Note on Methodology

DETECTING BEES IN CHERRY FLOWERS USING TIMELAPSE IMAGES AND FOUNDATIONAL MODELS

Jane Devlin[1], Fraser MacFarlane[2], Alison Karley[2], Fabio Manfredini[1,3], Dominic Williams[2,*]

[1]*School of Biological Sciences, University of Aberdeen, Zoology Building, Tillydrone Avenue, Aberdeen, AB24 2TZ, UK*
[2]*James Hutton Institute, Errol Road, Dundee, DD2 5DA, UK*
[3]*Department of Food, Environmental and Nutritional Sciences, University of Milan, Via Celoria 2, 20133, Milano, Italy*

Journal of Pollination Ecology, 41(4), 2026, pp 28--39
DOI: 10.26786/1920-7603(2026)897

Received 20 August 2025, accepted 5 February 2026

*Corresponding author: dominic.williams@hutton.ac.uk

Abstract—Bees perform important pollination services in fruit crops such as sweet cherry (*Prunus avium*). Growers will often introduce managed bees to supplement natural pollinators. Monitoring pollinating insects is important to understand the impact augmenting pollinators has on fruit yield, particularly in relation to June drop which is a major cause of yield instability in the cherry industry. Timelapse cameras allow for continuous monitoring of flowers, but manual analysis of the generated footage is very time consuming. Timelapse imaging combined with automated image processing methods, is a valuable tool in studying the role bee pollination plays in fruit production.
We have developed a novel method for detecting bees in time lapse images, called BeeSAM2. This exploits both the zero-shot detector *Grounding DINO* and the foundational model *Segment Anything 2*. Promising results are achieved with the method being capable of detecting the bumblebee *Bombus terrestris* in images with a recall of 0.959 and precision of 0.991. These results are sufficiently accurate to deploy our method to quantify bee activity in cherry plantations, advancing the ability of researchers to monitor flower-pollinator interaction, and saving a significant amount of time during video processing.

Keywords—Bee detection, *Prunus Avium*, Computer vision, Image analysis, Pollinator detection, Timelapse photography

INTRODUCTION

Pollinators are very important for crop production worldwide (Klein et al. 2007; Gill et al. 2016; Potts et al. 2016). In commercially grown sweet cherry (*Prunus avium)* plantations, self-sterile cherry varieties are often grown, meaning that pollen from another compatible variety carried by bees or other insects is needed for successful pollination. Cherry trees suffer from a phenomenon known as "June Drop" where immature fruit is dropped by a tree, this is a major cause of yield instability in the cherry industry. The drivers of this process are not fully understood but it is thought that it may be related to pollination (Gatti 2024) (Mir et al. 2025). In particular the quality of pollination services such as visitation rate, duration, stigma contact and flower handling may all affect both initial fruit set and subsequent fruit drop. This has motivated the study of pollination in cherry trees. It is standard commercial practice amongst UK cherry growers to supplement naturally occurring pollinators with hives of buff-tailed bumblebee (*Bombus terrestris audax),* honeybees (*Apis melifera)* are also commonly used, thus we have focussed our work on pollination services provided by bees. While our work has focussed on the pollination of cherry the challenges associated with efficient pollinator monitoring is not limited to cherry production and is one easily applicable to a broad spectrum of scenarios and crops (Breeze et al. 2021).

There are a variety of methods that can be used to monitor pollination: these include trapping pollinators, walking transects, eDNA, and the use of cameras (Howard et al. 2021; Johnson et al. 2023; Van Klink et al. 2024). The use of cameras to monitor pollinator behaviour is an increasing area of interest (Steen 2017; Ratnayake et al. 2021); advantages of using cameras include allowing

precise monitoring of flowers without extensive periods of field work being required and lower pollinator disturbance by a human observer. Researchers have used a combination of timelapse photography and computer vision methods to analyse these types of data (Ngo et al. 2021; Bjerge et al. 2023a; Bjerge et al. 2023b; Ștefan et al. 2025). The use of motion sensor cameras to study pollinators encounters two major challenges; Firstly, poor detection due to the small size of insect pollinators, and secondly, false triggers from wind mediated flower movement. Consequently, continuous recording and post capture analysis of footage is used instead (Høye et al. 2025).

Several studies have used timelapse photography combined with automated detection methods to monitor insects. Ngo et al. 2021 used a camera system to monitor honeybees entering and leaving a beehive. They used a simple USB camera connected to a Raspberry Pi which continuously recorded bees at the entrance of the hive. Bees were classified according to whether they were carrying pollen or not using a model based on YOLOv3 (Redmon & Farhadi 2018), characterized by tiny architecture and achieving high accuracy (0.91 precision, 0.99 recall) for detection of bees with pollen loads. Ștefan et al. 2024 used smart phones mounted on tripods to monitor pollinator interactions with flowers. They recorded timelapse footage at 1 second intervals for one hour. No automation was used for image analysis and over 1,720 hours were spent labelling the 460,000 gathered images indicating the high time cost associated with analysing data if computer vision techniques are not used. Subsequently they trained YOLO models on their dataset with mixed results when attempting to classify species (Ștefan et al. 2025). Bjerge et al. 2023b used timelapse images of flowers to monitor insect visitation primarily by honeybees. They set up recording units containing USB web cameras connected to a Raspberry Pi recording images at 30 second intervals. Images were captured between the hours of 4:30 and 22:30. They used a combination of motion enhancement and the YOLOv5 (Jocher et al. 2020) object detection method to automatically detect insects in the images. The model was trained on 3,783 images of which 2,499 contained insects. The same group used a similar system without the addition of motion enhancement, aiming instead to classify the different insect pollinators detected using only YOLOv5 (Bjerge et al. 2023a).

One of the challenges in using object detection methods is the requirement for specific training and a frequent lack of generalisability with the models produced. This has prompted interest in producing "foundational" models which are able to generalise to new problems without the need for finetuning. Segment anything model (SAM) (Kirillov et al. 2023) aims to segment any objects in an image going from either point or bounding boxes to indicate the object. By prompting systematically across an image, it can detect all objects in said image. It has been used extensively in various imaging problems including medical image analysis (Zhu et al. 2024; Sengupta et al. 2025), crop image analysis for the detection of leaves (Williams et al. 2024), detecting the size of mason bee (*Osmia*) cocoons in images (Getz et al. 2024) and to help segment images of moths (Jain et al. 2024). Segment anything model 2 (SAM2) (Ravi et al. 2024) is an advancement on SAM which aims to track objects through video frames. An initial prompt is required which is used to select an object of interest and then this object is followed through all the frames in a video. It achieves state of the art results on several standard video segmentation benchmarks.

Alongside foundational models, there has also been research into zero-shot object detection methods, i.e. object detection that does not require training. Grounding DINO (Liu et al. 2024) is an example of this method. A text input is given, and matching objects in an image are detected. This method has been used extensively to solve a variety of problems. Mullins et al. 2024 tested different text prompts using both Grounding DINO and YOLO-World for a series of tasks on blueberry plants. They found Grounding DINO outperformed YOLO-World for most tasks and required fewer descriptive prompts, finding blueberries with the prompt "smooth blueberries" rather than "a small blue sphere". Grounding DINO has been used to help crop images of parasitoid wasps prior to taxonomic identification (Shirali et al. 2024). Grounding DINO and SAM have been combined by (Ren et al. 2024): here grounding dino was used as a zero-shot object detector and SAM was used to generate a segmentation of detected objects.

So far limited work has been done to incorporate zero-shot detectors or foundational image models to the task of insect pollinator detection. In this paper we present a novel method that combines Grounding DINO and SAM2 to detect bumblebee pollination events in timelapse images of cherry flowers: this method has the potential to be applied to other bee species or pollinating insects in different scenarios.

MATERIALS AND METHODS

EXPERIMENTAL SETUP AND PLANT STUDY SUBJECT

The experiment was carried out in two polytunnels with two rows of sweet cherry, *Prunus Avium,* trees in each. The polytunnels were located at the James Hutton Institute's Invergowrie farm site (56.456N, 3.066W). The trees were arranged in blocks of the same variety. Image data was gathered from the variety Kordia which was present in both tunnels. Kordia is a self-sterile variety requiring cross-pollination with a different compatible variety to develop fruit. In tunnel 1, three varieties, Kordia, Sweetheart and Regina, were grown in 5 plant plots, with a total of 45 trees in each row. Sweetheart and Regina are both compatible with Kordia for pollination. The second tunnel has four varieties, Kordia, Penny, Lapins, and Sweetheart grown in 4 plant plots. Penny, Lapins and Sweetheart are all compatible pollen donors for Kordia. At the time of the experiment the trees were 6 years old. Three commercial colonies of the bumblebee *Bombus terrestris audax* were added to the tunnels to provide pollinators for the cherry flowers mimicking standard commercial practice by cherry growers in Scotland. Insect netting was used to create three sections only accessible to commercial colonies (one colony per section) and three sections that were open to wild pollinators to enter: each section contained 15 trees, 5 each of Kordia, Sweetheart and Regina. Twelve Afidus timelapse ATL-200S cameras with a resolution of 1920x1080 (two per section) were mounted on wooden posts and pointed towards a single flower cluster on the target tree. The trees were randomly selected but each flower cluster was chosen to the fit criteria of being approximately 1.5m from the ground, at approximately the same stage of development, and be in a recordable position. They were set to take timelapse photos at a rate of 1 per second between 11am and 5pm, corresponding to apparent peak foraging times for commercial bumblebees in this environment: this was done to reduce the recording time with the aim of extending the battery life of cameras. The assessment of peak foraging times was based on authors preliminary observations during the trial. The data used in this study was gathered between 25/04/2023 and 28/04/2023. The short recording period was due to the fleeting viability window of cherry flowers.

MANUAL ANALYSIS OF TIMELAPSE IMAGES

The cameras stored the timelapse images as video files. These were then watched by an observer and any visits of bees to the target flower cluster was recorded with information on time of arrival and departure, duration, bee species, and flower identity. The purpose of this annotation was to record pollination events, so only bees that landed on the target flower clusters were recorded; bees which either failed to land on flowers or landed on flowers outside of target cluster were not recorded. This was a very time-consuming process – approximately 500 hours which motivated the development of an automated method to complete it more quickly in future applications.

BJERGE ET AL. 2023B DATASET

The method was also tested on an existing dataset published by Bjerge et al. 2023b. This data consisted of images of six different flower species gathered in a glasshouse in Flakkebjerg, Denmark of six different flowers species. Small hives of western honeybees (*Apis mellifera*) were added and images of honeybees on the target flowers were gathered at 30s intervals.

AUTOMATED BEE DETECTION METHODS

A number of methods were developed aiming to automatically detect bee visits from the timelapse footage with minimal human intervention. For all methods, the individual frames were extracted from the timelapse footage. Grounding DINO was used, setting a prompt as "bee" for the target object with a confidence threshold of 0.3 used for any object found. Results of testing of confidence intervals between 0.3 and 0.8 are presented in the results. A 0.3 threshold was found to be optimum, as the precision of Grounding DINO was found to be higher than recall. Using a lower threshold has the effect of

increasing the recall while reducing precision. We refer to this method in the results and discussion sections as GD_bee.

Further testing of additional prompts was carried out to determine the best prompt to use to detect bees in the images. Use of closely related prompts such as "wasp", "insect", "pollinator" produced similar results but slightly more missing observations. These alternative prompts were only tested via visual inspection on a small subset of the data. One of the challenges faced while detecting bees during a pollination event was occlusion caused by the bees landing on the flowers and being partially hidden within the flowers. To overcome this problem the prompt of "pollinating bees" was used to provide the model with additional context to improve results. Initial testing was also carried out for prompts of "bees in flowers" but any inclusion of flower resulted in detection of flowers without bees, so this was not continued. In the results and discussion sections this method is referred to as GD_pollinating_bee.

Segment Anything Model 2

Using segment anything 2 (SAM2) requires definition of an object to track on the first frame of the video. In our case most of the first frames did not contain bees. Therefore, a synthetic first frame was added to the beginning of the videos. This was created by taking the first frame of the video and then adding a bee taken from one of the subsequent images known to contain a bee. This bee was placed in the image 4 times (i.e. in the same image on 4 separate occasions) to enable tracking of multiple bees. An example of what this looks like is shown in Fig. 1. The bee was flipped for two of the instances. SAM2 was then prompted with 4 different objects one for each bee. Once the initial object was selected, SAM2 attempted to track the object through all the other frames in the video, and it maintained the ability to detect objects even after they had not appeared in the video for 100s of frames. While the purpose of the method was to track unique instances of an object (Ravi et al. 2024), we aimed to detect the presence/absence of bees rather than counting individual bees. In the results and discussion this method is called BeeSAM2.

In order to improve the results of this method a combination of Grounding DINO and SAM2 was used. The first step was to run Grounding DINO through all frames in the video. The prompt of "bee" was used, and we tested a confidence threshold of 0.3 and 0.6. The higher confidence threshold was used to minimise the number of false positives. In all frames with bees found by Grounding DINO, the bounding box produced by Grounding Dino was used to mark objects in the image for SAM2 to track. To assist SAM2 in videos where very few bees were detected by Grounding Dino, all videos were still supplemented with a synthetic image as the first frame. After labelling the bees in the images, the labels were propagated through the rest of the video using SAM2. In the results and discussion this method is known as GD_BeeSAM2.

EVALUATION OF RESULTS – OUR DATA

To quantify the success of the different methods we used the manually analysed timelapse footage data. This was footage where the time and duration of bees landing on target flowers had been recorded by an observer. The exact location of the bee in the image was not recorded by the manual observer, so it was assumed that any bees detected in the same frame where the observer recorded a bee on a target flower was a successful bee detection. The method developed here aimed to detect all bees, not just those on the target flowers, and so we produced an additional category of false positives: correct detection of bees that were not on target flowers. To quantify this component, an additional annotation was carried out for all the frames with a false positive bee detection recording, to discern whether they were truly a false positive or were a correct detection of a bee. Frames where no bees were detected were not checked again, so we did not quantify the proportion of false negatives for bees away from target flowers (as these were not recorded manually). The adjusted precision metric used is precision after this additional annotation of false positives was carried out.

We used the following measures to compare the different methods: recall, precision and adjusted precision. Recall is defined by the formula:

$$\text{Recall} = TP/(TP+FN)$$

where TP is true positives – i.e., bees on target flowers accurately detected - and FN is false

Figure 1: Example synthetic image with four bees added to it. The image was obtained by imposing the same picture of a bee in 4 different locations and 2 orientations on a background image representing a cluster of cherry flowers. The red box indicates the flowers of interest where manual observations were made of bees visiting flowers.

negatives - or number of bees on target flowers missed by model.

Precision is defined by the formula:

$$\text{Precision} = TP/(TP+FP)$$

where TP is bees on target flowers accurately detected and FP is false positives – i.e., number of bees in frames where there are no bees on target flowers (so including accurate detection of bees off target flowers).

Finally, Adjusted Precision is defined by the formula:

$$\text{Adjusted Precision} = (TP + FFP)/(TP + FFP + TFP)$$

where TP is bees on target flowers accurately detected, FFP is bees detected on non-target flowers and TFP is frames where detected bees do not correspond to a bee i.e. another object was detected that was not a bee.

Evaluation of results – data from Bjerge et al. 2023b

To allow for a comparison to be made with previous methods we also evaluated our method using the dataset from Bjerge et al. 2023b, which consists of timelapse images with labelled bee objects. We again evaluated performance on a frame level assuming that any bee detected in the frame was the same as the labelled bee in the same frame. Our method includes the creation of a synthetic image (Hinterstoisser et al. 2018) for the first frame with four bees added to an image as described above. Two different methods were used: the first involved using the same bee as in our study and adding this 4 times to the first frame of new footage; the second consisted of finding a new bee from the Bjerge study using Grounding DINO and adding it to the first frame of each video (again 4 times).

Results

Grounding DINO confidence testing results used the prompt "bee" are shown in Table 1. As expected, recall decreases with an increasing confidence threshold and precision increases. The adjusted precision is a lot higher than the precision indicating many of the bees detected by Grounding DINO had not been initially marker by manual observer. This probably due to the presence of bees away from the target flowers. Due to a high adjusted precision score of 0.972, a threshold of 0.3 this value was picked as the optimum value to maximise recall.

Table 1: Recall, precision and adjusted precision or GD_bee method testing different confidence thresholds.

Confidence threshold	Recall	Precision	Adjusted Precision
0.3	0.580	0.289	0.972
0.4	0.497	0.331	0.997
0.5	0.399	0.358	0.998
0.6	0.285	0.397	1
0.7	0.102	0.438	1
0.8	0.008	0.5	1

Table 2: Recall and precision of different prompt testing for Grounding Dino. The prompt Bee performs best having reasonable precision and recall over the different thresholds tested.

	Bee	Pollinating Bee	Pollinator	Bee in flower	Wasp	Insect
Recall 0.3	0.58	0.796	0.967	**0.984**	0.936	0.975
Precision 0.3	**0.289**	0.063	0.013	0.021	0.248	0.016
Recall 0.4	0.497	0.624	0	0.428	**0.802**	0.254
Precision 0.4	**0.331**	0.154	0	0.022	0.305	0.304
Recall 0.5	0.399	0.421	0	0.221	**0.662**	0.182
Precision 0.5	**0.358**	0.232	0	0.189	0.323	0.342
Recall 0.6	0.285	0.421	0	0.097	**0.503**	0.093
Precision 0.6	**0.397**	0.232	0	0.264	0.336	0.355

The results of testing different prompts for Grounding DINO are shown in Table 2. While the use of other prompts can improve the recall at a low confidence threshold this comes with a heavy hit to precision. A full analysis of false negatives was not carried out for the alternate prompts, but manual inspection of a few images indicated that low precision was mostly due to detection of inaccurate objects rather than presence of none target insects in footage. The second-best performing prompt was wasp, which had a lower precision but consistently higher recall, indicating Grounding DINO struggles to distinguish wasps and bees from each other.

The methods that we tested showed different measures of recall and precision (Table 3). Precision was low for all methods, with a max of 0.29 for GD_bee. This is due to the way we calculated precision and indicates the detection of many bees not located on the target flowers. We will instead focus on the adjusted precision scores as a more accurate reflection of the precision of the methods.

Table 3: Recall, precision and adjusted precision of the four versions of methods presented here. GD_BeeSAM2 has highest recall and adjusted precision.

	Recall	Precision	Adjusted Precision
GD_bee	0.580	0.289	0.972
GD_pollinating_bee	0.796	0.063	0.309
BeeSAM2	0.937	0.130	0.907
GD_BeeSAM2 (0.3 threshold)	0.968	0.128	0.938
GD_BeeSAM2 (0.6 threshold)	0.959	0.129	0.991

GD_bee had the lowest recall of 0.58 but had a high adjusted precision of 0.97. This showed that Grounding DINO 'out of the box' can be used to detect bees with reasonable accuracy. However, with a recall of only 0.58 it could not be confidently used to determine if a flower had been visited by a bee.

Changing the prompt to 'pollinating bee' resulted in an increase in the recall score to 0.80:

however, this resulted in a significant reduction of the adjusted precision score to 0.31, indicating that most of the objects found using this method were not bees, limiting its usability as a method.

BeeSAM2 had better results with a recall of 0.94 and an adjusted precision of 0.91. This showed good potential for this method. However, once it started to track an incorrect object it would generally continue to track it unless there was further intervention. A large part of the reason for lower precision when compared to GD_bee was the tracking of a leaf cluster instead of bees in one of the timelapse videos.

GD_BeeSAM2 with a threshold of 0.6 was the best performing method in terms of both recall (0.96) and adjusted precision (0.99). The recall was slightly higher when a threshold of 0.3 was used (0.97) but the adjusted precision score was reduced to 0.94 leading us to conclude a higher confidence threshold on Grounding DINO produces the best results. These scores imply that this method could reliably be used to monitor bee flower visitation with good confidence that not many bees would be missed. It is worth noting that in some of the frames where a bee was known to be on the flower, the bee was very heavily occluded and the presence of the bee was only known to the manual observer by watching the sequence of frames and seeing the flower position shift by the weight of the foraging bee.

Figure 2 shows the location of all detected bees in the image. It demonstrates the reason for the use of the adjusted precision score, many of the false positives are clustered around off target flowers indicating that they were likely to be correct identification of bees.

Assessment of our method with data from Bjerge et al. 2023b

The application of the GD_BeeSAM2 method to data from Bjerge et al. 2023b produced worse results than those observed for our dataset overall but still achieved recall and precision greater than 0.8. A slight improvement in performance, from 0.812 to 0.832 for precision and from 0.804 to 0.831 for recall, was seen when using bees from the Bjerge et al. 2023b dataset to generate the initial synthetic image (Table 4). This shows that the

Figure 2: Locations of bee detections marked on an example image. Orange dots represent true positives and blue dots represent false positives. The target flower cluster is at the centre of the image, bee visits on flowers in the background were not recorded. We can see here that many of the blue points are clustered around off-target flowers indicating that they may correspond to actual bee detections. The orange dots away from target flowers are due to there being multiple bees in the same frame, corresponding to bees that were detected off target flower in the same frame as a bee was present on target flower cluster.

Table 4: Recall and precision of Grounded Bee Sam2 on data from Bjerge et al. 2023b. Note here there is no adjusted precision as they annotated all bees present in their video file.

	Recall	Precision
GD_BeeSAM2	0.804	0.812
GD_BeeSAM2_alternativebee	0.831	0.832

creation of a new synthetic image from the relevant data improves the performance of our method, this may be important since the bees in Bjerge were primarily honeybees rather than the bumblebees present in our images. Our method may have performed worse on the Bjerge dataset as their data was only captured at 30 second intervals rather than the 1 second intervals we used, this removes the advantage of SAM2 being able to track bees between frames. These results are slightly worse but comparable to performance seen in the original paper where the authors achieved 0.888 recall and 0.897 precision on the same data set. The fact that our method closely matches the performance of a model trained specifically on the dataset highlights the potential of our model to generalise well and to be used on new unseen datasets.

Failure case examples

To understand the limitations of the methods proposed, it is useful to show some of the failure cases. One of the weaknesses of SAM2 is that it has no knowledge of the nature of the object it is tracking beyond the initial identification. Images a and b in Fig. 3 are an example of where a small bee on a flower was initially correctly identified, but in the following frames the tracking was transferred from the bee to the flower it was on. The method was initially designed to be interactive so this could be corrected with negative points, but of course this would not be suitable for fully automated detection of bees. Finetuning of the model may be able to overcome this issue but would reduce the generalisability of the method described here.

Another failure case is shown in panel c in Fig. 3 (produced by Grounding DINO). A few additional objects in the image were occasionally identified as bees - for example a cluster of leaves as shown in the picture. Once the prompt from Grounding DINO was given to SAM2 it would track that object indefinitely. In this case a single incorrect prompt from Grounding DINO was replicated for the rest of the frames of that video by SAM2. Increasing the confidence threshold used by Grounding DINO can help reduce the likelihood of this happening.

Looking at panels d, e and f in Fig. 3 we can see an example of a false negative. First, we can see a bee detected flying towards the flower (panel d); in the next frame (panel e) no bee is detected, as the bee has entered one of the flowers; finally, several frames later (panel f) the bee is detected again leaving the flower. If we manually observe the sequence, we can see that the bee is in the flower, so this sequence of frames was marked during manual classification as having a bee present; however, none of our models were able to detect this. Looking at several false negatives this appeared to be a consistent source of errors. This shows that while the model can detect some heavily occluded bees, it has some limitations compared to manual observers who can infer bee locations from a sequence in a way that our model cannot. This may cause problems for users trying to count how many bees are foraging on a specific flower, as it could result in the model considering a bee landing and then leaving the same flower as two separate visits.

Discussion

In this study we have developed a method that is able to detect bees on flowers in timelapse video footage. The method detects bees with good accuracy, over 0.95 recall and 0.99 precision on new data that we obtained specifically for this study. When applied to the data from Bjerge et al. 2023b a recall of 0.83 and precision of 0.83 was achieved. Our method performed worse than the Bjerge model which achieved a recall of 0.89 and precision of 0.90. However, it must be noted that these results were obtained without any training of our method on that dataset. In Bjerge et al. 2023b, the authors trained their model specifically on the data they used, requiring time consuming data labelling to be able to use their method. The method described here has no training or finetuning required beyond the creation of a synthetic first image, making it easily generalisable

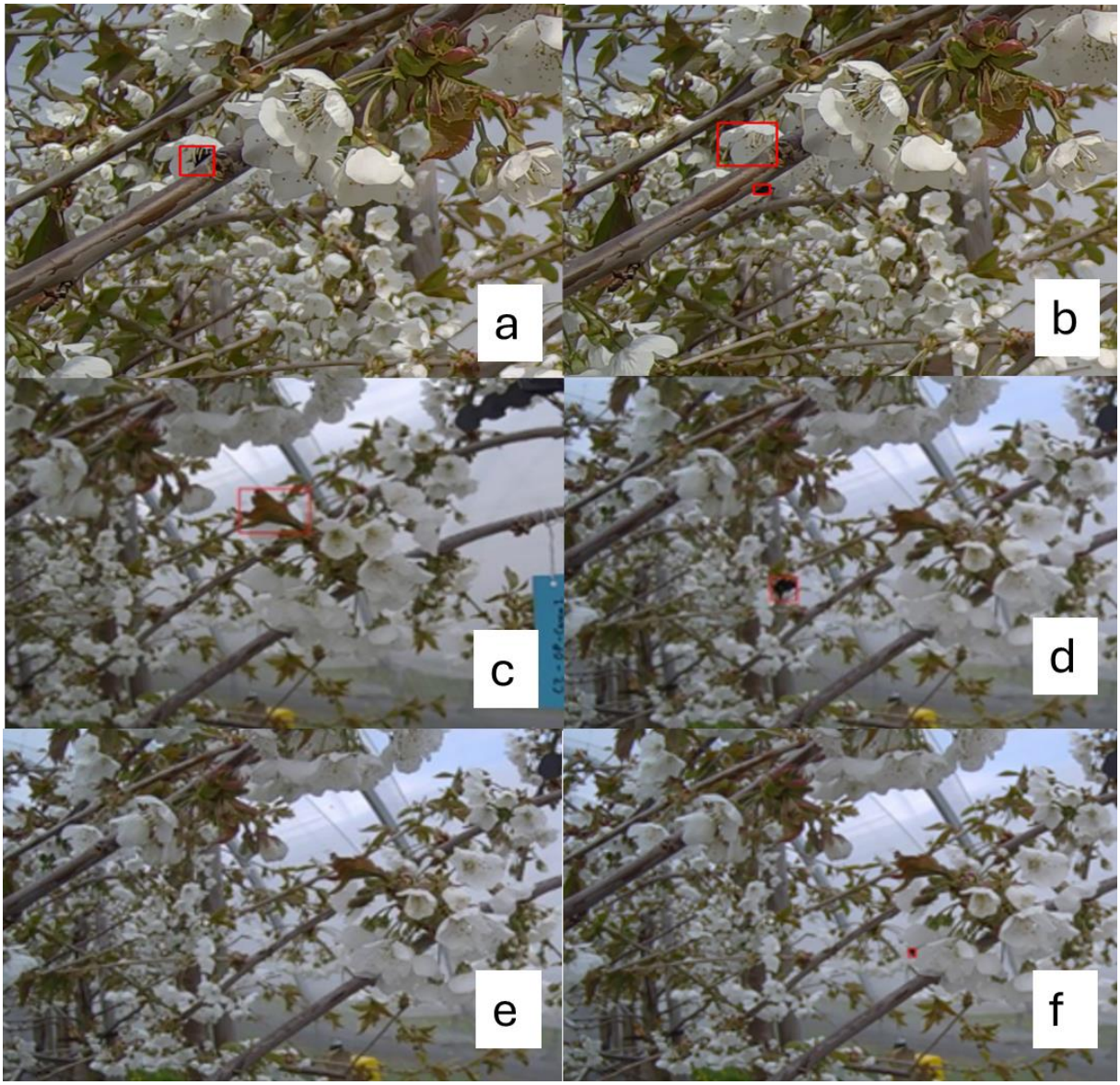

Figure 3: Example images showing failure cases. Images a and b show a sequence where a true positive detection (a bumblebee) becomes a false positive (a flower's petal). Image c shows an example of a false positive caused by the false detection of leaf buds as a bee. The sequence d, e and f show an example of a false negative due to a bee (initially correctly detected, panel d) becoming hidden within a flower in panel e and then re-emerging in panel f when exiting the flower.

to new settings. Their images included flowers from a range of different plant species and were recorded over a longer time period, so there may have been more variability in the data which could partly explain the relatively lower performance.

When using our method on data from Bjerge et al. 2023b we found that using bees from that study to create the synthetic image improved results. This indicates that further investigation into using 4 different bees for the initial synthetic image may improve results. Particularly if a wider variation of bee types were being investigated in a study. Given the very high recall and precision scores achieved on our data we did not feel the need to attempt this. It is also worth noting the differences in frequency of images gathered between the two datasets, our dataset was gathering images every second compared to every 30 seconds in the Bjerge dataset. The relative importance of SAM2 as opposed to Grounding DINO is probably altered by this. More frequent images allow for object tracking to happen frame to frame by SAM2 as bees will linger in shot across several frames. For datasets with longer time gaps, SAM2 may offer lower performance gains.

While we present here a combined method, we also compared the performance of SAM2 and Grounding DINO working separately. Comparing SAM2 and Grounding DINO we see SAM2 has much higher recall (0.94 v 0.58) and slightly lower precision (0.91 v 0.97) then Grounding DINO. This indicates that giving SAM2 a single image of a bee to track enables it to be used to detect examples in subsequent images. Adding in more examples of bees using Grounding DINO improved the performance of SAM2 showing the two models work best in combination.

Other studies have focussed on monitoring the entrance and exit of honeybees to/from hives. In such studies, additional information on whether the bees were entering or exiting the colony and the presence of pollen baskets was also required. Our model, as it stands, is not able to provide this information. Our output is a segmentation of the entire bee, with no information on the direction of movement. A second model, able to detect different parts of a bee, would be required to get additional information. However, considering that our model tracks bees through images, with sufficient video footage it may be able to track bees through frames of a video and therefore infer movement direction that way instead.

Our aim in this study was to monitor individual flowers to quantify how many times bees had visited a particular flower. This could be done by assigning an area of the image to a particular flower and then counting the number of times a bee was detected in this area. This would potentially also detect bees flying between flowers, but it would still provide a good guide to flower visitation which in turn could then be linked to future fruit drop. One potential addition to our model would be to add a classifier that can categorize bees based on whether they are in flight or have landed on a flower. This would help determine which flowers are undergoing pollination in the images. Nevertheless, it is still possible to get a rough estimation of the two scenarios with our method, as the output from our model includes multiple detections of bees during a single visit to a flower because they stay in the same place for longer. While in contrast bees in flight will appear in only a few frames. This can be seen in Fig. 2 which shows the location of detected bees. This approach could also be used to estimate the duration a bee is present on a flower, a factor that seems to be positively correlated – until a certain point – with successful pollination.

A limitation of our work is that it cannot distinguish between different species of pollinators as others have done using YOLO based models (Bjerge et al. 2023a; Ștefan et al. 2025). If species level determination of pollinators was required, an additional image classification model capable of distinguishing pollinator species would need to be trained. Our SAM2 based method could be used to assist this by detecting the relevant pollinators which would then need to be classified. For datasets where pollinators are infrequently present, this could substantially reduce the time needed for data annotation. The creation of more benchmark datasets for insect detection would make this process easier and allow for better comparison between different techniques (Schneider et al. 2023).

While we have focussed on cherry trees in this study, the method we present could be applied to other contexts as shown by its successful application to data from different flowers in Bjerge et al. 2023b. It would be of particular interest to any researchers who want to study the fine interaction between pollinators and flowers. For example, how environmental factors may influence which flowers pollinators favour or identify specific plant features that increase floral attractiveness to pollinators.

Conclusion

A method called BeeSAM2 has been developed to monitor bees on cherry flowers using timelapse imaging and automated image analysis. A high level of accuracy was achieved, showing that the method is suitable for use by scientists wishing to monitor insect pollination. The model was not specifically trained on the data from our images, making it easily generalisable to new unseen data. It has been tested on data from Bjerge et al. 2023b where we found accuracy could be slightly improved by generated a new synthetic image specific to the data set, this is a lot less effort than producing an annotated dataset needed to train a custom model on new data.

Unlike other previous studies, this method does not attempt to distinguish pollinators to a species level. This is partly due to the trade-offs

required to create a generalisable model that can be applied to any species of insect, and because the experimental setup implemented to obtain the initial dataset, resulted in the overwhelming majority of pollinators in the footage being *Bombus terrestris audax* workers.

ACKNOWLEDGEMENTS

The research was supported by The Anthony and Margaret Johnston Centre for Doctoral Training in Plant Sciences, Charles Sutherland Sponsorship travel grant, Internal Funding of The University of Aberdeen Grants Academy to Pump-Prime Research and Research Networks; Project Title: 'Implementing the use of remote camera tools for the study of plant-pollinator interactions', Scottish Society of Crop Research, and the Scottish Government's Rural and Environment Science and Analytical Services Division (RESAS) through the strategic research program (2022-2027).

AUTHOR CONTRIBUTION

DW, AK, FM and JD designed the experimental setup; JD collected the data; DW and FM designed the methodology; DW and JD analysed the data; DW led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

GENERATIVE AI DISCLOSURE STATEMENT

No generative AI tools were used in the creation of this manuscript.

DATA AVAILABILITY STATEMENT

Data is available here https://www.ebi.ac.uk/biostudies/studies/S-BSST2341
Code is available here Dom3442/BeeSAM2

REFERENCES

Bjerge K, Alison J, Dyrmann M, Frigaard CE, Mann HM, Høye TT (2023a) Accurate detection and identification of insects from camera trap images with deep learning. PLOS Sustainability and Transformation 2:e0000051. https://doi.org/10.1371/journal.pstr.0000051

Bjerge K, Frigaard CE, Karstoft H (2023b) Object detection of small insects in time-lapse camera recordings. Sensors 23:7242. https://doi.org/10.3390/s23167242

Breeze TD, Bailey AP, Balcombe KG, Brereton T, Comont R, Edwards M, Garratt MP, Harvey M, Hawes C, Isaac N (2021) Pollinator monitoring more than pays for itself. Journal of Applied Ecology 58:44-57. https://doi.org/10.1111/1365-2664.13755

Gatti G (2024) Investigating the causes of late fruit drop in 'Regina'sweet cherry (*Prunus avium*). [Dissertation thesis], Alma Mater Studiorum Università di Bologna. https://doi.org/10.48676/unibo/amsdottorato/11144

Getz MP, Best LR, Melathopoulos AP, Warren TL (2024) The establishment and potential spread of *Osmia cornuta* (Hymenoptera: Megachilidae) in North America. Environmental Entomology 53:1147-1156. https://doi.org/10.1093/ee/nvae100

Gill RJ, Baldock KC, Brown MJ, Cresswell JE, Dicks LV, Fountain MT, Garratt MP, Gough LA, Heard MS, Holland JM (2016) Protecting an ecosystem service: approaches to understanding and mitigating threats to wild insect pollinators. Advances in ecological research 54: 135-206. https://doi.org/10.1016/bs.aecr.2015.10.007

Hinterstoisser S, Lepetit V, Wohlhart P, Konolige K (2018) On pre-trained image features and synthetic images for deep learning Proceedings of the European Conference on Computer Vision (ECCV) Workshops, pp 0-0. https://doi.org/10.1007/978-3-030-11009-3_42

Howard SR, Nisal Ratnayake M, Dyer AG, Garcia JE, Dorin A (2021) Towards precision apiculture: Traditional and technological insect monitoring methods in strawberry and raspberry crop polytunnels tell different pollination stories. PLoS One 16:e0251572. https://doi.org/10.1371/journal.pone.0251572

Høye TT, Montagna M, Oteman B, Roy DB (2025) Emerging technologies for pollinator monitoring. Current Opinion in Insect Science:101367. https://doi.org/10.1016/j.cois.2025.101367

Jain A, Cunha F, Bunsen M, Pasi L, Viklund A, Larrivée M, Rolnick D (2024) A machine learning pipeline for automated insect monitoring. arXiv preprint arXiv:2406.13031. https://doi.org/10.48550/arXiv.2406.13031

Jocher G, Stoken A, Borovec J, Changyu L, Hogan A, Diaconu L, Poznanski J, Yu L, Rai P, Ferriday R (2020) ultralytics/yolov5: v3. 0. Zenodo. https://doi.org/10.5281/zenodo.3983579

Johnson MD, Katz AD, Davis MA, Tetzlaff S, Edlund D, Tomczyk S, Molano-Flores B, Wilder T, Sperry JH (2023) Environmental DNA metabarcoding from flowers reveals arthropod pollinators, plant pests, parasites, and potential predator–prey interactions while revealing more arthropod diversity than camera traps. Environmental DNA 5:551-569. https://doi.org/10.1002/edn3.411

Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, Xiao T, Whitehead S, Berg AC, Lo W-Y (2023) Segment anything. arXiv preprint

arXiv:2304.02643. https://doi.org/10.1109/ICCV51070.2023.00371

Klein A-M, Vaissière BE, Cane JH, Steffan-Dewenter I, Cunningham SA, Kremen C, Tscharntke T (2007) Importance of pollinators in changing landscapes for world crops. Proceedings of the Royal Society B: Biological Sciences 274:303-313. https://doi.org/10.1098/rspb.2006.3721

Liu S, Zeng Z, Ren T, Li F, Zhang H, Yang J, Jiang Q, Li C, Yang J, Su H (2024) Grounding dino: Marrying dino with grounded pre-training for open-set object detection European Conference on Computer Vision. Springer, pp 38-55. https://doi.org/10.1007/978-3-031-72970-6_3

Mir MM, Mir M, Iqbal U, Mushtaq I, Rehman MU, Iqbal R, Parveze MU, Khan SQ, Rather GH, Banday SA (2025) The Impact of Pollination Requirements in Sweet Cherry: A Systemic Review. Journal of Plant Growth Regulation:1-19. https://doi.org/10.1007/s00344-025-11642-6

Mullins CC, Esau TJ, Zaman QU, Toombs CL, Hennessy PJ (2024) Leveraging Zero-Shot Detection Mechanisms to Accelerate Image Annotation for Machine Learning in Wild Blueberry (*Vaccinium angustifolium* Ait.). Agronomy 14:2830. https://doi.org/10.3390/agronomy14122830

Ngo TN, Rustia DJA, Yang E-C, Lin T-T (2021) Automated monitoring and analyses of honey bee pollen foraging behavior using a deep learning-based imaging system. Computers and Electronics in Agriculture 187:106239. https://doi.org/10.1016/j.compag.2021.106239

Potts SG, Imperatriz-Fonseca V, Ngo HT, Aizen MA, Biesmeijer JC, Breeze TD, Dicks LV, Garibaldi LA, Hill R, Settele J (2016) Safeguarding pollinators and their values to human well-being. Nature 540:220-229. https://doi.org/10.1038/nature20588

Ratnayake MN, Dyer AG, Dorin A (2021) Towards computer vision and deep learning facilitated pollination monitoring for agriculture Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2921-2930. https://doi.org/10.1109/CVPRW53098.2021.00327

Ravi N, Gabeur V, Hu Y-T, Hu R, Ryali C, Ma T, Khedr H, Rädle R, Rolland C, Gustafson L (2024) Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714. https://doi.org/10.48550/arXiv.2408.00714

Redmon J, Farhadi A (2018) Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767. https://doi.org/10.48550/arXiv.1804.02767

Ren T, Liu S, Zeng A, Lin J, Li K, Cao H, Chen J, Huang X, Chen Y, Yan F (2024) Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159. https://doi.org/10.48550/arXiv.2401.14159

Schneider S, Taylor GW, Kremer SC, Fryxell JM (2023) Getting the bugs out of AI: Advancing ecological research on arthropods through computer vision. Ecology Letters 26:1247-1258. https://doi.org/10.1111/ele.14239

Sengupta S, Chakrabarty S, Soni R (2025) Is SAM 2 better than SAM in medical image segmentation? Medical Imaging 2025: Image Processing, vol. 13406. SPIE, pp 666-672. https://doi.org/10.1117/12.3047370

Shirali H, Hübner J, Both R, Raupach M, Reischl M, Schmidt S, Pylatiuk C (2024) Image-based recognition of parasitoid wasps using advanced neural networks. Invertebrate Systematics 38 (6): IS24011. https://doi.org/10.1071/IS24011

Steen R (2017) Diel activity, frequency and visit duration of pollinators in focal plants: in situ automatic camera monitoring and data processing. Methods in Ecology and Evolution 8:203-213. https://doi.org/10.1111/2041-210X.12654

Ștefan V, Stark T, Wurm M, Taubenböck H, Knight TM (2025) Successes and limitations of pretrained YOLO detectors applied to unseen time-lapse images for automated pollinator monitoring. Scientific Reports 15:30671. https://doi.org/10.1038/s41598-025-16140-z

Ștefan V, Workman A, Cobain JC, Rakosy D, Knight TM (2024) Utilising affordable smartphones and open-source time-lapse photography for pollinator image collection and annotation. Journal of Pollination Ecology 37:1-21. https://doi.org/10.26786/1920-7603(2025)778

Van Klink R, Sheard JK, Høye TT, Roslin T, Do Nascimento LA, Bauer S (2024) Towards a toolkit for global insect biodiversity monitoring, vol. 379. The Royal Society, (1904): 20230101. https://doi.org/10.1098/rstb.2023.0101

Williams D, Macfarlane F, Britten A (2024) Leaf only SAM: A segment anything pipeline for zero-shot automated leaf segmentation. Smart Agricultural Technology 8:100515. https://doi.org/10.1016/j.atech.2024.100515

Zhu J, Hamdi A, Qi Y, Jin Y, Wu J (2024) Medical sam 2: Segment medical images as video via segment anything model 2. arXiv preprint arXiv:2408.00874. https://doi.org/10.48550/arXiv.2408.00874

This work is licensed under a Creative Commons Attribution 4.0 License. ISSN 1920-7603