

6 Supplementary information

This supplementary section provides additional examples related to the annotation strategy used in the dataset generation process.

6.1 Supplementary Note S1. Occlusion in Fruit Annotation

Protocol

This supplementary note complements the complete annotation protocol provided in the associated INRAE data repository (DOI: 10.57745/NORPF1) by detailing the specific rules used to handle fruit occlusion. Occlusion is a key source of complexity in the proposed dataset and represents one of the main factors contributing to annotation ambiguity in fruit detection. In dense orchard canopies, fruits are frequently partially hidden by leaves, branches, or neighboring fruits, making both their presence and spatial extent difficult to determine. As visually incomplete fruit instances must be distinguished from background structures and surrounding vegetation, the dataset is particularly challenging for object detection models. The examples shown in Supplementary Figure 12 illustrate the annotation strategy adopted for partially occluded fruits. Fruits were annotated only when sufficient visual evidence remained available to reliably identify the object and delineate its bounding box consistently. Severely occluded or visually ambiguous fruits were excluded to avoid introducing uncertain ground-truth annotations.

6.2 Supplementary Note S2. RT-DETR baseline architecture

RT-DETR combines a convolutional neural network backbone with an efficient hybrid encoder that refines the backbone features through attention-based intra-scale feature interaction and CNN-based cross-scale feature fusion. This design captures multi-scale visual information and global context while maintaining real-time inference capability. Within the CNN-based cross-scale feature-fusion module, the fused features are

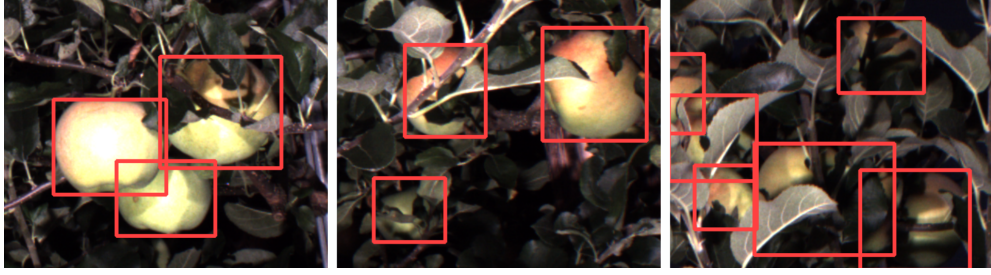


Fig. 12 Representative examples of ground-truth annotations under fruit occlusion conditions. These examples highlight the challenging nature of the dataset, where fruits may be partially hidden by leaves, branches, or neighboring fruits. Partially occluded fruits were annotated when a visible surface region and a distinguishable portion of the rounded fruit shape remained identifiable, whereas severely occluded or visually ambiguous fruits were omitted from annotation.

refined using RepBlocks. Each RepBlock combines parallel 3×3 and 1×1 convolution–batch-normalization branches, whose outputs are summed and passed through a SiLU activation. During inference, these branches can be reparameterized into a single 3×3 convolution, thereby simplifying the computational graph. The Transformer decoder processes a fixed set of object queries, each representing a potential object instance, and predicts for each query a bounding box (x, y, w, h) , an object class, and a confidence score. During training, predictions are assigned to ground-truth annotations through one-to-one bipartite matching using the Hungarian algorithm. The optimization loss combines classification, L_1 bounding-box regression, and IoU-based localization terms, following the original RT-DETR implementation. An overview of the architecture is provided in Supplementary Figure 13.

6.3 Supplementary Note S3. Stage-dependent detection

performance of apple flowering organs across BBCH stages

This supplementary note provides a stage-wise analysis of flowering-stage detection performance to better understand the lower baseline results observed for floral organs in the main manuscript. We examine how BBCH progression, object scale, and false-negative patterns jointly contribute to detection difficulty under orchard conditions.

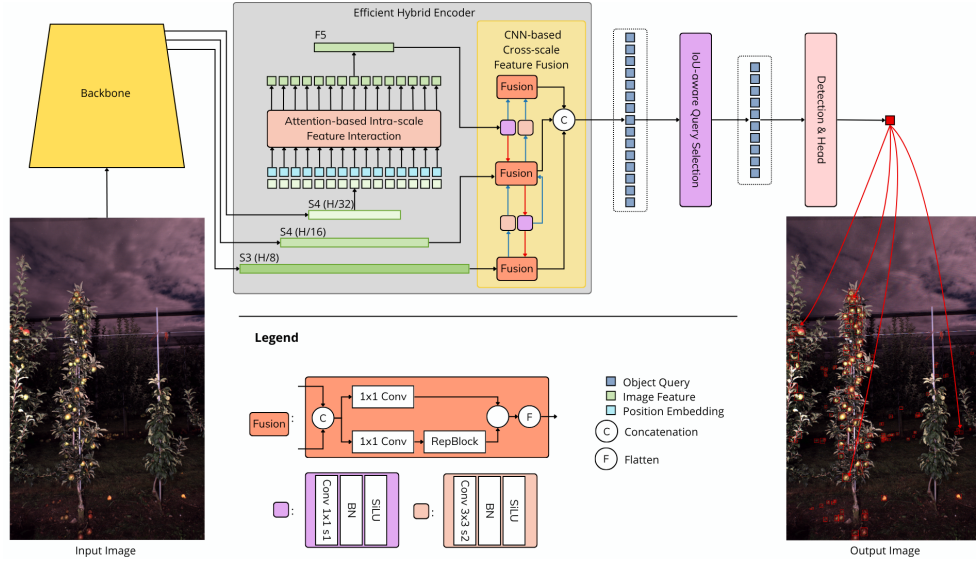


Fig. 13 Overview of the single-class RT-DETR framework from Zhao et al. (2024). The input orchard image is first processed by a backbone to extract multi-scale image features. These features are then refined by an efficient hybrid encoder, which combines attention-based intra-scale feature interaction with CNN-based cross-scale feature fusion. IoU-aware query selection is used to select relevant object queries, which are finally passed to the detection head to generate single-class bounding-box predictions.

Figure 14 shows that the F1 score follows a non-monotonic pattern across BBCH flowering stages. The similar scores observed at stages 00 and 51 reflect their close visual appearance, corresponding respectively to dormant wood buds and green wood buds. Performance increases from 53 to 56 as leaves progressively unfold and provide clearer structural cues. The drop at stage 57 may be explained by its compact morphology, corresponding to a single closed pink flower bud with limited visual information. The score then increases from 59 to 65 as the floral cluster becomes more distinctive, reaching its maximum at 65 when all flowers are open. The decrease at 67 and 69 is likely due to petal fall and visual ambiguity between late flowering stages, since 67 may appear close to 69 depending on the viewing angle.

The relationship between mean bounding box area and F1 score is presented in Figure 15. A strong positive correlation was observed between object size and detection performance, with a Pearson correlation coefficient of $r = 0.8157$. This result

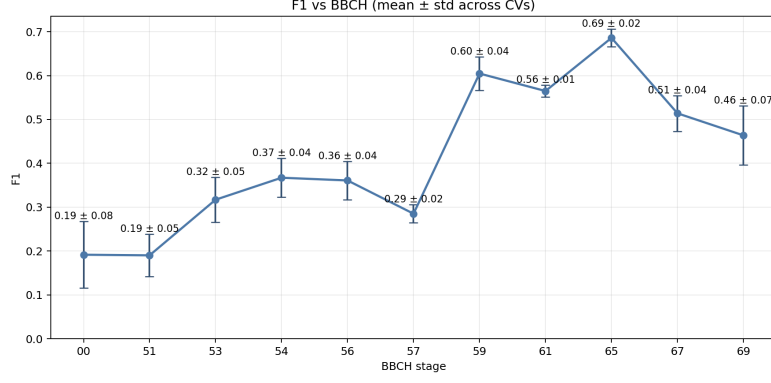


Fig. 14 Per-stage F1 score for apple floral organ detection across BBCH flowering stages on the test set. Points represent the mean F1 score across cross-validation folds, and error bars indicate the standard deviation. Overall, detection performance was lowest during the early flowering stages (00–51), increased toward intermediate-to-advanced flowering stages, reached its highest value at stage 65, and then decreased toward the end of the flowering period (67–69), although a local drop was observed at stage 57.

indicates that object scale constitutes a major factor affecting detectability in orchard conditions. Small objects from early flowering stages contain fewer discriminative pixels and are more sensitive to localization inaccuracies, image resizing, and partial occlusions. Conversely, larger organs from intermediate flowering stages provide richer visual structures that facilitate feature extraction and object localization.

Interestingly, the correlation between object size and performance is not strictly monotonic. Although stages 67 and 69 correspond to relatively large objects, their performance decreased compared with stage 65. Their lower performance suggests that detectability is not solely driven by object size. At BBCH 67, the coexistence of flowers with and without petals creates strong appearance variability, and the petal-bearing regions may be difficult to observe depending on viewpoint. At BBCH 69, petals have fallen, leaving mainly pedicels and reproductive structures. As a result, the object remains physically large, but its visually discriminative surface is reduced, providing fewer informative pixels for reliable recognition.

The false-negative analysis shown in Figure 16 reveals strong variations in detection difficulty across BBCH stages. The highest number of false negatives was observed

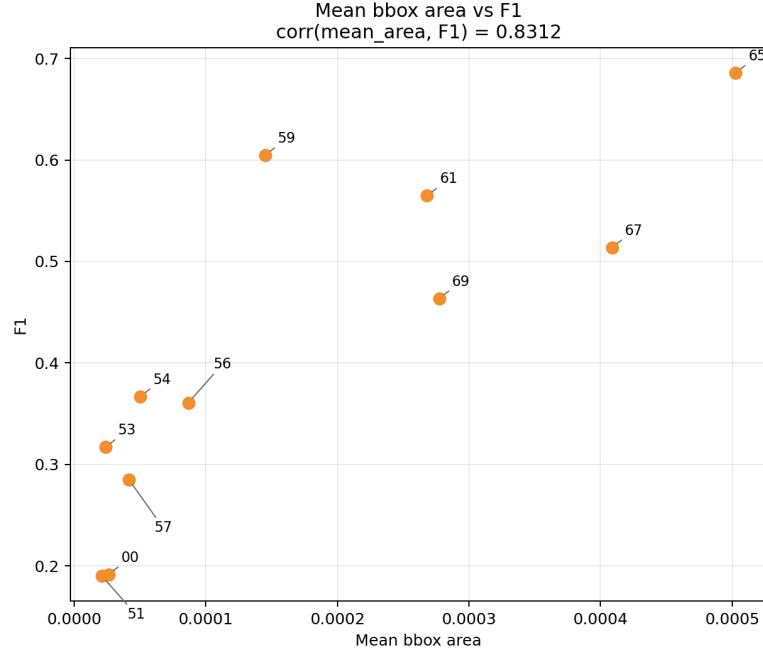


Fig. 15 Relationship between mean bounding box area and F1-score across BBCH stages. A strong positive correlation ($r = 0.8157$) was observed between object size and detection performance, highlighting the influence of organ scale on detectability.

at stage 56 (11.82 ± 10.13 false negatives per image), which corresponds to closed pink flower buds. At this stage, organs remain small, which increases localization ambiguity. Stage 00 exhibited the highest variability (9.17 ± 35.10) because stage 00 corresponds to wood buds that are difficult to distinguish when images mainly contain woody branches. From stages 59 to 69, the number of false negatives progressively decreased, suggesting that later-stage organs were less frequently missed as their physical size increased. However, this trend mainly reflects recall-related behavior and does not necessarily imply higher overall detection performance. In particular, stages 67 and 69 may still be difficult to recognize reliably because their visually discriminative surface is reduced or highly variable, which can affect precision and F1 score.

Overall, the results show that the visual heterogeneity associated with BBCH progression substantially affects single-class detection performance. This effect is most

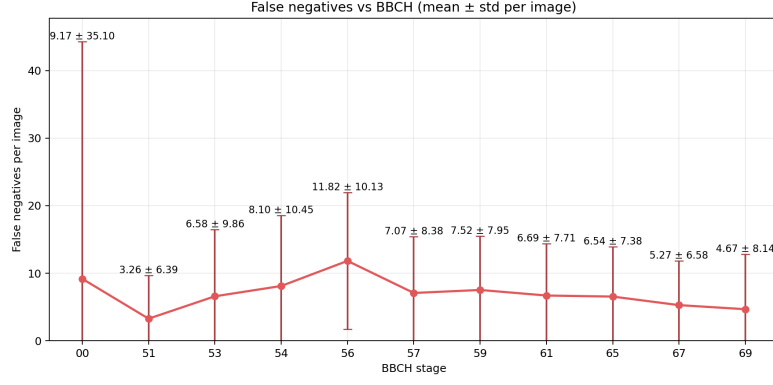


Fig. 16 Evolution of the average number of false negatives per image across BBCH stages. Early and intermediate stages exhibit higher false negative rates, while later stages show improved detection stability as organs become larger and more spatially separated. Error bars represent the standard deviation across images.

pronounced for flowering-stage detection, where small object scale, high organ density, occlusion, and strong context dependence jointly increase localization errors and false negatives. By contrast, fruitlets and mature fruits represent more object-centric targets, with more compact and homogeneous morphologies, clearer boundaries, and greater visual separability from the background, resulting in more consistent localization across acquisition conditions.