Fruit Detection in the Wild: The Impact of Varying Conditions and Cultivar

Michael Halstead[1] Simon Denman[2], Clinton Fookes[2], Chris McCool[1]

Abstract—**Agricultural robotics is a rapidly evolving research field due to advances in computer vision, machine learning, robotics, and increased agricultural demand. However, there is still a considerable gap between farming requirements and available technology due to the large differences between cropping environments. This creates a pressing need for models with greater generalisability.**

We explore the issue of generalisability by considering a fruit (sweet pepper) that is grown using different cultivar (subspecies) and in different environments (field vs glasshouse). To investigate these differences, we publicly release three novel datasets captured with different domains, cultivar, cameras, and geographic locations. We exploit these new datasets in a singular and combined (to promote generalisation) manner to evaluate sweet pepper (fruit) detection and classification in the wild. For evaluation, we employ Faster-RCNN for detection due to the ease in which it can be expanded to incorporate multitask learning by utilising the Mask-RCNN framework (instance-based segmentation). This multi-task learning technique is shown to increase the cross dataset detection F1-Score from 0.323 **to** 0.700**, demonstrating the potential to reduce the requirements of new annotations through improved generalisation of the model. We further exploit the Faster-RCNN architecture to include both super- and sub-classes, fruit and ripeness respectively, by incorporating a parallel classification layer. For sub-class classification considering the percentage of correct detections, we are able to achieve an accuracy score of** 0.900 **in a cross domain evaluation. In our experiments, we find that intra-environmental inference is generally inferior, however, diversifying the data by using a combination of datasets increases performance through greater diversity in the training data. Overall, the introduction of these three novel and diverse datasets demonstrates the potential for multi-task learning to improve cross-dataset generalisability while also highlighting the importance of diverse data to adequately train and evaluate real-world systems.**

I. Introduction

Agricultural robotics is increasingly prevalent due to advances in a number of fields, including robotics, computer vision, and machine learning. These advances are partly driven by the requirements placed on farmers to produce crop that are both high in yield and quality; while also reducing labour costs which have been reported [1] to be one of the most cost-demanding factors in agriculture. Improvements in these farming metrics requires automated technologies such as weed management [2], and harvesting [3], [4]. In these fields, robotic vision and machine learning will play an integral role in ensuring successful integration into existing processes.

A clear bottleneck for applying state-of-the-art machine learning techniques in agriculture is the ability for a trained system to perform well in different conditions. In this paper, we evaluate the ability to detect sweet pepper under varying conditions including: illumination, camera type, location, and sub-class/colouring (cultivar). To achieve this, we build off the prior work in [5] and explore the potential to use multi-task learning to enhance cross-domain performance.

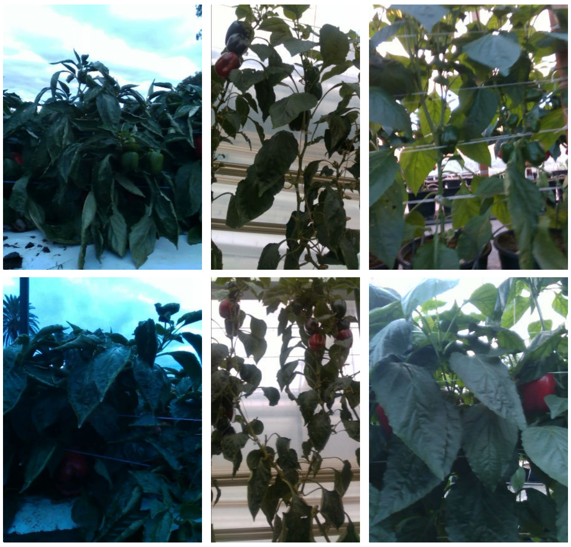


Fig. 1. Example images from each of the three datasets: (left column) QHDF field dataset; (middle column) BUP glass house dataset; and (right column) QHDP protected extended dataset.

To evaluate the performance of models in varying conditions we exploit three novel datasets and leverage these for both intra- and inter-environmental evaluations. To detect and classify objects we utilise the Faster-RCNN [6] framework in a similar method to [5], which provides bounding box locations of the fruit. We expand on this by incorporating multi-task learning using the Mask-RCNN framework [7] trained on instance based masks, resulting in pixel-wise segmentation. Multi-task learning, in this setting, demonstrates the ability to increase cross domain accuracy. This is enabled by exploiting the super-class similarity of the sweet pepper shape. In general sweet pepper maintain a similar overall appearance irrespective of differences in domain or environments such as differing colouration between cultivar (sub-species).

To enrich the information provided to agricultural workers, we also investigate classification of the ripeness (sub-class) of a sweet pepper. This property has the potential to greatly reduce labour costs through more intelligent workforce management (i.e. only assigning workers to fields that have high numbers of ripe fruit). We examine two methods of estimating

[1] These authors are with the University of Bonn, Bonn 53115, Germany {michael.halstead, cmccool}@uni-bonn.de

[2] These authors are with the Queensland University of Technology, Brisbane 4000, Australia {s.denman, c.fookes}@qut.edu.au

this property, quality as a super-class and quality as a parallel layer [5]. From this we show that classifying the sub-class using a parallel node increases performance by creating an undiluted representation of the super-class.

To enable these evaluations, large-scale datasets are required to both train and evaluate learning frameworks. Large and diverse datasets such as ImageNet [8] have proven beneficial to the research community due to their generalisation via the number of training samples. In agricultural research, however, existing datasets are often small and specific to one environment, such as grape counting [9], sweet pepper detection in a polytunnel [10], and cucumber detection [11]. The lack of available samples in these sets creates a gap between farming requirements and research output due in part to the individuality of each setting, such as the location, cultivar, or crop set up.

In this paper, we complete an analysis of sweet pepper detection in the wild, employing three novel datasets which we release publicly. Each dataset used in this evaluation represents a different domain. We exploit datasets collected in two unique geographical locations: Australia and Germany; and in three different set ups: field, polytunnel, and a glass house. Figure 1 shows an example from each of these datasets. The two QHD sets were collected in Australia by the **Q**ueensland University of Technology (QUT) with **H**orticulture Innovation Australia (HIA) and the **D**epartment of Agriculture and Fisheries (DAF). The final dataset (BUP) was collected in Germany by the University of Bonn.

To summarise, we perform an evaluation of sweet pepper detection, segmentation, and quality assessment. Evaluations show that while domain specific models yield higher performance on their source data, detection and segmentation of sweet pepper on unseen data is viable for agricultural robotics by exploiting multi-task learning. Our novel contributions include:

1) We improve cross domain generalisation through multi-task learning, and demonstrate the generality of the Mask-RCNN framework which when trained on the BUP dataset achieves impressive performance both within and across domains;
2) An analysis into the generalisation of the three new datasets and their combinations for detection using the Faster-RCNN framework, including the benefits of using a parallel classification layer for super- and sub-class detection;
3) The release of two novel QHD datasets captured on RealSense 200 cameras which contain RGB images, along with bounding boxes and sub-class annotations of each sweet pepper;
4) The release of a third sweet pepper dataset, BUP, captured using the Realsense 435i camera, containing raw bagfiles, registered RGB and depth images, and sub-class based instance segmentation masks.

II. Related Work

Crop detection and classification techniques have improved substantially since Nuske et al. [12] presented their grape yield estimation method using a radial symmetry transform. This early work, while sufficient for yield estimation, did have significant limitations with object occlusion.

Continuing with smaller objects, Hung et al. [13] proposed a yield estimation approach for almonds, and achieved impressive performance using a combination of a sparse auto-encoder and conditional random fields (CRF). However, once again occlusion was a limiting factor for accuracy.

In an effort to alleviate issues of occlusion, Zabawa et al. [9] transform segmentation and detection of grapes into a three class problem. Using neural networks they detect not only the location of the grapes themselves, but also the edges of the individual grape. This allows the grapes to be segmented into individual fruit, and counted for yield estimation. While limited data proved an issue for this technique, their overall accuracy, specifically in the face of multiple objects in a small space, was impressive.

McCool et al. [14] propose one of the earliest sweet pepper approaches. They segment each sweet pepper at the pixel level, using hand selected and sparse auto-encoder features fed to a CRF. Once again somewhat alleviating the occlusion problem this technique achieved results similar to a human.

Considering detection performance, these results were superseded by Sa et al. [10] and Halstead et al. [5]. In [10] Faster-RCNN [6], is employed to classify the object and it's bounds. They were able to achieve impressive results across a number of crops including sweet pepper. Expanding on [10], [5] sought to detect the "quality" (sub-class) of the crop using a parallel classification node added to the Faster-RCNN framework. This technique could detect both the super-class (sweet pepper) and the sub-class with high accuracy.

For the purpose of detecting apples at varying stages of the life cycle in an orchard, Tian et al. [15] evaluate different deep learning techniques for detection. Faster-RCNN and Yolo-V3 [16] are considered, with Yolo-V3 providing the best results. However while all of these techniques employ different detection routines, a common theme remains: the evaluation data is from the same domain as the training data. For a technique to be utilised on a robotic platform it requires generality as each farm, or field within a farm, has unique properties.

To evaluate the generalisability of these techniques datasets require variability in a number of areas including: lighting conditions, cultivar, and environment of capture. Early sweet pepper datasets include [10] and [5] which were captured in a single environment and cultivar with only a small number of samples. Models trained on this type of dataset capture the bias of both the environment and captured cultivar. In a similar manner [11] captured cucumber data in a greenhouse for a total of 522 images. While their technique employs data augmentation to significantly increase the volume of data it is still captured in a single environment. In an effort to increase the variability of their dataset [17] captures apple data in two separate orchards. While this data does have a number of similarities between environments (orchard setting and using the same camera) their dataset is complicated by the addition of night time capture under artificial light.

In many agricultural datasets the captured data is too small or specific to create models that are able to generalise. For robotic platforms to be usable across a broad range of farming environments more dynamic data is required. This data acts to ensure viability of detection, segmentation, or harvesting with the aim to reduce farming labour costs.

III. Datasets

This paper exploits three new publicly available datasets (see Figure 1 and Table I) and their combinations for detection and classification of sweet pepper. Each dataset contains unique properties (collection environment, camera used, illumination, occlusion, and cultivar) adding extra complexity and diversity during inference.

TABLE I
Number of images contained in each of the datasets used in this paper, where 'T', 'V', and 'E' represent the training, validation, and evaluation sets respectively.

Dataset	T	V	E	Height	Width	Camera
QHDF	509	604	470	640	480	RealSense 200
QHDP	345	86	256	640	480	RealSense 200
BUP	114	84	88	1280	720	RealSense 435i

All datasets are annotated for sweet pepper location (bounding box or instance based masks), super-class, and sub-class classification. Each dataset and their properties are described in the following sections.

A. QHD Datasets

In this paper we use two sweet pepper datasets collected in Australia by QUT as part of the DAF and HIA directives (denoted QHD[1]). The first dataset was collected under direct sunlight in a field situation (QHDF), and the second in a polytunnel (QHDP), providing some protection from the sun. Table II outlines the ground truth labels for each sub-class in these datasets. All cultivar in the QHD datasets consist of three sub-classes: green, mixed, and red.

TABLE II
Distribution of sweet pepper in each of the QHD datasets.

Dataset	Subset	Green	Mixed	Red
QHDF	training	1215	89	609
	validation	1389	94	716
	evaluation	1131	73	458
QHDP	training	782	170	718
	validation	208	34	155
	evaluation	956	190	528

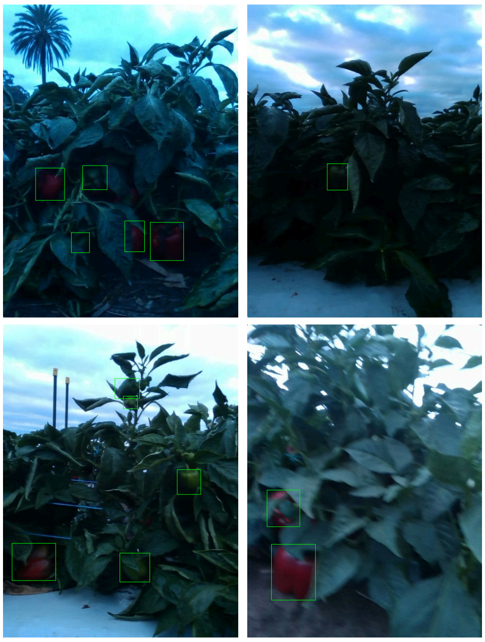


Fig. 2. Four example images with their respective bounding boxes from the QHDF dataset.

1) QHD Field: The QHD field (QHDF) dataset consists of two cultivar, *Warlock* and *SV6947*, each cultivar was planted in a single- and double-row plant configuration in outdoor field conditions. Due to the direct exposure to the sun plants in this domain are smaller in stature and with greater amounts of foliage to protect the fruit (as seen in Figure 2). The single- and double-row configurations also provide the potential for varying levels of foliage and occlusion. Data was collected using a RealSense 200 camera at resolutions of 640 × 480. Annotation of the sweet pepper location and their sub-classes was carried out by a single person with checks for ambiguity. Four examples from this dataset with their annotation are illustrated in Figure 2.

2) QHD Protected: The QHDP dataset is a super set of the data from [5] and was collected in the same manner, with a collection of examples shown in Figure 3. Annotation of the extra data was completed by a single individual with verification by a second. Slight variations exist between the original data and this super set, primarily due to the manual removal of foliage by farmers to make manual crop counting more efficient.

B. BUP

A novel, northern hemisphere based sweet pepper dataset is presented here: BUP (University of Bonn - Protected crop)[2]. The BUP dataset was captured in a glass house replicating a commercial setting at Campus Klein-Altendorf. Two different cultivar of sweet pepper were grown simultaneously during experiments: Mazurka (*Rijk Zwaan*) and Mavras (*Enza Zaden*). Mazurka mirrors the QHD datasets in ripening, however

[1] https://data.researchdatafinder.qut.edu.au/dataset/qut-hia-daf-capsicum-datasets

[2] http://agrobotics.uni-bonn.de/data/

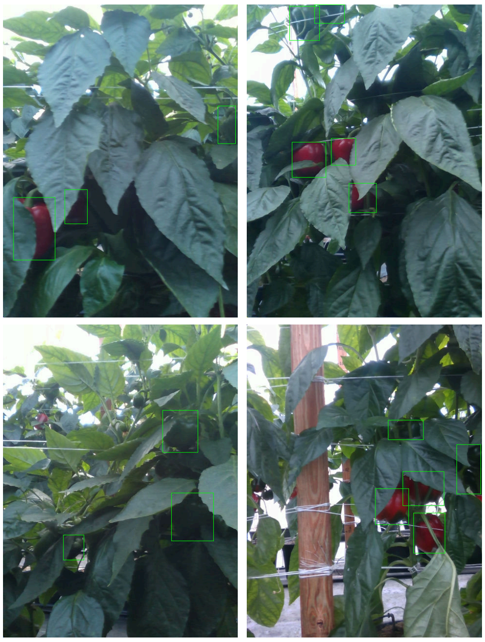


Fig. 3. Four example images with their respective bounding boxes from the QHDP dataset.

TABLE III
DISTRIBUTION OF THE SUB-CLASSES IN THE NEW BUP DATASET.

subset	Black	Green	Mixed	Red
train	1052	316	158	47
validation	450	400	79	62
evaluation	578	442	69	71

Mavras fruit ripens in four stages: green, black, mixed, red. The addition of the black sub-class adds further complexity when compared to the QHD counterparts. In addition, plants in the BUP dataset grow significantly taller with much sparser foliage due to the glass house setting.

The glass house for sweet pepper cultivation was arranged into six rows of approximately 40m in length each. Data was recorded into bagfiles using an Intel RealSense D435i camera at 30fps. For recording each row was separated into four equally spaced sections. Post processing was completed to align the depth and RGB images using the pyrealsense2[3] libraries (see Figure 4-(a) and -(g) for examples). The stored depth image is a *uint16* TIFF format file where 1mm is represented by each change in value.

For annotation, the glasshouse data was separated into three distinct sections: 1/3 training, 1/3 validation, and 1/3 evaluation. The separation of sections during recordings allowed for the data to be evenly split between each sub-set. Extending beyond bounding box regression, instance based masks are annotated. Annotation was completed by three individuals who annotated different images. A separate mask is included for each sub-class where zero denotes "background", and a numbered response indicates the presence of a sweet pepper, with examples outlined in Figure 4.

The introduced dataset contains the four sub-classes of sweet pepper with distributions shown in Table III. Illumination variations between rows create natural fluctuation in the data, creating a complicated and challenging dataset.

IV. IMPLEMENTATION AND EVALUATION PROTOCOLS

Following [5], we implement Faster-RCNN for fruit and sub-class detection and classification. Faster-RCNN was selected due to the ease in which multi-task learning can be explored by exploiting the Mask-RCNN framework. To evaluate the generality of these models we perform same and cross domain ("sweet pepper in the wild") analysis. Different experimental set ups, with and without the sub-class node, are evaluated. To show the versatility of the new BUP dataset we implement instance based segmentation of sweet pepper and evaluate it's performance across domains. Our results are presented using the standard F1-Score metric, for further detail refer to [5].

A. Hyperparameters

Faster-RCNN and Mask-RCNN models are based on the PyTorch implementations in TorchVision[4]. Training uses SGD as the optimizer, with a uniform learning rate of 0.001, momentum of 0.9, and weight decay of 0.0005. A batch size of four is used with mean normalisation. We train for 250 epochs and the results displayed here use the validation set for parameter selection. For evaluations with Mask-RCNN, we empirically found that a hidden layer of 256 neurons achieved optimal results. We do not employ data augmentation in our experiments as [10] outlined that high performance can be achieved with as few as 100 samples when fine tuning.

B. Super-Class

The following evaluations concentrate on super-class (sweet pepper) detection only. We evaluate the performance of both detection (Faster-RCNN) and segmentation (Mask-RCNN) for both intra- and inter-environment accuracy.

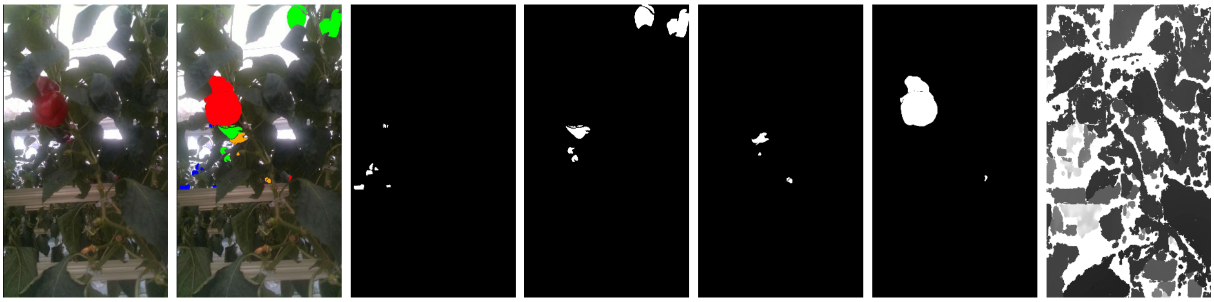
1) Detection: Super-class classification is considered by implementing the standard Faster-RCNN classification technique. In the "classification" only component of Figure 5, super-class classification only uses the unchecked classification and regression boxes.

In the two class problem we simplify the classification into "background" and "sweet pepper". With this protocol, all datasets can be directly compared.

2) Segmentation: In a harvesting situation, pixel wise segmentation has considerable benefits over bounding boxes alone. Segmentation allows for more accurate object localisation and picking [4]. The new BUP dataset contains instance based segmentation of the sub-classes in the dataset. We perform instance based segmentation using the two class problem setting (standard Faster-RCNN) from Section IV-B1 and the Mask-RCNN framework (see Figure 5, Segmentation and Classification).

[3] https://github.com/IntelRealSense/librealsense

[4] https://pytorch.org/tutorials/intermediate/torchvision_tutorial.html



(a) (b) (c) (d) (e) (f) (g)

Fig. 4. Example images from the BUP dataset: (a) is the raw image, (b) is a colourised version of the instance masks, (c)-(f) are representations of the instance masks for black, green, mixed, and red, and (g) is a quantized version of the depth image for visualisation.

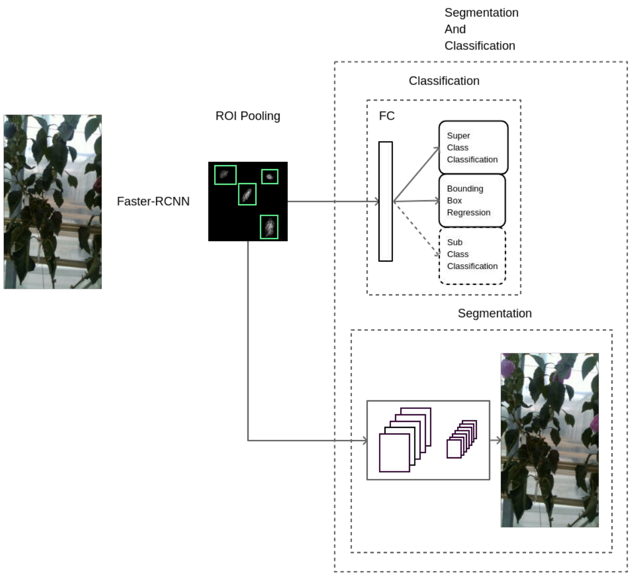


Fig. 5. Faster-RCNN and Mask-RCNN frameworks. In standard Faster-RCNN networks only the classification step is included, while Mask-RCNN incorporates both paths. The standard Faster-RCNN framework is extended to include sub class classification using a parallel layer directly classifying data off the final embedding.

Calculation of the F1-Score requires intersection over union (IoU) based precision recall scores. The BUP scores can be calculated directly from the segmentation ground truth contained in the dataset. However, for cross dataset analysis we leverage the bounding box ground truth contained in the QHD datasets and compare them to the masks estimated at inference. While bounding box comparisons may not accurately reflect system performance for segmentation (e.g. for Mask-RCNN) it does give a strong indication of the cross dataset generalisation using the BUP dataset for training.

C. Sub-Class

Sub-class evaluation is performed using two networks based on the Faster-RCNN framework. In each technique the primary goal is to detect not only the location of the sweet pepper but also to accurately detect the underlying sub-class.

1) N Super-Classes: Initially we implement the same network as that in Section IV-B1. However, instead of there being two classes (background and sweet pepper) we extend the sub-classes as super-classes. For this evaluation, for BUP we combine black and green into a single class (due to their visual similarity) and evaluate for: “background”, “green”, “mixed”, and “red”. For this problem we report the average F1-Score of the individual sub-classes.

2) Parallel Sub-Class: Sub-class classification evaluation follows [5] which introduced the parallel classification layer shown in Figure 5 (perforated lines representing sub-class classification). The technique allows the model to learn super- and sub-class classification layers. In this manner we are able to leverage all sweet pepper data to create a strong super-class model, while also classifying the sub-classes.

Two metrics are considered, the super-class F1-Score and the average of the confusion matrix (sub-classes), such that,

$$CAve = \frac{1}{N_r} \times \sum_{i}^{N_r} \frac{conf(i,i)}{\sum_{j}^{N_c} conf(i,j)}, \tag{1}$$

where N_r and N_c are the number of rows and columns respectively ($N_r = N_c$ as this is a confusion matrix), and the confusion matrix is represented by $conf(\cdot)$.

V. Evaluation

The experiments contained in this section follow those outlined in Section IV. Datasets are made up of the three primary sets and four combined sets (seven in total), listed in Table IV. Two BUP datasets are used, one which combines black and green (BUP3), and one that contains all four sub-classes (BUP4). In all experiments the BUP dataset images are downsized to the height of the QHD sets, with a final resolution of 640×360. For evaluations relating to BUP4, any dataset or combination thereof that does not include the ‘black’ sub-class is not included.

When referring to different combinations that use the BUP dataset, if the four is not designated then we are using the BUP3 version. For example QFBUP refers to using the three class version of the BUP dataset combined with QHDF, whereas, QFBUP4 refers to using the four class version of the BUP dataset with QHDF.

TABLE IV
THE COMBINED DATASETS AND THEIR ACRONYMS. THE FOUR IN BRACKETS INDICATES WHETHER THE BUP USES ALL THE SUB-CLASSES (BUP4), IF THE FOUR IS NOT DESIGNATED THEN WE USE BUP3.

Acronym	Combinations
QHDC	QHDF + QHDP
QFBUP(4)	QHDF + BUP(4)
QPBUP(4)	QHDP + BUP(4)
All(4)	QHDF + QHDP + BUP(4)

TABLE V
F1-SCORES OF THE TWO CLASS PROBLEM. TRAINING DATA IS IN THE ROWS AND EVALUATION DATA USED IN IN THE COLUMNS.

Evaluation Train	QHDF	QHDP	BUP	QHDC	QFBUP	QPBUP	All
QHDF	0.783	0.755	0.465	0.773	0.665	0.649	0.702
QHDP	0.696	0.877	0.508	0.783	0.629	0.747	0.725
BUP	0.323	0.647	0.775	0.463	0.456	0.691	0.523
QHDC	0.792	**0.879**	0.524	**0.836**	0.696	0.749	0.765
QFBUP	0.763	0.763	**0.780**	0.763	0.763	0.763	0.763
QPBUP	0.681	0.873	0.773	0.775	0.708	**0.828**	0.770
All	**0.795**	0.866	0.755	0.830	**0.777**	0.820	**0.810**

All evaluations consider two factors: how well a model targets the source dataset, and how well it generalises to unseen data. Tables display the detection F1-Scores where data used to train the models are presented in the rows, and the evaluation datasets are displayed in the columns. In each table we highlight the best performing model for each of the datasets.

A. *Two Class Detection*

The two class evaluation reports the performance on detection of sweet pepper only, i.e. there are no sub-classes.

For the uncombined datasets, we see from Table V that the target and source sets perform best, and domain shift between datasets is clearly visible. This is particularly evident when comparing the BUP and QHDF sets, where BUP to QHDF scores only 0.323. Sensor, illumination and cultivar variation are factors in this reduced performance.

The addition of data when combining datasets generally has a positive impact on performance. In the case of QHDC, we see an increase in performance on the BUP datset when compared to the score when QHDF is used to train the model. Overall, while these datasets are dissimilar in nature, their combination typically improves (or maintains) performance across all evaluations. A primary factor in the combined performance is the nature of the two class problem, which only detects based on the appearance of a sweet pepper and adding more samples diversifies the training set and improves performance.

B. *Instance Segmentation Evaluation*

As a further study into the two class problem we evaluate instance based segmentation using the BUP dataset which is trained using the Mask-RCNN model. This evaluation will ascertain whether multi-task learning impacts cross-domain inference.

This evaluation yields detection F1-Scores of 0.700, 0.837, and 0.789 on the QHDF, QHDP, and BUP datasets respectively. Interestingly the QHDP dataset obtains a higher F1-Score than the source dataset. The BUP score, of 0.789, refers to the instance-based segmentation (per-pixel) performance, whereas, the QHDF and QHDP scores refer to bounding box detection as discussed in Section IV-B2.

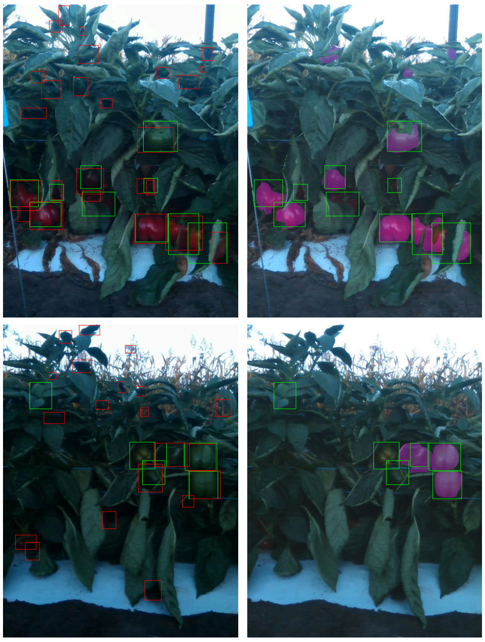


Fig. 6. Example images using the BUP data for training and evaluated on the QHDF dataset. The left column represents the 2 class Faster-RCNN where the green bounding boxes represent the ground truth and the red the detections. On the right we show the same image using Mask-RCNN, where the green bounding boxes are once again the ground truth and the magenta segmented regions are the output from the framework.

Of particular interest here is the severe reduction in domain shift experienced between the BUP and QHDF datasets. In Table V, the two class detection task, the achieved F1-Score was 0.323, considerably lower than the 0.700 achieved here. We attribute the considerable improvement to the influence of multi-task learning of both the segmentation and classification. The dual losses used during training are able to improve generalisability of the cross-domain models.

Based on these results it is evident that multi-task learning can play a significant role in cross domain localisation. Figure 6 clearly outlines one of the primary benefits witnessed, a reduction in false detections (when compared to two class detection).

As a qualitative analysis, we present a positive and negative example from the QHDF and QHDP datasets in Figure 7. A common issue in both datasets is the detection of foreign objects, like pots in the negative example. Objects such as these are not included in the BUP dataset and as such appear similar in shape and colour to the black sweet pepper found in the BUP dataset. In the difficult, heavily occluded QHDF dataset the BUP segmentation model is still able to perform reliably for sweet pepper segmentation.

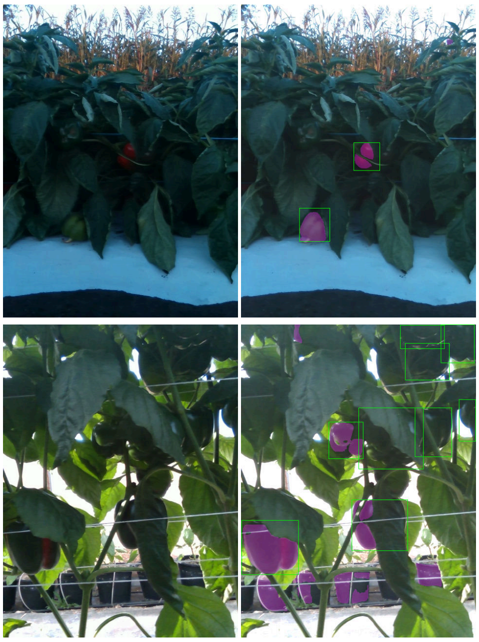


Fig. 7. Visualisation of a positive example from QHDF (left two images) and a negative example from QHDP (right two images). In each case the left image is the original, and the right image is the segmentation result.

TABLE VI
N-CLASS EVALUATION, REPORTED SCORES ARE BASED ON THE AVERAGE F1-SCORE OF THE THREE LABELS (GREEN, MIXED, AND RED).

Train \ Evaluation	QHDF	QHDP	BUP3	QHDC	QFBUP	QPBUP	All
QHDF	0.599	0.535	0.369	0.564	0.550	0.496	0.536
QHDP	0.565	0.723	0.535	0.649	0.559	0.658	0.625
BUP3	0.225	0.581	0.592	0.411	0.344	0.608	0.455
QHDC	**0.678**	**0.761**	0.500	**0.728**	0.636	0.689	0.690
QFBUP	0.619	0.614	0.562	0.621	0.663	0.638	0.641
QPBUP	0.539	0.735	**0.617**	0.645	0.598	0.718	0.655
All	0.665	0.747	0.610	0.717	**0.686**	**0.726**	**0.712**

C. N Super-Class Evaluation

The N super-class evaluation treats the sub-classes as super-classes. The average F1-Score for the N-classes is displayed in Table VI.

Comparing the two class and N-class models we see a considerable drop in performance. In the two class problem the QHDP best F1-Score was 0.877 compared to 0.723 in Table VI. This performance decrease can be attributed to the dilution of what constitutes a sweet pepper. It could be considered that as the number of sweet pepper in each class is increased the performance of the N-class problem also increases. Similar to the two class problem, it appears that the combination of data creates more general models. This is most evident when comparing QFBUP to the QHDP dataset. In the single dataset evaluations the BUP trained model is able to achieve the highest score of 0.581, where the combined set produces 0.614. This is an increase in performance and shows that generalised models are able to detect and classify sweet peppers from unseen domains.

D. Sub Class Evaluation

Based on the results in Sections V-A and V-C and the work in [5], we train models based on the super- and sub-class network. This approach uses a parallel node to train super- and sub-class classifiers, ensuring the maximum possible data available for each task.

The use of a parallel node for sub-class classification is vindicated by the results in Table VII. We see a marked improvement in F1-Score over the N-class results, and achieve results commensurate with the two class evaluation. Interestingly, a high F1-Score does not necessarily indicate the highest sub-class classification ($CAve$) performance and vice versa.

Overall the sub-class layer is able to accurately detect the colour in sweet pepper to a high degree of accuracy. This accuracy is generally higher when we train models based on all the data. In these evaluations we are consistently able to produce high F1-Scores and $CAve$ scores. This is highlighted by a $CAve$ score of 0.900 on the QFBUP variant.

Once again we see the benefit of combining the data where the *All* model is consistently able achieve the best $CAve$ scores. It should be noted that in the QHD datasets the green colour dominates, and including this in the combined sets acts to increase the $CAve$ score. This is evident when comparing to datasets that include the BUP dataset, particularly the single BUP model. However, overall we again see the value in combining datasets to create a more generic model, as this acts to increase not only the F1-Score but the $CAve$.

We also investigate the results of the four sub-class (BUP4) problem using the parallel layer. Table VIII indicates that the inclusion of the fourth class decreases sub-class detection, while the super-class performance remains similar. The addition of this extra sub-class adds complexity to the classification problem. Similar to the N-class problem in Section V-C where the sweet pepper was diluted through multiple super-classes, here we are diluting the sub-classes by splitting 'black' and 'green' back into their original labels. Another possible reason for this is noise in the annotations. Through thorough investigation during these evaluations it was noted that sweet pepper were both missed on occasion, and in some cases subjectively annotated. As the training, validation, and testing sets are were annotated by different individuals, subjectivity may be impacting results.

Once the combined datasets are used we see a marked increase in accuracy when compared to the BUP4 model alone. This is partly due to the increased availability of sweet peppers, and the significant increase in the green sub-class performance. Dataset combination positively impacts the generalisation of models, with significant improvements in both the sub- and super-class scores.

VI. CONCLUSION

In this paper we have explored cross-domain performance of state-of-the-art detection systems, applied to an agricultural setting (sweet pepper). Consistent with other research we show that exploring multiple domains during training improves overall performance. However, we found that the incorporation

TABLE VII
PARALLEL NODE EVALUATION BASED ON THREE SUB-CLASSES, THE F1-SCORE OF THE SUPER CLASS AND THE AVERAGE OF THE SUB-CLASS CONFUSION MATRIX ARE DISPLAYED.

Evaluation	QHDF		QHDP		BUP3		QHDC		QFBUP		QPBUP		All	
Train	F1	CAve	F1	CAve	F1	CAve	F1	CAve	F1	CAve	F1	CAve	F1	CAve
QHDF	0.778	0.721	0.725	0.694	0.442	0.701	0.751	0.706	0.654	0.757	0.620	0.709	0.682	0.720
QHDP	0.672	0.806	**0.874**	**0.858**	0.503	0.773	0.770	0.842	0.611	0.807	0.745	0.848	0.714	0.837
BUP3	0.314	0.777	0.645	0.758	0.762	0.715	0.452	0.825	0.441	0.652	0.686	0.767	0.510	0.701
QHDC	0.785	0.801	0.873	0.834	0.511	0.722	0.829	0.830	0.690	0.808	0.748	0.827	0.762	0.827
QFBUP	0.767	0.797	0.791	0.758	**0.769**	**0.678**	0.779	0.765	0.767	0.832	0.780	0.776	0.775	0.784
QPBUP	0.694	0.796	0.874	0.841	0.752	0.711	0.784	0.824	0.714	0.812	0.825	0.835	0.775	0.826
All	**0.787**	**0.813**	0.872	0.873	0.756	0.743	**0.830**	**0.858**	**0.774**	**0.900**	**0.826**	**0.872**	**0.811**	**0.865**

TABLE VIII
SUB CLASS EVALUATION, RESULTS ARE DISPLAYED AS A CONFUSION MATRIX WITH THE SUPER CLASS (BACKGROUND, SWEET PEPPER) F1-SCORE THEN THE AVERAGE SUB CLASS CONFUSION MATRIX RESULT.

Evaluation	BUP4		QFBUP4		QPBUP4		All4	
Train	F1	CAve	F1	CAve	F1	CAve	F1	CAve
BUP4	0.756	0.640	0.427	0.629	0.658	0.718	0.489	0.676
QFBUP4	0.760	0.617	0.771	0.825	0.769	0.775	0.772	0.804
QPBUP4	**0.761**	**0.621**	0.722	0.805	**0.831**	**0.840**	0.779	0.841
All4	0.752	0.636	**0.772**	**0.859**	0.825	0.837	**0.811**	**0.848**

of multi-task learning, extending the Faster-RCNN framework to the Mask-RCNN framework, greatly increases cross-domain performance from 0.323 to 0.700. Furthermore, qualitative research outlines the benefits of instance-based segmentation in a cross domain setting, particularly relating to a reduction in false detections. These evaluations outline the potential for models to generalise well across domains, effectively decreasing farming costs through increased workforce efficiency.

To enable this work, we have made use of three novel datasets which we publicly release. These datasets reflect challenging real-world agricultural conditions. Variations in domain are captured by each dataset with different colouration due to cultivar (sub-species), changes in environment (field, polytunnel, and glass house), location (northern and southern hemisphere), and cameras.

Future work should explore ways in which the extra depth modality captured in the dataset can further enhance performance. Similarly, the potential for domain adaptation techniques, in conjunction with mutli-task learning should also be explored.

ACKNOWLEDGEMENTS

This work was partially funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2070 – 390732324. It was also partly funded by Hort Innovation Australia, using the vegetable research and development levy and contributions from the Australian Government project (VG15024). We would also like to give special mention to Susanne Heisswolf, David Carey, Mary Firrell, Elio Jovicich, and Heidi Wittl from the Department of Agriculture and Fisheries in Australia for their contribution to the QHD dataset collections.

REFERENCES

[1] ABARES, "Farm financial performance," in *Australian Bureau of Agriculture and Resource Economics and Sciences (ABARES)*, 2018.

[2] D. Hall, F. Dayoub, J. Kulk, and C. McCool, "Towards unsupervised weed scouting for agricultural robotics," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 5223–5230.

[3] C. Lehnert, A. English, C. McCool, A. W. Tow, and T. Perez, "Autonomous sweet pepper harvesting for protected cropping systems," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 872–879, 2017.

[4] B. Arad, J. Balendonck, R. Barth, O. Ben-Shahar, Y. Edan, T. Hellström, J. Hemming, P. Kurtser, O. Ringdahl, T. Tielen, *et al.*, "Development of a sweet pepper harvesting robot," *Journal of Field Robotics*, 2020.

[5] M. Halstead, C. McCool, S. Denman, T. Perez, and C. Fookes, "Fruit quantity and ripeness estimation using a robotic vision system," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 2995–3002, 2018.

[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[9] L. Zabawa, A. Kicherer, L. Klingbeil, A. Milioto, R. Topfer, H. Kuhlmann, and R. Roscher, "Detection of single grapevine berries in images using fully convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[10] I. Sa, Z. Ge, F. Dayoub, B. Upcroft, T. Perez, and C. McCool, "Deepfruits: A fruit detection system using deep neural networks," *Sensors*, vol. 16, no. 8, p. 1222, 2016.

[11] X. Liu, D. Zhao, W. Jia, W. Ji, C. Ruan, and Y. Sun, "Cucumber fruits detection in greenhouses based on instance segmentation," *IEEE Access*, vol. 7, pp. 139 635–139 642, 2019.

[12] S. Nuske, S. Achar, T. Bates, S. Narasimhan, and S. Singh, "Yield estimation in vineyards by visual grape detection," in *IROS*, 2011.

[13] C. Hung, J. Nieto, Z. Taylor, J. Underwood, and S. Sukkarieh, "Orchard fruit segmentation using multi-spectral feature learning," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 5314–5320.

[14] C. McCool, I. Sa, F. Dayoub, C. Lehnert, T. Perez, and B. Upcroft, "Visual detection of occluded crop: For automated harvesting," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 2506–2512.

[15] Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, and Z. Liang, "Apple detection during different growth stages in orchards using the improved yolo-v3 model," *Computers and electronics in agriculture*, vol. 157, pp. 417–426, 2019.

[16] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[17] X. Liu, D. Zhao, W. Jia, W. Ji, and Y. Sun, "A detection method for apple fruits based on color and shape features," *IEEE Access*, vol. 7, pp. 67 923–67 933, 2019.