agronomy

MDPI

Article

A Grape Dataset for Instance Segmentation and Maturity Estimation

Achilleas Blekos *, Konstantinos Chatzis, Martha Kotaidou, Theocharis Chatzis, Vassilios Solachidis, Dimitrios Konstantinidis and Kosmas Dimitropoulos

Institute Information Technologies (ITI), Centre for Research and Technology Hellas (CERTH), 57001 Thessaloniki, Greece; kohatzis@iti.gr (K.C.); marthakn@iti.gr (M.K.); hatzis@iti.gr (T.C.); vsol@iti.gr (V.S.); dikonsta@iti.gr (D.K.); dimitrop@iti.gr (K.D.)

* Correspondence: achilleasmplekos@iti.gr

Abstract: Grape maturity estimation is vital in precise agriculture as it enables informed decision making for disease control, harvest timing, grape quality, and quantity assurance. Despite its importance, there are few large publicly available datasets that can be used to train accurate and robust grape segmentation and maturity estimation algorithms. To this end, this work proposes the CERTH grape dataset, a new sizeable dataset that is designed explicitly for evaluating deep learning algorithms in grape segmentation and maturity estimation. The proposed dataset is one of the largest currently available grape datasets in the literature, consisting of around 2500 images and almost 10 k grape bunches, annotated with masks and maturity levels. The images in the dataset were captured under various illumination conditions and viewing angles and with significant occlusions between grape bunches and leaves, making it a valuable resource for the research community. Thorough experiments were conducted using a plethora of general object detection methods to provide a baseline for the future development of accurate and robust grape segmentation and maturity estimation algorithms that can significantly advance research in the field of viticulture.

Keywords: grape maturity estimation; grape-instance segmentation; object detection

Citation: Blekos, A.; Chatzis, K.; Kotaidou, M.; Chatzis, T.; Solachidis, V.; Konstantinidis, D.; Dimitropoulos, K. A Grape Dataset for Instance Segmentation and Maturity Estimation. *Agronomy* **2023**, *13*, 1995. https://doi.org/10.3390/agronomy13081995

Academic Editor: Valerio Cristofori

Received: 15 June 2023
Revised: 21 July 2023
Accepted: 25 July 2023
Published: 27 July 2023

1. Introduction

In recent years, due to climate change, it has become clear that the agricultural sector faces significant challenges in providing food of high quality and sufficient quantity to feed the entire population. In viticulture, specifically, the abrupt changes in temperature and humidity have contributed to the high occurrence of various grape diseases, such as downy mildew, powdery mildew, black rot, and botrytis, which can significantly affect the production and quality of table grapes. As a result, it is imperative for the research community to develop cost-effective methods that can automatically identify vulnerable-to-disease grape clusters before they are infected. These methods should be capable of accurately detecting grape clusters and estimating their maturity, as high sugar levels are accompanied by an increased risk of disease development, so farmers can optimize their harvests and enhance the overall quality of their produce. Through the robust estimation of grape maturity levels, such methods can also inform farmers about the correct time to harvest their grapes, contributing to the improvement of the quality and quantity of grapes sent to consumers.

Fortunately, advances in machine learning and sensor technologies have provided important tools for the development of automated methods for grape-bunch identification and maturity estimation. Regarding grape-bunch identification, early works relied on traditional machine learning techniques, such as support vector machines (SVMs) [1,2] and artificial neural networks (ANNs) [3], as well as image-filtering approaches, such as thresholding, mathematical morphology operators, and edge detection [4–7].

Recently, the success of deep learning in various computer vision applications has also led to its adoption for grape-bunch identification [8]. Powerful object detection models, such as Fast R-CNN [9], Faster R-CNN [10], Mask R-CNN [11], and YOLO [12], have been employed in the context of grape segmentation, achieving remarkable accuracy and real-time performance. These models initially employ convolutional neural networks (CNNs) to identify image regions as candidates for the location of the object of interest, and then they classify the objects in the candidate image regions into specific classes.

In [13], the authors employed Mask R-CNN, YOLOv2 [14], and YOLOv3 [15] for accurate grape segmentation and showed that Mask R-CNN achieved superior performance compared to the other networks. In a similar fashion, the authors of [16] employed YOLOv4 [17] to detect and quantify the grape yield in vineyards and found that YOLOv4 primarily detected large grape bunches due to occlusions of smaller grape bunches by leaves. Through extensive experimental results, the authors of [18] showed that YOLOv5x can achieve optimal performance in correctly detecting grape bunches, while the authors of [19] concluded that YOLOv5s can meet the precision and real-time performance requirements for grape-cluster detection. Finally, the authors of [20] employed more advanced YOLO network architectures (i.e., YOLOv5x6, YOLOv7-e6e, and YOLOR-CSP-X) to detect and classify grape bunches as healthy or damaged. Additionally, the authors created two new publicly available datasets for grape-bunch detection and classification with the same 910 original images and 1066 annotated grape bunches, augmented through scaling, rotation, translation, and blurring operations, ultimately leading to around 10 K images and more than 11.6 K grape bunches.

On the other hand, Shen et al. [21] introduced a novel method to accurately identify veraison in colored wine grapes under natural field-growing conditions. Initially, the authors utilized a semantic segmentation model to effectively eliminate the irrelevant background, and then a Mask R-CNN pipeline, which incorporates anchor parameter optimization, was utilized to further enhance the accuracy and robustness of the grape identification process. Finally, to deal with issues regarding grape-cluster overlapping and occlusions, Chen et al. in [22] employed the post-processing algorithm of Mask R-CNN, enhanced by the linear weighting method of the soft non-maximum suppression (Soft-NMS) algorithm [23], achieving significant performance improvements. To improve real-time performance in grape bunch detection, the authors of [24] combined a Swin Transformer [25] with YOLOv5. Their model was tested on two different grape varieties, 'Chardonnay' and 'Merlot', and under different conditions, including two weather conditions, two berry maturity stages, and three sunlight intensities. A comparison with other common detectors, such as Faster R-CNN, YOLOv4, and YOLOv5, revealed the superior performance that can be achieved when a Transformer network is combined with an object detection network. However, the complicated environment of vineyards, with varying illumination conditions and occlusions, and the lack of large annotated public datasets can significantly affect the performance of these methods.

Regarding grape maturity estimation, there are two main methods used in viticulture to estimate the maturity level of grapes. The first one is the traditional method, where we measure the sugar level using a specific instrument that is usually calibrated in Brix or pH, called a refractometer [26]. However, this specialized sensor is costly and requires farmers with technological knowledge to handle it. The second and most cost-effective solution involves the processing of grape images using machine learning and Artificial Intelligence (AI) techniques. A comprehensive review of these techniques in viticulture, along with a new dataset (i.e., GrapeCS-ML) that contains images of grape varieties at different stages of development, together with the corresponding ground-truth data obtained through chemical analysis (i.e., pH and Brix), can be found in [27].

To estimate grape maturity, color images have been utilized and processed using standard machine learning algorithms, such as artificial neural networks [28,29], random forests [30], unsupervised clustering [31], and CNNs [32–35]. For ground-truth maturity levels, some methods in the literature have utilized grape color and shape information [28,30,31,36], the

knowledge of the harvesting week [33], and well-known chemical indices, such as total soluble solids (TSSs), titratable acidity (TA), and pH [29]. Attention mechanisms have also been employed, achieving significant performance improvements in grape maturity estimation. An improved SM-YOLOv4 algorithm that utilizes YOLOv4, as well as Mobilenetv3 and the SENet attention mechanism as the backbone feature extraction network, was proposed in [34] to increase inference speed, improve robustness, and make the network more lightweight. In addition, a novel Mask R-CNN-based algorithm that utilizes three different attention mechanism modules (i.e., squeeze-and-excitation attention (SE), convolutional block attention (CBAM), and coordinate attention (CA)) in its ResNet backbone was proposed in [35].

Other works have utilized hyperspectral or multispectral images to improve classification accuracy and robustness. In [37], partial least-squares regression (PLSR) and neural networks were utilized to estimate the quality of grapes from sugar content predictions based on hyperspectral imaging. The results suggest that the combination of hyperspectral imaging with appropriate chemometric techniques or machine learning algorithms can lead to satisfactory generalization for vintages not employed in model training. Similarly, the authors of [38] proposed a CNN-based method to predict two important enological parameters (i.e., sugar level and pH) from hyperspectral images and achieved excellent overall performance, even for different varieties or vintages that were not employed during training. On the other hand, dimensionality reduction methods were studied in [39] for the prediction of sugar content from hyperspectral images of wine grape berries. The authors combined hyperspectral imaging with neural networks, achieving good generalization capacity across different datasets.

Similarly, the authors of [40] tested a multilayer perceptron (MLP) and a 3D-CNN network in a novel multispectral dataset of grape images, coupled with measurements such as weight, anthocyanin content, and Brix index, achieving very high accuracy. In [41], data from canopy reflectance sensors, as well as Unmanned Aerial Vehicle (UAV) and satellite images, were processed using open-source AutoML techniques for robust prediction of grape-quality attributes. On the other hand, absorbance and fluorescence data of Cabernet Sauvignon grape samples were employed to predict relevant grape indices in terms of technological, phenolic, and flavor maturity in [42]. Finally, in [43], four different machine learning algorithms (i.e., PLSR, random forest regression, support vector regression (SVR), and CNN) were employed to estimate grape ripeness in Brix from hyperspectral data received from a contact probe spectrometer. The authors concluded that the CNN model outperformed the other algorithms in most test cases.

Despite the high performance of the aforementioned deep learning algorithms in grape segmentation and maturity estimation, there is still room for improvement so that these algorithms can be used in real-life applications. In principle, deep learning algorithms usually require more data than traditional machine learning techniques and thus they are heavily dependent on large, annotated datasets to achieve high accuracy, robustness, and generalization. To address this need, several datasets encompassing different data types (e.g., images, sugar levels, hyperspectral information) have been proposed in the literature, but they remain too small to be successfully employed for deep learning training. This work proposes one of the largest annotated grape datasets to date, called the CERTH grape dataset, whose purpose is to be used for the training and optimization of both grape segmentation and maturity estimation algorithms. The images in the dataset depict table grapes from the Crimson Seedless variety, which were captured under different environmental conditions and with significant occlusions between grape clusters and leaves. Thus, the size and the challenging nature of the CERTH grape dataset can substantially benefit any machine learning method trained and evaluated on it. As a result, the main contributions of this work are as follows:

- The new CERTH grape dataset is one of the largest grape segmentation and maturity estimation datasets in the literature. It consists of around 2.5 K images, captured under varying illumination and viewing conditions, with strong occlusions between grape bunches and leaves. The dataset contains almost 10 K annotated grape bunches,

which are further classified into three maturity classes (i.e., immature, semi-mature, and mature).
- A plethora of state-of-the-art object detection methods are trained and evaluated on the CERTH grape dataset, providing a baseline for future research and showcasing the challenging nature of the CERTH grape dataset.

2. Materials and Methods

This section initially presents the well-known publicly available image-based grape segmentation and maturity estimation datasets and then introduces the proposed CERTH grape dataset, which aims to overcome the limitations of the currently available datasets. Finally, this section presents state-of-the-art object detection algorithms, which are trained and evaluated in the tasks of grape segmentation and maturity estimation.

2.1. Available Datasets

Several publicly available grape datasets, comprising different types of data, such as images, genetic data, environmental and weather data, and chemical data (e.g., Brix, pH), have been developed to support research in grape segmentation and maturity estimation. The *Grape CS-ML database* [27], released by Charles Sturt University in 2018, consists of five datasets showcasing 15 grape varieties at different stages of development, accompanied by size and Macbeth color references. The Embrapa Wine Grape Instance Segmentation Dataset (WGISD) [44], released in 2019, contains 300 images with annotated grape clusters, including bounding boxes and binary masks. The AI4EU Grape Dataset [45], published in 2021, contains 250 images of Tempranillo grapes with bounding box annotations. The wGrapeUNIPD-DL dataset [46], available since 2022, comprises 373 images of various grape varieties captured at different phenological stages across six Italian vineyard locations. GrapesNet [47], published in 2023, offers four datasets containing RGB and RGB-D images of grape bunches, facilitating tasks such as grape segmentation and weight prediction. Representative examples of images appearing in the aforementioned datasets are presented in Figure 1.

Despite the importance of these datasets for grape segmentation and maturity estimation, there are some shortcomings. Most of the currently available datasets have been developed for grape segmentation, and only the Grape CS-ML database can be used for maturity estimation. However, the Grape CS-ML database consists of several smaller datasets that depict a lot of different grape varieties with their own distinct characteristics. These datasets are annotated based on different criteria (i.e., color, shape, size, sugar levels) for maturity estimation. Additionally, the relatively small number of images in most of the other datasets can significantly affect the performance of deep learning-based grape segmentation algorithms.

2.2. CERTH Grape Dataset

To leverage the scarcity of large annotated grape datasets and overcome the shortcomings of the currently available grape datasets, we introduce in this work a new dataset, named the CERTH grape dataset. The aim is to advance computer vision and machine learning research in the field of viticulture by providing valuable annotated data for developing and refining algorithms for accurate grape segmentation, yield prediction, and, most importantly, maturity estimation. The proposed dataset consists of 2502 high-resolution images captured from a vineyard cultivating the 'Crimson Seedless' table grape variety during the 2022–2023 development and harvesting period.

The data collection process involved the use of an iPhone 11 Pro smartphone, positioned strategically between the rows of vines, to capture grape images at distances ranging from approximately 1 to 2 m. This careful arrangement ensured optimal visualization of the grape bunches and their surroundings. To maintain consistency and uniformity within the dataset, all captured images were subsequently scaled to a resolution of 2160 $\times$ 3840 pixels, ensuring the same high level of detail and clarity for each image in the dataset.

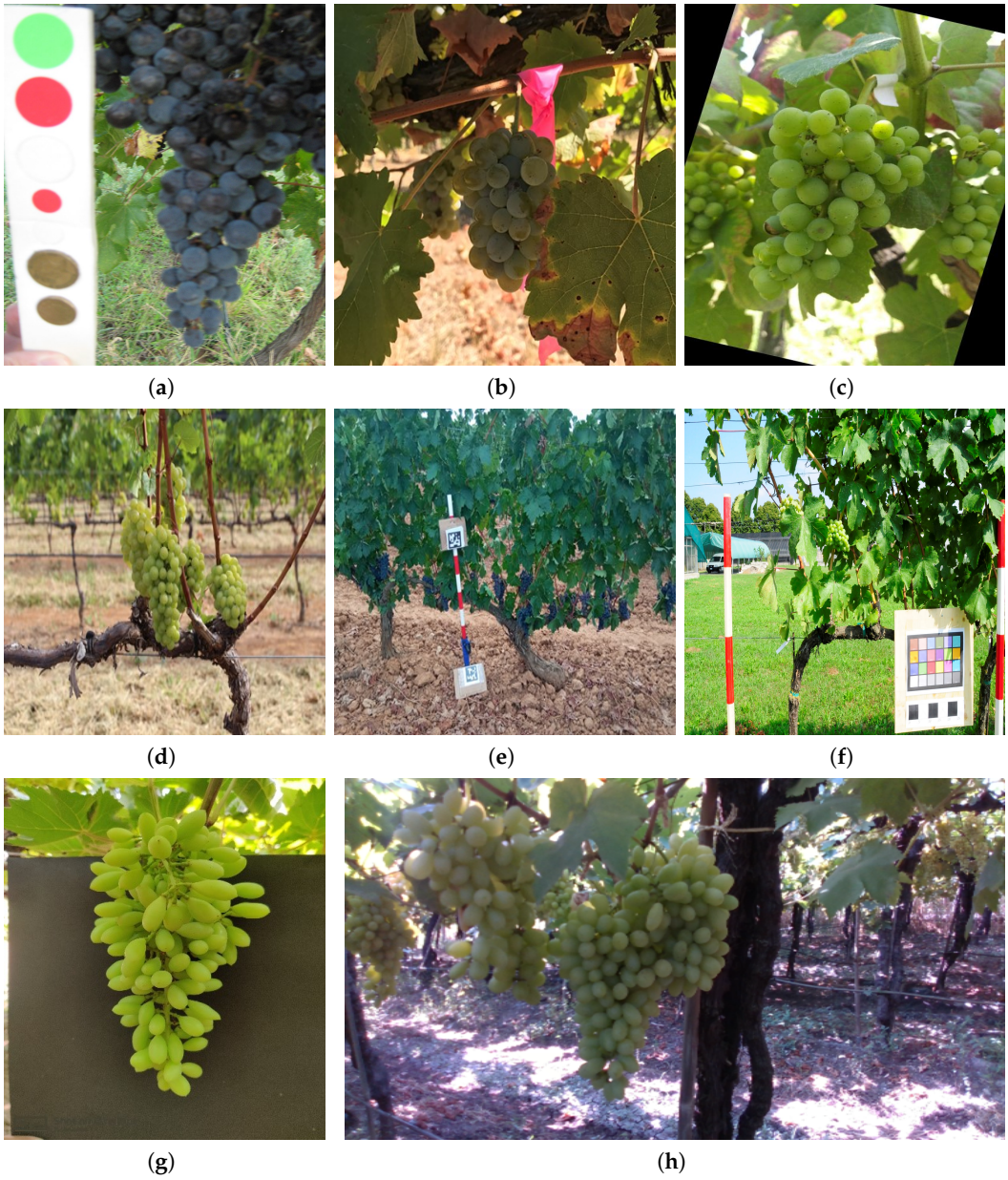


Figure 1. Representative images from currently available datasets. (**a**) Dataset 1 of the Grape CS-ML dataset, (**b**) Dataset 5 of the Grape CS-ML dataset; (**c**) Grapevine dataset [20], (**d**) Embrapa WGISD dataset, (**e**) AI4EU Grape Dataset, (**f**) wGrapeUNIPD-DL dataset, (**g**) Dataset 1 of GrapesNet, (**h**) Dataset 3 of GrapesNet.

Afterward, the images in the CERTH grape dataset were meticulously annotated by human experts using the advanced Ritm segmentation tool [48], which offers advanced functionalities to aid in the accurate and efficient annotation of grape bunches while also utilizing cutting-edge algorithms and models to automate the process of segmenting objects. During the annotation procedure, all grape bunches were identified and segmented from their surroundings using detailed object masks, as shown in Figure 2. Additionally, each bunch was categorized into three distinct classes (i.e., immature, semi-mature, and mature) based on the degree of grape maturity, as identified by the color of the grapes in the bunch. In the maturity annotation procedure, assistance was provided by agronomists who received information regarding the week of grape development and the time distance from harvest. As a result, grapes in the immature class were early in their development phase, grapes in the mature class were close to the harvesting season, and grapes in the

semi-mature class were in the intermediate period when changes in the color of the grapes from yellow to red had initiated.

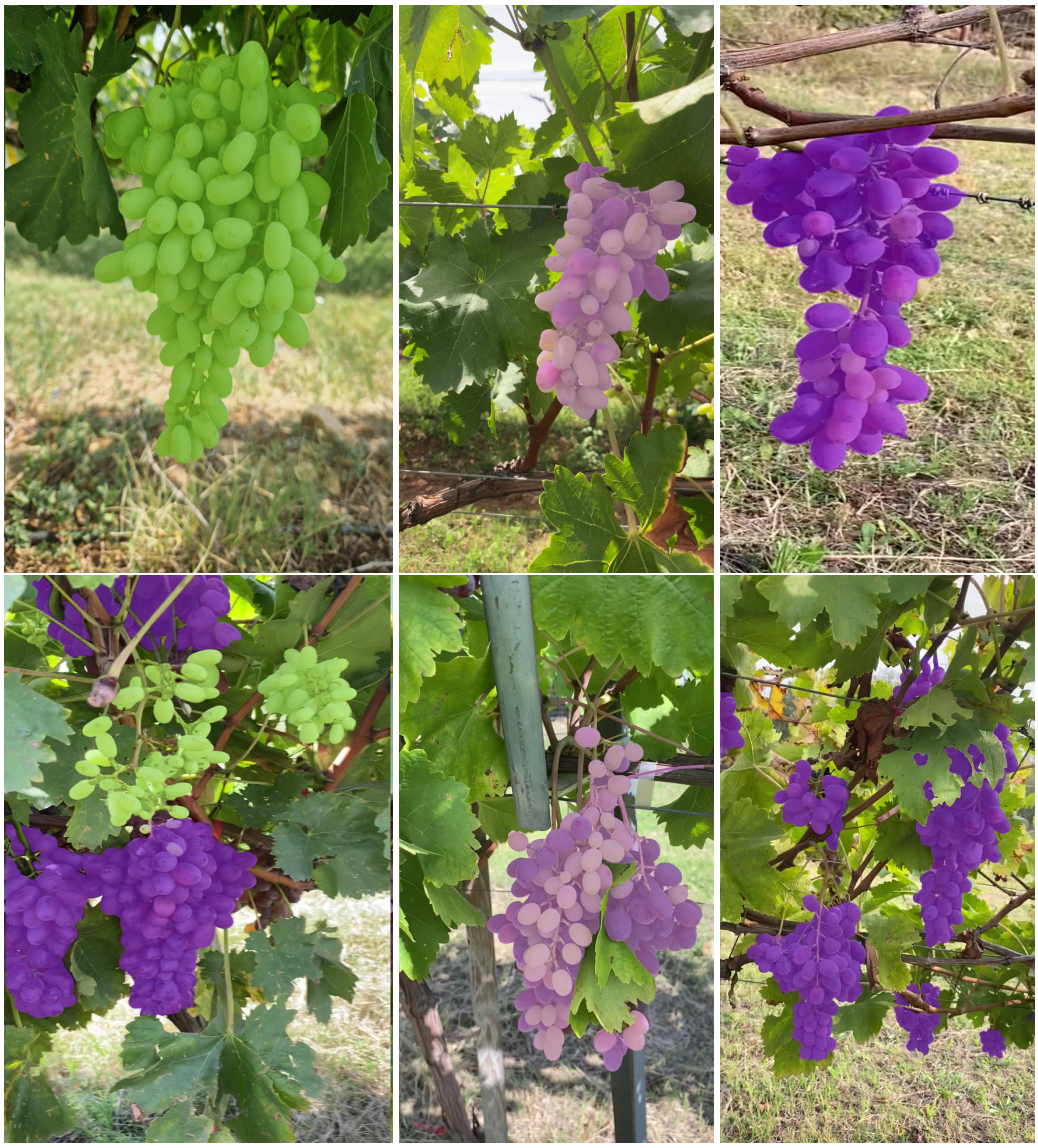


Figure 2. Images from the CERTH grape dataset annotated with masks and maturity levels.

After the annotation procedure, the CERTH grape dataset consisted of 9832 labeled grape bunches, extracted from the 2502 images. The grape images were then split into training, validation, and test sets, consisting of 2000, 251, and 251 images, respectively. The distribution of the grape bunches and their maturity levels in the images in the proposed dataset can be seen in Table 1. In the table, it can be seen that most of the labeled grape bunches belonged to the mature class, whereas the other two classes had a similar, smaller number of labeled bunches, close to 1500. Due to the capturing setup and environmental conditions, the CERTH grape dataset was designed to be a challenging dataset for machine learning techniques. The proposed dataset exhibits different view angles, camera focus conditions, and illumination variations (Figure 3), as well as significant occlusions between grape bunches and between bunches and leaves (Figure 4).

Table 1. Distribution of labeled grape bunches per maturity level in the CERTH grape dataset.

	Immature	Semi-Mature	Mature	Total
Training set	1278	1099	5582	7959
Validation set	200	155	559	914
Test set	190	144	625	959
Total	1668	1398	6766	9832

Figure 3. Images of grape bunches with varying camera focuses and illumination conditions.

Figure 4. Images of grape bunches with significant occlusions between grape bunches and leaves.

Furthermore, to facilitate the comprehensive evaluation of the performance of the algorithms in single-instance classification, 100 images were taken from the test set of the CERTH grape dataset and cropped to depict a single grape bunch per image. This new single-instance/one-class subset consisted of 36, 43, and 21 images that depict grape bunches from the immature, semi-mature, and mature classes, respectively. This subset provides a unique opportunity to assess the robustness and reliability of the object detection models in accurately identifying and classifying objects in a highly specific context.

Finally, a comparison of the proposed dataset and other publicly available image-based grape datasets is presented in Table 2. It can be concluded that the proposed dataset is the only other dataset, apart from the Grape CS-ML database, that can be used for both grape segmentation and maturity estimation. However, the Grape CS-ML database depicts different grape varieties with their own distinct characteristics, whereas the CERTH grape dataset depicts a single grape variety. Additionally, most of the available grape

segmentation datasets comprise a relatively small number of images, which can significantly inhibit the performance of the deep learning algorithms trained on them. Therefore, due to its size, uniformity (i.e., depiction of a single grape variety), and annotated maturity levels, the proposed dataset can be considered suitable for the training and evaluation of deep learning models in the tasks of grape segmentation and maturity estimation.

Table 2. Comparison of CERTH grape dataset and publicly available image-based grape datasets. GS: Grape segmentation; ME: Maturity estimation.

Name	Year	Number of Images	Grape Varieties	Modality	Applicability
Grape CS-ML (Datasets 1–4) [27]	2018	2016	15	RGB, Macbeth color, size	GS, ME
Grape CS-ML (Dataset 5) [27]	2018	62	1	RGB, Sugar level	GS, ME
Embrapa WGISD [44]	2019	300	5	RGB	GS
AI4EU Grape Dataset [45]	2021	250	1	RGB	GS
wGrapeUNIPD-DL dataset [46]	2022	271	4	RGB	GS
GrapesNet [47]	2023	2129	1	RGB-D	GS
Grapevine bunch (condition) detection [20]	2023	910 *	6	RGB	GS
CERTH grape dataset	2023	2502	1	RGB, Maturity levels	GS, ME

* The original 910 images were augmented to 10 K images using various morphological and blurring operations.

2.3. Deep Learning Methods

To train and evaluate the performance of deep learning algorithms in the tasks of grape segmentation and maturity estimation using the proposed CERTH grape dataset, the MMDetection toolbox [49] was utilized. MMDetection is a versatile object detection toolbox based on PyTorch that provides a wide range of pre-trained models, algorithms, and modules for object detection and instance segmentation tasks. A training pipeline was developed in this work that included 13 pre-trained object detection and instance segmentation models taken from the library of MMDetection and fine-tuned on the augmented CERTH grape dataset in a similar fashion to the COCO training pipeline.

2.3.1. YOLOv3

YOLOv3 [15] consists of Darknet-53 as the backbone network, followed by detection layers with anchor boxes at different scales, allowing for accurate predictions of objects of various sizes and different classes in an image. The output of YOLOv3 includes bounding boxes, object classes, and confidence scores, specifying the location, size, and class of the objects in an image, as well as the confidence of the model in its prediction. In this work, two YOLOv3 variants, namely YOLOv3-320 and YOLOv3-608, differing in the size of the input image (i.e., 320 and 608 pixels, respectively), were evaluated.

2.3.2. YOLOv8

YOLOv8 [50] utilizes an anchor-free model with a decoupled head, enabling independent processing of objectness, classification, and regression tasks. This design empowers each branch to focus on its specific task, ultimately improving the overall accuracy of the model. In this work, we utilized YOLOv8n, which consists of 225 layers, with a depth and

width of 0.33 and 0.25, respectively, and prioritizes speed and efficiency. In addition, we utilized YOLOv8s, which consists of 225 layers, with a depth and width of 0.33 and 0.5, respectively, and strikes a balance between performance and model size.

2.3.3. Mask R-CNN

Mask R-CNN [11] consists of three key components: (a) the backbone network, which combines a ResNet [51] and a Feature Pyramid Network (FPN) [52] to extract features at different scales; (b) the Region Proposal Network (RPN), which proposes candidate object regions as bounding box coordinates and binary classes (foreground or background); and (c) the Region of Interest (ROI) Align component that processes the candidate regions to perform class prediction, bounding box refinement, and pixel-wise mask generation for each detected object instance. In this study, we evaluated the performance of Mask R-CNN-R50 and Mask R-CNN-R101, which use ResNet-50 and ResNet-101 backbones, respectively.

2.3.4. Cascade Mask R-CNN

Cascade Mask R-CNN [53] enhances the original Mask R-CNN architecture through the use of cascades. The cascades divide the task into multiple stages with different IoU thresholds, sequentially training each stage using the output of the previous stage as the training set. This approach improves object detection and instance segmentation by adapting to varying IoU thresholds. In this study, we evaluated Cascade Mask R-CNN-R50 and Cascade Mask R-CNN-R101, which use ResNet-50 and ResNet-101, respectively, as the backbones.

2.3.5. Hybrid Task Cascade (HTC)

HTC [54] is inspired by Cascade Mask R-CNN and introduces direct connections between mask branches in the cascade for robust information flow and the progressive refinement of masks. HTC also includes a new branch for pixel-wise semantic segmentation, trained jointly with the other branches. By incorporating semantic segmentation information, HTC achieves improved predictions, especially in complex backgrounds. In this study, we evaluated HTC with ResNet-50 and ResNet-101 backbones, resulting in HTC-R50 and HTC-R101, respectively.

2.3.6. Mask R-CNN with Swin Transformer

The Swin Transformer [25] is a versatile backbone architecture that splits the input image into non-overlapping patches and then processes these patches using a self-attention mechanism to identify the content of the image. Mask R-CNN and Swin can be combined by allowing Swin to replace the backbone network of Mask R-CNN, enabling more efficient and effective feature extraction by addressing the known limitations of traditional CNN backbones, such as handling large-scale images and effectively capturing global contexts. In this work, we utilized the well-known Swin tiny and Swin small network architectures as the backbones in Mask R-CNN, giving rise to the Mask R-CNN-Swin(T) and Mask R-CNN-Swin(S) architectures, respectively.

2.3.7. Mask2Former with Swin Transformer

Mask2Former [55] is a unified model designed for various image segmentation tasks, including panoptic, instance, and semantic segmentation. It incorporates masked attention to extract localized features within predicted mask regions, enabling the model to focus on relevant areas and capture fine-grained details. In this study, we utilized Mask2Former with Swin small as the backbone, giving rise to the Mask2Former-Swin(S) architecture.

3. Results and Discussion

This section presents the experimental results related to the evaluation of a plethora of state-of-the-art object detection methods in the task of grape segmentation and maturity estimation using the CERTH grape dataset. These experiments aim to provide a baseline

for the development and evaluation of future object detection algorithms. To this end, the 13 different object detection methods introduced in Section 2.3 were tested on the multiple-instance/multi-class CERTH grape dataset, as well as on the single-instance/one-class subset of the CERTH grape dataset. As far as the implementation details are concerned, all the network architectures utilized were the default ones provided by the MMDetection toolbox, with the differences lying in the number of classes of the head, which was set to 3 (i.e., the number of maturity levels). As a result, all object detection methods were trained to identify both the location and maturity level of grape bunches in a similar fashion to the COCO models, which identify both object boundaries and classes. The RPN anchors were the default values, configured to have one scale of a size of 8 pixels and three aspect ratios, which were 0.5, 1.0, and 2.0. In addition, Mask2Former's object queries were set to 100.

Table 3 presents the sizes of the tested object detection methods. In Table 3, it can be observed that the YOLOv8 networks were the most lightweight, with less than 12 M parameters. On the other hand, Cascade Mask R-CNN-R101 and HTC-R101 were the largest networks, with more than 96 M parameters. Such large networks usually require large datasets for effective training, and the aim of this work was to provide one such dataset for the tasks of grape segmentation and maturity estimation.

Table 3. Number of parameters for the tested object detection methods.

Method	Parameters	Method	Parameters
YOLOv3-320	61.53 M	Cascade Mask R-CNN-R101	96.02 M
YOLOv3-608	61.53 M	HTC-R50	77.16 M
YOLOv8n	3.01 M	HTC-R101	96.15 M
YOLOv8s	11.12 M	Mask R-CNN-Swin(T)	47.38 M
Mask R-CNN-R50	43.76 M	Mask R-CNN-Swin(S)	69.11 M
Mask R-CNN-R101	62.75 M	Mask2Former-Swin(S)	68.71 M
Cascade Mask R-CNN-R50	77.03 M		

The CERTH grape dataset was split into training, validation, and test sets, with each set representing 80%, 10%, and 10% of the dataset for each class, ensuring equal distribution of samples per class among all sets. The performance of the methods was measured using the standard metric of mean average precision (mAP) with predictions of either bounding boxes or masks, depending on the capabilities of the specific method. Mean average precision was computed by taking the average precision for the confidence threshold of 0.3 and IoU thresholds ranging from 0.5 to 0.95 with an increment of 0.05. In addition, precision, recall, and F1-score metrics were utilized for predictions with an IoU threshold of 0.5 and a confidence threshold of 0.3. Tables 4 and 5 present the experimental results of the different object detection methods on the test set of the CERTH grape dataset (i.e., multiple-instance/multi-class scenario) and on the single-instance/one-class subset of the CERTH grape dataset, respectively, in the form of bounding box/mask performance.

From the experimental results in Tables 4 and 5, it can be concluded that the optimal performance was achieved by the Mask2Former-Swin(S) algorithm. More specifically, Mask2Former-Swin(S) significantly outperformed all the other algorithms by at least 6% in both scenarios, achieving an mAP of 51.5% and 83.7% in mask prediction in the multiple-instance/multi-class and single-instance/one-class scenarios, respectively. Of significance was the performance of the YOLOv8s algorithm, which was able to achieve the second-highest accuracy in the prediction of bounding boxes for both scenarios, even overcoming the accuracy of the Mask R-CNN-Swin(S) algorithm. Moreover, the comparison of the Mask R-CNN algorithm with a backbone of ResNet or the Swin Transformer showed that the Swin Transformer was capable of better modeling the input image compared to ResNet, effectively leading to an overall improvement in the accuracy of the Mask R-CNN algorithm in grape segmentation and maturity estimation by at least 2% in the multiple-instance/multi-class scenario and 1.6% in the single-instance/one-class scenario. Finally,

among all the versions of the YOLO algorithms, YOLOv8 achieved the best results, with a significant accuracy margin of at least 11% compared to YOLOv3.

Table 4. Comparison of different state-of-the-art algorithms in the multiple-instance/multi-class scenario of the CERTH grape dataset for grape segmentation and maturity estimation. Results appear in the form of bounding box/mask performance.

Method	mAP	Precision	Recall	F1-Score
YOLOv3-320	31.7%/ -	0.57/-	0.84/-	0.68/-
YOLOv3-608	36.0%/-	0.71/-	0.87/-	0.71/-
YOLOv8n	45.4%/-	0.68/-	0.93/-	0.78/-
YOLOv8s	46.3%/-	0.68/-	0.95/-	0.79/-
Mask R-CNN-R50	42.9%/40.4%	0.64/0.64	0.87/0.87	0.74/0.74
Mask R-CNN-R101	44.7%/42.2%	0.64/0.67	0.87/0.88	0.74/0.76
Cascade Mask R-CNN-R50	45.5%/42.0%	0.65/0.67	0.84/0.85	0.73/0.75
Cascade Mask R-CNN-R101	45.9%/42.1%	0.64/0.66	0.85/0.86	0.73/0.75
HTC-R50	46.5%/42.6%	0.66/0.67	0.92/0.93	0.77/0.78
HTC-R101	46.9%/42.9%	0.65/0.67	0.93/0.94	0.77/0.78
Mask R-CNN-Swin(T)	45.4%/43.3%	0.66/0.69	0.92/0.93	0.77/0.79
Mask R-CNN-Swin(S)	45.6%/44.1%	0.68/0.71	0.95/0.95	0.79/0.81
Mask2Former-Swin(S)	52.3%/51.5%	0.68/0.73	0.95/0.96	0.79/0.83

Table 5. Comparison of different state-of-the-art algorithms in the single-instance/one-class scenario of the CERTH grape dataset for grape segmentation and maturity estimation. Results appear in the form of bounding box/mask performance.

Method	mAP	Precision	Recall	F1-Score
YOLOv3-320	59.0%/-	0.85/-	0.99/-	0.91/-
YOLOv3-608	59.1%/-	0.88/-	0.99/-	0.93/-
YOLOv8n	79.3%/-	0.89/-	0.99/-	0.94/-
YOLOv8s	79.3%/-	0.90/-	0.99/-	0.95/-
Mask R-CNN-R50	66.7%/63.4%	0.89/0.87	0.99/0.98	0.94/0.92
Mask R-CNN-R101	76.0%/71.6%	0.90/0.90	0.99/0.99	0.94/0.95
Cascade Mask R-CNN-R50	75.2%/68.1%	0.88/0.88	0.99/0.99	0.93/0.93
Cascade Mask R-CNN-R101	73.2%/66.9%	0.85/0.85	0.99/0.99	0.92/0.91
HTC-R50	74.9%/67.3%	0.87/0.87	0.99/0.99	0.93/0.93
HTC-R101	76.1%/69.4%	0.89/0.89	0.99/0.99	0.94/0.94
Mask R-CNN-Swin(T)	73.4%/69.5%	0.90/0.88	0.99/0.98	0.94/0.93
Mask R-CNN-Swin(S)	78.7%/73.2%	0.89/0.89	0.99/0.99	0.94 /0.94
Mask2Former-Swin(S)	85.7%/83.7%	0.92/0.93	0.99/0.99	0.96/0.96

Similar conclusions can be drawn for the other metrics. Mask2Former-Swin(S) achieved the highest F1-score, with a value of 0.79 for the bounding boxes and 0.83 for the masks. From the recall and precision values, it can be seen that most of the grape bunches were correctly identified by all the object detection algorithms (recall ranged from 0.84 to 0.96). However, the algorithms tended to make several mistakes through detections that were not actual grape bunches (i.e., false positives), with precision ranging from 0.57 to 0.73. These false positives were usually parts of grape bunches that were identified by the object detection algorithms as individual instances and rarely background objects that were identified as grape bunches. This observation verifies an inherent issue in grape segmentation concerning the correct delineation of individual grape bunches, especially when such bunches are occluded by other bunches or background objects.

Additionally, for the four state-of-the-art mask-prediction object detection algorithms (i.e., Cascade Mask R-CNN-R101, HTC-R101, Mask R-CNN-Swin(S), and Mask2Former-Swin(S)), a different presentation of the results based on the maturity level of the grape bunches was performed. The aim of this experiment was to analyze the performance of a few of the best-tested object detection algorithms in terms of their ability to correctly classify

the maturity levels of the grape bunches and use these results as a baseline for evaluations in future research. Tables 6 and 7 present the performance of the four object detection algorithms in the task of grape maturity level estimation in the multiple-instance/multi-class and single-instance/one-class scenarios, respectively.

Table 6. Grape maturity level estimation in the multiple-instance/multi-class scenario of the test set of the CERTH grape dataset.

	Method			
	Cascade Mask R-CNN-R101	**HTC-R101**	**Mask R-CNN-Swin(S)**	**Mask2Former-Swin(S)**
Maturity Level	**mAP BBox/Mask**			
Immature	41.4%/35.4%	41.3%/35.1%	42.5%/39.3%	47.8%/44.5%
Semi-mature	34.3%/32.9%	38.4%/35.6%	33.8%/33.2%	43.7%/42.6%
Mature	61.9%/58.1%	61.0%/57.8%	60.4%/59.8%	65.5%/66.3%
Total	45.9%/42.1%	46.9%/42.9%	45.6%/44.1%	52.3%/51.5%

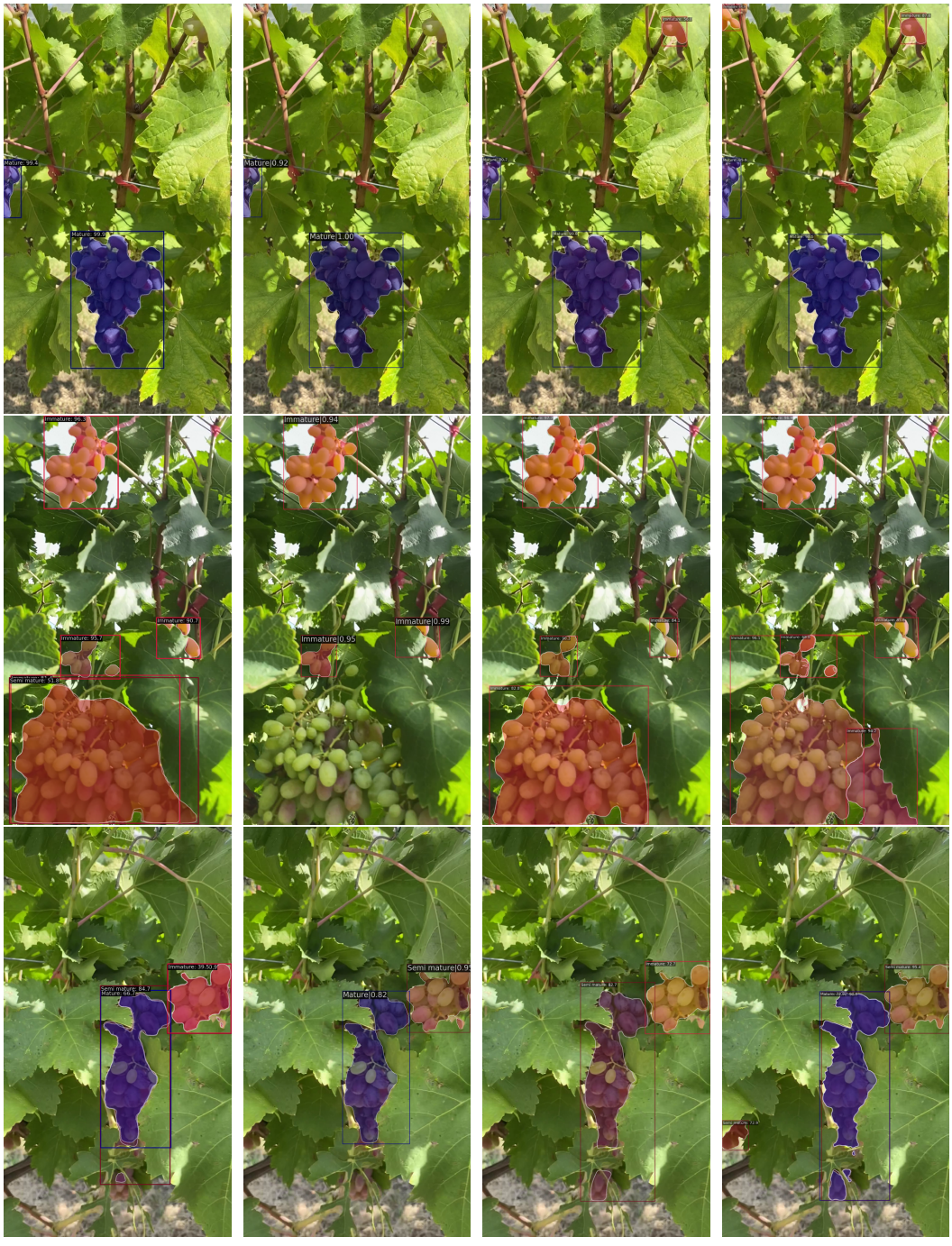
Table 7. Grape maturity level estimation in the single-instance/one-class scenario of the CERTH grape dataset.

	Method			
	Cascade Mask R-CNN-R101	**HTC-R101**	**Mask R-CNN-Swin(S)**	**Mask2Former-Swin(S)**
Maturity Level	**mAP BBox/Mask**			
Immature	62.7%/55.5%	66.8%/58.4%	62.2%/59.4%	83.5%/77.8%
Semi-mature	80.2%/70.8%	88.7%/78.0%	71.3%/71.1%	91.1%/86.2%
Mature	74.1%/72.0%	72.9%/69.8%	76.6%/76.3%	82.5%/87.2%
Total	73.2%/66.9%	76.1%/69.4%	78.7%/73.2%	85.7%/83.7%

From the results in Tables 6 and 7, it can be concluded that the Mask2Former-Swin(S) achieved the best performance in every maturity level category, outperforming all the other algorithms by a significant margin. A comparison of the maturity level estimation results revealed that all the tested algorithms achieved the highest accuracy for the mature class, whereas the lowest accuracy was achieved for the semi-mature class. This can be attributed to the color discrepancies among the mature red grapes, the immature yellow grapes, and the leaves, which led the yellow grapes to be more easily blended with the leaves compared to the red grapes. On the other hand, the existence of semi-mature grape bunches with red and yellow grapes in the same bunch posed challenges for the object detection algorithms in accurately segmenting them from the background. Moreover, the high accuracy for the mature class can also be attributed to the fact that the samples for this class were five times greater than the samples for the other two classes. These results further support the need for large, annotated datasets to enhance the accuracy and robustness of grape segmentation and maturity estimation algorithms.

Figure 5 presents a few images from the CERTH grape dataset, along with the predictions made by the four state-of-the-art object detection algorithms, namely Cascade Mask R-CNN-R101, HTC-R101, Mask R-CNN-Swin(S), and Mask2Former-Swin(S). From these images, a few important observations can be made regarding the performance of the algorithms in grape segmentation and maturity estimation. Although Cascade Mask R-CNN-R101 estimated the maturity levels of the grape bunches with high accuracy, it faced problems with the segmentation of small bunches, and it made several predictions for the same grape bunch. Similarly, HTC-R101 failed to segment small bunches, and it also had issues with detecting the large grape bunch in the image in the second row of Figure 5. On the other hand, Mask R-CNN-Swin(S) managed to identify even small grape bunches in the images but at the expense of incorrectly estimating the maturity levels of some

grape bunches. Finally, Mask2Former-Swin(S) successfully detected all grape bunches, even ones of small sizes, in the images, and correctly predicted the maturity levels of most grape bunches, achieving the best accuracy among all the state-of-the-art algorithms on the CERTH grape dataset. However, Mask2Former-Swin(S) incorrectly split the large grape bunch in the image in the second row of Figure 5 into three sub-bunches.



(**a**) (**b**) (**c**) (**d**)

Figure 5. Predictions in 3 images of the CERTH grape dataset using (**a**) Cascade Mask R-CNN-R101, (**b**) HTC-R101, (**c**) Mask R-CNN-Swin(S), and (**d**) Mask2Former-Swin(S).

From Figure 5, it is clear that almost all tested algorithms faced difficulties in accurately detecting and correctly estimating the maturity of grape bunches in the images. This inability stems from the challenging nature of the CERTH grape dataset, which contains

images with significant illumination variations, grape bunches captured under different viewing angles and focal lengths, and occlusions between grape bunches and between grape bunches and leaves. These challenges are further enhanced by the complex environment of the vineyard and the trellis system that is used for systematic cultivation.

Finally, Figure 6 presents two confusion matrices, which describe the maturity estimation results of the best-performing algorithm (i.e., Mask2Former-Swin(S)) in the multiple-instance/multi-class and single-instance/one-class scenarios. From the results, it can be seen that the most challenging class for correct classification was the semi-mature class. This can be attributed to the fact that the color of grapes can be both yellow and red, which posed challenges to the performance of the object detection algorithm. Of significant interest are the results in the confusion matrix of the multiple-instance/multi-class scenario regarding the misclassifications of the background. From these results, it can be determined that more than 12% of the immature grape bunches were not correctly identified, a percentage that fell to almost 8.6% in the case of the mature grape bunches. These results can be attributed to the yellow color of the immature grape bunches, which can be easily mistaken for the leaves in the complex environment of a vineyard, whereas the red mature grape bunches can be better distinguished from the background.

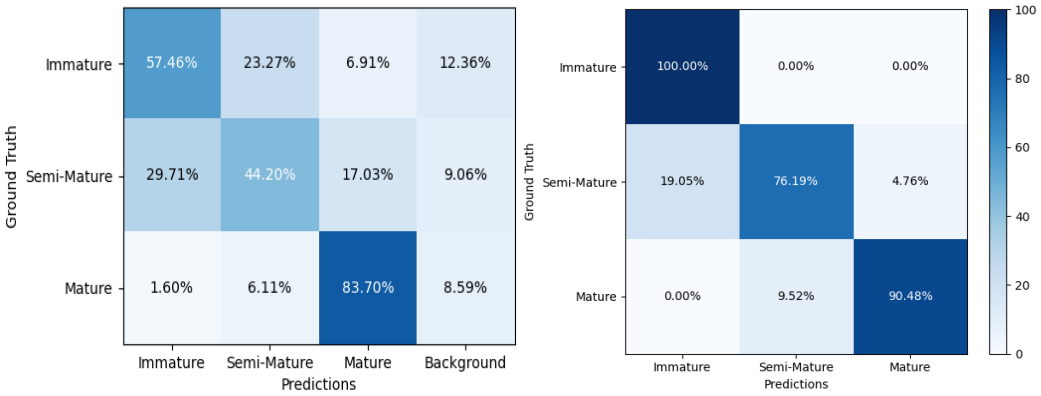


Figure 6. Confusion matrix of maturity estimation results from Mask2former-Swin(S) in the multiple-instance/multi-class and single-instance/one-class scenarios.

The aforementioned results reveal that the CERTH grape dataset is quite challenging due to the large number of grape bunches, the illumination variations, and the occlusions between grape bunches and between bunches and leaves. The results presented in other works show that even traditional machine learning algorithms can achieve high performance on publicly available datasets. For instance, the authors of [27] employed several traditional machine learning algorithms (e.g., SVM, k-NN, classification trees, etc.) for grape detection on the Grape CS-ML dataset, achieving high accuracy of more than 80% and 85% for white and red grape varieties, respectively. In future work, they stated the need to collect more data and experiment with more complex deep learning networks to facilitate research in viticulture. On the other hand, the authors of [20] tested three YOLO networks for grape bunch identification, concluding that the best performance was achieved by YOLOv7, with a precision of 98%, a recall of 90%, an F1-score of 94%, and an mAP of 77%. Similarly, the authors of [56] tested their proposed model, consisting of a Transformer network and a multi-scale feature extractor inspired by YOLO, on the wGrapeUNIPD-DL dataset, achieving a precision of 85.7%, a recall of 62.3%, an F1-score of 72.2%, and an mAP of 72.8%. The same authors tested their model on the Embrapa WGISD dataset, achieving a precision of 88.6%, a recall of 78.3%, an F1-score of 83.1%, and an mAP of 87.7%. These results demonstrate the enhanced performance of the object detection methods tested on the currently available datasets, with mAPs higher than 70%. The results on the Grapevine bunch (condition) detection and Embrapa WGISD datasets reveal high accuracy and recall results, meaning that the tested object detection methods can easily detect the actual grape bunches without producing a lot of false positives. On

the other hand, the results on the wGrapeUNIPD-DL dataset reveal a high precision of 78.3%, with a mediocre recall of 62.3%, meaning that the tested object detection methods do not identify a lot of false positives but fail to recognize the actual grape bunches.

In comparison to previous works, the best-performing model on the proposed dataset achieved an mAP of 52.3%, a precision of 68%, and a recall of 95%. This means that the model recognized almost every annotated grape bunch but also produced a lot of false positives. When the task of grape bunch segmentation was simplified by cropping a single grape bunch in each image (single-instance/one-class scenario), thus removing most of the background objects and occlusions, the performance of all object detection methods significantly improved, with Mask2Former-Swin(S) achieving an mAP of 85.7% and an F1-score of 96%. However, in real-life applications, the complex environment of a vineyard with varying lighting conditions and strong occlusions can pose significant challenges to a grape segmentation and maturity estimation algorithm, thus demonstrating the importance of developing challenging datasets that can thoroughly test the performance of such algorithms.

4. Conclusions

Leveraging the need for large and challenging public datasets in viticulture, this work introduces the CERTH grape dataset, which is one of the largest grape segmentation and maturity estimation datasets in the literature. The proposed dataset consists of around 2.5 K images, captured under varying illumination conditions and viewing angles, and almost 10 K heavily occluded grape bunches annotated with their detailed masks and maturity levels. Experimental results using a plethora of state-of-the-art object detection methods demonstrate the challenging nature of the proposed dataset and provide a baseline for the development of accurate and robust grape segmentation and maturity estimation algorithms in the future, thus laying the groundwork for significant advances in the field of viticulture. In future work, we aim to employ the proposed dataset, as well as other indicators (e.g., weather data), to develop automatic methods for predicting grape diseases at the early stages of infection, thus contributing to grape quality control and safeguarding measures.

Author Contributions: Conceptualization, V.S., D.K. and K.D.; software, A.B. and T.C.; formal analysis, A.B., K.C., M.K. and T.C.; resources, A.B. and T.C.; data curation, A.B., K.C., M.K. and T.C.; writing—original draft preparation, A.B., K.C., M.K. and T.C.; writing—review and editing, V.S., D.K. and K.D.; visualization, A.B., K.C. and T.C.; supervision, K.D.; project administration, K.D.; funding acquisition, K.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the General Secretariat for Research and Technology under grant agreement No. KMP6-0284680 "GraDA: Development of a molecular and environmental data collection system in order to ensure quality in the value chain of table grapes".

Data Availability Statement: The grape dataset will be made publicly available after the publication of this work.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Liu, S.; Whitty, M. Automatic grape bunch detection in vineyards with an SVM classifier. *J. Appl. Log.* **2015**, *13*, 643–653. [CrossRef]
2. Bruni, V.; Dominijanni, G.; Vitulano, D. A Machine-Learning Approach for Automatic Grape-Bunch Detection Based on Opponent Colors. *Sustainability* **2023**, *15*, 4341. [CrossRef]
3. Behroozi-Khazaei, N.; Maleki, M.R. A robust algorithm based on color features for grape cluster segmentation. *Comput. Electron. Agric.* **2017**, *142*, 41–49. [CrossRef]
4. Aquino, A.; Diago, M.P.; Millán, B.; Tardáguila, J. A new methodology for estimating the grapevine-berry number per cluster using image analysis. *Biosyst. Eng.* **2017**, *156*, 80–95 . [CrossRef]
5. Aquino, A.; Millan, B.; Diago, M.P.; Tardaguila, J. Automated early yield prediction in vineyards from on-the-go image acquisition. *Comput. Electron. Agric.* **2018**, *144*, 26–36. [CrossRef]

6. Millan, B.; Velasco-Forero, S.; Aquino, A.; Tardaguila, J. On-the-go grapevine yield estimation using image analysis and boolean model. *J. Sensors* **2018**, *2018*, 1–14. [CrossRef]
7. Liu, S.; Zeng, X.; Whitty, M. A vision-based robust grape berry counting algorithm for fast calibration-free bunch weight estimation in the field. *Comput. Electron. Agric.* **2020**, *173*, 105360. [CrossRef]
8. Mohimont, L.; Alin, F.; Rondeau, M.; Gaveau, N.; Steffenel, L.A. Computer Vision and Deep Learning for Precision Viticulture. *Agronomy* **2022**, *12*, 2463. [CrossRef]
9. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015.
10. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *arXiv* **2016**, arXiv:1506.01497
11. He, K.; Gkioxari, G.; Dollar, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017.
12. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* **2016**, arXiv:1506.02640.
13. Santos, T.T.; de Souza, L.L.; dos Santos, A.A.; Avila, S. Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. *Comput. Electron. Agric.* **2020**, *170*, 105247. [CrossRef]
14. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
15. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
16. Sozzi, M.; Cantalamessa, S.; Cogato, A.; Kayad, A.; Marinello, F. Grape yield spatial variability assessment using YOLOv4 object detection algorithm. In *Precision Agriculture'21*; Wageningen Academic Publishers: Wageningen, The Netherlands, 2021; pp. 193–198.
17. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
18. Sozzi, M.; Cantalamessa, S.; Cogato, A.; Kayad, A.; Marinello, F. Automatic bunch detection in white grape varieties using YOLOv3, YOLOv4, and YOLOv5 deep learning algorithms. *Agronomy* **2022**, *12*, 319. [CrossRef]
19. Zhang, C.; Ding, H.; Shi, Q.; Wang, Y. Grape Cluster Real-Time Detection in Complex Natural Scenes Based on YOLOv5s Deep Learning Network. *Agriculture* **2022**, *12*, 1242. [CrossRef]
20. Pinheiro, I.; Moreira, G.; Queirós da Silva, D.; Magalhães, S.; Valente, A.; Moura Oliveira, P.; Cunha, M.; Santos, F. Deep Learning YOLO-Based Solution for Grape Bunch Detection and Assessment of Biophysical Lesions. *Agronomy* **2023**, *13*, 1120. [CrossRef]
21. Shen, L.; Chen, S.; Mi, Z.; Su, J.; Huang, R.; Song, Y.; Fang, Y.; Su, B. Identifying veraison process of colored wine grapes in field conditions combining deep learning and image analysis. *Comput. Electron. Agric.* **2022**, *200*, 107268. [CrossRef]
22. Chen, Y.; Li, X.; Jia, M.; Li, J.; Hu, T.; Luo, J. Instance Segmentation and Number Counting of Grape Berry Images Based on Deep Learning. *Appl. Sci.* **2023** *13*, 6751. [CrossRef]
23. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS–improving object detection with one line of code. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5561–5569.
24. Lu, S.; Liu, X.; He, Z.; Zhang, X.; Liu, W.; Karkee, M. Swin-Transformer-YOLOv5 for Real-Time Wine Grape Bunch Detection. *Remote Sens.* **2022**, *14*, 5853. [CrossRef]
25. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.
26. Schorn-García, D.; Giussani, B.; García-Casas, M.J.; Rico, D.; Martin-Diana, A.B.; Aceña, L.; Busto, O.; Boqué, R.; Mestres, M. Assessment of Variability Sources in Grape Ripening Parameters by Using FTIR and Multivariate Modelling. *Foods* **2023**, *12*, 962. [CrossRef]
27. Seng, K.P.; Ang, L.M.; Schmidtke, L.M.; Rogiers, S.Y. Computer Vision and Machine Learning for Viticulture Technology. *IEEE Access* **2018**, *6*, 67494–67510. [CrossRef]
28. Zuñiga, A.; Mora, M.; Oyarce, M.; Fredes, C. Grape maturity estimation based on seed images and neural networks. *Eng. Appl. Artif. Intell.* **2014**, *35*, 95–104. [CrossRef]
29. Bazinas, C.; Vrochidou, E.; Kalampokas, T.; Karampatea, A.; Kaburlasos, V.G. A Non-Destructive Method for Grape Ripeness Estimation Using Intervals Numbers (INs) Techniques. *Agronomy* **2022**, *12*, 1564. [CrossRef]
30. Cavallo, D.P.; Cefola, M.; Pace, B.; Logrieco, A.F.; Attolico, G. Non-destructive and contactless quality evaluation of table grapes by a computer vision system. *Comput. Electron. Agric.* **2019**, *156*, 558–564. [CrossRef]
31. Hernández, S.; Morales, L.; Urrutia, A. Unsupervised Learning for Ripeness Estimation From Grape Seeds Images. *Int. J. Smart Sens. Intell. Syst.* **2017**, *10*, 1–19. [CrossRef]
32. Xinguang, W.; Wu, L.; Ge, D.; Yao, M.; Bai, Y. Prediction of the Maturity of Greenhouse Grapes Based on Imaging Technology. *Plant Phenomics* **2022**, *2022*, 1–14. [CrossRef]
33. Ramos, R.P.; Gomes, J.S.; Prates, R.M.; Simas Filho, E.F.; Teruel, B.J.; dos Santos Costa, D. Non-invasive setup for grape maturation classification using deep learning. *J. Sci. Food Agric.* **2021**, *101*, 2042–2051.
34. Qiu, C.; Tian, G.; Zhao, J.; Liu, Q.; Xie, S.; Zheng, K. Grape Maturity Detection and Visual Pre-Positioning Based on Improved YOLOv4. *Electronics* **2022**, *11*, 2677. [CrossRef]

35. Li, Y.; Wang, Y.; Xu, D.; Zhang, J.; Wen, J. An Improved Mask RCNN Model for Segmentation of (Vitis labruscana) Grape Bunch and Detection of Its Maturity Level. *Agriculture* **2023**, *13*, 914. [CrossRef]
36. Kangune, K.; Kulkarni, V.; Kosamkar, P. Grapes Ripeness Estimation using Convolutional Neural network and Support Vector Machine. In Proceedings of the 2019 Global Conference for Advancement in Technology (GCAT), Bangaluru, India, 18–20 October 2019; pp. 1–5. [CrossRef]
37. Gomes, V.M.; Fernandes, A.M.; Faia, A.; Melo-Pinto, P. Comparison of different approaches for the prediction of sugar content in new vintages of whole Port wine grape berries using hyperspectral imaging. *Comput. Electron. Agric.* **2017**, *140*, 244–254. [CrossRef]
38. Gomes, V.; Mendes-Ferreira, A.; Melo-Pinto, P. Application of Hyperspectral Imaging and Deep Learning for Robust Prediction of Sugar and pH Levels in Wine Grape Berries. *Sensors* **2021**, *21*, 3459. [CrossRef] [PubMed]
39. Silva, R.; Melo-Pinto, P. A review of different dimensionality reduction methods for the prediction of sugar content from hyperspectral images of wine grape berries. *Appl. Soft Comput.* **2021**, *113*, 107889. [CrossRef]
40. Navarro, P.J.; Miller, L.; Díaz-Galián, M.V.; Gila-Navarro, A.; Aguila, D.J.; Egea-Cortines, M. A novel ground truth multispectral image dataset with weight, anthocyanins, and Brix index measures of grape berries tested for its utility in machine learning pipelines. *GigaScience* **2022**, *11*, giac052. [CrossRef] [PubMed]
41. Kasimati, A.; Espejo-García, B.; Darra, N.; Fountas, S. Predicting Grape Sugar Content under Quality Attributes Using Normalized Difference Vegetation Index Data and Automated Machine Learning. *Sensors* **2022**, *22*, 3249. [CrossRef] [PubMed]
42. Armstrong, C.E.; Gilmore, A.M.; Boss, P.K.; Pagay, V.; Jeffery, D.W. Machine learning for classifying and predicting grape maturity indices using absorbance and fluorescence spectra. *Food Chem.* **2023**, *403*, 134321. [CrossRef]
43. Kalopesa, E.; Karyotis, K.; Tziolas, N.; Tsakiridis, N.; Samarinas, N.; Zalidis, G. Estimation of Sugar Content in Wine Grapes via In Situ VNIR-SWIR Point Spectroscopy Using Explainable Artificial Intelligence Techniques. *Sensors* **2023**, *23*, 1065. [CrossRef] [PubMed]
44. Santos, T.; de Souza, L.; dos Santos Andreza; Avila, S. Embrapa Wine Grape Instance Segmentation Dataset—Embrapa WGISD Zenodo. 2019. The Building of the WGISD Dataset was Supported by the Embrapa SEG Project 01.14.09.001.05.04, Image- Based Metrology for Precision Agriculture and Phenotyping, and the CNPq PIBIC Program (Grants 161165/2017-6 and 125044/2018-6). Available online: https://zenodo.org/record/3361736 (accessed on 27 July 2023). [CrossRef]
45. Morros, J.R.; Lobo, T.P.; Salmeron-Majadas, S.; Villazan, J.; Merino, D.; Antunes, A.; Datcu, M.; Karmakar, C.; Guerra, E.; Pantazi, D.A.; et al. AI4Agriculture Grape Dataset. Zenodo. 2021. Available online: https://zenodo.org/record/5660081 (accessed on 27 July 2023). [CrossRef]
46. Sozzi, M.; Cantalamessa, S.; Cogato, A.; Kayad, A.; Marinello, F. wGrapeUNIPD-DL: An open dataset for white grape bunch detection. *Data Brief* **2022**, *43*, 108466. [CrossRef]
47. Barbole, D.K.; Jadhav, P.M. GrapesNet: Indian RGB & RGB-D vineyard image datasets for deep learning applications. *Data Brief* **2023**, *48*, 109100. [CrossRef]
48. Sofiiuk, K.; Petrov, I.A.; Konushin, A. Reviving Iterative Training with Mask Guidance for Interactive Segmentation. *arXiv* **2021**, arXiv:2102.06583.
49. Chen, K.; Wang, J.; Pang, J.; Cao, Y.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Xu, J.; et al. MMDetection: Open MMLab Detection Toolbox and Benchmark. *arXiv* **2019**, arXiv:1906.07155.
50. Jocher, G.; Chaurasia, A.; Qiu, J. YOLO by Ultralytics. 2023. Available online: https://github.com/ultralytics/ultralytics/ (accessed on 15 May 2023).
51. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
52. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
53. Cai, Z.; Vasconcelos, N. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *arXiv* **2019**, arXiv:1906.09756.
54. Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. Hybrid Task Cascade for Instance Segmentation. *arXiv* **2019**, arXiv:1901.07518.
55. Cheng, B.; Misra, I.; Schwing, A.G.; Kirillov, A.; Girdhar, R. Masked-attention Mask Transformer for Universal Image Segmentation. *arXiv* **2022**, arXiv:2112.01527.
56. Su, S.; Chen, R.; Fang, X.; Zhu, Y.; Zhang, T.; Xu, Z. A Novel Lightweight Grape Detection Method. *Agriculture* **2022**, *12*, 1364. [CrossRef]