (Z_t(a) - Z_t(b))^2. \quad (3)$$

Equation (3) shows that an error in a depth value changes all three coordinates after back-projection. This is the reason why the pixel-wise error of a depth image and the geometric error of a distance estimate can have different rankings across methods.

Because arbitrary point pairs in the real sequences lack independent high-precision distances, the primary evaluation uses a pseudo-reference. For a target frame, a temporal neighborhood is aggregated while invalid values are ignored, followed by an optional 3 × 3 spatial median. The primary construction is temporal median with half-window 30 followed by spatial median. Sensitivity is tested with temporal mean at half-window 30 and temporal median at half-windows 20 and 30. A pseudo-reference is not an absolute physical truth: it supports relative comparison only, and agreement with its own construction operator must not be interpreted as independent accuracy.

The spatial median method replaces the depth value by the median in a local neighborhood. The temporal mean method calculates an average across the selected temporal window. The temporal median method replaces the average by a robust median. The spatial-temporal median method combines local and temporal robustness.

For each target frame, valid pixel pairs are sampled after applying the same valid-mask policy to the processed depth map and the pseudo-reference. For every pair, the Euclidean distance is calculated using formula (2). The processed distance is then compared with the reference distance. Errors are aggregated across sampled pairs, frames, and random seeds. The primary metrics are root mean square error and mean absolute error of the recovered distances:

$$RMSE = \sqrt{\left(\frac{1}{N}\sum_{i=1}^{N}(D_i - D_i^{ref})^2\right)}, \quad (4)$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|D_i - D_i^{ref}|. \quad (5)$$

The complete evaluation pipeline is summarized in Figure 1.

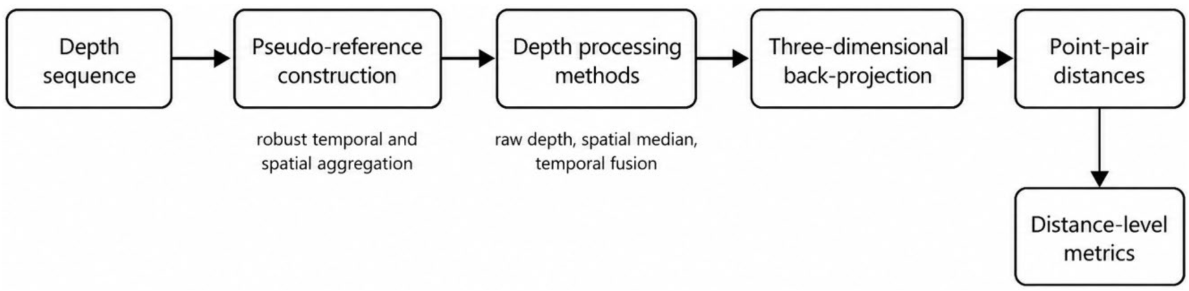

Figure 1 – Evaluation scheme for depth processing methods at the level of three-dimensional distances
Рисунок 1 – Схема оценки методов обработки глубины на уровне трехмерных расстояний

The roles of the compared methods are summarized in Table 1.

Table 1 – Compared methods and their role in the experimental protocol
Таблица 1 – Сравниваемые методы и их роль в экспериментальном протоколе

Method	Operation	Role in comparison
RAW	No processing	Baseline sensor output
Spatial median	Single-frame local median	Suppression of isolated local outliers
Temporal mean	Windowed temporal average	Use of temporal redundancy with sensitivity to outliers
Temporal median	Windowed temporal median	Robust temporal fusion
Spatial-temporal median	Spatial median and temporal median	Strong classical baseline

The primary configuration uses a pseudo-reference half-window of 30 frames, an evaluation half-window of 10 frames for temporal methods, 2000 sampled point pairs per frame, a depth limit of 10 m, 595 target frames per Freiburg1 sequence, and five random seeds. The dynamic test uses 60 target frames and three seeds. All methods use the same metric conversion, invalid-value policy, pair coordinates, and camera intrinsics within each sequence.

The main experimental parameters are presented in Table 2.

Table 2 – Main parameters of the experimental protocol
Таблица 2 – Основные параметры экспериментального протокола

Parameter	Value
Dataset	Technical University of Munich depth sequences: rgbd_dataset_freiburg1_desk and rgbd_dataset_freiburg1_xyz
Evaluated frames	595 target frames per sequence
Depth conversion	Sixteen-bit depth maps were converted to meters with scale factor 5000
Maximum accepted depth	10 m
Pseudo-reference construction	Temporal median with half-window 30 followed by spatial median filtering
Default temporal half-window	10 frames for the main comparison of temporal methods
Spatial median kernel	3 × 3 neighborhood
Point pairs per frame	2000 randomly sampled valid point pairs
Random seeds	0, 1, 2, 3, 4
Main metrics	Distance-level root mean square error and mean absolute error between reconstructed point-pair distances
Additional dynamic sequence	TUM Freiburg3 walking_static; 723 depth frames, 60 target frames
Pseudo-reference sensitivity	Median N = 30 + spatial median; mean N = 30 + spatial median; median N = 20 + spatial median
Dynamic stability gate	At least 80 % valid support and $q90 - q10 \le \max(0.05\ \text{m}, 0.03q50)$, estimated at 21 uniformly spaced timestamps
Controlled stress test	180 × 240 synthetic depth; sharp 1.5/4 m or 4/8 m boundary; 0–24 px motion; 2–10 % dropout; range-dependent noise
Cost benchmark	640 × 480 float32 frames; single-process CPU; in-memory timing; disk I/O excluded

Several implementation details are important for reproducibility. All depth maps are first converted to meters before any filtering or fusion operation. Invalid pixels are not filled by arbitrary constants; they are excluded from aggregation and from pair sampling. This rule prevents the evaluation from being distorted by artificial zero-depth points. The same mask policy is used for the processed map and for the pseudo-reference, so that every distance comparison is based on a pair that is valid in both representations.

The pseudo-reference is constructed independently of the method being evaluated. This separation is necessary because otherwise a method could be favored by using its own output as the comparison basis. The temporal median is selected as the main robust operator because it is resistant to isolated frame-level fluctuations. The spatial median is applied only as a local correction to reduce isolated depth outliers while keeping the method transparent and reproducible.

The random sampling of point pairs is repeated for several seeds. This is necessary because the set of valid points changes from frame to frame, and a single random sample may overrepresent a small region of the image. Averaging across seeds reduces the influence of this sampling variability and makes the comparison between methods more reliable, because all methods are evaluated on the same sequence structure and the same sampling logic.

The protocol therefore combines common units, a common valid-mask policy, repeated pair sampling, reference variation, and an explicit dynamic-region test. These safeguards reduce sampling and implementation bias but do not turn the pseudo-reference into an external metrological standard; the remaining limitation is quantified below.

Additional validation and sensitivity analyses. Pixel-versus-geometric comparison uses the same five processed maps and the same primary pseudo-reference. For each sequence, pixel MAE and distance RMSE are compared by Pearson correlation and Spearman rank correlation. Reference sensitivity is evaluated on 50 target frames per Freiburg1 sequence with 2000 pairs and three seeds; the rank of each method is recomputed for every reference construction.

For Freiburg3 walking_static, full-window aggregation is retained, but an evaluation pixel is marked temporally stable only when at least 80 % of the observations are valid and $q90 - q10 \leq \max(0.05\text{ m}, 0.03q50)$. Quantiles are estimated at 21 uniformly spaced timestamps. Results are reported both for all pixels and for the stable subset; the latter is a coverage-limited relative test, not a ground-truth measurement of moving regions.

The controlled stress test uses a synthetic depth field with analytically defined independent ground truth at 180 × 240 pixels, with a rectangular foreground and a sharp depth discontinuity. Standard scenes use 1.5 and 4 m surfaces; far scenes use 4 and 8 m. Across 61 frames, the foreground moves by 0, 6, or 24 pixels. Gaussian noise follows $\sigma(z) = cz^2$ with $c = 0.001$ or 0.003, invalid returns are 2 % or 10 %, and ten random seeds are used. This design isolates boundary motion, distance-dependent noise, and pseudo-reference error against the independent synthetic ground truth. It approximates degraded ranging but is not presented as a sensor-specific low-light experiment. No real low-illumination sequence was included; therefore, the reported results do not validate robustness under low-light sensing conditions.

Computational complexity is analyzed for image height H, width W, spatial kernel $k \times k$, and temporal support $T = 2N + 1$. Temporal mean requires $O(THW)$; a generic sorting-based temporal median requires $O(THW \log T)$, although selection implementations can reduce the average factor; spatial median requires $O(HWk^2)$; and the spatial-temporal median combines the spatial cost across T frames with temporal aggregation. The minimum float32 input-stack memory is 4THW bytes. Wall-clock timings were measured on a 12th Gen Intel Core i9-12900H system with 16 GB RAM under Windows 11, using Python 3.12.13 and NumPy 2.3.5 in a single-process implementation with in-memory 640 × 480 arrays; disk I/O was excluded.

Results

Tables 3 and 4 present the overall distance-level comparison under the default setting. On the desk sequence, TemporalMean achieves the lowest distance root mean square error, followed by TemporalMedian and SpatialTemporalMedian. On the xyz sequence, the same ordering is observed. Spatial median filtering gives a higher root mean square error than the raw baseline on both sequences. Compared with RAW, TemporalMean reduces the distance-level root mean square error by 24.07 % on the desk sequence and by 23.66 % on the xyz sequence.

Table 3 – Desk sequence: distance-level comparison under the default setting, half-window N = 10
Таблица 3 – Последовательность desk: сравнение методов на уровне расстояний при половине окна N = 10

Method	RMSE mean (m)	RMSE std (m)	MAE mean (m)	RMSE gain vs RAW (%)
Temporal mean	0.20233	0.09047	0.12002	24.07
Temporal median	0.23345	0.10492	0.12250	12.39
Spatial-temporal median	0.23696	0.10236	0.12422	11.08
RAW	0.26647	0.13672	0.14760	0.00
Spatial median	0.27288	0.13612	0.15150	−2.41

Table 4 – XYZ sequence: distance-level comparison under the default setting, half-window N = 10
Таблица 4 – Последовательность xyz: сравнение методов на уровне расстояний при половине окна N = 10

Method	RMSE mean (m)	RMSE std (m)	MAE mean (m)	RMSE gain vs RAW (%)
Temporal mean	0.23039	0.08528	0.12540	23.66
Temporal median	0.27090	0.10971	0.12739	10.23
Spatial-temporal median	0.27714	0.10820	0.13026	8.16
RAW	0.30178	0.12412	0.14907	0.00
Spatial median	0.31117	0.12205	0.15412	−3.11

The temporal window has a decisive influence. Figure 2 shows that the errors relative to the primary pseudo-reference decrease as the window grows. Temporal median at N = 30 reaches 0.02185 m on desk and 0.02862 m on xyz. However, the reference itself is constructed with temporal median at N = 30 followed by a spatial median, so these very low values primarily diagnose operator self-similarity and are not independent evidence of superiority. Consequently, N = 10 remains the primary cross-method comparison, and the N = 30 sweep is interpreted only together with the reference-sensitivity and independent-ground-truth tests.

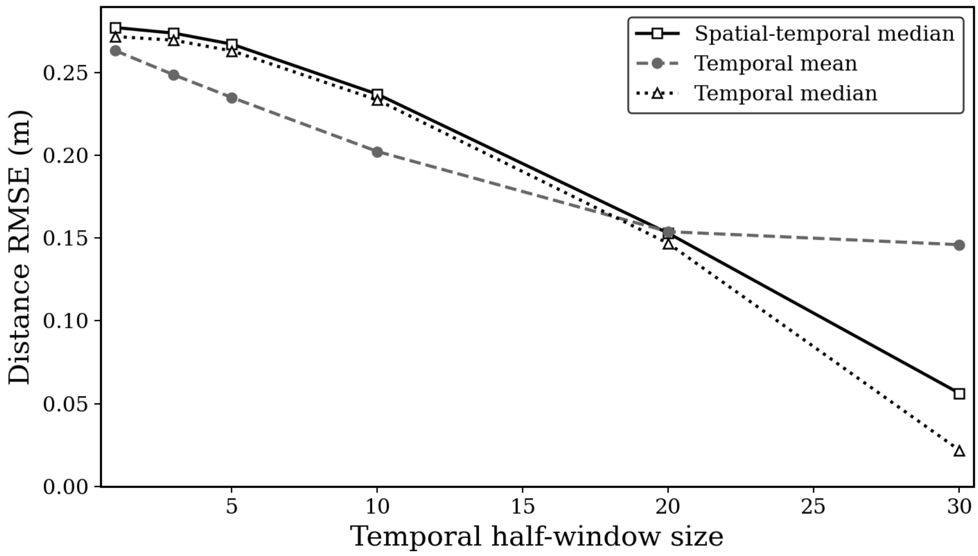

Figure 2 – Desk sequence: effect of temporal half-window size on distance-estimation error
Рисунок 2 – Последовательность desk: влияние размера временного окна на ошибку оценки расстояний

Direct comparison of pixel and geometric metrics. Figure 3 and Table 5 compare pixel MAE and distance RMSE on exactly the same outputs. Pearson correlations are 0.864 for desk and 0.812 for xyz, while Spearman rank correlations are 0.70 and 0.90. Thus, the metric families are related but not interchangeable: temporal median is best by pixel MAE on both sequences, whereas temporal mean is best by distance RMSE. This direct rank reversal is the evidence for retaining the geometric protocol.

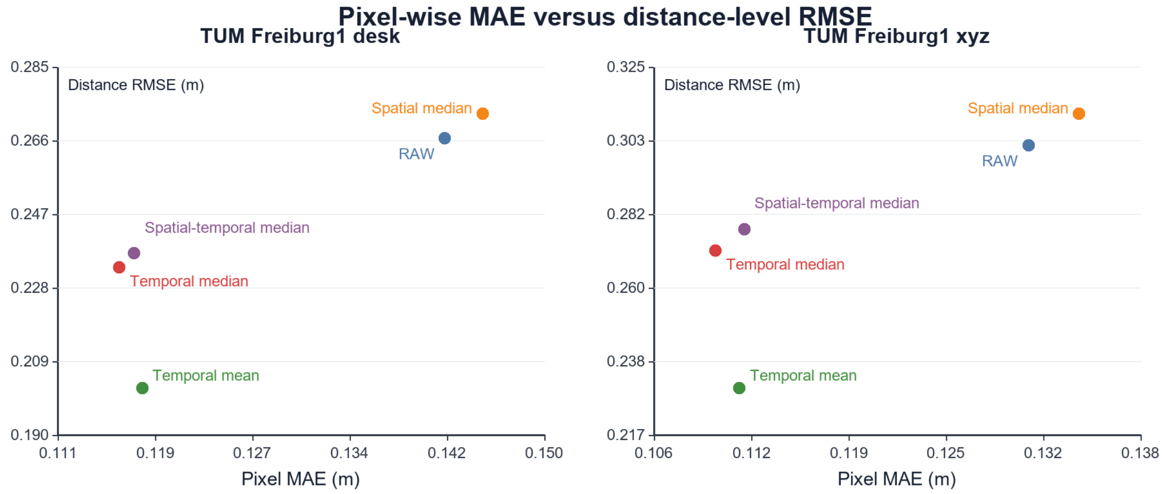

Figure 3 – Pixel MAE versus three-dimensional distance RMSE for the same processed data
Рисунок 3 – Сопоставление пиксельной MAE и RMSE трехмерных расстояний для одних и тех же обработанных данных

Table 5 – Direct comparison of pixel and geometric rankings
Таблица 5 – Прямое сопоставление пиксельного и геометрического ранжирования

Sequence	Method	Pixel MAE (m)	Pixel rank	Distance RMSE (m)	Distance rank
desk	Temporal median	0.11579	1	0.23345	2
desk	Spatial-temporal median	0.11698	2	0.23696	3
desk	Temporal mean	0.11763	3	0.20233	1
desk	RAW	0.14214	4	0.26647	4
desk	Spatial median	0.14518	5	0.27288	5
xyz	Temporal median	0.10988	1	0.27090	2
xyz	Temporal mean	0.11145	2	0.23039	1
xyz	Spatial-temporal median	0.11177	3	0.27714	3
xyz	RAW	0.13061	4	0.30178	4
xyz	Spatial median	0.13393	5	0.31117	5

Pseudo-reference sensitivity and dynamic-scene validation. Across the three pseudo-reference constructions, temporal mean is ranked first and temporal median second on both

Freiburg1 sequences (Table 6). The invariant ranking shows that the primary N = 10 conclusion is not tied to one median-reference operator. It does not, however, validate absolute physical accuracy.

Table 6 – Ranking stability under three pseudo-reference variants
Таблица 6 – Устойчивость ранжирования к трем вариантам псевдоэталона

Sequence	Method	Mean RMSE across references (m)	Mean rank	Worst rank
desk	RAW	0.25493	3.0	3
desk	Spatial median	0.26334	4.0	4
desk	Temporal mean	0.17971	1.0	1
desk	Temporal median	0.21425	2.0	2
xyz	RAW	0.27731	3.0	3
xyz	Spatial median	0.28603	4.0	4
xyz	Temporal mean	0.19915	1.0	1
xyz	Temporal median	0.24310	2.0	2

The dynamic sequence produces a deliberate stress case (Table 7). On all pixels, RAW is ranked first relative to the unregistered pseudo-reference and temporal median fourth. After the stability gate, which retains only 18.92 % of pixels, temporal median becomes first and temporal mean second. This rank reversal indicates that slow or non-rigid motion inside a long window contaminates reference values near moving objects. The stable-mask result is therefore a restricted relative comparison; moving regions require motion compensation or an external reference.

Table 7 – Dynamic Freiburg3 walking_static sequence
Таблица 7 – Динамическая последовательность Freiburg3 walking_static

Method	All-pixel RMSE (m)	Rank	Stable-pixel RMSE (m)	Rank	Stable support (%)
RAW	0.66107	1	0.11226	3	18.92
Spatial median	0.72892	2	0.11466	4	18.92
Temporal mean	0.73064	3	0.05897	2	18.92
Temporal median	0.77671	4	0.02569	1	18.92

The controlled scenes quantify the mechanism (Table 8). Under static standard conditions, the rank correlation between independent-ground-truth and pseudo-reference rankings is only 0.17 because the spatial-temporal median resembles the reference operator. With 6-pixel and 24-pixel motion, the correlations are 0.59 and 0.53, and the pseudo-reference boundary MAE rises from 0.00444 m to 0.00644 m and 0.01057 m. In the 4–8 m high-noise moving case it reaches 0.04120 m, while the stability gate retains 10.36 % of pixels. Thus, reference bias is concentrated at moving discontinuities and becomes a coverage-versus-validity trade-off.

Table 8 – Pseudo-reference sensitivity in controlled scenarios
Таблица 8 – Чувствительность псевдоэталона в контролируемых сценариях

Scenario	**Best by independent GT**	**Best by pseudo-reference**	**Spear man ρ**	**Reference RMSE (m)**	**Boundary MAE (m)**
Static, standard	Temporal mean	Spatial-temporal median	0.17	0.01602	0.00444
Slow motion (6 px)	Temporal median	Spatial-temporal median	0.59	0.01612	0.00644
Fast motion (24 px)	Temporal median	Spatial-temporal median	0.53	0.01636	0.01057
Far, slow motion, high noise	Spatial-temporal median	Spatial-temporal median	0.95	0.02825	0.04120

The benchmark in Table 9 shows that temporal support is also a resource parameter. At $N = 10$, temporal mean requires 56.2 ms per frame, temporal median 217.5 ms, and the unoptimized spatial-temporal median 3155.6 ms. At $N = 30$, minimum input-stack memory rises to 71.48 MiB and the corresponding times are 137.1, 610.9, and 8884.7 ms. The measurements characterize the stated Python/NumPy implementation and are not hardware-independent real-time guarantees.

Table 9 – Computational cost at 640 × 480 pixels
Таблица 9 – Вычислительные затраты при 640 × 480 пикселей

Method	**N**	**T**	**Time (ms)**	**Min. stack (MiB)**	**Asymptotic cost**
Spatial median	0	1	154.3	1.17	$O(HWk^2)$
RAW	0	1	0.3	1.17	$O(HW)$
Temporal mean	10	21	56.2	24.61	$O(THW)$
Temporal median	10	21	217.5	24.61	$O(THW \log T)$*
Temporal mean	30	61	137.1	71.48	$O(THW)$
Temporal median	30	61	610.9	71.48	$O(THW \log T)$*
Spatial-temporal median	10	21	3155.6	24.61	$O(THWk^2 + THW \log T)$*
Spatial-temporal median	30	61	8884.7	71.48	$O(THWk^2 + THW \log T)$*

* Generic sorting bound; selection-based implementations can reduce the average temporal-median factor.

Discussion

The direct comparison supports a measurement-oriented protocol without claiming that pixel metrics are uninformative. Correlations are positive, but the best method changes from temporal median under pixel MAE to temporal mean under distance RMSE. The reason is geometric: depth residuals at different image coordinates affect all components of the back-projected point and therefore the distance between a pair. A complete evaluation should report pixel, structural or boundary, and downstream geometric criteria together.

For the primary N = 10 comparison, temporal mean improves distance RMSE by 24.07 % on desk and 23.66 % on xyz, and remains first under all three pseudo-reference constructions. This is stronger evidence than the original N = 30 result because it does not rely on exact equality between the evaluated temporal operator and the reference window. The complexity results also show why a larger or more robust window is not automatically preferable in real-time systems: its latency and memory grow with temporal support, and spatial-temporal median is especially costly.

The pseudo-reference remains the main limitation. Controlled synthetic experiments with independent ground truth show that its method ranking can diverge sharply under operator self-similarity and moving boundaries. On Freiburg3 walking_static, the stability gate changes the ranking but retains only 18.92 % of pixels; in the far, high-noise moving simulation it retains 10.36 %. Hence, stable masking improves interpretability at the cost of coverage and cannot validate moving objects. Dynamic areas require optical-flow or pose-based registration, scene-flow compensation, or an external calibrated reference. The added public dynamic sequence and controlled high-noise tests broaden the evidence base, but they do not replace sensor-specific experiments under low illumination, reflective materials, or very long range. Such calibrated tests are a priority for future work on practical dToF enhancement [23], sparse direct-ToF completion [24], and lightweight RGB-ToF fusion [25].

Conclusion

Scientific contribution. The paper introduces a reproducible distance-level evaluation protocol, directly demonstrates that pixel and geometric rankings can differ on the same processed data, quantifies pseudo-reference sensitivity under reference variation and scene motion, and links robustness to time and memory complexity.

Empirical findings. At N = 10, temporal mean reduces distance RMSE by 24.07 % on desk and 23.66 % on xyz and remains ranked first for three pseudo-reference variants. Pixel MAE instead selects temporal median. Dynamic and controlled tests show that moving boundaries can reverse pseudo-reference rankings; the N = 30 temporal-median result is therefore retained only as a self-similarity diagnostic, not as independent evidence of accuracy.

Engineering recommendations and limitations. For aligned, predominantly static RGB-D sequences under the tested conditions, temporal mean at moderate support offers the best accuracy-cost trade-off. Robust median fusion is appropriate when outliers dominate and temporal stability is verified. Unregistered dynamic regions should be masked or motion-compensated, and any absolute metrological claim requires an external calibrated reference. Future work should add sensor-specific low-light, reflective-surface, and long-range measurements and optimized parallel implementations.

REFERENCES / СПИСОК ИСТОЧНИКОВ

1. Khoshelham K., Oude Elberink S. Accuracy and resolution of Kinect depth data for indoor mapping applications. *Sensors*. 2012;12(2):1437–1454. https://doi.org/10.3390/s120201437

2. Nguyen Ch.V., Izadi Sh., Lovell D. Modeling Kinect sensor noise for improved 3D reconstruction and tracking. In: *2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission, 13–15 October 2012, Zurich, Switzerland*. IEEE; 2012. P. 524–530. https://doi.org/10.1109/3DIMPVT.2012.84
3. Sarbolandi H., Lefloch D., Kolb A. Kinect range sensing: structured-light versus time-of-flight Kinect. *Computer Vision and Image Understanding*. 2015;139:1–20. https://doi.org/10.1016/j.cviu.2015.05.006
4. Tomasi C., Manduchi R. Bilateral filtering for gray and color images. In: *Sixth International Conference on Computer Vision, 04–07 January 1998, Bombay, India*. IEEE; 1998. P. 839–846. https://doi.org/10.1109/ICCV.1998.710815
5. He K., Sun J., Tang X. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013;35(6):1397–1409. https://doi.org/10.1109/TPAMI.2012.213
6. Lin B.-Sh., Su M.-J., Cheng P.-H., et al. Temporal and spatial denoising of depth maps. *Sensors*. 2015;15(8):18506–18525. https://doi.org/10.3390/s150818506
7. Or-El R., Rosman G., Wetzler A., et al. RGBD-fusion: real-time high precision depth recovery. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 07–12 June 2015, Boston, MA, USA*. IEEE; 2015. P. 5407–5416. https://doi.org/10.1109/CVPR.2015.7299179
8. Ronneberger O., Fischer Ph., Brox Th. U-Net: convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015: 18th International Conference: Part III, 05–09 October 2015, Munich, Germany*. Cham: Springer; 2015. P. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28
9. Sterzentsenko V., Saroglou L., Chatzitofis A., et al. Self-supervised deep depth denoising. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV), 27 October – 02 November 2019, Seoul, South Korea*. IEEE; 2019. P. 1242–1251. https://doi.org/10.1109/ICCV.2019.00133
10. Sturm J., Engelhard N., Endres F., et al. A benchmark for the evaluation of RGB-D SLAM systems. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, 07–12 October 2012, Vilamoura-Algarve, Portugal*. IEEE; 2012. P. 573–580. https://doi.org/10.1109/IROS.2012.6385773
11. Shabanov A., Krotov I., Chinaev N., et al. Self-supervised depth denoising using lower- and higher-quality RGB-D sensors. In: *2020 International Conference on 3D Vision (3DV), 25–28 November 2020, Fukuoka, Japan*. IEEE; 2020. P. 743–752. https://doi.org/10.1109/3DV50981.2020.00084
12. Shivakumar Sh.S., Nguyen T., Miller I.D., et al. DFuseNet: deep fusion of RGB and sparse depth information for image guided dense depth completion. In: *2019 IEEE Intelligent Transportation Systems Conference (ITSC), 27–30 October 2019, Auckland, New Zealand*. IEEE; 2019. P. 13–20. https://doi.org/10.1109/ITSC.2019.8917294
13. Пьявченко А.О., Ильченко А.В. Метод пространственной локализации статических объектов по данным датчика глубины и RGB-камеры. *Известия ЮФУ. Технические науки*. 2018;(1):271–284. https://doi.org/10.23683/2311-3103-2018-1-271-284
Pyavchenko A.O., Ilchenko A.V. Method of the static objects spatial localization according to depth sensor data and RGB camera. *Izvestiya SFedU. Engineering Sciences*. 2018;(1):271–284. (In Russ.). https://doi.org/10.23683/2311-3103-2018-1-271-284
14. Вохминцев А.В., Пачганов С.А. Алгоритм одновременной навигации и составления карты мобильным роботом на основе итеративного алгоритма ближайших точек и дескриптора, вычисляемого в круглом скользящем окне. *Вестник Югорского государственного университета*. 2018;14(3):49–56.

https://doi.org/10.17816/byusu2018049-56
Vokhmintcev A.V., Pachganov S.A. The algorithm of simultaneous navigation and mapping for mobile robot based on the iterative algorithm of the nearest pixels and descriptor calculated in a circular moving window. *Yugra State University Bulletin*. 2018;14(3):49–56. (In Russ.). https://doi.org/10.17816/byusu2018049-56

15. Осман В., Громов В.С. Исследование системы навигации для мобильных роботов на основе одновременной локализации и построения карты. *Научно-технический вестник информационных технологий, механики и оптики*. 2020;20(3):371–376. https://doi.org/10.17586/2226-1494-2020-20-3-371-376
Othman W., Gromov V.S. Research of visual simultaneous localization and mapping-based navigation system for mobile robots. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*. 2020;20(3):371–376. (In Russ.). https://doi.org/10.17586/2226-1494-2020-20-3-371-376
16. Wu H., Fu K., Zhao Y., et al. Joint self-supervised and reference-guided learning for depth inpainting. *Computational Visual Media*. 2022;8(4):597–612. https://doi.org/10.1007/s41095-021-0259-z
17. Yang Q., Cui W., Zheng Zh., et al. Unsupervised depth completion based on RGB image and sparse depth map. In: *EITCE 2024: Proceedings of the 2024 8th International Conference on Electronic Information Technology and Computer Engineering, 18–20 October 2024, Haikou, China*. New York: ACM; 2024. P. 152–157. https://doi.org/10.1145/3711129.3711157
18. Zhang Y., Guo X., Poggi M., et al. CompletionFormer: depth completion with convolutions and vision transformers. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 17–24 June 2023, Vancouver, BC, Canada*. IEEE; 2023. P. 18527–18536. https://doi.org/10.1109/CVPR52729.2023.01777
19. Wang Y., Li B., Zhang G., et al. LRRU: long-short range recurrent updating networks for depth completion. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV), 01–06 October 2023, Paris, France*. IEEE; 2023. P. 9388–9398. https://doi.org/10.1109/ICCV51070.2023.00864
20. Tang J., Tian F.-P., An B., et al. Bilateral propagation network for depth completion. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16–22 June 2024, Seattle, WA, USA*. IEEE; 2024. P. 9763–9772. https://doi.org/10.1109/CVPR52733.2024.00932
21. Park J., Li Y.-J., Kitani K. Flexible depth completion for sparse and varying point densities. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16–22 June 2024, Seattle, WA, USA*. IEEE; 2024. P. 21540–21550. https://doi.org/10.1109/CVPR52733.2024.02035
22. Yan Zh., Lin Y., Wang K., et al. Tri-perspective view decomposition for geometry-aware depth completion. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 16–22 June 2024, Seattle, WA, USA*. IEEE; 2024. P. 4874–4884. https://doi.org/10.1109/CVPR52733.2024.00466
23. Xiang J., Zhu X., Wang X., et al. DEPTHOR: depth enhancement from a practical light-weight dToF sensor and RGB image. In: *2025 IEEE/CVF International Conference on Computer Vision (ICCV), 19–25 October 2025, Honolulu, HI, USA*. IEEE; 2025. P. 6101–6111. https://doi.org/10.1109/ICCV51701.2025.00576
24. Kim H., Wang R., Yao Ch., et al. Dense metric depth completion from sparse direct time-of-flight sensors. In: *2026 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 03–07 June 2026, Denver, CO, USA*. IEEE; 2026. P. 36518–36528.
25. Li Y., Lou L., Tang Y., et al. LiteSense: lifting lightweight ToF with RGB for high-resolution metric depth estimation. In: *2026 IEEE/CVF Conference on Computer Vision*

and Pattern Recognition (CVPR), 03–07 June 2026, Denver, CO, USA. IEEE; 2026. P. 5783–5792.

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Сяохань Юй, аспирант, Российский университет дружбы народов имени Патриса Лумумбы, Москва, Российская Федерация.
e-mail: yxhpro@yandex.ru

Xiaohan Yu, Postgraduate, Peoples' Friendship University of Russia named after Patrice Lumumba, Moscow, the Russian Federation.

Статья поступила в редакцию 10.05.2026; одобрена после рецензирования 27.07.2026; принята к публикации 17.08.2026.

The article was submitted 10.05.2026; approved after reviewing 27.07.2026; accepted for publication 17.08.2026.