applied sciences

Review

A Review of Recent Advances and Challenges in Grocery Label Detection and Recognition

Vânia Guimarães [1,2], Jéssica Nascimento [1,3], Paula Viana [1,4] and Pedro Carvalho [1,4,*]

[1] Centre for Telecommunications and Multimedia at INESC TEC, Institute for Systems and Computer Engineering, Technology and Science, 4200-465 Porto, Portugal
[2] Faculty of Sciences, University of Porto, 4169-007 Porto, Portugal
[3] Faculty of Engineering, University of Porto, 4200-465 Porto, Portugal
[4] Instituto Superior de Engenharia do Porto (ISEP), School of Engineering, Polytechnic of Porto, 4249-015 Porto, Portugal
* Correspondence: pedro.m.carvalho@inesctec.pt; Tel.: +351-222094000

Abstract: When compared with traditional local shops where the customer has a personalised service, in large retail departments, the client has to make his purchase decisions independently, mostly supported by the information available in the package. Additionally, people are becoming more aware of the importance of the food ingredients and demanding about the type of products they buy and the information provided in the package, despite it often being hard to interpret. Big shops such as supermarkets have also introduced important challenges for the retailer due to the large number of different products in the store, heterogeneous affluence and the daily needs of item repositioning. In this scenario, the automatic detection and recognition of products on the shelves or off the shelves has gained increased interest as the application of these technologies may improve the shopping experience through self-assisted shopping apps and autonomous shopping, or even benefit stock management with real-time inventory, automatic shelf monitoring and product tracking. These solutions can also have an important impact on customers with visual impairments. Despite recent developments in computer vision, automatic grocery product recognition is still very challenging, with most works focusing on the detection or recognition of a small number of products, often under controlled conditions. This paper discusses the challenges related to this problem and presents a review of proposed methods for retail product label processing, with a special focus on assisted analysis for customer support, including for the visually impaired. Moreover, it details the public datasets used in this topic and identifies their limitations, and discusses future research directions of related fields.

Keywords: retail; grocery products; computer vision; object detection; object recognition; text detection; text recognition; product label analysis

check for updates

Citation: Guimarães, V.; Nascimento, J.; Viana, P.; Carvalho, P. A Review of Recent Advances and Challenges in Grocery Label Detection and Recognition. *Appl. Sci.* **2023**, *13*, 2871. https://doi.org/10.3390/app13052871

Academic Editor: António J. R. Neves

Received: 13 January 2023
Revised: 11 February 2023
Accepted: 15 February 2023
Published: 23 February 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

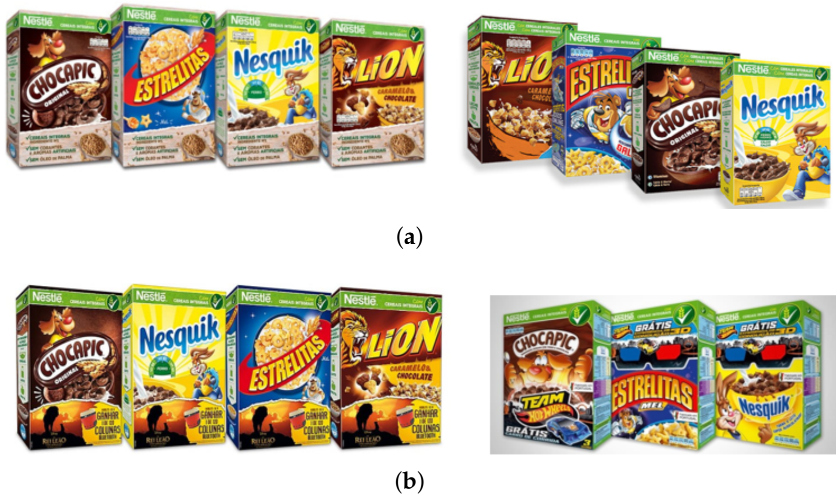
1. Introduction

The retail industry has been integrating technological innovations throughout its product chain in order to reduce costs and, cumulatively, improve customer experience, since increasing profit margins and attracting customers are the primary goals of any entrepreneur. In the retail context, the automatic detection and recognition of products has allowed a more efficient use of resources and revolutionised the way customers buy. They want the purchase process to be simple and fast; therefore, they value payout and tools that allow them to find desired items, know the product's availability and avoid payment queues. Automatic self-checkout systems are the solution to fill the last need. Real-time inventory management, in particular automatic shelf monitoring, is also possible with computer vision tools, with out-of-stock shelves being detected in real-time by capturing images of racks. By the same philosophy, there is an opportunity to verify if the product displays and store layout follow the plan known as a planogram. Self-service technology is being applied to create autonomous stores and to permit a self-guided shopping experience, and product recognition devices can add value to the customer experience by assisting them in the correct purchase of the product.

In a grocery environment, food and nutrition labelling are extremely important, as the information allows consumers to make more conscious food choices, appropriate to their needs and preferences, contributing to the correct storage, preparation and consumption of food. Although interpreting grocery labels may not be an easy task, either due to a lack of food literacy of consumers to understand some of the available information or due to other types of limitations, including visual clutter. This motivates the development of auxiliary tools, democratizing access to information and developing inclusive strategies; particularly, visually impaired people have strong limitations in carrying out daily activities, which include grocery shopping independently. Several studies have addressed this issue, focusing on the detection or recognition of specific products, often under unique conditions. Automatic grocery product detection and recognition in a real environment raise many challenges [1,2], which are even more relevant in the case of images captured by visually impaired people as this disability prevents them from understanding if the image includes the product or if it is legible and readable. As a result, the items can appear in arbitrary poses and perspectives, cropped, with partial occlusions, with great differences of illumination, with reflection due to glossy product packages, at various distances from the camera, in a cluttered background, or in a blurred image due to camera shake, among others. This will have an impact on the performance of the implemented solution as the quality of the image is the basis for successful product recognition.

When taking photos or filming grocery items, some methods apply image- or frame-processing strategies to select or improve image quality, such as multi-frame super-resolution techniques. Others frameworks use a system beep to warn the user when a product or label is detected. Beyond the problem of the quality of the image, recognition of similar logos and different products with identical appearances is challenging. Identifying subtle details is critical for a fine-grained product classification, since small variations in packaging are very common among products of the same category. Classification between a product's subcategory can be a difficult task, even for humans, for example, when identifying several flavours of cereals, differences in the quantity of shower gel, or the type of hair of the same shampoo brand. Intraclass product recognition systems must differentiate such minor variations in an uncontrolled environment. Furthermore, new products are launched frequently and their appearance is likely to change over time to attract consumers' attention; also, due to marketing strategies, the same product may present slight variations in the packaging (e.g., during a promotional campaign) or the items may be integrated to be sold together. Figures 1 and 2 represent these common challenges for fine-grained classification.

Deep learning and Convolution Neural Networks (CNN) have been shown to be successful in object recognition, much like in many other application scenarios. Different levels of automatically learned features have proven to be more accurate and discriminatory than manually extracted features. Although deep learning methods are extensively used, there are two main drawbacks: (1) the performance is impaired by small training sets and (2) when the model is trained to learn new classes/tasks the previously learned classes/tasks are not preserved. In the grocery context, retraining the complete network with all classes when a new product appears in the market may be unfeasible since even the smallest supermarket has thousands of products available [3]. The wide variety of articles is, on the other hand, a constraint in creating an adequate dataset. Deep neural networks outperform traditional detection and recognition techniques, but they are often bottlenecked by limited datasets. The amount of labelled data is as important as the quality. Ideally, the images of each product should be taken in a real environment, from different perspectives and lighting conditions, in multiple retail stores. Manually annotating the object's location and providing, in detail, the attributes of each product (e.g., brand, flavour, type, size, etc.) are expensive tasks. It becomes even more cumbersome when all the textual information on the package is to be provided. For all these reasons, databases are limited and under-represented. Some researchers try to propose adaptable frameworks, and several systems introduce data augmentation during the training, e.g., models [4,5] generate synthetic context with the generative adversarial network (GAN) [6], some resort to transfer learning [7] and others [8,9] used a reference database with images of products taken under a controlled environment from several points of view for the training phase to mitigate this gap. The match classification approach is also convenient to handle new grocery products. Even after the minimisation of this issue, there are still challenges with the object's detection and identification, given the domain shift, complex background, image quality, intra-class similarity and limitations of computational resources. For recognizing grocery products, visual features are no longer the only option. Today, textual information—or a combination of the two—is already commonplace.



(**a**)

(**b**)

Figure 1. Illustration of differences introduced by marketing strategies, namely differences in flavours between same parent brands (**a**), and small variations in the packaging of the same product (**b**).

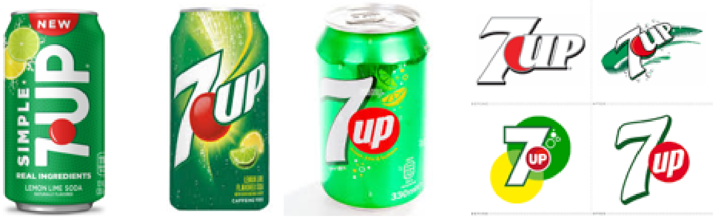


Figure 2. Same brand with a different appearance and orientation. Logos extracted from LogoDet-3K dataset [10].

Regardless of the techniques that each method chooses to overcome the different challenges, the system should assist the buyer throughout the process of selecting and taking the product. In other words, it should detect and recognise the products displayed on the racks, based on images or videos captured by the user, and guide him to acquire the correct product. After determining the product category, the client should be provided with more information about subcategories, such as sub-brands, flavours, types and quantities. All this relevant information is available on the label, although sometimes not in the best way. Therefore, in the next stage, the system should detect the label and identify the regions of interest (RoI), which could be logos, alphanumeric characters, words and symbols. The extracted RoI is to be processed to translate useful information. Finally, the classifier identifies the product based on text features or recognised texts. A conceptual pipeline summarizing these steps is represented in Figure 3.

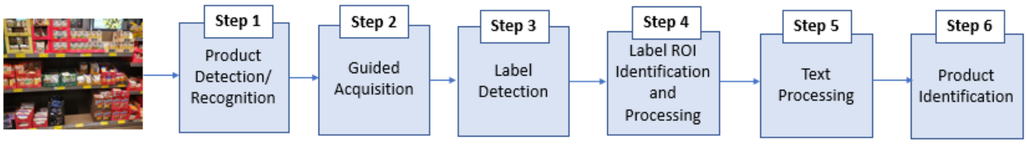


Figure 3. Conceptual pipeline of assisted retail product label analysis.

In the literature, there are some reviews of product recognition approaches in retail stores. The authors of [11], focused on stock tracking and planogram matching, and described classification systems of products on shelves; however, only traditional methods were covered. In [12], the authors conducted a comprehensive review of the detection methods of retail products. Despite including deep learning methods, the feature descriptors are predominantly hand-crafted. A more recent paper [1] presented a review of publications based on deep learning retail product recognition, without emphasizing the difficulties and solutions raised in the detection of objects in the supermarket. This extensive survey includes several areas of retail recognition, such as self checkout, stock management, planogram and products on shelves, but label information was not considered. The addition of text information has improved the ability to distinguish between visually identical products, but the research based on this perspective of retail product recognition is just starting. Therefore, reviewing grocery label analysis systems and complementary information about scene text recognition can help advance this new field.

This paper provides a complete overview of suggested approaches for grocery product label processing, with a particular emphasis placed on assisted analysis for customer support, discussing the main problems and shortcomings and suggesting future research directions. It also offers a general panorama of technological applications. The execution of new applications requires adequate datasets; hence, we included an extensive list of publicly available datasets that can be used by researchers. Unlike other previous reviews, this paper addresses techniques for the full pipeline, from product detection on the shelf to label processing, including individual object selection automatic image acquisition and label detection.

The organization of the rest of the paper is as follows: Section 2 identifies some existing applications of retail product recognition. Section 3 illustrates techniques of pre-processing and image selection. Section 4 describes the available datasets, both those related to grocery products and those related to text recognition. Section 5 reviews the works for grocery product detection and recognition, while Section 6 introduces an overview of product label analysis. Finally, Section 7 analyses the current challenges and offers some guidelines for future research.

2. Technological Review

The use of technology has evolved and the use of smartphones in daily routine has increased significantly. The extraordinary capabilities of these devices revolutionised, among others, the applicability of computer vision systems. The broad use of smartphones in our daily lives makes mobile applications accessible, easy to implement and portable, and has boosted the emergence of an ever-increasing number of apps. Nevertheless, the introduction of grocery product recognition systems is not limited to mobile phones. In this section, we briefly described different kinds of applications that have incorporated a grocery product recognition system. These are mainly divided into two methods: barcode-based systems (Section 2.1); label-based systems (Section 2.2).

2.1. Barcode Based System

A barcode is a unique identifier for each product, which is currently easily readable and with few probabilities of error. Hence, most grocery product recognition applications use this technology to identify the product. Usually, the user points the reader (e.g., a camera) at the barcode until the software detects it. The reader typically translates the information and searches for the item in a database. The database is not stored in the device, so these apps require an internet connection to identify items. Another drawback is that the barcode is placed in an arbitrary position which leads to the constant need to look up the location of the barcode. Some barcode-based applications that use object detection/recognition algorithms are listed below.

RoboCart [13]

A robotic shopping assistant for the visually impaired proposed in 2005 that guides users through grocery aisles and identifies the products by a barcode reader. Among some limitations, it is not an easily portable device.

Seeing AI [14]

An app launched by Microsoft, designed for blind and visually impaired users. Users can use it to recognise documents, products, people, scenes and currency, among other features. The identification of a product is possible through a barcode scanner. Since locating the barcode is challenging for people with limited or no vision, the app provides a location guide. The user should rotate the product, and the movement is guided by a sequence of tones that grows more rapidly as he gets closer to the barcode. When the product is identified, the system provides some available information, such as name, weight and ingredients.

Yuka [15]

This app makes consumers aware of the product labels, and ultimately allows them to make healthier choices, through the item's barcode. For that purpose, it provides detailed information about the product and a score out of 100 for food items, and classification into four risk categories for cosmetics. There is a possibility to activate the offline mode, with the constraint of downloading the product database to the mobile phone, reducing the memory space and later becoming outdated.

Open Food Facts [16]

A collaborative project designed to help consumers decipher food labels and make more conscious food choices regarding the impact on their health and the planet. Beyond the brands, ingredients and allergens, the consumer can also find the Nutri-score, the NOVA rank, and the carbon footprint. The application uses a barcode scanner to identify the product and provide some information.

2.2. Label-Based System

Previously described solutions rely on an internet connection, which may not be possible. This problem can be overcome by using computer vision algorithms that recognise the detected label based on visual features or textual information.

Lookout—Assisted Vision [17]

An application released by Google with the purpose of helping visually impaired people explore their surroundings. The system enables six kinds of activities: image description, text reading, food package identification, page capture, currency identification and object information. The system has two ways to identify the food package: via a barcode scanner, and the front of a food package recognition. The user should put the product in front of the camera to use the food label reader; if the product is in the wrong position, the system will warn the user to move the product until the front of the item is clearly visible. By downloading the database, the recogniser can work offline, with the inconvenience of reducing free memory space. The main limitation of this application is that it only recognises the front side of the package.

Alexa Accessibility—Show and Tell on the Echo Show [18]

Amazon Echo Show is a smart display, similar to a laptop, which is integrated with the virtual assistant Alexa to help with daily tasks. Show and Tell is an accessible Alexa feature designed to identify grocery items using the Echo Show's camera. With the product in front of the camera, the user only has to ask "What am I holding?" to activate the functionality. When the product is not identified, Alexa tells the user to turn the item, to find further rich information. Additionally, the beep sound system guides the user in the correct placement of the product. It may also provide a brief description of the product, stored in a database. Since the Show and Tell feature is only available on the Echo Show, its usefulness is restricted to home usage.

OrCam MyEye [19]

In addition to identifying products from barcode recognition, it also identifies products by reading the text label. The system applied in product recognition is the same as in reading books and newspapers; it will not provide information beyond legible words. The user must point to the product/text that he wants to be recognised. Without visual information and guidance on product placement, the possibility of the system not recognizing the words greatly increases.

Wine Searcher [20]

As the name suggests, it is an app for the wine and spirit industry. The aim is to inform the user about the specifications of the wine (the grape variety, the regions it comes from and the producers), the critics' score and, especially, it offers a price comparison tool using a search engine in online mode. The identification of the product is achieved by scanning the label. The system functionality is limited to a category of products enabling a focused recognition, but preventing applicability to other areas. On the other hand, the application was designed for visual people, so there are no additional features to help users obtain a good-quality image.

Amazon GO [21]

A chain of automatic stores that uses thousands of CCTV cameras, computer vision and machine learning methods to analyse if the customer picks up or puts back an item on the shelf and identifies what the item is. Even with a sophisticated system, its results are not entirely reliable, requiring auxiliary systems such as Bluetooth and weight sensors on shelves.

Technological advances have led to the appearance of products that try to respond to the needs of consumers and retailers. Nevertheless, existing limitations highlight the importance of additional research. Table 1 summarises the comparison of the aforementioned technologies.

Table 1. Comparison between existing technological applications.

Applications	Barcode Detection	Label Detection	Guidance	Mode	Portable
RoboCart [13]	√	×	×	RFID	×
Seeing AI [14]	√	×	√	Both	√
OrCam MyEye [19]	√	√	√	Offline	√
Yuka [15]	√	×	×	Both	√
Open Food Facts [16]	√	×	×	Online	√
Lookout [17]	√	√	√	Both	√
Alexa Echo Show [18]	√	√	√	Online	×
Wine Searcher [20]	×	√	√	Online	√
Amazon GO [21]	×	√	×	Online	×

3. Image Pre-Processing and Selection

The input image data affect subsequent steps of an object detection and recognition solution. In particular, when captured by non-experts or visually impaired people, the pre-processing and selection of the images are important steps. Image pre-processing techniques can be divided into two main groups: corrections, e.g., of lighting, noise or colour; enhancements, e.g., illumination, blur or focus.

In [22], the authors analysed the calibration parameters (camera distortion correction and gamma correction) of Micro Aerial Vehicles (MAVs) and Unmanned Aerial Vehicles (UAVs) and image parameters (quantization, compression, resolution, colour model, additional channels). By analysing each parameter individually, they could demonstrate the different impacts each one has on object detection. While some require more memory and others do not perform noticeably differently, camera resolution was discovered to be crucial to this task. Cropping, filtering, rotating and flipping images is frequently performed during pre-processing. Typically, these tasks are completed manually, with a bulk of images undergoing the same transformations. To automatise this process, the researchers of [23] created a deep reinforcement learning framework with an agent in charge of deciding which, and if, an image needs additional transformation, and created an environment that performs these changes. They concluded that this method assisted in automating the pre-processing of different data types, primarily images.

To process fast low-light images, Chen et al. [24] proposed an end-to-end learning approach employing a fully convolutional network. The outcomes demonstrated effective noise suppression and accurate colour transformation. Real-time processing was achieved for low-resolution images but not for full high-resolution ones.

Ledgi et al. [25] proposed a Generative Adversarial Network for image super-resolution (SR) called SRGAN. The authors achieved this by constructing a perceptual loss function for photo-realistic single-image super-resolution using a ResNet architecture and GAN concepts. Compared to other approaches, this solution produced good visual and objective metrics results.

The burst image technique captures several frames that can be combined to produce a higher-resolution image when used with super-resolution algorithms. The BIPNet [26] framework aims to improve and restore burst images. In contrast to other approaches [27,28] that assume that the scene is static and create a solution to remove noise and enhance low-resolution images, this algorithm focuses on motion. The previously aligned burst image features are concatenated using pseudo-burst features fusion. In contrast to [29], which uses a registration-based super-resolution method on each frame to fuse into a high-resolution image, it uses an edge-boosting burst alignment module to prevent mismatching features. The BIPNet framework is effective for tasks requiring high resolution, low-light enhancement, and denoising. This method has a low computational complexity, which is advantageous when using a mobile device.

An approach to denoise burst images captured from a mobile phone was proposed by Godard et al [30]. A fully convolutional deep neural network was used for each frame after it had been stabilised. An improved image was then produced by combining the data from each one using a parallel recurrent network. Despite their promising results, their deep learning strategy was computationally expensive and required considerable memory and processing power.

Lecouat et al. [31] conducted a fast and low-memory-requirement method. They proposed a high dynamic range and image super-resolution reconstruction using raw image bursts as input . This enhanced photos with low-light conditions, noise, camera shake and moderate object motion. This approach improved low-light, noisy, camera-shake and moderately moving object photos.

Another approach is a burst super-resolution transformer (BSRT) [32]. It receives a low-resolution raw capture from a smartphone and tries to correct the noise and misalignment issues with the raw data. To resolve this, the authors implemented a Pyramid Flow-Guided Deformable Convolution Network to help with the performance of the alignment and to decrease the noise. To improve the algorithm's efficiency even more, they added Swin Transformer blocks and groups to enhance the performance of the burst super-resolution task. Results of this algorithm demonstrated that they excelled over earlier approaches, such as HighRes-net [33] and deep burst super-resolution [27].

An adaptive feature consolidation network (AFCNet) [34] for multi-frame super-resolution serves as another illustration of the use of transformers in combination with burst super-resolution. It extracted multi-scale local-global features using an encoder–decoder transformer, which improved feature alignment. A multi-image super-resolution (MISR) that also integrates transformers is TR-MISR [35]; it was applied to the fusion of low-resolution image features. This solution lessened the limitations of multi-image super-resolution transformers.

To maintain high-resolution representations with low-resolution inputs, Zamir et al. [36] proposed a method for real image restoration and enhancement using multi-scale residual blocks. Their approach entails extracting semantically richer and more spatially precise features from three parallel fully convolutional streams, which will exchange and aggregate information using an attention-based mechanism called selective kernel feature fusion. The proposed method outperforms other image denoising, enhancement and super-resolution algorithms, producing images with reduced noise, sharp edges, smooth homogeneous regions, improved colour reproduction and a natural and vivid appearance with appropriate contrast.

Nguyen et al. [37] considered combining three algorithms: Radial Brightening according to [38]; Contrast Limited Adaptive Histogram Equalization (CLAHE) [39]; and Retinex [40] for multi and single-scale to find the best optimal image pre-processing techniques for image enhancement. They ran 15 tests to see how each algorithm and the order in which they were combined affected the enhancement task before applying them to Canny Edge detection. According to reported results, CLAHE-based combinations performed better because they improved detection efficiency, while Retinex reduced sharpness and Brightening showed no discernible change.

EnlightenGAN [41] is a deep-learning-based technique that uses GANs to produce enhanced images from low-light inputs. According to the research results, it performed better than alternative approaches in subjective and objective metrics. Before running an ImageNet-pretrained ResNet-50 classifier, the authors tested the model for the extremely dark [42] dataset due to the increasing interest in image pre-processing for boosting object classification. The findings indicate that the image enhancement task improved classification by 22.02% to 40.92%.

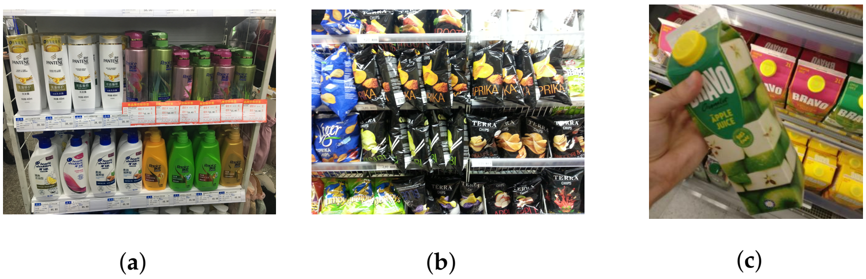
Given that labels contain textual information that may have to be processed, it is also important to know how to make the text on the captured image more readable. Fuelled by the increased interest in safeguarding historical documents, Koshy et al. [43] tested pre-processing methods on digitalised receipts. Their analysis included thresholding, morphology, and blurring methods for pre-processing.

4. Related Datasets

The development and capability of an algorithm and the evaluation of its performance are strongly dependent on the availability of data representative of the target scenario. Datasets of the real-world domain are of greater interest because these data can increase a model's efficiency and robustness. For this paper's target scenario, we considered two main groups of datasets: those intended for the detection and recognition of grocery products and the ones oriented to the detection and recognition of texts in the wild.

4.1. Datasets for Grocery Product Detection and Recognition

Several image datasets have been made available, targeting different aspects or scenarios related to grocery product detection or recognition. Figure 4 presents some examples of possible images. Next, these datasets are briefly described and a comparison is summarised in Table 2.



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

Figure 4. Image examples of a set of publicly available datasets: (**a**) CAPG dataset [44]; (**b**) GroZi 3.2k dataset [45]; (**c**) GroceryStore dataset [46].

Products-6K [8] provides images from Greek supermarkets. The dataset contains 12,917 studio-quality images associated with 6348 stock-keeping units (SKUs). The reference images were captured in a clean environment with white background, while 373 query images of 104 classes were captured under real supermarket conditions, taken by a mobile phone camera. Products can be displayed on the shelf or in hand, with photos taken from different sides of the product. A fine-grained description and some textual information are provided, but texts were not labelled by human annotators and text location information is missing. The annotations, whether class or textual, are in Greek as well as the description found on the packages themselves.

GroZi-120 [47] consists of 120 Swiss market product categories that are very different from each other. The training set includes 676 images selected from the web containing products captured under ideal conditions. Training images contain just one single product instance viewed from different front-side perspectives. The test set has 4973 frames of 29 real videos from grocery racks to represent the challenging natural environment. The product images that appeared in the frames were cropped to create a query image set. As these images are of products on the shelf, only the front view was captured and the low resolution of the images adds another degree of difficulty. Ground-truth includes the image class and bounding box.

GroZi-3.2k [45] contains image of Swiss products on supermarket racks. Inside the food category, there are 27 coarse classes and 3235 training images. Training images collected from the web are taken in a controlled environment with a white background. Query set images are composed of 680 high-quality images captured from 5 real-life retail stores using a mobile phone. Ground-truth includes relative coordinates of the bounding boxes and product labels, such as coffee, cereals and milk.

Grocery Store [46] includes images of fruits, vegetables and carton items captured with a smartphone camera in different Swedish grocery stores. The dataset also includes a studio-quality image per item. For food label reading purposes, 1745 images of carton items (e.g. juice, milk, yoghurt) can be used. Most of the products are held by a hand and a few of them are displayed on shelves. A total of 31 fine-grained classes are annotated, which includes the super category and the brand with the flavour/type. Other files are accessible, one with a brief description in English and the other with more detailed information in Swedish, such as ingredients, nutrition values, manufacturer, volume and manufacturing country. It should be noted that the dataset is designed for the classification and recognition of objects and not for their detection.

Freiburg Groceries Dataset [48] consists of 4947 images of 25 coarse grocery classes. The training images were taken predominantly in stores, but some in apartments and offices in Germany, using four different camera phones. In addition to having photos from several domains, the products are seen from various perspectives and light conditions. To have a variable degree of clutter, images contain one or several instances of one of 25 classes. As annotated classes, we have cereal, pasta and cake, among others. Data are labelled only at the image level.

D2S dataset [49] is a dataset designed for automatic checkout machines or inventory systems. It contains images of German products (such as fruits, vegetables, cereal packets, pasta and bottles) from 60 categories placed on a table to simulate the real environment. The training set includes 4380 images that involve one or more products of a single class with a clean background. There are 16,620 test images with a more complex background, showing single or multiple objects of different classes. The overlap of some objects causes partial occlusion and each scene is acquired with three different lighting conditions. The annotations include label pixels for semantic segmentation methods and bounding box coordinates for bounding box detection. The supercategory and subcategory are also annotated.

Unitail-Det [50] aims to support weel-aligned grocery product detection on shelves. A total of 11,744 quality images of shelves in supermarkets were captured, achieving 1,740,037 quadrilateral annotations. Quadrilateral bounding boxes are more accurate in covering objects. Additionally, another set of 500 images of 3024 × 4032 resolution is provided with another domain. The number of quadrilateral instances annotated is 37,071. There is a large diversity of products, including the most common grocery, delicatessen, textiles, electronics, and household products and medicines, among others.

Unitail-OCR [50] was created to sustain retail product recognition through product matching via robust reading. In the gallery of products to be recognised, there are 1454 fine-grained products with frontal photos. Among these products, there are 10,709 labelled text regions located and 7565 legible word transcriptions. For testing purposes, 10k images of products were cropped, as well as the text on the packaging. In total, 18,972 text regions were detected, but only 13,416 are legible and have been transcribed for text recognition. A vocabulary list is provided with all words presented. The characteristics of this dataset meet the existing needs in terms of detection and text recognition of supermarket products.

RPC dataset [5] consists of 83,739 images of Chinese products; 53,739 single-product exemplary images for training; 30,000 checkout images for validation and testing. Each training image is captured in controlled conditions with four cameras from different views. Designed for automatic checkout applications, the test images contain several products placed on a table. The images are grouped into three degrees of clutter, according to the number of items presented. A hierarchical structure annotation of 200 fine-grained product categories can be coarsely categorised as 17 meta-classes. Essentially designed for the automatic checkout scenario, the dataset can be used for detection and counting tasks.

CAPG Grocery Product Dataset [44] contains 102 grocery products of 3 types of products belonging to 5 different brands: box-like packaged products (69 classes); bag-packaged products (15 classes); tube-packaged products (18 classes). For each class, a photo is taken under controlled conditions as the training set. Within each class, there can be sub-classes representing different sizes or flavours of that product; therefore, the variance of the visual features among sub-classes is small. The dataset is adequate for fine-grained grocery product recognition, being present in 177 sub-classes. As the testing set, 234 images of products on racks from 2 Chinese stores were collected.

RP2K [51] is a large dataset that contains 2395 unique SKUs of the Chinese market that can be grouped into 7 coarse categories. In that sense, products of different sizes, flavours/types are classified as another class. Like the previous dataset, it was created for fine-grained image classification, but with substantial annotations that include class, category, brand, flavour/type, size and shape. Since the real environment is also a concern, 384,311 cropped images of retail products on the shelves were extracted in physical supermarkets.

Store shelf images and product images for retail [52] is a public dataset available on the Kaggle website. It represents 100 categories of retail products. For each one, 3 photos with different perspectives were taken with a product placed on the supermarket floor, which makes a total of 300 photos. Another 3153 images were captured from supermarket shelves, taking into account the lighting conditions and visibility to provide legible photos. No data annotations were made available.

Open Food Facts dataset [16] is the base of the mobile application with the same name. The users can add photos and information about food products. Therefore, it is a dataset which is continually being updated. To date, there are approximately two and a half million photos captured under real or ideal conditions. There are products from almost every corner of the globe, but they are predominantly from France, the United States and Spain. For each item, it tries to provide as much information as possible, which includes brand, category, quantity, packaging, ingredients, origins of ingredients, allergens, traces, additives, added vitamins, minerals, amino acids and others, as well as the Nutrition Score, Nova classification that categorises the degree of processing of foods, and Eco Score.

SKU-110K [3] is a dataset designed to support product detection in a densely packed setting. The images contain thick retail shelves with identical items. The location of items is labelled by bounding boxes. Wanting to represent a variety of situations, the images contain thousands of supermarkets, which are located in the United States, Europe and East Asia. The dataset is divided into training, validation and test set, with 8232, 587 and 2940 images, respectively. As the photographers are not subject to any viewing settings, the images present many scales, viewing angles, lighting conditions and noise levels.

WebMarket [53] dataset is available on Kaggle and contains 300 images of retail racks. It was designed for object detection; therefore, only the products' location is labelled.

Outside the supermarket context, there are also logo-oriented databases that can help train the model to be more robust in the presence of small variations of those logos. Most of these datasets are divided into categories, food brands being the obvious option. Some that might be interesting are Logo-2K+ [54], LogoDet-3K [10], FoodLogoDet-1500 [55] and WebLogo-2M dataset [56].

4.2. Scene Text Recognition Datasets

Less attention has been given to the processing of text in grocery or retail labels. Hence, there is a lack of information and specific datasets about product labels, forcing the text reader modules to be trained on scene text datasets, as shown in Figure 5. The benchmark datasets to support scene text recognition (STR) can be divided into three groups: synthetics, regular, and irregular datasets.

Synthetic datasets were created because labelling and its verification are very time-consuming. Artificial distortions of the image are often applied to obtain a higher quantity of realistic images and, automatically, the annotation of object localization follows the same distortion. Next, the most relevant synthetic datasets are described.

MJsynth [57] is a large-scale dataset for horizontal text recognition that contains 8.9 M English word box images. The process of generating synthetic data is the following: (1) font rendering; (2) border and shadow rendering; (3) base colouring; (4) projective distortions; (5) natural data blending; (6) noise. The generated word samples have a fixed height of 32 pixels and a variable width according to the length of the word.

SynthText [58] contains 800,000 generated scene images that contain several horizontal text instances displayed on multi-perspectives. The text samples used to create these images were extracted from an English text dataset that includes nouns, numbers, symbols and punctuation marks. Initially, data were designed for scene text detection by labelled text region location, but with labelled extracted text regions, SynthText has also been used for scene text recognition. Label information is rich, annotated at the text-strings-, word- and character-level bounding boxes. The data have more than 7 M cropped word boxes and almost 30 M characters.

Table 2. Comparison between grocery and logo datasets. The letter “R” means reference images, and “Q” queries images. The indicated resolution is estimated.

Datasets	#Images	#Classes	Objective	Labels	Language	Resolution
Products-6K [8]	12,917 R 373 Q	6348 R 104 Q	Classify products on the shelf or in a hand by matching system.	SKU classes, image class Textual information	Greek	800 × 800 R 3024 × 4032 Q
GroZi-120 [47]	676 R 4973 Q	120	Detected and classify products on the shelf by matching system.	Image class, Bounding box	Swiss	183 × 162
GroZi-3.2k [45]	3235 R 680 Q	27	Detected and classify products on the shelf by matching system.	Object superclass, Bounding box	Swiss	421 × 500 R 3264 × 2448 Q
Grocery Store [46]	1745	31	Classify subcategory products essentially in hand and get detailed information of the product.	Image sub-class, small description and detailed information	Swedish	348 × 348
Freiburg [48]	4947	25	Classify super-category product essentially on the shelf. Photos contain one category.	Image class	German	256 × 256
D2S-Densely Segmented [49]	4380 R 19,620 Q	60	Designed for automatic checkout or inventory system. Detect and classify products on a table.	Instance segmentation Bounding box Object super-class Object subclass	German	1920 × 1440
Unitail-Det [50]	11,744 + 500	1,740,037 + 37,071	Detect all products displayed in a supermarket.	Quadrilateral bounding boxes	English	1216 × 1600 3024 × 4032
Unitail-OCR [50]	1454 R 10,000 Q	1454	Classify cropped image products by three steps: text detection, text recognition and product matching.	Image class Quadrilateral text location. Text transcription	English	194 × 504
RPC [5]	53,739 R 30,000 Q	200	Designed for automatic checkout or inventory system. Detect and classify products on a table.	Bounding box SKU class Hierarquical classes	Chinese	1817 × 1817
CAPG Grocery [44]	18 R 236 Q	18	Detected and classify products on the shelf by matching system.	Bounding box SKU class	Chinese	261 × 600 R 4032 × 3024 Q
RP2K [51]	384,311	2395	Classify cropped image products capture on the shelf.	SKU class	Chinese	153 × 251
Store images for Retail [52]	300 R 3153 Q	100	Grocery product classification	No labels	English [a]	2272 × 1704 R 757 × 568 Q
Open Food Facts dataset [16]	+2.5 M	+2.5 M	Image classification and extensive product information	SKU class, Additional information such as common name, allergens, Nutri-score, Nova score	Multi-lingual	width × 400
SKU-110K [3]	11,762	110,712	Detect all objects on shelves	Bounding box	Multi-lingual	-
WebMarket [53]	300	-	Detect all objects on shelves	Bounding box	English	2272 × 1704

[a] The language is not mentioned. The observed images contain English products/logos.

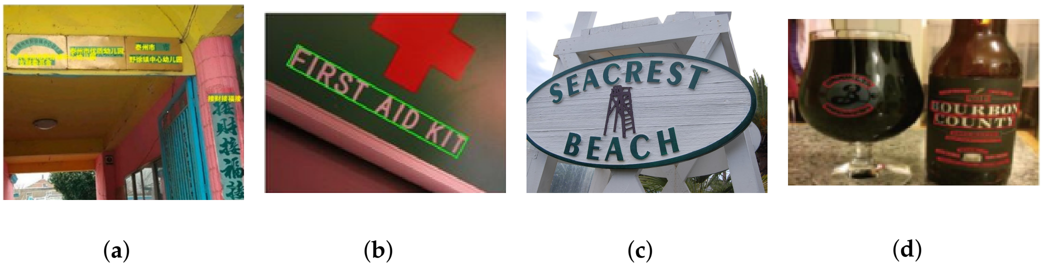


Figure 5. Image examples of publicly available datasets for scene text recognition. (**a**) ICDAR19-LSVT [59]. (**b**) MSRA-TD500 [60]. (**c**) CUTE80 [61]. (**d**) SCUT-CTW1500 [62].

Regular STR datasets collect several text images of the real world that are relatively easy to detect and recognise. Some images are noisy, with variations in illumination and scale or low resolution, but all text instances have horizontal orientation and are easily separated.

IIIT5K-Word [63] contains 5000 cropped words extracted from images crawled from Google Image Search. Some examples of query words used in the search engine are “billboards”, “signboard”, “house numbers”, “house name plates” and “movie posters”. The images are split into 2000 words for training and 3000 for testing. For each photo, two lexicon lists are provided: a 50-word lexicon and a 1000-word lexicon. IIIT5K-Word is labelled at the word and character level.

Street View Text [64] is a dataset that contains 350 images downloaded from Google Street View, searched by English business names. Hence, the most frequent words belong to business signage. The dataset was designed for text detection and word recognition, but all words in the photo are not referenced in the ground-truth label. A lexicon list is given for each image and the aim is to find words entered in the lexicon, which may correspond to the scenario where a blind person searches for supermarket products with a grocery list which is the lexicon. In selecting images, the focus was on obtaining frontal texts by minimizing the skew angles. Some of the images are noisy, blurry and of low-resolution. The median height of the images is 55 pixels, but image quality varies greatly. From 350 images, 101 are for training and 249 for testing. There are 725 words labelled: 211 for the training set and 514 for the test set.

ICDAR2003 [65] was created for the ICDAR 2003 Robust Reading competitions to support the reading of texts of natural scenes. It contains 509 images, 258 for training and 251 for testing. Bounding boxes surround text instances, and ground-truth text location is given. For the word recognition task, there are 1156 word images in the training set and 1110 images for evaluation. Excluding words less than three characters and ones that contain non-alphanumeric characters, the result is 867 text regions. Character recognition is also possible by ground-truth character labels; there are 6185 characters for training and 5430 for testing.

ICDAR2013 [66] had three challenges. For reading text in Born-Digital Images (Web and E-mail), 561 images were collected with a minimum size of 100 × 100 pixels, with 420 used for training and 141 for evaluation. The dataset is labelled for text location, text segmentation and word recognition. For the last task, words with fewer than three characters were excluded, resulting in 3564 cropped words in the training set and 1439 in the test set.

Char74k [67] collects 1922 images of sign boards, hoardings and advertisements taken in the streets of India. Some photos were taken of products in supermarkets and shops, with English and Kannada texts present in this dataset. Focusing on the English dataset, 12,503 characters were annotated and 4798 were labelled as bad images due to excessive occlusion, low resolution or noise. In character classification, they distinguish upper and lower cases, which makes a total of 62 classes with the inclusion of digits. They also created a complementary English dataset with 3410 hand-printed characters generated by

55 volunteers. A third English dataset was synthetically generated by 254 different fonts in 4 styles (normal, bold, italic and bold+italic), reaching 62,922 characters.

As labelled grocery products have texts of countless positions, shapes and densities, the aforementioned datasets are unsuitable for the scope of this paper. On the other hand, irregular datasets cover complicated text scenes, such as curved texts, with different perspectives, orientations and positions.

ICDAR 2015 [68] is a dataset addressed to Challenge 4 of the ICDAR2015 Robust Reading Competition. About 1500 images were acquired with Google Glasses without paying attention to the position, perspective or quality. As a result, the dataset includes blurry, noisy and low-quality images. As in real scenarios, texts may have different styles, arbitrary shapes and orientations, and even be curved and illegible. The dataset was designed for text localization and word recognition. A total of 4468 training instances were labelled with a quadrilateral bounding box, and its transcription is provided when the word is readable. Otherwise, the illegible word is annotated as “###”. For testing the model, the dataset contains 2077 text images..

MSRA-TD500 [60] consists of 500 images captured by a pocket camera in indoor and outdoor scenes. Each image contains Chinese, English or a combination of both languages. Texts have different sizes, fonts, directions and colours. The primary application of this dataset is multidirectional text detection, where text lines are confined by rotatable rectangles.

SVT Perspective [69] is a dataset based on the StreetViewText dataset, which consists of collecting photos from Google Street View, using English business names as a means of searching. SVT Perspective contains images of the same addresses, but the selection focus on non-frontal texts, choosing texts viewed from different perspectives and of arbitrary orientations, summing to 238 images. They preserved the same lexicon of SVT and labelled only words presented in the specific lexicon for that image. In total, 639 words were annotated using quadrilaterals. The heights of cropped words vary from 9 to 330 pixels.

CUTE80 [61] consists of 80 curved text images for natural scene image recognition. Therefore, images suffer complex backgrounds and variations in text scale, font, orientations and perspective. A set of polygon points for each curved text line labels 288 word instances.

COCO-Text dataset [70] is based on the MS COCO dataset, which contains images of complex everyday scenes. The COCO-Text dataset contains non-text images and legible and illegible text images. It is the first large-scale dataset for text detection and recognition in natural scenes, including 63,686 images with 239,506 labelled text regions in the new version. For every word, the segmentation mask is annotated. Moreover, each word is categorised into three attributes: machine-printed vs. handwritten, legible vs. illegible, and English vs. non-English.

Total-Text [71] dataset was intended to make a more significant contribution to the availability of curve-orientated texts, providing 1555 images that include horizontal, multi-oriented and curved texts. The dataset provides labels at word level by polygon bounding box. Ful annotations are provided for text detection, recognition and segmentation of 11,459 English text instances.

SCUT-CTW1500 [62] is a multi-lingual dataset that contains English and Chinese texts. At least one curved text appears in 1500 images from different web sources and texts have arbitrary shapes and orientations. In the last version of annotations, Chinese texts are labelled with “###” because their weight is not significant. Polygon bounding box coordinates are provided at the setence level.

VinText [72] is a dataset of Vietnamese scene images. The 2000 images contain text instances from several perspectives, sizes, orientations and shapes, providing about 56 k quadrilateral bounding boxes and word-level transcriptions. According to the authors, Vietnamese script is similar to Portuguese script, and the transfer learning technique might be interesting to Portuguese scenarios.

ICDAR 2017-MLT [73] is a multi-lingual text dataset with 7200 training images, 1800 validation images and 9000 test images. The photos were extracted from various sources and represent a high diversity of domains, such as street views and pictures in microblogs. Variety is also presented in text format, including horizontal, multi-perspective and vertical texts. Chinese, Japanese, Korean, English, French, Arabic, Italian, German and Indian languages are equally represented in this dataset. For each language, there are at least 2000 images; however, each image may contain more than one language. Ground-truth quadrilateral bounding box is provided at the word level. Text recognition and script identification are possible because each recognizable word is associated with a script class and the transcription. There are 84,868 training word images and 97,619 test instances for these tasks.

ICDAR 2019—RRC-MLT [73] is also a multi-lingual dataset that consists of 20,000 images of 10 different languages, where for each language 2000 photos were selected. In addition to the nine languages mentioned, Bengali is now presented. The training set contains 10,000 images with bounding box coordinates and another 10,000 images compose the test set. For the text recognition task, 89,177 cropped text instances are provided as training images and 102,462 as test images. ICDAR 2019 is not only distinguished from ICDAR 2017 by having added another language to the list but also by being more challenging. The weight of curved and vertical images is higher than in ICDAR 2019. They continue to have texts of different sizes, aspect ratios and directions.

ICDAR 2019—Art [74] joins images of Total-Text, SCUT-CTW1500, curved texts of LSVT and 7111 newly collected images with high arbitrary shape, multi-oriented and curved text instances. The aim is to support models to be robust against diverse text formats and language variants, since there are Chinese and Latin texts. There are 10,166 images divided into 5603 training images and 4563 for evaluation. Text instances are annotated at the word level, enclosed by quadrilateral or polygon bounding boxes. The dataset is labelled for text detection, recognition and spotting tasks.

ICDAR2019—LSVT [59] is a large dataset containing photos captured from different mobile phones in the streets of China. There are 50,000 fully annotated images and 400,000 training images with weak annotations. Photos are considered fully annotated when text instances are transcribed and the corresponding ground-truth location is also provided. On the other hand, they are weakly annotated when they only have the transcription of the text of interest in these images.

The properties of the aforementioned scene text datasets are summarised in Table 3.

Table 3. Comparison between Scene Text Datasets.

Datasets	#Training / Test Images	#Text Instances	Annotation	Orientation	Curved	Language
MJSynth [57]	∼8.9 M	∼8.9 M	Word	Multi-oriented	Not curved	English
SynthText [58]	∼800 k	∼8 M	Text-string/ Word/ Character	Multi-oriented	Not curved	English
III5K-Words [63]	380/740	2000/3000	Word/ Character	Horizontal	Not curved	English
Street View Text [64]	101/249	211/514	Word	Horizontal	Not curved	English
ICDAR2003 [65]	258/251	1156/1110	Word/ Character	Horizontal	Not curved	English
ICDAR2013 [66]	420/141	3564/1439	Word/ Character	Horizontal	Not curved	English
Char74k [67]	-	∼78,000	Character	Horizontal	Few curved	Latin
ICDAR2015 [68]	1000/500	4468/2077	Word	Multi-oriented	Few curved	English
MSRA-TD500 [60]	300/200	-	Word	Multi-oriented	Few curved	Chinese and English
SVT Perspective [69]	238	639	Word	Multi-oriented	Few curved	English
VinText [72]	2000	∼56,000	Word	Multi-oriented	Few curved	Vietnamese
CUTE80 [61]	80	288	Word	Multi-oriented	Curved	English
COCO-Text [70]	43,686/20,000	118,309/27,550	Word	Multi-oriented	Curved	English
Total-Text [71]	1555/300	111,666/293	Word	Multi-oriented	Curved	English
SCUT-CTW1500 [62]	1000/500	7683/3068	Word	Multi-oriented	Curved	Chinese and English
ICDAR2017—MLT [73]	9000/9000	84,868/97,619	Word	Multi-oriented	Curved	9 Languages
ICDAR2019—MLT [73]	10,000/10,000	89,177/102,462	Word	Multi-oriented	Curved	10 Languages
ICDAR2019—Art [74]	6603/4563	50,029/48,426	Word	Multi-oriented	Curved	Chinese and English
ICDAR2019—LSVT [59]	430,000/20,000	-	Word	Multi-oriented	Curved	Chinese and English

5. Product Detection and Recognition

Humans can easily locate and identify objects of interest (such as people, animals, buildings or cars) in an image or video. Object detection and recognition are computer vision tasks that intend to reproduce this ability. The interest in this area dates back more than three decades and the progress of objection detection and recognition can be divided into two periods based on the main strategies: traditional handcrafted-feature-based methods and deep-learning-based methods. Manual feature extraction suffers from a lack of robustness due to the diversity of aspects of the class and strongly depends on the experience of the researcher. The problem becomes more challenging when dealing with a large number of different classes. Convolutional Neural Networks (CNN), a type of deep learning algorithm, outperformed traditional techniques by learning robust and high-level feature representations of an image.

Object classification, particularly in the paper's target scenario, can be integrated into a hierarchical system. Coarse classes include objects of macro-categories; on the other hand, fine-grained classes refer to objects of sub-categories. In the retail context, Coca-Cola can be classified as a drink, as a soft drink, as its brand (Coca-Cola) or with more details, e.g., as Coca-Cola Zero and Coca-Cola Zero 330 mL. The class level provided by a model is imposed by the annotation levels of the dataset. Moreover, the model's architecture is not always able to correctly identify products with great detail. The first studies were mostly designed to predict coarse classes, while more recent papers were focused on identifying products according to a fine-grained classification. This section reviews relevant retail/grocery product-recognition proposals.

Michele Merler et al. [47] tested three traditional feature extraction methods (histogram of colour, SIFT [75] and Haar-like features) for product detection and recognition, and created the GroZi-120 dataset for training and test purposes. The small number of classes and great shape variability among classes can help the model performance. A more extensive dataset, the GroZi-3.2k, was designed in [45]. The same authors proposed and evaluated a model that extracts SIFT feature descriptors for each grid of the image and compared them with reference images. A voting algorithm provides a ranking classification based on votes over grid patches of several sizes. This approach reduced the noise of compared cross-domain images. The localization of products in shelves is performed with deformable dense pixel matching, and the final classification is through a genetic algorithm optimization that uses the top N ranking. In [76], SIFT features were incorporated in a hybrid context-aware model to detect and classify fine-grained products displayed on the shelves. The approach combined a context-free visual classifier with a graphical model. An SVM was used as a classifier and Hidden Markov Model and Conditional Random Fields as a graphical model. The underlying idea is that the arrangement of products follows a plan, i.e., similar products are usually together. A different approach was proposed by Yörük et al. [77] that intended to identify the product and estimate the pose. The method compares SURF features from a query to a database of model features. Then, with fewer computational steps, a refined Hough transform simultaneously detects, recognises and estimates the pose of grocery products.

Santra et al. [78] proposed an annotation-free machine vision system that locates products on the shelves. Based on features extracted by the BRISK descriptor, their proposal generated region proposals with a two-stage exemplar-driven region proposal algorithm, relying on a single example or image templates. Subsequently, a CNN classified each region proposal. Finally, the overlapping and ambiguous region proposals were removed by a greedy non-maximal suppression strategy. The context information of retail stores used to estimate the scales sometimes produced inferior results than the physical dimension of the product template. This proposal had a high computational weight and the authors assumed that the images were taken with the camera almost parallel to the rack's face. Franco et al. [2] compared the bag of visual words [79] representation with Convolution Neural Network features (in particular AlexNet [80]) using the GroZi-120 dataset. The results showed that deep-neural-network-based features are usually more effective in more complex scenarios.

The results, through the application of deep learning techniques in several areas, along with the increasing accessibility of large datasets and efficient computational resources, are leading to greater research based on these methods. Some researchers contributed to detecting products on shelves, others tried to improve the recognition results of grocery images, but end-to-end pipelines have also been studied. Given the state-of-the-art results obtained through the use of deep learning techniques, the following sub-section focuses on an overview of deep-learning-based product detection and recognition methods.

5.1. Product Detection Based on Deep Learning

Grocery product detection is intended to locate items in a grocery store without categorisation. Detecting a few large objects in an image is easier than detecting grocery products in dense racks since there are many small objects. ScaleNet [81] used the ResNet feature extractor and reduced the searching space of scales by estimating the object scale and using it to guide object proposal generation in supermarket images. The object proposal detection is based on the SharpMask method, which requires annotation at pixel level. Goldman et al. [3] presented an alternative strategy aiming to overcome the challenge posed by common multiple overlapping bounding boxes in densely packed scenes. The CNN detector estimates the bounding box, the objectness and the Soft-IoU score as a measure of detection quality. The Expectation-Maximization (EM) unit uses the previous score and detections to select the correct location. The model was evaluated in the SKU-110K dataset and showed that even the best results were still significantly saturated. Santra et al. [82] introduced an R-CNN detector, replacing the greedy non-maximal suppression with a novel graph-based non-maximal suppression that obtained the best proposal region by combining classification scores and the product classes of the overlapping region proposals. In [50], the authors proposed a retail product detection based on DenseBox-style [83] with a Feature Pyramid Network (FPN). The authors re-defined the centre of quadrilateral bounding boxes computing Quad-Centerness (QC) and introduced the Soft-Scale (SS) algorithm to detect objects of arbitrary shape. The corner refinement module was added to improve heatmap prediction and the location coordinates. The Quad-Centerness (QC), Soft-Scale (SS) and corner refinement module (CRM) gradually improved the mean average precision (mAP). To reduce the need to collect a large amount of data, a one-stage one-shot detector—OS2D—was proposed in [84]. It joined the localization and recognition phases, applying a fully convolutional TransformNet to extract an image's features. A dense correlation and a feed-forward geometric transformation model were used to match and align features. The bilinear resample served to finalise the training. The assessment was performed using the GroZi-3.2k dataset [45] and the authors concluded that the OS2D model outperformed some reliable baselines.

A fair comparison between these detection models is hard, since the methods were assessed using different test datasets and distinct metrics. Table 4 demonstrates the specific approach implemented in each method.

Table 4. Comparison of grocery product detection methods. In the RetailDet, the evaluation was made in two test sets, the origin-domain, which is represented as OD, and cross-domain as CD.

Methods	Annotations	Dataset	Scores	Metrics
ScaleNet [81]	Mask based/segmentation	MS COCO	0.578	AR@1k
Goldman et al. [3]	Bounding box	SKU-110K	0.492	AP
Santra et al. [82]	Bounding box	GroZi-3.2 WebMarket GroZi-120	0.802 0.755 0.448	F1
RetailDet [50]	Bounding box	SKU110k UnitailDet	0.590 OD:0.587, CD:0.509	mAP
OS2D [84]	Bounding box	GroZi-3.2k	0.850	mAP

5.2. Product Recognition Based on Deep Learning

Object recognition enables classifying products already detected or images with a single product. In [85], an updated version of [86] was presented with the contextual and visual features computed by CNN models, such as ResNet50, Alexnet or VGG. The deep class embedding, strengthened by the product's visual appearance and its relative position on the shelf, is reparametrised by a CRF-based method. The proposed model was designed for the fine-grained classification of numerous classes. A framework to recognise fine-grained product images of different source and target domains was proposed by [87] with the CNN model containing two modules: an adversarial module to handle the cross-domain scenario and a self-attention module that captures the most discriminative image regions to increase accuracy. The results were compared with domain adaption methods, excluding known grocery recognition models. Santra et al. [88] suggested a fine-grained classifier that combined object-level and part-level information of the product images. The novel reconstruction-classification network (RC-Net) extracted a more general product feature, being robust to a range of store lighting levels. The discriminative features were searched in an unsupervised manner and encoded using convolutional Long Short-Term Memory (LSTM). The R-CNN classified the object based on both feature-level.

In [46], the authors presented preliminary benchmark results for the Grocery Store dataset (GSD). The experiments showed that using the joint DenseNet-169 features of both iconic and natural images as inputs to the variational autoencoder canonical correlation analysis (VAE-CCA) [89] was better than only using the extracted features of natural images. For the classification, the SVM was trained on the latent representations. Ciocca et al. [90] used the previous model as the baseline method. They suggest supervised and unsupervised frameworks based on the three hierarchy class labels of GSD. In both, a multi-task learning DenseNet-169 network was used to extract features for the corresponding class level. The supervised model was trained over ImageNet and fine-tuned on the GSD dataset with data augmentation. The best result (83%) was obtained using a cubic SVM on the features extracted from the last average pooling layer before the network split. For the unsupervised method, they linked the same features of the supervised model to the Affinity Propagation algorithm [91]. Both frameworks have difficulty adapting to other datasets. To overcome the scarcity of image data, Domingo et al. [92] studied the implementation of Siamese neural networks in grocery classification, assessing four one-shot learning architectures. The datasets contain few images per category or even one iconic image per class. The integration of ResNeXt-101 with LOMO descriptor into the Siamese network achieved good results, obtaining an F1-score of 89.1%. Low computational cost can be achieved with lightweight descriptors.

In [93], Wang et al. proposed a self-attention mechanism to eliminate the noise when performing the destruction and construction of image knowledge to grocery classification in the Retail Product Checkout dataset [5]. Other fine-grained classification methods fall short compared to self-attention-based destruction and construction learning (SADCL). This application achieved an accuracy above 80% with low computational costs. As shown in Table 5, each approach uses a different dataset and metric to analyse the results, thus making a proper comparison difficult.

5.3. End-to-End Product Classification Based on Deep Learning

Object detection and object recognition can be combined to create effective end-to-end product identification. To overcome the obstacles placed by the scarcity of relevant data, Leonid et al. [94] proposed a classification model with just one training example per class. The initial detection and classification were performed based on a non-parametric probabilistic model trained on the limited data. In the second phase, a CNN-based model used the previous information to produce fine-grained classification refinements. Synthetic data were created from a few training examples to train the CNN model and the final output combines the scores of the first model with the corresponding CNN confidences. A coarse-to-fine approach was proposed in [44] with product regions detected by recurring features and classified into coarse classes. For a fine-grained classification, they employed attention maps built by SIFT features to guide the CNN classifier to focus on fine discriminative details. By using one-shot learning, the identification is based on feature matching. The system could adjust to new product classes without having to retrain the classifier.

Table 5. Comparison of grocery product recognition methods

Methods	System	Classification	Dataset	Scores	Metrics
Goldman and Goldberger [85]	Sequence of products on the rack	Fine-grained/ ultrafine-grained	Private dataset	0.129	Mean Error
Wang, Y. et al. [87]	Matching	Fine-grained	Private dataset	0.422	Avg Accuracy
Santra et al. [88]	Matching	Fine-grained	Grocery Products WebMarket GroZi-120	0.797 0.740 0.440	F1
Klasson et al. [46]	Matching	Hierarchical	Grocery Store	0.804	Accuracy
Ciocca et al. [90] Supervised	Matching	Hierarchical	Grocery Products	0.904	Accuracy
Ciocca et al. [90] Unsupervised	Matching	Hierarchical	Grocery Products	0.929	Accuracy
Wang, W et al. [93]—VGG-16	Matching	Fine-grained	RPC	0.787	Accuracy
Wang, W et al. [93]—ResNet-50	Matching	Fine-grained	RPC	0.814	Accuracy
Domingo et al. [92]	Matching (Siamese)	Fine-grained	Grocery Store	0.891	F1

In [95], the global feature description of the detected region was compared to a database of descriptions of iconic product images. YOLOv5 was used to extract the proposal regions and naive Bayes (NB) similarity search provided the product classification. The recognition results were improved with a combination of shape, size and colour features encoded by Fisher vectors and the Dirichlet function. The mAP of the proposed model achieved 58% on GroZi-120 dataset. Using a one-shot learning approach, [4] located objects on the shelves with a CNN-based detector, such as YOLOv2 [96]. The global product descriptors of reference and query images are obtained by computing MAC (maximum activations of convolutions) features from the CNN *embedder* pre-trained on ImageNet to produce embedding vectors. The recognition was obtained by pairwise means K-NN similarity search. As the training and query image domain are different, later in [97], the same authors incorporated GAN with adversarial behaviour in the *embedder* to deal with domain shift, to augment the training set size and for the CNN to learn a stronger embedding function. Ankit Sinha et al. [98] refined the framework suggested in [4]. The YOLO-based detector was replaced by a Faster RCNN integrated with a Feature Pyramid Network (FPN) to enhance multi-scale object detection. Intending to deploy the model on an edge device with little memory space, they used a lightweight ResNet-18-based product recognition model instead of the more complex VGG-16 model. This framework was trained over a large set of images created by augmentation techniques. They were able to introduce a time-efficient framework. Table 6 summarizes the quantitative performance of grocery detection and recognition methods on the GroZi-3.2K dataset. Since the model proposed in [95] was just evaluated on the SKU-110K dataset, their results are shown to provide a guideline.

Table 6. Evaluation and comparison of grocery detection and recognition methods on GroZi-3.2K.mAP.

Methods	System	GroZi-3.2K		SKU-110K	
		AP	AR	AP	AR
Leonid et al. [94]	Region matching	0.447	-	-	-
Geng et al. [44]	Matching	0.739	-	-	-
Tonioni et al. [4]	Matching	0.735	0.827	-	-
Tonioni et al. [97]	Matching	0.842	-	-	
Sinha et al. [98]	Multi-scale CNN-based	0.478	0.581	-	-
Gothai et al. [95]	Matching	-	-	0.580	0.710

6. Product Label Analysis

Various techniques can be used to identify grocery products. There are systems based on barcode readers, visual appearance and/or text labels. The label provides information that can be paramount to classify the product or to afford detailed product characteristics. Consequently, some researchers have already started to use textual information in their product classification models. In [99], employing optical character recognition (OCR) techniques, a histogram of words for each class is built from training images, which provides the frequency of words in a given class. This word histogram measures the confidence of the product's name introduced by the user in a corresponding class and is used to rank the possible classes. However, the extracted text information is only used to identify the coarse class of each product of a shopping list. The fine-grained product classification is based on clusters of mid-level discriminative patches. Different model versions were tested in edge devices, although the accuracy was always less than 62%.

Previous methods applied visual and text features independently. Nevertheless, product identification, especially those with similar appearance, can benefit from textual information. The experiments conducted by Marcus Klasson et al. [9] support the idea that combining visual features and text descriptions improves the classification model's performance, especially in distinguishing between visually similar items. If for any constraint only one type of feature can be employed, models using visual features achieved better results. They also concluded that introducing features of the referenced images is essential to achieve higher classification performance since visual information is clearer. For image grocery product classification, they used a multi-view generative model, Variational Canonical Correlation Analysis (VCCA) learns a shared latent space for four views of natural and referenced images and product descriptions. Referenced images help to separate the items based on their colour and shape, whereas text information joins the groceries based on their ingredients and flavour. Despite the good results achieved, its implementation on a mobile device has not been tested. A different approach was presented in [8], proposing the fusion of visual and textual descriptors. Google Cloud Vision's OCR mechanism extracted textual information of reference and query images. The former information was concatenated with the product description provided in the dataset. According to a distance metric between the two sets of words, all index images are ranked in descending order, and then the list is re-ranking by the similarity between visual information. The use of inverse order of descriptors is also suggested. The limitations of the model were related to the differences between referenced and query images. Usually, the referenced images only contain the front of the package. If the shopper takes a photo of another side, the system has difficulty identifying the item correctly. There is also the possibility of having a database of side views of the product and the system assigns a class that belongs to the same brand or category, but not the correct SKU because the minor details are frequently on the frontal view. Even when the natural photo captures the frontal view, the layout may differ from the referenced image because of marketing strategies. The system proposed in [9] can handle these obstacles better because the latent representations are extracted from multi-view encoders. Nevertheless, when the model is implemented with only frontal view packages, the model presented by Georgiadis et al. [8] has better performance. In [50], the authors resort to a similar strategy. The Unitail framework was initially used to calculate the cosine similarity between the query and gallery images based on visual features. Then, text similarity scores were computed to determine the best class between the highest and second-highest rank list. Consequently, the model compared the encoded features of both images and used the Hungarian algorithm [100] to compute the similarity between variable text sequences. The product matching system was not trained end-to-end, so errors were accumulated and propagated by the text detector and recogniser. To mitigate propagated errors, the model exploits intermediate feature vectors from the text recogniser.

The following frameworks only use text information to identify grocery products. The system becomes very independent of the text that is provided, how visible it is in the frame and how it is formatted when only text information is employed for grocery classification. Rachid et al. [7] suggest the use of label texts to avoid training models on a large amount of image data. The proposed system verifies the information of the barcode reader with the product classifier in a self-checkout system through natural language processing (NLP) techniques. Textual information is extracted by the pre-trained Google Vision API. After several experiments, they concluded that combining word and character embeddings improves the model's accuracy. The classification based on these text embeddings was higher using a CNN-based model instead of LSTM or random multimodel deep learning (RMDL). The images in the source domain were in RGB format while in the target domain they are monochromatic. Given the domain shift, the authors evaluated transfer learning in NLP [101]. Comparing the performance of the CNN with Global Vectors (GloVe) embeddings trained on a large corpus with a pre-trained BERT [102] model available on [103], the results showed that the BERT was more robust for the domain change. The quality of the image and text description will determine how well the model performs. Products without text information or scarce information are frequently misclassified. The algorithm also performs poorly when there is partial or total text occlusion. Specifically for package label detection and transcription, the authors of [104] proposed the pre-trained SegLink [105] to detect text instances. Assuming that the texts are regular and horizontal, read from left to right, its transcription results from the combination of CNN and RNN networks, trained with Connectionist Temporal Classification (CTC) loss function. They compared two versions, the Bidirectional LSTM and the Bidirectional Gated Recurrent Unit (GRU). The models achieved similar recognition rates, although BGRU has the advantage of being a lightweight method. In addition to the limitation of recognising regular texts, dense background environments also pose challenges with the main difficulties for text transcription identified as the character size and the font. The results could be influenced by the CTC-based approach, since they are not as effective as attention-based methods. The model of Prabu et al. [106] applied the YOLOv5 [107] to detect a grocery product and the detected item passes through the text detector to obtain product information. The method could detect regular and irregular texts with a complex background, non-uniform spacing and different texts in a single image. The more important contributions to the robustness of the model are the use of ResNet50 + FPN, the use of an algorithm to select the centring point in the text centre line instead of picking a random point, and finally, the introduction of a post-processing technique named Width-Height-based Bounding Box Reconstruction (WHBBR) to enclose the starting and the ending characters. The detected text is cropped and then sent to Selective Context Attentional Text Recognizer (SCATTER) [108], a text recognition model. The framework is computationally expensive during training, but can detect and recognise objects in real-time and efficiently during testing. Given the lack of adequate databases for this problem, the performance of each module was tested separately in different benchmark datasets for each task. Text detection and recognition tasks were evaluated in scene text datasets.

The application of label detection and recognition techniques in the grocery context is relatively recent. In this sense, the evaluated methods are often based on approaches used in the detection and recognition of text in natural images. Therefore, with the intention of contributing to the development of package label analysis, we briefly review deep learning methods applied to scene text images. Traditional methods are excluded from this survey due to their limitations in more complex scenes and low-quality images. Similar to the previous section, the methods are organised into text detection, text recognition and end-to-end STR.

6.1. Text Detection

Following common taxonomies, text detection algorithms were divided into regression-based and segmentation-based methods. Table 7 summarises the comparisons of the text detection methods.

6.1.1. Regression-Based Methods

Considering text instances as objects, regression-based methods predict candidate coordinates to define the text regions. Given the physical limitations of bounding boxes, the frameworks introduce post-processing steps to handle multi-oriented or arbitrary-shape text. The papers [109,110] suggested the rotation of Region Proposal Network (RRPN), predicting the angle of inclination. Rotational anchor improves the performance of the model when compared with horizontal box prediction. Naturally, the additional boxes generate greater computing costs, but they remain near to the Faster R-CNN. RRD [111] learns the rotation-sensitive features to improve the results, while TextBoxes++ [112] combined with CRNN boosts the accuracy of arbitrary-oriented text detection. EAST [113] is a fast and accurate detector which predicts multi-oriented words or lines, without intermediate steps. Vertical and curved text are not detected by this method. SegLink [105] detects text segments and then links these segments to obtain the final text region. The method proposed in [114], SegLink++, was adjusted to detect dense and arbitrary-shaped scene texts with an attractive and repulsive link system. CounterNet [115] incorporated further steps to improve the arbitrary-shape text detection, emphasizing the adaptive-RPN and local orthogonal texture-aware module. Zhang et al. [116] joined the CNN with Graph Convolution Network (GCN) to predict small rectangular components and link to their neighbours to also handle arbitrary-shape text. A different approach was proposed in [117], where the text instances are established in the Fourier domain. MOST [118] uses a module to adjust the receptive field and the instance-wise IoU loss to deal with texts of different scales and aspect ratios.

6.1.2. Segmentation-Based Methods

Even with post-processing techniques, regression-based methods have important difficulties in dealing with arbitrary shape text. For this reason, arbitrary shape text detectors have adopted segmentation-based methods to represent text regions that rely on instance-level and pixel-level features. For more complex text scenes, these methods usually outperform regression-based methods. Nevertheless, they require complex and time-consuming post-processing steps to produce the final detection result, especially to separate adjacent text instances very close to each other.

Table 7. Quantitive results of text detection methods on ICDAR15 and SCUT-CTW1500 datasets. "MS" is the abbreviation of multi-scale inputs. FPS are indicative values since the methods run in different environments. Bold numbers represent the best results.

		ICDAR2015				SCUT-CTW1500				Text Style	
	Methods	**Recall**	**Precision**	**F-Measure**	**FPS**	**Recall**	**Precision**	**F-measure**	**FPS**	**Multi Orientation**	**Curved**
Regression-based	*R^2CNN* [109]	79.68%	85.62%	82.54%	0.44	-	-	-	-	√	×
	RRPN [110]	73.23%	82.17%	77.44%	-	-	-	-	-	√	×
	RRD [111]	79.00%	85.60%	82.20%	6.5	-	-	-	-	√	×
	RRD+MS [111]	80.00%	88.00%	83.80%	-	-	-	-	-	√	×
	EAST+PVANET2x [113]	73.47%	83.57%	78.20%	13.2	-	-	-	-	√	×
	TextBoxes++ [112]	76.70%	87.20%	82.90%	11.6	-	-	-	-	√	×
	TextBoxes++_MS [112]	78.50%	87.80%	82.90%	2.3	-	-	-	-	√	×
	MOST [118]	87.30%	89.1%	88.2%	10	-	-	-	-	√	×
	SegLink [105]	76.80%	73.10%	75.00%	-	-	-	-	-	√	×
	SegLink++ [114]	80.30%	83.70%	82.00%	7.1	79.80%	82.80%	81.30%	-	√	√
	CounterNet [115]	86.10%	87.60%	86.90%	3.5	84.10%	83.70%	83.90%	4.5	√	√
	Zhang et al. [116]	84.69%	88.53%	86.56%	-	83.02%	85.93%	84.45%	-	√	√
	FCENet [117]	84.20%	85.10%	84.60%	-	80.70%	85.70%	83.10%	-	√	√
	FCENet+DCN [117]	82.60%	90.10%	86.20%	-	83.40%	87.60%	85.50%	-	√	√
Segmentation-based	TextSnake [119]	80.40%	84.90%	82.60%	1.1	63.40%	65.40%	64.40%	-	√	√
	LOMO [120]	83.50%	91.30%	87.20%	-	69.60%	**89.20%**	78.40%	-	√	√
	LOMO+MS [120]	87.60%	87.80%	87.70%	-	76.50%	85.70%	80.80%	-	√	√
	PSENet-1s [121]	84.50%	86.90%	85.70%	1.6	79.70%	84.80%	82.20%	8.4	√	√
	CRAFT [122] [a]	84.30%	89.80%	86.90%	-	81.10%	86.00%	83.50%	-	√	√
	PAN [123]	81.90%	84.00%	82.90%	**26.1**	81.50%	85.50%	83.50%	**58.1**	√	√
	DBNet [124]	83.20%	91.80%	87.30%	12	80.20%	86.90%	83.40%	22	√	√
	DBNet++ [125]	83.90%	90.90%	87.30%	10	82.80%	87.90%	85.30%	21	√	√
	MOSTL [126]	84.56%	**92.50%**	**88.35%**	5	-	-	-	-	√	√
	RSCA [127]	82.70%	87.20%	84.90%	23.3	83.30%	86.60%	85.00%	30.4	√	√
	Wang et al. [128]	-	-	-	-	80.52%	86.91%	83.59%	25.1	√	√
	SAST [129]	87.09%	86.72%	86.91%	-	77.05%	85.31%	80.97%	27.63	√	√
	SAST+MS [129]	**87.34%**	87.55%	87.44%	-	81.71%	81.19%	81.45%	-	√	√
	Long et al. [130]	-	-	-	-	**87.44%**	84.56%	**85.97%**	-	√	√

[a] The paper indicates that the best result of FPS is 8.6, independently of the dataset.

TextSnake [119] predicts text of any shape considering text instances as a series of sequence disks. The geometry attributes of text instances are estimated via a Fully Convolutional Network (FCN). LOMO [120] also considered geometric properties to locate the characters and obtain quadrangular text regions. PSENet [121] implemented the progressive scale expansion network to solve the problem of closely adjacent text instances. CRAFT [122] was based on character awareness, predicting individual characters and their affinity to construct word bounding boxes. PAN [123] is a light computational detector that used a Pixel Aggregation network to aggregate text pixels precisely. Another real-time text detector is DBNet [124] which contains a Differentiable Binarization module to simplify the post-processing and achieve better results, being improved with DBNet++ [125]. The Adaptive Scale Fusion module was additionally introduced, enabling greater robustness of the model and maintaining its efficiency. Naiemi et al. [126] proposed a lightweight model, called MOSTL, to detect irregular text, including curved and vertical texts. The framework contains an improved ReLU layer (i.ReLU) and an improved inception layer which extract more valuable text information. The RSCA [127] network also had the ability to detect curved texts in real time. The strategy was to apply local-context-aware upsampling to be effective with less computational cost and dynamic text-spine labeling for simplifying label generation and assignment. Wang et al. [128] proposed an innovative and robust framework combining position and semantic information to avoid degrading segmentation-based methods due to inaccurate annotations of text pixels. The SAST model suggested in [129] integrates high-level object knowledge and low-level pixel information in a single shot to overcome the difficulty of separating adjacent text instances. The model detects scene text of arbitrary shapes with high accuracy and efficiency. A state-of-the-art unified detector was proposed in [130]. Without complex post-processing, the model detects text as masks and posteriorly groups them into clusters.

Automatic text detection is challenging, namely when dealing with blur images, crowded backgrounds, multi-orientations, arbitrary shapes and font size variations. Some methods proposed axis-aligned bounding boxes or polygons to better enclose the text. Considering the background information that is still present in more irregular texts, other researchers proposed segmentation methods that adapt to arbitrary shape texts. The segmentation approach has the disadvantage of being vulnerable to false detection when two text instances are near one another. Despite the use of several strategies to address these issues, some text instances are not detected, some regions are wrongly classed as containing text, text regions are incorrectly separated into words, or individual portions are erroneously linked. Further studies are required in the text detection area.

6.2. Text Recognition

In this subsection, we briefly describe methods that recognise both regular and irregular texts. ASTER [131], an updated version of RARE [132], employed a rectification module based on Spatial Transform Networks to rectify irregular texts and recognise arbitrary shape texts. ScRN [133] achieved better results with the symmetry-constrained rectification module. Nevertheless, the curved text with terminal letters that are almost horizontally oriented and close to the edges of the image causes trouble for the rectification module. CA-FCN [134] indicated the characters at the pixel level. The character attention mechanism, joined with the word formation module, predicts the position and recognises the script. However, the performance tends to suffer without a sequence learning approach. The aforementioned issue is addressed in [135]. The model extracts high-level 2D spatial features and transforms them into a 1D feature sequence. Using word-level annotation, Li et al. [136] applied the two-dimensional attention module to handle the complex spatial layout of irregular text. To tackle the attention drift and inappropriate threshold selection on segmentation maps, Wan et al. suggested the TextScanner [137]. The algorithm has the advantage of being fast and adequate for long text. Character-level annotations can aid TextScanner to be more efficient when pre-training on artificial data. Considering that text recognition depends on visual perception information and high-level text semantic context understanding, the SRN network [138] incorporated the global semantic reasoning module (GSRM), parallel visual attention module (PVAM) and visual-semantic fusion decoder (VSFD). DPAN [139] differs from the previous parallel-decoupled encoder–decoder framework, designing a dual parallel attention network to alleviate visual misalignment in hard samples. CDistNet [140] is a transformer-based encoder–decoder framework that integrates character feature interactions among visual, semantic and position spaces for recognising more difficult texts. Linguistic knowledge proved to contribute to the refinement of character sequences; therefore, the MATRN model [141] combined language-aware visual and semantic features, exploring the multiple combinations of multi-modal processes with bi-directional fusion to enhance each feature. Visual and linguistic features were combined in [142] to be robust to different language texts. The suggested model (S-GTR) is a Graph Convolutional Network for textual reasoning that joins the pixels based on their spatial context similarity. In order to use bidirectional language context, the methods implemented a two-stage ensemble approach. To overcome the time constraints of these methods, Bautista and Atienza designed a simple model that uses permutation language modelling, called PARSeq [143]. The context-free and context-aware autoregressive inference is unified and refined using bidirectional context. In contrast to sequence-to-sequence and segmentation-based approaches, Cai et al. [144] proposed a classification-based model with similar results, the CSTR. The model's merits include simplicity and a lack of costly character-level annotations.

Multi-orientation, perspective distortion, diversity of the internal properties of texy regions (font, size, shape, space), variable length and multilingual content are only a few of the difficulties that text recognition must overcome. The introduction of a list of words for each image was initially suggested to improve the accuracy of results. Given the limited applicability, some researchers incorporate language models to expand the number of possibilities and rectify the predictions according to prior knowledge. The strategy increases the computational costs and limits the application for that specific language. Besides multi-lingual text, the long arbitrarily shaped text is another unsolved challenge that requires more data to investigate this subject. A balance between speed and accuracy is always required. The segmentation-based method is an effective and simple approach, but pixel-level annotation is necessary. Sequence-based recognition is another popular method that produces good outcomes but has the drawback of being complex. Table 8 provides a more detailed comparison of the state-of-the-art text recognisers' performance on benchmark datasets with irregular text. None of the displayed results were generated from a lexicon.

Table 8. Accuracy results of text recognition methods on ICDAR15, SVT Perspective and CUTE80. "Anno." is short for required annotations.

Methods	Anno.	ICDAR15	SVTP	CUTE80
ASTER [131]	word	0.761	0.785	0.795
ScRN [133]	word, char	0.784	0.811	0.906
CA-FCN+data [134]	word, char	-	-	0.799
CAPNet [135]	word, char	0.766	0.788	0.868
SAR [136]	word	0.788	0.864	0.896
TextScanner [137]	word, char	0.835	0.848	0.916
SRN [138]	word	0.827	0.851	0.878
DPAN [139]	word	0.855	0.890	0.919
CDistNet [140]	word, char	0.860	0.887	0.934
MATRN model [141]	word, char	0.828	0.906	0.935
S-GTR [142]	word	0.873	0.906	0.947
PARSeq [143]	word	0.896	0.957	0.983
STN-CSTR [144]	word	0.820	0.862	-

6.3. Text Spotting

Text spotting systems perform text detection and recognition in an end-to-end way. The base idea is to share the same CNN feature extractor with the detector and classifier. For fast-oriented text spotting, FOTS [145] applied the RoIRotate operator from convolution feature maps. Like the MOSTL text detector, Naiemi et al. [146] proposed a unified framework that also used improved ReLU and inception blocks. The new LWDP algorithm enhanced the character recognition results. The suggested approach could handle many font sizes and text orientations, including vertical texts. TextDragon [147], inspired by TextSnake, used the differentiable operator named RoISlide for connecting arbitrary-shaped text detection and recognition. In ABCNet [148] model, the Bezier curve was introduced to describe arbitrary text instances and BezierAlign to extract accurate features. The model was refined, generating an accurate real-time text spotter. The ABCNet V2 [149] considered the bidirectional multi-scale features, inserted the character attention module without requiring character-level annotation and integrated an adaptive end-to-end training strategy. The ARTS [150] framework enhanced these end-to-end methods, propagating the recognition loss back into the detection branch with auto-rectification module. Mask TextSpotter [151] was a segmentation model influenced by Mask R-CNN. The detection and recognition are at the character level. Since character-level annotations are typically unavailable in public databases, Mask TextSpotter V3 [152] addressed this limitation, introducing the Segmentation Proposal Network (SPN). Both methods incorporated the RoIAlign to conserve more precise information. Another model inspired by Mask R-CNN is suggested in [153]. Considering the authors that demonstrate that feature rectification degrades the performance of irregular shape texts, the feature rectification step was ignored, with it being enough to separate text from the background. MANGO [154] introduced the position-aware mask attention module to perceive the position of text instances, allowing it to retain the global spatial features. Character-level annotations are unnecessary, and the RoI operation to link text detection and recognition was removed. Another framework free from RoI operations is the TExt Spotting TRansformers (TESTR) [155]; furthermore, heuristics-driven post-processing techniques are excluded in the single-encoder dual-decoder framework. The model is suitable for both the polygonal and Bezier curve annotations. To overcome the challenge of determining the space between words, Wang et al. proposed the AE TextSpotter [156] framework, which learns both linguistic and visual representations. Swin-TextSpotter [157], a transformer-based model, unified the two tasks with the recognition conversion mechanism, leveraging the synergy of both. This contrasts with approaches that typically share the backbone to integrate detection and recognition.

The recent text-spotting methods combined text detection and text recognition problems in order to optimise these two related tasks in a unified pipeline. When the system works independently, the text detection errors interfere with text recognition, but in a cooperative system, the text recogniser may reduce the false detections or even enhance the limits of text region. The most efficient technique of information sharing across the modules has been researched, and the first developments are currently being made. Table 9 depicts the performance of several text spotting algorithms on the ICDAR2015 dataset, using three types of lexicons, and Table 10 deepens the comparison of models on the Total Text dataset.

Table 9. Comparison of end-to-end text recognition in ICDAR2015. The acronyms "S", "W" and "G" represent the "Strong", "Weak" and "Generic" lexicon, respectively.

Methods	**ICDAR15 - End-to-End**		
	S	**W**	**G**
FOTS [145]	0.811	0.759	0.608
TextDragon [147]	0.825	0.783	0.651
ABCNet+MS [148]	-	-	-
ABCNet V2 [149]	0.827	0.785	0.730
ARTS [150]	0.815	0.773	0.687
Mask TextSpotter [151]	0.830	0.777	0.735
Mask TextSpotter V3 [152]	0.833	0.781	0.742
Unconstrained [153]	0.855	0.819	0.699
MANGO [154]	0.818	0.789	0.673
TESTR [155]	0.852	0.794	0.736
AE TextSpotter [156]	-	-	-
SwinTextSpotter [157]	0.839	0.773	0.705

Table 10. Comparison of end-to-end text recognition in Total-Tex dataset. For more details of the networks, please consult the respective papers.

Methods	**Detection**	**End-to-End**		**FPS**
	F-Measure	**None**	**Full**	
FOTS [145]	0.440	0.322	0.359	-
TextDragon [147]	0.803	0.488	0.748	-
ABCNet+MS [148]	-	0.695	0.784	6.9
ABCNet V2 [149]	0.870	0.704	0.781	10
ARTS-RT [150]	0.803	0.659	0.781	28.0
ARTS-S [150]	0.865	0.771	0.851	10.5
Mask TextSpotter [151]	0.613	0.529	0.718	4.8
Mask TextSpotter V3 [152]	-	0.712	0.784	-
Unconstrained [153]	0.864	0.707	-	-
MANGO [154]	-	0.729	0.836	4.3
TESTR-Bezier [155]	0.880	0.716	0.833	5.5
TESTR-Polygon [155]	0.869	0.733	0.839	5.3
AE TextSpotter [156]	-	-	-	-
SwinTextSpotter [157]	0.880	0.743	0.841	-

7. Challenges and Opportunities

The automatic detection and recognition of grocery items has many applications with high economic impact. Moreover, the ability to analyse product labels is becoming increasingly relevant mostly as a way to improve the customer experience, which is expected to translate to higher sales and fidelity, but also as a means to assist people with impairments. Nevertheless, the process is highly challenging due to the inherent characteristics of the scenario such as visual clutter or high density in shelves. Moreover, when focusing on the label, processing the information introduces additional challenges due to small text size or blurry appearance. The difficulties associated with these scenarios typically increase when we descend the hierarchy of classification. Fine-grained classification is more challenging due to intra-class variance and inter-class similarity. Grocery items of the same macro-category share similar visual characteristics in size, shape, colour and inscriptions. On the other hand, products of the same brand are frequently only recognised through slight packaging variations. This is the case of products that differ by flavour or quantity. When these distinctions only occur in small areas, there is no guarantee that images taken from a particular angle can capture the mentioned differences. Moreover, several environmental conditions, including lighting, background and occlusions, may significantly impact product recognition, particularly with micro-categories. Consequently, the recognition model should be able to identify the nuances of packages and be agnostic to environmental factors. The items typically follow an established layout in a grocery store. Those who belong to a general category are placed together on the racks. As a result, along with local features, context may serve as a guideline for identifying similar products on shelves.

The grocery market is constantly evolving with products launched frequently and their appearance may change over time to appeal to consumers. Hence, in addition to being robust, the recognition model must also be flexible enough to adapt or retrain itself with less effort when a new product or packaging appears. Training a CNN with just new classes is not feasible because it cannot preserve the past learned classes; on the other hand, the frequent retraining of the complete network with all classes is impractical, given the existence of thousands of products. To handle this issue, incremental learning or one-shot learning are two possible approaches. Within incremental learning methods, there is the fine-tuning approach, which consists of using the pre-trained network and adjusting the old parameters to adapt to the new data. However, this technique risks learning new classes while forgetting the old ones. In the retraining process, freezing some layers and setting up adequate parameters, such as a small learning rate, are important strategies to keep the old knowledge. Model configuration is a tricky task and becomes even more challenging with the constantly growing number of classes. Recently, knowledge distillation has been explored to handle the class incremental problem [158–160]. The foundation of one-shot learning for computer vision tasks is Siamese neural networks (SNN); these special types of CNNs are trained to evaluate the distance between features in two input images. SNNs have the enormous benefit of not requiring intensive retraining to identify new classes after having been trained on large datasets. Regarding speed and accuracy, the SNNs may outperform other types of methods. Nevertheless, in terms of memory and computational resources, they are expensive since they need to train two models. Osokin et al. [84] proposed a one-shot object detector, and Domingo et al. [92] classified grocery products with Siamese neural networks. This topic might attract increasing interest when additional datasets are released.

The huge diversity of grocery items poses a challenge to developing a suitable dataset. Although deep networks perform better than conventional detection and recognition methods, they are sometimes constrained by scarcity and under-represented data. Collecting grocery images captured in the real environment and subsequently annotating the dataset are laborious tasks. Therefore, most available datasets represent a small number of classes and/or contain restricted images per class. This limitation damages the efficiency of the model and creates obstacles to being expanded to real-world applications. In that sense, to increase the size of the training set, some frameworks apply data augmentation [78,90,98], and other researchers suggest generating realistic samples by GAN to achieve reliable results [5,97]. Another strategy widely used to minimise the problem of the lack of adequate datasets is the compilation of grocery product photos taken under a controlled environment, occasionally from a range of perspectives, that act as a reference of the class. Almost all datasets designed for image identification used this approach and, as a result, most retail recognition methods rely on template-matching-based techniques. This technique involves the interpretation of all positions of the template and measuring how well the template and the query image match at each location. The process is repeated for each template, which is very time-consuming. Instead of a feature extraction match, the comparison can be made with feature vectors of two convolutional nets using SNNs. They also have high computational costs. In that sense, further investigations are required in order to develop an accurate and efficient model that only needs a single image to classify.

Whether the system uses template-matching-based techniques or Siamese networks, it will likely be challenging to correctly identify the class due to minor differences in the same product. Even when comparing identical products, the orientation, illumination, reflection, resolution and other aspects of the shelf image may differ from those of the reference image; the training data environment does not match the deployment data environment, and the model's performance decreases naturally. The domain shift issue can also be minimised with a GAN algorithm since it is trained to learn how to convert iconic samples into apparent in-store images. Then, the model, trained over more data, is able to extract rich invariant features. For cross-domain detection and recognition, transfer learning is commonly employed, with a vast majority of frameworks using a model pre-trained on ImageNet, such as [4,44,90,98]. In [90], after training the network on ImageNet, the model was fine-tuned on the Grocery Store dataset [46] for the domain adaption. Wang et al. [87] design a CNN model that combines two technical modules. The adversarial module handles the domain shift by gradually reducing the disparity between different domains and the self-attention module captures discriminative image features that are essential for fine-grained recognition. Other domain generalization techniques should be researched, including the possibility of adapting strategies defined for other application scenarios [161].

In the presence of noise and perturbations, Deep Convolutional Neural Networks and transformer-based architectures are two effective solutions, with the latter requiring even more sufficient data [162]. A robust model is constructed at the cost of heavy frameworks that are computationally expensive. These characteristics may be unfeasible in the implementation of real-time models on edge devices, but techniques for image pre-processing can be used to enhance model performance. Finding a fair balance between accuracy and efficiency is crucial, though. For example, Sinha et al. [98] avoid template-matching-based object detection because it is time-consuming; it selected a light backbone, the ResNet-18, and extracted multi-scale features to improve the product location.

Product label analysis is a particular case of scene text recognition. Recent studies focused on providing an efficient solution to STR in tough environments, such as texts with arbitrary shape and orientation, various scales and text fonts, blur, distortion, occlusion, multilingual texts or even with a complex background. In fact, in a grocery environment, all these challenging tasks may occur. Independently of the deep learning model's architecture, a considerable amount of adequate data is crucial to train and assess the model. Given the scarcity of high-quality samples and annotations, creating a synthetic dataset based on the limited data available is a possible strategy. For example, in [163], the authors suggested an efficient framework to generate realistic images via a 3D graphics engine. Subsequently, the realistic and real-world sample distribution between training and testing should be evaluated. Another path could be the implementation of the GAN algorithm. Luo et al. [164] separated the text instance from the wild background benefiting from a generative adversarial architecture. Moreover, the design of effective data augmentation approaches could be a sustainable solution to overcome the accessible data problem. On the other hand, developing unsupervised algorithms, such as [165,166], can be further studied to take advantage of unlabelled real-world data. The scarcity of appropriate datasets is not solely a result of the lack of images taken in-store, but also due to packages written in different languages. Text recognition algorithms are frequently trained to identify English texts, and performance in other languages is critical. The first steps are being taken in the creation of multi-lingual datasets and, once again, synthetic multi-lingual datasets are an interesting solution to overcome this limitation.

The performance of a text recogniser can be affected by complex text images. As image preprocessing techniques can potentially improve the results, even of a robust model, analysis of the integration of these two components can be relevant. Another interesting factor is the performance of end-to-end systems. Researchers are focused on text spotting development that benefits from a joint optimisation of detector and recogniser. Nevertheless, some limitations need to be addressed, such as the effective way to link and communicate text detection and recognition and to enhance joint optimization. This is a recent area that requires more investigation. Another area that has drawn attention is the introduction of a language model. The verse of NLP methods in STR to acquire linguistic information has yielded promising results [101,143], with the disadvantage of increased computational costs. The effective way to fuse visual and linguistic information while maintaining high efficiency still raises doubts.

The most significant difficulties in identifying grocery products are noted along with potential research directions:

- **Poor image quality**, as blurring and perspective distortion can be enhanced with image preprocessing techniques.
- **Lack of datasets** can be overcome by creating synthetic datasets, applying data augmentation or selecting one-shot or unsupervised learning.
- **Shift environment** between training data and test data requires strategies such as GAN network, data augmentation and transfer learning.
- **New classes** are regularly added and the use of incremental learning or one-shot learning prevents the model from being completely retrained.

Ultimately, we conclude that grocery product recognition is still a challenging task because of intra-class variance and inter-class similarity; new data arriving frequently; lack of huge quality data; cross-domain between training samples and test samples; complex background; noise and perturbation of query images; irregular texts; multi-lingual texts, and the need for a light model to perform in real-time. More research is clearly required in this field in order to make progress, hence some guidelines for each problem were suggested.

Author Contributions: Conceptualization, P.C., P.V., V.G. and J.N.; methodology, P.C. and P.V.; validation, V.G., J.N., P.V. and P.C.; investigation, V.G. and J.N.; writing—original draft preparation, V.G. and J.N.; writing—review and editing, V.G., J.N., P.V. and P.C.; visualization, V.G., J.N., P.V. and P.C.; supervision, P.C.; funding acquisition, P.C. and P.V. All authors have read and agreed to the published version of the manuscript.

Funding: This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The work presented in this paper has been supported by the European Commission under project WATSON (grant agreement 101084265-WATSON).

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

References

1. Wei, Y.; Tran, S.N.; Xu, S.; Kang, B.H.; Springer, M. Deep Learning for Retail Product Recognition: Challenges and Techniques. *Comput. Intell. Neurosci.* **2020**, 2020, 8875910. [CrossRef]
2. Franco, A.; Maltoni, D.; Papi, S. Grocery product detection and recognition. *Expert Syst. Appl.* **2017**, *81*, 163–176. [CrossRef]
3. Goldman, E.; Herzig, R.; Eisenschtat, A.; Goldberger, J.; Hassner, T. Precise detection in densely packed scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–17 June 2019; pp. 5227–5236.
4. Tonioni, A.; Serra, E.; Di Stefano, L. A deep learning pipeline for product recognition on store shelves. In Proceedings of the 2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS), Sophia Antipolis, France, 12–14 December 2018; pp. 25–31.
5. Wei, X.S.; Cui, Q.; Yang, L.; Wang, P.; Liu, L. RPC: A Large-Scale Retail Product Checkout Dataset. *arXiv* **2019**, arXiv:1901.07249.
6. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the Advances in Neural Information Processing Systems, Montréal, QC, Canada, 8–13 December 2014; pp. 2672–2680.
7. Oucheikh, R.; Pettersson, T.; Löfström, T. Product verification using OCR classification and Mondrian conformal prediction. *Expert Syst. Appl.* **2022**, *188*, 115942. [CrossRef]
8. Georgiadis, K.; Kordopatis-Zilos, G.; Kalaganis, F.; Migkotzidis, P.; Chatzilari, E.; Panakidou, V.; Pantouvakis, K.; Tortopidis, S.; Papadopoulos, S.; Nikolopoulos, S.; et al. Products-6K: A Large-Scale Groceries Product Recognition Dataset. In Proceedings of the 14th PErvasive Technologies Related to Assistive Environments Conference, Corfu, Greece, 29 June–2 July 2021; pp. 1–7.
9. Klasson, M.; Zhang, C.; Kjellström, H. Using variational multi-view learning for classification of grocery items. *Patterns* **2020**, *1*, 100143. [CrossRef] [PubMed]

10. Wang, J.; Min, W.; Hou, S.; Ma, S.; Zheng, Y.; Jiang, S. LogoDet-3K: A Large-scale Image Dataset for Logo Detection. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2020**, *18*, 1–19. [CrossRef]
11. Melek, C.G.; Sonmez, E.B.; Albayrak, S. A survey of product recognition in shelf images. In Proceedings of the 2017 International Conference on Computer Science and Engineering (UBMK), Antalya, Turkey, 5–7 October 2017; pp. 145–150. [CrossRef]
12. Santra, B.; Mukherjee, D.P. A comprehensive survey on computer vision based approaches for automatic identification of products in retail store. *Image Vis. Comput.* **2019**, *86*, 45–63. [CrossRef]
13. Kulyukin, V.; Gharpure, C.; Nicholson, J. RoboCart: Toward robot-assisted navigation of grocery stores by the visually impaired. In Proceedings of the 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, Edmonton, AB, Canada, 2–6 August 2005; pp. 2845–2850. [CrossRef]
14. Seeing-AI. Available online: https://www.microsoft.com/en-us/ai/seeing-ai (accessed on 30 December 2022).
15. Yuka. Available online: https://yuka.io/en/ (accessed on 30 December 2022).
16. Open Food Facts. Available online: https://github.com/openfoodfacts (accessed on 30 December 2022).
17. Lookout—Assisted Vision. Available online: https://play.google.com/store/apps/details?id=com.google.android.apps.accessibility.reveal&hl=en_GB&gl=US (accessed on 30 December 2022).
18. Identify Products with Your Echo Show. Available online: https://www.amazon.com/gp/help/customer/display.html?nodeId=G5723QKAVR8Z9S26 (accessed on 30 December 2022).
19. OrCam MyEye. Available online: https://www.orcam.com/en/ (accessed on 30 December 2022).
20. Wine-Searcher. Available online: https://www.wine-searcher.com/wine-searcher (accessed on 30 December 2022).
21. Amazon Go. Available online: https://www.amazon.com/ref=footer_us (accessed on 30 December 2022).
22. Varga, L.A.; Koch, S.; Zell, A. Comprehensive Analysis of the Object Detection Pipeline on UAVs. *Remote Sens.* **2022**, *14*, 5508. [CrossRef]
23. Minh, T.N.; Sinn, M.; Lam, H.T.; Wistuba, M. Automated Image Data Preprocessing with Deep Reinforcement Learning. *arXiv* **2018**, arXiv:1806.05886.
24. Chen, C.; Chen, Q.; Xu, J.; Koltun, V. Learning to See in the Dark. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3291–3300. [CrossRef]
25. Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. Photo-realistic single image super-resolution using a generative adversarial network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4681–4690.
26. Dudhane, A.; Zamir, S.W.; Khan, S.; Khan, F.S.; Yang, M.H. Burst Image Restoration and Enhancement. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–20 June 2022; pp. 5749–5758. [CrossRef]
27. Bhat, G.; Danelljan, M.; Van Gool, L.; Timofte, R. Deep Burst Super-Resolution. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 9205–9214. [CrossRef]
28. Bhat, G.; Danelljan, M.; Yu, F.; Van Gool, L.; Timofte, R. Deep Reparametrization of Multi-Frame Super-Resolution and Denoising. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 2440–2450. [CrossRef]
29. Wronski, B.; Garcia-Dorado, I.; Ernst, M.; Kelly, D.; Krainin, M.; Liang, C.K.; Levoy, M.; Milanfar, P. Handheld Multi-Frame Super-Resolution. *ACM Trans. Graph.* **2019**, *38*, 1–18. [CrossRef]
30. Godard, C.; Matzen, K.; Uyttendaele, M. Deep Burst Denoising. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 560–577.
31. Lecouat, B.; Eboli, T.; Ponce, J.; Mairal, J. High Dynamic Range and Super-Resolution from Raw Image Bursts. *ACM Trans. Graph.* **2022**, *41*, 1–21. [CrossRef]
32. Luo, Z.; Li, Y.; Cheng, S.; Yu, L.; Wu, Q.; Wen, Z.; Fan, H.; Sun, J.; Liu, S. BSRT: Improving Burst Super-Resolution with Swin Transformer and Flow-Guided Deformable Alignment. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022; IEEE Computer Society: Los Alamitos, CA, USA, 2022; pp. 997–1007. [CrossRef]
33. Deudon, M.; Kalaitzis, A.; Goytom, I.; Arefin, M.R.; Lin, Z.; Sankaran, K.; Michalski, V.; Kahou, S.E.; Cornebise, J.; Bengio, Y. HighRes-net: Recursive Fusion for Multi-Frame Super-Resolution of Satellite Imagery. *arXiv* **2020**, arXiv:2002.06460.
34. Mehta, N.; Dudhane, A.; Murala, S.; Zamir, S.W.; Khan, S.; Khan, F.S. Adaptive Feature Consolidation Network for Burst Super-Resolution. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), New Orleans, LA, USA, 19–20 June 2022; pp. 1278–1285. [CrossRef]
35. An, T.; Zhang, X.; Huo, C.; Xue, B.; Wang, L.; Pan, C. TR-MISR: Multiimage Super-Resolution Based on Feature Fusion With Transformers. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2022**, *15*, 1373–1388. [CrossRef]
36. Zamir, S.W.; Arora, A.; Khan, S.; Hayat, M.; Khan, F.S.; Yang, M.H.; Shao, L. Learning Enriched Features for Real Image Restoration and Enhancement. In Proceedings of the Computer Vision—ECCV 2020, Glasgow, UK, 23–28 August 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 492–511.
37. Nguyen, T.P.H.; Cai, Z.; Nguyen, K.; Keth, S.; Shen, N.; Park, M. Pre-processing Images using Brightening, CLAHE and RETINEX. *arXiv* **2020**, arXiv:2003.10822.

38. Mehrnejad, M.; Albu, A.B.; Capson, D.; Hoeberechts, M. Towards Robust Identification of Slow Moving Animals in Deep-Sea Imagery by Integrating Shape and Appearance Cues. In Proceedings of the 2014 ICPR Workshop on Computer Vision for Analysis of Underwater Imagery, Stockholm, Sweden, 24 August 2014; pp. 25–32. [CrossRef]
39. Reza, A.M. Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement. *J. VLSI Signal Process. Syst. Signal, Image Video Technol.* **2004**, *38*, 35–44. [CrossRef]
40. Parthasarathy, S.; Sankaran, P. An automated multi Scale Retinex with Color Restoration for image enhancement. In Proceedings of the 2012 National Conference on Communications (NCC), Kharagpur, India, 3–5 February 2012; pp. 1–5. [CrossRef]
41. Jiang, Y.; Gong, X.; Liu, D.; Cheng, Y.; Fang, C.; Shen, X.; Yang, J.; Zhou, P.; Wang, Z. EnlightenGAN: Deep Light Enhancement Without Paired Supervision. *IEEE Trans. Image Process.* **2021**, *30*, 2340–2349. [CrossRef]
42. Loh, Y.P.; Chan, C.S. Getting to know low-light images with the Exclusively Dark dataset. *Comput. Vis. Image Underst.* **2019**, *178*, 30–42. [CrossRef]
43. Koshy, A.; MJ, N.B.; Shyna, A.; John, A. Preprocessing Techniques for High Quality Text Extraction from Text Images. In Proceedings of the 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT), Chennai, India, 25–26 April 2019; pp. 1–4. [CrossRef]
44. Geng, W.; Han, F.; Lin, J.; Zhu, L.; Bai, J.; Wang, S.; He, L.; Xiao, Q.; Lai, Z. Fine-Grained Grocery Product Recognition by One-Shot Learning. In Proceedings of the 26th ACM International Conference on Multimedia, MM '18, Seoul, Republic of Korea, 22–26 October 2018; Association for Computing Machinery: New York, NY, USA, 2018; pp. 1706–1714. [CrossRef]
45. George, M.; Floerkemeier, C. Recognizing products: A per-exemplar multi-label image classification approach. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 440–455.
46. Klasson, M.; Zhang, C.; Kjellström, H. A Hierarchical Grocery Store Image Dataset with Visual and Semantic Labels. In Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 7–11 January 2019.
47. Merler, M.; Galleguillos, C.; Belongie, S. Recognizing Groceries in situ Using in vitro Training Data. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8. [CrossRef]
48. Jund, P.; Abdo, N.; Eitel, A.; Burgard, W. The Freiburg Groceries Dataset. *arXiv* **2016**, arXiv:1611.05799.
49. Follmann, P.; Böttger, T.; Härtinger, P.; König, R.; Ulrich, M. MVTec D2S: Densely Segmented Supermarket Dataset. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Springer International Publishing: Cham, Switzerland, 2018; pp. 581–597.
50. Chen, F.; Zhang, H.; Li, Z.; Dou, J.; Mo, S.; Chen, H.W.; Zhang, Y.; Ahmed, U.; Zhu, C.; Savvides, M. Unitail: Detecting, Reading, and Matching in Retail Scene. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–24 October 2022.
51. Peng, J.; Xiao, C.; Wei, X.; Li, Y. RP2K: A Large-Scale Retail Product Dataset for Fine-Grained Image Classification. *arXiv* **2020**, arXiv:2006.12634.
52. India, A. Store Shelf Images and Product Images for Retail. 2022. Available online: https://www.kaggle.com/datasets/amanindiamuz/store-shelf-images-and-product-images-for-retial?select=url (accessed on 30 December 2022).
53. WebMarket. Available online: https://www.kaggle.com/datasets/manikchitralwar/webmarket-dataset (accessed on 30 December 2022).
54. Wang, J.; Min, W.; Hou, S.; Ma, S.; Zheng, Y.; Wang, H.; Jiang, S. Logo-2K+: A Large-Scale Logo Dataset for Scalable Logo Classification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, New York, USA, 7–12 February 2020.
55. Hou, Q.; Min, W.; Wang, J.; Hou, S.; Zheng, Y.; Jiang, S. FoodLogoDet-1500: A Dataset for Large-Scale Food Logo Detection via Multi-Scale Feature Decoupling Network. In Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 20–24 October 2021.
56. Su, H.; Gong, S.; Zhu, X. Weblogo-2m: Scalable logo detection by deep learning from the web. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 270–279.
57. Jaderberg, M.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Synthetic Data and Artificial Neural Networks for Natural Scene Text Recognition. In Proceedings of the Workshop on Deep Learning, NIPS, Montréal, QC, Canada, 12–13 December 2014.
58. Gupta, A.; Vedaldi, A.; Zisserman, A. Synthetic data for text localisation in natural images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2315–2324.
59. Sun, Y.; Ni, Z.; Chng, C.K.; Liu, Y.; Luo, C.; Ng, C.C.; Han, J.; Ding, E.; Liu, J.; Karatzas, D.; et al. ICDAR 2019 Competition on Large-Scale Street View Text with Partial Labeling—RRC-LSVT. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20–25 September 2019 2019; pp. 1557–1562.
60. Yao, C.; Bai, X.; Liu, W.; Ma, Y.; Tu, Z. Detecting texts of arbitrary orientations in natural images. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1083–1090.
61. Risnumawan, A.; Shivakumara, P.; Chan, C.S.; Tan, C.L. A robust arbitrary text detection system for natural scene images. *Expert Syst. Appl.* **2014**, *41*, 8027–8048. [CrossRef]
62. Liu, Y.; Jin, L.; Zhang, S.; Luo, C.; Zhang, S. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recognit.* **2019**, *90*, 337–345. [CrossRef]
63. Mishra, A.; Karteek, A.; Jawahar, C.V. Scene Text Recognition using Higher Order Language Priors. In Proceedings of the BMVC, Surrey, UK, 3–7 September 2012.

64. Wang, K.; Babenko, B.; Belongie, S.J. End-to-end scene text recognition. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 1457–1464.
65. Lucas, S.M.M.; Panaretos, A.; Sosa, L.; Tang, A.; Wong, S.; Young, R. ICDAR 2003 robust reading competitions. In Proceedings of the Seventh International Conference on Document Analysis and Recognition, Edinburgh, UK, 6 August 2003; pp. 682–687.
66. Karatzas, D.; Shafait, F.; Uchida, S.; Iwamura, M.; Bigorda, L.G.i.; Mestre, S.R.; Mas, J.; Mota, D.F.; Almazàn, J.A.; de las Heras, L.P. ICDAR 2013 Robust Reading Competition. In Proceedings of the 2013 12th International Conference on Document Analysis and Recognition, Washington, DC, USA, 25–28 August 2013; pp. 1484–1493. [CrossRef]
67. de Campos, T.E.; Babu, B.R.; Varma, M. Character recognition in natural images. In Proceedings of the International Conference on Computer Vision Theory and Applications, Lisbon, Portugal, 5–8 February 2009.
68. Karatzas, D.; Gomez-Bigorda, L.; Nicolaou, A.; Ghosh, S.; Bagdanov, A.; Iwamura, M.; Matas, J.; Neumann, L.; Chandrasekhar, V.R.; Lu, S.; et al. ICDAR 2015 competition on Robust Reading. In Proceedings of the 2015 13th International Conference on Document Analysis and Recognition (ICDAR), Tunis, Tunisia, 23–26 August 2015; pp. 1156–1160. [CrossRef]
69. Phan, T.Q.; Shivakumara, P.; Tian, S.; Tan, C.L. Recognizing Text with Perspective Distortion in Natural Scenes. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, NSW, Australia, 1–8 December 2013; pp. 569–576. [CrossRef]
70. Veit, A.; Matera, T.; Neumann, L.; Matas, J.; Belongie, S.J. COCO-Text: Dataset and Benchmark for Text Detection and Recognition in Natural Images. *arXiv* **2016**, arXiv:1601.07140.
71. Chng, C.K.; Chan, C.S. Total-Text: A Comprehensive Dataset for Scene Text Detection and Recognition. In Proceedings of the 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 9–15 November 2017; Volume 1, pp. 935–942.
72. Nguyen, N.; Nguyen, T.; Tran, V.; Tran, M.T.; Ngo, T.D.; Nguyen, T.H.; Hoai, M. Dictionary-Guided Scene Text Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 7383–7392.
73. Nayef, N.; Patel, Y.; Busta, M.; Chowdhury, P.N.; Karatzas, D.; Khlif, W.; Matas, J.; Pal, U.; Burie, J.C.; Liu, C.L.; et al. ICDAR2019 Robust Reading Challenge on Multi-lingual Scene Text Detection and Recognition—RRC-MLT-2019. In Proceedings of the 2019 International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 20–25 September 2019; pp. 1582–1587.
74. Chng, C.K.; Liu, Y.; Sun, Y.; Ng, C.C.; Luo, C.; Ni, Z.; Fang, C.; Zhang, S.; Han, J.; Ding, E.; et al. ICDAR2019 Robust Reading Challenge on Arbitrary-Shaped Text (RRC-ArT). *arXiv* **2019**, arXiv:1909.07145.
75. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
76. Baz, I.; Yoruk, E.; Cetin, M. Context-aware hybrid classification system for fine-grained retail product recognition. In Proceedings of the 2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP), Bordeaux, France, 11–12 July 2016; pp. 1–5.
77. Yörük, E.; Öner, K.T.; Akgül, C.B. An efficient hough transform for multi-instance object recognition and pose estimation. In Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), Cancun, Mexico, 4–8 December 2016; pp. 1352–1357.
78. Santra, B.; Shaw, A.K.; Mukherjee, D.P. An end-to-end annotation-free machine vision system for detection of products on the rack. *Mach. Vis. Appl.* **2021**, *32*, 56. [CrossRef]
79. Zhang, J.; Marszałek, M.; Lazebnik, S.; Schmid, C. Local features and kernels for classification of texture and object categories: A comprehensive study. *Int. J. Comput. Vis.* **2007**, *73*, 213–238. [CrossRef]
80. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
81. Qiao, S.; Shen, W.; Qiu, W.; Liu, C.; Yuille, A. Scalenet: Guiding object proposal generation in supermarkets and beyond. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1791–1800.
82. Santra, B.; Shaw, A.K.; Mukherjee, D.P. Graph-based non-maximal suppression for detecting products on the rack. *Pattern Recognit. Lett.* **2020**, *140*, 73–80. [CrossRef]
83. Huang, L.; Yang, Y.; Deng, Y.; Yu, Y. DenseBox: Unifying Landmark Localization with End to End Object Detection. *arXiv* **2015**, arXiv:1509.04874.
84. Osokin, A.; Sumin, D.; Lomakin, V. OS2D: One-Stage One-Shot Object Detection by Matching Anchor Features. In Proceedings of the proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.
85. Goldman, E.; Goldberger, J. CRF with deep class embedding for large scale classification. *Comput. Vis. Image Underst.* **2020**, *191*, 102865. [CrossRef]
86. Goldman, E.; Goldberger, J. Large-Scale Classification of Structured Objects using a CRF with Deep Class Embedding. *arXiv* **2017**, arXiv:1705.07420.
87. Wang, Y.; Song, R.; Wei, X.S.; Zhang, L. An adversarial domain adaptation network for cross-domain fine-grained recognition. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Snowmass village, Colorado, USA, 1–5 March 2020; pp. 1228–1236.
88. Santra, B.; Shaw, A.K.; Mukherjee, D.P. Part-based annotation-free fine-grained classification of images of retail products. *Pattern Recognit.* **2022**, *121*, 108257. [CrossRef]

89. Wang, W.; Lee, H.; Livescu, K. Deep Variational Canonical Correlation Analysis. *arXiv* **2016**, arXiv:1610.03454.
90. Ciocca, G.; Napoletano, P.; Locatelli, S.G. Multi-task learning for supervised and unsupervised classification of grocery images. In Proceedings of the International Conference on Pattern Recognition, Milan, Italy, 10–15 January 2020; 2021; pp. 325–338.
91. Dueck, D.; Frey, B.J. Non-metric affinity propagation for unsupervised image categorization. In Proceedings of the 2007 IEEE 11th International Conference on Computer Vision, Rio de Janeiro, Brazil, 17–22 June 2007; pp. 1–8. [CrossRef]
92. Duque Domingo, J.; Medina Aparicio, R.; González Rodrigo, L.M. Improvement of One-Shot-Learning by Integrating a Convolutional Neural Network and an Image Descriptor into a Siamese Neural Network. *Appl. Sci.* **2021**, *11*, 839. [CrossRef]
93. Wang, W.; Cui, Y.; Li, G.; Jiang, C.; Deng, S. A self-attention-based destruction and construction learning fine-grained image classification method for retail product recognition. *Neural Comput. Appl.* **2020**, *32*, 14613–14622. [CrossRef]
94. Karlinsky, L.; Shtok, J.; Tzur, Y.; Tzadok, A. Fine-grained recognition of thousands of object categories with single-example training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 22–29 October 2017; pp. 4113–4122.
95. Gothai, E.; Bhatia, S.; Alabdali, A.; Sharma, D.; Raj, B.; Dadheech, P. Design Features of Grocery Product Recognition Using Deep Learning. *Intell. Autom. Soft Comput.* **2022**, *34*, 1231–1246. [CrossRef]
96. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 22–29 October 2017; pp. 6517–6525. [CrossRef]
97. Tonioni, A.; Di Stefano, L. Domain invariant hierarchical embedding for grocery products recognition. *Comput. Vis. Image Underst.* **2019**, *182*, 81–92. [CrossRef]
98. Sinha, A.; Banerjee, S.; Chattopadhyay, P. An Improved Deep Learning Approach For Product Recognition on Racks in Retail Stores. *arXiv* **2022**, arXiv:2202.13081.
99. George, M.; Mircic, D.; Soros, G.; Floerkemeier, C.; Mattern, F. Fine-grained product class recognition for assisted shopping. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Santiago, Chile, 7–13 December 2015; pp. 154–162.
100. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [CrossRef]
101. Pettersson, T.; Oucheikh, R.; Lofstrom, T. NLP Cross-Domain Recognition of Retail Products. In Proceedings of the 2022 7th International Conference on Machine Learning Technologies (ICMLT), Rome, Italy, 11–13 March 2022; pp. 237–243.
102. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
103. BERT. Available online: https://huggingface.co/docs/transformers/model_doc/bert (accessed on 30 December 2022).
104. Georgieva, P.; Zhang, P. Optical character recognition for autonomous stores. In Proceedings of the 2020 IEEE 10th International Conference on Intelligent Systems (IS), Varna, Bulgaria, 28–30 August 2020; pp. 69–75.
105. Shi, B.; Bai, X.; Belongie, S. Detecting oriented text in natural images by linking segments. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2550–2558.
106. Selvam, P.; Koilraj, J.A.S. A Deep Learning Framework for Grocery Product Detection and Recognition. *Food Anal. Methods* **2022**, *15*, 3498–3522. [CrossRef]
107. Jocher, G. ultralytics/yolov5: V3.1—Bug Fixes and Performance Improvements. 2020. Available online: https://github.com/ultralytics/yolov5 (accessed on 20 February 2023). [CrossRef]
108. Litman, R.; Anschel, O.; Tsiper, S.; Litman, R.; Mazor, S.; Manmatha, R. Scatter: Selective context attentional scene text recognizer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11962–11972.
109. Jiang, Y.; Zhu, X.; Wang, X.; Yang, S.; Li, W.; Wang, H.; Fu, P.; Luo, Z. R2CNN: Rotational Region CNN for Orientation Robust Scene Text Detection. *arXiv* **2017**, arXiv:1706.09579.
110. Ma, J.; Shao, W.; Ye, H.; Wang, L.; Wang, H.; Zheng, Y.; Xue, X. Arbitrary-oriented scene text detection via rotation proposals. *IEEE Trans. Multimed.* **2018**, *20*, 3111–3122. [CrossRef]
111. Liao, M.; Zhu, Z.; Shi, B.; Xia, G.s.; Bai, X. Rotation-sensitive regression for oriented scene text detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5909–5918.
112. Liao, M.; Shi, B.; Bai, X. TextBoxes++: A Single-Shot Oriented Scene Text Detector. *IEEE Trans. Image Process.* **2018**, *27*, 3676–3690. [CrossRef] [PubMed]
113. Zhou, X.; Yao, C.; Wen, H.; Wang, Y.; Zhou, S.; He, W.; Liang, J. East: An efficient and accurate scene text detector. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5551–5560.
114. Tang, J.; Yang, Z.; Wang, Y.; Zheng, Q.; Xu, Y.; Bai, X. SegLink++: Detecting Dense and Arbitrary-shaped Scene Text by Instance-aware Component Grouping. *Pattern Recognit.* **2019**, *96*, 106954. [CrossRef]
115. Wang, Y.; Xie, H.; Zha, Z.; Xing, M.; Fu, Z.; Zhang, Y. ContourNet: Taking a Further Step Toward Accurate Arbitrary-Shaped Scene Text Detection. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA 1–5 March 2020; pp. 11750–11759.
116. Zhang, S.X.; Zhu, X.; Hou, J.B.; Liu, C.; Yang, C.; Wang, H.; Yin, X.C. Deep relational reasoning graph network for arbitrary shape text detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 1–5 March 2020; pp. 9699–9708.

117. Zhu, Y.; Chen, J.; Liang, L.; Kuang, Z.; Jin, L.; Zhang, W. Fourier Contour Embedding for Arbitrary-Shaped Text Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 21–25 June 2021; pp. 3122–3130.
118. He, M.; Liao, M.; Yang, Z.; Zhong, H.; Tang, J.; Cheng, W.; Yao, C.; Wang, Y.; Bai, X. MOST: A multi-oriented scene text detector with localization refinement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8813–8822.
119. Long, S.; Ruan, J.; Zhang, W.; He, X.; Wu, W.; Yao, C. Textsnake: A flexible representation for detecting text of arbitrary shapes. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 20–36.
120. Zhang, C.; Liang, B.; Huang, Z.; En, M.; Han, J.; Ding, E.; Ding, X. Look More Than Once: An Accurate Detector for Text of Arbitrary Shapes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, 16–20 June 2019; pp. 10552–10561.
121. Wang, W.; Xie, E.; Li, X.; Hou, W.; Lu, T.; Yu, G.; Shao, S. Shape robust text detection with progressive scale expansion network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9336–9345.
122. Baek, Y.; Lee, B.; Han, D.; Yun, S.; Lee, H. Character Region Awareness for Text Detection. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 9357–9366.
123. Wang, W.; Xie, E.; Song, X.; Zang, Y.; Wang, W.; Lu, T.; Yu, G.; Shen, C. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8440–8449.
124. Liao, M.; Wan, Z.; Yao, C.; Chen, K.; Bai, X. Real-time scene text detection with differentiable binarization. In Proceedings of the AAAI conference on artificial intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11474–11481.
125. Liao, M.; Zou, Z.; Wan, Z.; Yao, C.; Bai, X. Real-Time Scene Text Detection With Differentiable Binarization and Adaptive Scale Fusion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 919–931. [CrossRef]
126. Naiemi, F.; Ghods, V.; Khalesi, H. MOSTL: An Accurate Multi-Oriented Scene Text Localization. *Circuits, Syst. Signal Process.* **2021**, *40*, 4452–4473. [CrossRef]
127. Li, J.; Lin, Y.; Liu, R.; Ho, C.M.; Shi, H. RSCA: Real-time Segmentation-based Context-Aware Scene Text Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2349–2358.
128. Wang, Z.; Silamu, W.; Li, Y.; Xu, M. A Robust Method: Arbitrary Shape Text Detection Combining Semantic and Position Information. *Sensors* **2022**, *22*, 9982. [CrossRef] [PubMed]
129. Wang, P.; Zhang, C.; Qi, F.; Huang, Z.; En, M.; Han, J.; Liu, J.; Ding, E.; Shi, G. A Single-Shot Arbitrarily-Shaped Text Detector based on Context Attended Multi-Task Learning. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France 21–25 October 2019.
130. Long, S.; Qin, S.; Panteleev, D.; Bissacco, A.; Fujii, Y.; Raptis, M. Towards End-to-End Unified Scene Text Detection and Layout Analysis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 1049–1059.
131. Shi, B.; Yang, M.; Wang, X.; Lyu, P.; Yao, C.; Bai, X. Aster: An attentional scene text recognizer with flexible rectification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 2035–2048. [CrossRef]
132. Shi, B.; Wang, X.; Lyu, P.; Yao, C.; Bai, X. Robust scene text recognition with automatic rectification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4168–4176.
133. Yang, M.; Guan, Y.; Liao, M.; He, X.; Bian, K.; Bai, S.; Yao, C.; Bai, X. Symmetry-constrained rectification network for scene text recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–November 2019; pp. 9147–9156.
134. Liao, M.; Zhang, J.; Wan, Z.; Xie, F.; Liang, J.; Lyu, P.; Yao, C.; Bai, X. Scene Text Recognition from Two-Dimensional Perspective. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 8714–8721. [CrossRef]
135. Long, S.; Guan, Y.; Bian, K.; Yao, C. A new perspective for flexible feature gathering in scene text recognition via character anchor pooling. In Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Virtual Barcelona, 4–8 May 2020; pp. 2458–2462.
136. Li, H.; Wang, P.; Shen, C.; Zhang, G. Show, attend and read: A simple and strong baseline for irregular text recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 28–29 January 2019; Volume 33, pp. 8610–8617.
137. Wan, Z.; He, M.; Chen, H.; Bai, X.; Yao, C. TextScanner: Reading Characters in Order for Robust Scene Text Recognition. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 28–29 January 2019.
138. Yu, D.; Li, X.; Zhang, C.; Liu, T.; Han, J.; Liu, J.; Ding, E. Towards accurate scene text recognition with semantic reasoning networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 12113–12122.
139. Fu, Z.; Xie, H.; Jin, G.; Guo, J. *Look Back Again: Dual Parallel Attention Network for Accurate and Robust Scene Text Recognition, Proceedings of the 2021 International Conference on Multimedia Retrieval (ICMR '21), Taipei, Taiwan, 21–24 August 2021*; Association for Computing Machinery: New York, NY, USA, 2021; pp. 638–644. [CrossRef]
140. Zheng, T.; Chen, Z.; Fang, S.; Xie, H.; Jiang, Y.G. CDistNet: Perceiving Multi-Domain Character Distance for Robust Text Recognition. *arXiv* **2021**, arXiv:2111.11011.

141. Na, B.; Kim, Y.; Park, S. Multi-modal text recognition networks: Interactive enhancements between visual and semantic features. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–24 October 2022; pp. 446–463.
142. He, Y.; Chen, C.; Zhang, J.; Liu, J.; He, F.; Wang, C.; Du, B. Visual Semantics Allow for Textual Reasoning Better in Scene Text Recognition. *arXiv* **2021**, arXiv:2112.12916.
143. Bautista, D.; Atienza, R. Scene Text Recognition with Permuted Autoregressive Sequence Models. In Proceedings of the Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part XXVIII; Springer: Berlin/Heidelberg, Germany, 2022; pp. 178–196. [CrossRef]
144. Cai, H.; Sun, J.; Xiong, Y. Revisiting Classification Perspective on Scene Text Recognition. *arXiv* **2021**, arXiv:2102.10884.
145. Liu, X.; Liang, D.; Yan, S.; Chen, D.; Qiao, Y.; Yan, J. Fots: Fast oriented text spotting with a unified network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18-23 June 2018; pp. 5676–5685.
146. Naiemi, F.; Ghods, V.; Khalesi, H. A novel pipeline framework for multi oriented scene text image detection and recognition. *Expert Syst. Appl.* **2021**, *170*, 114549. [CrossRef]
147. Feng, W.; He, W.; Yin, F.; Zhang, X.Y.; Liu, C.L. TextDragon: An End-to-End Framework for Arbitrary Shaped Text Spotting. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9075–9084.
148. Liu, Y.; Chen, H.; Shen, C.; He, T.; Jin, L.; Wang, L. Abcnet: Real-time scene text spotting with adaptive bezier-curve network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 1–5 March 2020; pp. 9809–9818.
149. Liu, Y.; Shen, C.; Jin, L.; He, T.; Chen, P.; Liu, C.; Chen, H. ABCNet v2: Adaptive Bezier-Curve Network for Real-Time End-to-End Text Spotting. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 8048–8064. [CrossRef]
150. Zhong, H.; Tang, J.; Wang, W.; Yang, Z.; Yao, C.; Lu, T. ARTS: Eliminating Inconsistency between Text Detection and Recognition with Auto-Rectification Text Spotter. *arXiv* **2021**, arXiv:2110.10405.
151. Lyu, P.; Liao, M.; Yao, C.; Wu, W.; Bai, X. Mask TextSpotter: An End-to-End Trainable Neural Network for Spotting Text with Arbitrary Shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *43*, 532–548.
152. Liao, M.; Pang, G.; Huang, J.; Hassner, T.; Bai, X. Mask textspotter v3: Segmentation proposal network for robust scene text spotting. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 706–722.
153. Qin, S.; Bissacco, A.; Raptis, M.; Fujii, Y.; Xiao, Y. Towards Unconstrained End-to-End Text Spotting. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 4703–4713.
154. Qiao, L.; Chen, Y.; Cheng, Z.; Xu, Y.; Niu, Y.; Pu, S.; Wu, F. MANGO: A Mask Attention Guided One-Stage Scene Text Spotter. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA 7–12 February 2020.
155. Zhang, X.; Su, Y.; Tripathi, S.; Tu, Z. Text Spotting Transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 9519–9528.
156. Wang, W.; Liu, X.; Ji, X.; Xie, E.; Liang, D.; Yang, Z.; Lu, T.; Shen, C.; Luo, P. Ae textspotter: Learning visual and linguistic representation for ambiguous text spotting. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 457–473.
157. Huang, M.; Liu, Y.; Peng, Z.; Liu, C.; Lin, D.; Zhu, S.; Yuan, N.J.; Ding, K.; Jin, L. SwinTextSpotter: Scene Text Spotting via Better Synergy between Text Detection and Text Recognition. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 4583–4593.
158. Hao, Y.; Fu, Y.; Jiang, Y.G.; Tian, Q. An End-to-End Architecture for Class-Incremental Object Detection with Knowledge Distillation. In Proceedings of the 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 8–12 July 2019; pp. 1–6.
159. Yang, D.; Zhou, Y.; Zhang, A.; Sun, X.; Wu, D.; Wang, W.; Ye, Q. Multi-view correlation distillation for incremental object detection. *Pattern Recognit.* **2022**, *131*, 108863. [CrossRef]
160. Zhang, L.; Du, D.; Li, C.; Wu, Y.; Luo, T. Iterative Knowledge Distillation for Automatic Check-Out. *IEEE Trans. Multimed.* **2021**, *23*, 4158–4170. [CrossRef]
161. Capozzi, L.; Barbosa, V.; Pinto, C.; Pinto, J.R.; Pereira, A.; Carvalho, P.M.; Cardoso, J.S. Toward Vehicle Occupant-Invariant Models for Activity Characterization. *IEEE Access* **2022**, *10*, 104215–104225. [CrossRef]
162. Bhojanapalli, S.; Chakrabarti, A.; Glasner, D.; Li, D.; Unterthiner, T.; Veit, A. Understanding Robustness of Transformers for Image Classification. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 10211–10221.
163. Long, S.; Yao, C. UnrealText: Synthesizing Realistic Scene Text Images from the Unreal World. *arXiv* **2020**, arXiv:2003.10608.
164. Luo, C.; Lin, Q.; Liu, Y.; Jin, L.; Shen, C. Separating Content from Style Using Adversarial Learning for Recognizing Text in the Wild. *Int. J. Comput. Vis.* **2020**, *129*, 960–976. [CrossRef]

165. Coates, A.; Carpenter, B.; Case, C.; Satheesh, S.; Suresh, B.; Wang, T.; Wu, D.J.; Ng, A.Y. Text detection and character recognition in scene images with unsupervised feature learning. In Proceedings of the 2011 International Conference on Document Analysis and Recognition, Beijing, China, 18–21 September 2011; pp. 440–445.
166. Gupta, A.; Vedaldi, A.; Zisserman, A. Learning to Read by Spelling: Towards Unsupervised Text Recognition. In Proceedings of the 11th Indian Conference on Computer Vision, Graphics and Image Processing, Hyderabad, India, 18–22 December 2018.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.