Small Object Detection in Remote Sensing Images and Videos
An overview of methods and solutions

Stamatios Orfanos
mtn2211
stamatisorfanos99@gmail.com

Abstract

Small Object Detection in Remote Sensing Images and Videos is a critical challenge in the field of Computer Vision. As visual data continues to proliferate across various industries, the ability to accurately detect and locate small objects within images is essential for military, traffic, civilian, sports, and numerous other applications. This article dives into the intricacies of data augmentation, super-resolution, and feature extraction as vital components in the field. It navigates the landscape of detection algorithms, from the efficiency of Single-Shot Detectors to the precision of multi-shot counterparts. Anchoring these discussions are the emerging trends, including the integration of spectral imaging and the dynamic interplay between anchor-based and anchor-free approaches. The ongoing advancements are shaping a future where detection systems exhibit enhanced accuracy, efficiency, and adaptability, with the incorporation of spectral-spatial data emerging as a promising solution for more nuanced identification of small objects.

1 Introduction

Object detection is a fundamental task in computer vision. When given an image, object detection aims at finding where and what each object instance is. From the application perspective, object detection can be grouped into two types: **generic object detection** and **domain-specific detection**. The first type aims at detecting different types of visual objects under a unified framework, while the purpose of the second type is the detection under specific application scenarios, such as face detection, traffic sign detection, pedestrian detection, remote sensing target detection and so on. Even though impressive results have been achieved on large and medium objects in large-scale detection benchmarks, the performance on small or tiny objects is far from satisfactory. Compared with large and medium objects, the small objects are more difficult to detect accurately, because of four main difficulties. Firstly, small objects have low resolution and insufficient features. Secondly, the span of object-scale is large and multiple scales coexist. Thirdly, the examples of small objects are scarce and lastly the categories for small objects are imbalanced for the majority of datasets.

Remote sensing involves the collection and interpretation of information about an area from a distance. It relies on sensors, such as satellites, aircraft, and drones, to capture data in the form of images or other measurements. These data are invaluable for a wide range of applications, including environmental monitoring, agriculture, forestry, urban planning, and disaster management. Remote sensing allows us to obtain vital information about an area, such as land cover, temperature, and topography, which can aid in resource management and decision-making.

2 Fundamentals of Small Objects Detection

2.1 Definition of Small Objects

The concept of small or tiny objects seeks to elucidate the scale of these objects or the proportion of pixels they occupy within the entire image. There are two main ways to define small objects. The first way is the use **relative size**. According to the definition of Society of Photo-Optical Instrumentation Engineers [SPIE], if the object size is less than 0.12% of the original image, it is regarded as a small object. Following the same principle Krishna and Jawahar [1] showed that an object is considered small if it occupies only a tiny portion of the image, which is less than 1% of the image area. Namely, the bounding box of a small object should cover less than 1% of the original image. The second way of defining a small object by using the **absolute size**, where a small object has size less than 32x32 pixels defined in MS-COCO dataset or 16x16 pixels defined in USC-GRAD-STDdb [2].

Additionally the definition of small object may differ based on the domain of the application. For instance, in medical imaging, small objects could be microorganisms or sub-cellular structures, while in surveillance, they might include small items like weapons or specific clothing details and in some cases with aerial or satellite imaging a vessel may fit the definition as well. Using the last point as a reference, contextual perspective is an important matter in this field. The perceived size of an object can also depend on its context within the image.

2.2 Use Cases and Applications

As mentioned above, small object detection is analyzed in a variety of domains that are currently looking for a better performance in small object detection. Starting with the domain of medical imaging, the detection of small objects can be a crucial matter concerning the quality life of relevant patients. Small objects like tumours, lesions can serve as early indicators of serious medical conditions, such as cancer as per Karthick Prasad Gunasekaran et al. [54]. Detecting these small anomalies with high accuracy is crucial for timely diagnosis and treatment, often saving lives.

It is known that nowadays, due to the climate problems, the significance of efficient agriculture and accurate environmental monitoring as per Garioud A. et al. [50] have emerged as a critical challenges for humankind. In terms of agriculture, small object detection is aiming to optimise crop management, by counting individual plants, monitoring crop health, and managing pests and diseases at a fine-grained level. Following the same principle environmentalists and researchers use small object detection in environmental monitoring in order to efficiently track the movement of small animals, such as birds or insects, and identifying micro-plastics in natural environments, which is vital for preserving ecosystems and biodiversity.

In addition, forensics, surveillance and security as per Liu C. et al. [49] are domains that small object detection is vital and many solutions are being explored. In forensic science, small object analysis is instrumental in criminal investigations, like fingerprints, ballistic evidence, or trace amounts of substances, meanwhile surveillance systems heavily rely on small object detection to monitor public spaces, transport hubs, and critical infrastructure. The task includes identifying small objects such as concealed weapons, suspicious packages, or individuals in crowded scenes. Precise small object detection is essential for ensuring public safety and security. A great use case for small object detection is autonomous driving, where identifying small objects like pedestrians, cyclists, and road signs is essential for making real-time decisions and avoiding collisions, making small object detection a critical component of autonomous driving technology.

In these diverse use cases, the accurate detection and analysis of small objects are essential for achieving specific goals and objectives, ranging from helping in the medical field to improving the efficiency of agricultural practices and ensuring public safety and environmental preservation. Small object detection techniques continue to advance to meet the unique challenges posed by each application, making it a pivotal area of research in computer vision.

3 Data Collection and Preprocessing

3.1 Diverse and Representative Datasets

Datasets have played a critical role in object detection and all machine learning endeavours. They not only provide the data for data-driven algorithms, but also enable comparison between different object detection algorithms. Unfortunately the scarcity of diverse and comprehensive datasets, poses a significant challenge in the field of small object detection in computer vision. To overcome this data scarcity problem, the exploitation of self-supervised techniques has gained prominence [54]. Most researchers have to evaluate their small or tiny object detection methods on the datasets built by themselves or extracted from large datasets such the datasets found in the table 3.1.

Dateset	Description	Published/Year	Objective
[DOTA]	Includes 2086 images, 188282 instances, and 15 common categories. Each image is of the size about 4000x4000 pixels	CVPR/2018	Aerial Image Object Detection
SDOTA	Images range from 800x800 to 4000x4000 pixels, including 227656 instances covering four classes, those being small-vehicle, storage tank, ship and large-vehicle.	J-STARS/2021	Small Object Detection
[PASCAL-VOC]	Versions VOC2007 and VOC2012, where both mid-scale datasets with 20 categories.	IJCV/2010	Small Object Detection
[MS-COCO]	MS-COCO contains images annotated with rich information about objects, their attributes, and their relationships in a wide variety of real-world scenes.	ECCV/2014	Object Segmentation
[Cityscapes]	Cityscapes dataset is a widely used benchmark in computer vision research, particularly for semantic urban scene understanding	CVPR/2020	Object Segmentation
[ImageNet]	ImageNet consists of millions of labelled images for tasks like image classification and object recognition.	2018	Object Segmentation Localisation
[Stanford Drone Dataset]	The dataset includes annotated images and videos to facilitate the development and evaluation of algorithms for tasks such as object detection.	ECCV/2016	Aerial Image Object Detection
[xVIEW]	xView is one of the largest publicly available datasets of overhead imagery, by Defense Innovation Unit Experimental and the National Geospatial-Intelligence Agency.	DIUx xView/2018	Aerial Image Object Detection
[ADE20K]	ADE20K contains images covering a wide variety of scenes, with pixel-level annotations for over 150 object categories.	2019	Semantic Segmentation
[CityPersons]	This is a subset of the Cityscapes dataset including images of pedestrians in the city	CVPR/2017	Pedestrian Detection
TinyCityPersons	TinyCityPersons is constructed through down-sampling CityPersons by 4x4.	WACV/2020	Pedestrian Detection

Table 1: Overview of some popular detection datasets about small objects

3.2 Annotation and Labeling

Annotation and labeling are fundamental steps in creating high-quality datasets for small object detection. However, obtaining accurate annotations for small objects can be particularly challenging, and there are several reasons for this challenge.

Firstly, the availability of suitable datasets with precise annotations for small object detection is limited. Small objects, by their nature, may have limited visibility and may be easily overshadowed by the surrounding context. Capturing these small objects with precision requires meticulous annotation, often at a pixel-level or with highly accurate bounding boxes. As a result, finding existing datasets with comprehensive and accurate annotations for small objects can be a daunting task, especially for niche or specialised domains.

Secondly, the technology used for data collection, such as cameras, evolves over time. New camera models with improved resolutions capabilities are regularly introduced, leading to variations in image quality and characteristics. This continuous evolution can pose challenges in creating consistent datasets for small object detection, as annotating images from different camera sources and generations may require adjustments in labeling techniques to maintain accuracy. As a result, researchers and practitioners in the field carefully consider these challenges when embarking on small object detection projects and take the necessary steps to address them.

3.3 Data Augmentation and Preprocessing

3.3.1 Data Augmentation

Data is at the core of any deep learning model. Insufficient training samples are every so often answerable for deprived performances in deep learning solutions to problems. Data augmentation is a technique that can be utilised to extend the size of dataset required by deep learning models for training through artificially created data. Currently, small object detection techniques (Duan et al., 2019; Akshatha et al., 2023) address the challenge of detecting small objects by leveraging multi-layer features. However, they reshape features from various layers to the same size before aggregating them, inevitably introducing information loss. It can be broadly divided into four categories

1. Geometric Transformations (e.g. scaling, rotating, flipping, cropping, padding etc.)
2. Colour Transformations (changing contrast, brightness, hue, saturation, noise in an image)
3. Random occlusion (for instance random cutout, erase, hide and seek, grid mask)
4. Schemes based on deep learning

Traditional data augmentation techniques, while effective for many computer vision tasks, can pose unique challenges when applied to small object detection. Small objects are inherently sensitive to perturbations, and augmentations that work well for larger objects may not be suitable in this context. Common augmentations like rotation, scaling, and translation, when applied aggressively, can easily cause small objects to disappear, overlap, or lose their structural integrity. Additionally, small objects may become indistinguishable from background clutter, resulting in increased false negatives. Therefore, striking a balance between augmenting the data to enhance model generalisation and preserving the integrity of small objects is a non-trivial task. Specialised techniques and considerations are often required to ensure that data augmentation effectively benefits small object detection models without introducing detrimental changes that hinder their performance.

Starting off with YOLOv4 [15], that uses the mosaic augmentation for small object detection in images for the first time. They join four images into one single image. Consequently, the objects in the joined image appear at a smaller scale than the original image. This kind of aug- mentation is conductive to ameliorating the detection of small objects in images. Also it is important to point that there is a chance of lost information or features of the original data, because these techniques attempt to change the geometry or lighting conditions of the images, making the task of small object detection even harder in some cases.

Deep learning methods nowadays seem to produce credible results, but in the case of small object detection two major difficulties persist. Common datasets contain only a few images with small objects, that also leads to lack of diversity in the locations of those small objects. To tackle these problems Kisantal et al. [36] proposed oversampling those images containing small objects as a solution for the first issue and copy-pasting small objects multiple times in each image as a solution for the second issue. However, there are limitations since it is not advisable to simply paste the cropped object randomly in the drone captured image, where an example would be a car flying in the sky. Thus, Chen et al. [42] introduce a novel adaptive data augmentation strategy called adaptive resampling AdaResampling to logically augment the data.

Unlike the methods above, Chen et al. [43] put forward a novel feedback-driven data provider called Stitcher. In the Stitcher, images are resized into smaller components and then stitched into the same size as regular images. Stitched images inevitably contain smaller objects, which would be beneficial to guide next-iteration update by utilising the loss statistics as feedback.

Capturing a substantial number of new images in any domain is generally regarded as a challenging endeavor. In this context, the utilization of data augmentation methods emerges as a time and cost-saving solution. Unlike the alternative approach of altering model architecture, which introduces additional complexity to inference processes, potentially slowing down models, data augmentation strategies circumvent such intricacies. Notably, these strategies do not impose any extra burden on inference complexity. Designing effective augmentation strategies for object detection poses a greater challenge compared to classification tasks. Despite this, data augmentation schemes have garnered relatively less research attention. This might be attributed to the perception that they contribute less to detection performance improvement and exhibit suboptimal transferability.

3.3.2 Super-Resolution

Super resolution can be applied as a preprocessing step to improve the quality of input images before they are fed into object detection algorithms. This enhancement can help object detectors to perform better, especially when dealing with low-quality or pixelated images, as it can potentially make objects more distinguishable and recognisable. A high-resolution image is of great benefit to small object detection because it provides more refined details about the original scene.

The Generative Adversarial Networks (GANs) [5] can be used to rebuild high-resolution images and it has achieved great progress in image super-resolution [6]. The basic architecture of a GAN consists of a generator and a discriminator. The generator yields super-resolved images to fool the discriminator while the discriminator attempts to distinguish the real images from fake images produced via the generator.

Possibly the first time a GAN was used in small object detection with the goal of providing better information was by Li et al. [7]. The proposed perceptual GAN model improves small traffic sign detection by generating super-resolved representations for small objects in order to narrow the representation difference between small and large objects. The details of the perceptual GAN are shown in 1. Specifically, the generator is a deep residual network which takes the low-level features as the input to capture more details for super-resolved representation. Multiple residual blocks in the generator are employed to learn the residual representation between small objects and similar large objects.

The discriminator takes the features of large object and the super-resolved representation of small object as inputs and splits into adversarial branch and perception branch. The adversarial branch contains three fully connected layers followed by a sigmoid layer, which is utilised to distinguish the generated super-resolved region of small objects from similar large objects. The perception branch contains two fully connected layers followed by two output sibling layers, which are employed for bounding box regression and classification respectively to justify the detection accuracy from the generated super-resolved representation.

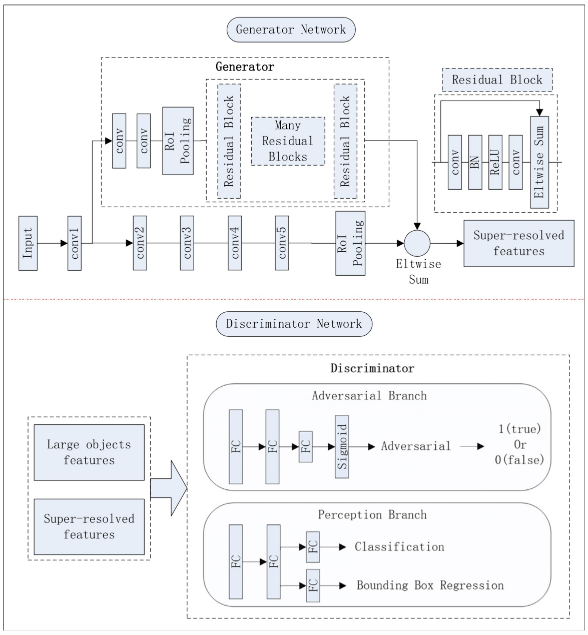

Figure 1: Perceptual GAN Architecture

To bridge the gap from the initial application of GAN and super-resolution in small object detection to the latest advancements, researchers have continually pushed the boundaries of innovation. The solutions provided can be separated in four abstract categories. Firstly some solutions implement Super-resolution by using the filter and refinement approach. Having performance in mind, the low resolution image is used to produce the Regions Of Interest (ROIs) and as a second step, those regions are super-resolved to get the best possible results, an idea found at Yang et al. [8] PKG, Liu et al. [41] TFPGAN and Yang et al. [14] QueryDet for image data and Bosquet et al. [1] STDnet for videos.

Secondly there are methods that focus on finding the differences between the features of small and big objects and how to use that difference in order to enhance the features in the small object detection. In this category the work of Pang et al. [9] JSC-Net, Zhang et al. MTGAN and Gu et al. [13] GDL stand out. Thirdly some methods use both the features of low and high resolution data with different depth of CNN's in order to retain the information needed from each case. Solutions that follow this process are Bai et al. [10], Krishna and Jawahar [4], Liu et al. [12] (HRDNet). Fourthly methods that attempt to super-resolve the features of the images like the work of Noh et al. [11] Feature SR.

The GAN-based approach effectively enhances the detail information of image and in principle, it can be applied to any kind of generator without devising specific architecture. Training a GAN presents challenges in achieving convergence as the interplay between the generator and discriminator must be delicately navigated, and an imbalance in their learning dynamics can impede the network from reaching a stable and effective state. Besides, the rewards of samples produced by the generator during the training process are limited, which will affect the further improvement of detection performance to a certain extent.

4 Feature Extraction and Representation

As mentioned in the introduction the small objects are more difficult to detect accurately, because small objects have low resolution and insufficient features. Effective feature extraction and representation are critical in small object detection, as they enable the model to focus on the relevant details while filtering out noise and irrelevant information.

Detecting objects in different scales is challenging in particular for small objects. A pyramid of the same image can be used at a different scale to detect objects. However, processing multiple scale images is time consuming and the memory demand is too high to be trained end-to-end simultaneously. Alternatively, a pyramid of features is created and used for object detection. However, feature maps closer to the image layer composed of low-level structures that are not effective for accurate object detection.

4.1 Feature Pyramid Network

Feature Pyramid Network-FPN is a feature extractor designed with accuracy and speed in mind, Tsung-Yi Lin et al. [44]. It replaces the feature extractor of detectors like Faster R-CNN and generates multiple feature map layers or multi-scale feature maps with better quality information than the regular feature pyramid for object detection. FPN consists of a bottom-up and a top-down pathway.

The top-down pathway uses higher resolution features by upsampling spatially coarser, but semantically stronger, feature maps from higher pyramid levels. These features are then enhanced with features from the bottom-up pathway via lateral connections. Each lateral connection merges feature maps of the same spatial size from the bottom-up path- way and the top-down pathway. The bottom-up feature map is of lower-level semantics, but its activations are more accurately localized as it was subsampled fewer times.

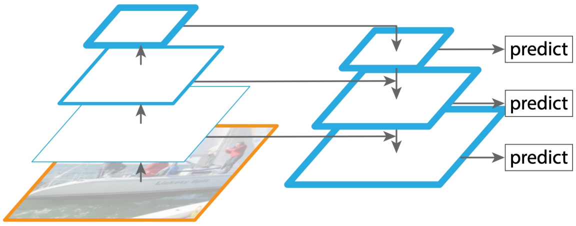

Figure 2: Top-down Architecture with skip connections

The bottom-up pathway is the usual convolutional network for feature extraction. Ascending in the spatial resolution decreases. With more high-level structures detected, the semantic value for each layer increases. SSD makes detection from multiple feature maps. However, the bottom layers are not selected for object detection. They are in high resolution but the semantic value is not high enough to justify its use as the speed slowdown is significant. So SSD only uses upper layers for detection and therefore performs much worse for small objects.

4.2 Spectral, Spatial, and Temporal Features

Tracking small objects amidst complex scenes poses greater challenges due to their variable appearances and limited feature representation. Transformer networks have been successful in object tracking in RGB videos, but transformer trackers typically utilize a single deep layer feature for object tracking. Small objects are highly squeezed on the deep layer with lower spatial resolution and limited discriminated information, which is crucial for small object tracking.

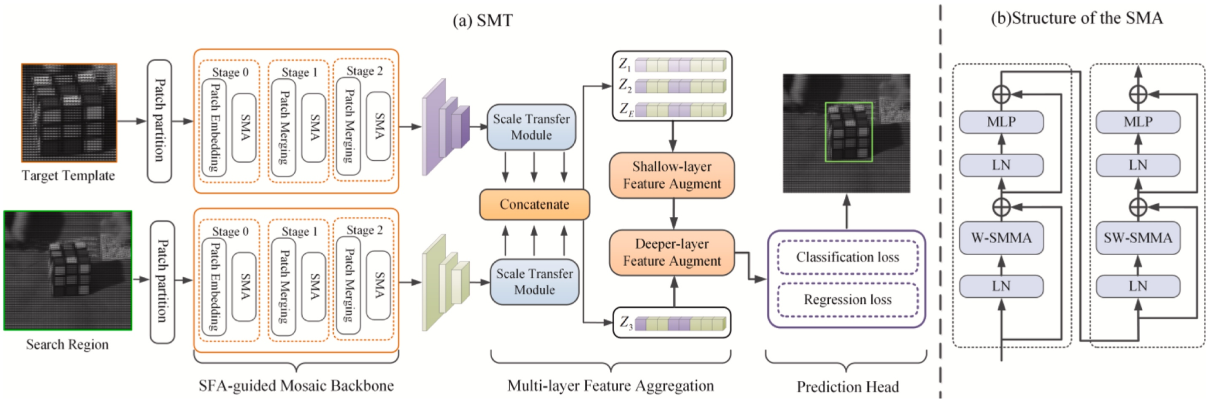

Figure 3: Spectral Mosaic Transformer

Spectral imaging, leveraging spectral and spatial information to characterize the material properties of objects, offers enhanced object feature discrimination compared to conventional visual imaging, making it an ideal choice for this task. Currently, small object detection techniques address the challenge of detecting small objects by leveraging multi-layer features. However, they reshape features from various layers to the same size before aggregating them, inevitably introducing information loss.

To tackle the stated issues Lulu Chen et al. [45] presented an end-to-end Spectral Filter Array(SFA) guided Mosaic Transformer (SMT) for tracking small objects in mosaic spectral video. SMT consists of three computing modules: SFA-guided mosaic backbone (SMB), Multi-layer Feature Aggregation (MFA), and Prediction head 3 shows the acquisition of a mosaic spectral image with snapshot spectral imaging. This indicates that spectral information in mosaic spectral image is position sensitive, providing a great solution for both the classification and localisation challenges of small object detection.

5 Detection Algorithms

Two main types of detectors have emerged as prominent solutions in addressing the problem of Small Object Detection, being the single-shot detectors and multi-shot detectors. Both types of detectors play crucial roles in the evolving landscape of small object detection, each offering distinct advantages depending on the specific requirements of the application.

5.1 Single-Shot Detectors

Single-shot detectors, exemplified by models like You Only Look Once (YOLO) and Single-Shot MultiBox Detectors (SSD), are known for their efficiency and real-time processing capabilities. They perform object localization and classification in a single pass, making them suitable for applications that require low-latency responses.

The key idea behind YOLO et al. [46] is to divide the input image into a grid and predict bounding boxes and class probabilities for each grid cell. The network outputs a fixed number of bounding boxes along with their associated class probabilities and confidence scores. The confidence score reflects how certain the algorithm is about the predicted bounding box containing an object. YOLO also employs anchor boxes to improve the accuracy of bounding box predictions for objects of different sizes and aspect ratios.

Single Shot MultiBox Detector (SSD) et al. [47] is another popular object detection algorithm designed for real-time applications by [] The main components of SSD include a base convolutional neural network (CNN) that extracts features from the input image and a set of auxiliary convolutional layers at different scales. These auxiliary layers are responsible for predicting bounding boxes and class scores for objects of varying sizes. By incorporating features from multiple scales, SSD is able to detect objects of different sizes more effectively than some earlier methods.

One of the key advantages of SSD is its ability to handle a diverse set of object scales in a single pass. It utilizes a set of default bounding boxes (prior boxes) at different aspect ratios and scales

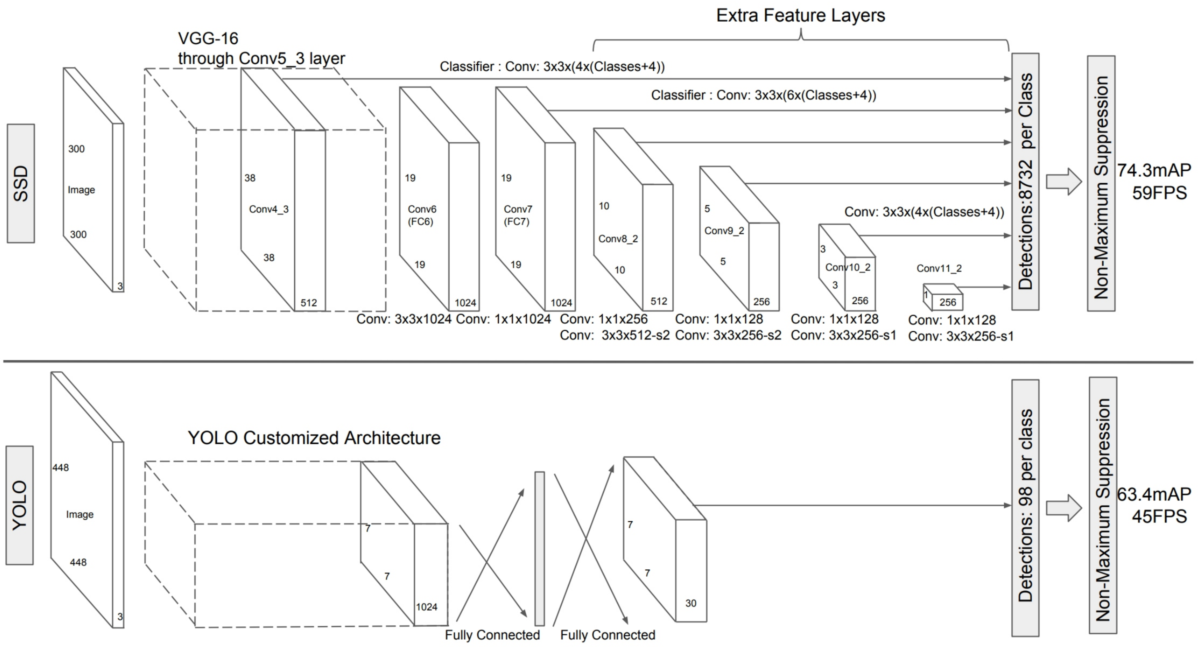

Figure 4: Single-Shot Detection Models Architecture

to predict the final bounding boxes. This makes SSD well-suited for detecting objects of different sizes in a variety of scenarios.

The advantages of using a Single-Shot model are the following:

1. Efficient and fast, suitable for real-time applications.
2. Simultaneous localization and classification in a single pass.
3. Good for detecting small objects with varying aspect ratios.

While the main disadvantages of using those methods are:

1. May struggle with accurately localizing small objects due to coarse grid.
2. Can have lower precision compared to two-stage detectors.
3. Sensitive to object scale changes within the same image.

5.2 Multi-shot Detectors

On the other hand, multi-shot detectors, represented by architectures such as Faster R-CNN et al. [48], adopt a two-stage approach involving region proposals and subsequent refinement.

Faster R-CNN consists of two main components: a Region Proposal Network (RPN) and a Fast R-CNN detector. The RPN generates region proposals or candidate bounding boxes for potential objects in the image, and the Fast R-CNN detector refines and classifies these proposals.

The RPN is responsible for suggesting potential object regions in an image. It operates on the convolutional feature maps generated by a shared convolutional backbone network. The RPN predicts bounding box proposals and assigns objectness scores to these proposals. The proposals with high objectness scores are then passed on for further processing.

The Fast R-CNN takes the region proposals from the RPN and refines them for more accurate localization. It also predicts the class of the objects within these regions. The shared convolutional backbone is utilized to extract features from the proposed regions, and these features are used for both bounding box regression and class prediction.

Just like the Single-Shot models the Multi-Shot detection models have positives and negatives:

1. Achieves high detection accuracy, especially for small objects.

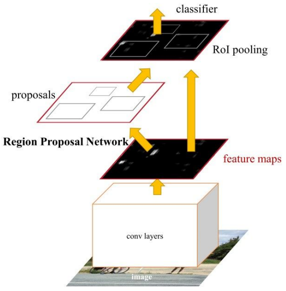

Figure 5: Faster R-CNN Architecture

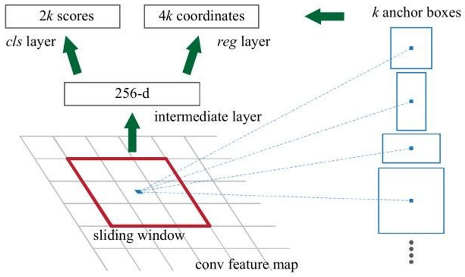

Figure 6: Enter Caption

2. Two-stage approach allows for more accurate localization.

3. Effective feature representation through region proposals.

The negative aspects being:

1. Slower compared to single-shot detectors.

2. More complex architecture and training process.

3. Can be sensitive to changes in object scale and aspect ratio.

5.3 Anchor-Based and Anchor-Free Approaches

Anchor Mechanism

The anchor is widely adopted by most of the existing detectors. Faster R-CNN [2] introduces the Region Proposal Network (RPN) to generate proposals. The RPN is based on anchors, which are predefined regions of different sizes and aspect ratios to handle multiple scales. The RPN produces the coordinates of the bounding boxes and their corresponding categories, namely object and background. Finally, given the output of the RPN and the last feature map of the feature extractor, the bounding box and category of the object are determined by a fully-connected classification network.

Krishna and Jawahar [4] formulate finding that appropriate sizes of the anchor boxes mathematically and perform detailed experiments to reveal the effectiveness in their choice. By introducing the expected max overlapping (EMO) score, the authors in [25] calculate the expected max intersection over union (IoU) between anchor and object. They find the smaller stride of the anchor (SA) is, the higher EMO score achieves, statistically leading to improved average max IoU of all objects. It is noted that a smaller SA can sample more high-quality samples well capturing the small objects, which is of help for both detector training and testing.

Zhang et al. [27] develop an anchor-based face detector, which only outputs a single high-resolution feature map with small anchors, to specifically learn small faces and train it via a new hard image mining scheme which automatically adjusts training weights on images according to their difficulties. It is necessary to tile massive dense anchors on high-resolution feature map for pursuing high recall. However, it results in an extreme imbalance of category that drastically impacts the classification task in the detection framework.

An adaptive anchor tiling strategy, like MetaAnchor [28] and Guided Anchor [29], is proposed to shrink search space efficiently. Specifically, Yang et al. [28] present a novel anchor mechanism called MetaAnchor for object detection. Unlike many previous detectors model anchors by a predefined manner, anchor functions in the MetaAnchor could be dynamically produced from the arbitrary customised prior boxes. In this way, they empirically find that MetaAnchor is more robust to anchor settings and bounding box distributions.

Zhang et al. [26] propose an asymmetric multistage network (AMS-Net), which considers the asymmetry of a pedestrian's body shape in small-scale pedestrian detection. The rectangular anchors are utilised to produce various rectangular proposals that have a height greater than the width. Besides, asymmetric rectangular convolution kernels are adopted to capture the compact features for the pedestrian body, as seen at the figure below 7.

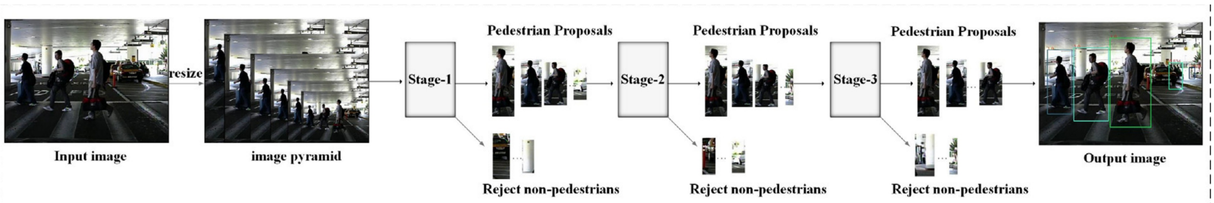

Figure 7: Asymmetric multi-stage network

Anchor-Free Mechanism In addition to the anchor-based methods, some researchers discard the prior anchors and they utilise anchor-free approaches for object detection. Law et al. [31] propose CornerNet, a new method for object detection. They detect an object bounding box as paired keypoints, namely the top-left corner and the bottom-right corner, using a single CNN. They also introduce corner pooling to help the network localise corners better. Unlike CornerNet, Lu et al. [32] introduce a new detection framework called Grid R-CNN, which employs a grid guided localisation mechanism for accurate object detection. Instead of utilising only two independent points, they devise a multi-point supervision formulation to encode more clues so as to reduce the influence of inaccurate prediction of specific points.

A well-devised region proposal strategy can take advantage of limited anchor size and anchor amount, reduce computational cost in producing interested region, and efficiently detect small objects. Hence, these well-designed methods adopt predefined anchors to enumerate possible locations, scales and aspect ratios for the search of the objects, which are conducive to the detection of small objects to a certain extent. These approaches are anchor box free, as well as proposal free. The usage of anchor introduces a large number of hyper-parameters, which makes the network hard to train. Besides, improper usage of anchor could cause imbalance between the positive and negative samples of small objects, which makes the model pay more attention to large objects. This is not beneficial for the detection improvement of small objects. The existing anchor design is difficult to substantially balance the contradiction between the detection accuracy and the calculation cost for small objects.

The main advantages of Anchor-based and Anchor-Free detection algorithms are the following:

1. **Anchor-based**: Explicitly considers object scales and aspect ratios.

2. **Anchor-free**: More flexible in handling object size variations.

while the disadvantages are:

1. **Anchor-based**: Choosing appropriate anchors can
be challenging.

2. **Anchor-free**: May require more complex network architectures.

3. Training complexity and computational requirements.

6 Challenges and Trends in the Field

6.1 Challenges

As mentioned in the sections above, a multitude of challenges must be addressed in the realm of small object detection. These challenges encompass various aspects, each posing unique hurdles to the development of effective detection models.

Firstly, the limited spatial information associated with small objects in an image presents a significant obstacle. These objects often lack distinct features and can be easily overshadowed by the background or other elements, leading to false negatives. Additionally, the issue of occlusion compounds the complexity of small object detection, as these objects are more prone to being partially or completely obscured by other elements or obstacles. Conventional detection methods may falter when faced with the task of identifying occluded small objects.

Scale variability further complicates the matter, as small objects can exhibit significant size variations within the same image. Robust algorithms and architectures are required to detect small objects across a broad range of scales. The selection of appropriate anchors and aspect ratios also emerges as a challenge, with poor choices potentially resulting in missed detections or high false-positive rates.

Moreover, the demand for computational efficiency in small object detection adds another layer of complexity. Real-time or near-real-time processing is often necessary, requiring a delicate balance between detection accuracy and computational resources—especially in applications with limited resources.

The process of data annotation for small objects introduces its own set of challenges. Accurately annotating small objects in datasets is a time-consuming and error-prone task, emphasizing the need for meticulous annotation to train reliable models. Lastly, achieving generalization across various scenes, lighting conditions, and object variations is a non-trivial task in small object detection, underscoring the importance of robust training methodologies. In essence, addressing these challenges is crucial for advancing the capabilities of small object detection systems.

6.2 Trends in the Field

In the dynamic landscape of small object detection, several notable trends are shaping the evolution of detection methodologies, ushering in new possibilities and addressing existing challenges.

Firstly, the field is witnessing a profound impact from ongoing advancements in deep learning. The continuous refinement of deep learning architectures, coupled with innovative pretraining techniques, has led to the emergence of more powerful models. Noteworthy examples include EfficientDet and Deformable ConvNets, which are gaining popularity for their efficacy in addressing the intricacies of small object detection.

Another significant trend involves the rise of anchor-free approaches. Unlike traditional methods reliant on predefined anchors, these novel approaches offer increased adaptability to the variations inherent in small objects. This adaptability is proving crucial in enhancing detection performance. Attention mechanisms have also taken center stage in the quest for improved small object detection accuracy. Models incorporating self-attention mechanisms and feature fusion techniques demonstrate promise by enabling focused analysis of relevant details within an image.

The utilization of spectral-spatial data has emerged as a prominent trend in small object detection, enabling more accurate and nuanced identification of diminutive objects by leveraging both spectral and spatial information within the data

Lastly a noteworthy development is the emergence of few-shot and zero-shot learning techniques. These approaches aim to empower models to detect small objects with minimal or no labeled examples, alleviating the burden of data annotation while expanding the applicability of detection models. In tandem with these advancements, the importance of robust evaluation metrics is gaining recognition.

7 Conclusion

In conclusion, the landscape of small object detection is marked by a multitude of challenges that demand innovative solutions for continued progress in the field. The difficulties posed by limited spatial information, occlusion, scale variability, anchor and aspect ratio selection, semantic context integration, data annotation, and generalization underscore the complexity of detecting small objects within images. Addressing these challenges requires a holistic approach that combines advancements in algorithm design, data annotation methodologies, better feature extraction and computational efficiency. As the demand for small object detection grows across various applications, from surveillance systems to autonomous vehicles, the resolution of these challenges becomes paramount to ensuring the reliability and effectiveness of detection systems.

Simultaneously, the field is witnessing promising trends that hold the potential to reshape the small object detection landscape. The adoption of anchor-free approaches, attention mechanisms, transfer learning from pretrained models, few-shot and zero-shot learning methodologies, efficient hardware acceleration, and the development of robust evaluation metrics contribute to the adaptability and generalization capabilities of detection systems. As researchers and practitioners continue to explore these trends, they pave the way for a future where small object detection systems are more accurate, efficient, and adaptable to the dynamic nature of real-world scenarios. The utilization of spectral-spatial data, recognized as a prominent trend, emerges as a promising solution that, while currently in its early stages, holds significant potential for more accurate and nuanced identification of diminutive objects. As these advancements unfold, the trajectory of small object detection systems points toward increased accuracy and adaptability, positioning them as indispensable tools in various domains.

References

[1] B. Bosquet, M. Mucientes, V.M. Brea. 2020. *STD-net: exploiting high resolution feature maps for small object detection*, *Eng. Appl. Artif. Intell.*, 91 (2020), Art. no. 103615.

[2] S. Ren, K. He, R. Girshick, J. Sun. 2015. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*, *Advances in Neural Information Processing Systems*, Montreal, Quebec, Canada, 2015.

[3] T.-Y. Lin, P. Goyal, R.B. Girshick, K. He, P. Dollar. 2017. *Focal Loss for Dense Object Detection*, *IEEE International Conference on Computer Vision*, Venice, Italy, 2017.

[4] H. Krishna, C.V. Jawahar. 2017. *Improving Small Object Detection*, *Asian Conference on Pattern Recognition*, Nanjing, China, 2017.

[5] I.J. Goodfellow, et al. 2014. *Generative Adversarial Nets*, *Neural Information Processing Systems*, Montreal, Quebec, Canada, 2014.

[6] C. Ledig, et al. 2017. *Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network*, *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017.

[7] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, S. Yan. 2017. *Perceptual Generative Adversarial Networks for Small Object Detection*, *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017.

[8] Z. Yang, et al. 2019. *Prior Knowledge Guided Small Object Detection on High-Resolution Images*, *IEEE International Conference on Image Processing*, Taipei, Taiwan, 2019.

[9] Y. Pang, J. Cao, J. Wang, J. Han. 2019. *JCS-Net: joint classification and super-resolution network for small-scale pedestrian detection in surveillance images*, *IEEE Trans. Informat Forensics Security*, 14 (12), 3322–3331.

[10] Y. Bai, Y. Zhang, M. Ding, B. Ghanem. 2018. *Finding Tiny Faces in the Wild With Generative Adversarial Network*, *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[11] J. Noh, W. Bae, W. Lee, J. Seo, G. Kim. 2019. *Better to Follow, Follow to Be Better: Towards Precise Supervision of Feature Super-Resolution for Small Object Detection*, *IEEE International Conference on Computer Vision*, Seoul, South Korea, 2019.

[12] Z. Liu, G. Gao, L. Sun, Z. Fang. 2021. *HRD-Net: High-resolution Detection Network for Small Objects*, *IEEE International Conference on Multimedia and Expo*, Shenzhen, China, 2021.

[13] Y. Gu, J. Li, C. Wu, W. Jia, J. Chen. 2020. *Small Object Detection by Generative and Discriminative Learning*, *International Conference on Pattern Recognition*, Milan, Italy, 2020.

[14] C. Yang, Z. Huang, N. Wang. 2021. *Query-Det: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection*, *arXiv:2103.09136v1*, 2021.

[15] A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao. 2020. *YOLOv4: optimal speed and accuracy of object detection*, 2020, Art. no. arXiv:2004.10934.

[16] P. Hu, D. Ramanan. 2017. *Finding Tiny Faces*, *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017.

[17] C. Zhang, T. Li, S. Guo, N. Li, Y. Gao, K. Wang. 2019. *Aggregation Connection Network For Tiny Face Detection*, *IEEE International Joint Conference on Neural Network*, Budapest, Hungary, 2019.

[18] X. Tang, D.K. Du, Z. He, J. Liu. 2018. *Pyramid-Box: A Context-Assisted Single Shot Face Detector*, *European Conference on Computer Vision*, Munich, Germany, 2018.

[19] Z. Li, X. Tang, J. Han, J. Liu, R. He. 2019. *PyramidBox++: High Performance Detector for Finding Tiny Face*, *arXiv:1904.00386v1*, 2019.

[20] J. Liu, et al. 2019. *Multi-component fusion network for small object detection in remote sensing images*, *IEEE Access*, 7 (2019), 128339–128352.

[21] M. Hong, S. Li, Y. Yang, F. Zhu, Q. Zhao, L. Lu. 2021. *SSPNet: Scale Selection Pyramid Network for Tiny Person Detection from UAV Images*, *arXiv:2107.01548v1*, 2021.

[22] S. Liu, D. Huang, Y. Wang. 2018. *Receptive Field Block Net for Accurate and Fast Object Detection*, *European Conference on Computer Vision*, Munich, Germany, 2018.

[23] Y. Li, Y. Chen, N. Wang, Z.-X. Zhang. 2019. *Scale-Aware Trident Networks for Object Detection*, *IEEE International Conference on Computer Vision*, Seoul, South Korea, 2019.

[24] T.-Y. Lin, P. Dollár, R.B. Girshick, K. He, B. Hariharan, S.J. Belongie. 2017. *Feature Pyramid Networks for Object Detection*, *IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, 2017.

[25] C. Zhu, R. Tao, K. Luu, M. Savvides. 2018. *Seeing Small Faces From Robust Anchor's Perspective*, *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[26] S. Zhang, X. Yang, Y. Liu, C. Xu. 2020. *Asymmetric multi-stage CNNs for small-scale pedestrian detection*, *Neurocomputing*, 409, 12–26.

[27] Z. Zhang, W. Shen, S. Qiao, Y. Wang, B. Wang, A.L. Yuille. 2020. *Robust Face Detection via Learning Small Faces on Hard Images*, *IEEE Winter Conference on Applications of Computer Vision*, Snowmass Village, CO, USA, 2020.

[28] T. Yang, X. Zhang, Z. Li, W. Zhang, J. Sun. 2018. *MetaAnchor: Learning to Detect Objects with Customized Anchors*, *Advances in Neural Information Processing Systems*, Montréal, Canada, 2018.

[29] J. Wang, K. Chen, S. Yang, C.C. Loy, D. Lin. 2019. *Region Proposal by Guided Anchoring*, *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019.

[30] K. Duan, D. Du, H. Qi, Q. Huang. 2020. *Detecting small objects using a channel-aware deconvolutional network*, *IEEE Trans. Circuits Syst. Video Technol.*, 30 (6), 1639–1652.

[31] H. Law, J. Deng. 2018. *CornerNet: Detecting Objects as Paired Keypoints*, *European Conference on Computer Vision*, Munich, Germany, 2018.

[32] X. Lu, B. Li, Y. Yue, Q. Li, J. Yan. 2019. *GridR-CNN*, *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019.

[33] S. Luo, X. Li, R. Zhu. 2019. *SFA: small faces attention face detector*, *IEEE Access*, 7 (2019), 171609-171620.

[34] C. Gao, W. Tang, L. Jin, Y. Jun. 2020. *Exploring Effective Methods to Improve the Performance of Tiny Object Detection*, *European Conference on Computer Vision Workshops*, Glasgow, UK, 2020.

[35] B. Singh, L.S. Davis. 2018. *An Analysis of Scale Invariance in Object Detection SNIP*, *IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018.

[36] M. Kisantal, Z. Wojna, J. Murawski, J. Naruniec, K. Cho. 2019. *Augmentation for Small Object Detection*, *The 9th International Conference on Advances in Computing and Information Technology*, Sydney, Australia, 2019.

[37] J. Wu, C. Zhou, Q. Zhang, M. Yang, J. Yuan. 2020. *Self-Mimic Learning for Small-scale Pedestrian Detection*, *ACM International Conference on Multimedia*, Seattle, WA, USA, 2020.

[38] K. Chen, et al. 2019. *Towards Accurate One-Stage Object Detection With AP-Loss*, *IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019.

[39] Q. Qian, L. Chen, H. Li, R. Jin. 2020. *DRLoss: Improving Object Detection by Distributional Ranking*, *IEEE Conference on Computer Vision and Pattern Recognition*, Seattle, WA, USA, 2020.

[40] H. Zhang, H. Chang, B. Ma, S. Shan, X. Chen. 2019. *Cascade RetinaNet: Maintaining Consistency for Single-Stage Object Detection*, *British Machine Vision Conference*, Cardiff, UK, 2019.

[41] D. Liu, Z.-Q. Zhao, W. Tian 2020. *TFPGAN: Tiny Face Detection with Prior Information and GAN*, *presented at the International Conference on Intelligent Computing*, Bari, Italy, 2020.

[42] C. Chen 2019. *RRNet: A Hybrid Detector for Object Detection in Drone-Captured Images*, *presented at the IEEE International Conference on Computer Vision Workshops*, Seoul, South Korea, 2019.

[43] Y. Chen 2019. *Stitcher: feedback-driven data provider for object detection*, *Comput. Res. Reposit. (2020) arXiv:2004.12432v1.*

[44] Tsung-Yi Lin 2017 CVPR *Feature Pyramid Networks for Object Detection*, *Computer Vision and Pattern Recognition (2017) arXiv:1612.03144.*

[45] Lulu Chen 2023 ISPRS *FSFA-guided mosaic transformer for tracking small objects in snapshot spectral imaging*, *ISPRS Journal of Photogrammetry and Remote Sensing*

[46] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi 2016 *You Only Look Once: Unified, Real-Time Object Detection*, *Computer Vision and Pattern Recognition 9 May 2016 arXiv:1506.02640*

[47] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, Alexander C. Berg 2016 *SSD: Single Shot MultiBox Detector*, *Computer Vision and Pattern Recognition 29 Dec 2016 arXiv:1512.02325*

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun 2016 *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*, *Computer Vision and Pattern Recognition 6 Jan 2016 arXiv:1506.01497*

[49] Liu, C., Yang, L. 2021 *Tracker evaluation for small object tracking*, *Computer Vision and Pattern Recognition 6 Jan 2016 arXiv:1506.01497*

[50] Garioud A., Valero, S., Giordano, S., Mallet, C. 2021 *Sentinel time series for continuous vegetation monitoring. Rem. Sens. Environ. 263, 112419.*

[54] Jiaxu Leng, Yihui Ren, Wen Jiang, Xiaoding Sun, Ye Wang 2020 *Realize your surroundings: Exploiting context information for small object detection*, *Neurocomputing 19 December 2020*

[54] Karthick Prasad Gunasekaran 2023 *Leveraging object detection for the identification of lung cancer*, *Computer Vision and Pattern Recognition 25 May 2023 arXiv:2305.15813v1*