The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLVIII-1/W5-2025
13th International Conference on Mobile Mapping Technology (MMT 2025)
"Mobile Mapping for Autonomous Systems and Spatial Intelligence", 20–22 June 2025, Xiamen, China

Improving Image Alignment in vineyard environment with deep learning image matching

Andrea Maria Lingua [1], Stefania Manca[1], Francesca Gallitto[1], Filiberto Chiabrando[2]

[1] DIATI, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129, Italy– (andrea.lingua, stefania.manca, francesca.gallitto@polito.it)
[2]DAD, Politecnico di Torino, Viale Mattioli 39, 10125, Italy – filiberto.chiabrando@polito.it

Keywords: Deep Image Matching vineyard, vSLAM, UAV, fine-tuning.

Abstract

Globalisation has accelerated the spread of invasive agricultural pests, including *Popillia japonica Newman*, introduced to Italy in 2014. This species has caused severe damage to vineyards, highlighting the need for efficient detection methods. Manual identification, though accurate, is time-consuming and labour-intensive. This study explores a computer vision (CV)-based approach using Near-Infrared (NIR) imagery captured by Uncrewed Aerial Systems (UAS) to detect adult *Popillia* specimens. Conducted in two vineyards in northern Italy, the project aims to develop a standardised and replicable monitoring protocol. CV-based detections are validated by entomologists and integrated into a Geographic Information System (GIS) to generate prescription maps for targeted drone-based pesticide application.
However, traditional feature extraction and matching (FEM) algorithms, such as SIFT, SURF, and ORB, struggle in vineyard environments due to repetitive structures (seriality of fixed components, such as poles, supports, etc) and limited NIR texture. These limitations hinder image alignment, especially in the absence of geodetic-grade GNSS and high-precision IMU data. To address this, the study replaces FEM methods with deep image matching (DIM) techniques like SuperPoint and DISK for feature extraction, paired with SuperGlue for graph-based matching. Applied within a visual SLAM (vSLAM) framework, these deep learning models significantly improve image connectivity and alignment. Experimental results, supported by a fine-tuned SuperPoint model trained on vineyard datasets from the DANTE2 project, demonstrate up to 90% alignment improvement over conventional methods. This work presents a robust, scalable solution for accurate pest mapping in viticulture, contributing a fine-tuned PyTorch model to the scientific community.

1. Introduction

Globalisation and climate change have significantly accelerated the spread of invasive pests, threatening agricultural systems and global food security. A prime example is the *Popillia japonica Newman* beetle (Figure 1), native to Japan but inadvertently introduced into Europe and North America. Since its detection in Northern Italy in 2014, this species has caused substantial damage, particularly in vineyards, due to its voracious feeding habits and rapid population growth. Adult beetles skeletonize leaves and damage fruit, while larvae feed on root systems, resulting in extensive crop loss and economic burden. Traditional monitoring and containment efforts are labour-intensive and often inadequate for timely intervention.
To enhance pest surveillance and management, increasing attention has been directed toward automated, image-based detection systems supported by Computer Vision (CV) and Artificial Intelligence (AI). Among available technologies, image-based approaches are especially promising due to their accessibility, scalability, and capacity for morphological species identification. However, real-world conditions such as repetitive vineyard structures, visually complex natural backgrounds and insects size (very close images within a distance of 2-3 m are required), where insects exhibit colour and texture similarities with foliage, pose significant challenges for conventional image matching techniques. Compounding this, Popillia japonica exhibits a highly reflective green exoskeleton that shares spectral similarity with foliage in the visible spectrum, complicating visual differentiation using RGB imagery and the.
To mitigate these challenges, this study explores using near-infrared (NIR) imagery acquired via Uncrewed Aerial Systems (UAS, Longhi V. et al, 2024). NIR imaging improves contrast between the beetles and the vegetation: plants exhibit high reflectivity in the NIR band, whereas the beetles appear significantly darker due to their low NIR reflectance (Figure 2). (Matrone f. et al, 2024) demonstrates that also point clouds can be used for classification problems and defines procedures for Enhancing explainability of deep learning models.

Figure 1.An adult insect of Popillia Japonica on a vine plant.

In complex agricultural environments, cellular and telecommunication networks are often lacking, resulting in the unavailability of internet connectivity. Consequently, GNSS receivers embedded in UAS cannot receive RTK corrections, which are essential for high-precision definition of external orientation parameters of acquired images. To overcome this limitation, it is sometimes necessary to deploy dedicated GNSS base stations acting as master units; however, this approach is expensive and requires expert users. As an alternative, accurate photogrammetric reconstruction of the sweeping paths is required to enable reliable insect identification.

It is therefore necessary to follow the standard photogrammetric workflow, which involves aligning the images to form a photogrammetric block, identifying ground control points (GCPs), and subsequently orienting the entire block. However, image alignment often presents challenges, as conventional image matching and feature extraction algorithms (SIFT, ORB, SURF, …) tend to perform poorly in such very complex environments.

Figure 2.An example of an NIR image of Popillia insects in the studied vineyards.

There are numerous applications of artificial intelligence in precision agriculture (Pádua, L. et al., 2022, Baldaccini, M. et al., 2024, Ramyaa, R. et al., 2024, Tsouros, D. et al., 2023, Williams, T. et al., 2022) addressing a wide range of thematic aspects, such as vineyard growth assessment, health status monitoring, and the detection of diseases like downy mildew, using both visible spectrum and multispectral (including near-infrared) imagery. However, the existing literature lacks studies focusing on integrating photogrammetric techniques in this context, particularly concerning the critical issue of image alignment and external orientation parameter estimation.
To effectively register and align NIR images over time and across different acquisition angles, the Deep Image Matching algorithms (Morelli L. et al., 2024) could potentially allow possible solutions using e.g. SuperPoint/SuperGlue (also with lightglue version, De Tone D. et al., 2018, Sarlin P. E. et al., 2020) or DISK/SuperGlue (Tyszkiewicz M.J. et al., 2020).
This contribution focuses on enhancing the reliability of image alignment in vineyard environments through the application of deep image matching (DIM) techniques. It demonstrates how SuperPoint/SuperGlue and DISK/SuperGlue can overcome the limitations of more traditional feature-matching methods (SIFT–like), using Bunfdle Block Adjustment and a vSLAM approach.

2. Study area

In the past year, the authors have proposed a study (Longhi et al., 2024) to evaluate a CV algorithm's effectiveness in identifying adult Popillia specimens using Near-Infrared sensors on Uncrewed Aerial Systems (UAS).
The project, conducted in two vineyards in northern Italy, intends to establish a replicable and standardised data acquisition protocol for future monitoring activities. Manual counting performed by entomologists validates insects detected by the CV-based method. In a GIS environment, prescription maps are generated in near real-time to identify where the vineyard is most affected and to guide the drone spraying treatment only on the areas in which the threshold is exceeded. For a correct spatialization of the insect locations, the exterior orientation parameters of NIR images have to be known: in the absence of a geodetic GNSS antenna and an accurate IMU on board of drone, a Structure from Motion procedures has to be applied to realize the image alignment, the Ground Control Point plotting and the bundle block adjustment.
At the beginning of the project, two areas of comparable sizes for each vineyard were chosen and classified based on the type of treatment they would receive. The two areas have similar surfaces (about 1 ha in Briona and Ghemme) and were defined based on a 3D model generated by aerial photogrammetry at the beginning of the project (time 0 [T0] acquisition) using DJI mat4rice 300 and P1 digital camera (Table 1). The vineyard in Briona is located on flat ground, while the one in Ghemme is on a slope, a feature to consider when planning a UAV survey.

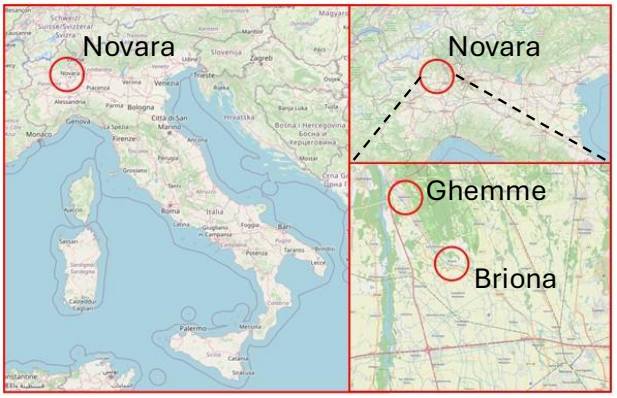

Figure 3.The study area: in the bottom right picture, Ghemme and Briona towns are highlighted by red circles.

As previously noted, the imagery was captured in the Near-Infrared (NIR) spectral band. Data was collected via aerial surveys using a DJI Mavic 2 Pro drone equipped with a Sentera single NIR sensor (see Table 1). The acquisition protocol was specifically designed to include flights with the camera tilted at approximately 45°, oriented perpendicularly to the vine rows. Flights were conducted at a low altitude, between 2 and 3 meters above the canopy level, and at a speed of roughly 2 m/s. Since Popillia japonica tends to remain still in the early morning hours and becomes active only after sunrise, all image acquisition missions were scheduled between 6:00 and 8:00 a.m. These flight parameters enabled the collection of high-resolution, close-range images of the pest (Brusco et al., 2023)

UAS	DJI Matrice 300	DJI Mavic 2 Pro
Sensor	Zenmuse P1 - RGB	Sentera Single - NIR
Resolution	8192 × 5460	1248 × 950
Focal Length	35 mm	4.14 mm
Pixel Size	4.39 × 4.39 µm	3.75 × 3.75 µm
Flying altitude	26.7 m	2.5 m
GSD	4.1 mm/pix	2.2 mm/pix

Table 1. Main specifications of sensors and UAS.

This study is framed within the DANTE project, which aims to develop a precision agriculture framework for monitoring and treating Popillia japonica using UAS technologies. The method used for extract and spatialise the insects has been described in (Longhi V. et al, 2024).

3. Problem

In a recent study (Longhi et al., 2024), the authors investigated the effectiveness of a computer vision (CV) algorithm for detecting adult specimens of Popillia japonica using Near-Infrared (NIR) imagery acquired via Uncrewed Aerial Systems (UAS). Conducted across two vineyards in northern Italy, the study aimed to establish a replicable and standardised protocol for pest monitoring, integrating UAS-based imaging, CV-driven insect identification, and GIS-based prescription mapping. Insect detections, validated through manual counting by entomologists, were spatially analysed to generate targeted treatment maps that direct drone-based pesticide application only to vineyard sub-areas exceeding infestation thresholds.

A key challenge in this workflow lies in accurately georeferencing the detected insect positions within the vineyard. In the absence of a geodetic-grade GNSS receiver and a high-precision Inertial Measurement Unit (IMU) onboard the drone, direct recovery of the exterior orientation parameters (EOPs) for the captured NIR images is not feasible. Therefore, Structure-from-Motion (SfM) techniques must be employed to reconstruct the camera poses and 3D scene geometry. This process relies heavily on the detection and matching of robust visual key points across overlapping images.

However, traditional feature extraction and matching (FEM) algorithms, such as SIFT, SURF, and ORB, often fail to perform reliably under these conditions. The vineyard environment poses multiple challenges: a) repetitive and self-similar structures (e.g., vine rows, support poles, trellis systems), b) limited texture in NIR spectral bands, and c) low inter-image baseline variation due to the drone's flight altitude and trajectory. These factors result in insufficient and ambiguous key point matches, leading to disconnected image blocks, failed bundle block adjustments, and ultimately, incomplete or imprecise 3D reconstructions. As demonstrated in Figure 4, these issues are particularly evident when in the process are employing conventional SfM pipelines, such as those implemented in traditional SfM software generally based on SIFT-like approach.

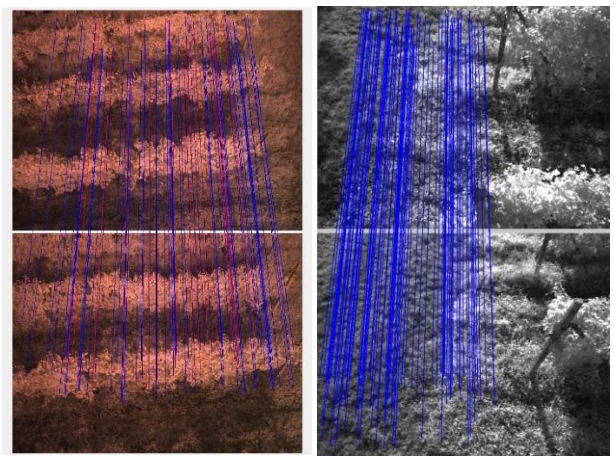

Figure 4. 2 examples of bad solution of image alignment using CV feature extraction and matching

As illustrated in the following Figure 5, two examples of NIR image alignment are provided for the vineyards of Gemme and Briona. In both cases, approximately 1300 images were collected. As is outlined in the images there are large parts of vineyard without aligned images (more than 45%), consequently a large amount of data is missed and according to this results the proposed monitoring methodology of the *Popilla Japonica* failed.

4. Methodology

4.1 Result and discussion of direct application of DIM algorithm in vineyard environment

To overcome these limitations, in the present research different DIM techniques were analysed and tested. Hereafter a list of the employed techniques with a short description is reported:

1- Superpoint+Superglue. Superpoint is a fully convolutional neural network designed for real-time interest point detection and descriptor extraction in images. It was introduced in 2018 (DeTone et al., 2018) The method uses a self-supervised training strategy, beginning with synthetic data where a detector is trained on simple geometric shapes. Through a process called homographic adaptation, this model is fine-tuned on real images without requiring manually labeled keypoints. SuperPoint outputs both keypoint locations and corresponding descriptors in a single forward pass. Superglue is a deep learning model introduced in 2020 (Sarlin et al., 2020), designed to match sets of local features between pairs of images by jointly finding correspondences and rejecting non-matchable points. The model takes as input the key points and descriptors extracted from two images and processes them through an attentional graph neural network.

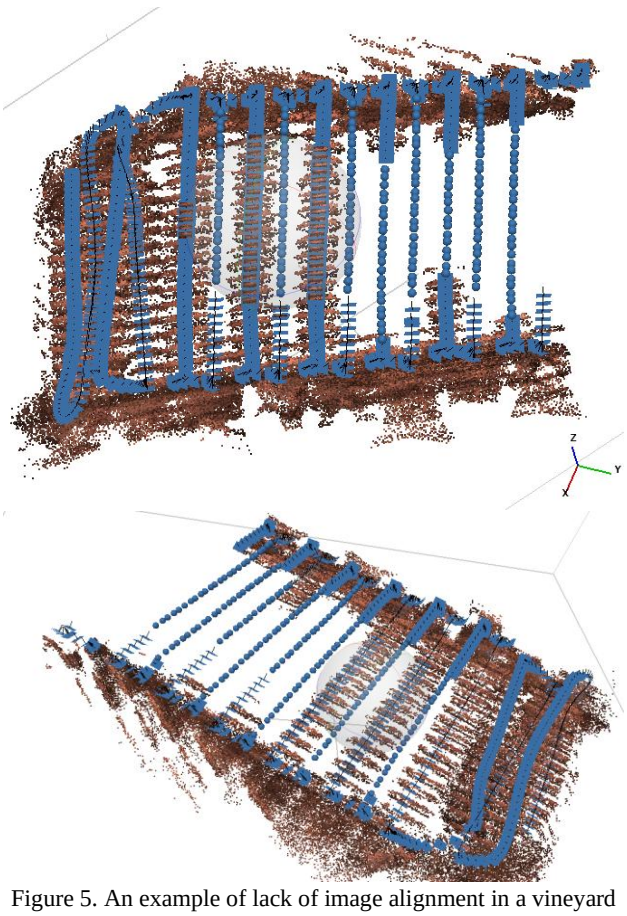

Figure 5. An example of lack of image alignment in a vineyard (Briona, up and Ghemme, down)

2- Superpoint+Lightglue. Lightglue builds on the groundwork established by Superglue, refining key architectural elements to improve efficiency and accuracy (Lindenberger et al., 2023). It introduces an adaptive strategy that dynamically adjusts the model's computational depth according to the complexity of the image pair. For image pairs with high visual similarity, the model

can accelerate processing by halting computations early, whereas more challenging pairs are given deeper processing to maintain precise matching.

3- *Superpoint+Lightglue Fast* implements a particular settings of Lightglue with Flash attention (Dao T. et al., 2022) and lower adaptive thresholds.

4- *Disk + Superglue*
Disk (DIScrete Keypoints) is a deep learning model introduced in 2020 (Tyszkiewicz et al., 2020), the model addresses the challenge of learning local feature frameworks in an end-to-end fashion, which is often difficult due to the discrete nature of selecting and matching sparse keypoints. To overcome this, Disk leverages principles from reinforcement learning, specifically policy gradients, to optimize the detection and matching of key points directly for the number of correct matches.

The results have been compared with more traditional FEM algorithms (well known):

5- *SIFT+kornia matcher*
6- SIFT-like commercial software;

All the tests (1-5) were conducted using the Deep Image Matching toolbox developed by Fondazione Bruno Kessler (FBK), including the open-source DIM codes, with 3 strategies S1, S2, LR):
- strategies S1 and S2 simulate the FEM step of the vSLAM solution, supposing that there is an overlap between sequential images of 1 image (only 1 pair for each image) or 2 (2 pairs for each image, 3 images are concatenated);
- the strategy LR selects the possible pairs using a preliminary matching with subsampled low-res images.

There is the exception of the final case (6), for which a commercial SIFT-like software was employed only with automatic pair selection based on generic preselection.
The six algorithms and the three strategies have been applied in 4 cases

A. 20 images with a good solution of commercial SIFT-like software (algorithm 6), figure 6 and Table 2;
B. 20 images with a poor solution of commercial SIFT-like software (algorithm 6), figure 7 and Table 3;
C. an entire vineyard (1274 images) with a good solution of commercial SIFT-like software (algorithm 6), figure 9 and Table 4;
D. an entire vineyard (1214 images) with a poor solution of commercial SIFT-like software (algorithm 6), figure 10 and Table 5.

Cases A and B have been developed to evaluate the various algorithms in detail, with matching plotting graphs and single pair analysis to understand the causes of good or poor solutions; cases C and D have been applied to check the results in a whole vineyard area.
The default parameters of processing of DIM-FBK have been used except the maximum number of selected points, which was set without any limitation. For SuperPoint algorithms, we used the pretrained model `superpoint_v1.pth` (De Tone D., et al., 2018). Tables 1, 2, 3, 4, 5 and 6 contain the results of numerous processes in terms of:

(a) used strategy, S;
(b) number of aligned images, Al. im. and %;
(c) number of good ties Points, Tie Pts;
(d) mean residual of reprojection errors, mean res [μm]
(e) max residual of reprojection errors, max res [μm];
(f) average tie points multiplicity, mt.

In Table 2 (Case A), which represents a scenario with good initial alignment, the commercial SIFT-like software performs optimally, generating the highest number of tie points (23,269), and achieving the lowest mean reprojection error of 0.36 μm. This indicates that under favourable imaging conditions with sufficient texture and geometric structure, traditional SIFT-based pipelines yield highly reliable results in both image alignment and geometric accuracy.

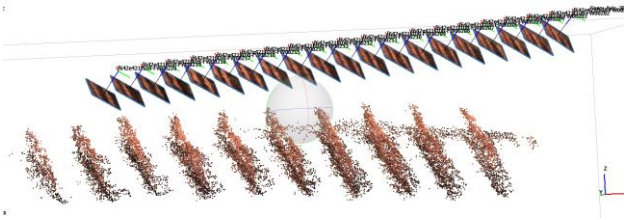

Figure 6. case A: 20 images with a good solution of commercial sift-like software (20/20 aligned images)

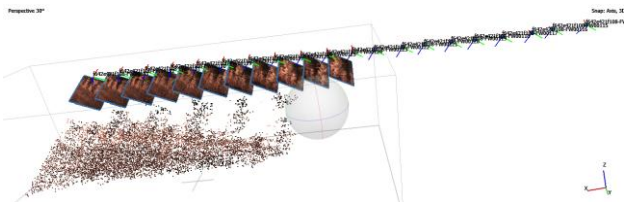

Figure 7. case B: 20 images with a poor solution of commercial sift-like software (11/20 aligned images)

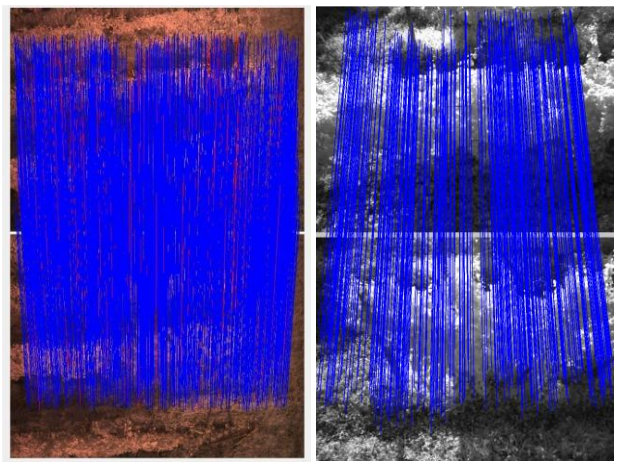

Figure 8. Superpoint/superglue applied to the vineyard images

Among the deep learning-based methods, DISK combined with LightGlue produces the densest match set, with tie points exceeding 48,000, although it comes with a slightly higher mean reprojection error (1.1–1.2 μm), suggesting a potentially noisier result that would benefit from outlier filtering. Superpoint combined with Superglue and LightGlue yields moderate tie point counts (ranging from 7,000 to 11,000) and consistent reprojection errors around 1.2–1.4 μm, demonstrating strong robustness and stability even if not surpassing the classical solution in this optimal context. The matching times for these learning-based models remain within an acceptable range (3.1 to 3.8), underscoring their practical feasibility. Overall, this case confirms that while deep learning approaches are competitive,

SIFT remains the most effective under optimal vineyard imaging conditions.

In Table 3 (Case B), representing a more challenging scenario with poor initial alignment, the limitations of the SIFT-like method become apparent. It aligns only 11 out of 20 images and produces fewer tie points (8,350), underscoring its vulnerability in environments with low texture and visually ambiguous NIR content. In contrast, the deep learning methods demonstrate superior resilience: both Superpoint + Superglue and Superpoint + LightGlue manage to align 17 out of 20 images and yield tie point counts between approximately 4,800 and 6,100, albeit with slightly higher reprojection errors (1.2–1.4 µm). This trade-off is acceptable given the improvement in image connectivity and overall reconstruction integrity. DISK + LightGlue shows inconsistent performance, with only 3 images aligned under the S1 strategy but significantly better results (11–13 aligned images and over 18,000 tie points) under the S2 and LR strategies, indicating its heavy dependence on effective pair selection mechanisms. The increase in reprojection error for DISK (1.6–1.7 µm) under poor conditions may stem from overmatching in visually repetitive or low-texture areas. Thus, Table 3 highlights the robustness and adaptability of deep learning-based matchers in degraded imaging contexts, with Superpoint + LightGlue standing out as a balanced and reliable performer, and DISK showing high potential when guided by intelligent pairing strategies.

	S (a)	Al. im. (b)	Tie Pts (c)	Mean res [µm] (d)	Max res [µm] (e)	Mt (f)
SIFT-like	-	20	23269	0.36	6.5	2.4
SIFT+ Kornia	S1	20	5343	0.51	4.0	3.0
	S2	20	5672	0.57	3.46	3.3
	LR	20	4689	0.51	3.9	3.3
Superpoint+ superglue	S1	20	9787	1.2	3.9	3.4
	S2	20	11201	1.4	4.0	3.3
	LR	20	8897	1.2	3.9	3.3
Superpoint+ lightglue	S1	20	7259	1.3	4.1	3.4
	S2	20	10369	1.4	4.3	3.3
	LR	20	11473	1.3	4.5	3.3
Superpoint+ lightglueFast	S1	20	7273	1.3	4.1	3.4
	S2	20	9039	1.3	4.3	3.3
	LR	20	4913	1.2	3.8	3.1
DISK+ lightglue	S1	20	48173	1.1	4.8	3.7
	S2	20	49069	1.2	4.3	3.8
	LR	20	48733	1.1	4.2	3.8

Table 2. case A: 20 images case with good initial solution

	S (a)	Al. im. (b)	Tie Pts (c)	Mean res [µm] (d)	Max res [µm] (e)	Mt (f)
SIFT-like	L	11	8350	0.4	6.1	2.2
SIFT+ Kornia	S1	10	1758	0.7	3.8	2.9
	S2	11	1443	1.0	3.9	3.2
	LR	11	1142	0.7	3.4	3.3
Superpoint+ superglue	S1	17	1844	1.2	3.5	2.2
	S2	17	4784	1.4	4.2	2.9
	LR	17	5002	1.4	3.9	2.8
Superpoint+ lightglue	S1	17	4835	1.3	4.1	2.6
	S2	17	5908	1.4	4.1	2.7
	LR	17	6134	1.3	4.0	2.8
Superpoint+ lightglueFast	S1	17	4952	1.3	4.2	2.3
	S2	17	5356	1.3	3.7	2.7
	LR	17	5559	1.3	4.1	2.6
DISK+ lightglue	S1	3	2886	1.1	3.6	2.4
	S2	11	18832	1.7	3.9	3.3
	LR	13	25367	1.6	4.1	3.8

Table 3. case B: 20 images case with poor initial solution

These results were confirmed by the application to the entire vineyard, as summarized in Tables 4 and 5. The significant improvement in alignment completeness using DIM algorithms is clearly evident, with values ranging from approximately 72–79% of images successfully aligned, compared to 48–57% obtained using SIFT-like algorithms, representing an improvement of about 150%. A significant portion of the images, approximately 1/4 still fail to align, resulting in substantial gaps in vineyard coverage.

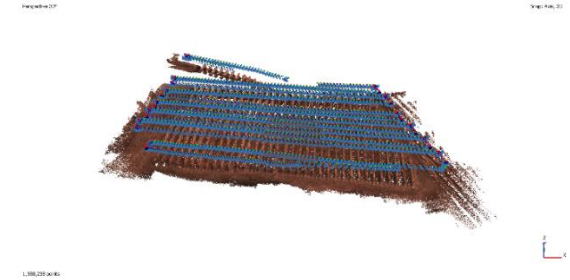

Figure 9. Case C: the whole vineyard of Briona with good initial solution (27.07.02024).

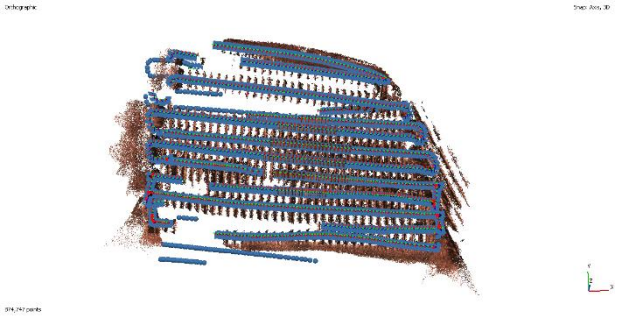

Figure 10. Case D: the whole vineyard of Briona with poor initial solution (02.08.2024).

Total images: 1369	S (a)	Al. im.,% (b)	Tie Pts [kp] (c)	Mean res [µm] (d)	Max res [µm] (e)	Mt. (f)
SIFT-like	-	1341, 98%	1380	0.15	6.39	3.4
SIFT+ Kornia	S1	1180, 86%	315	0,76	4,80	3,3
	S2	1245, 91%	353	0,85	4,15	3,6
	LR	1310, 96%	307	0,76	4,68	3,6
Super point+ superglue	S1	1240, 91%	607	1,56	4,68	3,5
	S2	1320, 96%	739	1,82	4,80	3,4
	LR	1355, 99%	603	1,59	4,68	3,4
Super point+ lightglue	S1	1220, 89%	443	1,69	4,92	3,6
	S2	1325, 97%	687	1,82	5,16	3,5
	LR	1342, 98%	770	1,77	5,40	3,5
Superpoi nt+lightg lue Fast	S1	1215, 89%	442	1,69	4,92	3,6
	S2	1310 ,96%	592	1,65	5,16	3,5
	LR	1340, 98%	329	1,56	4,56	3,3
DISK+ lightglue	S1	1211, 88%	2917	1,43	5,76	3,9
	S2	1330, 97%	3263	1,56	5,16	4,0
	LR	1350, 99%	3289	1,43	5,04	4,0

Table 4. case C: a whole vineyard with good initial solution

The processing times for the entire vineyard (Case C and Case D) were measured and normalised against the execution time of the commercial SIFT-like software, with the results summarised in Table 6. It can be observed that:

- the SuperPoint + SuperGlue pipeline is approximately 2–5 times slower than the SIFT-like algorithms;
- the DISK + LightGlue algorithm exhibits substantially higher processing times, ranging from 4 to over 20 times slower, particularly during the matching phase, likely attributable to the large number of tie-points extracted;

- SuperPoint combined with LightGlue and LightGlueFast exhibits processing times comparable to those of SIFT-like algorithms (albeit marginally higher), particularly when constrained to the S2 strategy;
- Applying DIM algorithms, the residual (d) and (e) increase in relevant way (2 tims about) but remain entirely acceptable for precision agriculture applications, particularly for insect detection, using the threshold of about 2.5 μm for the mean residual.

Considering the results summarised in Tables 6 and 7, in terms of both performance gains and computational efficiency, SuperPoint-based approaches (especially when paired with the LightGlue matcher) emerge as the most advantageous solutions under the S2 strategy (sequential, overlap 2) for vSLAM implementations.

Total images: 1210	S (a)	Al. im., % (b)	Tie Pts [kp] (c)	Mean res [μm] (d)	Max res [μm] (e)	Mt (f)
SIFT-like	L	693, 56%	693	0,16	11,2	3,2
SIFT+ Kornia	S1	592, 48%	104	1,05	4,56	3,2
	S2	665, 54%	87	1,50	4,68	3,5
	LR	702, 57%	73	1,25	4,08	3,6
Superpoint+ superglue	S1	920, 74%	309	1,56	4,20	2,3
	S2	950, 77%	493	1,82	5,04	3,0
	LR	965, 78%	580	1,82	4,68	2,9
Superpoint+ lightglue	S1	922, 73%	382	1,69	4,92	2,7
	S2	952, 77%	470	1,82	4,92	2,8
	LR	980, 79%	601	1.75	4,80	2,9
Superpoint+lightglueFast	S1	915, 74%	386	1,69	5,04	2,4
	S2	947, 76%	492	1,65	4,44	2,8
	LR	975, 78%	598	1,68	4,92	2,7
DISK+ lightglue	S1	406, 33%	391	1,43	4,32	2,5
	S2	710, 57%	2116	2,21	4,68	3,5
	LR	810, 72%	2837	2,08	4,92	4,0

Table 5. case D: a whole vineyard with a poor initial solution

Algorithm/strategy	**S1**	**S2**	**LR**
SIFT-like	--	--	1,00
SIFT+ Kornia	0,86	0,86	1,53
Superpoint+ superglue	2,44	***4,56***	4,67
Superpoint+ lightglue	0,72	***1,00***	1,97
Superpoint+ lightglueFast	0,69	***0,97***	1,75
DISK+ lightglue	3,42	5,50	20,83

Table 6. Processing time in both cases C and D

5. Methods for retraining the SuperPoint model in vineyard environment

To try to improve the performance, the self-supervised fine-tuning of a pretrained SuperPoint network has been realised by enforcing two consistency objectives over a pair of images: the original frame and a randomly warped counterpart.
Notably, this self-supervised framework obviates the need for any a priori ground-truth correspondences; the network learns keypoint and descriptor consistency solely from the original–warped image pairs without requiring pre-matched point annotations.
Let I be an input image and $I_w = H(I)$ its homographically transformed version, where H is a random perturbation matrix. Both images are passed through the shared SuperPoint encoder, yielding:

- Detection heatmaps $semi(I) \in \mathbb{R}^{65\times H\times W}$ and $semi(I_w)$.
- Descriptor volumes $desc(I) \in \mathbb{R}^{256\times H\times W}$ and $desc(I_w)$.

It is possible to compute:

1. Detection Loss
$\mathcal{L}_{det} = MSE(semi(I)_{0:63}, H^{-1}(semi(I\_w)_{0:63}))$ enforcing that the first 64 channels (keypoint scores) are consistent under the inverse warp.
2. Descriptor Loss
$\mathcal{L}_{desc} = MSE(desc(I), H^{-1}(desc(I\_w)))$ enforcing that the dense feature vectors agree under the same spatial transformation.

The total loss is a weighted sum:
$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{det} + \lambda_2 \mathcal{L}_{desc}$.

Gradients of $\mathcal{L}_{total}$ are back-propagated to update all SuperPoint parameters, thus adapting both the keypoint detector and descriptor extractor to the target domain. In this application, we used: λ_1 =0.55 and $\lambda 2$ = 0.45.
The self-supervised fine-tuning has been practically realised with 1000 images randomly extracted from the about 2600 images for the 2 vineyards, using a notebook in Google Colab, generating the `SuperpointVineyardModel_v1.pth`.
The processing time for fine tuned was about 20 hours.

5.1 Results and discussion of the new retrained model

This model has been applied to case D, trying to improve the completeness of the results using the Superpoint algorithms with the *sequentia*l strategy with *overlap 2*, obtaining the Fine Tuned (FT) solution. The result, summarised in Table 7, was compared with the S2 solutions of superglue algorithms obtained with the general-purpose pretrained model superpoint_v1.pth (De Tone D., et al., 2018).

Total images: 1210	S (a)	Al. im., % (b)	Tie Pts [kp] (c)	Mean res [μm] (d)	Max res [μm] (e)	Mt (f)
Superpoint+ superglue	S2	950, 77%	493	1,82	5,04	3,0
	FT	1090, 88%	860	1.79	5.82	3.2
Superpoint+ lightglue	S2	952, 77%	470	1,82	4,92	2,8
	FT	1124, 91%	950	1.89	5.01	3.0
Superpoint+ lightglueFast	S2	947, 76%	492	1,65	4,44	2,8
	FT	1115, 90%	940	1.76	4.51	3.0

Table 7. The results using the retrained (RT) model compared with S2 solution in Table 5

It can be observed that:

- The number of tie points increases significantly, by approximately a factor of two;
- the proportion of successfully aligned images increases markedly from 76–77% to 88–91%;
- This corresponds to a relative improvement of approximately 18% compared to the non–fine-tuned solution;
- only 9–12% of images remain unaligned;
- the residual (d) and (e) increase slightly but remain entirely acceptable for precision agriculture applications, particularly for insect detection, considering the threshold of about 2.5 μm for the mean residual and 7 μm for maximum residual.

Processing times increase by approximately 40%, owing to the extraction of roughly twice as many tie points compared to the non–fine-tuned solution.

6. Conclusions

The application of DIM algorithms in a precision agriculture environment for investigating infestation by the invasive species *Popillia japonica* in vineyard has demonstrated clear efficacy.
Where imaging conditions are optimal, traditional SIFT-based methods deliver the best results, achieving full alignment, the highest tie point counts, and the lowest reprojection errors. However, deep learning methods like DISK + LightGlue and Superpoint variants also perform well, offering a viable alternative with good robustness and only slightly lower accuracy.
Traditional SIFT-like approaches struggle significantly under poor imaging conditions with low texture and repetitive patterns. Deep learning-based methods (DIM), especially Superpoint + LightGlue and Superpoint + Superglue, demonstrate superior alignment completeness and resilience, making them better suited for challenging real-world vineyard monitoring. DISK + LightGlue performs well but relies heavily on advanced pairing strategies to be effective.
DIM algorithms enable the alignment of over 75 % of images even under challenging conditions, representing a 150 % improvement compared to SIFT-like solutions. Employing the original pretrained model and a sequential strategy with overlap-2, SuperPoint-based approaches emerge as the optimal solution in terms of both alignment performance and computational efficiency, especially when paired with LightGlue matchers.
The authors have developed an original fine-tuning procedure to optimise both keypoint detection and descriptor extraction, producing a retrained model that achieves over 90 % image alignment in complex scenarios, demonstrating the solution's efficacy and efficiency.

Acknowledgments

This study is funded by the project the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR FAIR, CUP n. E13C22001800001) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). The project uses data acquired for the project INTERREG MAIN10ANCE.
The data has been acquired in collaboration with the project DANTE2 funded by Regione Piemonte (Italy, CUP n.J75G23000060002) and involves the following partners: the Department of Environment, Land and Infrastructure Engineering (DIATI) of Polytechnic of Turin, the Department of Agricultural, Forestry and Food Sciences (DiSAFA) of the University of Turin, Regione Piemonte and "Consorzio di Tutela Nebbioli dell'Alto Piemonte".
This manuscript reflects only the authors' views and opinions; neither the European Union nor the European Commission can be considered responsible for them.

References

Baldaccini, M., Francesconi, A., & Neri, B. (2024). Precision agriculture for wine production: A machine learning approach to link weather conditions and wine quality. Heliyon, 10(5), e24076795.

Brusco, D., Belcore, E., Piras, M. (2023). Popillia Japonica Newman Detection Through Remote Sensing and AI Computer Vision. In: 2023 IEEE Conference on AgriFood Electronics (CAFE), IEEE, Torino, Italy, pp. 50–54.

Dao T., Fu D. J., Ermon S., Rudra A., Ré C. (2022). FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness. arXiv.

De Tone, D., Malisiewicz, T., & Rabinovich, A. (2018). Superpoint: Self-supervised interest point detection and description. In Proceedings of the IEEE conference on computer vision and pattern recognition workshops (pp. 224–236).

DJI Agras MG-1P RTK. URL https://www.dji.com/it/mg-1p. (26 March 2024).

DJI Matrice 300 RTK. URL https://enterprise.dji.com/it/matrice-300 (26 March 2024).

Lindenberger, P., Sarlin, P. E., & Pollefeys, M. (2023). Lightglue: Local feature matching at light speed. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 17627–17638).

Longhi, V., Martino, A., Lingua, A. M., Maschio, P. F., & Belcore, E. (2024). Monitoring the spread of a pathogenic insect on vineyards using UAS. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-1-2024, 443–450.

Matrone, F., Paolanti, M., Frontoni, E., & Pierdicca, R. (2024). Enhancing explainability of deep learning models for point cloud analysis: a focus on semantic segmentation. *International Journal of Digital Earth*, *17*(1). https://doi.org/10.1080/17538947.2024.2390457

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. Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci., XLVIII-2/W4-2024, 309–316.

Pádua, L., Vanko, J., Marques, P., Oliveira, R., & Sousa, J. J. (2022). Computer Vision and Deep Learning for Precision Viticulture: A Review. Agronomy, 12(10), 2463.

Ramyaa, R., Varadharajan, V., & Elavarasan, R. M. (2024). Vineyard Zoning and Vine Detection Using Machine Learning in Unmanned Aerial Vehicle Imagery. Remote Sensing, 16(3), 584.

Sarlin, P. E., DeTone, D., Malisiewicz, T., & Rabinovich, A. (2020). Superglue: Learning feature matching with graph neural networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4938–4947).

Tsouros, D. C., Stergiou, C., Bibi, S., Sarigiannidis, P., & Bouloumpasis, I. (2023). A Deep Learning Approach for Precision Viticulture, Assessing Grape Maturity. Sensors, 23(19), 8611.

Tyszkiewicz, M., Fua, P., & Trulls, E. (2020). DISK: Learning local features with policy gradient. Advances in Neural Information Processing Systems, 33, 14254–14265.

Williams, T., Byrne, J., & Pound, M. (2022). End-to-End Deep Learning for Directly Estimating Grape Yield from Ground-Based Imagery. arXiv preprint arXiv:2208.02394.