ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XI-2-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission II, 4–11 July 2026, Toronto, Canada

Evaluation of Recent AI-based Point Matching Algorithms Applied on Aerial Images

Pablo d’Angelo*, Franz Kurz, Alaa Eddine Ben Zekri, Reza Bahmanyar

Remote Sensing Technology Institute, German Aerospace Center (DLR)
Münchener Str. 20, 82234 Oberpfaffenhofen, Germany
{pablo.angelo, franz.kurz, alaa.benzekri, reza.bahmanyar}@dlr.de

Keywords: Matching, Aerial imagery, Ground control point, Image orientation, Evaluation

Abstract

Accurate image matching is essential for the precise orientation of airborne imagery, yet modern feature matchers are rarely evaluated on real aerial data with great temporal, seasonal, and radiometric changes. For this study, we introduce the AerialRefMatch dataset, which comprises 51 challenging aerial images and corresponding true-ortho reference data. We benchmark classical and deep learning–based matching algorithms on AerialRefMatch, considering two scenarios: matching original images and matching approx-orthorectified images generated using GNSS/IMU orientations. For each method, image-based ground control points are derived and used for single-image pose estimation; accuracy is assessed via independent checkpoints. Results show that directly matching on original images is very difficult: fewer than 14% of images can be oriented with pixel-level accuracy. When approx-orthorectification is used, performance improves substantially. JamMa, SIFT, and SuperPoint+LightGlue achieve pixel-level accuracy for up to 30% of images, with JamMa being most robust on difficult cases and SIFT-based variants being more precise on the easier ones. Deep detector-free models such as ELoFTR and RoMa are less accurate but more robust to the original images than other models. Overall, state-of-the-art deep learning-based matchers still struggle with large rotations, scale differences, and semantic differences, and strongly benefit from prior image orientation knowledge and lack sub-pixel precision.
The AerialRefMatch dataset can be downloaded here: https://www.dlr.de/en/eoc/aerial-ref-match

1. Introduction

Image matching is central to remote sensing and supports a wide range of photogrammetric applications. In recent years, many new point-based image matching algorithms have been introduced, mostly in the computer vision domain. In this paper, we apply several of these methods to a challenging aerial image dataset and evaluate their performance in a photogrammetric context. Examining these methods on real-world aerial data with challenging temporal and appearance changes highlights their practical applicability and limitations, offering guidance for future photogrammetric processing pipelines.

The main use case we considered is the estimation of precise image orientation, including interior and exterior camera parameters, for images acquired with high-resolution, airborne camera systems equipped with high-quality GNSS/IMU units. Accurate image orientation is essential for producing orthorectified imagery and for precisely geolocating semantic information extracted from original, non-rectified images. It is also necessary for many other downstream photogrammetric tasks.

Most approaches for image orientation utilize tie points between images and ground control points connecting an image point (x, y) with a 3D point (X, Y, Z) in object space for precise absolute orientation. For products requiring high absolute accuracy, for example, orthophotos from official mapping agencies, ground control point (GCP) coordinates are usually measured using differential GPS on the ground and manually matched with the corresponding image coordinates. While this procedure provides high quality GCPs, it requires manual interaction, which is less suitable for some applications, for example, require frequent monitoring, or areas where the ground cannot be accessed. We thus focus on using reference ortho images instead of GPS measurements as a source for ground control. Our paper evaluates different feature-based point matching algorithms for accurately matching aerial images to reference images. At first glance, this task may seem simpler than the wide-baseline matching problems typically addressed in close-range photogrammetry and computer vision. However, it can still be challenging due to differences in time and appearance or partial scene changes, even when high-quality prior image orientations are provided by sophisticated GNSS/IMU systems.

For our experiments, we evaluate classical SIFT-based matching and several recent deep learning approaches, including SuperPoint+LightGlue (Sarlin et al., 2020), SE2LoFTR (Bökman and Kahl, 2022), RoMa (Edstedt et al., 2024), MatchAnything variants (He et al., 2025), and JamMa (Lu and Du, 2025) using their publicly released pretrained models and standardized inference settings. For each method, correspondences are estimated and subsequently filtered using RANSAC (Fischler and Bolles, 1981). We evaluated the methods on two tasks: one with original aerial images and the other with approx-orthorectified aerial images using given GNSS/IMU information, where scale and rotational differences were removed.

1.1 Related works

Image matching supports a wide range of mapping and monitoring applications (Ma et al., 2019). One of the core tasks is image registration, which involves aligning multi-temporal or multisensor satellite and aerial images. This allows for consistent analysis over time or across modalities. A closely related task is change detection, which involves comparing observations taken at different times to identify urban growth, environmental alterations, or the impact of disasters (Cheng et al., 2024). Mosaicking and orthorectification involve stitching together overlapping aerial photographs to create seamless,

* Corresponding author: Pablo.Angelo@dlr.de

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-2-2026-1-2026

georeferenced orthomaps suitable for large-area mapping. In three-dimensional applications, stereo-based terrain reconstruction derives elevation models and surface structures that are essential for topographic and urban analyses (Aguilar et al., 2022). Georeferencing and co-registration link imagery to precise geographic coordinates, ensuring positional accuracy for Geographic Information System (GIS) integration, map correction, and autonomous navigation. Cross-view matching extends these efforts by matching ground-level or oblique perspectives with overhead imagery, such as aerial or satellite images, for localization, urban modeling, and joint scene interpretation (Durgam et al., 2024).

1.1.1 Classical approaches Classical image matching methods for remote sensing rely on handcrafted features and geometric consistency to establish correspondences between aerial or satellite images. Approaches such as Scale-Invariant Feature Transform (SIFT) (Lowe, 2004), Speeded-Up Robust Features (SURF) (Bay et al., 2006), Affine SIFT (SIFT) (Morel and Yu, 2009), and Binary Robust Invariant Scalable Keypoints (BRISK) (Leutenegger et al., 2011) first extract keypoints that are distinctive and remain stable under scale, rotation, and moderate viewpoint changes. These keypoints are then matched across image pairs using similarity measures, such as Euclidean or Hamming distances. These measures are often accelerated using algorithms like Fast Library for Approximate Nearest Neighbors (FLANN) (Muja and Lowe, 2009). Then geometric refinement is performed using methods such as Random Sample Consensus (RANSAC) (Fischler and Bolles, 1981), which estimates transformations while rejecting outlier correspondences, and Least Squares Matching (LSM) (Gruen, 1985), which aligns images with sub-pixel accuracy. These techniques are widely used for image registration, mosaicking, and change detection, providing reliable, interpretable results (Fan et al., 2019). However, their performance degrades under large viewpoint shifts, in low-texture areas, and in the presence of radiometric differences. These limitations have motivated the development of learning-based approaches that can automatically extract more robust and invariant representations (Chen et al., 2014, Ma et al., 2019).

1.1.2 Approaches based on deep learning Early feature-matching with Deep Learning (DL) aimed to jointly learn keypoint detection and description. SuperPoint (DeTone et al., 2017) introduced a self-supervised convolutional network that predicts both interest point locations and descriptors, creating a foundation for later models. SuperGlue (Sarlin et al., 2020) formulated matching as a differentiable optimization problem using a graph neural network with alternating self- and cross-attention layers to reason over keypoint context within and across images. LightGlue (Lindenberger et al., 2023) restructured the transformer-based approach of SuperGlue with adaptive inference and dynamic pruning, matching SuperGlue's accuracy while being substantially faster and more flexible across feature backbones. LoFTR (Sun et al., 2021) introduced a detector-free framework for dense correspondences using a coarse-to-fine transformer pipeline, enabling robust matching even in challenging regions. SE2LoFTR (Bökman and Kahl, 2022) added SE(2)-equivariance for better rotation and translation robustness. ELoFTR (Wang et al., 2024) improved efficiency and accuracy with aggregated attention and a two-stage correlation layer for sub-pixel matching. RoMa (Edstedt et al., 2024) advanced dense matching by combining DINOv2 (Oquab et al., 2023) features with ConvNet and a transformer-based match decoder for robust multimodal correspondences. Furthermore, MatchAnything (He et al., 2025) extended ELoFTR and RoMa with large-scale multi-domain training for universal, domain-agnostic matching across diverse image types and viewpoints. JamMa (Lu and Du, 2025) introduced an ultra-lightweight matcher using the Mamba state-space model to efficiently capture global dependencies between image pairs. FmCFA (Liao et al., 2025) proposed a multimodal framework with critical feature attention, improving correspondence accuracy across diverse imaging modalities.

2. Dataset

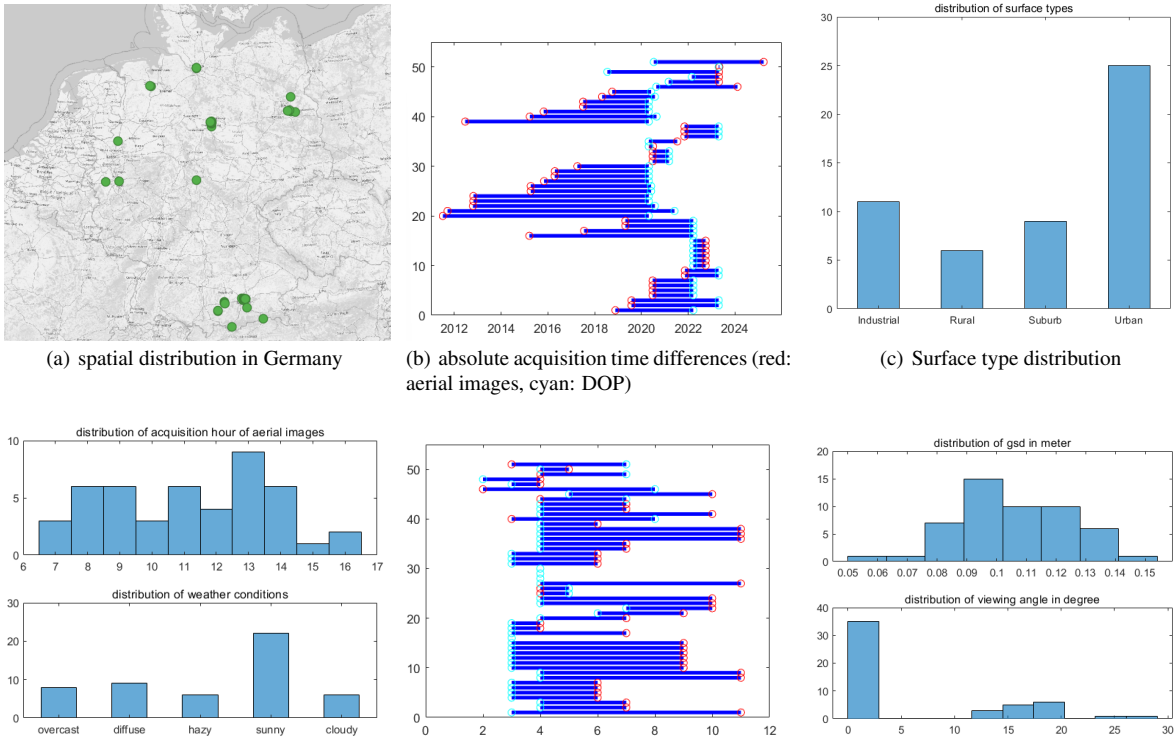
The AerialRefMatch dataset consists of 51 not-projected original images and 51 corresponding digital ortho images (DOP) overlaid with digital terrain/surface models (DTM/DSM). The aerial images were acquired with a maximum variety with respect to the illumination- and viewing configuration, the spatial distribution, the observed scenery, the flight heights and acquisition times. The images in the dataset were acquired over the German cities Berlin, Brunswick, Cologne, Garmisch-Partenkirchen, Hamburg, Landsberg, Kaufbeuren, Munich, Münster, Oldenburg, Wolfenbüttel and some rural areas outside of cities. The original images have Ground Sampling Distance (GSD) values ranging from 6–14 cm/pix, while the DOPs have GSD of 20 cm/pix. Individual image sizes range from 17 to 22 Mpx. In Figure 1, the properties of the original imagery with respect to positions, surface types, flight heights, GSDs, viewing angles, acquisition times, and weather conditions are visualized. The difference between the acquisition times of nearly all images and the DOPs is more than one year, which leads to differences in the images due to new buildings, renewed road markings, etc. The seasonal, weather-related and acquisition hour differences are also remarkable, as they cause variations in the images for example due to differences in shadows, leafy trees and traffic density. The viewing angle of the aerial images influences the number and size of occlusions caused by buildings and trees. These occlusions do not occur in DOPs because they are processed with a perfect nadir view (true ortho images). In summary, the dataset offers a wide range in the scenes depicted and is therefore a good basis for validating various matching methods.

In Figure 2, samples for image pairs (original image, DOP) are shown. For illustration, the black lines show the ground truth checkpoints either measured manually or derived via BRISK matching followed by a least-squares fine-matching.

2.1 Dataset components in the experiments

In our experiments, we consider several dataset components for different processing stages, including:

- Original (O): Non-projected RGB images as acquired by airborne camera systems. Each image has GNSS-based position and orientation information from onboard inertial systems. While the parameters of the internal orientation were previously determined by calibration, they can still be adapted to the respective recording situation through self-calibration. The original images contain various rotations as well as varying resolutions and slight perspective distortions.
- Digital Elevation Model (DEM): The coarse resolution (30 m/pix GSD) worldwide available Copernicus DEM (C et al., 2024) are used for the projection of original images to Approx-Ortho images A.



(a) spatial distribution in Germany

(b) absolute acquisition time differences (red: aerial images, cyan: DOP)

(c) Surface type distribution

(d) acquisition hour and weather distribution

(e) Seasonal acquisition time differences (red: aerial images, cyan: DOP)

(f) gsd and off-nadir viewing angle distribution

Figure 1. Statistics of the data set with number of images on the y-axis for histograms (c, d, f) and with image ids for time differences (b, e).

- Approx-Ortho (A): Approximately orthorectified images using approximate GNSS/IMU orientations and DEM. Distortions, rotations, and varying resolutions are removed by using the approximately orthorectified images, which are north-oriented, at the same scale and account for terrain effects. Thus, they make maximum use of the metadata available in the aerial survey use case, and should significantly ease the image matching problem.
- DOP (D): DOP images serve as reference and originate from the official surveying authorities and were taken at different points in time. They have a resolution of 20 cm/pix and are true-ortho, i.e. the aerial images were projected onto a DSM. On flat terrain, an accuracy of 20 cm is assumed. Each DOP has full overlay with the original images and was cropped according to the footprint of the aerial image, taking into account an additional frame of 100 meters in each direction. This extends the search space for the matching procedures, but is the same size for each image.
- Digital Surface Model (DSM): originate from the official surveying authorities and were acquired with LIDAR and reach 10 cm accuracy. DSMs are used for deriving the height Z for image-based ground control points (iGCP) at a position X and Y matched in D

2.2 Checkpoints

To evaluate the quality of the matched points and the image orientations derived from them, independent checkpoints are required. These were manually measured between the reference images and the original images. As these points must be of the highest quality and consistent over time and with the LiDAR DSM, only those on stable ground features and far from large discontinuities were measured. Care has been taken to measure points with sub-pixel accuracy, by using circular or intersection of linear features. This proved challenging, particularly for scenes that consisted almost entirely of vegetation. Due to time constraints, manual measurements were only performed for 21 images.

For the remaining images, the preferred method for tie points was BRISK (Leutenegger et al., 2011) matching followed by local least-squares fine matching, delivering sub-pixel accuracy. As this method failed when applied on original images, a workaround involving matching Approx-Ortho and reference images was applied. Starting from the tie points matched by BRISK between Approx-Ortho and DOP images, the coordinates in the original image are calculated by inverting the orthoprojection. These tie points between O and D are refined to sub-pixel accuracy using bidirectional least-squares matching, and points where fore/back matching results differ by 0.2 pixels are rejected as outliers.

3. Evaluation Method

3.1 Image matching

Image matching is used to extract tie points defined as

$$T = ((x^D, y^D), (x^O, y^O)) \tag{1}$$

between images D and O. We evaluate two matching scenarios, first only with GNSS information, and second with heavily GNSS/IMU information to ease that matching task.

Figure 2. Image samples of original O and DOP D showing big visual and contextual differences of the dataset caused by e.g. viewing differences (#1,#3,#4), seasonal differences (#5,#8), temporal decorrelation (#2,#6) and difficult surface types (#7).

3.1.1 Matching with only using GNSS information Here, we only use the GNSS information and focal length to extract the footprint of the reference image D. Other than that, no additional image or camera information is utilized and no scale correction is applied. To account for inaccuracies in orientation and positioning, a buffer of 100 m is taken into account when calculating the footprint. Tie points are extracted for each matching algorithm m. Outlier filtering is performed using MAGSAC (Barath et al., 2020) resulting in a set of inlier tie points T_m^O. Using the corresponding ground coordinates (X^D, Y^D) and the height retrieved from the DSM, iGCPs are obtained by

$$iGCP_m^O = ((X_i^D, Y_i^D, DSM(X_i^D, Y_i^D)), (x_i^O, y_i^O)). \quad (2)$$

3.1.2 Matching with using GNSS and IMU information All measured GNSS/IMU data is used to calculate Approx-Ortho images A, i.e. scale and rotational differences relative to D are removed. The remaining systematic difference between D and A is mostly a global x, y shift due to inaccuracies of the approximate orientations, which eases the image matching problem. It further allows the use of a simple row/column shift model during RANSAC, requiring only one tie point for model fitting. Instead of limiting the number of RANSAC iterations, we perform exhaustive testing of all points, with a fixed threshold of 20 pixels. The resulting tie points between DOP and Approx-Ortho images

$$T_D^A = ((x_i^D, y_i^D), (x_i^A, y_i^A)) \quad (3)$$

cannot be directly used for pose estimation as (x^A, y^A) are related to the Approx-Ortho image. The approx-orthorectification process is thus reversed by projecting T_D^A into the original image O using approximate GNSS/IMU image orientation K^*, the ground coordinates (X^A, Y^A) from the Approx-Ortho image and the height Z^A extracted from the DEM used during approximate orthorectification. Finally, we obtain iGCPs by

$$iGCP_m^A = ((X_i^D, Y_i^D, DSM(X_i^D, Y_i^D)), K^*(X_i^A, Y_i^A, Z_i^A)). \quad (4)$$

3.2 Image orientation

Image orientation parameters K_m of the original images are calculated using the iGCPs of each matching method m, which are fed into a pose estimation for each single image. Image bundles were not considered because images are widely distributed across Germany and camera setups differ. Neighbouring images are not considered at this stage, which would stabilise the image orientation step and increase the accuracy. For pose estimation, we applied least-squares optimisation with iterative outlier removal. First, the residues at the tie points are calculated and then filtered before a second round of optimisation is run. The a priori weights for the interior orientation were set relatively high to avoid additional distortion caused by outliers. Other a priori weights, like for the GCPs itself, are set equally for each matching method and image. Measurements for the projection centre position and image attitudes are introduced as additional observations in pose estimation.

Image orientation is then performed separately for each original image, resulting in refined calibration parameters (focal length, distortion parameters, principal point, boresight misalignments) and exterior orientations (position of the projection center, image attitude angles).

3.3 Evaluation metrics

Contrary to the general deep learning image matching comparisons, which mainly evaluate camera pose error or homography-based error (Lindenberger et al., 2023), we are most interested in good co-registration of the images O with the DOP reference imagery D, as this, for example, allows us to precisely geolocate semantic information extracted from the airborne imagery. This is similar to the evaluation done on some medical image datasets in MatchAnything (He et al., 2025), where homography or non-rigid warping is evaluated using ground truth checkpoints, but with much higher thresholds for error, i.e., thresholds for good matches are between 10 and 25 pixels, depending on the dataset.

We evaluate the matching algorithms for their absolute accuracy and precision, by computing a 2D ortho projection error e using image orientations K_m computed from the $iGCP$ derived using the different matching method m. First, object coordinates of the independent check points CP are obtained by forward projecting the checkpoints' image coordinates

$$(X_i^*, Y_i^*) = K_m^{-1}(x_i^O, y_i^O, DSM) \quad (5)$$

onto the DSM. (X_i^*, Y_i^*) represent the object coordinates that would have been obtained by orthorectification. Their difference to the ground truth coordinates of the checkpoint (X_i^D, Y_i^D) defines the 2D ortho projection error

$$e_i = (X_i^D - X_i^*, Y_i^D - Y_i^*) \quad (6)$$

We expect that precise image orientations show errors e that are similar or smaller than the resolution of the reference imagery D.

In addition, we evaluate the relative consistency of the matched points by predicting using the same procedure on the $iGCP$ instead of the independent checkpoints CP.

3.4 Matching methods

We used the implementations provided by the Deep Image Matching framework (Morelli et al., 2024). Four methods were tested: SIFT, SuperPoint + LightGlue (SP+LG), SE2LoFTR, and RoMa. All methods were used with their pretrained weights and original inference configurations, which are available in the framework, except for increasing the maximum number of feature points to 15000, as we utilized full resolution images without downscaling. Additionally, we modified the toolbox to save image coordinates in float32 instead of float16, to preserve the sub-pixel accuracy of SIFT, which would otherwise be rounded away due to the limited precision of the half datatype. For SIFT, we extracted keypoints and descriptors using the default OpenCV implementation. Then, we obtained feature correspondences through nearest-neighbor matching. For the approx-orthorectified use case, we additionally used our own SIFT implementation in which we disabled rotation and scale invariance SIFT upright (SIFT-UP). For SP+LG, interest points and descriptors were first generated with SuperPoint.

Then, the matching stage was performed with LightGlue using the default model checkpoint and parameters. SE2LoFTR and RoMa were executed as detector-free models and received input image pairs at their original resolution, as specified in the configuration files. The confidence thresholds and other inference parameters were kept at their default values for all methods.

We used the MatchAnything framework (He et al., 2025) as a unified backbone for deep feature matching, running the authors' checkpoints with default inference setup and no fine-tuning. To retain more candidate matches, especially in low-texture regions, we set a low confidence threshold of 0.1. We refer to these results as MA ELoFTR and MA RoMa. We evaluated JamMa using the authors' inference script and default parameters. Since JamMa processes concatenated image pairs, both images were resized to a 1400-pixel long side and padded to 1400×1400 to match spatial dimensions and fit GPU memory. Final correspondences were projected back to the original image coordinates for accuracy.

Each method produced a set of putative matches that were refined using RANSAC, as described in Section 3.1.

4. Results and discussion

The algorithms given in the previous section were applied to the dataset. Results for different matching algorithms and strategies are shown in Table 1. Matching of the original images O was the most challenging use case, and no algorithm could find points that allowed the orientation of all images. For the pre-rectified images A, success rates were much higher, and two algorithms could find points that allowed the orientation of all images. We consider images where the 2D MAE (mean absolute error) of the checkpoints is smaller than the GSD (0.2 m) of the reference imagery well oriented. In computer vision, 3 pixels are often considered acceptable, but for our application, this translates into a 0.6m MAE, which is too high for precision mapping applications. Values are nevertheless reported for completeness, and might be helpful for other applications with lower requirements on the absolute accuracy of the mapping products. The mean X,Y and Euclidean error of all checkpoints provide an idea of the offset and spread of the estimated image orientations.

When matching the original images O, most methods fail completely to provide points that allow very precise image orientation, only SP+LG allowed orientation of 7 out of 51 images. RoMa was successful on 3 images and JamMa and SIFT on 2 images. Thus, none of the evaluated methods provided satisfactory performance for precise image orientation in our use-case. Some methods, such as SP+LG and JamMa cannot handle rotated images well, as they were trained on MegaDepth which does not include significant image rotations, but still were able to match some image pairs with less rotational difference. If reduced accuracy is acceptable, RoMa performed best and could orient 29% of the images with an MAE less than 0.6 m. The Match Anything trained versions of ELoFTR and RoMa performed worse than the original version.

Figure 3 plots the checkpoint MAE values for all evaluated matchers. This visualisation shows that for the original imagery, only very few images can be oriented with an MAE < 0.2 m. Here, RoMa shows an interesting behavior: while it is not the most precise, most images can still be

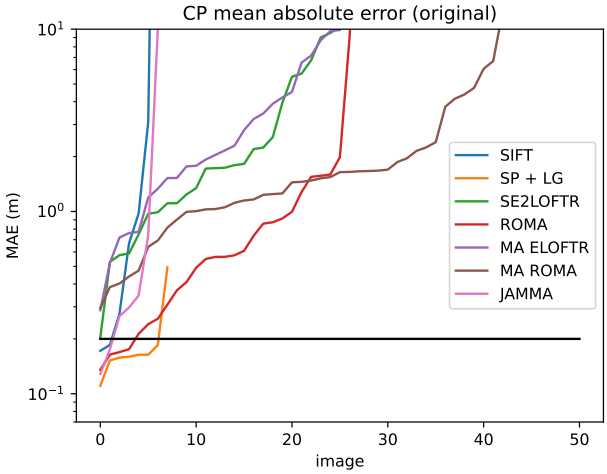


Figure 3. Per image checkpoint MAE for the different matchers for matching original and 20 cm reference imagery.

roughly oriented with the two RoMa variants, especially with the MatchAnything-trained RoMa version, at the expense of errors in the 1-5 m range. It thus still finds coarse approximate points for hard cases, but not with the accuracy required for good image orientation.

Moving to the easier, approx-orthorectified use case, JamMa, SP+LG and SIFT-UP managed to precisely orient 15 images, followed by SIFT with 10. Notably, ELoFTR and RoMa fall behind, both in the original and the MatchAnything trained version. When checking the success rate for images with 0.6m MAE, which roughly corresponds to an accuracy of 3 pixels, it can be seen that SIFT is the least robust method of these three. Here JamMa, followed by SP+LG take the lead, showing that for harder pairs, they still find image orientations that might be useful for some applications for 82 % of the images, i.e., 18 % can still not be oriented with suitable quality.

Figure 4 shows that most matchers show an exponential loss of accuracy with increasing image matching difficulty. For easy pairs both variants of the SIFT matcher produce the most accurate results, but JamMa is very close, and continues to work when matching difficulty increases. SIFT-UP, a modified SIFT variant without scale and rotation invariance, continues to perform consistently when the plain SIFT starts to break down. When utilizing the prior information, even SIFT provides results comparable to state-of-the-art deep learning-based methods. SP+LG also shows a competitive performance, while the RoMa produces less accurate results, especially for the easier images. ELoFTR seems to be the least accurate of the evaluated deep learning based matching methods.

Most recent algorithms perform matching between image pairs only, thus we focus our work on algorithms that match two images. Methods based on keypoint separate detection and matching steps, such as SIFT allow chaining of pairwise matches to multi ray matches, but this is not possible for newer methods such as JamMa. Extending these methods to multi-image matching would allow the use of multi-image bundle adjustment and could improve the accuracy of the evaluation procedure, however this was outside the scope of this paper and is left to further work.

It is important to note that the ground truth of our dataset relies on the accuracy of DOP D and DSM imagery used. Errors

Method	images with MAE $< 0.2m$ [%]	$< 0.6m$ [%]	processed images [%]	MAE XY [m]	RMSE XY [m]	rel iGCP MAE [m]
Original imagery						
SIFT	3.9	5.9	13.7	0.26	0.51	5.90
SP + LG	**13.7**	15.7	17.6	0.16	0.18	3.32
SE2LOFTR	0.0	7.8	56.9	1.86	7.48	2.79
ROMA	7.8	**29.4**	56.9	0.63	5.65	1.11
MA ELOFTR	0.0	3.9	52.9	2.98	21.88	0.90
MA ROMA	0.0	9.8	94.1	2.37	18.72	1.17
JAMMA	3.9	9.8	13.7	1.93	4.77	1.36
Approx-Ortho imagery						
SIFT	19.6	49.0	68.6	0.54	2.26	0.44
SP + LG	**29.4**	**82.4**	98.0	0.25	0.39	1.16
SE2LOFTR	2.0	29.4	96.1	0.74	1.68	1.51
ROMA	3.9	66.7	98.0	0.34	0.49	0.85
MA ELOFTR	2.0	25.5	92.2	0.80	1.76	0.91
MA ROMA	5.9	64.7	100.0	0.45	1.44	0.64
JAMMA	**29.4**	**82.4**	100.0	0.25	0.38	1.08
SIFT-UP	**29.4**	72.5	98.0	0.23	0.41	0.99

Table 1. Matching results for both matching of the original and coarsely approx-orthorectified imagery against 0.2m reference imagery. The percentage of scenes with MAE less than 0.2m, shown in column 2, reports successful image orientation. We additionally report values for 0.6 m (3 pixels) in column 3. The number of scenes for which the matchers provided points are given in column 4. For these scenes, mean absolute error and RMSE of all checkpoints are reported. The last column shows the iGCP MAE of the image orientation procedure, which gives an overall impression on the relative precision of the matchers.

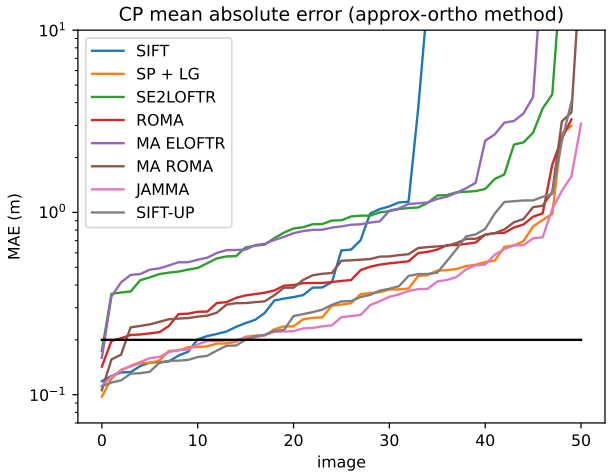


Figure 4. Per image checkpoint MAE for matching approx-orthorectified vs 20 cm reference imagery.

in the true ortho-rectification procedure and the LiDAR DSM lead to errors in the ground coordinates of both CP and $iGCP$ and thus influence the results. In addition, any error in the measurement of the CP image coordinates is attributed to the matching methods. The image threshold MAE < 0.2 m is thus relatively strict. However, these error sources affect all matching methods in the same way, and become more relevant when evaluating sub-pixel precise matching algorithms, which except for SIFT none of the evaluated algorithms are. Figures 3 and 4 still clearly show the ranking of the methods, and allows to draw conclusions about their relative performance.

4.1 Conclusions

The results of this paper show no clear winner, but some trends. In the original image case, where images exhibit scale and rotation differences, none of the methods achieve satisfactory results on the test dataset. SuperPoint+LightGlue successfully oriented 14% of the images at high quality. This is interesting, especially since SuperPoint+LightGlue was trained on the MegaDepth dataset and does not work well in the presence of large image rotations. RoMa was the most robust method for the original images use case, but it does not provide good accuracy. When using GNSS/IMU orientations and interior camera calibration to perform image matching on coarsely orthorectified images, JamMa, SuperPoint+LightGlue, and SIFT-UP (a SIFT method without rotation and scale invariance) worked well for up to 30% of the test images. For difficult cases, JamMa showed the best MAE metrics, while for well-matched pairs, the SIFT-based methods performed best.

This study was a first step in evaluating state-of-the-art matching methods on aerial imagery with large temporal and thus image-content differences. Results show that current methods still strongly benefit from prior knowledge, such as GNSS/IMU orientations. Future work could focus on training promising networks with data more relevant to our application domain; for example, the well-performing JamMa was trained only on MegaDepth and thus cannot handle stronger rotations, which occur in our dataset. Additionally, sub-pixel precision is another area for improvement. Simply applying local least-squares fine-matching to the tie points of deep learning matchers often fails, as the points are not necessarily located in areas with sufficient local image corners, or scene changes prevent local least-squares matching from converging. For more difficult cases, investigating matching other primitives, such as lines or semantic objects, instead of points, could be an interesting alternative.

References

Aguilar, M. A., Jiménez-Lao, R., Aguilar, F. J., 2022. Assessment of stereo-extracted DSM from WorldView-3 over different land covers depending on the imaging geometry. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, V-2-2022, 31–38.

Barath, D., Noskova, J., Ivashechkin, M., Matas, J., 2020. Magsac++, a fast, reliable and accurate robust estimator. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Bay, H., Tuytelaars, T., Gool, L. V., 2006. SURF: Speeded up robust features. *European Conference on Computer Vision (ECCV)*, Springer, 404–417.

Bökman, G., Kahl, F., 2022. A case for using rotation invariant features in state of the art feature matchers. *CVPRW*.

C, B., C, L.-V., C, G., P, G., L, H., D, G., S, T., V, H.-C., S, R., A, C., HI, R., P, S., 2024. Novel approach for ranking DEMs: Copernicus DEM improves one arc second open global topography. *IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING*, 62, 1-22. https://ieeexplore.ieee.org/document/10440392.

Chen, C., Ma, J., Zhao, J., Zhou, J., Yuille, A. L., 2014. The Registration of Multimodal Remote Sensing Images via Fully Convolutional Networks. *Remote Sensing*, 6(12), 12087–12103.

Cheng, G., Huang, Y., Li, X., Lyu, S., Xu, Z., Zhao, H., Zhao, Q., Xiang, S., 2024. Change Detection Methods for Remote Sensing in the Last Decade: A Comprehensive Review. *Remote Sensing*, 16(13), 2355.

DeTone, D., Malisiewicz, T., Rabinovich, A., 2017. SuperPoint: Self-Supervised Interest Point Detection and Description. *CoRR*, abs/1712.07629. http://arxiv.org/abs/1712.07629.

Durgam, A., Paheding, S., Dhiman, V., Devabhaktuni, V., 2024. Cross-View Geo-Localization: A Survey. *IEEE Access*, 12, 192028-192050.

Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M., 2024. RoMa: Robust Dense Feature Matching. *IEEE Conference on Computer Vision and Pattern Recognition*.

Fan, B., Kong, Q., Zhang, Z., Wu, F., 2019. A Performance Evaluation of Local Features for Remote Sensing Image Matching. *IEEE Geoscience and Remote Sensing Letters*, 16(5), 732–736.

Fischler, M. A., Bolles, R. C., 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, 24(6), 381–395.

Gruen, A., 1985. Adaptive Least Squares Correlation: A Powerful Image Matching Technique. *South African Journal of Photogrammetry, Remote Sensing and Cartography*, 14(3), 175–187.

He, X., Yu, H., Peng, S., Tan, D., Shen, Z., Bao, H., Zhou, X., 2025. Matchanything: Universal cross-modality image matching with large-scale pre-training. *Arxiv*.

Leutenegger, S., Chli, M., Siegwart, R. Y., 2011. BRISK: Binary robust invariant scalable keypoints. *IEEE International Conference on Computer Vision (ICCV)*, 2548–2555.

Liao, Y., Wu, X., Liu, J., Liu, P., Pan, Z., Duan, Q., 2025. Fm-CFA: A Feature Matching Method for Critical Feature Attention in Multimodal Images. *Scientific Reports*, 15(1), 6640.

Lindenberger, P., Sarlin, P.-E., Pollefeys, M., 2023. LightGlue: Local Feature Matching at Light Speed. *ICCV*.

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

Lu, X., Du, S., 2025. Jamma: Ultra-lightweight local feature matching with joint mamba.

Ma, J., Jiang, J., Zhang, J., Zhang, J., Li, X., 2019. A Review of Feature-Based Image Registration Techniques for Remote Sensing Applications. *Information Fusion*, 46, 1–20.

Morel, J.-M., Yu, G., 2009. ASIFT: A New Framework for Fully Affine Invariant Image Comparison. *SIAM Journal on Imaging Sciences*, 2(2), 438–469.

Morelli, L., Ioli, F., Maiwald, F., Mazzacca, G., Menna, F., Remondino, F., 2024. Deep-Image-Matching: A Toolbox for Multiview Image Matching of Complex Scenarios. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-2/W4-2024, 309–316. https://github.com/3DOM-FBK/deep-image-matching.

Muja, M., Lowe, D. G., 2009. Fast approximate nearest neighbors with automatic algorithm configuration. *International Conference on Computer Vision Theory and Applications (VISAPP)*, 331–340.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2023. Dinov2: Learning robust visual features without supervision.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. SuperGlue: Learning feature matching with graph neural networks. *CVPR*.

Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X., 2021. LoFTR: Detector-Free Local Feature Matching with Transformers. *CVPR*.

Wang, Y., He, X., Peng, S., Tan, D., Zhou, X., 2024. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. *CVPR*.