Towards a Framework for Benchmarking Dense 3D Displacement Estimation Approaches for Geomonitoring Using Long-Range TLS Data

Nicholas Meyer[1]*, Tomislav Medic[1], Andreas Wieser[1]

[1] Institute of Geodesy and Photogrammetry, ETH Zurich, 8093 Zurich, Switzerland - <first>.<last>@geod.baug.ethz.ch

Keywords: Terrestrial laser scanning, point clouds, optical flow, cross-correlation, salient feature tracking, deformation estimation.

Abstract

Accurate and spatially dense 3D displacement estimation can contribute to a better understanding of geomorphological processes, while long-range terrestrial laser scanning (LR-TLS) has emerged as a promising technique for generating such observations. However, selecting the most effective algorithms for dense 3D displacement estimation remains challenging due to the lack of benchmarking. This study introduces an open and extensible benchmarking framework for 3D displacement estimation and provides an initial validation through a systematic comparison of representative 2D projection-based and 3D point cloud–based methods for estimating 3D displacements from LR-TLS scans. The evaluation includes 252 combinations of algorithmic and hyperparameter configurations, covering cross-correlation, optical flow, and salient feature tracking approaches, as well as the 3D displacement estimation method F2S3. All methods were benchmarked on a single common LR-TLS dataset, using sparse GNSS and manually derived displacements as ground truth. Results show that F2S3 achieves the highest agreement with the ground truth, while the top-performing configurations of the 2D approaches reach comparable accuracy, albeit slightly lower than that of F2S3. Our findings further highlight key sensitivities of current methods to parameter choices and data characteristics. The presented open and extensible evaluation framework enables reproducible performance assessment and could provide a foundation for future large-scale benchmarking and further development of 3D displacement estimation techniques for LR-TLS data.

1. Introduction

Policy making and risk-mitigation for geohazards require a deep understanding of geomechanical processes. Such understanding relies on multi-sensor observations and their use to calibrate and validate physics-based models for simulation and prediction that approximate real-world behavior (Casagli et al., 2017). There is a push towards developing geomonitoring strategies that provide observations of 3D displacements with high temporal and spatial resolution. Besides 3D vector fields of displacement estimates, obtaining their derivatives (velocities and accelerations) is of high relevance for correct process modeling, but also for early warning (Verruijt, 1995). However, numerical differentiation amplifies noise, reinforcing the need for high-accuracy, low-noise displacement estimates. Among the available geomonitoring sensors, terrestrial laser scanners (TLS) appear particularly well-positioned to meet these requirements. They can deliver dense, accurate 3D point clouds at reasonably high temporal resolution. In particular, repeated long-range terrestrial laser scanning (LR-TLS) from a single strategically chosen viewpoint offers a suitable monitoring strategy due to the possibility of covering large measurement volumes, a common requirement in geomonitoring (Lindenbergh et al., 2025).

Estimating geomaterial displacements by comparing point clouds from multiple measurement epochs has been extensively studied, as reported in the review by Abellan et al. (2016). Much research effort related to TLS- and, in particular, LR-TLS-based monitoring was invested in point cloud time series analysis, mostly analyzing spatial (3D) distribution of 1D displacement estimates (Lindenbergh et al., 2025). The most prominent example is M3C2 (Lague et al., 2013), which has become a standard for geomonitoring due to its robustness; however, in its base form, it is limited to estimating displacements along surface normals. Only a subset of the related literature analyzed the potential for 3D displacement estimates, out of which even a smaller subset investigated the methods of deriving dense 3D vector fields. However, based on the above reasons, 3D displacement vector fields are of high relevance for geohazard management. We thus focus on them, herein, considering displacement estimates to be *dense* if they tend towards representing a 3D displacement vector for each measured point in the point clouds.

The literature comprises two classes of methods that were used to estimate dense 3D displacement vector fields from LR-TLS data: *3D methods* that operate directly on the 3D point clouds, and *2D methods* that primarily operate on 2D image representations or projections of point clouds. Among the 3D methods, arguably the most relevant representatives are F2S3 (Gojcic et al., 2020), M3C2 with DMD (Williams et al., 2021) and Piecewise-ICP—first introduced as Piecewise Alignment Method (PAM) in Teza et al. (2007). F2S3, a deep learning-based approach that establishes point-wise correspondences in a high-dimensional feature space, and M3C2 with DMD, a heuristics-based approach that establishes point-wise correspondences in the Euclidean space, both natively provide dense 3D displacement estimates, converging towards the goal of per-point estimates for all points within the point cloud. Piecewise-ICP, a class of approaches with multiple realizations (e.g., also in Pfeiffer et al., 2018; Walicka et al., 2021), establishes correspondences per *point cloud patch*. This provides originally much sparser displacement estimates, but per-point estimates can be derived therefrom at high computational costs. All these 3D methods showed promising results on the data used to first present the algorithms in the literature. However, to the best of our knowledge, only F2S3 and Piecewise-ICP were successfully used in multiple geomonitoring studies beyond that.

* Corresponding author

The 2D methods differ in three main aspects: (i) the modality chosen for the image representation of the point clouds, (ii) the algorithm chosen for correspondence estimation, and (iii) the projection from 3D point cloud to 2D image. The most established image representations in TLS-based geomonitoring are hillshade images (Holst et al., 2021; Fey et al., 2015; Carle et al., 2023). Other modalities used in this context in the literature are (i) the norm of 2D gradients of the ranges (Travelletti et al., 2014), and (ii) laser backscatter intensities (Medic et al., 2023).

The correspondence search algorithms can be classified into (i) salient point (or region) feature detectors and descriptors (Holst et al., 2021; Hosseini et al., 2023), (ii) patch-wise image cross-correlation (Carle et al., 2023; Fey et al., 2015; Aati et al., 2022), and (iii) optical flow (Carle et al., 2023; Chanut et al., 2021). Within these categories, only the optical flow natively produces very dense (per-pixel) 3D displacement estimates; cross-correlation can produce them at high computational costs, and the salient feature-based algorithms unavoidably provide sparser estimates.

Regarding projection, a natural choice for LR-TLS with repeated scans from the same location, nearly-upright scanner, and no data close to the zenith or nadir is to use a plate carrée projection of the scanner’s spherical coordinate system, i.e., to directly use the horizontal and vertical angles of the point cloud as 2D image coordinates. Alternative projections used in the literature for point cloud–based geomonitoring are orthographic projections onto a horizontal or best-fit plane (Holst et al., 2021). Depending on the topography, point cloud acquisition process, required sensitivity and further aspects, many other projections—classical map projections onto developable surfaces, or projections onto more general surfaces and derivation of image coordinates from the parameter lines on those surfaces—are conceivable. In most cases, the projected 3D points will not coincide with a regular grid in the 2D image space, and the pixel values need to be obtained from the projected values by interpolation to the desired image resolution (grid size). For this step, all 2D-interpolation algorithms with or without smoothing can theoretically be applied, e.g., nearest-neighbor interpolation, cubic spline surfaces, or barycentric interpolation.

From all mentioned variants, so far only one was tested for LR-TLS scans in geomonitoring—the combination of range image gradients, cross-correlation and plate carée projection with linear interpolation (Travelletti et al., 2014). The other algorithms were either applied on point clouds of different origin, e.g., photogrammetric reconstruction, or in different monitoring applications (including lab-based simulations).

The presented methods have not yet been assessed in a systematic and comparative manner. Most of them were evaluated only in their original publications, sometimes on a single study site and with costly tuning of implementation- and site-specific settings. Consequently, it remains unclear how robustly they transfer to other geomonitoring sites with different characteristics, and how sensitive the results are to the parameter settings and implementation choices. In other words, we currently lack a comparison that goes beyond the original demonstrations and that reflects the realistic perspective of someone who wants to choose and apply available methods to a new scene without extensive, method-specific fine-tuning.

An important prerequisite for displacement estimation from multi-epoch TLS data is accurate co-registration of the point clouds. Even for repeated measurements from a fixed scanner position, residual misalignment may arise from atmospheric effects, instrumental influences, and scene changes, and can propagate into the subsequent displacement estimates. In addition, long-range TLS observations are affected by range-dependent noise and varying radiometric conditions, which can influence both 2D image-based and 3D point-based correspondence estimation. In this study, we do not aim to isolate or optimize these upstream effects individually; rather, we compare displacement estimation methods under a common preprocessing pipeline in order to assess their relative performance on the same LR-TLS data.

Ideally, a comparison would comprehensively benchmark all relevant methods on a diverse selection of datasets. As a first step towards that goal, in this study, we compare a set of publicly available 3D displacement estimation algorithms, on a single LR-TLS geomonitoring dataset. Within this scope, we conduct a comparatively broad evaluation of 2D projection–based methods and relate their performance to a single representative 3D method, F2S3, which presently stands as a state-of-the-art solution for dense per-point 3D displacement estimation.

To facilitate a transparent and repeatable benchmarking process, we have developed a modular Python-based framework designed for the systematic evaluation of 3D displacement estimation approaches. Rather than providing a static comparison, this framework enables automated hyperparameter sweeping and standardized performance assessment across diverse methods. In this work, we present an initial validation of the framework using a LR-TLS dataset and a selected subset of available 2D and 3D algorithms. By exploring a wide range of parameter combinations, we emulate the challenges faced by future users and provide a baseline for the relative performance, robustness, and usability of these approaches. This work represents a first step toward a community-driven benchmark, providing a code base that allows researchers to integrate and test additional datasets or emerging algorithms within a consistent pipeline.

2. Methods

This section introduces the algorithms employed in this study. The 2D (projection-based) and 3D methods are described in Sections 2.1 and 2.2, respectively. The evaluation metrics are presented in Section 2.3.

2.1 2D methods

The 2D algorithms investigated within this paper share the same core structure. We derive 3D displacement estimates from point cloud data of two epochs by (i) projecting each cloud into a 2D representation, (ii) rasterizing the projected points into a set of (feature) images, (iii) computing (dense) 2D correspondences between the epochs in the selected feature images, and (iv) projecting the 2D correspondences back to 3D to obtain per-point displacement vectors. Each of these steps has various possible implementations.

Since this study focuses on LR-TLS from a single location of the scanner, we adopt a plate carrée projection, which preserves the full information content in this configuration. Alternative projection schemes may become advantageous for different acquisition setups (e.g., multi-station TLS, mobile mapping, or airborne laser scanning), but their comparative evaluation is

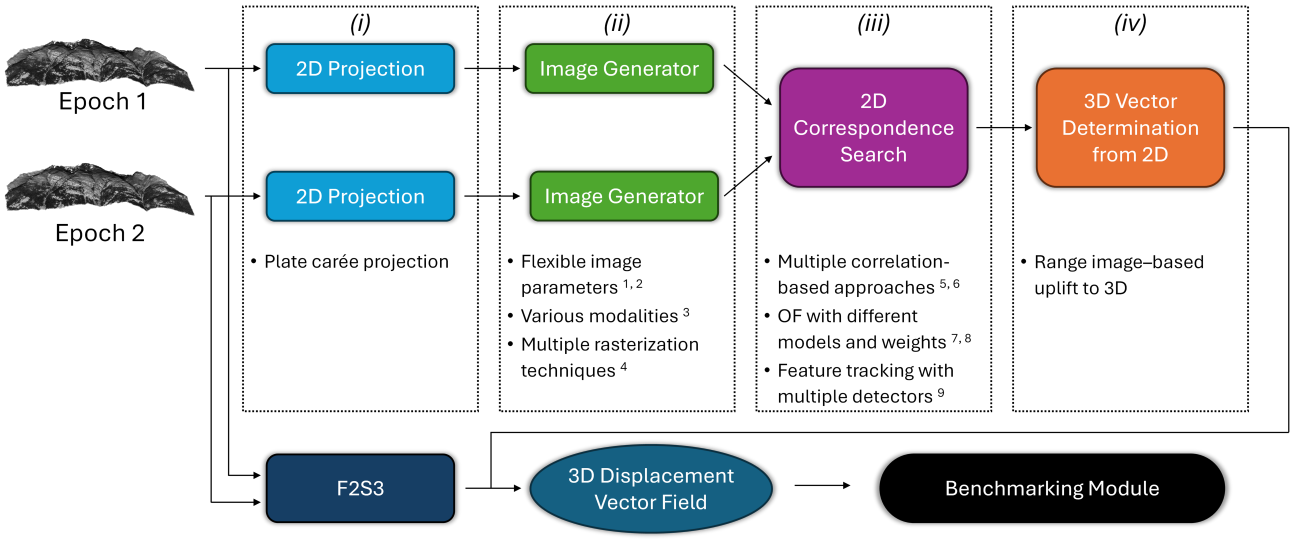

Figure 1. Method overview. Roman numerals *(i)–(iv)* denote the processing steps outlined in Section 2.1, while superscript numerals 1–9 correspond to the hyperparameters listed in Table 1.

beyond the scope of this work and is deferred to future research. We project several modalities in parallel, namely range, and intensity. Magnitude of gradients, hillshade, and an intensity/hillshade composite are generated as additional modalities subsequently from the 2D images created in the rasterization step.

For the rasterization step, we implemented nearest-neighbor, linear, cubic, and barycentric interpolation. The interpolation strategy defines how points are aggregated within image pixels and allows flexible control over parameters such as image size and effective ground sampling distance (or angular step size per pixel in the case of the chosen projection). Based on the selected image size and step size, the 2D projected points are tiled into non-overlapping regions. The angular step size and the image size form part of the hyperparameter search space explored in this study.

For the 2D correspondence computation step, the modalities and the algorithm must be selected. Herein, we investigate only the use of a single modality at a time and treat the chosen modality as a hyper-parameter. We leave the investigation of cross-modal algorithms which use different modalities jointly (without requiring to pre-process them into one as in our intensity/hillshade composite) for future research. We cover all three approaches for correspondence estimation outlined in Section 1: cross-correlation, optical flow, and salient feature tracking. Cross-correlation methods are well established in photogrammetry and remote sensing, as the classical foundation for dense image matching. We implemented two variants within our framework: a simple normalized cross-correlation (NCC) baseline, and COSI-Corr (Leprince et al., 2007), a widely used cross-correlation algorithm originally developed for measuring ground displacements from optical satellite imagery. Both methods operate on the previously generated images using window-based matching to estimate integer or sub-pixel displacements. The window size serves as the primary hyperparameter and controls the local support area used for similarity computation. The NCC implementation serves as a baseline for evaluating relative performance and parameter sensitivity within our framework. COSI-Corr extends the basic approach through a multi-scale coarse-to-fine scheme, defined by the number of pyramid levels, and by employing a configurable step size that determines the spatial sampling interval of correlation windows. These parameters influence the trade-off between computational cost, matching robustness, and spatial resolution and are included in the hyperparameter exploration conducted in this study.

Optical flow methods were evaluated to extend the analysis toward dense, learning-based correspondence estimation. The selected models represent a balanced spectrum from well-established baselines to recent transformer-based state-of-the-art approaches, allowing us to assess both absolute performance and generalization behavior across architectural paradigms. RAFT and its lightweight variant RAFT-small (Teed and Deng, 2020) serve as strong baselines and established reference methods in the computer vision community, employing iterative refinement and dense correlation volumes to achieve accurate flow estimation. FlowFormer++ (Shi et al., 2023) was included as a representative state-of-the-art architecture, introducing transformer-based global context modeling that significantly improves accuracy on complex and large-displacement motion. GMFlow (Xu et al., 2022) adopts a similar attention-driven formulation but prioritizes computational efficiency, making it suitable for evaluating performance–speed trade-offs. CRAFT (Sui et al., 2022) combines convolutional and attentional mechanisms to hierarchically refine flow across scales, while MaskFlowNet (Zhao et al., 2020) incorporates explicit occlusion reasoning to handle visibility changes between epochs.

All models were integrated through the *ptlflow* framework (v0.4.1, Morimitsu, 2021), which provides standardized access to a wide variety of machine learning–based optical flow implementations and (multiple) pre-trained weights for each model from established benchmark datasets. In our experimental design, each combination of model architecture and pretrained checkpoint is considered a distinct configuration within the broader hyperparameter sweep, allowing us to assess performance variation across different correspondence estimation strategies systematically. Classical optical flow meth-

ods such as Horn–Schunck (Horn and Schunck, 1981) and Lucas–Kanade (Lucas and Kanade, 1981) were not implemented explicitly, as their optimization principles are effectively embedded within these modern learning-based formulations. RAFT, for example, serves as a learned analogue of Horn–Schunck's variational refinement, while CRAFT parallels the pyramidal Lucas–Kanade approach in a learned feature space. Transformer-based models such as FlowFormer++ and GMFlow further generalize these formulations through global correspondence optimization in attention space. Together, these architectures span the major design families in contemporary optical flow research, enabling a comprehensive evaluation of their suitability for 3D displacement estimation from rasterized point cloud features.

Finally, salient feature tracking represents a sparse matching strategy that complements the dense methods described above. We evaluated the common scale- and rotation-invariant local feature detectors/descriptors SIFT (Lowe, 2004) and KAZE (Alcantarilla et al., 2012), to identify repeatable keypoints across epochs. Correspondences were established using nearest-neighbor matching in descriptor space, followed by Lowe's ratio test to suppress ambiguous matches. The ratio threshold was treated as a hyperparameter, allowing us to explore the trade-off between match density and reliability. Although inherently sparse, these correspondences provide valuable benchmarks, as the selected descriptors are agnostic to global positioning and are expected to be better suited for capturing large or locally discontinuous displacements than the correlation and optical flow approaches.

Table 1. Overview of all hyperparameters included in the evaluation, together with the value sets sampled during the parameter sweep. Roman numerals *(i)–(iv)* denote the processing steps outlined in Section 2.1, while superscript numerals 1–9 correspond to algorithmic overview in Figure 1.

Image generation *(ii)*	
Image size[1]	600x600, 900x600, 1000x500, 1200x600, 600x800
Ang. step size[2]	3 mgon, 6 mgon, 12 mgon
Modality[3]	magnitude of gradients, hillshade, intensity, intensity/hillshade composite
Interp. meth.[4]	linear, nearest neighbor, cubic, barycentric
Cross-Correlation *(iii)*	
NCC[5]	
Window size	15, 35, 75, 115, 145
COSI-Corr[6]	
Window size	32, 64, 128
Levels	1, 2, 3
Step size	1, 4, 8, 16
Optical flow *(iii)*	
Model[7]	RAFT, RAFT small, FlowFormer++, CRAFT, GMFlow, MaskFlownet
Pre-trained[8]	FlyingThings3D, MPI Sintel
Salient Feature Tracking *(iii)*	
Detector [9]	SIFT, KAZE
Lowe's Ratio	0.6, 0.75, 0.85
F2S3	
Outlier ref.	True, False
Voxel ds size	0 (no downsampling), 0.2, 0.4, 1, 2

The 2D correspondence estimation provides 2D shift vectors (u, v) for either every pixel (correlation-based, optical flow) or for selected ones (keypoints of salient feature tracking) of the epoch 1 image, indicating the corresponding location in the epoch 2 image. These are the 2D representations of the 3D displacement vectors. The reprojection step (iv) lifts these vectors up to 3D. It inverts the initial 2D projection and thus depends on it directly. However, the pixels in the images do not correspond bijectively to the related point clouds and thus, similar to the rasterization after projection from 3D to 2D, many ways are conceivable to lift the 2D information up to 3D, now. Herein, we use the range image of the first epoch to lift the pixels from the image of the first epoch back into 3D space and thus obtain the starting points of the 3D vectors. Using the epoch 2 range image, we interpolate the range for the end points of the 3D vectors. Together with the angular offsets, derived from (u, v) and the angular step size, this yields the 3D coordinates of the end points. So far, this reprojection yields vectors starting and ending at points which are not part of the original point clouds. So, we finally interpolate the displacement vectors in 3D starting at the points of the original epoch 1 point cloud. For the dense correspondence methods, we assign the vector of the respective nearest starting point. The same vector could thus be assigned to several points. In contrast, for the sparse salient feature matches, we associate each lifted vector to the nearest neighbor of the lifted start point. Each lifted vector is thus only assigned to a single point in the original point cloud, ensuring a one-to-one assignment that reflects the limited spatial density of the 2D feature estimates.

2.2 3D methods

While the geomonitoring community frequently employs methods like M3C2 and its derivatives, they were excluded from this study either because they are limited to 1D displacement estimates along surface normals or because they currently lack a public open-source implementation. To provide a representative 3D baseline for the framework, F2S3 was selected as it natively generates 3D displacement vectors and provides a reproducible code base. Local geometric descriptors are computed for both point clouds, and for each point in the first epoch a nearest-neighbor search in the feature space of the descriptors identifies its most similar counterpart in the second, yielding an initial dense, albeit noisy, 3D displacement field. To improve reliability, F2S3 applies an outlier-removal step based on the assumption that motion within each local neighborhood can be approximated by a rigid body. Points whose displacements deviate from the learned RANSAC-like consensus are discarded, and neighborhoods without a valid consensus are excluded entirely. An optional refinement step then transforms all points within neighborhoods where a valid consensus was found using the rigid-body parameters estimated for that neighborhood. The change of coordinates modeled by the transformation corresponds to the estimated displacement vectors.

While F2S3 has demonstrated promising results, practical applications have also shown that its performance is sensitive to the geometric characteristics of the input point clouds. In particular, the algorithm couples the neighborhood size to the local point density, which generally yields stable results but also makes density a key control parameter. In this study, we therefore include both the voxel-grid downsampling size (as an optional pre-processing step) and the optional refinement step as hyperparameters.

2.3 Evaluation Metrics

We assess the quality of the generated 3D displacement estimates against a set of sparse reference values. They consist of (a) displacements extracted from highly accurate (sub-cm-accuracy) coordinate time series of permanent GNSS-stations, and (b) manually annotated reference displacements with an accuracy corresponding approximately to the point spacing. These were obtained by registering small patches (radius $<15\,\mathrm{m}$) between the two epochs using ICP. The estimated rigid transformation was then applied to the centroid of the epoch 1 patch, and the displacement between the transformed and original centroid provided the reference value (e.g. Raffl et al., 2019). For each point cloud region i corresponding to one reference estimate $\vec{g}_i$, a representative displacement vector $\vec{e}_i$ is obtained by computing the median of the x-, y-, and z-components of all displacement estimates within a patch of radius $15\,\mathrm{m}$ around the GNSS antenna or annotated patch center. The representative displacement vectors are compared to their corresponding reference vectors using three metrics: the relative vector error (RVE), the absolute magnitude error (AME), and the angle deviation (AD).

$$\mathrm{RVE}_i = \frac{|\vec{g}_i - \vec{e}_i|}{|\vec{g}_i|} \tag{1}$$

$$\mathrm{AME}_i = ||\vec{g}_i| - |\vec{e}_i|| \tag{2}$$

$$\mathrm{AD}_i = \arccos\left(\frac{\vec{g}_i \cdot \vec{e}_i}{|\vec{g}_i||\vec{e}_i|}\right) \tag{3}$$

For all three metrics, smaller values indicate better agreement between estimated and reference displacements. RVE captures both magnitude and directional deviation, and its normalization by the ground-truth magnitude allows for meaningful comparison across regions with different reference displacement magnitudes. AME quantifies the absolute difference of the displacement magnitudes, and AD measures the angular deviation between estimated and reference displacement vectors. While RVE provides a balanced, scale-independent assessment, AME and AD are generally more intuitive to interpret, as they directly reflect metric and angular discrepancies, respectively. Each algorithm and hyperparameter configuration is scored using these metrics by calculating the respective median $\mathrm{RVE}_{\mathrm{med}}$, $\mathrm{AME}_{\mathrm{med}}$ and $\mathrm{AD}_{\mathrm{med}}$ over all reference patches $\mathrm{i} = 1 \ldots N$. Subsequently, we assess specific groups of algorithms and choices, e.g., to compare the performance achieved using different modalities or difference correspondence algorithms, by studying the distribution of these medians.

3. Data Set

For this study, we selected two LR-TLS scans acquired, about two years apart, during previous research on rock glacier dynamics in the Mattertal valley, Switzerland. The region of interest spans an area of about $3\,\mathrm{km}$ by $2\,\mathrm{km}$ and heights from $1500\,\mathrm{m}$ to $3400\,\mathrm{m}$ above sea level. The lower third of the area is characterized by vegetation consisting primarily of bushes and trees, whereas the upper slopes are dominated by coarse rock debris. The scans were acquired using a Riegl VZ-4000 terrestrial laser scanner from a single, fixed position on the opposite side of the valley resulting in distances between 1900 and $3800\,\mathrm{m}$. The scans were taken with an angular sampling of about $5.6\,\mathrm{mgon}$ ($0.09\,\mathrm{mrad}$). Further details on the study area and data acquisition are provided in Medic et al. (2024).

We co-registered the two scans using ICP. We acknowledge that co-registration is a critical component of point cloud–based displacement analysis and that, even for repeated scans from a fixed installation, residual misalignment may arise due to atmospheric influences, scanner characteristics, and actual scene changes. In applications focused on absolute displacement recovery, deliberate masking or otherwise excluding potentially unstable areas prior to registration would be preferable. In the present study, however, our primary aim is to compare the relative performance of different displacement estimation methods under a common preprocessing workflow. We therefore applied the same co-registration strategy to all evaluated methods. Residual registration errors may still affect the absolute agreement with the GNSS and manually derived references and should be considered when interpreting the reported error magnitudes.

For the selected epochs five permanent GNSS stations located within the region of interest provide independent reference displacement measurements. A previous analysis by Moeller et al. (2023) reported standard deviations of approximately $2\,\mathrm{mm}$ to $4\,\mathrm{mm}$ for the GNSS displacement estimates, confirming their suitability as high-precision reference data. In addition to the GNSS measurements, we manually annotated eight well distributed regions of interest as described in Section 2.3. Together, these provide 13 reference patches, each associated with a single 3D displacement vector. The magnitudes of these reference vectors range from $0.3\,\mathrm{m}$ to $9.5\,\mathrm{m}$.

The reference locations are not randomly distributed across the scene. GNSS stations are installed at instrumented sites with good visibility, and manually annotated patches were selected where reliable registration and feature identification were feasible. Consequently, the reference data are biased toward well-observable areas with favorable geometric and radiometric characteristics, while regions with low texture, vegetation, or complex geometry are underrepresented. The reported performance should therefore be interpreted as reflecting method behavior primarily in these well-observable parts of the scene.

4. Results

We evaluated 252 distinct combinations of the methods and hyperparameters listed in Table 1. For computational reasons, the search for combinations was not exhaustive. Instead, we randomly sampled (without replacement) from the hyperparameter choices given in the table. We chose these values as sensible based on our understanding of the implemented methods and the tackled geomonitoring task. We assume that this realistically approximates the perspective of a practitioner who wants to choose and apply available methods to a new scene without extensive prior knowledge. Due to differences in the number of tunable parameters across methods, more hyperparameter configurations were evaluated for some algorithms than for others; the differences in numbers of combinations (see e.g., Table 2) thus do not reflect an intentional weighting. For brevity, we report in this section only the impact of main algorithmic choices and hyperparameters that had a notable influence on the metrics described in Section 2.3.

Figure 2 illustrates the distribution of the $\mathrm{RVE}_{\mathrm{med}}$ scores (based on Eq. 1), for the four investigated categories of correspondence estimation methods: correlation, optical flow, salient

Table 2. $\mathrm{RVE_{med}}$ scores across performance percentiles grouped by the primary correspondence estimation methods, along with the $\mathrm{AME_{med}}$ and $\mathrm{AD_{med}}$ between estimated and ground-truth displacement vectors (underline denotes best per percentile).

		$\mathbf{RVE_{med}}$					$\mathbf{AME_{med}}$ [cm]			$\mathbf{AD_{med}}$ [deg]		
Method	**#**	**5%**	**25%**	**50%**	**75%**	**95%**	**Top**	**Median**	**MAD**	**Top**	**Median**	**MAD**
F2S3	7	0.19	0.20	0.21	0.22	0.31	10	11	0.5	6.2	6.8	0.12
Salient Features	18	0.25	0.34	0.52	3.57	99.26	9	19	9.3	6.2	20.3	9.82
Correlation	151	0.30	0.41	0.53	0.65	0.83	11	34	16.4	7.6	13.0	3.81
Optical Flow	75	0.28	0.41	0.56	0.66	0.78	16	39	11.8	8.7	14.1	4.17

features, and F2S3. As shown, F2S3 consistently exhibits the lowest error values, while correlation, optical flow, and salient features display similar distributions in the lower to mid percentiles, a similar performance floor, and median error scores approximately twice those of F2S3. Table 2 complements this visual scoring based on $\mathrm{RVE_{med}}$ values by indicating also the distribution of $\mathrm{AME_{med}}$, and $\mathrm{AD_{med}}$; concretely, the 5th and 50th percentiles are shown as representative for the top- and mid-performing method-hyperparameter combinations, and the MAD of the $\mathrm{AME_{med}}$ and $\mathrm{AD_{med}}$ as an additional quantification of the spread.

F2S3 achieves the highest overall displacement estimation accuracy across all evaluated metrics. It also exhibits the lowest sensitivity to hyperparameter choices, likely due to its comparatively small set of tunable parameters. The 2D correspondence estimation methods, by contrast, yield lower accuracy on average and a noticeably higher variability in performance across hyperparameter configurations. When focusing on the top-performing configurations, the 2D methods reach broadly similar accuracy levels amongst each other, although their relative ranking varies depending on whether the evaluation is based on relative ($\mathrm{RVE_{med}}$) or absolute ($\mathrm{AME_{med}}$ and $\mathrm{AD_{med}}$) metrics. Drawing firm conclusions about these differences in metrics would require a more detailed analysis at the level of individual reference patches, which lies beyond the scope of this study. Compared to the mid-performing configurations, however, the top-performing 2D methods lie much closer to F2S3, narrowing the overall accuracy gap between 2D and 3D approaches.

While Figures 2 and Table 2 present a statistical comparison across all method–hyperparameter configurations, aggregated per correspondence estimation category and based on the single representative estimated displacement vector for each patch, Figure 3 illustrates the spatial distribution of RVE scores for all per-point displacement estimates within one reference patch,

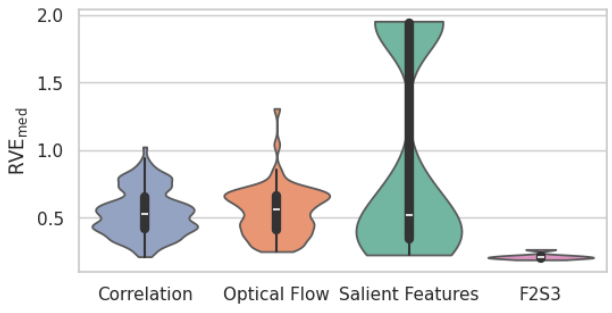

Figure 2. Distribution of $\mathrm{RVE_{med}}$ scores for all evaluated configurations, grouped by the primary correspondence estimation method (correlation, optical flow, salient feature, and F2S3 approaches). For visibility the values are clipped based on $k = 6$ times the MAD values.

shown for the top-performing configuration of each method category.

Across all four method categories, the resulting displacement estimates for the selected patch are generally in good agreement, particularly for the central and visually distinct boulder structure. Both the optical flow and correlation-based approaches yield high estimate densities, achieving near-complete coverage of the epoch 1 points, since they do not apply any explicit outlier identification or removal. F2S3 produces a similarly dense field of displacement estimates within the central structure but attains a substantially lower overall coverage (approximately 30%) over the entire patch due to its explicit outlier rejection, which retains only correspondences deemed reliable. In contrast, the salient feature approach yields a markedly sparse set of estimates with only seven estimates over the entire patch containing 7126 points, reflecting its reliance on distinct keypoints rather than dense correspondence estimation. Visually, the optical flow results appear smoother than those from cross-correlation, though it remains unclear whether this reflects an inherent smoothing effect of the method or genuinely reduced noise in the estimates. A more conclusive assessment of these effects would require access to dense ground-truth reference data, which is currently not available.

Table 3 provides a detailed overview of the error values for the different implementations of the correlation methods, salient feature descriptors, and optical flow models. For the correlation-based approaches, NCC slightly outperforms COSI-Corr across all performance brackets, albeit at a substantially higher computational cost. For the salient feature meth-

Table 3. $\mathrm{RVE_{med}}$ scores for the different model and algorithm variants within the 2D correspondence methods, shown across performance percentiles. Results are grouped by correlation-based, salient feature, and optical flow approaches with underline denoting best in group per percentile and **bold** marking overall best.

		$\mathbf{RVE_{med}}$				
Method	**#**	**5%**	**25%**	**50%**	**75%**	**95%**
NCC	35	**0.24**	0.33	0.42	0.51	0.73
COSI-Corr	116	0.33	0.45	0.56	0.71	0.84
KAZE	9	0.28	0.34	0.50	1.89	8.28
SIFT	9	**0.24**	0.47	0.55	6.53	311
MaskFlownet	9	0.26	**0.28**	**0.35**	**0.43**	3.27
FlowFormer++	15	0.31	0.43	0.49	0.70	0.88
CRAFT	21	0.28	0.46	0.57	0.64	0.71
RAFT	14	0.34	0.49	0.59	0.66	**0.69**
RAFT small	5	0.37	0.42	0.64	0.75	0.77
GMFlow	11	0.43	0.52	0.64	0.68	0.78

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-1-2026-381-2026 | © Author(s) 2026. CC BY 4.0 License.

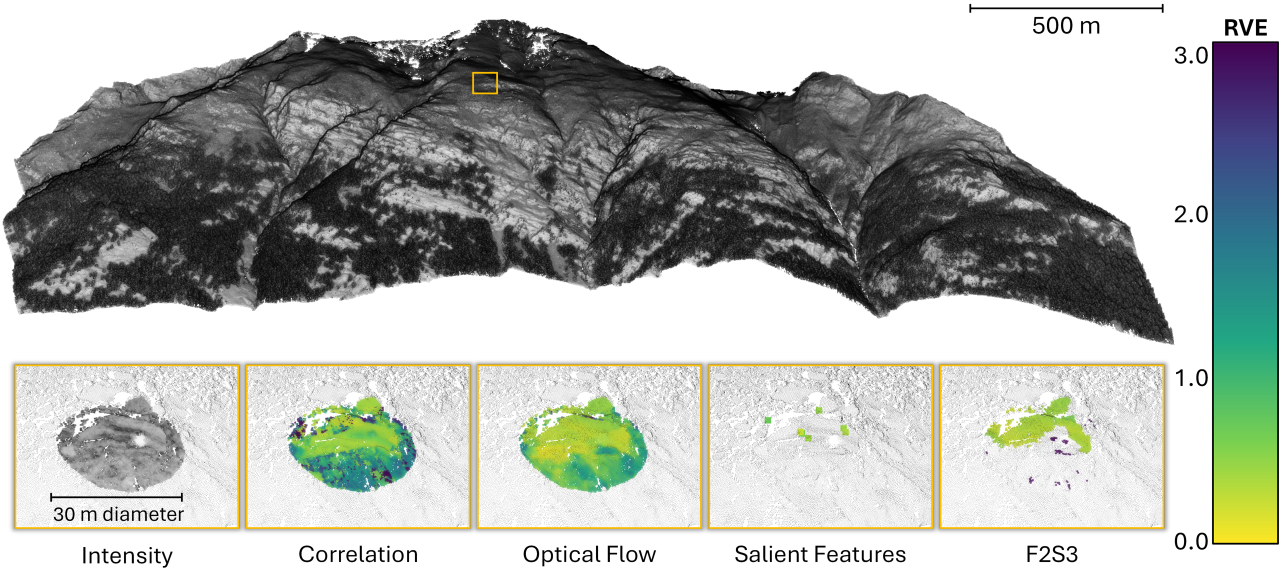

Figure 3. Point cloud of the study area with five overlaid zoom-ins showing one reference patch. The insets visualize point colors by intensity and by RVE obtained from the best performing (lowest $\mathrm{RVE}_{\mathrm{med}}$) configurations per correlation, optical flow, salient feature, and F2S3 method.

ods, KAZE and SIFT achieve comparable results in the top- to mid-performing configurations, although KAZE shows a tendency to be more robust to hyperparameter tuning. Among the optical flow models, MaskFlowNet outperforms the other architectures in the top- to mid-performing configurations but also produces the poorest results among the lowest-performing configurations. Overall, the choice of optical flow model has a pronounced influence on accuracy, underscoring the need for careful model selection and parameter tuning.

Finally, Table 4 summarizes the impact of the most influential image-generation parameters on the error scores. For the analyzed modalities, it is apparent that geometrically derived quantities yield better performance relative to the radiometric ones. The magnitude of the range-gradient vector (MRGV) achieves the best results overall, while hillshade and the intensity/hillshade composite (IHcomp) perform similarly. Intensity alone performs slightly worse than the other three options. This is likely explained by the higher variability of radiometric values between epochs, caused by changing atmospheric or environmental conditions. Common image pre-processing steps such as histogram matching might mitigate these effects but were outside the scope of the present study.

Table 4. $\mathrm{RVE}_{\mathrm{med}}$ scores across performance percentiles grouped for different image modalities, angular step sizes, and rasterization interpolation schemes with underline denoting best per hyperparameter per percentile.

		$\mathbf{RVE_{med}}$				
Parameter	**#**	**5%**	**25%**	**50%**	**75%**	**95%**
MRGV	26	0.24	0.35	0.48	0.60	0.73
Hillshade	78	0.28	0.38	0.50	0.64	0.80
IHcomp	68	0.28	0.41	0.53	0.68	0.89
Intensity	72	0.33	0.46	0.60	0.72	0.87
3 mgon	80	0.26	0.37	0.45	0.56	0.74
6 mgon	74	0.29	0.36	0.48	0.62	0.74
12 mgon	90	0.40	0.55	0.68	0.78	0.89
Linear	36	0.28	0.36	0.47	0.59	0.76
NN	26	0.30	0.38	0.50	0.63	0.73
Barycentric	156	0.28	0.42	0.53	0.69	0.86
Cubic	26	0.43	0.50	0.64	0.73	4.84

Regarding angular step size, both 3 and 6 mgon settings yield similar results, whereas 12 mgon underperforms. In relation to the original scan resolution, 6 mgon corresponds approximately to a 1:1 mapping between point in the point cloud and pixel, while 3 mgon and 12 mgon represent a doubling and halving of the effective image-to-scan resolution, respectively. This indicates that the potential accuracy gains from upsampling are small, while the computational advantages of downsampling come at a clear cost to the quality of the displacement estimates.

When comparing the different rasterization interpolation schemes, the choice among linear, nearest-neighbor (NN), and barycentric interpolation shows little influence on the results, whereas cubic interpolation yields noticeably poorer performance, and, therefore, should be avoided. However, again, a more detailed analysis of the impact of these parameters on the per-point displacement vector estimates would unavoidably require dense reference estimates.

5. Discussion

A central insight is that 2D projection–based methods, despite their simplicity and computational efficiency, can approach the accuracy of state-of-the-art 3D correspondence estimation when appropriately tuned. Their best-performing configurations achieve errors only slightly higher than those of F2S3, and in some local settings may perform on par depending on the evaluation metric. This suggests that 2D approaches remain highly relevant, particularly when combined with the rapid

progress in computer vision research, including optical flow models, feature tracking, and transformer-based architectures.

The results also illustrate that performance variability is strongly tied to hyperparameter choices. Whereas F2S3 shows relatively stable behavior due to its small number of tunable parameters, 2D methods exhibit a broad range of outcomes, underscoring the importance of systematic parameter exploration rather than relying on defaults or manual tuning. Feature selection emerged as particularly influential: the geometric image representations such as MRGV provided markedly better stability and accuracy than the intensity image, which was more susceptible to scan-to-scan variability. Similarly, the angular pixel spacing affected the performance in a predictable manner, with moderate upsampling having only a small benefit and downsampling degrading accuracy.

Several limitations of the present study underscore the need for further research to advance reliable benchmarking of displacement estimation methods. The evaluation relied on a single LR-TLS dataset, which—although representative of many geomorphological monitoring scenarios—cannot capture the diversity of terrain types, surface roughness conditions, and displacement regimes encountered in practice. Likewise, accuracy was assessed using sparse reference measurements, which provide reliable but limited sampling of the true displacement field. Generating dense ground-truth remains challenging and will likely require specialized field campaigns or physically realistic simulation environments that approximate true TLS acquisition geometry and noise characteristics. Simulation frameworks such as HELIOS++ (Winiwarter et al., 2022) enable the generation of physically realistic TLS point clouds with fully controlled acquisition conditions and dense ground-truth displacements. This would allow systematic evaluation across a wider range of scenarios than is feasible with field data alone, while also helping to mitigate the bias toward well-observable regions inherent in manually collected reference data. Integrating such datasets would complement real-world benchmarking and enable more comprehensive analysis under controlled conditions.

Another limitation is the scope of the evaluated methods. While the selection covers the major families of 2D algorithms and includes a representative dense 3D estimator, a larger number of promising techniques exist in the literature. Expanding the comparison will require community-wide participation and, importantly, (open-source) availability of implementations with standardized interfaces and output formats. Such shared infrastructure would enable broader benchmarking efforts and ensure that new methods can be evaluated under consistent conditions.

Finally, integrating confidence and uncertainty quantification into the estimation process represents an important frontier. In the present study, uncertainties arise not only from the correspondence estimation itself, but also from upstream factors such as co-registration and the quality of long-range TLS data, including range-dependent noise, varying point density, and atmospheric influences. While we applied a consistent preprocessing pipeline across all methods to enable relative comparison, residual registration errors and data noise may still affect the absolute agreement with the reference measurements. Correlation-based methods already output confidence measures, but these were not incorporated into the present pipeline; optical flow and learning-based methods rarely provide them at all. Embedding uncertainty estimation directly into the algorithms—or combining it with consensus-based filtering strategies such as those used in F2S3—could substantially improve the interpretability and operational value of displacement estimates.

Overall, the results demonstrate both the strengths and the limitations of current methods and highlight the need for reproducible, comparable, and transparent evaluation practices within the geomonitoring community.

6. Conclusion

This work introduces a unified Python-based framework for the systematic evaluation of dense 3D displacement estimation methods from LR-TLS data. By enabling structured hyperparameter exploration, consistent scoring across heterogeneous algorithms, and reproducible experimental workflows, the framework provides a practical foundation for community-wide benchmarking efforts.

Using this tool, we compared several representative 2D and 3D correspondence estimation approaches on a real LR-TLS dataset. F2S3 showed the highest accuracy and the lowest sensitivity to hyperparameters, while top-performing configurations of 2D methods achieved comparable accuracy, illustrating their continued relevance and strong potential for future improvement. The analysis also identified key sensitivities to feature choice, angular pixel spacing, and model architecture.

The main contribution of this work lies in establishing a reproducible and extensible benchmark structure. Expanding the comparison to additional datasets, incorporating denser reference information, broadening the range of evaluated algorithms, and integrating uncertainty quantification will be important next steps. We hope that the provided framework will serve as a catalyst for coordinated community benchmarking and foster the development of more robust, reliable, and transferable displacement estimation tools for geomonitoring applications.

Acknowledgment

The data collection of the data sets used within this study was financially supported by the Swiss Federal Office of the Environment (BAFU, grant 20.0011.PJ /6DDBCAD6D).

Declaration of generative AI and AI-assisted technologies

During the preparation of this work, the authors used ChatGPT to assist in code development and to improve readability and language. All content generated using this tool was subsequently reviewed and edited by the authors, who take full responsibility for the final version of the publication.

Code Availability

The code for the methods and experiments presented in this work will be released as open source at `https://github.com/gseg-ethz/geodispbench3d`.

References

Aati, S., Milliner, C., Avouac, J.-P., 2022. A new approach for 2-D and 3-D precise measurements of ground deformation from optimized registration and correlation of optical images and ICA-based filtering of image geometry artifacts. *Remote Sensing of Environment*, 277, 113038.

Abellan, A., Derron, M.-H., Jaboyedoff, M., 2016. “Use of 3D Point Clouds in Geohazards” Special Issue: Current Challenges and Future Trends. *Remote Sensing*, 8(2), 130.

Alcantarilla, P. F., Bartoli, A., Davison, A. J., 2012. KAZE Features. A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, C. Schmid (eds), *Computer Vision – ECCV 2012*, Springer Berlin Heidelberg, Berlin, Heidelberg, 214–227.

Carle, E., Sirguey, P., Cox, S. C., 2023. Measuring landslide-driven ground displacements with high-resolution surface models and optical flow. *Computers & geosciences*, 178, 105378.

Casagli, N., Frodella, W., Morelli, S., Tofani, V., Ciampalini, A., Intrieri, E., Raspini, F., Rossi, G., Tanteri, L., Lu, P., 2017. Spaceborne, UAV and ground-based remote sensing techniques for landslide mapping, monitoring and early warning. *Geoenvironmental Disasters*, 4(1), 9.

Chanut, M.-A., Gasc-Barbier, M., Dubois, L., Carotte, A., 2021. Automatic identification of continuous or non-continuous evolution of landslides and quantification of deformations. *Landslides*, 18(9), 3101–3118.

Fey, C., Rutzinger, M., Wichmann, V., Prager, C., Bremer, M., Zangerl, C., 2015. Deriving 3D displacement vectors from multi-temporal airborne laser scanning data for landslide activity analyses. *GIScience & Remote Sensing*, 52(4), 437–461.

Gojcic, Z., Zhou, C., Wieser, A., 2020. F2S3: Robustified determination of 3D displacement vector fields using deep learning. *Journal of Applied Geodesy*, 14(2), 177–189.

Holst, C., Janßen, J., Schmitz, B., Blome, M., Dercks, M., Schoch-Baumann, A., Blöthe, J., Schrott, L., Kuhlmann, H., Medic, T., 2021. Increasing spatio-temporal resolution for monitoring alpine solifluction using terrestrial laser scanners and 3D vector fields. *Remote Sensing*, 13(6), 1192.

Horn, B. K., Schunck, B. G., 1981. Determining optical flow. *Artificial Intelligence*, 17(1-3), 185–203. https://linkinghub.elsevier.com/retrieve/pii/0004370281900242.

Hosseini, K., Reindl, L., Raffl, L., Wiedemann, W., Holst, C., 2023. 3D landslide monitoring in high spatial resolution by feature tracking and histogram analyses using laser scanners. *Remote Sensing*, 16(1), 138.

Lague, D., Brodu, N., Leroux, J., 2013. Accurate 3D comparison of complex topography with terrestrial laser scanner: Application to the Rangitikei canyon (N-Z). *ISPRS Journal of Photogrammetry and Remote Sensing*, 82, 10–26. https://linkinghub.elsevier.com/retrieve/pii/S0924271613001184.

Leprince, S., Barbot, S., Ayoub, F., Avouac, J.-P., 2007. Automatic and Precise Orthorectification, Coregistration, and Subpixel Correlation of Satellite Images, Application to Ground Deformation Measurements. *IEEE Transactions on Geoscience and Remote Sensing*, 45(6), 1529–1558. https://ieeexplore.ieee.org/document/4215064/.

Lindenbergh, R., Anders, K., Campos, M., Czerwonka-Schröder, D., Höfle, B., Kuschnerus, M., Puttonen, E., Prinz, R., Rutzinger, M., Voordendag, A. et al., 2025. Permanent terrestrial laser scanning for near-continuous environmental observations: Systems, methods, challenges and applications. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 100094.

Lowe, D. G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91–110.

Lucas, B. D., Kanade, T., 1981. An iterative image registration technique with an application to stereo vision. *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI’81, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 674–679.

Medic, T., Ruttner, P., Holst, C., Wieser, A., 2023. Keypoint-based deformation monitoring using a terrestrial laser scanner from a single station: case study of a bridge pier. *Proceedings of the 5th Joint International Symposium on Deformation Monitoring-JISDM 2022*, Editorial de la Universitat Politècnica de València, 167–175.

Medic, T., Shi, N., Meyer, N., Wieser, A., 2024. Geomonitoring Using Long-Range TLS in Swiss Alps. *SIG 2024 - Proceedings of the International Symposium on Engineering Geodesy*, Croatian Geodetic Society, Zagreb, Croatia.

Moeller, G., Medic, T., Aichinger-Rosenberger, M., Schmid, L., Wieser, A., Rothacher, M., 2023. Alpine Metrology Lab: Geomonitoring Using Long- Range TLS and Permanent GNSS. *Ingenieurvermessung 23: Beiträge zum 20. Internationalen Ingenieurvermessungskurs Zürich*, Wichmann, Zurich, Switzerland.

Morimitsu, H., 2021. Pytorch lightning optical flow. `https://github.com/hmorimitsu/ptlflow`.

Pfeiffer, J., Zieher, T., Bremer, M., Wichmann, V., Rutzinger, M., 2018. Derivation of three-dimensional displacement vectors from multi-temporal long-range terrestrial laser scanning at the Reissenschuh landslide (Tyrol, Austria). *Remote Sensing*, 10(11), 1688.

Raffl, L., Wiedemann, W., Wunderlich, T., 2019. Non-signalized Structural Monitoring using Scanning Total Stations. *Proceedings of the 4th Joint International Symposium on Deformation Monitoring (JISDM), Athens, Greece*.

Shi, X., Huang, Z., Li, D., Zhang, M., Cheung, K. C., See, S., Qin, H., Dai, J., Li, H., 2023. FlowFormer++: Masked Cost Volume Autoencoding for Pretraining Optical Flow Estimation. arXiv:2303.01237 [cs].

Sui, X., Li, S., Geng, X., Wu, Y., Xu, X., Liu, Y., Goh, R., Zhu, H., 2022. CRAFT: Cross-Attentional Flow Transformer for Robust Optical Flow. arXiv:2203.16896 [cs].

Teed, Z., Deng, J., 2020. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. arXiv:2003.12039 [cs].

Teza, G., Galgaro, A., Zaltron, N., Genevois, R., 2007. Terrestrial laser scanner to detect landslide displacement fields: a new approach. *International Journal of Remote Sensing*, 28(16), 3425–3446.

Travelletti, J., Malet, J.-P., Delacourt, C., 2014. Image-based correlation of Laser Scanning point cloud time series for landslide monitoring. *International Journal of Applied Earth Observation and Geoinformation*, 32, 1–18.

Verruijt, A., 1995. *Computational geomechanics*. 7, Springer Science & Business Media.

Walicka, A., Pfeifer, N., Borkowski, A., Jóźków, G., 2021. An automatic method for the measurement of coarse particle movement in a mountain riverbed. *Measurement*, 174, 109029.

Williams, J. G., Anders, K., Winiwarter, L., Zahs, V., Höfle, B., 2021. Multi-directional change detection between point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 172, 95–113.

Winiwarter, L., Esmorís Pena, A. M., Weiser, H., Anders, K., Martínez Sánchez, J., Searle, M., Höfle, B., 2022. Virtual laser scanning with HELIOS++: A novel take on ray tracing-based simulation of topographic full-waveform 3D laser scanning. *Remote Sensing of Environment*, 269.

Xu, H., Zhang, J., Cai, J., Rezatofighi, H., Tao, D., 2022. GMFlow: Learning Optical Flow via Global Matching. arXiv:2111.13680 [cs].

Zhao, S., Sheng, Y., Dong, Y., Chang, E. I.-C., Xu, Y., 2020. MaskFlownet: Asymmetric Feature Matching with Learnable Occlusion Mask. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6277–6286. arXiv:2003.10955 [cs].