Aerial Image-based Inter-day Registration for Precision Agriculture

Aerial Image-based Inter-day Registration for Precision Agriculture

Chen Gao*[1], Franz Daxinger*[1], Lukas Roth[3], Fabiola Maffra[1], Paul Beardsley[2], Margarita Chli[1], Lucas Teixeira[1]
Vision For Robotics Lab - ETH Zürich, Switzerland and the University of Cyprus[1],
Unity Technologies[2], Institute of Agricultural Sciences - ETH Zürich, Switzerland[3]

***Abstract*— Satellite imagery has traditionally been used to collect crop statistics, but its low resolution and registration accuracy limit agricultural analytics to plant stand levels and large areas. Precision agriculture seeks analytic tools at near single plant level, and this work explores how to improve aerial photogrammetry to enable inter-day precision agriculture analytics for intervals of up to a month.**

Our work starts by presenting an accurately registered image time series, captured up to twice a week, by an unmanned aerial vehicle over a wheat crop field. The dataset is registered using photogrammetry aided by fiducial ground control points (GCPs). Unfortunately, GCPs severely disrupt crop management activities. To address this, we propose a novel inter-day registration approach that only relies once on GCPs, at the beginning of the season.

The method utilises LoFTR [1], a state-of-the-art image-matching transformer. The original LoFTR network was trained using imagery of outdoor urban areas. One of our contributions is to extend LoFTR's training method, which uses matching images of a static scene, to a dynamic scene of plants undergoing growth. Another contribution is a thorough evaluation of our registration method that integrates intra-day crop reconstruction with earlier-day scans in a seven degree-of-freedom alignment. Experimental results show the advantage of our approach over other matching algorithms and demonstrate the importance of retraining using crop scenes, and a training method customised for growing crops, with an average registration error of 27 cm across a season.

I. Introduction

Smart Farming is an invaluable development in agriculture. Traditional agriculture has been facing challenges such as soil erosion, nutrient runoff causing algal blooms and dead zones in surrounding bodies of water, a fall in soil organic carbon which affects fertility, and loss of biodiversity [2]. Smart Farming aims to address such challenges while still maintaining economic food production [3], [4].

Examples of Smart Farming include autonomous robots and drones for spraying, harvesting, and weeding, plus AI for farm management such as recommending the dates for irrigation and harvesting. Underpinning all of these technologies is the need for monitoring and automatic analysis of crops. Traditionally, satellite imagery has been the dominant imaging modality [5], with a typical image resolution of $\sim$10 meters per pixel, producing crop analysis at the scale of whole fields. Precision agriculture, a subdiscipline of smart farming, works at a finer scale [6] and often utilises aerial or ground imagery. A high spatial resolution is not only important for targeted maintenance like spraying of individual plants, but also because the recommended dates for crop treatments and harvesting can vary by days across the fields of even a relatively small farm of $\sim$50 Hectares. This paper addresses the problem of analysing crop imagery at the level of individual plants.

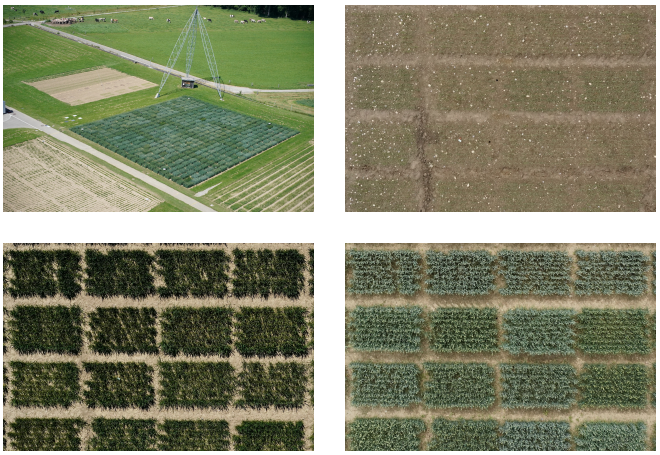


Fig. 1: **Top left:** an overview of the ETH plant research station in Switzerland. **Other images:** The same location on three different days, demonstrating the drastic changes in shape and appearance that happen in agricultural environments under growth.

Unmanned Aerial Vehicles (UAVs) are an important and affordable technology for Smart Farming, providing high spatial resolution imagery. The challenge which arises is how to match UAV imagery over time. A crop field contains self-similar structures leading to ambiguity in matching image features. Furthermore, growing plants exhibit changes in shape, height, and appearance. In addition, there is environmental change due to changing ambient lighting, and changing soil appearance due to moisture or crop management measures. Examples of environment visual changes can be seen in Figure 1. Traditional descriptors fail to find robust correspondences for such applications, and consumer-grade UAVs' GPS is not accurate enough for plant-level registration.

To tackle these problems, we present a novel inter-day registration framework to account for the growth of plants,

This work has been partly funded by the European Research Council (ERC), as part of the project SkEyes (Grant agreement no. 101089328) and by Unity Technologies. Data collection was supported by Innosuisse in the framework for the project 'Trait spotting' [Grant Number: KTI P-Nr 27059.2 PFLS-LS].

Code and Dataset: github.com/VIS4ROB-lab/interday_crop_registration
Video: https://youtu.be/RItJI8JfZsQ
*equal contribution

changes in appearance and the UAV GPS noise. In summary, the contributions of this work are the following:

- The application of a state-of-the-art matching algorithm LoFTR to agricultural imagery, including a demonstration of the value of retraining the network with agricultural data using our height compensation method.
- A novel approach to utilise LoFTR while also handling temporal change and the demonstration of matching for scans of wheat fields taken up to a month apart.
- We are publishing our code and the dataset of UAV imagery for wheat fields across two growing seasons, including meta-data for the camera pose and GCP Ground Control Points, as an asset for the community.

II. Related Work

Agricultural robots, ground-based and drones, are a route to boosting productivity and reducing costs while also enabling the adoption of more sustainable agriculture without requiring intensive and costly manpower to achieve it. Two key technologies are visual localization and image matching, allowing robots to capture and compare images of crops over time. Both topics have been investigated extensively for man-made environments but are still challenging for agricultural environments.

Visual Localization for Precision Agriculture: The problem of matching images of a crop field across time is addressed by Dong et al. [7]. Using a SLAM setup, they fuse data from different sensors to obtain a 4D model of a crop field and show that their approach is able to produce accurate models as long as the visual appearance does not change dramatically. Similarly, Marks et al. [8] use bundle adjustment and template matching at the plant level. [9] also use plant-level characteristics to perform registration. Another interesting approach to generating correspondences across time is presented by Chebrolu et al. [10]. To minimize challenges arising from changes in the visual appearance of the environment, they assume that crops are distinguishable and planted in rows such that they are able to use the field geometry and geometric feature descriptor. With this approach, they can collect statistics on a single plant level. The work by Kim et al. [11] introduces a methodology to improve the geometric registration of multi-temporal digital surface models without using Ground Control Points (GCPs). To accurately register these models, they rely on elevation invariant features, which can be found in areas next to the crop field with no growth (e.g. streets). While these approaches directly tackle the problem of localization and matching in the agriculture context, they usually make strong assumptions about the crops, making their use limited to very specific scenarios.

Generic Visual Localization: State-of-the-art visual localization approaches [12], [13], [14], [15], [16], [17] typically rely on a 3D scene representation and 2D-3D matches between a query image and the 3D representation to compute a camera pose for the query using a PnP solver in a RANSAC scheme. 2D-3D matches are commonly established using local features, and while hand-crafted features, such as SIFT [18], are widely used for matching images captured at similar conditions, they consistently fail when the scene appearance changes [19]. As such, deep features, such as SuperPoint [20] and LoFTR [1] have been largely used to tackle the problem of visual localization under extreme changes in appearance (e.g. time of the day, weather, seasons as well as human activity and occlusions) [13], [21]. Global descriptors, such as NetVLAD [22], are commonly used in an image retrieval step to improve the scalability of structure-based methods. Sarlin et al. [13] propose a state-of-the-art hierarchical approach for visual localization that leverages both global descriptors and local features, and scales well with large environments. The authors use NetVLAD and SuperPoint to train a small network using multi-task distillation, while feature matching is performed using SuperGlue [23]. More recently, scene coordinate regression methods determine the 2D-3D correspondences using random forests [24] or CNNs [25], [26]. In contrast, absolute pose regression methods [27] forego 2D-3D matching and train a network to predict a camera pose directly from an image.

All these methods typically assume that the scene is at least partially static, and the same features can be re-detected over time. However, crops grow over time, changing both the geometry and appearance of a field of crop plants. Furthermore, ambient light will vary anywhere between bright sunlight and the diffuse illumination of an overcast day, while field appearance is affected by dry versus wet conditions. In this work, we analyse the performance of state-of-the-art methods that were designed for static man-made scenes, modifying and applying them to growing crops. Specifically, we propose a geometrically correct training pipeline for re-training a learning-based method, LoFTR.

III. Method

This section presents an overall approach for creating 3D models of a crop field on different days and precisely aligning them between each other to enable precision agriculture tasks. We describe the core of our alignment procedure, which is a new training approach for enabling a state-of-the-art transformer-based feature-matching algorithm to be trained in a non-static agricultural environment.

A. Registration Pipeline

Figure 2 shows our overall registration approach. Our approach assumes that a UAV has been used to perform flights capturing GPS-tagged images suitable for photogrammetry-based 3D reconstruction. This means that a continuous sequence of images with high frontal and lateral overlap without blur was captured. We also assume that the vehicle has standard onboard GPS, which means that direct alignment using GPS coordinates is not possible, unlike when a more accurate RTK-GPS is onboard.

Firstly, each set of images from the same day, i.e. intra-day, is reconstructed using any photogrammetry software. We use COLMAP [28] for this task. It uses a traditional pipeline using SIFT features, RANSAC and incremental reconstruction. It works when the images are taken within an hour with

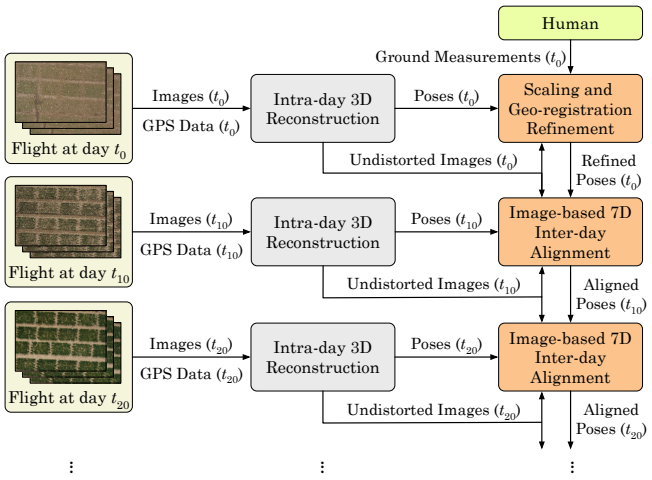


Fig. 2: The registration pipeline - Intra-day 3D reconstruction is performed individually for each day and after an inter-day alignment is deployed to enable analysis on the plants in the ground, the number of days between flights is application dependent. Here we show an example of ten days between flights.

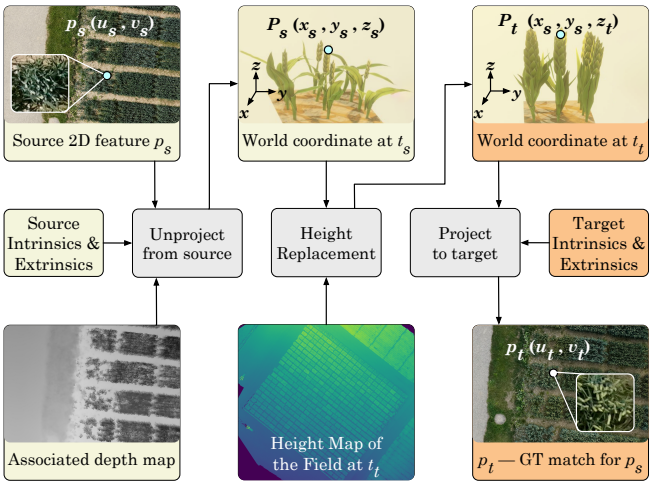


Fig. 3: Generation of ground truth matching pairs across time with height change. For a particular 2D feature p_s on an image from t_s, its ground truth (GT) match, p_t, on a certain image from t_t is found.

very similar illumination conditions. The resulting model is roughly geo-referenced, but this is potentially several meters off and non-metric.

In order to allow smart farming analytics [29], the first flight of the growing season, from day t_0, can be refined using human tuning and ground measurements, such as scalar constraints and ground control points [29]. Using a different UAV with RTK-GPS only for this flight is also an alternative. The flights in the next days after t_0, do not need human tuning or ground measurements. Our inter-day alignment approach is used instead. We indicate the day that the flight was performed as t_n. In this notation, the index n is the number of days since t_0.

This inter-day alignment assumes that the two models, target and source, to be aligned are at least roughly geo-registered. So for every image in the source, the feature matching is computed for the 10 closest images from the target. These matching are collectively used to compute the pose of the source image in the target model, following the approach in [13]. The poses of the source model images at the target model are used for the 7D alignment, i.e., translation, rotation and scale, using the method in [30].

At the core of our method is the inter-day feature matching that is detailed in the next section.

B. Inter-day feature matching

Our inter-day matching is done using a state-of-the-art matching transformer [1], LoFTR. Instead of sequentially performing image feature detection, description, and matching, LoFTR first establishes coarse-level pixel-wise dense matches and later refines them at a finer level using Transformer's self and cross attention layers. This enables LoFTR to produce dense matches even in low-texture areas where conventional feature detectors struggle to produce repeatable interest points. This out-of-the-box matching algorithm is trained on ScanNet [31], an indoor RGBD dataset, or MegaDepth [32], a large dataset of outdoor touristic places in the work built using photogrammetry. However, both datasets picture little to no natural environments. Therefore, we retrain LoFTR using a new dataset that will be presented in the next section. Unfortunately, the original training method cannot be directly applied, as it assumes a static scene, which is not suitable for our dynamic agricultural application. As a result, we introduce a refined training pipeline. Specifically, we introduce a new supervision method for creating ground truth matches during training to account for dynamic changes such as crop growth.

Our updated training pipeline assumes the availability of RGB images, high-quality poses, and a 3D mesh of the area for each dataset captured on different days. First, we compute a metric depth map for each RGB image using a custom Vulkan-based renderer. A height map of the whole area is also computed for the whole area for later use.

Figure 3 illustrates the procedure for generating ground-truth matches during training for 2D features. For every 2D feature p_s in the source image from day t_s, its corresponding 3D landmark $P_s(x_s, y_s, z_s)$ is computed using camera intrinsics, extrinsics, and the depth in the same pixel location in the associated depth map. We then replace the z-component of the landmark with the height value z_t, which is obtained from the height map of the target dataset at day t_t at the same x and y coordinates (x_s, y_s). This compensates for the height changes in the crop over time. The resulting adjusted 3D landmark P_t is then projected onto the target image from t_t, yielding a reliable ground truth correspondence p_t.

In summary. our method of retraining LoFTR adapts to the dynamic agricultural environment, whereas the original training focused on static scenes. This is achieved via a unique dataset and a custom training pipeline that accounts for height variations between different days, enabling accurate inter-day feature matching.

IV. Dataset

A. Agricultural experiment

The presented dataset is from flight campaigns conducted in 2018-2019 at the ETH plant research station in Eschikon-

Lindau, Switzerland (47.449°N, 8.682°E, 520 m.a.s.l.). Each flight campaign consists of a sequence of mapping flights that are used to monitor an agricultural experiment (35 x 40 m) over the entire growing season (from sowing to harvest). The experiments were breeding-related winter wheat experiments, which are further described in [33], [4].

B. Flight campaigns

On average, 40 mapping flights were performed per growing season using a DJI Matrice 600 Pro drone equipped with a Sony ILCE-9 camera with a full-frame sensor of 6000 × 4000 pixels. The flight height was 28 m, the flight speed was 1.8 m/s, the end overlap was 92%, and the side overlap was 75%. Ground sampling distance (GSD) was below 3 mm/pixel. The raw data can be found [34]

C. Ground reference for evaluation

Fields were prepared with crosswise ground control point (GCP) arrangements, and GCP positions were measured using a GNSS differential global position system (Trimble R10) with swipos-GIS/GEO RTK (real-time kinematic) correction, resulting in an accuracy of 8 mm horizontally and 15 mm vertically. At the four corners of the field, coded round GCPs with a 0.5 m diameter were used. At all other positions, uncoded squared GCPs sized 0.2 × 0.2 m were placed.

D. Preprocessing

Images were processed per mapping flight with the structure-from-motion (SfM) software Agisoft PhotoScan and Metashape. For the sparse point cloud processing, the keypoint limit was set to 40,000. Automatic GCP detection was used in an automated two-step process: first, coded GCPs were detected and matched with known coordinates using coded names, leading to a roughly georeferenced point cloud. Second, uncoded GCPs were detected and matched with the known coordinates of the closest uncoded GCPs, resulting in a precisely georeferenced point cloud. The SfM process was continued with dense point cloud processing to get a digital surface model. The digital elevation model (DEM) was generated using a GSD of 3 mm [29].

V. Experiments

This section presents practical and experimental aspects of our research. First, we present the implementation details of retraining LoFTR using our proposed supervision method for inter-day feature matching on the crop dataset in Section IV. We then compare the retrained LoFTR against state-of-the-art feature-matching algorithms in different temporal distances between query and reference flight data. Finally, we experiment with the full proposed registration pipeline over the whole season, considering the cumulative error while variating the matching algorithm.

A. Implementation details of retraining LoFTR on the crop dataset

We select a total of 12 dates from the 2018 dataset, spaced approximately 10 days apart, covering the entire growing season from March 22, 2018, to July 4, 2018. Images are cropped to 3000 × 2000 pixels around the original image center. To generate ground truth image pairs for training, for each image in the selected dataset, we created pairs with others that were less than 23 days apart temporally and had an overlap greater than 50% spatially on the covered ground area. This results in images from each date being typically paired with images from three target dates: the source date itself, a date roughly 10 days later, and another around 20 days later. We generate 15,910 image pairs, using 80% for training and the remaining 20% for validation.

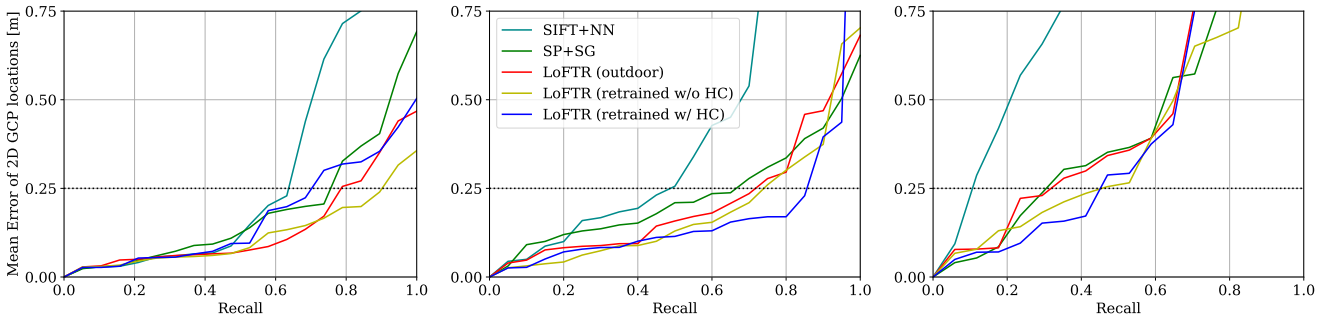
The generated image pairs then undergo LoFTR's training pipeline, utilizing ground truth matches as supervision. We employed two versions of supervision: one following the original approach of LoFTR [1], where ground truth matching pairs were generated by directly projecting 3D landmarks associated with 2D features into the target image frame, and the other implementing the method described in Section III-B, which replaced the z-component of the 3D landmark with corresponding height from the height map of the target dataset. The height map we use is created using the dense point cloud. It has a resolution of 6000 × 6000 pixels, where each pixel represents 1 cm × 1 cm on the ground.

Based on these ground truth matches, we trained two LoFTR models with both versions of supervision respectively, with training conducted on an NVIDIA Titan X. We initiated training using outdoor weights trained with the dual softmax technique. All images are further resized to 360 × 240 to accommodate a batch size of 4. Training consists of 30 epochs for both versions, and the weights from the epoch achieving the highest precision on the validation set were used for subsequent testing. All not specified settings followed the original paper [1].

B. Evaluation of Inter-Day Feature Matching Algorithms

In this section, we assess the effectiveness of our retrained LoFTR model as an inter-day feature matcher. Specifically, we evaluate its capacity to align a query model, which possesses only rough geo-referencing with a potential error of several meters, with a reference model refined through ground truth image positions. To provide a comprehensive evaluation, we compare our LoFTR retrained with both versions of supervision against three state-of-the-art feature matching algorithms:

1) **SIFT+NN:** SIFT feature extractor [18] (resized to 1600 pixels), with Nearest Neighbor matcher (mutual check, ratio threshold 0.8).
2) **SP+SG:** Superpoint feature extractor [20] (4096 keypoints, NMS radius 3, resized to 1024 pixels), with SuperGlue matcher [23] (outdoor model).
3) **LoFTR (outdoor):** Dense feature matching with the out-of-the-box LoFTR [1] (outdoor model, dual softmax, resized to 1024 pixels).
4) **LoFTR (retrained w/o HC):** Our retrained LoFTR without height change (resized to 1024 pixels).
5) **LoFTR (retrained w/ HC):** Our retrained LoFTR with height change (resized to 1024 pixels).



(a) 5 ± 2 days prior to the query date (b) 15 ± 2 days prior to the query date (c) 30 ± 2 days prior to the query date

Fig. 4: Comparison of our inter-day feature matching method against state-of-the-art approaches at varying time intervals between the query and reference models. Lower values indicate superior performance. Following the convention from previous model alignment tasks like [35], we use Error-Recall graphs to display the mean error between the 2D GCP locations of the aligned query model and the ground truth. The legend is presented solely on the middle plot to enhance readability, and consistent colors are used across all plots for method identification.

	Error of GCP locations, 3D [m]					Error of GCP locations, 2D [m]				
	SIFT+NN	SP+SG	LoFTR (outdoor)	LoFTR (retrained w/o HC)	LoFTR (retrained w/ HC)	SIFT+NN	SP+SG	LoFTR (outdoor)	LoFTR (retrained w/o HC)	LoFTR (retrained w/ HC)
t_0 - t_{16}	0.147±0.032	**0.047**±0.013	0.178±0.029	0.161±0.027	0.166±0.026	0.100±0.040	**0.039**±0.014	0.079±0.031	0.075±0.028	0.077±0.030
t_{16} - t_{33}	0.183±0.073	0.393±0.064	0.115±0.031	0.079±0.028	**0.058**±0.023	0.170±0.085	0.200±0.094	0.075±0.031	0.058±0.032	**0.048**±0.022
t_{33} - t_{49}	0.318±0.059	0.590±0.068	0.421±0.122	0.190±0.062	**0.168**±0.056	0.268±0.052	0.306±0.120	0.246±0.100	0.136±0.049	**0.124**±0.043
t_{49} - t_{65}	0.832±0.156	0.700±0.072	0.931±0.154	**0.363**±0.075	0.402±0.097	0.651±0.237	0.418±0.178	0.904±0.178	0.347±0.087	**0.342**±0.093
t_{65} - t_{82}	-	0.957±0.140	1.009±0.135	1.128±0.243	**0.783**±0.173	-	**0.708**±0.230	0.963±0.144	0.764±0.375	0.750±0.178
t_{82} - t_{96}	-	1.099±0.132	1.099±0.290	1.261±0.241	**0.275**±0.132	-	0.781±0.314	1.035±0.318	0.937±0.428	**0.273**±0.130
t_{96} - t_{114}	-	1.200±0.116	1.179±0.281	1.425±0.230	**0.309**±0.103	-	0.868±0.296	1.009±0.353	0.974±0.444	**0.266**±0.119
mean	-	0.712	0.705	0.658	**0.308**	-	0.474	0.616	0.470	**0.269**

TABLE I: Evaluation of our time-sequential registration pipeline, starting with the initial model constructed from images on t_0 (March 13, 2019), geo-registered using ground truth image locations. Subsequent model alignments rely exclusively on our inter-day alignment method, with the previously aligned model serving as the reference. Reported are the mean 3D and 2D errors of GCP locations for the aligned models. '-' indicates when an algorithm fails in the respective registration task.

Given that we trained our models on the 2018 dataset, we aim to evaluate their performance on a different year. To achieve this, we assess the registration performance by varying the time intervals between the query and reference models, utilizing the crop dataset captured in 2019. The process for constructing the testing data is as follows:

1) Randomly select 21 query dates from the 2019 dataset.
2) For each query date, reconstruct a query model from the available images, which is then geo-referenced using GPS locations. As true GPS data from flights is unavailable, we simulate GPS locations by introducing random Gaussian noise (with a variance of 5 meters and a zero mean) to the 3D coordinates of the ground truth image locations provided in the dataset.
3) For each query model, create three reference reconstructions, each corresponding to a specific time frame relative to the query date, approximately 5 days, 15 days, and 30 days before the query date, respectively. For each time frame, select a reference date with a difference to the query date within the range of $k \pm 2$ days ($k = 5, 15, 30$), and construct a reference model using images from that date.
4) Apply the registration pipeline to align the query and reference models.

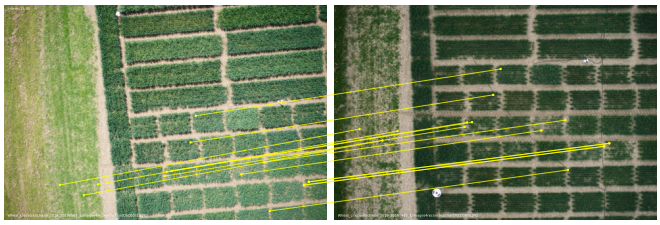
To assess the performance of our inter-day feature matching method, we calculated the mean errors of Ground Control Points (GCPs), specifically the 2D error (xy-component of the GCP location error), for each aligned query model against the ground truth.

Figure 4 illustrates the Error-Recall graphs for all three time frames. We select a threshold of 0.25 meters, considering the average planting spacing of winter wheat, which falls approximately within the range of 17.5-35 cm [36]. When the query model and reference model were closely spaced (3-7 days apart), all algorithms achieved relatively low ground error within the 0.25-meter threshold. Notably, our LoFTR retrained without height change exhibited a slight advantage, likely because the plant heights exhibited minimal appearance and height change during this period.

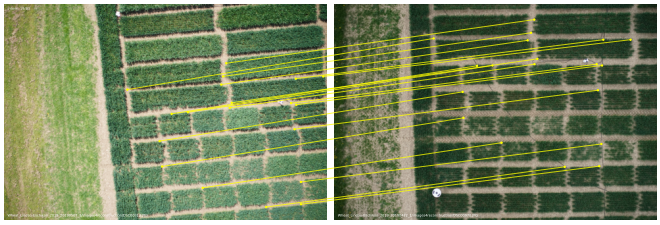
As the time interval increased to around 15 days, a noticeable disparity in performance emerged. SIFT+NN, for example, primarily looking for invariant features related to field geometry, struggled to achieve only approximately 50% recall at the 0.25-meter ground error threshold. In contrast, our retrained LoFTR with height change excelled, achieving approximately 85% recall—more than 10% higher than the second-best performer. Furthermore, our method consistently maintained lower errors for most recall levels below 0.25 m, indicating its effectiveness in achieving minimal errors at comparable recall rates. We believe that this advantage is because, during a period of around 15 days, the visual appearance changes moderately, allowing our method to capitalize on its height-aware matching approach. It strikes a balance, achieving accurate inter-day feature matching

without being hindered by excessive changes or underutilized when changes are minimal, making it the most effective choice for this critical time frame.

As the time interval increased to approximately one month, all methods experienced a significant drop in recall rates, falling below 50%. While our retrained model with height-compensated still outperformed others, the recall rate dropped to a point where the registration effectiveness may be compromised. Additionally, the gap between LoFTR with and without height change remained relatively small, possibly due to significant appearance changes hindering the establishment of correspondences between 2D features of a 3D landmark.

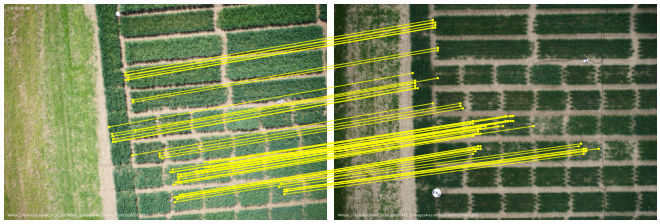


(a) SIFT+NN

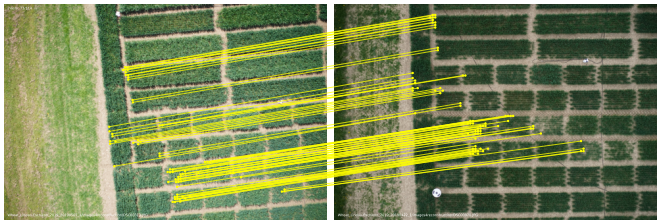


(b) SP+SG

(c) LoFTR (outdoor)



(d) LoFTR (retrained w/o HC)



(e) LoFTR (retrained w/ HC)

Fig. 5: Matching between image pairs taken 15 days apart using different algorithms.

Figure 5 showcases matching results for all five algorithms using example image pairs taken 15 days apart with significant appearance changes. SIFT+NN, SP+SG, and the out-of-the-box LoFTR identified relatively few matches, while our LoFTR retrained on the crop dataset found significantly more matches. Notably, the model considering height change found even denser matches, demonstrating its ability to handle considerable appearance variations.

C. Evaluation of sequential registration across the season

In this section, we leverage our proposed inter-day feature matching method within the pipeline outlined in Figure 2, which is equivalent to actual use in precision agriculture throughout a crop-growing season. We sequentially execute the registration pipeline, refining only the first model with human interference, specifically GCP. The remaining days rely on our inter-day alignment approach.

For simplicity, we denote the first date, March 13, 2019, as t_0, and subsequent flight days as t_n, where n represents the number of days since t_0. We consider intervals between flights of approximately 16 days, which align with the time frame where our retrained LoFTR exhibited optimal performance in the previous subsection.

We report the 3D and 2D error of GCP locations of the aligned models in Table I. Results reveal that apart from the initial alignment at t_0-t_{16}, where our method was outperformed by SP+SG, our approach consistently excelled, or fell within one standard deviation of the best-performing algorithm's error in most cases. Importantly, our approach does not rely on specific field geometries, as demonstrated by our choice of training set from 2018 and testing set from 2019, where field geometry varied. This demonstrates the potential to train our method on data from a given year and apply it to subsequent years, enabling continuous plant monitoring for smart farming.

VI. Conclusion

This paper has described feature matching over time in a wheat field captured from a UAV camera resolving features in the millimetre range. The matching problem is challenging because of the self-similarity of single plants and the changing appearance of the environment and plant stands over time, hence matching ambiguity. We utilised LoFTR for matching and demonstrated an improvement in accuracy by retraining the original network with crop imagery.

While our approach contributes to improved performance in various scenarios, we acknowledge the need for further investigation and refinement to enhance the model's effectiveness in handling more prolonged temporal changes, extending beyond a month. Furthermore, experiments were done on imagery of a wheat crop as a starting point. For other crops, the visual appearance and morphological changes with time will differ. Specific management parameters such as sowing density and row spacing may introduce additional variation. Consequently, adapting the methodology to new crops will be required. Another area for improvement is

distinguishing features on the ground from those on the plant canopy level to enhance the registration in the vertical direction, thereby allowing for tracking of the plant's size.

REFERENCES

[1] J. Sun, Z. Shen, Y. Wang, H. Bao, and X. Zhou, "Loftr: Detector-free local feature matching with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8922–8931.

[2] E. S. D. Centre, "Linking soil degradation processes, soil-friendly farming practices and soil-relevant policy measures," https://esdac.jrc.ec.europa.eu/projects/SOCO/FactSheets/EN2009, accessed: 2023.

[3] A. Walter, R. Finger, R. Huber, and N. Buchmann, "Smart farming is key to developing sustainable agriculture," *Proceedings of the National Academy of Sciences*, vol. 114, no. 24, pp. 6148–6150, 2017.

[4] L. Roth, D. Fossati, P. Krähenbühl, A. Walter, and A. Hund, "Image-based phenomic prediction can provide valuable decision support in wheat breeding," *Theoretical and Applied Genetics*, vol. 136, no. 7, p. 162, June 2023.

[5] A. Mora, T. M. Santos, S. Łukasik, J. M. Silva, A. J. Falcão, J. M. Fonseca, and R. A. Ribeiro, "Land cover classification from multi-spectral data using computational intelligence tools: A comparative study," *Information*, vol. 8, no. 4, p. 147, 2017.

[6] R. Gebbers and V. Adamchuk, "Precision agriculture and food security," *Science*, vol. 327, pp. 828 – 831, 2010.

[7] J. Dong, J. G. Burnham, B. Boots, G. Rains, and F. Dellaert, "4d crop monitoring: Spatio-temporal reconstruction for agriculture," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 3878–3885.

[8] E. Marks, F. Magistri, and C. Stachniss, "Precise 3d reconstruction of plants from uav imagery combining bundle adjustment and template matching," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 2259–2265.

[9] H. Pan, F. Hétroy-Wheeler, J. Charlaix, and D. Colliaux, "Multi-scale space-time registration of growing plants," in *2021 International Conference on 3D Vision (3DV)*, 2021, pp. 310–319.

[10] N. Chebrolu, T. Läbe, and C. Stachniss, "Robust long-term registration of uav images of crop fields for precision agriculture," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3097–3104, 2018.

[11] T. Kim, J. Park, C. Lee, Y. Yun, J. Jung, and Y. Han, "Multi-temporal orthophoto and digital surface model registration produced from uav imagery over an agricultural field," *Geocarto International*, pp. 1–24, 2022.

[12] F. Maffra, L. Teixeira, Z. Chen, and M. Chli, "Real-time wide-baseline place recognition using depth completion," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1525–1532, 2019.

[13] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 716–12 725.

[14] P.-E. Sarlin, A. Unagar, M. Larsson, H. Germain, C. Toft, V. Larsson, M. Pollefeys, V. Lepetit, L. Hammarstrand, F. Kahl, *et al.*, "Back to the feature: Learning robust camera localization from pixels to pose," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3247–3257.

[15] V. Panek, Z. Kukelova, and T. Sattler, "Visual localization using imperfect 3d models from the internet," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 175–13 186.

[16] J. Surber, L. Teixeira, and M. Chli, "Robust visual-inertial localization with weak gps priors for repetitive uav flights," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 6300–6306.

[17] F. Maffra, L. Teixeira, Z. Chen, and M. Chli, "Loop-closure detection in urban scenes for autonomous robot navigation," in *2017 International Conference on 3D Vision (3DV)*, 2017, pp. 356–364.

[18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, p. 91–110, nov 2004.

[19] C. Toft, W. Maddern, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, T. Pajdla, *et al.*, "Long-term visual localization revisited," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 4, pp. 2074–2088, 2020.

[20] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[21] P.-E. Sarlin, M. Dusmanu, J. L. Schönberger, P. Speciale, L. Gruber, V. Larsson, O. Miksik, and M. Pollefeys, "Lamar: Benchmarking localization and mapping for augmented reality," in *European Conference on Computer Vision*. Springer, 2022, pp. 686–704.

[22] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[23] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.

[24] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon, "Scene coordinate regression forests for camera relocalization in rgb-d images," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2930–2937.

[25] S. Tang, C. Tang, R. Huang, S. Zhu, and P. Tan, "Learning camera localization via dense scene matching," in *Computer Vision and Pattern Recognition (CVPR)*, 2021.

[26] Q. Yan, J. Zheng, S. Reding, S. Li, and I. Doytchinov, "Crossloc: Scalable aerial localization assisted by multimodal synthetic data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 358–17 368.

[27] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixe, "Understanding the limitations of cnn-based absolute camera pose regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3302–3312.

[28] J. L. Schönberger, "Robust methods for accurate and efficient 3d modeling from unstructured imagery," Ph.D. dissertation, ETH Zurich, 2018.

[29] L. Roth, M. Camenzind, H. Aasen, L. Kronenberg, C. Barendregt, K.-H. Camp, A. Walter, N. Kirchgessner, and A. Hund, "Repeated Multiview Imaging for Estimating Seedling Tiller Counts of Wheat Genotypes Using Drones," *Plant Phenomics*, vol. 2020, no. 3729715, 2020.

[30] J. L. Schönberger and J.-M. Frahm, "Structure-from-motion revisited," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[31] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[32] Z. Li and N. Snavely, "Megadepth: Learning single-view depth prediction from internet photos," in *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[33] L. Roth, C. Barendregt, C.-A. Bétrix, A. Hund, and A. Walter, "High-throughput field phenotyping of soybean: Spotting an ideotype," *Remote Sensing of Environment*, vol. 269, p. 112797, 2022.

[34] L. Roth and A. Hund, "Trait spotting wheat and soybean dataset: Rgb mapping flights." ETH Zurich, 2024, no. 10.3929/ethz-b-000660039.

[35] H. Izadinia and S. M. Seitz, "Scene recomposition by learning-based icp," in *CVPR*, 2020.

[36] M. Abichou, B. de Solan, and B. Andrieu, "Architectural response of wheat cultivars to row spacing reveals altered perception of plant density," *Frontiers in Plant Science*, vol. 10, 2019.