GIMA
Geographical Information Management and Applications

Counting Apples: Deep Learning-based Fruit Yield Estimation from High-Resolution UAV Imagery

Author: Haris Gkikas

Supervisors: Chenglong Zhang, Lammert Kooistra & João Valente

Responsible Professor: Ron van Lammeren

Universiteit Utrecht
TU Delft
University of Twente.
Wageningen UR

Abstract

Yield prediction is crucial for optimizing apple orchards. Recent research on the application of unmanned aerial vehicles (UAVs) remote sensing and convolutional neural networks (CNNs) object detection techniques, demonstrated a great potential for improved yield estimations. However, several major challenges exist in CNN-based yield estimation in orchards using UAV platforms, including illumination variance, occlusion conditions and the small scale of the fruits — as appear in aerial scenery. In addition, the UAVs data-acquisition ability is hampered by various factors, including exposure times, environmental conditions and sensor-related limitations, which can introduce blurriness and other optical distortions in the obtained images. In aim to overcome these challenges, this thesis deploys a single image super-resolution (SISR) method based on a generative adversarial network (GAN), for the enhancement of UAV images prior to the CNN-based detection of the fruits. In specific, the Real-ESRGAN was applied, due to the high perceptual accuracy it offers, ease of use, and low possibility of artifacts generation. To test the proposed method, a novel RGB UAV dataset was constructed. For the evaluation phase, image quality metrics were used, followed by a fruit detection comparison between two YOLOv5-based detectors — one trained on the super-resolved dataset and the other on the original. Results showed the effectiveness of the proposed method, where the detection rates for Yolov5 trained and employed in the super-resolved dataset, increased by 7.06% for precision, 30.77% for recall, and 18.92% for F1 score, compared to the YOLOv5 trained and employed on the original non-enhanced dataset. Moreover, the scores on image quality metrics showed that the proposed method can effectively reconstruct problematic UAV datasets, in comparison with other SISR methods. Concluding, enhancing the dataset with a SISR network can result to higher detection rates.

Keywords: yield estimation; deep learning; object detection; UAV; super-resolution

Acknowledgements

I would like to express my deepest gratitude to my supervisors, Chenglong Zhang, Dr. ir. Lammert Kooistra and Dr. João Valente, for their guidance. The insights, suggestions and encouragement provided by Chenglong, were invaluable and deeply appreciated. He taught me skills that I can carry with me in my professional career. The constructive feedback and directions provided by Dr. ir. Lammert Kooistra, were vital and building blocks to my critical thinking. The knowledge provided by Dr. João Valente, helped me to challenge my ideas, and spark new ones.

Τέλος, θα ήθελα να ευχαριστήσω ιδιαιτέρως τους φίλους μου, οι οποίοι υπήρξαν ένα ανεκτίμητο στήριγμα για εμένα, και τους γονείς μου, των οποίων οι κόποι και οι θυσίες θα με συνοδεύουν πάντοτε.

Haris Gkikas, June 2022

Contents

List of Figures

List of Tables

Nomenclature

Abbreviations

Abbreviation	Definition
CNN	Convolutional Neural Network
DL	Deep Learning
DNN	Deep Neural Network
ESRGAN	Enhanced Super-Resolution Generative Adversarial Network
GAN	Generative Adversarial Network
HR	Higher-Resolution
IoU	Intersection over Union
k-NN	k-Nearest Neighbors
LR	Lower-Resolution
MISR	Multi Image Super-Resolution
NIQE	Naturalness Image Quality Evaluator
PI	Perceptual Index
PIQE	Perception based Image Quality Evaluator
PSNR	Peak Signal-to-Noise Ratio
R-CNN	Region-based Convolutional Neural Network
ReLU	Rectified Linear Unit
RRDB	Residual-in-Residual Dense Blocks
SR	Super-Resolution
SRCNN	Single image Super-Resolution Convolutional Neural Network
SRGAN	Super-resolution Generative Adversarial Network
SSIM	Structural Similarity Index
SISR	Single Image Super-Resolution
SSD	Single-Shot MultiBox Detector
SVM	Support-Vector Machine
UAV	Unmanned Aerial Vehicle
UGV	Unmanned Ground Vehicle
YOLO	You Only Look Once

1
Introduction

1.1. Background

Obtaining accurate yield estimations is considered an elemental part of modern agriculture, allowing farmers to efficiently allocate natural resources, confidently negotiate pricing, schedule labour and timely take decisions. Currently, manual yield estimation techniques are the industry standard, typically performed based on historical data, weather conditions, and manual counting in multi-sampling locations (Q. Wang et al., 2012). These techniques are identified as labour intensive, inefficient and expensive (Kanwal et al., 2019). Algorithmic advances and the availability of low-cost computational resources, enabled convolutional neural networks (CNNs) to emerge as accurate and reliable detection systems for automatic fruit yield estimation. Compared to traditional machine learning approaches, CNNs achieve state-of-the-art performance on the detection and localization of apple fruits (Guo et al., 2016; Wang & He, 2021).

Thanks to the miniaturization of sensors and stimulated by the demand for inexpensive information-rich imagery, unmanned aerial vehicles (UAVs), commonly referred as drones, have sparked the interest as preferable imaging systems (Apolo-Apolo et al., 2020; Mital, Singh & Sharma, 2020). The major benefit of UAVs is their ability to acquire imagery of high spatial resolution in a timely manner. Compared to ground-sensing platforms, e.g. unmanned ground vehicles (UGVs) and human-mounted devices, UAVs not only show outstanding data-acquisition speed, but exceptional mobility and manoeuvrability, as terrain independent units. Unlike other aerial platforms, light-weight multi-rotor UAVs are considerably affordable and can hover over the point of interest.

However, several major challenges exist in CNN-assisted yield estimation using UAV platforms. First, the proportion of fruits visually available for detection is often limited, largely depended on pruning practises and the density of foliage. Even in more transparent foliage structures, apples are usually occluded by various obstacles, such as branches, leaves and other apples (Liu et al., 2019; Apolo-Apolo et al., 2020; Gené-Mola et al., 2021). Second, apples appear small in UAV-based scenes, frequently occupying less than 1% percent of the total image area. Third, UAVs data acquisition ability is sensitive to various internal, i.e. sensor limitations, exposure times etc., and external factors, i.e. wind, fog, rain etc., often introducing various distortions and artifacts. Moreover, due to limited battery duration, the coverage of large areas in short time is requiring a flight of higher altitude. This results to reduced spatial resolution, and thus a loss in essential information captured in the imagery. All these issues hamper significantly the precise detection and localization of apples, challenging the exact per tree count of fruits.

Many authors address these issues by either performing various modifications on the CNNs architectures, or using a combination of sensors — which can be proven expensive, while few studies focus on enhacning the primary data prior to the detection. The recent years, super-resolution (SR) techniques based on neural networks, have gained increased attention within the field of remote sensing (González et al., 2019; Lei et al., 2020). Research on their use, both for satellite and aerial platforms, showed excellent results on increasing the initial spatial resolution of imaging systems (Ferdous et al., 2019; Gonzalez et al., 2019; Courtrai et al., 2020; Lei et al., 2020; Pashaei et al., 2020). Especially in scenarios, where the datasets suffer from several distortion and the re-acquisition can be proven challenging, expensive and/or time consuming. Recently, deep learning-based SR techniques, dominated other SR approaches, providing remarkable achievements on benchmark datasets (Chen et al., 2022). However, the evaluation and exploration of deep learning SR techniques as assistance in UAV-based apple yield estimation, is urgently needed.

In this paper, aiming to tackle the challenges of object detection on UAV apple datasets, a super-resolution augmentation method based on generative adversarial networks (GAN) is implemented and evaluated, both in terms of image reconstruction quality and as an aid to the CNN-based detection. Furthermore, YOLOv5-based detectors are trained on the reconstructed super-resolved and original datasets, respectively, and tested based on their detection performance, aiming to evaluate the success of the proposed method.

1.2. Research Questions

Based on the previous statements, the following main research question can be drawn:

- **Can apple detection on UAV datasets by means of DL, benefit from SR data enhancement?**

Based on the main research question, three sub-questions are constructed:

- What is the process to create SR images, to be used effectively in object detection?
- Which SR method produce images of higher quality?
- Is object detection based on the SR-enhanced dataset superior in terms of precision, recall and F1 score, compared to the non-SR detection?

1.3. Research Goal

The goal of this thesis is to explore the capability of SR techniques in combination with a deep learning object detection algorithm, for apple yield estimation. The demand for computational resources and time will be discussed, to determine if and how deep learning-based image enhancement can be improve the object detection accuracy, not only in apple orchard management but in the general field of remote sensing.

2

Related Work

2.1. Convolutional Neural Networks (CNNs) for Apple Detection

Any automatic yield estimation process begins with the precise detection and localization of fruits. Over the years, researchers have used a variety of machine vision systems and sensors for fruit detection. Early works utilized the binary space, i.e. black-and-white (B/W), for image processing, detecting fruits based on their shapes and textures (Cardenas-Weber et al., 1991; Edan et al., 2000). However, the lack of colour information, one of the most prominent characteristics of fruits, led to major disadvantages. Soon researchers started to utilize the RGB space for implementing a variety of computer vision algorithms, including k-nearest neighbors (k-NN), Otsu's thresholding and support-vector machines (SVMs) (Stern et al., 2010; Linker et al., 2012). However these techniques built upon identifying simple image features such as colours and edges, while missing essential features such as pixel correlation and the spatial position of the fruits (Li et al., 2021). These advanced features are essential for robust defections under the varying conditions found on apple orchards, e.g. high illumination variance, different levels of occlusion etc..

Deep neural networks (DNNs) based learning, or commonly called deep learning (DL), became the mainstream approach in fruit detection and localization. They utilize raw data to automatically discover patterns without human intervention, based on a prior training process, where are asked to learn the mappings of the input to the intended output (LeCun et al., 2015). This training process can be performed based on a supervised or unsupervised way, where the former is considered the most common for object detection. As the name suggest, the DL model requires direct supervision, where a set of examples by a human user, i.e. the training set, is provided to the algorithm during the training phase. Each example contains an annotation, i.e. label, which indicates the unique features and the location of the object of interest, in respect to the total image space. Eventually, the model learns and predicts new annotations to previously unseen datasets. Currently, five fundamental deep learning architectures have been proposed and implemented successfully, identified as: stacked autoencoders (SAEs), deep belief networks (DBNs), recurrent neural networks (RNNs), convolutional neural networks (CNNs) and generative adversarial networks (GANs) (Xia, 2019). Among them, the CNNs are recognised as the most successful for apple detection and localization (Apolo-Apolo et al., 2020; Li et al., 2022; Yan et al., 2021).

Inspired by the structure of the mammalian visual cortex, convolutional neural networks (CNNs) are the most utilized architecture in the field of modern computer vision (Pang & Cao, 2019), achieving state-of-the-art performances in object detection, image classification, image segmentation and enhancement (Xia, 2019). CNNs for object detection, have been applied to a vast variety of domains, from remote sensing (L. Zhang et al., 2016), to medical imaging (Litjens et al., 2017).

The basic structure of a typical CNN is composed by the convolutional layers, the pooling layers, and the fully-connected layers. The convolutional layers are regarded as the core elements of CNNs, and described as the responsible mechanism for extracting the image features. In specific, the these layers convolve the input image with one or multiple image filters, referred as kernels, producing a new representation of the input image, i.e. feature maps. By utilizing different types of kernels, different outputs are obtained. In most CNN architectures, the first convolutional layers are responsible for extracting features like edges and textures. Figure 2.1, depicts visual examples of feature maps, prior to the last pooling operation.

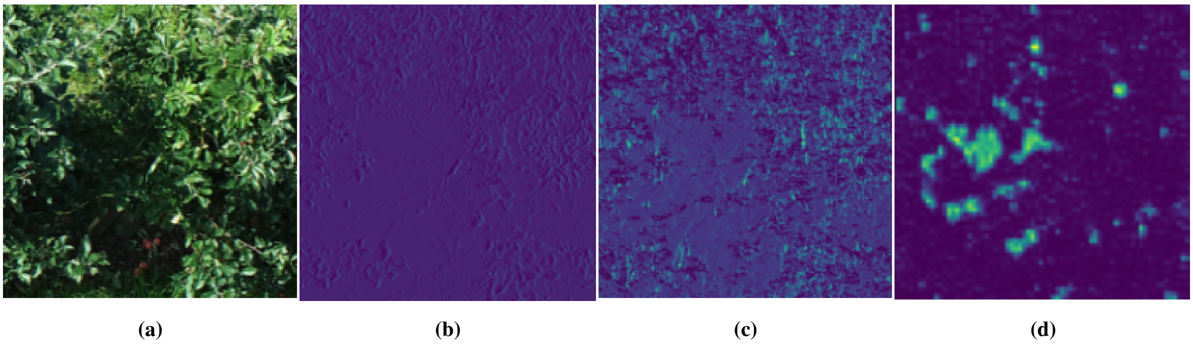


Figure 2.1: Feature maps from a YOLO detector. From left to right: a) is the input image, b) a feature map which appears to capture edges, c) a feature which appears to capture vertical gradients, and d) a feature map, which appears to capture large round-shape areas.

Following the convolutional layers, the pooling layers are summarizing the feature maps as collected by the kernels, by reducing their spatial dimensionality, forming a collection of extracted features. After multiple consecutive convolutional and pooling layers, the feature map of the last pooling layer, —i.e. the total collection of the extracted features—, is fed to the fully-connected layers, that are responsible to generate a probability where and what type of objects are present on the image. During the training phase, the fully-connected layers assign weights to the extracted image features, i.e. —which image features are more relevant for a specific class of object, for example round gradients and red color to apples–, and biases, that ensure even if the object of interest is not present, the CNN can predict that no object is present.

Numerous CNN-based detectors have been proposed by the DL community, categorized to two main classes, as: (1) two-stage method method and (2) one-stage method. The family of two-stage method detectors approach object detection as a two-step process. In the first step, regions proposals are generated with the possible object of interest, and then these regions are given as inputs to the trained fully-connected layers. Popular examples of two-stage detectors include: R-CNN (Girshick et al., 2013), Fast R-CNN (Girshick, 2015) and Faster R-CNN (Ren et al., 2015). Research on apple detection showed that two-stage detectors are capable of outstanding performance. In specific, Bargoti and Underwood (2017) proposed a crop yield estimation pipeline for mangoes and apples, based on a ground-based 360° RGB camera and a Faster R-CNN. They achieved high detection performance, with precision scores of a range 93.3% to 95.8%. Apolo-Apolo et al. (2020), utilized an RGB orthomosaic, constructed by UAV imagery, and implemented a fine-tuned Faster R-CNN model to detect apples with on-average precision of 93%. However, it is widely recognized that two-stage detectors show relative slow detection speed (Pang & Cao, 2019).

Different than two-stage detectors, one-stage detectors simultaneously predict the class and location of possible objects. Popular examples, include the single shot multibox (SSD)(W. Liu et al., 2015) and the you only look once (YOLO) (Redmon et al., 2016) algorithms, where in comparison with R-CNN, Fast R-CNN and Faster R-CNN, are much more simpler and fast, achieving comparable detection performance on common object detection benchmark datasets. Their ability for high-level detection performance and speeds can be observed in a variety of research works. Although not concerning apple detection, but UAV datasets, Benjdira et al. (2018) by comparing Faster R-CNN and YOLOv3 for car detection on UAV dataset, demonstrated that YOLOv3 outperforms Faster R-CNN in sensitivity and processing time, although they are comparable in the precision metric. Morevoer, Zhao et al. (2021) in their work concerning wheat spike detection from UAV imagery, showed that an improved version of the YOLOv5 algorithm can achieve state-of-the-art results compared to various two-stage and one-stage detectors. In addition, Zhu et al. (2021), using a modified YOLOv5, achieved remarkable results on the VisDrone 2020, a benchmark dataset containing UAV imagery with small objects. The evidence suggest that new generations of one-stage detectors, such as modified versions of the YOLOv5 algorithm, can achieve state-of-the-art results in UAV-based small objects detection. For this reason, the YOLOv5 was chosen as the method of preference for the detection purposes of this research.

2.1.1. YOLOv5 for Apple Detection

The YOLOv5, instead of picking region proposals, splits the image in a k x k grid, thus achieving extremely fast detection speeds–being up to 140 frames per second (Ya et al., 2021). Compared to other CNN-based detectors, e.g. Faster R-CNN, the YOLOv5 weight file is small —around 3.7M for the smallest variation, indicating that YOLOv5 is an ideal model for deployment in embedded devices for real-time detection purposes, e.g. fruit-picking robots and UAVs (Wang & He., 2021). In addition, the YOLOv5 models are considered very efficient in multi-scale prediction, enabling the algorithm to handle various sizes of objects (Yang et al., 2022). As discussed previously, apples usually vary greatly in size as seen in UAV imagery, therefore multi-scale detection ensures that the detector model can mitigate the changes shapes and sizes. This fact, in addition with the state-of-the-art performance in apple detection, compared with two-stage methods (Wang & He, 2021), plus the fast detection speed and the lightweight aspect, were considered as the main reasons for the selection of YOLOv5s for this research.

2.2. Generative adversarial networks (GANs) for Super-Resolution

2.2.1. Image Super-Resolution

The general term super-resolution (SR) describes a class of techniques aiming to reconstruct an image of lower resolution (LR), to an improved version of higher resolution (HR). Such techniques, not only increase the number of pixels of an image—which can be also achieved by simple resizing, but also maintain the semantic information contained in the image or even enhance them. The following figure depicts a visual example of this basis, where the LR input Figure 2.2a can be resized to Figure 2.2b, resulting to considerable increase in the number of pixels, but compared to the SR reconstruction Figure 2.2c, lacks essential details, such as edges and fine-textures.

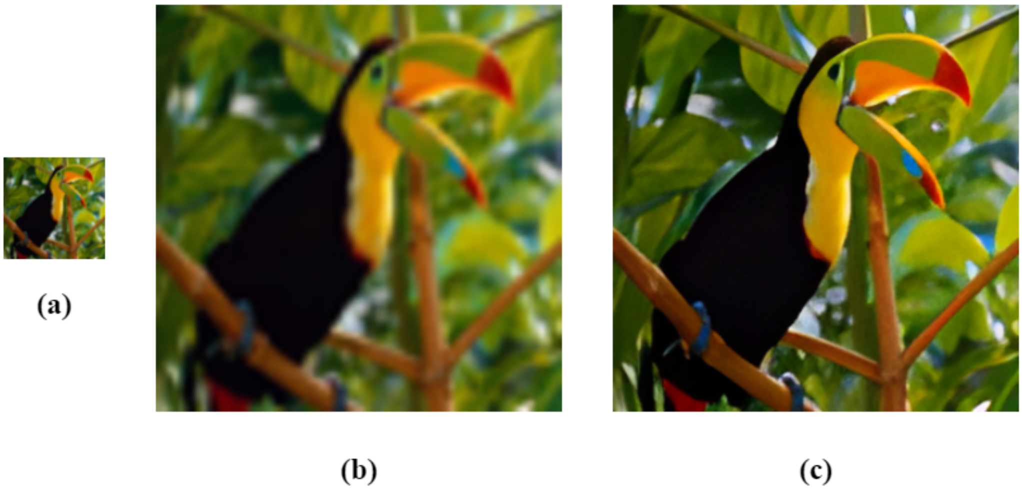


Figure 2.2: Input 72 x 72 px image (a), the 'Bird' from Set5 (Bevilacqua et al., 2012), as 288 x 288 px output, of: resizing (b), and super-resolution (c).

In principle, SR frameworks lie foundation on the assumption that a LR image I_{LR} is a reduced version of the HR image I_{HR}, both in terms of pixel and information quantity, modelled as Equation 2.1:

$$I_{LR} = D\,(I_{HR}, \partial), \tag{2.1}$$

where, D denotes the degradation function responsible for reducing the HR image I_{HR} to the LR version I_{LR}, and ∂ depicts the input parameters of the degradation function. Under this context, the SR operation g is equivalent to the inverse of degradation function D^{-1}, modelled as Equation 2.2:

$$g(I_{LR}, \delta) = I_{SR} \approx I_{HR}, \tag{2.2}$$

where, I_{LR} is the LR, the input parameters of SR function g, and I_{SR} an approximation of the ideal HR image I_{HR}. Nevertheless, the exact degradation function is unobtainable, since only the LR version is given, ending to infinite possibilities of HR reconstructions and forming an extremely ill-posed problem (Bevilacqua et al. 2012; Pashaei et al. 2020).

Existing methods for SR are categorized into single-image super-resolution (SISR) and multi-image super-resolution (MISR), according to the different number of LR images utilized for the reconstruction of the HR,—a single image of the scene for SISR, multiple images of the scene for MISR (Clabaut et al., 2021; Li, Pei & Zeng., 2021). MISR methods are recognized to hold a significant advantage over SISR, on the basis of information provided for resolving the SR problem (Salvetti el al.,2020). However, acquiring multiple LR images of the objects of interest can be proven cost-inducing and ill-timed process, requiring extended computational resources and storage space (Pashaei et al., 2020). Additionally, in some cases, obtaining multiple LR images of the objects-of-interest is impossible. This research aims for high-performance but considerably timely and cost-effective automatic yield-estimation systems, as such, SISR suits as the approach of preference.

Recently, deep learning (DL) models dominated the field of SISR, leading to a staple of state-of-the-art results, surpassing previous methods, e.g. bicubic interpolation and Lanczos resampling — which use predefined mathematical formulas, especially when the up-scalling factor increases (Yang et al., 2018). As learning-based methods, DL models adapt and learn statistical relationships between HR and LR images, making the SR task an inference problem. Similar to object detection, supervised and unsupervised approaches exist for DL SISR training. Since the supervised models have gained a substantial attention the recent years by the research community, in addition to their recent state-of-the-art performance (Chen et al., 2022), are chosen as the approach aiming to be explored in this report.

2.2.2. GANs

Based on supervised training, a variety of DL SISR architectures have been proposed. Spearheading among them, Generative adversarial networks (GANs), is a promising class of deep learning frameworks, widely used in the field of computer visions for tasks such as super-resolution, object detection, image generation, image-to-image translation, image-to-text translation etc. (Alqahtani, Kavakli-Thorne & Kumar, 2019). Their basic architecture consists of two networks: the generator model and the discriminator model. The learning process of GANs can be thought as a two-player mini-max game, where in terms of image-generation, the generator model can be thought of as analogous to a team of counterfeiters, trying to produce fake currency, i.e. a synthetic copy of the original image, while the discriminative model is analogous to the police, trying to detect the counterfeit currency, i.e. to discriminate the synthetic image from the original (Goodfellow et al., 2014). In a typical GAN, the generator tries to create samples with similar distribution to the original data, using random noises $\mathbf{z}$ from a Gaussian distribution $p_z(\mathbf{z})$. During the learning process, the discriminator receives samples both from the original dataset and from the generator, i.e. the reconstructed dataset, trying to predict the probability whether the sample belongs to the original or the reconstructed dataset. In conjunction, the generator tries to produce realistic samples fooling the discriminator. This adversarial min-max problem between the generator and the discriminator , is introduced by Goodfellow et al. (2014), as function $V(G_\theta, D_\theta)$ (Eq. 2.3):

$$\min_{G_\theta} \max_{D_\theta} V(G_\theta, D_\theta) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D_\theta(x)] + \mathbb{E}_{z\sim p_z(z)}[1 - \log D_\theta(G_\theta(z))], \tag{2.3}$$

where, both G_θ and D_θ are trained simultaneously, and in the state, D_θ can not recognize the real from the synthetic data, the learning process has been achieved.

Representative examples of GANs architectures used for SISR, are the SRGAN (Ledig et al., 2016) and ESRGAN (X. Wang et al., 2018). The SRGAN utilizes: 1) a generator, which is composed by two convolutional layers of 3×3 kernels and 64 feature maps, followed by batch-normalization layers, with ParametricReLU as the activation function, and 2) a discriminator, which contains eight convolutional layers with an increasing number of 3 × 3 kernels, increasing by a factor of 2, starting from 64 to 512 kernels (Ledig et al., 2016). The ESRGAN improves on the original SRGAN, by modifying the generator in two parts: (1) removing all batch-normalization layers, and (2) introducing the concept of Residual-in-Residual Dense Blocks (RRDBs), which increases the capacity of the network, eases the training process, and improves the quality of the generated images (Wang et al., 2018), and by replacing standard discriminator, with the Relativistic average Discriminator (RaD), based on the discriminator proposed by Jolicoeur-Martineau (2018).

The use of ESRGAN in the field of remote sensing have been characterized as particular successful, being one of the most popular architectures utilized in the domain (Buddha et al., 2019; Burdziakowski, 2020; Ye et al., 2022). Examples include, the work of Pashaei et al. (2020), where by employing an ESRGAN on a UAV dataset, managed to construct higher-quality digital surface models (DSM) from lower-resolution images based on Structure from Motion (SfM) photogrammetry. In a further SISR experiment, Clabaut et al. (2021), confirmed that domain-specific ESRGANs can efficiently reconstruct satellite and aerial imagery. Furthermore, Rabi et al., (2020), revealed that by training an edge-enhance ESRGAN in an end-to-end manner with detector, an increase in the detection rates of small objects in satellite imagery can be achieved. However, the majority of these work are reporting results in artificially created lower-resolution space, and not directly in real-world source imagery. In an effort to bring the results of ESRGAN in real-life, various researchers utilize several pre-processing steps for the training of the models. A recent example, is the work of Velumani et al., (2021), in which they employed an image-to-image translation network for synthesizing training pairs with realistic degradations. By using these pairs to train an ESRGAN, managed to effectively enhance source RGB UAV datasets, and elevate the detection rates of maize species. Yet, such pre-processing steps can be proven time-consuming and unpractical to be used for precision agriculture purposes. Recently, the creators of ESRGAN, improved the original architecture to handle real-life degradations, for practical SISR restorations. The new network, by the name of Real-ESRGAN (X. Wang et al., 2021), achieved outstanding results in blind SISR image-restoration, —where the term blind refers to a model not trained on a given dataset, to effectively reconstruct it. Inspired by these results, the model chosen as the DL SISR method is the Real-ESRGAN.

2.2.3. GANs Learning Process & Degradation Modelling

Training of supervised GANs requires the source images of a given dataset to be artificially degraded in a lower-resolution counterpart, with a fixed degradation function, a process known as degradation modelling, as an approximation to the unknown ground-truth degradation function. By using these pairs, in this case denoted as I_{src} for the native source images and I_{LR} as the artificially reduced LR versions, the model can learn to recognize how various image features are expressed in the lower resolution space, based on the corresponding I_{src} samples. Based on this learning, the model can reconstruct an unseen native input image, —where the reconstruction is denoted as I_{SR}—, as close approximations of the I_{HR} image of the ideal higher-resolution space, as modelled in (Eq. 2.2).

Typical supervised deep learning SR frameworks, model the degradation function D as a combination of bicubic downsampling and Gaussian blur kernel k, as Equation 2.4:

$$D = (I_{src} \otimes k) \downarrow_s, \tag{2.4}$$

where, $\otimes$ is the convolution between the given higher-resolution image I_{src} and the blur kernel k_g, and $\downarrow_s$ the downsampling operation. This specific kind of degradation modelling, referred as classical degradation, poses significant problems. Assumptions that the I_{LR} is a product of a convolution and bicubic downsampling, does not correspond to real-life. To demonstrate the tremendous extent of potential degradations, known issues related with UAVs sensors, include: blurriness introduced by the motion of the platform, various radiometric and geometric distortions due to camera optics, chromatic aberrations, edge vignetting, artifacts introduced by JPEG compression etc. (Whitehead & Hugenholtz, 2014; Mittal, Singh & Sharma., 2020), without excluding a combination of all the aforementioned. Various researchers adopted more sophisticated degradation modelling processes, incorporating JPEG distortion and additive Gaussian or Poisson noise (A. Liu et al., 2021), as Equation 2.5:

$$D \;=\; [(I_{src} \otimes k) \downarrow_s + \mathbf{n}]_{JPEG}, \tag{2.5}$$

where, the additive noise **n**, the Gaussian blur kernel k, and JPEG compression are the primary degradation factors. A visual comparison between a classical degradation model and a sophisticated one, is given in Figure 2.3.

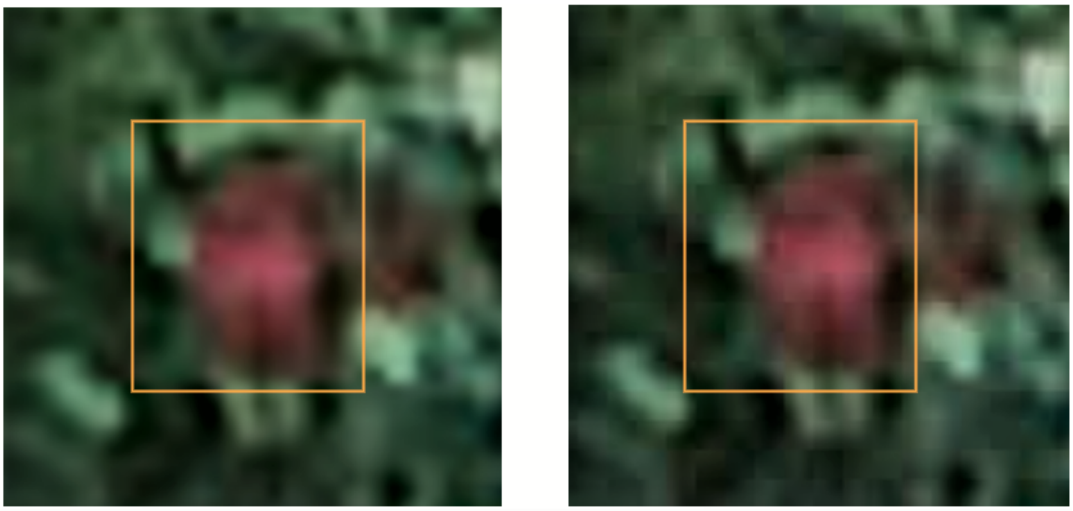


Figure 2.3: From left to right: An apple as product of classical degradation model, and as a product of a high-order degradation model. Zoom in the rectangles, to note the pixel-mosaic effect introduced by JPEG compression on the right image, in comparison with the smoothness of bicubic downsampling on the left image.

2.2.4. Performance Evaluation

Performance assessment of different SISR approaches is currently performed using image quality metrics (IQMs). In order to be considered appropriate, IQMs should stem on the fundamental definition of SR, —a process which effectively increases the size of a given image, while maintaining or enhancing the semantic information contained—, and evaluate the reconstructions based on: (a) the deviation in terms of image statistics in reference with the source image, e.g. expression of colors and position of shapes, referred as reconstruction accuracy, (b) the introduction of distortions and artifacts, as perceived by human viewers, referred as perceptual quality, and (c) the impact on the performance for the task that are utilized, e.g. object detection.

Following this triptych, various IQMs have been proposed for SISR, which are mainly divided into subjective methods, usually employing user studies, and objective methods, which rely solely on computation models (Z. Wang et al., 2004). While subjective methods are considered the most accurate, can be proven expensive, time-consuming and ineffective — as users are typically exposed to a limited number of methods and/or limited number of images per method (Blau et al. 2018). These observations led to the adoption of objective methods as the mainstream of SISR performance assessment. However in practice, most research works introduce visual examples between different SISR methods, to be compared by the readers.

Various objective IQMs have been proposed for SISR assessment, categorized to full-reference (FR), reduced-reference (RR), or no-reference (NR). Among these, FR-IQMS and NR-IQMs are mostly utilized in SISR research. FR-IQMs compare the SR reconstruction of the LR image, with the ground-truth HR image. Prime examples of FR-IQMs in SISR, are the mean square error (MSE), peak signal-to-noise ratio (PSNR) and structure similarity index (SSIM). In general, FR-IQMs can predict more easily the reconstruction accuracy, compared to RR-IQMs and NR-IQMs, since more reference information are available. However, are recognized as not optimal to predict the perceptual quality (Blau et al., 2018). Furthermore, as ground-truth HR images are not available in real-world scenarios, FR-IQMs are often used for dynamical monitoring the training process of DL SISR methods, assisting in hyper-parameterization. In comparison with FR-IQMs, NR-IQMs capture more efficiently unique characteristics of human perception, and are heavily utilized for capturing perceptual quality (Sajjadi et al., 2017; Blau et al., 2018; Haris et al., 2018; Bai et al., 2018). Usually, such approaches are not characterized by the terms FR/RR/NR-IQMs, thus will be presented separately

Task-based assessments utilize the SR reconstructions as image materials for a specific problem and evaluate their performance in comparison with the non-SR images. These techniques mainly focus on whether SR reconstructions can maintain or enhance the semantic information contained on the lower-resolution images. In the domain of remote sensing, the tasks used for evaluation can range from, e.g. assessing the deviation of photogrametrically-derived digital terrain models (DTM) from the ground-truth elevation via root-mean-square error (RMSE), to object detection performance evaluation via object detection metrics, e.g. F1 score. However, different tasks require different needs in terms of reconstruction accuracy or perceptual quality. For example, leaves disease classification is a task that requires high reconstruction accuracy (Wen et al., 2020), while monitoring land change in satellite imagery by simple visual investigation, requires sufficient perceptual quality. The task-based approach for the evaluation of this research is object detection performance.

3

Methods & Materials

3.1. Overview

This study aims to improve apple detection performance on UAV imagery. Toward this objective, the enhancement of UAV imagery prior to detection is proposed via a single image super-resolution (SISR) generative adversarial network (GAN). A prerequisite step for this task, is training the GAN module to match the distribution of the higher-resolution images, by providing the corresponding examples of lower-resolution images. However, in real-life applications, including this study, high-resolution images are unavailable. In addition, GAN methods usually require a considerable number of HR/LR training pairs. While this can be resolved though transfer-learning, —i.e. the technique that utilizes the knowledge gained on an older task, for the training of a new model in a new task —, the LR counterparts should resemble products from real-world degradations to achieve adequate SR performance.

To overcome these issues, the Real-ESRGAN model (X. Wang et al., 2021) is employed, which is capable to synthesize training pairs based on a sophisticated degradation modelling process —similar to real-world degradations. Furthermore, the weights from a model pre-trained on abundant high-quality natural images are used for the fine-tuning of a new model on a limited number of carefully selected distortion-free UAV images, acquired from the dataset. With the domain-trained Real-ESRGAN available, a part of the UAV dataset is super-resolved and used as training material for a YOLOv5s detector. For evaluation of the proposed method, non-reference image quality metrics (NR-IQMs) are utilized for the assessment of the reconstructions, followed by an apple detection comparison between two YOLOv5s models, trained on super-resolved and native lower-resolution datasets respectively (Figure 3.1). In the subsequent sections, information the datase, the architectures and the evaluation procedures are presented.

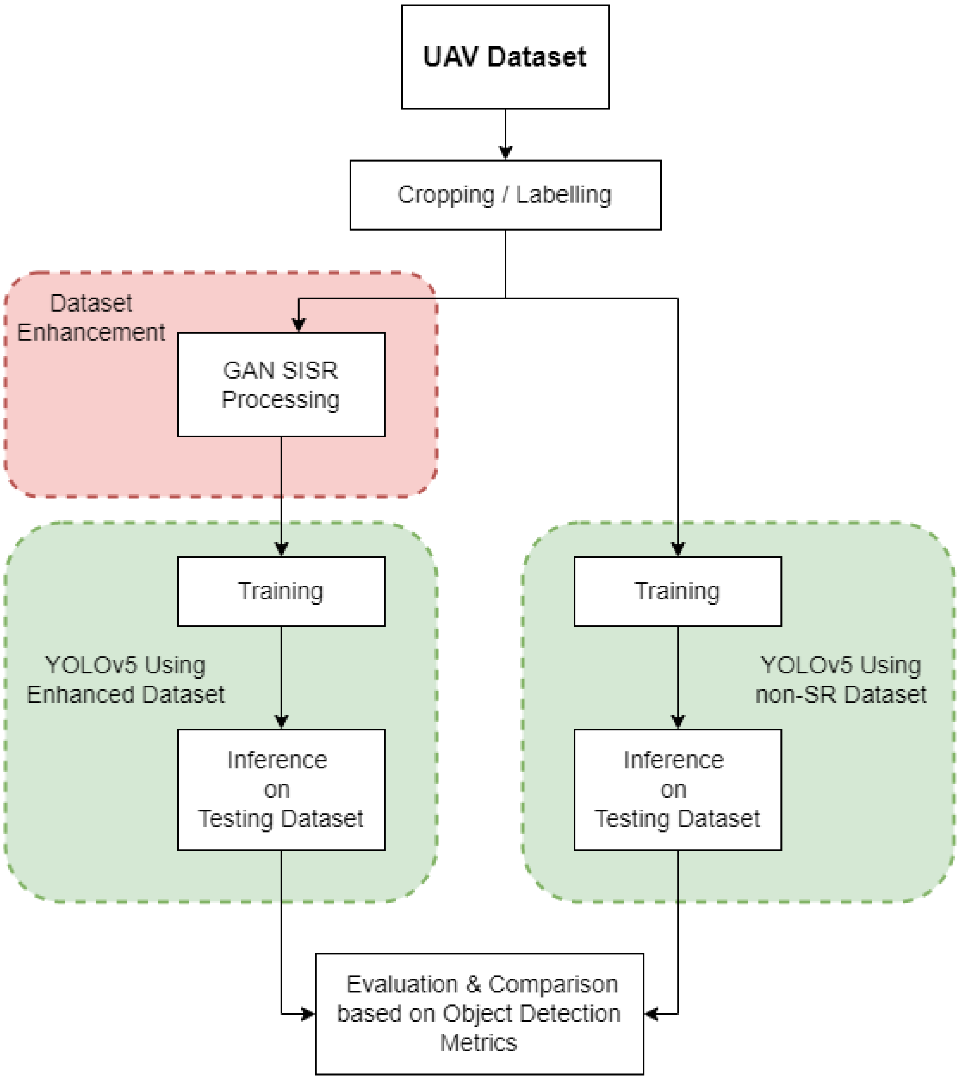

UAV Dataset
Cropping / Labelling
Dataset Enhancement
GAN SISR Processing
Training
YOLOv5 Using Enhanced Dataset
Inference on Testing Dataset
Training
YOLOv5 Using non-SR Dataset
Inference on Testing Dataset
Evaluation & Comparison based on Object Detection Metrics

Figure 3.1: The pipeline for the object detection evaluation of the study. First, the UAV dataset is cropped to tiles and annotated. Afterwards, the annotated images are duplicated, where the one duplicate is super-resolved, while the other is not. Final, two YOLOv5 detector are trained on the super-resolved and the original datasets, respectivelly, where their perfomance is evaluated based on obejct detection metrcis.

3.2. Study Site & Data Collection

The data acquisition was conducted in an apple orchard located in Randwijk, Netherlands (51°56'16.8" N, 5°42'24.5" E) (Figure 3.2a). The apple trees (*Malus domestica*) were planted in 2007, and belong to 'Elstar' variety. The orchard size is 0.47 ha, consisted of 14 rows with 55 trees per row, NW SE oriented. Each row is separated by 3m, and each tree within row by 1.1m. Average tree height is approximately 3 m. The orchard follows conventional farming practices. The UAV platform selected for the imagery acquisition was the quadrotor DJI Phantom 4 Pro with an embedded Real Time Kinematic (RTK) module (DJI Technology Co., Ltd., Shenzhen, China). The camera deployed was 1" CMOS sensor, with an effective pixel count of 20M, lens FOV of 84°, focal length of 8.8 mm and focal ratio of f/4.5 to f/11. To test the applicability of our proposed method under a variability of flight conditions, the UAV platform conducted various flights at different heights and camera angles. The overlapping setting was 80% for all flights and the lighting conditions were constant. Table 1 lists the flight parameters for each survey used in this study. The datasets acquired from each flight were mixed, forming a total of 435 RGB images of 5472 x 3648 px initial resolution, stored in JPEG format. The apples captured in the datasets, were often occluded (Figure 3.2b), while many images exhibited blurriness and other artifacts (Figure 3.2c). All these conditions were considered favourable for the testing of the proposed method.

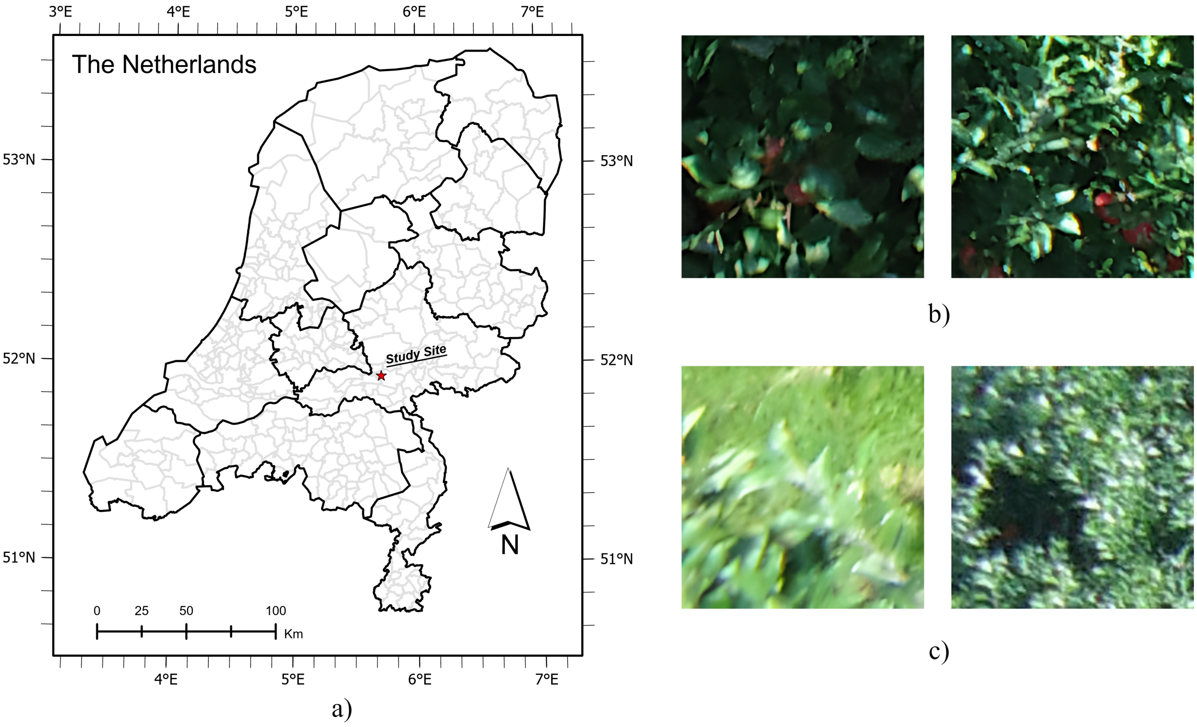


Figure 3.2: The study site in Randwijk, Netherlands (a). Examples of occluded fruits (b), by branches and leaves. Examples of blur images (c).

3.3. Implementation of Real-ESRGAN for Super-Resolution

3.3.1. Degradation Modelling

As explained in previous section (Section 2.2.3), DL SISR networks learning process is facilitated by synthesizing training pairs, where a high-resolution source image (I_{src}) is reduced to a low-resolution counterpart (I_{LR}). Several state-of-the-art SISR networks, e.g. ESRGAN (X. Wang et al., 2018), EDSR (Lim et al., 2017b), RCAN (Y. Zhang et al., 2018) etc., assume that I_{LR} are products of bicubic down-sampling. However, these approaches can lead to the introduction of artifacts which can hamper the detection process. To better model real-world degradations, the Real-ESRGAN used in this study, relies on synthesizing lower-resolution counterparts of the source images, by a sophisticated deterioration process by the name high-order degradation modelling (Eq. 3.1):

$$I_{LR} = D^n(I_{src}) = (D_n \times \cdots \times D_2 \times D_1)(I_{src}), \tag{3.1}$$

where, a n-order model involves n repeated classical deterioration models D (Eq. 2.5), where each D model incorporates blur kernels, added noise, random resizing and JPEG compression, but with different hyper-parameters each time. In addition, a novel *sinc* filter aiming to add edge overshoot and ringing artifacts, — common during digital transmission of the images (Hu et al., 2014) —, is introduced during the blurring stage and the last step of the process. The *sinc* filter kernel k is expressed as:

$$\mathbf{k}(i,j) \;=\; \frac{\omega_c}{2\pi\sqrt{i^2 \times j^2}}\mathrm{J}_1(\omega_c\sqrt{i^2 \times j^2}), \tag{3.2}$$

where (i, j) is the kernel coordinates; $omega_c$ is the cutoff frequency, and J_1 is the first order Bessel function. The *sinc* filters are employed in the blurring operation and the last step of the I_{LR} synthesis. The Real-ESRGAN modelling process was considered ideal for the case of UAVs, as these platforms suffer from a variety of distortions, which can degrade the acquired images in unpredicted and complicated ways.

3.3.2. Architecture

Generator: The Real-ESRGAN network employs the same generator with ESRGAN. In specific, the generator module is composed by a network of 23 residual-in-residual dense blocks (RRDBs), without batch normalization (BN). Each block is consisted by 5 convolutional layers, having 64 small kernels of 3 x 3 size, and Leaky ReLu is set as the activation function. In addition, aiming to reduce the demand for computational resources, the Real-ESRGAN network uses the pixel-unshuffle, an operation that reduces spatial size and increase channel size, before feeding the images to the generator. This operation allows the integration of up-scalling factors x2 and x1, in addition to x4, as found in the ESRGAN. Figure 3.3 depicts the structure of the generator network.

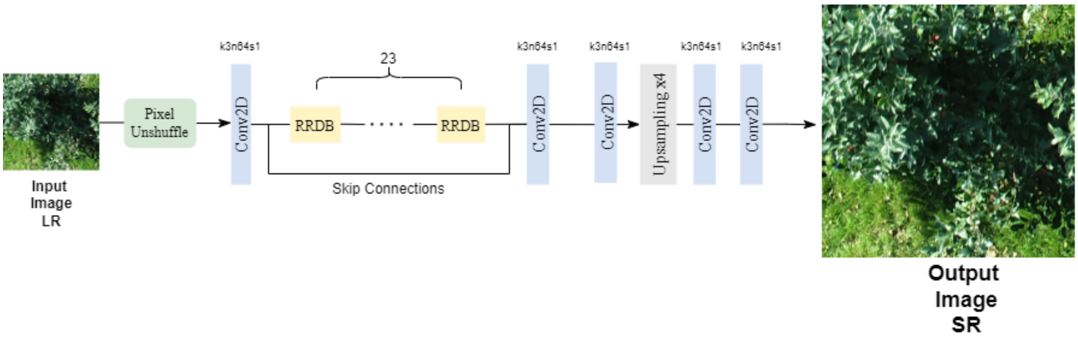


Figure 3.3: The architecture of the generator network, where in each convolutional layers, the k, n and s denote kernel size, number of feature maps and stride. For scale factors of x2 and x1, the pixel unshuffle operation is employed.

Discriminator: As the discriminator network, a U-Net architecture with spectral normalization (SN) regularization is utilized, with shortcut connections. The U-Net allows the network to be trained on LR samples with complex real-world degradations, while adding SN regularization stabilizes the training of GANs (X. Wang et al., 2021). The network is consisted by 10 convolutional layer, with various alternated kernel sizes, and Leaky ReLu is set as the activation function.

Both networks are trained to solve the min-max problem, as introduced in Equation 2.3, where the generator loss is set as:

$$L_{\mathbf{G}} \;=\; \mu L_{percep} + \kappa L_{gan-G} + \gamma L_1, \tag{3.3}$$

where, L_{percep} is the perceptual loss as proposed by Johnson et al.(2016), but before the activation layers, L_{gan-G} is the adversarial loss of the generator, and $L_1 = \mathbb{E}_{I_{LR}} \parallel G(I_{SR}) - I_{src} \parallel_1$ is the content-loss that evaluates the L_1 distance between the reconstruction I_{SR} of the synthesized I_{LR} and the source I_{source} image. For, further details about the total loss function, readers can review X. Wang et al. (2018).

3.3.3. Training

Training any GAN architecture from scratch is a timely and computational expensive process, despite the availability of processing power in the form of graphic processor units (GPUs). For this reason, the weights of pre-trained generator and discriminator networks on the DIV2K (Aggustsson and Timofte, 2017), Flickr2K (Lim et al., 2017a) and OutdoorSceneTraining (X. Wang et al., 2018) datasets were used (available on: https://github.com/xinntao/Real-ESRGAN). For the training of the model, 50 images 5472 × 3648 px of the UAV dataset were selected, and split to 608 x 608 px patches, resulting to a total of 2700 tiles. While various SISR networks utilize smaller sized patches for computational efficiency, it was found that very deep convolutional networks with wider receptive field tend to benefit from larger patches (Pashaei et al., 2020). In order to retrieve effectively high-quality examples without blurriness introduced by the motion of the UAV or the wind, a blur detector using the variance of the Laplacian was used (Pech-Pacheco et al., 2000). In case the variance is lower than a pre-defined threshold, then the images are characterized as blurry. The threshold was set to 3000, and the operation was implemented using the OpenCV built-in function. To ensure that no distortions are presented, the resulted patches were thoroughly visually inspected (Figure 3.4). After the two procedures, the initial samples were refined to 1885 samples. The dataset was split to 80/20 for training and validation. To increase the sample size, simple data augmentation techniques involving rotation of 90 and 180 degrees were implemented.

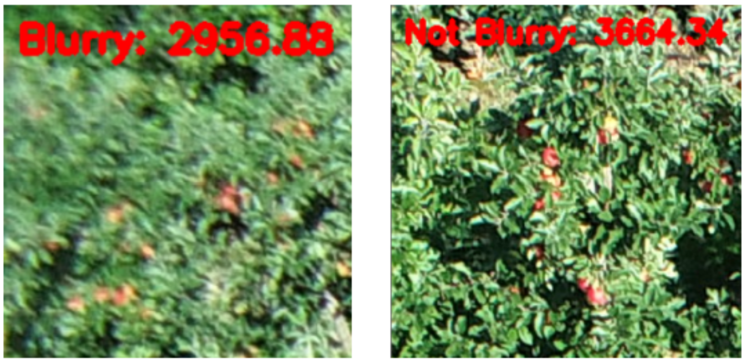


Figure 3.4: A screenshot during the data selection process, using the Laplacian variance and visual inspection.

Both generator and discriminator networks were fine-tuned, with learning rate 0.0005, batch size of 6, for 100000 iterations. Adam optimization was employed, with $\beta 1$=0.9 and $\beta 2$=0.999. The generator was trained using the loss function (Equation 3.3), with α=1, β=0.1 and γ= 0.1. Two degradation models were employed, involving the introduction of blurriness, random resize, noise and JPEG compression, following the parameters recommended by the authors of the Real-ESRGAN (X. Wang et al., 2021). The process carried out on the cloud-based virtual machine Google Colab Pro (Google LLC, Mountain View, CA, USA), which provides an Nvidia Tesla P100 GPU (NVIDIA, Santa Clara, CA, USA), with 3584 CUDA cores, and took approximately 2 days and 4 hours.

3.4. Implementation of YOLOv5 for Apple Detection

The YOLOv5 family incorporates four main architectures: YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x. Their only difference is the number of feature extraction modules and convolutional kernels. All four architectures are consisted by three components: 1) the backbone network, 2) the neck network, and 3) the head module. The backbone is a CNN which extracts the image features for each grid cell by multiple convolutions and poolings, generating four layers of feature maps with different sizes: 152 × 152 pixels, 76 × 76 pixels, 38 × 38 pixels, and 19 × 19 pixels (Xu et al., 2021). Then the neck network, i.e. path aggregation network (PANet), aggregates these different-sized feature maps, aiming to acquire more contextual information, producing new feature maps. Finally, the head module detects and classifies the objects, using the feature maps produced by the neck network. For this research, the YOLOv5s model was used (available on: https://github.com/ultralytics/yolov5).

3.4.1. Baseline YOLOv5

Before evaluating the detection on super-resolved images, a baseline should be established based on the native lower-resolution dataset. For that reason, images from the original 5472 × 3648 px UAV dataset, —not used for the training and validation of the SR model, were selected and cropped to non-overlapping tiles of 608 x 608 px, aiming to reduce processing time. A data cleaning process followed, ensuring that the tiles contain mostly apples and not scenes of solely grass or objects of non-interest. A total of 1233 tiles was gathered and annotated using the open-source labelling tool 'LabelImg' (available on: https://github.com/tzutalin/labelImg, under the YOLO format. The images were annotated using one class, i.e. apples. Hazy or blurred positive samples were excluded. The dataset was split to 80/20 for training and validation, i.e 1001 for training, and 232 for validation. Data augmentation techniques were applied to increase the datasets size, including rotations of 90° and 180° degrees, and gamma correction of value ± 0.5.. After the process, the training data consisted of a total 4004 tiles, and the validation a total of 928. The testing set of the study was created manually, by the remaining tiles, selecting images with a considerable variation in lighting conditions, size of fruits and distortions. It resulted to a total of 50 tiles of 608 x 608 px resolution.

For training the detector on the native dataset, a pre-trained model on the MS COCO dataset was used, to initialize the weights and decrease the training time. The initial learning rate was set to 0.0001, with batch size of 32, weight decay of 0.0001, and the stochastic gradient descent (SGD) was used as optimization strategy. The network was trained for 300 epochs, under the PyTorch framework, on the cloud-based virtual machine Google Colab Pro (Google LLC, Mountain View, CA, USA). The procedure lasted approximately 9 hours.

3.4.2. Super-Resolution YOLOv5

Having the baseline model ready, the exact same training, validation and testing datasets were firstly super-resolved by up-scalling factors of x2, –as the optimal up-scalling factor for detection (Rabi et al. 2020; Velumani et al., 2021)–, increasing the images size to 1216 x 1216 px. Furthermore, the dataset was cropped again to 608 x 608 px, preserving the annotations, using a Python script (available on: https://github.com/slanj/yolo-tiling). Resulting to 4004 training samples. Tiles featuring no labels were discarded, to ensure that the majority of the datasets contains images with apples, reducing the sample to 3623 annotated tiles. Figure 3.5 illustrates the super-resolution and cropping procedures. Furthermore, the datasets were manually inspected to ensure that the labels were corresponding to apples, and not background objects. It was found that annotated apples in the middle of the images, were split to half during the cropping process, resulting to labels (Figure 3.6).

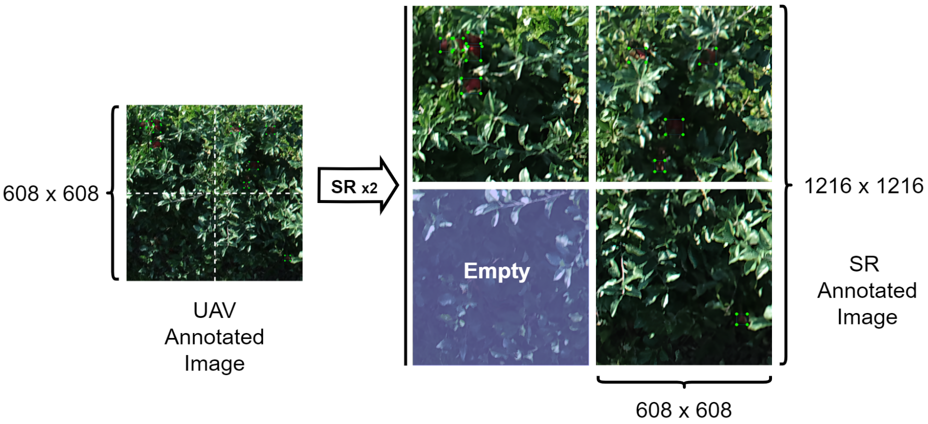


Figure 3.5: The cropping process after SISR, where the x2 image of 1216 x 1216 px is cropped again to 608 x 608 px, and the tiles with empty annotations are discarded. Zoom for better view of the green bounding boxes.

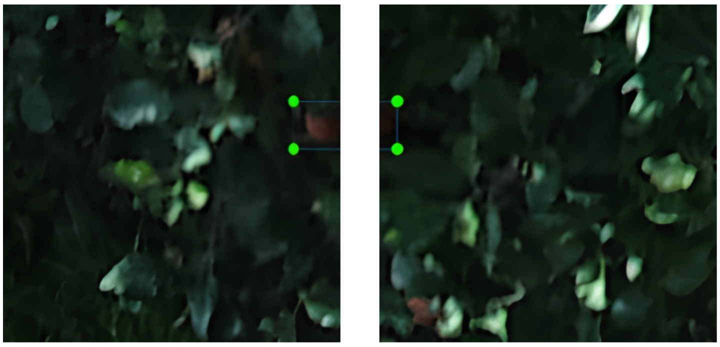


Figure 3.6: An example of an error induced by the cropping procedure. In this case, as the apple was in the center of the image, after the cropping it resulted to two tiles, the one empty representing a false label.

Data augmentation techniques applied, in the same manner as the baseline model. For the testing dataset, no cropping applied, and the inference was performed on the native resolution, i.e. 1216 x 1216 px, in order to investigate the effect of large-sized high-resolution images in detection performance. For training the SR detector, a pre-trained model on the MS COCO dataset was used. The initial learning rate was set to 0.0001, with batch size of 32, weight decay of 0.0001, and the stochastic gradient descent (SGD) was used as optimization strategy. The network was trained for 300 epochs, under the PyTorch framework, on the cloud-based virtual machine Google Colab Pro (Google LLC, Mountain View, CA, USA). The procedure lasted approximately 15 hours. Table 3.1 summarizes the characteristics of the baseline and super-resolved datasets.

Model:	Baseline	Super-Resolved
Native Resolution	608 x 608	1216 x 1216
# of Training Images	4004	14451
# of Annotated Apples	52257	52238
Testing Resolution	608 x 608	1216 x 1216
# of Testing Images	50	50

Table 3.1: Details regarding the two object detection datasets.

3.5. Metrics

3.5.1. Super-Resolution

Assessing the reconstruction performance of the proposed method is a challenging process, as no reference ground-truth high resolution images exist. Therefore, full-reference metrics (FR-IQMs), such as the peak signal-to-noise ratio (PSNR) and structure similarity index (SSIM), cannot be applied. For this reason, three non-reference image quality metrics (NR-IQMs) are adopted for the image quality evaluation: the naturalness image quality evaluator (NIQE), the perception-based image quality evaluator (PIQE), and the perceptual index (PI). These measures are found to strongly correlate with the human perception, and especially PI is regarded as the closest (Blau et al., 2018).

As the NIQE and PIQE are based on complex mathematical formulations, for the sake of brevity will be not presented in this report. Readers interested in the detailed methematical formulations can refer to Mittal, Soundararajan and Bovik (2012) for the NIQE, Chan and Goldsmith (2000) for the PIQE. The PI is formulated as follows (Eq. 3.4), where the Ma stands for the NR-IQM of Ma et al. (2017):

$$\text{Perceptual index (PI)} = \frac{1}{2}((10 - \text{Ma}) + \text{NIQE}), \tag{3.4}$$

3.5.2. Apple Detection

For the task of detection, three common measures found in apple detection research, are adopted: the precision, recall and F1 score, formulated as Equations 3.5, 3.6 and 3.7, respectively:

$$\text{Precision (P)} = \frac{TP}{\text{TP + FP}} \tag{3.5}$$

$$\text{Recall (R)} = \frac{TP}{\text{TP + FN}} \tag{3.6}$$

Where, correctly detected apples are denoted as true positives (TP), correctly detected non-apple objects as true negatives (TN), falsely detected apples as false positives (FP), and non-apple objects detected as apples as false negatives (FN), and a F1 score as the result of:

$$\text{F1 Score} = \frac{2 * Precision * Recall}{\text{Precision + Recall}} \tag{3.7}$$

4

Results

In this chapter, the results of the evaluation procedure are presented. To validate the performance of the adopted method, this thesis employs for comparison: (1) a traditional SISR method, the bicubic interpolation, (2) a state-of-the-art GAN-based model which employs the classical degradation process, the ESRGAN (X. Wang et al., 2018) not fine-tuned in the dataset, and, (3) a Real-ESRGAN pre-trained on DIV2K, Flickr2K and OST, but not fine-tuned in the UAV dataset. With this comparison, it is aimed to be investigated if the domain-specific generator of the Real-ESRGAN can enhance the distorted images present on the dataset and effectively increase detection rates. Subsequent sections are divided accordingly, to: a) results of image quality evaluation, based on quantitative and qualitative methods, and b) the results of object detection performance, using the object detection metrics.

4.1. Generative Adversarial networks (GANs) for super-resolution (SR)

4.1.1. No-Reference Image Quality Metrics (IQMs)

For the evaluation of the SR perfomance based on image quality, a total of 200 tiles of 608 x 608 px resolution, not used during the training of the SR model, were selected from the UAV dataset. Table 4.1, depicts the averages of each method on NIQE, PIQE and PI metrics.

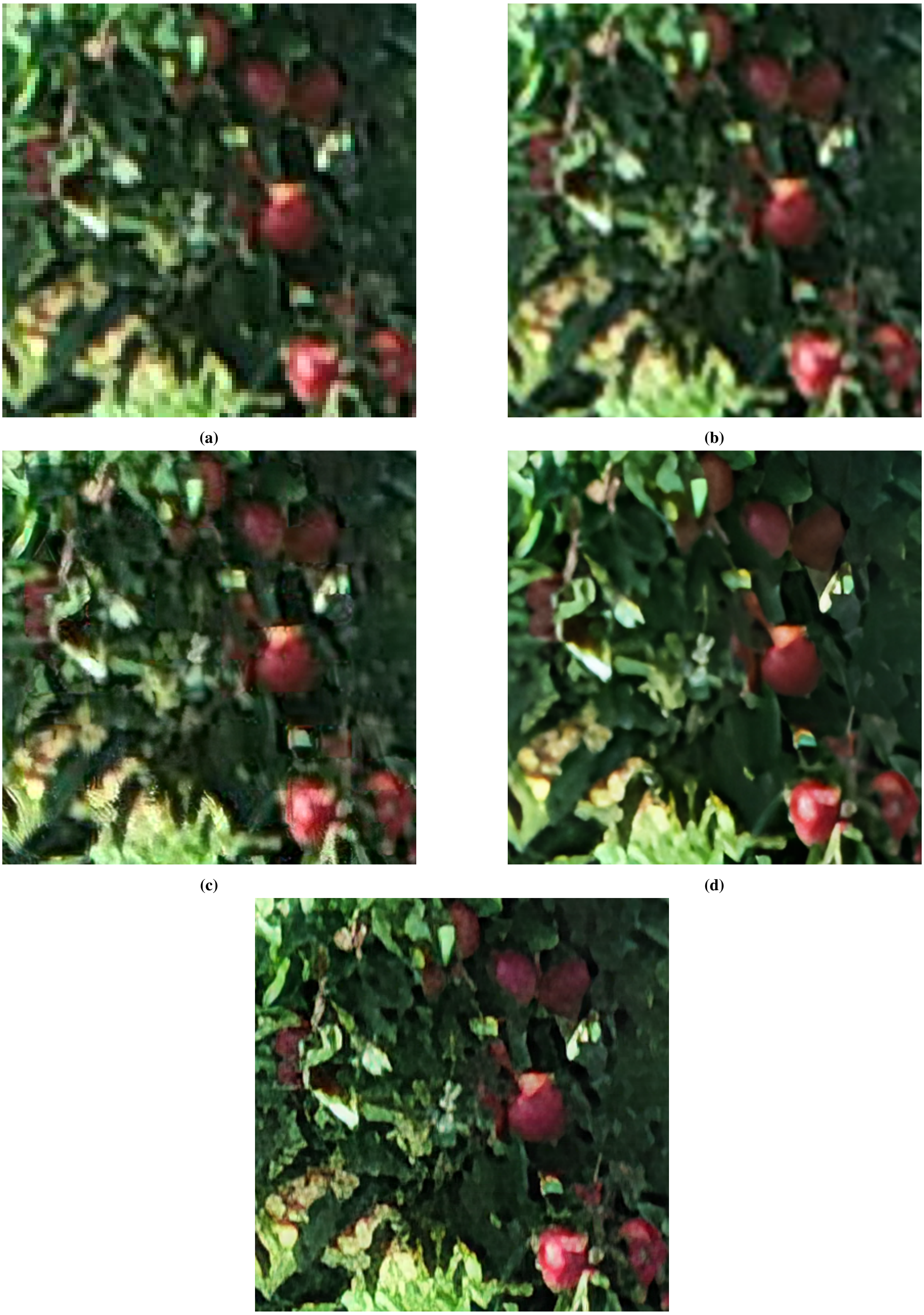
Table 4.1: Comparative results on the NR-IQMs. Note, the lower the values the higher the quality of the reconstructions.

Method	NIQE	PIQE	PI
Bicubic Interpolation	6.131	89.81	6.810
ESRGAN	5.916	**38.37**	6.451
Real-ESRGAN without fine-tuning	5.313	69.82	**3.785**
Real-ESRGAN with UAV fine-tuning	**4.991**	41.98	4.527

Among these IQMs, the two Real-ESRGAN models outerperfor bicubic interpolation and ESRGAN in PI. While, interestingly the Real-ESRGAN not fine-tuned in the dataset achieves higher performance in the PI index, compared the fine-tuned model. The ESRGAN model found to outerperform all the other approaches in the PIQE index.

4.1.2. Visual Comparison

Visual examples of image improvement between the different methods are presented in Figure 4.1. In comparison with the original image, all methods increased effectively the initial resolution. However, the enhancement of original images in terms of semantic information under visual examination, differs significantly.



(a)

(b)

(c)

(d)

(e)

Figure 4.1: Example of super-resolution (SR) reconstructions of the (a) original dataset: (b) Bicubic Interpolation , (c) ESRGAN, (d) Real-ESRGAN not fine-tuned in the dataset, and (e) the domain-specific Real-ESRGAN fine-tuned using a limited number of UAV imagery

The visual investigation reveled that deep learning based SISR methods produce images with finer details, compared to bicubic interpolation. Furthermore, the two Real-ESRGAN models exhibit high visual quality compared to, both the bicubic interpolation and ESRGAN, confirming the PI scores. However, the model fine-tuned in the UAV dataset introduces artifacts, compared to the model not fine-tuned in the dataset. This can be explained as the training images used for fine-tuning, despite being of the highest-quality for the given dataset, still exhibit considerable distortion compared to DIV2K images (Figure 4.2). This can lead the generator to recognize these distortions as high-resolution features, and in result to replicate them during the inference phase.

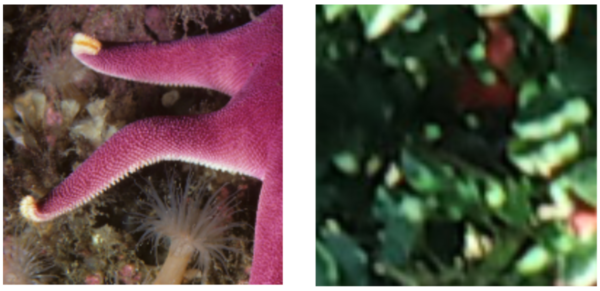


Figure 4.2: Zoomed at 400%. On the left: Image #0001 "Starfish" from the DIV2K dataset. On the right: an image used for the fine-tuning .

To confirm the statement that lower-quality training images can add artifacts in the reconstructions, an additional visual examination follows, aiming to explore if fine-tuning the models in blurry images can hamper the visual performance. In specific, three models are employed: a) a Real-ESRGAN that is not fine-tuned in the dataset, and trained only on DIV2K, b) the Real-ESRGAN fine-tuned on high quality UAV images, a retrieved using the Laplacian blur detector (See Section 3.3.3, and c) a Real-ESRGAN fine-tuned on a UAV dataset, that was not investigated for blurriness or other distortions. Figure 4.3), depicts the results.

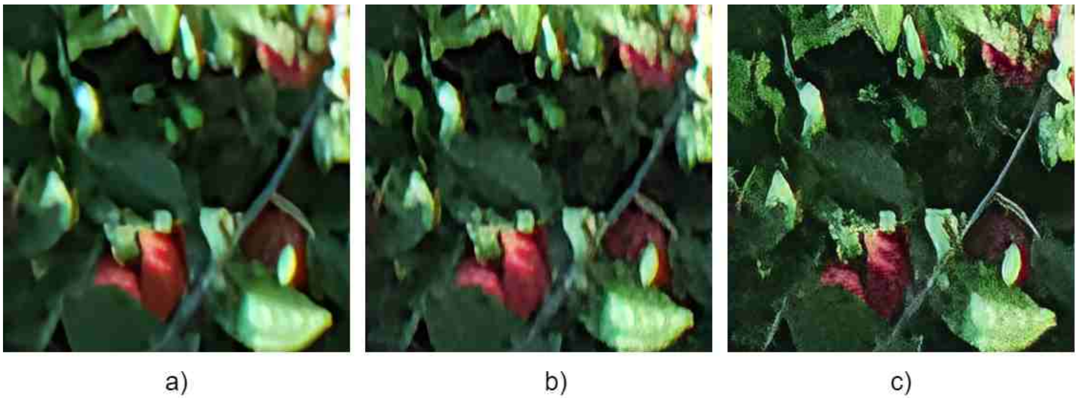


Figure 4.3: A visual example of the effect of training the model with image of adequate quality. In specific: a) the Real-ESRGAN trained only on natural images, e.g. DIV2K, Flickr2K and OST, results to higher visual quality, compared to the b) fine-tuned model of the research, —using the blur detector, and c) a model trained on low-quality UAV images from the dataset.

The visual examination of Figure 4.3, revealed indeed that training the GAN models on lower-quality images can result to poor visual performance. Nevertheless, the Real-ESRGAN models demonstrate superior reconstruction performance compared to traditional SISR methods, and GAN-based models which utilize a more simplistic degradation process (Figure2 4.1). However, the NR-IQMs and the visual investigation method used, are not suitable to assess the reconstruction performance. The metrics commonly utilized for assessing reconstruction performance on the SR research, are the peak signal-to-noise ration (PSNR) and structure similarity index (SSIM). Both metrics are full-reference IQMs. Since ground-truth higher-resolution images are not available, the PSNR and SSIM cannot be deployed.

4.2. YOLOv5s for Apple Detection

For the fruit detection performance, the same models used in the NR-IQMs evaluation and visual investigation were employed for comparison purposes. In specific, the models super-resolved the training, validation and testing dataset, with up-scalling factor of x2, following the pre-processing pipeline explained in Section 3.4.2. After, the super-resolution and cropping procedures, three additional YOLOv5 detectors were trained following the settings described in Section 3.4.2. For the inference of the testing dataset, the confidence threshold was set to 0.25, and intersection over union area (IoU) threshold to 0.25. Table 4.2 depicts the results of the detection.

Table 4.2: Comparative results on Precision, Recall and F1 score. Note, the higher the values the better the results.

Method	Operational Resolution	Precision	Recall	F1
YOLOv5 Baseline	608 x 608	0.85	0.65	0.74
YOLOv5 Bicubic Interpolation	1216 x 1216	0.86	0.74	0.80
YOLOv5 ESRGAN	1216 x 1216	0.88	0.81	0.84
YOLOv5 Real-ESRGAN without fine-tuning	1216 x 1216	**0.95**	0.73	0.83
YOLOv5 Real-ESRGAN with UAV fine-tuning	1216 x 1216	0.91	**0.85**	**0.88**

Good performances were observed when the models are trained on the higher-resolution reconstructions. In specific, the YOLOv5 trained and tested on the source dataset, i.e. without super-resolution scored the worst in all metrics. In comparison, the YOLOv5s trained on the dataset provided by the fine-tuned Real-ESRGAN, achieved the best performance. Furthermore, the YOLOv5 trained and tested on the ESRGAN dataset achieved higher score than the Real-ESRGAN not fine-tuned in the UAV dataset. Furthermore, the YOLOv5 trained and tested on the bicubic interpolation dataset, scored worse than the GAN-based SISR methods, but better than the source YOLOv5. Figure 4.4 shows an example of detection between the Baseline YOLOv5, i.e. the modle with lowest score, and the Real-ESRGAN YOLOv5 fine-tuned on the UAV dataset, i.e. the model with highest score.

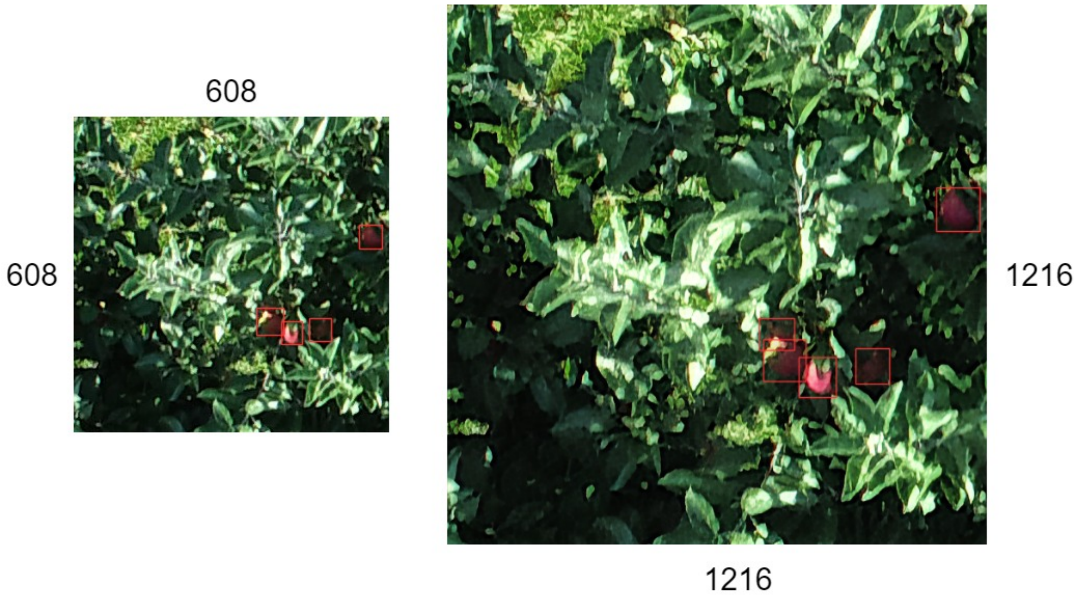


Figure 4.4: On the left, the model trained on source resolution (608 x 608 px), missed one occluded target. In comparison, the model trained on the Real-ESRGAN dataset (1216 x 1216 px), detected all targets.

Utilizing super-resolution prior to detection, shows higher detection performance, However, the testing for SISR methods performed on their native resolution, i.e. 1216 x 1216 px, which can proven computational demanding and decrease detection speed. To verify this statement the inference times between the models are compared. Table 4.3, shows the average time in milliseconds (ms) per tile, for the compared methods.

As the result indicate, detecting in texture-rich higher-resolutions can be proven slow. The baseline YOLOv5, processing tiles of 608 x 608 px, was the fastest as expected. This is an important factor in precision agriculture,

Method	Average Speed per Tile (ms)
Baseline	9.3
Bicubic Interpolation	13.6
ESRGAN	13.7
Real-ESRGAN without fine-tuning	13.7
Real-ESRGAN with UAV fine-tuning	13.3

Table 4.3: Comparative results on detection speed.

where fast and light-weight detection models are core technologies for fruit-picking robots (Yan et al., 2021)

5

Discussion

Super-resolution (SR) attracted immense interest by the scientific community, and was characterized as a powerful image pre-processing technique. Compared to other image enhancement approaches, e.g. block-matching and 3D filtering (BM3D) algorithms, enlarges the objects beyond the initial resolution of the given image, — a characteristic considerably valuable for the the task of UAV-based fruit detection. Nonetheless, in SR operations two main challenges occur: 1) the fundamental ill-possed challenge of the infinite number of potential super-resolution reconstructions, and 2) the hardship to define and use the appropriate SR model for a given dataset. The latter is reflected by the results of this research, where the GAN model trained under classical degradation modelling, resulted to reconstructions of lower quality, with a substantial amount of artifacts (Figure 4.1). In contrast, the GAN-based methods trained under more sophisticated degradation modelling. i.e. the Real-ESRGAN models, provided superior reconstructions under visual investigation. This is can be explained by deconstructing the high-order degradation modelling of the Real-ESRGAN models, used in this research. This degradation modelling approach, incorporated blur, resizing, additive noise and JPEG compression. The UAV dataset used in this study, although exhibited a variety of unknown degradations related to the imaging system, was highly affected by blurriness introduced by the motion of the UAV and loss of information due to JPEG compression (Figure 3.2b). Thus, the blur kernels and JPEG-compression simulation used in the training procedure, were highly beneficial for the reconstruction of images with high perceptual quality (Section 4.1.1. This pattern, where domain-specific degradation modelling yields superior reconstruction is also observed by Z. Zhang et al. (2022).

Moreover, the quality of the training material is also crucial. By comparing Figure 4.3, it can be seen that training HR images of low quality result to the introduction of artifacts and chromatic abbreviations. In particular, the Real-ESRGAN trained on abundant high-quality natural images, exhibited outstanding results compared to the other methods, in terms of perceptual quality. Confirming that supervised GAN-based methods require sufficient high-quality images to perform satisfactory. However, DL SISR models trained on off-domain datasets, can lead to the effect called domain-shift, where remote sensing datasets inherit characteristics found on natural scenes. This inheritance may lead to alterations in image statistics of the remote sensing images, and thus possible introduction of artifacts and distortions in the reconstructions. Such alterations, were not able to be quantified, as no FR-IQMs capable to capture the reconstruction accuracy were used, due to the absence of higher-resolution reference.

Interestingly, SR method used, i.e. bicubic interpolation, ESRGAN, Real-ESRGAN, the detection rates for the models utilized the SR as a prepossessing step were greater, compared to the model that detected images on the unprocessed dataset. In specific it was shown that by training the detector in super-resolved tiles of 608 x 608 px resolution, and performing the inference on super-resolved images of 1215 x 1216 px, an increase by 7.06% in precision, 30.77% in recall, and 18.92% on F1 score is observed (Table 4.2). This can be explained by the ability of SR to successfully reveal partial occluded and shaded fruits. However, a further explanation is that the detectors employed in this research were pre-trained on the MS COCO dataset, where the objects in this specific dataset are mostly large. Thus, the YOLOv5 models that utilized SR images, and as result larger objects, were able to perform better since they were fine-tuned based on the MS COCO weights. However, training from scratch is not recommend in UAV-based object detection (Apolo-Apolo et al., 2020; Zhu et al., 2021).

Another interesting observation, is related to detection performance of YOLOv5 model trained and tested on reconstructions produced by the Real-ESRGAN model, that was trained only on natural scenes. This YOLOv5 model scored lower than the domain-specific Real-ESRGAN and the ESRGAN (Table 4.2), — which used a classical degradation training approach. Although, reconstruction accuracy metrics are not available, it is known that the models exhibiting high PI values, fundamentally do not score well in reconstruction accuracy metrics such as the peak noise-to-signal ratio (PSNR) and structure similarity index (SSIM) (Blau et al., 2018). This evidence, can contribute to the hypothesis of domain-shift, where the reconstructions of the Real-ESRGAN trained only on natural scenes, might be altered in terms of image statistics. If true, its reconstructions show a variety of image features, inherited both fro the UAV dataset and the natural scenes, which might be proven for the YOLOv5 model to be captured accurately. However, as mentioned before such assumption can not be verified.

Finally, the YOLOv5s model trained and tested on the native UAV resolution, i.e. 608 x 608 px, found to exhibit the highest average detection speed per tile, of 9.3 ms. In comparison, the detectors trained and tested on higher-resolution SISR reconstructions, exhibited average speeds ranging from 13.3 to 13.7 ms. Suggesting that super-resolved images require additional time to be process by the YOLOv5. Nevertheless, the range of 13.3 to 13.7 ms is considered very minimal time-wise (Wang & He, 2021). This is a extremely relevant for the evaluation of the SISR methods in respect with their real-life application in yield-estimation, which time-efficiency is a big matter.

6

Conclusion

In this thesis, the problems associated with the detection of fruits on unmanned aerial vehicles (UAVs) datasets, has been approached by the utilization of single image super-resolution (SISR) methods. Results showed, that supervised generative adversarial networks (GANs) based SISR is a capable to effectively combat occlusion, blurriness and other optical distortions, elevating considerably the detection performance. In addition, as a cost-effective and timely approach, it can be proven efficient in situation when acquisition of high-quality images is expensive or infeasible. Concluding, while supervised SISR methods proven effective for the task of this thesis, further research should be implemented to investigate the relationship of domain-shift and fruit detection on remote sensing imagery.

Bibliography

Agustsson, E., & Timofte, R. (2017). Ntire 2017 challenge on single image super-resolution: Dataset and study. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Alqahtani, H., Kavakli, M., & Kumar Ahuja, D. G. (2019). Applications of generative adversarial networks (gans): An updated review. *Archives of Computational Methods in Engineering*, *28*. https://doi.org/10.1007/s11831-019-09388-y

Apolo-Apolo, O. E., Pérez-Ruiz, M., Martínez-Guanter, J., & Valente, J. (2020). A cloud-based environment for generating yield estimation maps from apple orchards using uav imagery and a deep learning technique. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.01086

Bai, Y., Zhang, Y., Ding, M., & Ghanem, B. (2018). Sod-mtgan: Small object detection via multi-task generative adversarial network. *ECCV*.

Bargoti, S., & Underwood, J. P. (2017). Image segmentation for fruit detection and yield estimation in apple orchards. *Journal of Field Robotics*, *34*(6), 1039–1060. https://doi.org/https://doi.org/10.1002/rob.21699

Benjdira, B., Khursheed, T., Koubaa, A., Ammar, A., & Ouni, K. (2018). Car detection using unmanned aerial vehicles: Comparison between faster r-cnn and yolov3.

Bevilacqua, M., Roumy, A., Guillemot, C., & Morel, M.-l. A. (2012). Low-complexity single-image super-resolution based on nonnegative neighbor embedding. *Proceedings of the British Machine Vision Conference*, 135.1–135.10. https://doi.org/http://dx.doi.org/10.5244/C.26.135

Burdziakowski, P. (2020). Increasing the geometrical and interpretation quality of unmanned aerial vehicle photogrammetry products using super-resolution algorithms. *Remote Sensing*, *12*(5). https://doi.org/10.3390/rs12050810

Cardenas-Weber, M., Stroshine, R. L., Haghighi, K., & Edan, Y. (1991). Melon material properties and finite element analysis of melon compression with application to robot gripping. *Transactions of the ASABE*, *34*, 920–0929.

Chan, R., & Goldsmith, P. (2000). A psychovisually-based image quality evaluator for jpeg images. *Smc 2000 conference proceedings. 2000 ieee international conference on systems, man and cybernetics. 'cybernetics evolving to systems, humans, organizations, and their complex interactions' (cat. no.0*, *2*, 1541–1546 vol.2. https://doi.org/10.1109/ICSMC.2000.886075

Chen, C., Shi, X., Qin, Y., Li, X., Han, X., Yang, T., & Guo, S. (2022). Blind image super resolution with semantic-aware quantized texture prior. https://doi.org/10.48550/ARXIV.2202.13142

Clabaut, É., Lemelin, M., Germain, M., Bouroubi, Y., & St-Pierre, T. (2021). Model specialization for the use of esrgan on satellite and airborne imagery. *Remote Sensing*, *13*(20). https://doi.org/10.3390/rs13204044

Courtrai, L., Pham, M.-T., & Lefèvre, S. (2020). Small object detection in remote sensing images based on super-resolution with auxiliary generative adversarial networks. *Remote Sensing*, *12*(19). https://doi.org/10.3390/rs12193152

Edan, Y., Rogozin, D., Flash, T., & Miles, G. (2000). Robotic melon harvesting. *IEEE Transactions on Robotics and Automation*, *16*(6), 831–835. https://doi.org/10.1109/70.897793

Ferdous, S. N., Mostofa, M., & Nasrabadi, N. M. (2019). Super resolution-assisted deep aerial vehicle detection. In T. Pham (Ed.), *Artificial intelligence and machine learning for multi-domain operations applications* (pp. 432–443). SPIE. https://doi.org/10.1117/12.2519045

Gené-Mola, J., Sanz-Cortiella, R., Rosell-Polo, J. R., Escolà, A., & Gregorio, E. (2021). In-field apple size estimation using photogrammetry-derived 3d point clouds: Comparison of 4 different methods considering fruit occlusions. *Computers and Electronics in Agriculture*, *188*, 106343. https://doi.org/https://doi.org/10.1016/j.compag.2021.106343

Girshick, R. (2015). Fast R-CNN. *2015 IEEE International Conference on Computer Vision (ICCV)*, 1440–1448. https://doi.org/10.1109/ICCV.2015.169

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2013). Rich feature hierarchies for accurate object detection and semantic segmentation. https://doi.org/10.48550/ARXIV.1311.2524

González, D., Patricio, M. A., Berlanga, A., & Molina, J. (2019). A super-resolution enhancement of uav images based on a convolutional neural network for mobile devices. *Personal and Ubiquitous Computing*. https://doi.org/10.1007/s00779-019-01355-5

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial networks. https://doi.org/10.48550/ARXIV.1406.2661

Guo, Y., Liu, Y., Oerlemans, A. A. J., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing*, *187*, 27–48.

Haris, M., Shakhnarovich, G., & Ukita, N. (2018). Task-driven super resolution: Object detection in low-resolution images.

Hu, S., Pizlo, Z., & Allebach, J. (2014). Jpeg ringing artifact visibility evaluation. *Proceedings of SPIE - The International Society for Optical Engineering*, *9016*, 90160E. https://doi.org/10.1117/12.2048594

Jolicoeur-Martineau, A. (2018). The relativistic discriminator: A key element missing from standard GAN. *CoRR*, *abs/1807.00734*. http://arxiv.org/abs/1807.00734

Justin, J., Alexandre, A., & Li, F.-F. (2016). Perceptual losses for real-time style transfer and super-resolution.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–44. https://doi.org/10.1038/nature14539

Ledig, C., Theis, L., Huszar, F., Caballero, J., Aitken, A. P., Tejani, A., Totz, J., Wang, Z., & Shi, W. (2016). Photo-realistic single image super-resolution using a generative adversarial network. *CoRR*, *abs/1609.04802*. http://arxiv.org/abs/1609.04802

Lei, S., Shi, Z., & Zou, Z. (2020). Coupled adversarial training for remote sensing image super-resolution. *IEEE Transactions on Geoscience and Remote Sensing*, *58*, 3633–3643.

Li, T., Feng, Q., Qiu, Q., Xie, F., & Zhao, C. (2022). Occluded apple fruit detection and localization with a frustum-based point-cloud-processing approach for robotic harvesting. *Remote Sensing*, *14*(3). https://doi.org/10.3390/rs14030482

Lim, B., Son, S., Kim, H., Nah, S., & Lee, K. M. (2017a). Enhanced deep residual networks for single image super-resolution. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Lim, B., Son, S., Kim, H., Nah, S., & Lee, K. M. (2017b). Enhanced deep residual networks for single image super-resolution. *CoRR*, *abs/1707.02921*. http://arxiv.org/abs/1707.02921

Linker, R., Cohen, O., & Naor, A. (2012). Determination of the number of green apples in rgb images recorded in orchards. *Computers and Electronics in Agriculture*, *81*, 45–57. https://doi.org/https://doi.org/10.1016/j.compag.2011.11.007

Litjens, G. J. S., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., van der Laak, J., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical image analysis*, *42*, 60–88.

Liu, A., Liu, Y., Gu, J., Qiao, Y., & Dong, C. (2021). Blind image super-resolution: A survey and beyond. *CoRR*, *abs/2107.03055*. https://arxiv.org/abs/2107.03055

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S. E., Fu, C., & Berg, A. C. (2015). SSD: single shot multibox detector. *CoRR*, *abs/1512.02325*. http://arxiv.org/abs/1512.02325

Ma, C., Yang, C.-Y., Yang, X., & Yang, M.-H. (2017). Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, *158*, 1–16. https://doi.org/https://doi.org/10.1016/j.cviu.2016.12.009

Mittal, A., Soundararajan, R., & Bovik, A. C. (2013). Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, *20*(3), 209–212. https://doi.org/10.1109/LSP.2012.2227726

Pang, Y., & Cao, J. (2019). Deep learning in object detection. In X. Jiang, A. Hadid, Y. Pang, E. Granger, & X. Feng (Eds.), *Deep learning in object detection and recognition*. Springer Singapore.

Pashaei, M., Starek, M. J., Kamangir, H., & Berryhill, J. (2020). Deep learning-based single image super-resolution: An investigation for dense scene reconstruction with uas photogrammetry. *Remote Sensing*, *12*(11). https://doi.org/10.3390/rs12111757

Pech-Pacheco, J. L., Cristóbal, G., Chamorro-Martínez, J., & Fernández-Valdivia, J. (2000). Diatom autofocusing in brightfield microscopy: A comparative study. *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, *3*, 314–317 vol.3.

Rabbi, J., Ray, N., Schubert, M., Chowdhury, S., & Chao, D. (2020). Small-object detection in remote sensing images with end-to-end edge-enhanced gan and object detector network. *Remote Sensing*, *12*, 1432. https://doi.org/10.3390/rs12091432

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, real-time object detection. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779–788. https://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. https://doi.org/10.48550/ARXIV.1506.01497

Sajjadi, M. S. M., Schölkopf, B., & Hirsch, M. (2017). Enhancenet: Single image super-resolution through automated texture synthesis. *2017 IEEE International Conference on Computer Vision (ICCV)*, 4501–4510.

Salvetti, F., Mazzia, V., Khaliq, A., & Chiaberge, M. (2020). Multi-image super resolution of remotely sensed images using residual attention deep neural networks. *Remote Sensing*, *12*(14), 2207. https://doi.org/10.3390/rs12142207

Stern, H., Burks, T., & Alchanatis, V. (2010). Low and high-level visual feature-based apple detection from multi-modal images. *Precision Agriculture*, *11*, 717–735. https://doi.org/10.1007/s11119-010-9198-x

Velumani, K., Lopez-Lozano, R., Madec, S., Guo, W., Gillet, J., Comar, A., & Frederic, B. (2021). Estimates of maize plant density from uav rgb images using faster-rcnn detection model: Impact of the spatial resolution.

Wang, Q., Nuske, S. T., Bergerman, M., & Singh, S. (2012). Automated crop yield estimation for apple orchards. *ISER*.

Wang, X., Xie, L., Dong, C., & Shan, Y. (2021). Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. https://arxiv.org/abs/2107.10833

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Loy, C. C., Qiao, Y., & Tang, X. (2018). Esrgan: Enhanced super-resolution generative adversarial networks. https://doi.org/10.48550/ARXIV.1809.00219

Wang, Z., Bovik, A., Sheikh, H., & Simoncelli, E. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1109/TIP.2003.819861

Wen, J., Shi, Y., Zhou, X., & Xue, Y. (2020). Crop disease classification on inadequate low-resolution target images. *Sensors*, *20*(16). https://doi.org/10.3390/s20164601

Whitehead, K., & Hugenholtz, C. (2014). Remote sensing of the environment with small unmanned aircraft systems (uass), part 1: A review of progress and challenges. *Journal of Unmanned Vehicle Systems*, *02*, 69–85. https://doi.org/10.1139/juvs-2014-0006

Xia, Z. (2019). An overview of deep learning. In X. Jiang, A. Hadid, Y. Pang, E. Granger, & X. Feng (Eds.), *Deep learning in object detection and recognition* (pp. 1–18). Springer Singapore. https://doi.org/10.1007/978-981-10-5152-4_1

Xu, R., Lin, H., Lu, K., Cao, L., & Liu, Y. (2021). A forest fire detection system based on ensemble learning. *Forests*, *12*(2). https://doi.org/10.3390/f12020217

Yan, B., Fan, P., Lei, X., Liu, Z., & Yang, F. (2021). A real-time apple targets detection method for picking robot based on improved yolov5. *Remote Sensing*, *13*(9). https://doi.org/10.3390/rs13091619

Zhang, L., Zhang, L., & Du, B. (2016). Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, *4*(2), 22–40. https://doi.org/10.1109/MGRS.2016.2540798

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., & Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. *CoRR*, *abs/1807.02758*. http://arxiv.org/abs/1807.02758

Zhang, Z., Tian, Y., Li, J., & Xu, Y. (2022). Unsupervised remote sensing image super-resolution guided by visible images. *Remote Sensing*, *14*(6). https://doi.org/10.3390/rs14061513

Zhao, J., Zhang, X., Yan, J., Qiu, X., Yao, X., Tian, Y., Zhu, Y., & Cao, W. (2021). A wheat spike detection method in uav images based on improved yolov5. *Remote Sensing*, *13*(16). https://doi.org/10.3390/rs13163095

Zhu, X., Lyu, S., Wang, X., & Zhao, Q. (2021). Tph-yolov5: Improved yolov5 based on transformer prediction head for object detection on drone-captured scenarios. https://doi.org/10.48550/ARXIV.2108.11539