IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Keypoint Semantic Integration for Improved Feature Matching in Outdoor Agricultural Environments

Rajitha de Silva[1] Jacob Swindell[1] Jonathan Cox[1] Marija Popović[2]
Cesar Cadena[3] Cyrill Stachniss[4] Riccardo Polvara[1]

***Abstract*— Robust robot navigation in outdoor environments requires accurate perception systems capable of handling visual challenges such as repetitive structures and changing appearances. Visual feature matching is crucial to vision-based pipelines but remains particularly challenging in natural outdoor settings due to perceptual aliasing. We address this issue in vineyards, where repetitive vine trunks and other natural elements generate ambiguous descriptors that hinder reliable feature matching. We hypothesise that semantic information tied to keypoint positions can alleviate perceptual aliasing by enhancing keypoint descriptor distinctiveness. To this end, we introduce a keypoint semantic integration technique that improves the descriptors in semantically meaningful regions within the image, enabling more accurate differentiation even among visually similar local features. We validate this approach in two vineyard perception tasks: (i) relative pose estimation and (ii) visual localisation. Our method improves matching accuracy across all tested keypoint types and descriptors, demonstrating its effectiveness over multiple months in challenging vineyard conditions.**

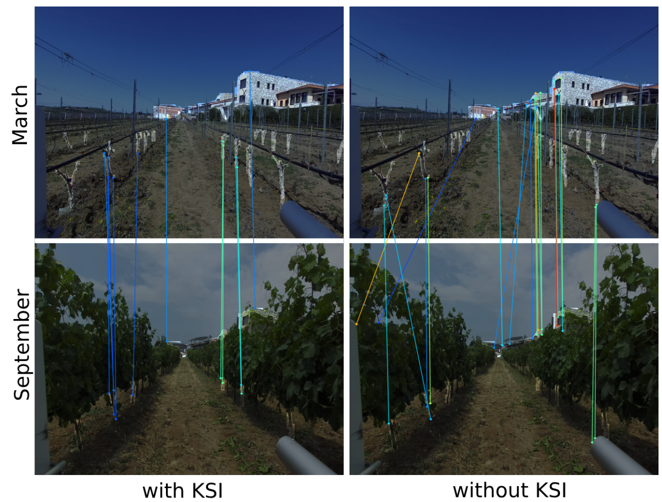

Fig. 1: KSI enhances the keypoint descriptors with corresponding semantic instances to minimise aliasing during keypoint matching in vineyards. Keypoint matches on semantic masks are visualised here across March and June.

I. Introduction

Modern agricultural systems require robust robotic solutions to automate labour-intensive tasks and enhance operational efficiency. Among these, vision-based perception gaining traction due to its affordability. For autonomous mobile robots operating in outdoor agricultural environments, perception systems must accurately detect and track visual features for tasks such as relative pose estimation and visual localisation. However, vineyards present repetitive structures—vine trunks, poles, water pipes—that complicate feature correspondence between consecutive frames. Both traditional and learned keypoint descriptors focus on local regions, lacking broader semantic context, which can lead to false-positive matches for spatially distant but visually similar keypoints. This challenge, known as perceptual aliasing, remains a major hurdle towards dependable vision-based navigation in outdoor settings.

In this paper, we address this challenge by exploiting the semantic and morphological properties of repetitive regions to reduce descriptor ambiguity and improve matching accuracy. Our approach enriches keypoint descriptors with semantic context to increase their distinctiveness, focusing on vineyard environments where repetitive structures are widely recognised yet rarely supplemented with semantic cues for descriptor differentiation [1]. Vine trunks, for instance, appear visually similar, producing nearly identical descriptors; however, morphological variations (Fig. 1) can be encoded with semantic embeddings to refine keypoint representations. This approach supports accurate keypoints matching across multiple months, even under significant seasonal changes.

Typical feature matching pipelines rely on heuristic filtering [1] or incorporate context at the matching stage [2], but they still struggle with uniform descriptors and repetitive spatial layouts [3], [4]. We propose keypoint semantic integration (KSI), a framework for enhancing the keypoint descriptors with semantic mask instances on which they are located. An autoencoder [5] compresses the semantic mask into a feature vector, generating a semantic embedding that enhances descriptors' distinctiveness. Consequently, keypoints on semantically meaningful regions gain enriched descriptors, while those in the background retain their original descriptors.

Our contributions are: (i) a semantic integration method, KSI, that mitigates perceptual aliasing in visually repetitive agricultural environments; (ii) an extensive ablation and evaluation study demonstrating the method's robustness under seasonal changes in vineyards; and (iii) the Semantic

[1]Rajitha de Silva, Jacob Swindell, Jonathan Cox and Riccardo Polvara are with Lincoln Centre for Autonomous Systems (L-CAS), University of Lincoln, UK. [2]Marija Popović is with MAVLab, TU Delft, Netherlands. [3]Cesar Cadena is with Robotics Systems Lab, ETH Zurich, Switzerland. [4]C. Stachniss is with the University of Bonn, Germany. (Corresponding author: Rajitha de Silva [1]odesilva@lincoln.ac.uk)

This work was supported by Engineering and Physical Sciences Research Council (EPSRC), UK Project: "GAIA: Ground-Aerial maps Integration for increased Autonomy outdoors" (EPSRC Reference: EP/Y003438/1).

Bacchus Long Term (SemanticBLT) dataset with panoptic segmentation annotations in vineyards. Code and dataset available at: `https://lcas.github.io/KSI/`.

II. Related Work

Visual feature extraction and matching are essential first steps in a traditional visual simultaneous localisation and mapping (SLAM) pipeline. Classic feature extractors, such as Scale Invariant Feature Transform (SIFT) [6] and Oriented FAST and Rotated BRIEF (ORB) [7], remain popular for their rotation and scale invariance and computational efficiency, while deep learning-based feature extractors like SuperPoint [8], and Reliable and Repeatable Detector Descriptor (R2D2) [9]) prioritise robustness, repeatability, and reliability. However, neither traditional nor deep learning–based descriptors inherently capture the semantic context of a keypoint, instead relying on local neighbourhood information of the keypoint. Variants such as Semantic SuperPoint [10] combine semantic information with descriptor learning, but must be retrained for each new semantic class, hindering generalisation capabilities. Other approaches integrate semantic cues at the matching stage or via heuristic filtering [1], adapt descriptors online based on the current environment [11], or exploit sequence information to address repetitive scenes in agricultural environments [12].

Keypoint matching in agricultural environments poses significant challenges due to the scarcity of unique landmarks, repetitive crop appearances, lightning changes, and uneven terrain [3], [4]. Seasonal changes and foliage growth [13] exacerbate these issues, leading to perceptual aliasing and matching errors. Classic matching algorithms, such as nearest neighbour search, struggle to address these issues effectively. Ways to address this include exploiting geometric structures of keypoints [3] or exploiting background knowledge of the agricultural domain [12]. Learning-based matches e.g., SuperGlue [2], LightGlue [14] harness graph neural networks (GNNs) or transformers to incorporate local context for more robust matching. Tanco et al. [15] highlight the limitations of traditional detection algorithms through an agriculture-specific feature descriptor and matching method. However, vineyard-specific solutions remain limited [16]. Papadimitriou et al. [1] address this issue by restricting keypoint extraction to semantically significant regions, such as vine trunks, which retain a relatively consistent appearance over time and remain visible despite foliage growth. Their approach, however, simply masks these regions rather than integrating explicit semantic context into the descriptors.

The use of semantic cues in keypoint detection and matching has been explored in a variety of contexts, including urban environments [17], remote sensing [18], robotic manipulation [19]. and agriculture [15], where semantic information is learned from different object classes. Yet most existing semantic keypoint extractors rely on a multi-task backbone [20] with separate heads for semantic labels and keypoint descriptors, limiting them to the classes encountered during training. Tanco et al. [15] observe that, while their approach is trained with a semantic decoding head, its performance in inferring semantic labels at test time remains suboptimal suggesting opportunities to better understand the impact of this multi-task setup on keypoint descriptors compared to standard descriptor-based methods. Xue et al. [21] introduced Semantic-guided Feature Detection and Description (SFD2), which learns selective keypoint detection based on 120 semantic classes during training. However, plants are treated as dynamic objects, and the feature extractor penalises keypoint generation on trees. As a result, SFD2 produces relatively fewer keypoints on plants, which may limit its suitability for outdoor agricultural environments. These limitations accentuate the need for a modular approach that extends easily to new semantic classes, improving generalisation in specialised contexts such as agriculture.

Datasets that represent vineyard-specific semantic classes are crucial for integrating semantic information into keypoint descriptors. The ARD-VO dataset [22] contains vineyard and olive grove images but offers no semantic segmentation and covers only August to October, when the foliage is fully developed. In contrast, the Bacchus Long Term dataset [23] spans from March to September, capturing conditions both before and after major foliage growth. Its initial release (BLT) lacked annotations, so we extend it to Semantic-BLT, adding images and matching semantic masks for training segmentation models. Large, multi-month datasets are critical for robust performance in low-light and dense vegetation [24]; Semantic-BLT fills this gap in vineyard environments.

Unlike existing methods, our approach provides a modular way to embed semantic cues in keypoint descriptors, supporting both classical and learned descriptors as well as diverse matching algorithms. We use a stand-alone panoptic segmentation model to produce instance masks, which are then converted into semantic embeddings. Unlike systems requiring retraining of keypoint extractors for specific semantic classes, ours requires only semantic segmentation, thus offering flexibility while boosting descriptor distinctiveness in semantically crucial vineyard regions. These enriched descriptors, in turn, enhance both pose estimation and visual localisation in vineyard environments.

III. Our Approach to Keypoint Semantic Integration

We generate semantically enhanced descriptors by adding semantic embeddings, generated with an autoencoder trained on instance masks, to the original descriptor. In our KSI architecture (Fig. 2), we apply these enhancements only to keypoints falling on semantic instance masks, while background keypoints retain their original descriptors. Although developed for vineyard environments, this method could generalise to other visually repetitive domains, such as orchards, forests, or scree slopes.

A. SemanticBLT Dataset

Vineyards undergo noticeable visual change, requiring a multi-month dataset for segmentation models to handle diverse foliage, lighting, and environmental conditions. To this end, the SemanticBLT dataset labels six classes (buildings,

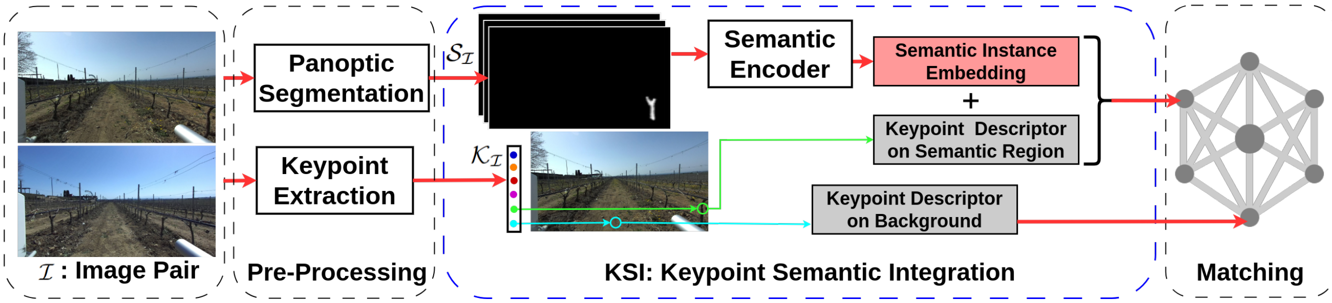

Fig. 2: Overview of proposed KSI framework: Input semantic instances $\mathcal{S}_\mathcal{I}$ and the set of keypoints $\mathcal{K}_\mathcal{I}$ are extracted from each image in the input image pair $\mathcal{I}$. The descriptors of keypoints on a semantic instance are combined with the corresponding semantic embedding to generate semantically enriched keypoint descriptors while the background descriptors are left unaltered. The semantically enriched descriptors and the background descriptors are matched together in a shared matcher.

pipes, poles, robots, trunks, and vehicles) across 1035 images captured from March to September in a single vineyard.

The data collection procedure employed the Thorvald robotic platform with a forward-facing ZED2 camera among other sensors. Whereas the BLT dataset [23] focused on lidar and unannotated image data, SemanticBLT offers semantically segmented imagery suited for vision-based tasks in repetitive agricultural environments.

B. Keypoint Extraction

The KSI architecture enriches existing keypoints by embedding semantic instance mask information in a modular manner compatible with various keypoint extractors. We use SuperPoint descriptors by default, but R2D2 [9], ORB [7], and SIFT [6] can also be integrated into SuperGlue-based feature matching [2]. In contrast, LightGlue matcher required pre-trained weights for each descriptor type, limiting its adaptability. From an image pair $\mathcal{I}$, we extract the keypoints $K_I = \{(pi, d_i)\}$, where $p_i \in \mathbb{R}^2$ represents the spatial coordinates and $d_i \in \mathbb{R}^D$ the corresponding descriptors.

C. Panoptic Segmentation in Vineyards

To capture raw semantic and spatial information, we apply a YOLOv9 [25] instance segmentation model trained on SemanticBLT to generate masks for six vineyard-relevant classes (buildings, pipes, poles, robots, trunks, vehicles). The panoptic segmentation component is modular, allowing the YOLOv9 model to be replaced with any other segmentation model without affecting the overall implementation. For KSI, we focus on trunks and buildings, whose distinct shapes and textures support accurate matching. In contrast, pipes and poles are excluded due to uniform shapes that often produce mismatches, while robots and vehicles are ignored since they are not guaranteed to remain static. Figure 3 illustrates the variation of all these semantic classes across multiple months. This panoptic segmentation step outputs a set of instance masks $\mathcal{S}_\mathcal{I}$ from the image pair $\mathcal{I}$.

D. Semantic Instance Encoding

We partition keypoints $\mathcal{K}_\mathcal{I}$ into two subsets: those on semantic masks $\mathcal{K}_\mathcal{S}$ and those in the background $\mathcal{K}_\mathcal{B}$. Each keypoint in $\mathcal{K}_\mathcal{S}$ is assigned a semantic embedding via a convolutional autoencoder $[f_{enc}, f_{dec}]$ trained on SemanticBLT instance masks. The encoder f_{enc} maps an instance masks s_i to a feature vector capturing morphological cues (shape, size, position). We add this embedding to the original descriptor, then L2-normalise $(N_{L2}(\cdot))$ to yield enhanced descriptors $\mathcal{K}'_\mathcal{S}$, preserving the original descriptor dimensions $D \times 1$:

$$\begin{aligned}\mathcal{K}'_\mathcal{S} &= \{(\mathbf{p}_i, \mathbf{d}'_i)\} \\ &= \{(\mathbf{p}_i, N_{L2}(\mathbf{d}_i + f_{enc}(\mathbf{s}_i))), \forall (\mathbf{p}_i, \mathbf{d}_i) \in \mathcal{K}_\mathcal{S}\}.\end{aligned} \tag{1}$$

E. Descriptor Matching

The KSI framework, illustrated in Fig. 2, produces two sets of keypoints – $\mathcal{K}_\mathcal{B}$ (background) and $\mathcal{K}'_\mathcal{S}$ (semantically enriched) – unified into a single descriptor collection $\widetilde{\mathcal{K}}$. Although many matchers (e.g. SuperGlue) assume a homogeneous set of descriptors, we concatenate the background and semantically enhanced descriptors for two images A, B:

$$\widetilde{\mathcal{K}}^A = [\mathcal{K}_\mathcal{B}{}^A, \mathcal{K}'_\mathcal{S}{}^A] \quad \text{and} \quad \widetilde{\mathcal{K}}^B = [\mathcal{K}_\mathcal{B}{}^B, \mathcal{K}'_\mathcal{S}{}^B],$$

and feed them in a single pass. Let $\mathbf{k}_i^A = (\mathbf{p}_i^A, \mathbf{d}_i^A)$ and $\mathbf{k}_j^B = (\mathbf{p}_j^B, \mathbf{d}_j^B)$ denote the keypoints (with positions and descriptors) in images A and B, respectively. Using a (learned or classical) similarity function Φ, we formulate matching as a bipartite assignment problem:

$$\hat{\mathcal{M}} = \underset{\mathcal{M} \subseteq \widetilde{\mathcal{K}}^A \times \widetilde{\mathcal{K}}^B}{\arg\max} \sum_{(\mathbf{k}_i^A, \mathbf{k}_j^B) \in \mathcal{M}} \Phi(\mathbf{d}_i^A, \mathbf{d}_j^B). \tag{2}$$

subject to each keypoint being matched at most once.

Here, $\Phi(\mathbf{d}_i^A, \mathbf{d}_j^B)$ represents the similarity score between descriptors $\mathbf{d}_i^A$ and $\mathbf{d}_j^B$. We employ SuperGlue for its context-based matching, although the above formulation also holds for other matchers (e.g. LightGlue). By consolidating both standard and semantically enriched descriptors into a single inference pass, our approach increases matching reliability while maintaining compatibility with standard visual-SLAM pipelines [26]. Further evaluation details are provided in Section V-B.

IV. Experimental Evaluation

Our KSI approach integrates semantic information into keypoint descriptors to mitigate perceptual aliasing in vineyards. Below, we present our experiments to demonstrate

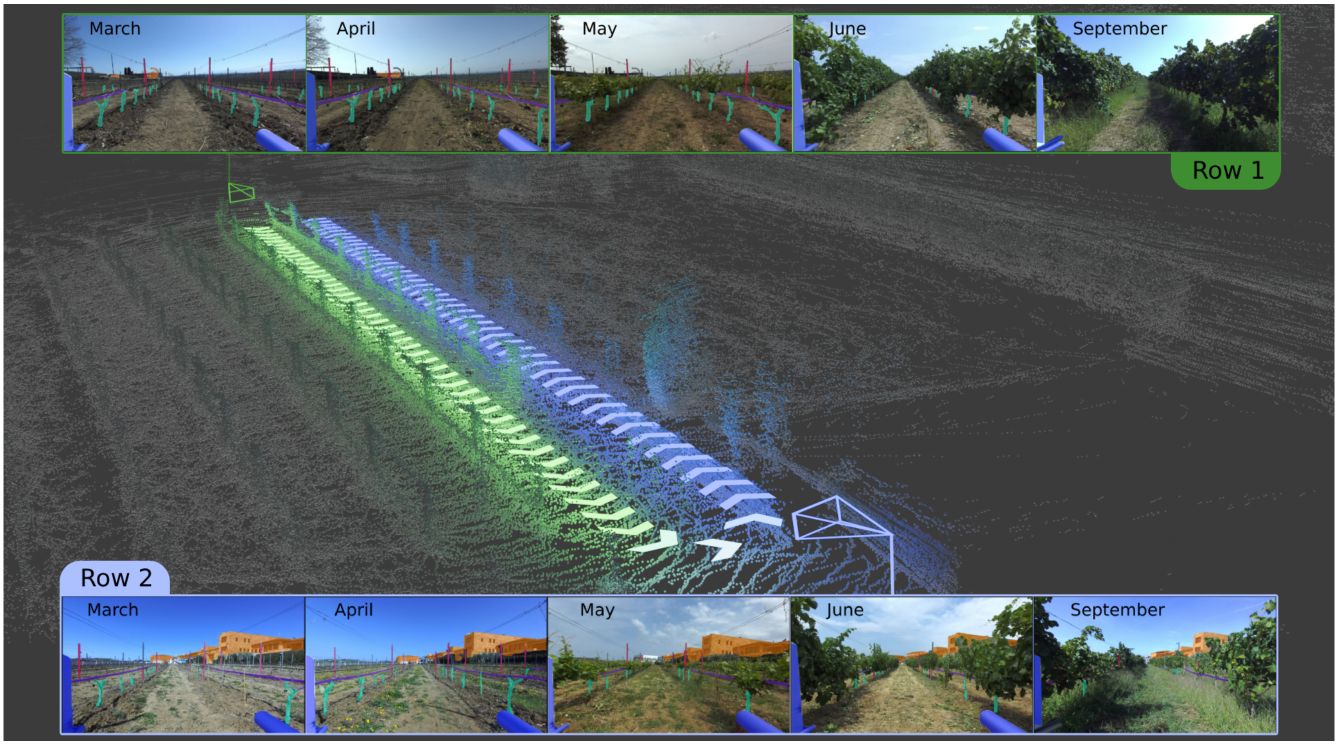

Fig. 3: Visualization of the vineyard loop path from which the data for experiments described in Section IV was colllected. The robot traverses row 1 in the forward direction, then returns along row 2, completing a closed-loop trajectory. Despite foliage occlusions, trunk and building classes remain visible across all months, rendering them ideal for enhancing keypoint descriptors. Buildings: Orange, Pipes: Purple, Poles: Red, Robot: Blue, Trunks: Cyan.

the capabilities of our method. The results support our key claims: (i) semantic enhancement of keypoint descriptors improves matching accuracy on image regions within masks derived from panoptic segmentation compared to their original form; (ii) leveraging the improved matching accuracy of semantically enriched descriptors leads to more reliable relative pose estimation during vineyard traversals; (iii) additional semantic context improves visual localisation accuracy in the same vineyard, outperforming traditional descriptors. These claims are backed up by our experimental evaluation.

A. Semantically Enhanced Keypoint Matching

The first experiment evaluates the performance of our approach, showing that semantically enhancing the keypoint descriptors improves matching accuracy in semantically significant vineyard image regions compared to the original descriptors. We deployed the KSI framework with SuperGlue, LightGlue and Nearst Neightbour (NN) to match features from SuperPoint, SIFT and SFD2 descriptors across image pairs sampled from sequences collected in March, April, May, June, and September within the same vineyard rows. We focus on keypoints in semantically significant regions (trunks and buildings) and report the percentage of correctly matched keypoints for each image pair. Because background keypoints lack semantic masks for descriptor enhancement, we do not include them in our analysis. As a baseline, we also performed feature matching without the KSI framework to quantify the semantic matching accuracy gains offered by our method.

To verify that a keypoint on a specific trunk or building in one frame was matched to the corresponding semantic instance in the next, we established semantic correspondences for each image pair. Using ground-truth rotation and translation, we projected the semantic mask from the first frame onto the second and aligned it spatially. The projected mask was then compared with all semantic instance masks in the subsequent frame using the intersection-over-union (IoU) metric. We applied a minimum IoU threshold of 0.1 to filter false positives, excluding smaller trunks beyond $\sim 4m$ and fragmented building masks caused by occlusion. Although this restricts evaluation to larger or closer semantic instances, it ensures fairness, as the same correspondences are applied to both our method and the baseline.

In Table I, we compare KSI with baselines that pair SuperPoint or SIFT descriptors and SuperGlue or LightGlue matchers. We additionally include SFD2, paired with its recommended nearest-neighbor matcher, as a baseline to enable comparison with a semantically enhanced feature descriptor approach. KSI outperforms the corresponding baselines in semantically significant regions in all five configurations, aside from the exceptions in March for SuperPoint + LightGlue and SFD2, where the difference is under 1%. The largest improvement appears in September, where SIFT + SuperGlue provides an 18.87% accuracy gain over its baseline. Overall accuracy tends to increase toward September, likely because heavier foliage reduces distinguishable background keypoints, making semantically enhanced keypoints more valuable. Notably, the highest average accuracy gain

TABLE I: Keypoint matching accuracy on semantic instances with and without KSI across different months.

Matcher	Descriptor	Avg. Accuracy Gain %	With KSI (%)					Without KSI (%)				
			March	April	May	June	September	March	April	May	June	September
Superglue	SuperPoint	06.70	**79.03**	**82.65**	**86.48**	**89.34**	**94.36**	75.96	76.54	80.35	80.56	84.95
	SIFT	14.58	**71.28**	**80.49**	**80.52**	**86.24**	**93.80**	60.39	65.30	68.03	70.80	74.93
LightGlue	SuperPoint	04.27	74.42	**80.43**	**84.28**	**87.28**	**93.38**	**75.80**	77.05	80.28	80.53	84.79
	SIFT	10.65	**70.87**	**74.97**	**81.07**	**86.06**	**88.39**	60.97	64.51	72.20	75.13	75.32
NN	SFD2	05.20	69.99	**77.24**	**80.96**	**87.88**	**95.35**	**70.10**	76.23	79.38	79.99	84.40

(14.58%) also arises with SIFT + SuperGlue, suggesting that classical descriptors can benefit even more from KSI than learned descriptors like SuperPoint. The results from SFD2 indicate that its semantically guided keypoint detection strategy performs less effectively than SuperPoint, likely due to its treatment of plants as short-term objects, which reduces the number of keypoints generated on plants. The highest absolute accuracy is achieved with by SuperPoint + SuperGlue, and thus we adopt this configuration for all subsequent experiments and ablations.

B. Relative Pose Estimation

The second experiment evaluates relative pose estimation with KSI-enabled keypoint matching, showing how our method improves pose estimation accuracy for image pairs during vineyard navigation. Accurate relative pose estimation is critical for real-world visual SLAM systems, particularly in agricultural robotics, where repetitive rows structures complicate localisation. By reducing pose errors, KSI-enabled keypoint matching improves trajectory estimation, thus enhancing navigation and task precision in structured environments.

We compute relative poses between image pairs by estimating the essential matrix via Random Sample Consensus (RANSAC) [27]. The ground truth poses are then used to set the translation scale in the estimated pose. Using these scaled relative pose estimates, we generate a trajectory for each month and compare it against the ground truth trajectory recorded by RTK-GPS. Each test sequence involves the robot traversing two vineyard rows in a loop – moving along one row, returning along another, and then repeating the forward pass – yielding a closed-loop path of approximately 153 m (illustrated in Figure 3). We employ the evo library [28] to calculate both the relative pose error (RPE) and the absolute pose error (APE) of the trajectory. RPE and APE results obtained from SuperGlue feature matching with KSI are compared against those produced without KSI.

The results in Table II show that KSI-enabled pose estimation consistently yields lower RPE and APE than the baseline. The largest performance gain appears in September, where the scarcity of distinguishable features makes pose retrieval especially challenging. This finding aligns with our motivation for semantic enrichment, demonstrating its effectiveness in visually limited scenarios.

C. Visual Localisation

Our third experiment demonstrates how KSI can accommodate cross-seasonal variations in vineyard environments, enabling robust visual localisation despite significant foliage growth. We uniformly sampled 136 images per row from two vineyard rows: row 1, located near the vineyard centre and containing mostly trunks, and row 2, closer to the edge, featuring both trunks and buildings (see Figure 3). Using the hierarchical localisation toolbox [29] for 6-degrees of freedom (DoF) visual localisation, we generate two point clouds from March images: one with SuperPoint+SuperGlue using KSI, and one without KSI. As shown in Table III, our approach delivers a 35.06% improvement in visual localisation accuracy for row 1, which offers fewer robust features. In contrast, accuracy in row 2 remained similar to the baseline. Hence, KSI significantly enhances cross-seasonal localisation in challenging vineyard scenes while maintaining similar performance under less demanding conditions.

We employed two metrics to assess visual localisation accuracy. First, we computed the Euclidean distance between the reference pose from March and the query pose for each image, then used the median translation error (MTE) over all images in each row. Second, we reported the recall, defined as the percentage of errors below thresholds of 0.5 m, 1 m, and 5 m, while discarding errors exceeding 1km as outliers. The results show that our KSI framework enhances accuracy under seasonal variations caused by foliage growth, especially in scenarios where repetitive structures limit robust feature identification. This is evident from the differing performance levels in row 1 and row 2, which reflects the varying difficulties of each environment.

D. Generalization to Non-Vineyard Environments

To evaluate the generalisation of our method beyond vineyard environments, we conducted experiments in a forestry setting. We collected a test sequence of a robot traversing approximately 50 m through a woodland and applied the PercepTreeV1 [30] detector for panoptic segmentation, using a single semantic class (trees) within the KSI pipeline. As shown in Fig. 4, the baseline keypoint matching accuracy on trees was 85.93% using SuperPoint descriptors with the SuperGlue matcher, while KSI improved this by 10.16%, reaching 94.66% on semantically significant regions. These results demonstrate that our method generalises effectively to non-vineyard outdoor environments such as forestry, which are typically challenging for traditional feature matching pipelines.

E. Runtime

To verify that our model can operate in real-time onboard the robot, we tested it on two devices: (1) a desktop with

TABLE II: Comparison of relative pose error (RPE) and absolute pose error (APE) across five months in a 159 m vineyard loop, using SuperPoint descriptors with SuperGlue matching. The KSI-enabled approach consistently reduces RPE and APE over the baseline, in some cases by more than half, with the largest improvement occurring in September.

Evaluation Method	Month : Mean (Standard Deviation) [Unit: cm]				
	March	April	May	June	September
Relative pose error with KSI	**17.32(11.49)**	**8.07(9.02)**	**5.85(5.71)**	**5.94(5.59)**	**3.28(11.20)**
Relative pose error without KSI	20.47(5.42)	12.53(8.39)	8.97(5.13)	9.41(5.52)	11.28(9.92)
Absolute pose error with KSI	**109.58(91.85)**	**117.23(76.21)**	**113.01(63.01)**	**150.95(78.05)**	**153.46(98.08)**
Absolute pose error without KSI	532.82(272.71)	398.24(290.34)	274.10(169.69)	358.52(151.41)	1013.72(804.68)
RPE Improvement, APE Improvement	15.39%, 79.43%	35.59%, 70.56%	34.78%, 58.77%	36.88%, 57.90%	70.92%, 84.86%

TABLE III: Visual localisation errors relative to March. The SuerPoint descriptors with SuperGlue matcher was used in this comparison. MTE: Median Translation Error in cm, R@X: Recall under X meters.

Metric	Row	Month (w/o KSI, with KSI)			
		April	May	June	September
MTE	1	23.75, **11.91**	21.09, **13.93**	24.47, **16.94**	513.34, **381.52**
	2	**06.35**, 06.53	07.90, **07.88**	06.78, **06.72**	13.49, **11.72**
R@0.5	1	94.85, **98.53**	**98.53, 98.53**	**63.24**, 61.03	26.47, **30.15**
	2	99.27, **100.00**	**100.00, 100.00**	**99.27**, 97.79	**91.18, 91.18**
R@1	1	**99.27**, 98.53	**98.53, 98.53**	**64.71**, 62.5	28.68, **30.88**
	2	**100.00, 100.00**	**100.00, 100.00**	**99.27**, 98.53	**94.85, 94.85**
R@5	1	99.27, **100,00**	99.27, **100.00**	**72.79**, 70.59	32.35, **35.29**
	2	**100.00, 100.00**	**100.00, 100.00**	**99.27, 99.27**	**96.32, 96.32**

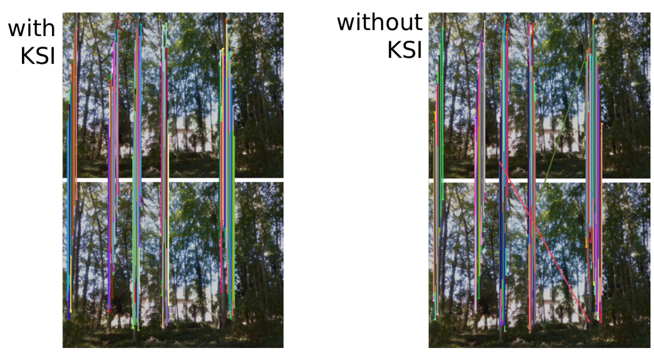

Fig. 4: Generalisation of KSI Pipeline Demonstrated in a Woodland.

an AMD Ryzen 7 7700 (16) @ 5.39 GHz CPU, NVIDIA GeForce RTX 4070 GPU and 32 GiB DDR5 5600 memory; (2) and a Jetson AGX Orin with an ARM Cortex-A78 2.2GHz CPU and NVIDIA Ampere 64GB (shared) GPU and memory.

Table IV presents the runtime results for both systems, including total time, runtime of the KSI framework and frame rates for the pipeline with SuperPoint and SuperGlue. Standard deviations are reported to indicate runtime variability across the evaluated months. The models are optimised using *TensorRT* to achieve faster inference runtime.

V. Ablation Study

We conducted two ablation studies to systematically evaluate and refine the KSI architecture, focusing on: (A) the impact of different semantic integration methods, and (B) the relevance of self and cross-attention during the matching stage. These studies were intended to optimise the KSI framework and to validate the design choices underpinning our proposed approach. The ablation study results are shown

TABLE IV: Runtime comparison for KSI on Jetson and a desktop computer per image pair. (ms: milliseconds, pps: pairs per second)

Component	Mean (SD)	
	Desktop	Jetson
KSI Framework (ms)	8.1 (2.3)	43.8.9 (11.5)
Total Time (ms)	67.8 (1.3)	212.3 (14.2)
Execution Rate (pps)	14.75 (0.28)	4.71 (0.31)

in the sub-tables of Table V, ordered concordantly with the text subsections.

A. Semantically Enhanced Descriptor Ablation

In this ablation study, we first examine whether adding semantic embeddings to descriptors yields better keypoint matching accuracy in semantically significant regions than simply concatenating them. We then assess the impact of normalisation applied both before and after the addition step, thereby validating and justifying our design choices for implementing normalisation within KSI.

1) Addition vs. Concatenation: Addition and concatenation are two common approaches to incorporate additional information into a feature descriptor. Each approach has its merits for semantic integration: concatenation provides an explicit structure for new semantic information, while addition preserves the original dimensionality and avoids the need for dimensionality reduction when used with a pre-trained matcher.

When designing feature descriptors for learned matches, one must respect dimensionality constraints, as many matchers are pre-trained on specific descriptor sizes. Ensuring compatibility without retraining is criticql for practical deployment. For example, in our proposed approach using SuperGlue – trained on 256×1 SuperPoint descriptors – we preserved this dimensionality in the concatenation method by compressing both the semantic masks and original descriptors into 128×1 embeddings using autoencoders. We then concatenated these compressed embeddings to form a 256×1 descriptor compatible with SuperGlue's pre-trained network, thereby preserving computational efficiency and avoiding additional matcher retraining.

In contrast, for the addition approach, we compressed the semantic masks into 256×1 embeddings and directly added them to the original 256×1 SuperPoint descriptors, followed by L2 normalisation of the resulting semantically enriched descriptor. As shown in Table V, addition outperforms concatenation within the KSI framework. In learned matchers

TABLE V: Ablation study on semantic-keypoint matching across months, evaluating integration strategies in KSI, normalisation configurations (SN: Semantic Normalisation, KN: Keypoint Normalisation), domain diversity (homogeneous vs. heterogeneous), and intra/inter-class matching distributions.

Configuration	**Month (%)**				
	March	April	May	June	September
Addition	**79.03**	**82.65**	**86.48**	**89.34**	**94.36**
Concat.	70.49	76.59	86.41	88.04	92.82
w/o KSI	75.96	76.54	80.35	80.56	84.95
SN:×, KN:×	78.93	82.65	86.38	89.29	94.22
SN:×, KN:✓	**79.03**	**82.65**	**86.48**	**89.34**	**94.36**
SN:✓, KN:×	77.36	82.35	85.32	88.51	93.00
SN:✓, KN:✓	77.36	82.35	85.19	88.41	93.02
Homogeneous	18.74	24.02	40.04	64.52	16.59
Heterogeneous	**79.03**	**82.65**	**86.48**	**89.34**	**94.36**
$\lvert\mathcal{K}'_{\mathcal{S}}{}^{A} \leftrightarrow \mathcal{K}'_{\mathcal{S}}{}^{B}\rvert$	14.30	13.47	9.52	3.84	3.07
$\lvert\mathcal{K}'_{\mathcal{S}}{}^{A} \leftrightarrow \mathcal{K}_{\mathcal{B}}{}^{B}\rvert$	0.14	0.21	0.11	0.04	0.04
$\lvert\mathcal{K}_{\mathcal{B}}{}^{A} \leftrightarrow \mathcal{K}_{\mathcal{B}}{}^{B}\rvert$	85.56	86.33	90.37	96.13	95.99
$\lvert\mathcal{K}_{\mathcal{B}}{}^{A} \leftrightarrow \mathcal{K}'_{\mathcal{S}}{}^{B}\rvert$	0.0	0.0	0.0	0.0	0.0

like SuperGlue and LightGlue, addition modifies the descriptor within its original space, aligning more naturally with the matchers' attention mechanisms. Concatenation, on the other hand, partitions the descriptor into separate visual and semantic components, potentially requiring the matcher to adjust to this altered structure. These distinctions can affect how effectively the matchers exploit the enhanced descriptors for feature matching.

2) Normalisation of Descriptors: In the KSI framework, we use an autoencoder to generate semantic embeddings, which are then added to the original descriptors. Both the autoencoder bottleneck output and the SuperPoint descriptor can optionally be L2-normalised: semantic embedding normalisation (SN) for the autoencoder bottleneck, and keypoint descriptor normalisation (KN) for the SuperPoint descriptor. This yields four possible normalisation configurations, all evaluated in this ablation study. To ensure consistency, we apply a mandatory L2 normalisation after the addition step so that the final descriptor sent to the matcher is always L2-normalised, as in Equation 1. Results in second sub-table of Table V show that the best performance is achieved when semantic embeddings are not normalised, while the SuperPoint descriptor is normalised. In this configuration, unnormalised semantic embeddings retain their magnitude, effectively shifting the SuperPoint descriptor in the descriptor space, while the final L2 normalisation guarantees the matcher always receives a properly normalised descriptor.

3) Keypoint Sampling: In feature matching pipelines, capping the maximum number of detected keypoints can reduce inference time. In SuperPoint's default configuration, this cap is 4096; however, all of our experiments ran without it, yielding 8000–9000 keypoints per image. We also evaluated configurations halving the cap from 4096 to 512, observing less than a 2% variation in keypoint matching accuracy within semantically significant regions. This stability is expected, as ambiguous keypoints typically occur in background areas (e.g., ground, foliage), whereas keypoints within semantic regions are comparatively robust.

B. Ablation on Descriptor Matching

1) Heterogeneous vs. Homogeneous Feature Matching: We compare keypoint matching accuracy between *heterogeneous* and *homogeneous* feature matching approaches. In the heterogeneous approach, the full keypoint set $[\mathcal{K}_{\mathcal{B}}, \mathcal{K}'_{\mathcal{S}}]$ containing both original and semantically enhanced descriptors is processed together in a single SuperGlue instance. In contrast, the homogeneous approach processes $\mathcal{K}_{\mathcal{B}}$ and $\mathcal{K}'_{\mathcal{S}}$ separately in two SuperGlue instances and then merges the resulting matches.

Although SuperGlue was not originally designed for heterogeneous matching (i.e., mixing different descriptor types), we adopt this approach in the KSI framework (Section III-E) and evaluate its performance. As shown in Table V, heterogeneous matching achieves a 53.59% higher keypoint matching accuracy in semantically significant regions compared to the homogeneous setup. This improvement stems from SuperGlue's graph neural network (GNN) context aggregation, which benefits from seeing all keypoints at once. By matching the full keypoint set in a single pass rather than splitting it, SuperGlue has access to a richer spatial context, leading to more accurate matches.

2) Cross Domain Matching: The last sub-table in Table V reports the matching accuracy for semantically significant keypoints from $[\mathcal{K}_{\mathcal{B}}{}^{A}, \mathcal{K}'_{\mathcal{S}}{}^{A}]$ in the first image A correspond to $[\mathcal{K}_{\mathcal{B}}{}^{B}, \mathcal{K}'_{\mathcal{S}}{}^{B}]$ in the second image B. The $|\mathcal{X} \leftrightarrow \mathcal{Y}|$ notation denotes the percentage of matches between set $\mathcal{X}$ in the image A and set $\mathcal{Y}$ in the image B. The matching accuracies across semantic and background domains are presented in Table V.

These results further indicates that our heterogeneous matching approach consistently assigns keypoints to their semantic relevance, even though some keypoints are semantically enriched with KSI. Only a small fraction of keypoints in semantically significant regions of the image are matched to background regions ($|\mathcal{K}'_{\mathcal{S}}{}^{A} \leftrightarrow \mathcal{K}_{\mathcal{B}}{}^{B}|$), while none of the background keypoints match semantically significant regions ($|\mathcal{K}_{\mathcal{B}}{}^{A} \leftrightarrow \mathcal{K}'_{\mathcal{S}}{}^{B}|$). Table V also shows how the number of keypoints in semantically significant areas decreases during months with dense foliage. As more foliage covers of the scene, fewer keypoints remain visible in these regions, thus reducing the $|\mathcal{K}'_{\mathcal{S}}{}^{A} \leftrightarrow \mathcal{K}'_{\mathcal{S}}{}^{B}|$ matches for those months.

These ablation studies validate our design choices around semantic integration, normalisation, sampling, and matching strategy. These findings confirm that incorporating semantic embeddings, applying selective normalisation, and adopting a heterogeneous matching approach significantly enhance keypoint matching accuracy.

VI. Conclusion

We introduced KSI, a novel method for embedding semantic information into keypoint descriptors to address perceptual aliasing and appearance changes in outdoor environments. By enhancing descriptors with semantic context, KSI increases robustness in environments such as vineyards where traditional descriptors often fail due to repetitive structures. Our approach encodes semantic instance masks derived from vineyard-specific objects, producing semantic embeddings that are added to the original descriptors. In addition, we presented the SemanticBLT dataset, designed for panoptic segmentation in vineyard environments. Evaluating our approach across multiple seasons with significant visual and temporal variations, we tested it with various keypoint descriptors and matches. In all descriptor-matcher combinations, our KSI framework achieved an overall 8.3% higher matching accuracy in semantically significant regions compared to the respective baselines. Moreover, semantic descriptor enrichment improved both relative pose estimation and visual localisation performance over multiple months, underscoring KSI's effectiveness in challenging agricultural environments.

REFERENCES

[1] A. Papadimitriou, I. Kleitsiotis, I. Kostavelis, I. Mariolis, D. Giakoumis, S. Likothanassis, and D. Tzovaras, "Loop closure detection and slam in vineyards with deep semantic cues," in *Proc. of the IEEE Intl. Conf. on Robotics & Automation (ICRA)*, 2022.

[2] 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 (CVPR)*, June 2020.

[3] 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 (RA-L)*, vol. 3, no. 4, pp. 3097–3104, 2018.

[4] K. Wang, L. Kooistra, Y. Wang, S. Vélez, W. Wang, and J. Valente, "Benchmarking of monocular camera uav-based localization and mapping methods in vineyards," *Computers and Electronics in Agriculture*, vol. 227, p. 109661, 2024.

[5] J. Zhai, S. Zhang, J. Chen, and Q. He, "Autoencoder and its various variants," in *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2018.

[6] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, 1999.

[7] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," in *Proc. of the IEEE/CVF Intl. Conf. on Computer Vision (ICCV)*, 2011.

[8] D. DeTone, T. Malisiewicz, and A. Rabinovich, "SuperPoint: Self-Supervised Interest Point Detection and Description," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] J. Revaud, P. Weinzaepfel, C. R. de Souza, and M. Humenberger, "R2D2: repeatable and reliable detector and descriptor," in *Proc. of the Conf. on Neural Information Processing Systems (NeurIPS)*, 2019.

[10] G. S. Gama, N. Dos Santos Rosa, and V. Grassi, "Semantic superpoint: A deep semantic descriptor," in *Proc. of. the Latin American Robotics Symposium (LARS)*, 2022.

[11] T. Guadagnino, X. Chen, M. Sodano, J. Behley, G. Grisetti, and C. Stachniss, "Fast Sparse LiDAR Odometry Using Self-Supervised Feature Selection on Intensity Images," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 3, pp. 7597–7604, 2022.

[12] L. Lobefaro, M. Malladi, O. Vysotska, T. Guadagnino, and C. Stachniss, "Estimating 4D Data Associations Towards Spatial-Temporal Mapping of Growing Plants for Agricultural Robots," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2023.

[13] I. Hroob, R. Polvara, S. Molina, G. Cielniak, and M. Hanheide, "Benchmark of visual and 3d lidar slam systems in simulation environment for vineyards," in *Proc. of the Towards Autonomous Robotic Systems (TAROS) Conference*, 2021.

[14] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[15] M. M. Tanco, G. Tejera, and J. M. Di Martino, "Learning agriculture keypoint descriptors with triplet loss for visual slam," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–10, 2023.

[16] H. Ding, B. Zhang, J. Zhou, Y. Yan, G. Tian, and B. Gu, "Recent developments and applications of simultaneous localization and mapping in agriculture," *Journal of Field Robotics*, vol. 39, no. 6, pp. 956–983, 2022.

[17] S. Kreiss, L. Bertoni, and A. Alahi, "Openpifpaf: Composite fields for semantic keypoint detection and spatio-temporal association," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 8, pp. 13498–13511, 2021.

[18] J. Cao, Y. You, C. Li, and J. Liu, "Tsk: A trustworthy semantic keypoint detector for remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[19] Z. Luo, W. Xue, J. Chae, and G. Fu, "Skp: Semantic 3d keypoint detection for category-level robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5437–5444, 2022.

[20] F. Xue, I. Budvytis, and R. Cipolla, "Sfd2: Semantic-guided feature detection and description," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[21] F. Xue, I. Budvytis, and R. Cipolla, "Sfd2: Semantic-guided feature detection and description," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5206–5216, 2023.

[22] F. Crocetti, E. Bellocchio, A. Dionigi, S. Felicioni, G. Costante, M. L. Fravolini, and P. Valigi, "Ard-vo: Agricultural robot data set of vineyards and olive groves," *Journal of Field Robotics*, vol. 40, no. 6, pp. 1678–1696, 2023.

[23] R. Polvara, S. Molina, I. Hroob, A. Papadimitriou, K. Tsiolis, D. Giakoumis, S. Likothanassis, D. Tzovaras, G. Cielniak, and M. Hanheide, "Bacchus long-term (blt) data set: Acquisition of the agricultural multimodal blt data set with automated robot deployment," *Journal of Field Robotics*, vol. 41, no. 7, pp. 2280–2298, 2024.

[24] F. Schmidt, C. Blessing, M. Enzweiler, and A. Valada, "ROVER: A Multi-Season Dataset for Visual SLAM," *arXiv preprint arXiv:2412.02506*, 2024.

[25] C.-Y. Wang, I.-H. Yeh, and H.-Y. Mark Liao, "Yolov9: Learning what you want to learn using programmable gradient information," in *Proc. of the Europ. Conf. on Computer Vision (ECCV)*, 2025.

[26] J. Cheng, L. Zhang, Q. Chen, X. Hu, and J. Cai, "A review of visual slam methods for autonomous driving vehicles," *Engineering Applications of Artificial Intelligence*, vol. 114, p. 104992, 2022.

[27] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, p. 381–395, 1981.

[28] M. Grupp, "evo: Python package for the evaluation of odometry and slam.." `https://github.com/MichaelGrupp/evo`, 2017.

[29] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk, "From coarse to fine: Robust hierarchical localization at large scale," in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[30] V. Grondin, J.-M. Fortin, F. Pomerleau, and P. Giguère, "Tree detection and diameter estimation based on deep learning," *Forestry: An International Journal of Forest Research*, 10 2022.