S^{3}AD: Semi-supervised Small Apple Detection in Orchard Environments

Robert Johanson*, Christian Wilms*, Ole Johannsen, and Simone Frintrop
Computer Vision Group, University of Hamburg, Germany
* denotes equal contribution
{firstname}.{lastname}@uni-hamburg.de

Abstract

Crop detection is integral for precision agriculture applications such as automated yield estimation or fruit picking. However, crop detection, e.g., apple detection in orchard environments remains challenging due to a lack of large-scale datasets and the small relative size of the crops in the image. In this work, we address these challenges by reformulating the apple detection task in a semi-supervised manner. To this end, we provide the large, high-resolution dataset MAD[1] comprising 105 labeled images with 14,667 annotated apple instances and 4,440 unlabeled images. Utilizing this dataset, we also propose a novel Semi-Supervised Small Apple Detection system S^{3}AD[2] based on contextual attention and selective tiling to improve the challenging detection of small apples, while limiting the computational overhead. We conduct an extensive evaluation on MAD and the MSU dataset, showing that S^{3}AD substantially outperforms strong fully-supervised baselines, including several small object detection systems, by up to 14.9%. Additionally, we exploit the detailed annotations of our dataset w.r.t. apple properties to analyze the influence of relative size or level of occlusion on the results of various systems, quantifying current challenges.

1. Introduction

Pre-harvest yield estimation is an integral part of agriculture for efficiently planning harvest, sales, transportation, and storage of crops [1, 7, 34, 38]. Yield estimation usually relies on labor-intensive manual counting in sample locations [1, 7, 34, 38], as well as weather information and historical data [1, 7, 38]. However, such estimations are inaccurate due to natural variances in fruit load, soil,- and light exposure [12, 34]. Recently, precision agriculture has received significant attention, leading to vision-based systems
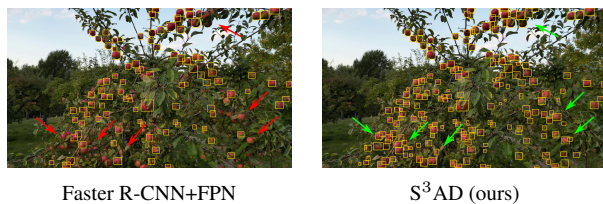
---

[1]MAD is available at www.inf.uni-hamburg.de/mad
[2]Code is available at www.github.com/RbtJhs/SSSAD

| Faster R-CNN+FPN | S^{3}AD (ours) |

Figure 1. Apple detection results using Faster R-CNN+FPN [18, 27] and our proposed, semi-supervised small apple detection system S^{3}AD with selective tiling on a test image of our MAD dataset. Red arrows denote missed apples, while green arrows denote the detection of previously missed apples by the other system.

for automated yield estimation [1, 7], quality control [37], or fruit picking [9].

An essential task in the development of such systems is the reliable detection of crops [8]. Despite recent advancements [1, 7, 8, 34, 37], this task remains challenging. For instance, the detection of apples in orchards remains difficult due to the complex environment caused by multiple factors: (i) the dense apple distribution, (ii) occlusion and shadow induced by other crops or leaves, and (iii) the small size of the apples compared to the trees. These effects are also visible from the example in Fig. 1 depicting a typical apple tree in an orchard environment. Especially challenging is the small apple size, since the performance of object detectors decreases significantly for small objects [4], due to the inherent down-sampling in CNNs and the limited GPU resources preventing the processing of entire high-resolution images. Moreover, the limited availability of data [3, 8] is another challenge in apple detection.

To address the aforementioned challenges, several approaches were proposed in general object detection. To cope with a limited amount of data, transfer learning [35] is used to utilize features learned on tasks with less annotation effort. Furthermore, semi-supervised systems were proposed to incorporate unlabeled data in object detection by consistency [10, 36] and pseudo-label methods [33, 41]. To improve the challenging detection of small objects, several directions were explored [5] including multiscale fea-

ture extraction [18, 19, 32, 42] to improve the feature representation of small objects by learning scale-specific features, attention-based methods [20, 44, 45] that aim to select relevant areas or features within the network, and tiling strategies [25, 39, 43] that improve the relative size of objects on input level.

In this paper, we address the challenging detection of apples in orchard environments by reformulating the problem in a semi-supervised manner. To tackle this problem, we propose a novel semi-supervised small apple detection system S^3AD focusing on small apples and a new large-scale apple detection dataset MAD for semi-supervised learning. MAD consists of 4,545 high-resolution images from apple orchards with 14,667 manually annotated apples across 105 images. The remaining images support semi-supervised learning. To address the problem of apple detection in a semi-supervised way and improving the challenging detection of small apples, we propose S^3AD. It is composed of three main modules: (i) an object detector trained with the semi-supervised pseudo-labeling framework Soft Teacher [41], which allows us to utilize the large amount of unlabeled data in our dataset, (ii) the TreeAttention module that leverages the contextual relationship between the apples and the tree crowns to localize regions of interest, and (iii) a selective tiling module that crops tiles from the regions of interest, enabling the object detector to utilize the full image resolution increasing the detection performance on small apples. In our extensive evaluation on two datasets, we show the strong results of S^3AD on small apples and apples across all sizes, utilizing contextual attention and selective tiling.

In summary, the contributions of this paper are threefold:

- We reformulate apple detection as a semi-supervised task limiting the labeling effort and publish a dataset, MAD, containing 105 labeled and 4,440 unlabeled high resolution-images with 14,667 manually annotated apples that facilitates the new formulation.

- We propose a novel apple detection system, S^3AD, that utilizes semi-supervised learning, contextual attention, and selective tiling to address the limited amount of labeled data and the small apple size.

- We validate our approach with an extensive evaluation of S^3AD on MAD and the MSU dataset, comparing it to strong fully-supervised small object detection systems and assessing the effect of three apple properties.

2. Related Work

This section briefly reviews related works on crop detection, crop detection datasets, and small object detection.

Crop Detection The detection of crops in precision agriculture is mainly based on variations of standard object detectors. For instance, [14] modify YOLOv3 for detection of the tomatoes, while [46] adapt SSD. [22] utilize Faster R-CNN and add image stitching as well as tiling steps to process rows of plants. For mango detection, [12] modify YOLOv2 and [29] propose a modified YOLOv4.

For apple detection, [3] and [6] employ Faster R-CNN with standard tiling and an occlusion-aware detection module, respectively. Moving to YOLO, [37] propose a YOLOv3 variation with a DenseNet backbone to detect apples in different growth stages. In [13], the performance of YOLOv3 is enhanced by pre- and post-processing steps. Recently, the authors of [11] enhanced YOLOv4 with non-local feature-level attention, and a convolution block attention module to detect apples in low-resolution images.

In contrast, we address apple detection in a semi-supervised manner, and focus on detecting small apples by introducing contextual attention and selective tiling.

Crop Detection Datasets Datasets for crop detection exist for various crops [21]. However, most crop detection datasets are limited in size. According to [8], the largest apple detection dataset contains 1,404 labeled images [34] with 7,065 annotated instances on low-resolution images. This also applies to the apple detection dataset by [3], containing 841 images with 5,765 labeled instances. The MinneApple dataset [8] contains more labeled instances, with 41,325 labeled instances in 1,001 images of moderate resolution (1280×720). Recently, the MSU Apple Dataset V2 [6] containing 14,518 annotated apples in 900 closeup images of apple tree crowns was proposed.

In contrast to these datasets, our proposed dataset MAD is well-suited for semi-supervised apple detection containing labeled and unlabeled images. Moreover, the dataset is larger than existing ones with 4,545 high-resolution images.

Small Object Detection The problem of small object detection has been addressed with different strategies. See [5] for an extensive survey. A popular strategy is the use of multiscale or scale-specific features to improve the representation of small objects. While [18] learn scale-specific features in a bottom-up and top-down manner, [19] recombine these features for an improved multiscale representation. [31] and [32] improve the training strategy of object detectors by reducing the intra-scale noise and move from a feature pyramid to an image pyramid. [23] improve the approach's efficiency with a coarse-to-fine detection. Improving the expressiveness of scale-specific features, [17] adapt the receptive field for different scales of objects. Recently, [42] employed a coarse-to-fine query-based detection mechanism on successively higher-resolution feature maps.

Table 1. Statistics of MAD. The columns show the number of images and the number of labeled instances per dataset split.

Split	Images	Labeled Instances
train	66	10089
val	12	1288
test	27	3290
unlabeled	4440	n/a

Figure 2. An example image (left) and the corresponding ground truth (right) from the test split of our dataset MAD.

Besides improved feature representation, several works utilize tiling [25, 39, 43] or super-resolution [2, 16, 24] to improve the spatial resolution of features. Another line of work uses attention mechanisms to highlight features or locations of small objects [20, 44, 45]. Our apple detection system S³AD is most related to the multiscale and tiling approaches. However, it is explicitly designed to address the detection of apples in a semi-supervised framework, e.g., utilizing domain knowledge to learn contextual attention.

3. Data Acquisition & Dataset

To facilitate our novel semi-supervised formulation of the apple detection task, we propose the Monastery Apple Dataset (MAD). The data was acquired in collaboration with a monastery in Bad Oldesloe, Germany. We collected video data from 16 trees of the monastery's apple orchard with a DJI-Mini 3 Pro drone that has a resolution of 4k (3840×2160). The data collection was conducted over the course of one month in September 2022 and during different lighting conditions to ensure diversity.

To generate the training, validation, and test splits, we first assigned the videos of 6 (training), 2 (validation), and 4 (test) trees to the respective splits. From the videos, we manually selected images for annotation to maximize data diversity. The training, validation, and test splits consist of 66, 12, and 27 frames with 10,089, 1,288, and 3,290 annotated instances as shown in Tab. 1. The apples were manually annotated with bounding boxes. An example image with ground truth from our dataset is shown in Fig. 2. Moreover, we automatically assigned three properties to each annotated apple representing the relative size, the level of occlusion, and the lighting conditions (see supp. mat. for details and statistics). This enables us, to evaluate the influence of such conditions on our system. The 4,440 un-labeled images in the training split are composed of frames from the 6 training trees that were not annotated and frames from the four remaining videos/trees that were not assigned to the other splits.

4. Method

This section introduces our novel **S**emi-**S**upervised **S**mall **A**pple **D**etection system S³AD that specifically addresses the challenges of small relative apple size and a limited amount of annotated data. S³AD consists of four steps, as shown in Fig. 3. Given an input image, our new TreeAttention module (see Sec. 4.1) utilizes contextual information to highlight the most promising image regions for apple detection. These regions are subsequently extracted by our selective tiling module, explained in Sec. 4.2, and cropped into a set of overlapping tiles. Tiling high-resolution images is advantageous since it increases the relative size of the apples and allows the object detector to process the image without initial down-sampling. Next, we train a semi-supervised Faster R-CNN object detector [27] with FPN backbone [18] in the Soft Teacher framework (see Sec. 4.3) [41]. The FPN backbone improves the detection of small apples, while the semi-supervised training in the Soft Teacher framework utilizes the unlabeled data. Finally, we merge the per-tile results in a filtering and reconstruction step, detailed in Sec. 4.4.

4.1. Tree Attention

In apple detection, we seek to detect the apples in the tree crown. Hence, given an image, S³AD only needs to search the area of the tree crown, which substantially reduces the search space. Generally, this is known as contextual information and frequently used in small object detection [4]. Therefore, we propose TreeAttention as contextual attention module in S³AD to focus processing on the most prominent tree crowns in the input image. We will use this attention in Sec. 4.2, to guide S³AD's selective tiling.

Our TreeAttention module is based on a simple residual U-Net architecture [28] comprising three stages with one residual block each in the encoder and decoder, respectively. The final result of our TreeAttention module is an attention map with values between 1 (high attention) and 0 (low attention) indicating the rough locations of apples. Since no precise localization is necessary in this step, we apply TreeAttention to the input images at a lower resolution for computational efficiency. For training TreeAttention, we use binary cross entropy loss and ground truth derived from the bounding box annotations of the respective dataset. Since the goal of TreeAttention is to highlight the tree crown, we compute the alpha shape with $\alpha = 100$ around the annotations of all apples in an image as ground truth. The alpha shape is a generalization of the convex hull and results in a tight-fitting hull around the bounding
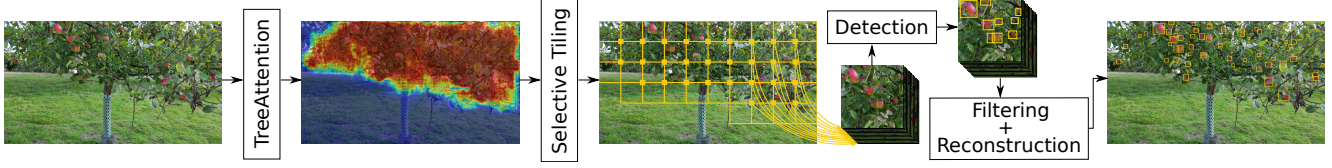
Figure 3. System figure of our proposed semi-supervised small apple detection approach S^{3}AD. First, an attention map is generated with our TreeAttention module. The attention map is used by our selective tiling module to crop the most promising image regions into a set of overlapping tiles, which are subsequently processed by a semi-supervised Faster R-CNN object detector. During filtering and reconstruction, the per-tile results are merged.

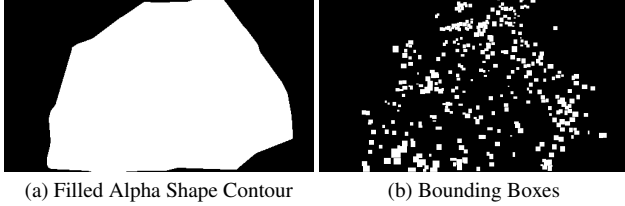(a) Filled Alpha Shape Contour   (b) Bounding Boxes

Figure 4. Example of the filled alpha shape (a) that is used to train TreeAttention and the corresponding bounding boxes (b).

boxes. An example mask, which is used in the training of TreeAttention, is shown in Fig. 4a with the corresponding bounding boxes depicted in Fig. 4b.

4.2. Selective Tiling

To improve the detection of the small apples in high-resolution images, S^{3}AD follows a tiling approach [15, 30, 40] and increases the relative size of the small apples for detection by extracting tiles. Subsequently, the object detector in S^{3}AD processes each tile in full resolution. Hence, our object detector can utilize all information available in the tiles and the small apples are less affected by downsampling in CNNs. Since standard tiling substantially increases the runtime, we utilize the attention maps generated by our TreeAttention module (see Sec. 4.1) to select which tiles are processed by S^{3}AD's detector.

Deriving the relevant tiles from an attention map, we first binarize the attention map with a threshold of $\tau = 0.3$ and up-sample it to the original image size. Next, a 800×800 sliding window with a stride of 400 pixels is moved over the binarized attention map. For each window, the amount of pixels with an attention value greater than τ is determined, and the tile is selected for further processing in S^{3}AD if the amount is larger than 20%. The overlap between tiles ensures that each apple is fully contained in at least one tile. Compared to processing all tiles, this approach results in a significant reduction of processed tiles and inference time of S^{3}AD, without losing detection performance (see Sec. 5.3).

4.3. Semi-supervised Detection

S^{3}AD can be used together with an arbitrary object detector. Due to the success in previous crop detection applications [3, 8, 22], we choose Faster R-CNN [27] with a Feature Pyramid Network (FPN) backbone [18]. Faster R-CNN is a two-stage object detector that first generates class-agnostic object proposals, while later classifying and refining the proposals. In S^{3}AD, the classification is binary (apple vs. background). The FPN backbone has been shown to improve detection results of Faster R-CNN on small objects [18], which fits the task of apple detection. As outlined above, we apply the detector to each selected tile, to improve the detection of small apples. Subsequently, the per-tile results are merged, as described in Sec. 4.4.

To utilize the large amount of unlabeled data in our proposed dataset MAD, we train S^{3}AD's Faster R-CNN detector in the semi-supervised Soft Teacher framework [41]. Soft Teacher utilizes two models for semi-supervised training: a teacher model and a student model. The teacher model learns to generate boxes from unlabeled data, while the student model is trained using a combination of labeled data and unlabeled data with boxes provided by the teacher model as pseudo ground truth. In this work, we use Faster R-CNN as our student model. Throughout the training, the teacher model is updated using the exponential moving average of the student model. Since the teacher generates thousands of box candidates, non-maximum suppression is applied to eliminate redundancy. Moreover, only candidates surpassing a given confidence score are considered as pseudo ground truth, aiming to reduce the number of false positives. We apply a confidence score threshold of 0.9 for the pseudo annotations.

4.4. Filtering & Reconstruction

After applying the semi-supervised Faster R-CNN detector on each selected tile from an image, S^{3}AD reconstructs the result for the entire image by filtering and merging the per-tile results and applying an image-level filtering step. First, S^{3}AD filters the per-tile results to remove detections for partially visible apples along the tile borders that are fully visible in another tile. Such partial apple detec-

tions can cause false positives and removal after merging the per-tile results based on an IoU threshold is difficult, since detections of partial apples will naturally not have a high overlap with detections for the entire apple. To solve this problem, S^3AD removes the detections along a 100-pixel wide border around each tile prior to reconstructing the full image-level results. Note that we do not remove detections along the image edges.

After the per-tile filtering step, S^3AD merges the individual detection results and subsequently applies non-maximum suppression with an IoU threshold of 0.5.

4.5. Training

The semi-supervised Faster R-CNN detector in S^3AD is trained in the Soft Teacher framework and is initialized with the semi-supervised MS COCO weights. For training on MAD, we use tiles from all labeled training images and 41,415 unlabeled image tiles selected using TreeAttention (see supp. mat. for details on selection and training of TreeAttention). Note that TreeAttention is trained as an individual system prior to training the actual detector. During training, the image tiles are randomly resized in a range between 600 and 1600, and grouped in batches of size 10. We use a data sampling ratio of 0.2 in the Soft Teacher framework. Hence, four unlabeled images are randomly selected for each labeled tile in a batch. The model learns to minimize a joint loss consisting of a supervised and a weighted unsupervised loss term (for more information, see [41]) and is trained for 100k steps using SGD with an initial learning rate, momentum, and weight decay of 0.001, 0.9, and 0.0001, respectively. Additionally, the learning rate is divided by 10 at 60k and 80k steps and the data sampling ratio is gradually decreased to 0 in the last 5,500 steps. To ensure high-quality pseudo annotations, the foreground threshold for using the teacher's pseudo ground truth is set to 0.9.

5. Evaluation

To validate the strength of S^3AD, we conduct an evaluation on our introduced dataset MAD. As outlined in Sec. 3, MAD covers complex orchard environments with apples of small relative size and only a limited amount of annotated training images. Additionally, we assess the generality of our approach w.r.t. other apple detection datasets by training and testing on the MSU dataset [6]. All results in this chapter are generated on the respective test splits.

On MAD, we compare our S^3AD to five strong fully-supervised baselines that specifically address the detection of small objects: Faster R-CNN [27] with FPN backbone [18], which is frequently used in crop detection approaches [3, 8, 22], PANet [19], SNIPER [32], AutoFocus [23], and Deformable DETR [47], a transformer-based object detector. A direct comparison to other crop detection approaches is impossible, due to the lack of publicly avail-

able implementations. For training the fully-supervised baselines, we only utilize the 66 annotated training images of MAD, containing 10,089 annotated apples. On MSU, we compare to the base system of S^3AD, which is Faster R-CNN+FPN. A comparison to O2RNet [6] proposed by the authors of MSU is not possible due to the lack of an available implementation. To match the semi-supervised setup of MAD, we train S^3AD and Faster R-CNN+FPN with annotations on only 10% of the training image, resulting in 93 annotated training images. Note that in MAD, only 1.5% of the training images are labeled.

To assess the quality of the apple detection results, we utilize the commonly used measures Average Recall (AR) and Average Precision (AP) based on 100 detections. While AR evaluates how many apples are found and how precisely they are located, AP also considers the false positives. Both metrics are regularly used for evaluating object detection [26, 27] and crop detection approaches [8, 12, 34, 40]. Supplementing these quantitative results, we also provide qualitative results.

5.1. Results on MAD

5.1.1 Overall Quantitative Results

Table 2 presents the quantitative results of S^3AD with and without selective tiling and the five fully-supervised baselines. From the results, it is clearly visible that we outperform all baselines. Compared to Faster R-CNN+FPN, the base system of S^3AD, the improvement is 10.9% in terms of AP by utilizing unlabeled data and even 14.9% by additionally introducing selective tiling. Compared to the other dedicated small object detection systems PANet and SNIPER, the improvement is up to 11.9% in terms of AP. S^3AD even outperforms the transformer-based system Deformable DETR by 9.9%. Only AutoFocus is able to stay competitive with S^3AD. Here the improvement is only 1.4% in terms of AP, but 11.5% in terms of AR, highlighting the strong recall of S^3AD. Moreover, AutoFocus is trained with a ResNet-101 backbone, while S^3AD is trained with a smaller ResNet-50 backbone. Finally, comparing S^3AD with and without tiling indicates the positive influence of our selective tiling on the results, with an improvement of 3.7%. Overall, our results show the strong positive effect of utilizing unlabeled data in the semi-supervised training and the benefit of our selective tiling.

5.1.2 Property-based Quantitative Results

After discussing the overall quantitative results, we will examine the behavior of the previously discussed detection systems based on a ResNet-50 backbone: S^3AD with and without our selective tiling, S^3AD's base system Faster R-CNN+FPN, PANet, and Deformable DETR. We examine the behavior w.r.t. three apple properties annotated in

Table 2. Quantitative results in terms of Average Precision (AP) and Average Recall (AR) for S³AD with and without our selective tiling, as well as five object detectors. The results are generated on the test split of our proposed dataset MAD.

System	Backbone	Semi-supervised	Selective Tiling	AP↑	AR↑
SNIPER [32]	ResNet-101			0.387	0.419
AutoFocus [23]	ResNet-101			0.417	0.419
Faster R-CNN+FPN [18,27]	ResNet-50			0.368	0.408
PANet [19]	ResNet-50			0.378	0.418
Deformable DETR [47]	ResNet-50			0.385	0.438
S³AD w/o tiling (ours)	ResNet-50	✓		0.408	0.448
S³AD (ours)	ResNet-50	✓	✓	**0.423**	**0.467**

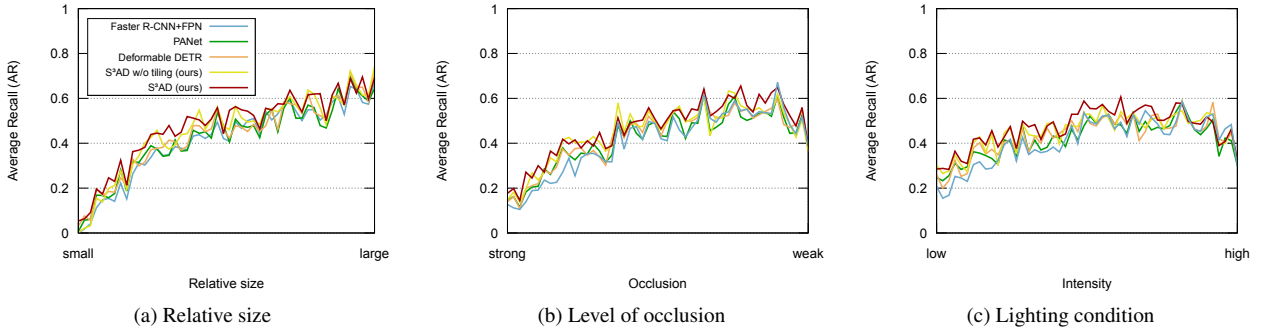(a) Relative size   (b) Level of occlusion   (c) Lighting condition

Figure 5. Apple detection results of S³AD with and without tiling, Faster R-CNN+FPN, PANet, and Deformable DETR in terms of AR for property-specific ranges. Each point on the curves represents a bin of 2% of the annotated apples in the test split of our dataset MAD.

MAD: relative size, level of occlusion, and lighting conditions. The AR-based results for different levels of each property are given in Fig. 5, where each data point represents 2% of the annotated apples on property-specific scales. We chose AR here, to show how many apples featuring the respective property are found without penalizing duplicates.

Analyzing the results w.r.t. the relative size of objects, it is clearly visible that S³AD almost consistently outperforms all other systems. Especially on small objects (left part in Fig. 5a), there is a stronger improvement over the other systems. Analyzing this improvement in more detail, we quantify the average improvement across the first third of the data points/bins (small objects) and the last third (large objects). S³AD with our tiling outperforms S³AD without tiling by 15% on small objects, while the improvement on large objects is only 1.7%. Similarly, the improvement over Faster R-CNN+FPN on small objects is 35.3%, while being only 7.4% on large objects. Despite the improvements, all system's results degrade with decreasing relative object size.

On the other two annotated properties, level of occlusion and lighting conditions, the systems behave similarly to each other. While S³AD almost consistently outperforms the other systems, all systems struggle with dark or bright

apples as well as strongly occluded apples similarly. While the results of S³AD degrade from weakly occluded apples (right third in Fig. 5b) to strongly occluded apples (left third in Fig. 5b) by 42.3%, the difference is 45.4% for PANet and 46.7% for Deformable DETR. Similarly, moving from well-illuminated apples (central third in Fig. 5c) to dark apples (left third in Fig. 5c), the drop in performance is 24.4% for S³AD and 26.5% for PANet.

Overall, this detailed evaluation highlights the positive influence of our proposed selective tiling in S³AD on the detection of small apples and the generally improved performance of S³AD on small apples compared to other systems. Moreover, from all three properties, we can derive and quantify challenging conditions for apple detection systems: small apples, strongly occluded apples, and dark or bright apples.

5.1.3 Qualitative Results

Besides the quantitative results, we also present qualitative results in Fig. 6 and Fig. 7. The results in Fig. 6 indicate that S³AD is able to improve the detection of small apples compared to Faster R-CNN+FPN and AutoFocus. For instance, in the upper example in Fig. 6 S³AD detects the small apple in the top left corner that Faster R-CNN+FPN
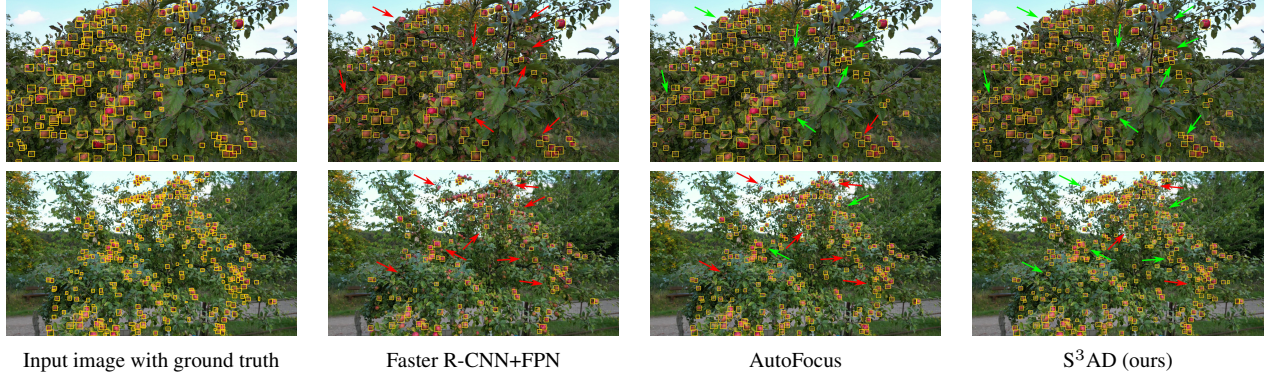
| Input image with ground truth | Faster R-CNN+FPN | AutoFocus | S³AD (ours) |

Figure 6. Qualitative results of S³AD, Faster R-CNN+FPN, and AutoFocus on the test split of our dataset MAD. Red arrows denote missed apples, while green arrows denote the detection of such apples by other systems.

misses. Moreover, several very small apples that are partially occluded are better detected by S³AD compared to Faster R-CNN+FPN and AutoFocus (see arrows). This also confirms the findings from Sec. 5.1.2 about the strong performance of S³AD on small apples and the generally better results of S³AD across all levels of occlusions. The lower example in Fig. 6 also supports these findings. However, in this example, it is also clearly visible that even S³AD still misses a few small apples, mainly due to occlusion (central red arrow) or insufficient lighting conditions (bottom right arrow). All three properties were also identified as major challenges for apple detection in Sec. 5.1.2.

Focusing on the influence of the selective tiling, Fig. 7 shows the results of S³AD with and without the selective tiling. As expected from the quantitative results, several small apples are missed without the selective tiling denoted by the red arrows. Multiple such examples are visible in the upper part and the right part of the lower example. For instance, the four small apples in the bottom right of the lower example, marked by an arrow, are all missed without our selective tiling. Across both examples, several small apples are missed due to strong occlusion. This is also slightly improved by applying the selective tiling, as the results in the top right of the upper example and the quantitative results in Fig. 5b indicate. Still, several strongly occluded apples are still missed despite selective tiling.

Overall, the qualitative results show the strong performance of our system and the substantial improvement using our selective tiling approach for the detection of small apples and to some extent also to strongly occluded apples.

5.2. Results on MSU

The results in the MSU dataset for our S³AD and its base system Faster R-CNN+FPN are presented in Tab. 3. Generally, the tendency of the results is similar to MAD, with S³AD outperforming Faster R-CNN+FPN. However, the margin between the methods is much larger on MSU
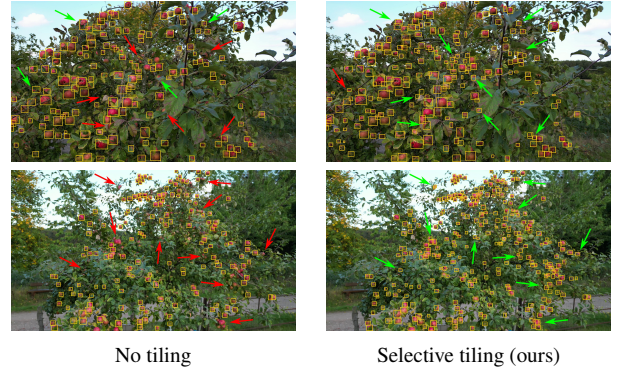

| No tiling | Selective tiling (ours) |

Figure 7. Qualitative results of S³AD with and without using selective tiling to improve the detection of small apples on the test split of our MAD dataset. See Fig. 6 for the input images with ground truth. Red arrows denote missed apples, while green arrows denote the detection of such apples by other systems.

compared to MAD. While the improvement in terms of AP from Faster R-CNN+FPN to S³AD was 14.9% on MAD, it is now up to 50.5%. To examine this effect in more detail, we also compare to S³AD without tiling to disentangle the effect of our selective tiling and the semi-supervised training. S³AD without tiling already outperforms Faster R-CNN+FPN by 44.2% in terms of AP, indicating the strong influence of semi-supervised learning in the absence of a large amount of annotated images. Still, the proposed selective tiling is effective on the MSU dataset as well, leading to another 4.4% improvement when added to S³AD, which is similar to the results on MAD. Comparing S³AD with selective tiling to a version of Faster R-CNN+FPN trained on the entire annotated training dataset of MSU (842 annotated images), reveals that despite using only around 11% of the annotated data, the result of S³AD reaches 90% of the fully supervised Faster R-CNN results (0.518 vs. 0.579). This highlights again the strength of the proposed semi-

Table 3. Quantitative results in terms of Average Precision (AP) and Average Recall (AR) for S^3AD with and without tiling, as well as Faster R-CNN+FPN on the test split of MSU.

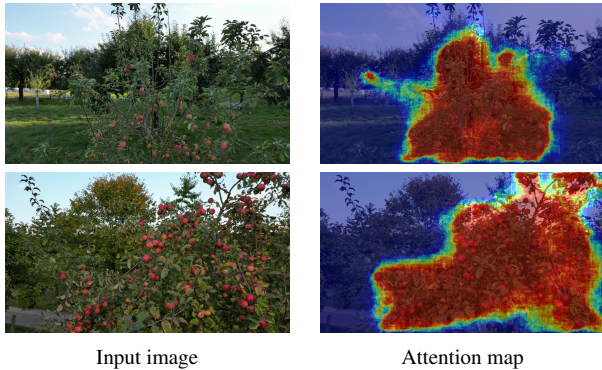| System	AP↑	AR↑
Faster R-CNN+FPN [18, 27]	0.344	0.439
S^3AD w/o tiling (ours)	0.496	0.574
S^3AD (ours)	**0.518**	**0.594**

Input image   Attention map

Figure 8. Attention maps generated by our TreeAttention module that focuses processing on the foreground tree crown. Red areas indicate high attention, while blue areas indicate low attention.

supervised training in S^3AD. Overall, the results on MSU show that S^3AD generalizes well to other apple detection datasets, while again displaying the strength of the semi-supervised training.

5.3. Analysis of Selective Tiling Approach

After presenting the main results of our system, we further analyze our selective tiling approach, which improves the detection of small apples, as shown in Fig. 5a.

5.3.1 Quality of TreeAttention

First, we present the results of our learned TreeAttention, which aims to select all relevant tiles of the foreground tree crown using contextual information. Overall, TreeAttention recalls 99.71% of the labeled apples, while processing only 62.47% of all tiles. The strong recall and the good precision are also well visible from the qualitative results in Fig. 8 depicting results of TreeAttention encoded as heat maps. While both tree crowns are completely covered (overlaid with red), most of the background is pruned (overlaid with blue). Hence, our TreeAttention is a high-quality pre-processing mechanism to effectively focus the processing, while preserving strong overall detection performance.

Table 4. Comparison of standard tiling and the proposed selective tiling in S^3AD in terms of AP and runtime.

Tiling approach	AP↑	Runtime[s]↓
Standard	0.423	1.98
Selective (ours)	0.423	1.51

5.3.2 Comparison of Tiling Approaches

Given the high-quality attention maps generated by TreeAttention, we are able to reduce the number of processed tiles substantially, as shown previously. Now, we compare our selective tiling with the standard tiling that extracts all considered tiles. Table 4 shows the result of this comparison for our proposed system in terms of AP and runtime. While the apple detection result in terms of AP is almost constant between the two tiling strategies, the runtime is reduced by 24.2% with our selective tiling approach. Therefore, selective tiling combines the benefits of standard tiling, improving the detection of small objects (see Sec. 5.1.2), with only a moderate increase in runtime.

6. Conclusion

In this paper, we reformulated the problem of apple detection in orchard environments as a semi-supervised detection task to improve the results in the absence of large-scale annotated training data. To this end, we collected MAD, a new apple detection dataset comprising a total of 4,545 high-resolution images and 14,667 annotated apples, including a large share of unlabeled data for semi-supervised training. Taking advantage of MAD, we proposed S^3AD, a novel semi-supervised apple detection system based on contextual attention, an efficient selective tiling approach, and a Faster R-CNN detector trained in a Soft Teacher framework. While the tiling approach improves the challenging detection of small apples utilizing the full resolution of the input images, S^3AD uses contextual attention to selectively process tiles and limit the additional runtime induced by tiling.

Through extensive evaluation on our novel dataset, which features complex orchard environments, we showed substantial improvements of S^3AD compared to strong baselines (up to 14.9%). Further results on the MSU dataset confirmed these improvements. By analyzing the results of S^3AD in more detail, we showed that especially on small objects, the selective tiling approach based on contextual information improved the performance, while limiting the additional runtime and maintaining a strong apple detection performance. Overall, this paper presents a new reformulation of the apple detection task and a novel, high-quality semi-supervised apple detection approach, addressing our new semi-supervised apple detection task.

Acknowledgments

References

[1] Peri Akiva, Kristin Dana, Peter Oudemans, and Michael Mars. Finding berries: Segmentation and counting of cranberries using point supervision and shape priors. In *Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 1

[2] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. SOD-MTGAN: Small object detection via multi-task generative adversarial network. In *European Conference on Computer Vision*, 2018. 3

[3] Suchet Bargoti and James Underwood. Deep fruit detection in orchards. In *International Conference on Robotics and Automation*, 2017. 1, 2, 4, 5

[4] Guang Chen, Haitao Wang, Kai Chen, Zhijun Li, Zida Song, Yinlong Liu, Wenkai Chen, and Alois Knoll. A survey of the four pillars for small object detection: Multiscale representation, contextual information, super-resolution, and region proposal. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2020. 1, 3

[5] Gong Cheng, Xiang Yuan, Xiwen Yao, Kebing Yan, Qinghua Zeng, Xingxing Xie, and Junwei Han. Towards large-scale small object detection: Survey and benchmarks. *Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2

[6] Pengyu Chu, Zhaojian Li, Kaixiang Zhang, Dong Chen, Kyle Lammers, and Renfu Lu. O2RNet: Occluder-occludee relational network for robust apple detection in clustered orchard environments. *Smart Agricultural Technology*, 5, 2023. 2, 5

[7] Nicolai Häni, Pravakar Roy, and Volkan Isler. Apple counting using convolutional neural networks. In *International Conference on Intelligent Robots and Systems*, 2018. 1

[8] Nicolai Häni, Pravakar Roy, and Volkan Isler. MinneApple: a benchmark dataset for apple detection and segmentation. *IEEE Robotics and Automation Letters*, 2020. 1, 2, 4, 5

[9] Yanbin Hua, Nairu Zhang, Xin Yuan, Lichun Quan, Jiangang Yang, Ken Nagasaka, and Xin-Gen Zhou. Recent advances in intelligent automated fruit harvesting robots. *The Open Agriculture Journal*, 2019. 1

[10] Jisoo Jeong, Seungeui Lee, Jeesoo Kim, and Nojun Kwak. Consistency-based semi-supervised learning for object detection. *Advances in Neural Information Processing Systems*, 2019. 1

[11] Mei Jiang, Lei Song, Yunfei Wang, Zhenyu Li, and Huaibo Song. Fusion of the YOLOv4 network model and visual attention mechanism to detect low-quality young apples in a complex environment. *Precision Agriculture*, 2022. 2

[12] Anand Koirala, Keryy B Walsh, Zhenglin Wang, and Cheryl McCarthy. Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'MangoYOLO'. *Precision Agriculture*, 2019. 1, 2, 5

[13] Anna Kuznetsova, Tatiana Maleva, and Vladimir Soloviev. Using YOLOv3 algorithm with pre-and post-processing for apple detection in fruit-harvesting robot. *Agronomy*, 2020. 2

[14] Mubashiru Olarewaju Lawal. Tomato detection based on modified YOLOv3 framework. *Scientific Reports*, 2021. 2

[15] Changlin Li, Taojiannan Yang, Sijie Zhu, Chen Chen, and Shanyue Guan. Density map guided object detection in aerial images. In *Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 4

[16] Jianan Li, Xiaodan Liang, Yunchao Wei, Tingfa Xu, Jiashi Feng, and Shuicheng Yan. Perceptual generative adversarial networks for small object detection. In *Computer Vision and Pattern Recognition*, 2017. 3

[17] Yanghao Li, Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection. In *International Conference on Computer Vision*, 2019. 2

[18] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Computer Vision and Pattern Recognition*, 2017. 1, 2, 3, 4, 5, 6, 8

[19] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Computer Vision and Pattern Recognition*, 2018. 2, 5, 6

[20] Xiaocong Lu, Jian Ji, Zhiqi Xing, and Qiguang Miao. Attention and feature fusion SSD for remote sensing object detection. *Transactions on Instrumentation and Measurement*, 70, 2021. 2, 3

[21] Yuzhen Lu and Sierra Young. A survey of public datasets for computer vision tasks in precision agriculture. *Computers and Electronics in Agriculture*, 2020. 2

[22] Yue Mu, Tai-Shen Chen, Seishi Ninomiya, and Wei Guo. Intact detection of highly occluded immature tomatoes on plants using deep learning techniques. *Sensors*, 2020. 2, 4, 5

[23] Mahyar Najibi, Bharat Singh, and Larry S Davis. AutoFocus: Efficient multi-scale inference. In *International Conference on Computer Vision*, 2019. 2, 5, 6

[24] Junhyug Noh, Wonho Bae, Wonhee Lee, Jinhwan Seo, and Gunhee Kim. Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection. In *International Conference on Computer Vision*, 2019. 3

[25] F Ozge Unel, Burak O Ozkalayci, and Cevahir Cigla. The power of tiling for small object detection. In *Computer Vision and Pattern Recognition Workshops*, 2019. 2, 3

[26] Joseph Redmon and Ali Farhadi. YOLOv3 An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 5

[27] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 2015. 1, 3, 4, 5, 6, 8

[28] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015. 3

[29] Arunabha M Roy and Jayabrata Bhaduri. Real-time growth stage detection model for high degree of occultation using DenseNet-fused YOLOv4. *Computers and Electronics in Agriculture*, 2022. 2

[30] Vít Růžička and Franz Franchetti. Fast and accurate object detection in high resolution 4k and 8k video using GPUs. In *High Performance Extreme Computing Conference*, 2018. 4

[31] Bharat Singh and Larry S Davis. An analysis of scale invariance in object detection - SNIP. In *Computer Vision and Pattern Recognition*, 2018. 2

[32] Bharat Singh, Mahyar Najibi, and Larry S Davis. SNIPER: Efficient multi-scale training. In *Advances in Neural Information Processing Systems*, 2018. 2, 5, 6

[33] Kihyuk Sohn, Zizhao Zhang, Chun-Liang Li, Han Zhang, Chen-Yu Lee, and Tomas Pfister. A simple semi-supervised learning framework for object detection. *arXiv preprint arXiv:2005.04757*, 2020. 1

[34] Madeleine Stein, Suchet Bargoti, and James Underwood. Image based mango fruit detection, localisation and yield estimation using multiple view geometry. *Sensors*, 2016. 1, 2, 5

[35] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International Conference on Artificial Neural Networks*, 2018. 1

[36] Peng Tang, Chetan Ramaiah, Yan Wang, Ran Xu, and Caiming Xiong. Proposal learning for semi-supervised object detection. In *Winter Conference on Applications of Computer Vision*, 2021. 1

[37] Yunong Tian, Guodong Yang, Zhe Wang, Hao Wang, En Li, and Zize Liang. Apple detection during different growth stages in orchards using the improved YOLO-V3 model. *Computers and Electronics in Agriculture*, 2019. 1, 2

[38] Qi Wang, Stephen Nuske, Marcel Bergerman, and Sanjiv Singh. Automated crop yield estimation for apple orchards. In *Experimental Robotics: The 13th International Symposium on Experimental Robotics*, 2013. 1

[39] Christian Wilms, Rafael Heid, Mohammad Araf Sadeghi, Andreas Ribbrock, and Simone Frintrop. Which airline is this? airline logo detection in real-world weather conditions. In *International Conference on Pattern Recognition*, 2021. 2, 3

[40] Christian Wilms, Robert Johanson, and Simone Frintrop. Localizing small apples in complex apple orchard environments. *arXiv preprint arXiv:2202.11372*, 2022. 4, 5

[41] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *International Conference on Computer Vision*, 2021. 1, 2, 3, 4, 5

[42] Chenhongyi Yang, Zehao Huang, and Naiyan Wang. Query-Det: Cascaded sparse query for accelerating high-resolution small object detection. In *Computer Vision and Pattern Recognition*, 2022. 2

[43] Fan Yang, Heng Fan, Peng Chu, Erik Blasch, and Haibin Ling. Clustered object detection in aerial images. In *International Conference on Computer Vision*, 2019. 2, 3

[44] Xue Yang, Junchi Yan, Wenlong Liao, Xiaokang Yang, Jin Tang, and Tao He. SCRDet++: Detecting small, cluttered and rotated objects via instance-level feature denoising and rotation loss smoothing. *Transactions on Pattern Analysis and Machine Intelligence*, 45(2), 2022. 2, 3

[45] Xue Yang, Jirui Yang, Junchi Yan, Yue Zhang, Tengfei Zhang, Zhi Guo, Xian Sun, and Kun Fu. SCRDet: Towards more robust detection for small, cluttered and rotated objects. In *International Conference on Computer Vision*, 2019. 2, 3

[46] Ting Yuan, Lin Lv, Fan Zhang, Jun Fu, Jin Gao, Junxiong Zhang, Wei Li, Chunlong Zhang, and Wenqiang Zhang. Robust cherry tomatoes detection algorithm in greenhouse scene based on SSD. *Agriculture*, 2020. 2

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*. 5, 6