



OPEN

DATA DESCRIPTOR

# Packed Fruits and Vegetables Visual Classification and Segmentation Benchmark

Svetlana Illarionova<sup>1,5</sup>✉, Sergey Nesteruk<sup>2,5</sup>, Tatiana Elina<sup>3</sup>, Sergey Bezzateev<sup>3</sup> & Evgeny Burnaev<sup>1,4</sup>

We introduce a novel dataset for visual recognition systems in retail automation, focusing specifically on fruits and vegetables. The dataset comprises 34 species and 65 varieties, featuring fairly balanced classes and including packed goods in plastic bags. We capture each sample from multiple viewpoints and provide additional annotations, such as object count and total weight. Furthermore, a subset of samples for each class includes segmentation masks. This dataset aims to overcome the limitations of current open-access datasets by providing a more comprehensive and diverse set of training data. A total of 72 annotators collected over 100,000 images of 370,000 objects across multiple shops and cities. Around 9,000 images have manual segmentation masks. To facilitate research in this area, we provide baseline results for zero-shot and supervised classification, instance segmentation, and object counting tasks. We also investigate the impact of packaging and background type on model performance. Ultimately, this dataset is designed to support the development of multitask models for visual recognition in offline retail settings.

## Background & Summary

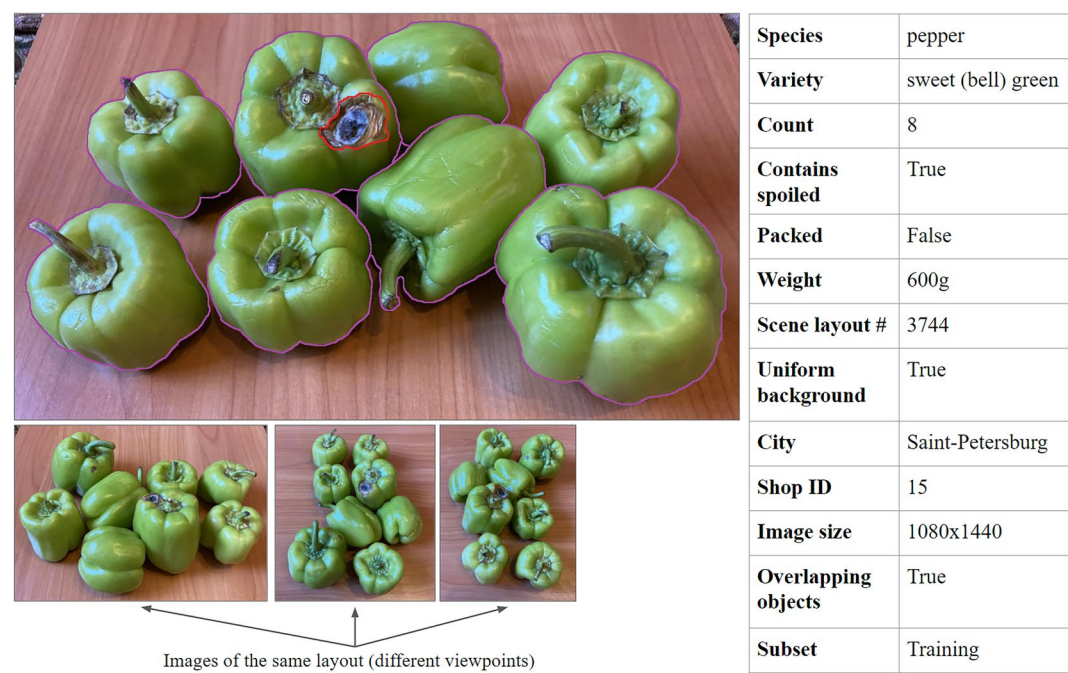
The development of accurate and robust visual recognition systems for retail automation has gained significant attention in recent years<sup>1</sup>. However, the lack of diverse and comprehensive datasets specifically tailored for this domain presents a challenge<sup>2</sup>. Existing datasets often have limitations such as a narrow range of classes, insufficient annotations, or a lack of real-world variability, hindering the advancement of research in this area. To address these gaps, we introduce PackEat, a novel dataset designed for visual recognition in retail automation, with a particular focus on fruits and vegetables.

In retail settings, the ability to accurately identify and classify fruits and vegetables is crucial for various applications, including inventory management<sup>3</sup>, quality control<sup>4,5</sup>, and customer experience enhancement<sup>6,7</sup>. However, this task poses several challenges due to the high intra-class variability and similarity between different species and varieties<sup>8</sup>. Furthermore, the presence of packed goods in plastic bags or other packaging adds another layer of complexity, as it alters the visual appearance of the products<sup>9</sup>.

Current open-access datasets often fall short in addressing these challenges. They may lack sufficient diversity in terms of classes, neglect packed goods, or fail to provide comprehensive annotations<sup>10</sup>. This limits the development and evaluation of robust models that can generalize well to different retail environments. To overcome these limitations, we designed PackEat as a comprehensive and diverse dataset, aiming to support the creation of adaptable and accurate visual recognition systems for retail automation.

In this work, we introduce PackEat, a novel dataset designed to advance visual recognition in retail automation, specifically for fruits and vegetables. PackEat addresses the limitations of current datasets by offering a diverse range of annotated images, including packed goods. The dataset was collected by 72 annotators across multiple cities, resulting in over 100,000 images with various annotations. It encompasses 34 species and 65 varieties of fruits and vegetables, including packed goods in plastic bags (Fig. 1). Each sample in the dataset is captured from multiple viewpoints, providing a comprehensive view of the objects.

<sup>1</sup>Skolkovo Institute of Science and Technology, Bolshoy Boulevard 30, bld. 1, 121205, Moscow, Russia. <sup>2</sup>Yandex, Lev Tolstoy Street, bld. 16, 119021, Moscow, Russia. <sup>3</sup>Saint Petersburg State University of Aerospace Instrumentation (SUAI), 67 Bolshaya Morskaya street, 190000, Saint-Petersburg, Russia. <sup>4</sup>Autonomous Non-Profit Organization Artificial Intelligence Research Institute (AIRI), 105064, Moscow, Russia. <sup>5</sup>These authors contributed equally: Svetlana Illarionova, Sergey Nesteruk. ✉e-mail: [s.illarionova@skoltech.ru](mailto:s.illarionova@skoltech.ru)



**Fig. 1** Examples of images from the dataset with the corresponding annotations.

One of the key features of PackEat is its hierarchical class structure, which includes both species and varieties. This structure allows for a more nuanced evaluation of model performance and enables the investigation of fine-grained recognition tasks. Furthermore, the dataset includes pixel-level annotations for instance segmentation, providing detailed information beyond object-level labels. The hierarchical class structure and pixel-level annotations make PackEat a comprehensive resource for research and development.

To establish a benchmark and facilitate future research, we provide baseline results for various tasks, including zero-shot and supervised classification, instance segmentation, and object counting. Experiments revealed that while zero-shot methods offer reasonable accuracy, supervised learning with the ConvNext model achieved the best performance. We also analyze the impact of factors such as packaging and background complexity on model performance. By sharing our dataset and baseline results, we aim to foster advancements in computer vision and retail automation, providing a valuable resource for the research community.

Additionally, we explored the impact of packaging and background type on model performance. Our findings indicate that packed objects and complex backgrounds present greater challenges for accurate recognition. This insight underscores the importance of considering these factors in model design and evaluation, particularly when dealing with real-world retail scenarios.

The experimental results indicate that the recognition of packed fruits and vegetables is a feasible task for supervised learning pipelines. However, there remains a significant performance gap for general-domain models, even when dealing with common classes. The introduction of packaging complicates the task, leading to a notable performance decline for zero-shot methods. On the other hand, supervised tuning effectively addresses this challenge, making the proposed dataset a valuable benchmark for evaluating and improving foundation visual-language models (VLMs). The dataset's inclusion of packaged items serves as a realistic representation of real-world scenarios, providing a more comprehensive assessment of model capabilities beyond idealized conditions.

The contributions of this work can be summarized as follows:

- We introduce PackEat, a large-scale dataset specifically designed for visual recognition in retail automation, with a focus on fruits and vegetables, including packed goods.
- We provide comprehensive baseline results for various tasks, including zero-shot and supervised classification, instance segmentation, and object counting, establishing a benchmark for future comparisons.
- We conduct extensive experiments to analyze the impact of factors such as packaging and background complexity on model performance, offering insights into the challenges and opportunities presented by the dataset.

In the following sections, we provide a detailed description of the PackEat dataset, present our experimental setup and baseline results, and discuss the implications and potential future research directions enabled by this dataset. Supplementary materials, including illustrations, dataset statistics can be found in ‘Retail dataset overview’ Section. Details about the tested models are in ‘Model details’ Section. Auxiliary results can be found in ‘Ablation results’ Section. The dataset is available and can be accessed and downloaded from<sup>11</sup>.

## Related Work

The development of visual recognition systems for retail automation has garnered significant research interest, leading to the creation of various datasets aimed at addressing the challenges in this domain. Numerous visual retail datasets have been proposed, focusing on the recognition of packed goods on shelves or in vending machines. This problem is relatively straightforward due to the consistent packaging of the same product, with brands maintaining brand recognition and changing designs infrequently. Moreover, the presence of barcodes simplifies recognition tasks<sup>12</sup>. However, the recognition of unpacked, weighted groceries, such as fruits and vegetables, presents a more complex challenge due to high intra-class variability. Distinguishing between varieties of the same species further complicates the task due to high inter-class similarity. The additional challenge of objects being packed in plastic bags exacerbates the difficulty of accurate recognition.

**Retail Datasets.** *Grocery Datasets.* Retail automation often entails auxiliary tasks beyond classification, such as object counting and detecting spoilage or rotten items<sup>4</sup>. While there exist datasets containing agricultural images of fruits, vegetables, and other edible plants in their natural environment, their direct applicability to retail automation is limited<sup>13</sup>. These datasets are typically designed for tasks like harvesting automation or treatment recommendations and often focus on a single species. Meta-datasets, which aggregate multiple existing datasets, offer a rich source of classes but may introduce biases if the data collection methods differ significantly. The VegFru dataset<sup>14</sup> is a substantial collection of images, encompassing 292 varieties of fruits and vegetables with a total of 160,000 images. However, it is important to note that the dataset was curated by parsing images from the Internet, and as a result, many of the images may not accurately represent the retail settings or the specific context in which the products are typically displayed and sold. This discrepancy could introduce biases or challenges when utilizing the dataset for retail automation tasks. Ensuring a representative and unbiased test set is crucial to prevent models from overfitting to background variations rather than learning to recognize target objects effectively. It has been demonstrated that when training on external datasets or parsed data, combining them with in-domain data can yield beneficial results<sup>15</sup>. In the referenced paper<sup>5</sup>, the authors introduce a dataset specifically focused on rotten fruits, providing a comprehensive description for each image. However, it is important to note that the dataset is relatively small. While there exist large datasets for food recognition<sup>16–18</sup>, the scope of this paper focuses specifically on the domain of fresh, unprepared fruits and vegetables in retail settings.

Several datasets have been introduced specifically for retail product recognition, but many of them are relatively small or created in artificial conditions with uniform backgrounds, limiting their applicability in real-world settings. Background substitution<sup>19</sup> and image synthesis<sup>20,21</sup> techniques can enhance their robustness<sup>2</sup>. The Fruits-360 dataset<sup>22</sup> is a substantial collection of fruit images designed for classification tasks. However, it is important to note that the images in this dataset have been processed to remove their backgrounds and are downsampled to a resolution of  $100 \times 100$  pixels. This preprocessing may limit the applicability of the dataset for tasks that require higher-resolution images or contextually relevant background information.

*Plastic Bags.* There is a scarcity of datasets with packed goods, and existing ones often provide only image-level annotations or localize the entire bag rather than individual objects within it. The dataset in<sup>23</sup> was gathered by the authors, featuring apples covered in separate plastic bags within a fruit orchard that differs greatly from retail environment. The dataset entitled “Fruits & Vegetable Detection for YOLOv4”<sup>29</sup> comprises goods in plastic bags and is better suited for groceries detection in retail stores. However, the provided annotation for the packed groceries does not include markup for individual objects. In<sup>24</sup>, the authors collected and annotated 1067 images including samples with groceries in plastic bags. However, the dataset has limited variety of backgrounds that makes it less appropriate for real life applications. Another example of a retail dataset with groceries in plastic bags is shown in<sup>25</sup>. However, the annotations in this dataset assume only an image classification task, and the collected samples also feature repetitive backgrounds. Some datasets artificially inflate their size by including slightly rotated or shifted versions of the same object, which adds limited value and may lead to overfitting<sup>22,26,27</sup>.

**Deep Learning for Retail.** Computer vision models have become indispensable for retail automation, particularly in the task of grocery recognition. While classic computer vision approaches relied on handcrafted features like SIFT<sup>28</sup> and HOG<sup>29</sup> for object detection and recognition<sup>30</sup>, deep learning has brought about a paradigm shift. Convolutional Neural Networks (CNNs), with their ability to automatically learn hierarchical representations, have been widely applied to grocery recognition. Transformers, known for their attention mechanisms, have also been explored for this task, as they excel in understanding global context. The evolution of computer vision models has enabled a range of tasks in retail automation, including classification (identifying grocery items), object detection (locating products on shelves), and instance segmentation (distinguishing individual items in a cluster). Generic classification algorithms, such as those based on CNN architectures like VGG<sup>31</sup>, ResNet<sup>32</sup>, Inception<sup>33</sup>, and ConvNext<sup>34</sup>, have been fine-tuned for grocery recognition, achieving accurate results. Metric learning<sup>35</sup>, which focuses on learning similarity between images, has also been employed, especially useful for identifying similar products or different packaging of the same item<sup>36,37</sup> and for product tracking<sup>38</sup>. Contrastive pretraining<sup>39</sup> enhances representation learning and is beneficial for recognizing products with subtle differences<sup>40</sup>. Additionally, VLMs, exemplified by LLaVA<sup>41</sup>, open up possibilities for tasks like grocery list interpretation and natural language-based product search.

Species	Total	Training	Testing	Packed	Non-packed	With uniform background	With complex background
apple	11573	9029	2544	5465	6108	4818	6755
apricot	2249	1883	366	1091	1158	872	1377
avocado	3413	2746	667	1646	1767	1410	2003
banana	3178	2426	752	1552	1626	1269	1909
beet	2227	1687	540	1091	1136	679	1548
cabbage	6388	4900	1488	2881	3507	2551	3837
carrot	2262	1714	548	1106	1156	1071	1191
corn	1595	1215	380	780	815	900	695
cucumber	3351	2577	774	1623	1728	1593	1758
daikon	979	744	235	479	500	531	448
garlic	1255	1007	248	586	669	603	652
grape	5150	3991	1159	2782	2368	1874	3276
grapefruit	405	405	0	213	192	393	12
kiwi	1942	1574	368	946	996	846	1096
lemon	2943	2479	464	1380	1563	1524	1419
lime	1499	1164	335	716	783	511	988
mango	851	659	192	421	430	464	387
melon	1005	877	128	240	765	498	507
onion	3081	2453	628	1490	1591	1280	1801
orange	2268	1895	373	998	1270	1104	1164
nectarine	4036	3264	772	38	3998	48	3988
pomelo	1988	1636	352	0	1988	0	1988
pear	2666	2170	496	380	2286	274	2392
pepper	8805	6922	1883	4305	4500	3411	5394
plum	2827	2444	383	1335	1492	900	1927
pomegranate	1349	1037	312	667	682	889	460
potato	1994	1522	472	971	1023	1046	948
pumpkin	984	748	236	240	744	0	984
radish	2275	1744	531	1074	1201	1165	1110
salad	2151	1635	516	1555	596	472	1679
tangerine	654	498	156	320	334	480	174
tomato	11879	9208	2671	5796	6083	4462	7417
watermelon	987	763	224	0	987	115	872
zucchini	3203	2473	730	1552	1651	1330	1873
Total	103412	81489	21923	45719	57693	39383	64029

**Table 1.** The number of images per species.

## Methods

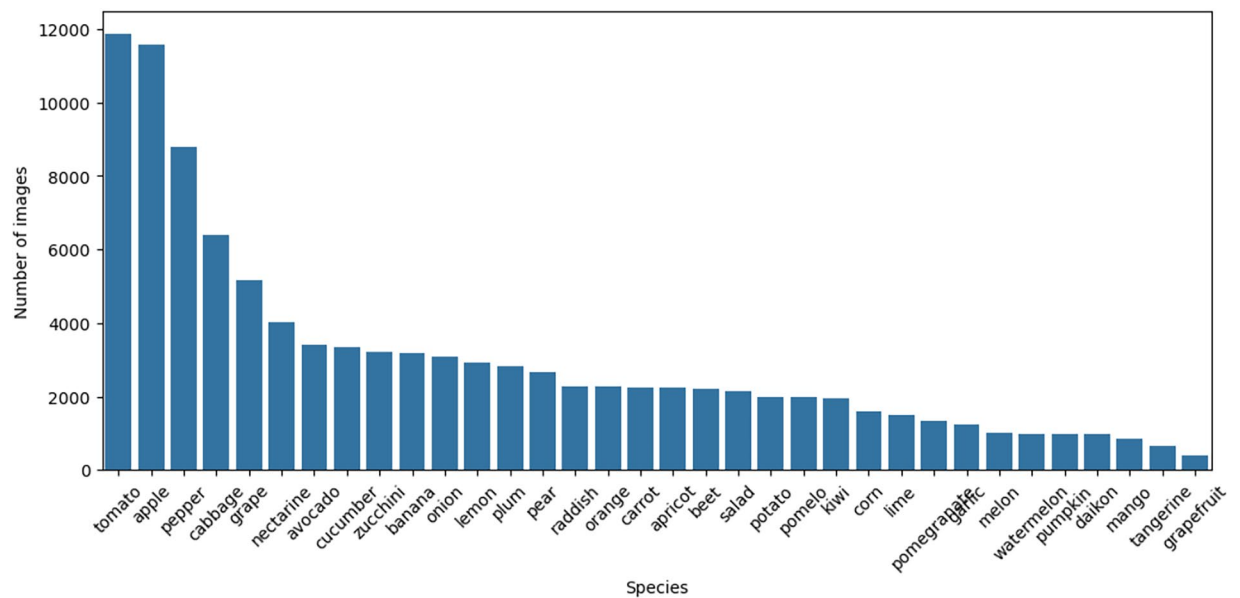
The primary objective of developing the PackEat benchmark was to provide a platform for evaluating the adaptability of models in retail automation scenarios. The key consideration was to create a fair and unbiased test subset. To achieve this, we employed a stratified sampling approach, where not only different samples were used for training and testing, but also samples collected by distinct annotators in different shops or cities were split to minimize bias. Furthermore, we ensured that the sets for different classes were balanced whenever possible.

**Data Acquisition.** For data acquisition, we engaged 72 annotators across 16 cities, and collected fruits and vegetables from 132 different shops (Table 1).

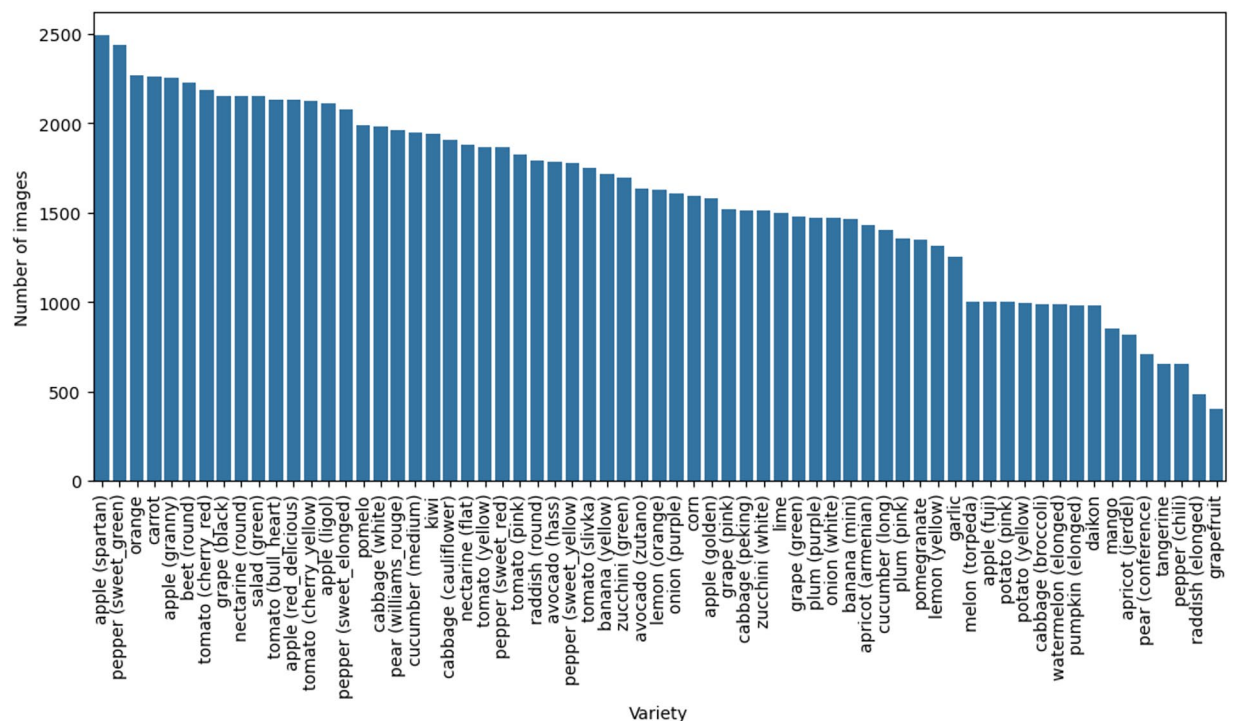
Initially, we selected 34 common fruit and vegetable species (Fig. 2). Subsequently, we determined their available varieties in mass markets, resulting in a total of 65 varieties (Fig. 3). Consequently, the target classes exhibit a hierarchical structure (Fig. 4). This approach allows us to control the difficulty of the recognition problem, as distinguishing between different species is generally easier than distinguishing between varieties of the same species.

The image collection process was designed to meet the following criteria:

1. Uniform shooting process.
2. Diverse shooting settings.
3. Recording available auxiliary information.
4. Balancing of the dataset with respect to varieties and auxiliary features, whenever possible.



**Fig. 2** Species count distribution.

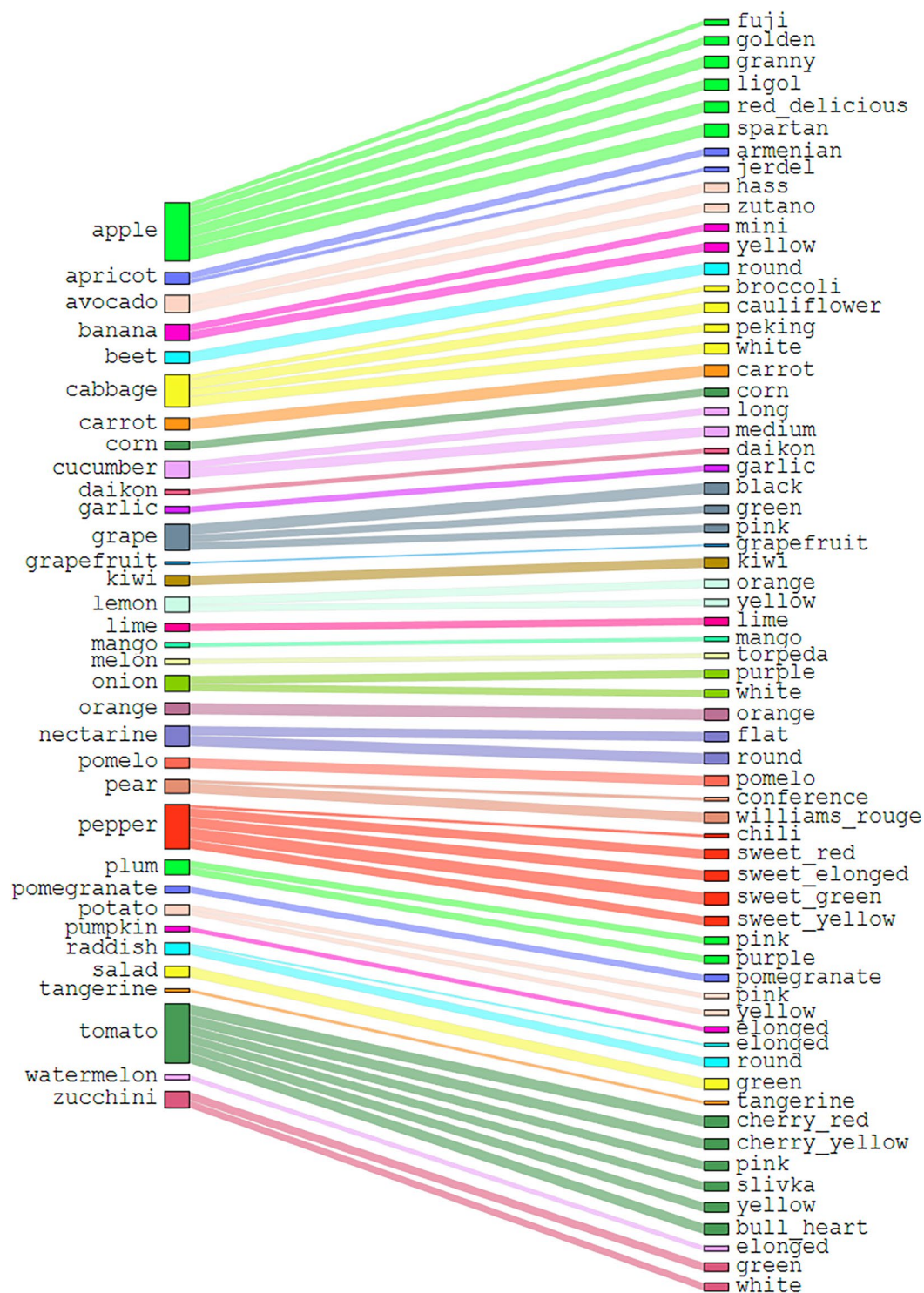


**Fig. 3** Varieties count distribution.

We collected images in portions, where each portion consists of a specified number of layouts for a single variety. A layout is defined as a fixed placement of objects, and each layout is imaged from four different view-points. To prevent data leakage during the splitting of training and testing subsets, we ensured that each annotator was assigned data portions from different varieties.

To maintain a uniform shooting process, we provided annotators with instructions to collect a fixed number of layouts with varying numbers of objects within each data portion. Specifically, we requested 15 layouts with four images each for the following scenarios: one object, two to three objects, four to six objects, and seven or more objects. Half of the layouts were required to feature packed objects, where possible. Exceptions included watermelons, which did not fit available plastic bags, and salad, which is typically sold in packaging. Additionally, every image portion included 12 top-view images with a filled box of the selected variety, Figure 5. Top-view images featuring single varieties of fruits or vegetables without background are used as supplementary





**Fig. 4** The relations between species and their varieties.

data for model development. Researchers may apply advanced augmentation techniques to generate additional training data by automatically delineating target object boundaries and copy-pasting them onto new backgrounds<sup>19</sup>. Since these images lack annotations, they are excluded from the testing subset of the dataset. However, they can be utilized in additional experiments to enhance training sample diversity by artificially generated samples of the same fruits or vegetables varieties. We instructed annotators to utilize a varied range of backgrounds, encompassing both shop and home interiors, in order to introduce environmental diversity into the image dataset. However, we exempted top-view images from this requirement, as they were captured at close range, thereby minimizing the presence of background elements.



**Fig. 5** Examples of top-view crowded images with single variety.

We incorporate images with varying orientations, specifically both landscape (horizontal) and portrait (vertical) formats, to ensure robustness and invariance of our model to image rotation.

**Data Annotation.** During the imaging process, annotators recorded the following metadata: object species and variety, layout ID, total weight of all objects, object count, and information about spoilage and packaging. Subsequently, each image was supplemented with additional metadata, including city, shop, image size, overlapping, and background type. Table 2 presents a comprehensive overview of the distribution of auxiliary annotations across different species. It includes valuable information such as the number of unique layouts, approximate number of objects, images with crowded scenes, cities represented, shops, and the number of annotators involved. It is important to note that the object count is approximate due to the potential reuse of the same objects across multiple layouts, which introduces a level of uncertainty in the exact count. This table provides insights into the richness and variability of the dataset, highlighting the diverse conditions under which the annotations were collected.

The overlapping flag enables the separation of images with numerous occlusions from those with evenly spread objects. The background type is categorized as either simple or complex, which, although a subjective measure, is shown to impact the performance of certain models in later experiments.

All images underwent a quality control process, where three assessors reviewed the species, varieties, and image quality, and one assessor evaluated the remaining criteria. Following this quality and duplicates check, some images were removed, resulting in a slightly smaller dataset than initially intended.

Following the formation of the classification dataset, we performed manual instance segmentation annotation on a carefully curated subset of images from each variety, selecting samples that exhibited the highest degree of diversity. On average, 135 images per variety were annotated, resulting in a total of approximately 9,000 annotated images.

In addition to the annotated images, we provide a comprehensive taxonomy for each variety, serving as supplementary information to facilitate further analysis and exploration. The taxonomy file serves as a valuable resource for domain specialists, enabling more precise identification of different fruit and vegetable varieties. Unlike common datasets that typically list only generic product names (e.g., “apple” or “pepper”), this detailed classification is essential for practical applications where distinguishing between specific varieties is crucial. Furthermore, the comprehensive descriptions can support agronomists in downstream tasks such as plant growth analysis and maintaining optimal growing conditions based on precise phenotypic.

**Relationship between species and varieties.** Figure 4 illustrates the relationships between species (depicted on the left) and their corresponding varieties (on the right). In cases where a species has only one variety, the variety name is identical to the species name. It is evident from the figure that the number of varieties varies across different species. The dataset predominantly includes apples and tomatoes, followed by peppers and cabbages. Consequently, there is a certain degree of imbalance in the sample numbers with respect to species representation (Fig. 2), although the number of samples per variety remains relatively balanced (Fig. 3).

Species	Number of layouts	Number of objects	Number of crowded images	Number of cities	Number of shops	Number of annotators
apple	2568	45676	1487	6	25	21
apricot	372	8530	80	2	5	4
avocado	674	14051	212	3	12	7
banana	760	12144	360	3	5	6
beet	546	8203	13	1	5	4
cabbage	1496	23219	460	4	17	15
carrot	550	8526	344	3	5	5
corn	384	5964	135	1	4	4
cucumber	782	12709	737	3	5	7
daikon	240	3568	176	2	6	2
garlic	254	5968	379	2	2	2
grape	1171	18421	813	3	9	10
grapefruit	101	1020	299	1	1	1
kiwi	371	10308	382	3	4	3
lemon	473	11272	239	3	12	7
lime	338	7586	75	2	5	3
mango	196	2499	183	2	7	2
melon	129	2894	81	1	4	2
onion	633	10900	241	2	6	6
orange	377	7460	226	2	4	4
nectarine	779	292	467	1	10	9
pomelo	356	356	136	2	4	4
pear	504	3234	380	4	7	7
pepper	1899	38036	1169	3	16	14
plum	393	10885	113	2	6	6
pomegranate	316	4787	67	1	3	2
potato	485	7384	92	2	5	4
pumpkin	240	4048	12	1	2	2
radish	541	9632	108	1	6	5
salad	523	6950	391	1	11	4
tangerine	161	2556	58	1	4	1
tomato	2690	48088	1791	6	23	23
watermelon	231	3188	143	3	3	3
zucchini	741	11528	350	3	8	6
Total	22174	371526	12199	16	100+	72

**Table 2.** The information about image annotations per species.

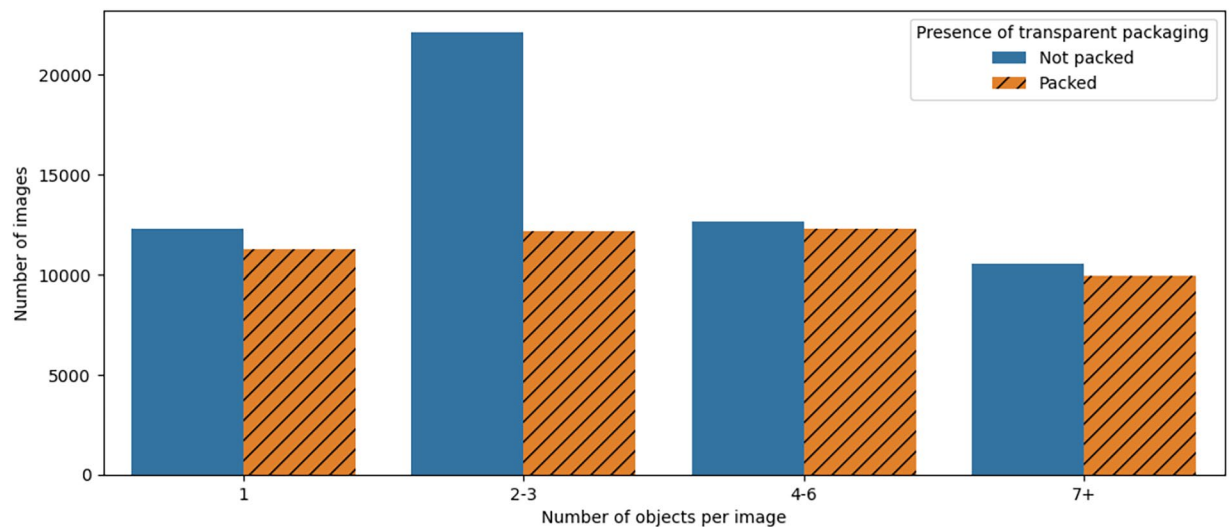
Table 1 presents detailed information regarding the image data for each species. The table includes the total number of images, as well as the numbers allocated for training and testing splits. Additionally, it provides insights into the number of samples packaged in plastic bags and those that are unpackaged. It is important to note that for certain species, obtaining images in bags proved challenging, resulting in a smaller number or absence of such samples. Our policy is to avoid splitting samples from a single portion into training and testing sets to prevent data leakage. As a supplementary feature, the table also includes information on the number of images with uniform and complex backgrounds, as subjectively assessed by the annotators.

**Plants Taxonomy.** Taxonomic hierarchies are often inconsistent and noisy across sources<sup>42,43</sup>, leading to challenges in reproducibility<sup>44</sup>. To address this, we provide common and scientific names for all varieties in our dataset, utilizing the basic taxonomic ranks: kingdom, phylum, class, order, family, genus, and species, with an additional “subfamily” taxon for specific plant categories. For species with numerous variants, such as bananas or those with hybrid parentage, we assign the genus as the most reliable taxon from the Catalogue of Life (CoL)<sup>45</sup>.

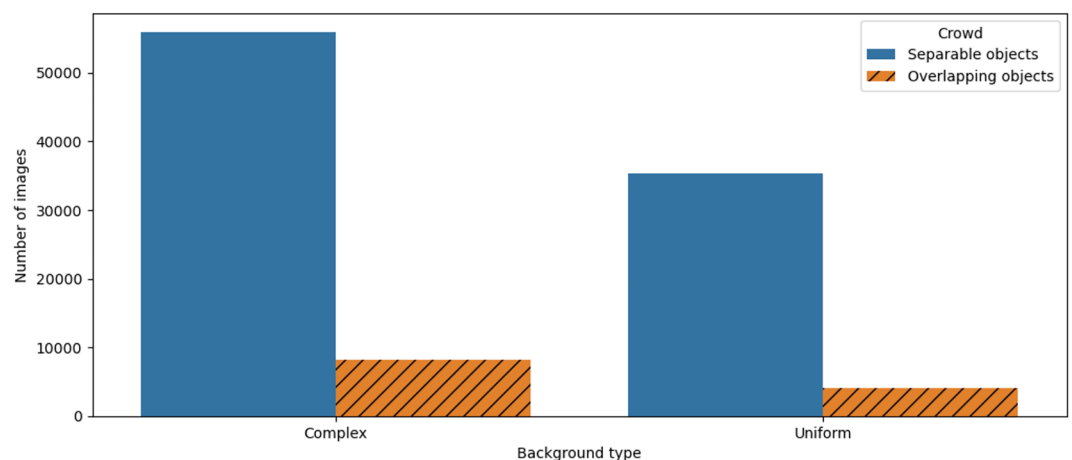
**Data Analysis.** *Dataset Statistics.* As illustrated in Figure 6, our dataset exhibits a balanced distribution of object counts and packaging presence across the images. Furthermore, we provide object counts for each image, enabling researchers to select the most relevant subsets for their specific requirements. The explicit annotation of packaging type, background type, and object count enhances the flexibility of the dataset, allowing for tailored selections.

Furthermore, as depicted in Figure 7, the dataset demonstrates a balanced representation of images with complex backgrounds and those with uniform or simple backgrounds. The variety of background types provides a diverse set of scenarios, mimicking real-world conditions and facilitating a comprehensive evaluation of object





**Fig. 6** The number of images with different object counts and different packaging.



**Fig. 7** The number of images with different complexity.

detection algorithms. It is noteworthy that the background type classification is subjective, as it was determined by the annotators.

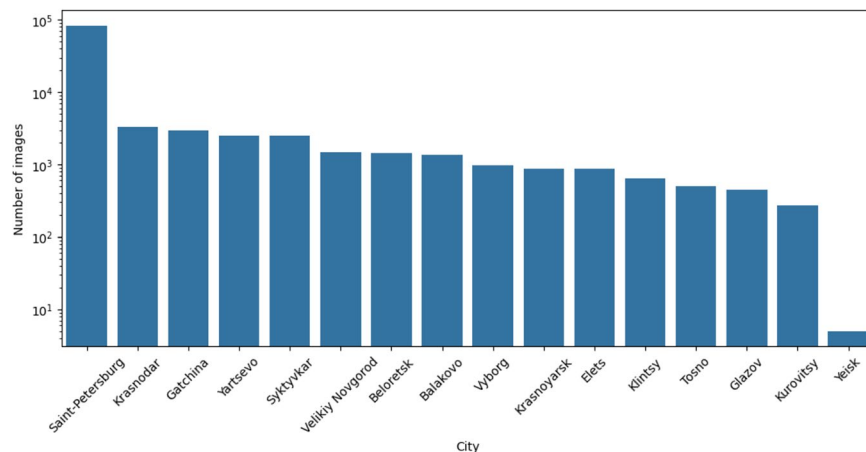
Similarly, the measure of image crowdedness is also subjective, with images considered to have overlapping objects if more than 30% of the area of some target objects occlude each other, as per the annotator instructions. The dataset is intentionally designed to have a limited number of examples with overlapping objects.

Figure 8 reveals insights into the geographical distribution of sample collection. The majority of samples were collected in Saint Petersburg, as indicated by the logarithmic scale. However, it is noteworthy that contributions from other cities are evenly distributed, ensuring a diverse representation of different environments and lighting conditions. This aspect enhances the generalization capabilities of models trained on the dataset, making them more adaptable to varying real-world scenarios. The dataset's diversity in terms of object counts, packaging, background complexity, and geographical distribution makes it a valuable resource for training and evaluating robust object detection models with broad applicability.

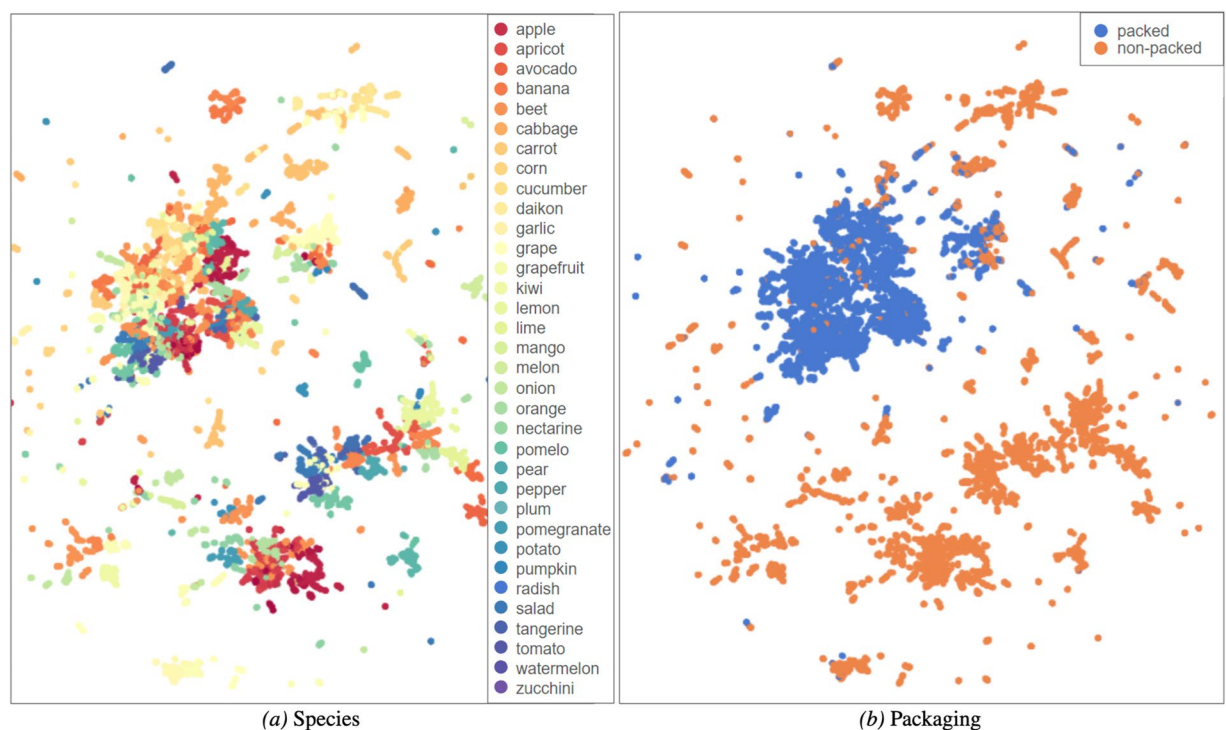
Figure 5 reveals that the majority of samples were collected in Saint-Petersburg, although the contributions from other cities are evenly distributed.

**CLIP Embeddings Analysis.** Prior to conducting experiments, we employ visualization techniques to gain insight into the complexity of tasks by analyzing CLIP embeddings<sup>39</sup>. Our primary objective is to understand the expected performance of foundation models pre-trained on general domains in zero-shot inference scenarios. In contrast to supervised training, where the latent space is optimized to fit the task-specific data, leading to improved class distinguishability, we focus on zero-shot performance as a proxy for model generalizability and the achievable quality without extensive data collection and annotation.

Figure 9 presents a visualization of latent space based on CLIP embeddings for the dataset, offering insights into how a model trained on general domain data discriminates between different varieties. CLIP embeddings



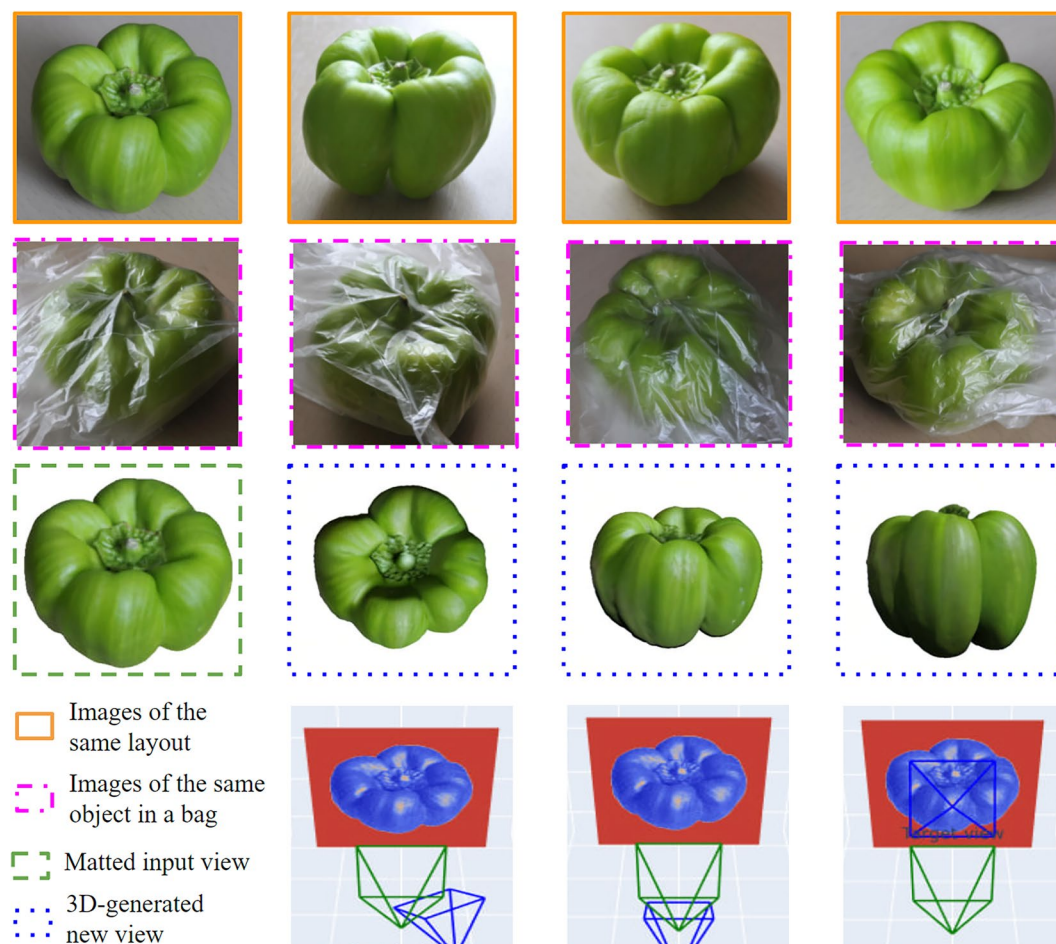
**Fig. 8** The number of images taken in different cities.



**Fig. 9** UMAP visualization of CLIP embeddings.

serve as a proxy measure to assess the model's ability to distinguish and categorize the varieties in our dataset. To facilitate visualization, we employ Uniform Manifold Approximation and Projection (UMAP)<sup>46</sup>, a dimension reduction algorithm that preserves global data structure while providing efficient run times. UMAP constructs a fuzzy topological representation of the high-dimensional CLIP embedding space and optimizes a low-dimensional representation to match this topology. This technique enables us to visualize and analyze the relationships between data points in a simplified 2-dimensional space. Therefore, Figure 9 presents two exemplary visualizations, specifically for species and packaging classification tasks. The visualization results suggest that, although most species are separable, their boundaries are not well-defined, and varieties are more challenging to distinguish. As anticipated, the embeddings of packed and non-packed objects exhibit significant differences, with packed object classes being more difficult to recognize.

Furthermore, we observe that CLIP embeddings struggle to separate images with varying object counts, indicating limited object counting capabilities. Nevertheless, a notable exception is observed with top-view images that exclusively feature multiple objects without any background context; these images are distinctly separated from other images.



**Fig. 10** Examples of image crops from the dataset and potential of generation of diverse 3D views.

Additionally, we verify that images from different cities, shops, and annotators exhibit distinct characteristics, thereby justifying the decision to split training and testing sets based on these factors to mitigate overfitting due to annotator and shop biases.

**Multi-View Data.** A crucial aspect of our dataset design involves incorporating four viewpoints for each layout, which enables us to investigate the impact of multi-viewpoint data on model performance. Specifically, we examine the effect of leveraging multiple viewpoints as a test-time augmentation (TTA) strategy to enhance model accuracy in Experiments section. Moreover, we propose additional use cases for this type of data. Figure 10 illustrates a single layout featuring, a single unpacked bell pepper, and a single packed bell pepper, demonstrating the potential to generate diverse 3D views from these layouts<sup>47</sup>. Orange border on the first row marks four images of the same layout. Purple border on the second row marks four photos of the same object packed in a plastic bag. The most left photo on the third row is an image from the dataset without any background. The rest of the photos on the third row are examples of 3D-generated samples based on the same object. Schemes on the last row show the difference between the original and 3D-generated views for the corresponding photos above them. This approach can provide an additional source of training samples, which can be exploited to further enrich the dataset and potentially improve model generalizability.

Figure 5 also presents illustrative examples of top-view images encompassing several varieties of produce. While these images can be advantageous for additional image augmentation scenarios, there may be instances where researchers prefer to exclude them from the training set to tailor the dataset to specific requirements. Consequently, these top-view images are not included in the testing set either, ensuring consistency and avoiding potential biases. The inclusion of such images in the dataset offers flexibility, allowing users to incorporate or omit them based on their particular research objectives and experimental designs.

### Data Records

The PackEat Dataset is published on Zenodo and is associated with record number 16901177<sup>11</sup>. All images are presented in the 'jpg' format. The segmentation images and annotations for them are stored in folders with extension 'segmentation'. To support reproducibility of results of machine learning models, we splitted the dataset into 'train' and 'test' folders for classification and into 'segmentation\_train', 'segmentation\_test' for

Supervision	Model	Species (34 classes)				Varieties (65 classes)			
		All	Packed	Not Packed	All Multiview	All	Packed	Not Packed	All Multiview
Zero-Shot	CLIP	0.43	0.28	0.58	<b>0.58</b>	<b>0.28</b>	<b>0.2</b>	0.36	<b>0.4</b>
	BioCLIP	0.3	0.12	0.48	0.46	0.1	0.04	0.15	0.17
	BioCLIP (taxons)	0.39	0.19	0.6	0.51	0.24	0.09	<b>0.4</b>	0.36
	LLaVA1.5	<b>0.58</b>	<b>0.45</b>	<b>0.73</b>	—	0.24	0.17	0.3	—
Linear Probing	CLIP	0.95	0.91	<b>0.99</b>	0.96	<b>0.96</b>	<b>0.94</b>	<b>0.98</b>	0.96
	BioCLIP	0.88	0.78	0.97	0.92	0.88	0.82	0.95	0.9
Supervised	ConvNext	<b>0.96</b>	<b>0.96</b>	0.96	<b>0.98</b>	0.95	<b>0.94</b>	0.96	<b>0.97</b>
Supervised with non-packet only	ConvNext	0.76	0.54	0.98	0.85	0.71	0.45	0.97	0.81

**Table 3.** Classification Baselines Results (F1-score).

segmentation, respectively. For the segmentation task, annotation is stored in ‘txt’ format to reduce the required memory. We also provided two files in ‘CSV’ format describing taxonomy for each variety of the collected data (‘taxonomy.csv’) and information for variety classification (‘variety\_classification.csv’).

Technical Validation

This section establishes baselines for the primary objectives for which the PackEat dataset was designed. We evaluate models with varying levels of supervision on our dataset, ranging from zero-shot to supervised learning. It is important to note that the term “baselines” refers to the utilization of widely adopted models with default hyperparameters and pipelines, without extensive tuning or data engineering. Consequently, it is anticipated that further optimization and data manipulation would yield improved outcomes. However, this aspect falls beyond the scope of the present paper, as our primary focus is to establish a benchmark for investigating model generalization capabilities.

**Model Details.** *Zero-Shot Scenario.* For the zero-shot scenario, we utilize CLIP-based models due to their renowned capability for effective image-text semantics matching in the general domain and efficient inference on GPUs or even CPUs. Firstly, we employ the OpenAI implementation of CLIP with a ViT-B/32 backbone<sup>48</sup>, using an image size of 224 × 224. We also experiment with BioCLIP, which is a variant of CLIP that utilizes a ViT-B/16 backbone and has been trained on a large corpus of biological data. BioCLIP is tested in two modes: common name mode and scientific mode. In common name mode, the answer options for species are the common fruit and vegetable names used in our dataset, such as ‘apple’. For varieties, we include the corresponding variety in brackets, such as ‘apple (golden delicious)’. In scientific mode, BioCLIP is provided with a string that includes the common name along with taxonomy details, such as ‘Malus domestica with common name apple (golden delicious)’. We found that providing the full taxonomy resulted in poorer performance. For example, ‘Plantae Tracheophyta Magnoliopsida Rosales Rosaceae Amygdaloideae Malus domestica with common name apple (golden delicious)’ did not yield better results. We share the complete taxonomy for each variety used in our experiments.

Another CLIP-based model we employ is LLaVA1.5<sup>41</sup>, which combines the CLIP ViT-L visual encoder and the LLaMA text decoder, trained for natural text instruction following. The image size used for LLaVA1.5 is 336 × 336. The prompt utilized for species classification is ‘What is on the image? Answer with a single option. Options are: ’, followed by a comma-separated list of species. For varieties classification, the prompt is modified to include the variety in brackets: ‘What is on the image? Answer with a single option including the specific variety in brackets. Options are: ’, followed by a comma-separated list of varieties.

*Linear Probing Scenario.* In the linear probing experiments, we apply logistic regression with an L2 penalty and a regularization strength of 0.316 to classify the embeddings obtained from both CLIP and BioCLIP visual encoders. This approach allows us to evaluate the effectiveness of the embeddings in representing the images for the classification task.

*Supervised Learning Scenario.* In the context of supervised learning, we employ a ConvNext-tiny model with an image size of 224 × 224. ConvNext is a state-of-the-art convolutional neural network architecture that has demonstrated impressive performance in various computer vision tasks. By utilizing the ConvNext-tiny variant, we aim to strike a balance between model complexity and computational efficiency. This model is trained in a fully supervised manner using the labeled images and their corresponding annotations from our dataset.

For the ConvNext model, we employ a learning rate of 5e-5, a batch size of 128, and utilize a 4-step gradient accumulation strategy. The training process is relatively brief, spanning only 6 epochs. To enhance the model’s performance and robustness, basic augmentation techniques are applied, including random rotations, horizontal flipping, and contrast adjustments.

**Classification Results.** *Experimental Setup.* In this experimental setup, we compare different pipelines for the classification problem under varying levels of supervision: zero-shot, linear probing, and supervised learning. The results of this comparison are presented in Table 3. Zero-shot classification refers to the scenario where the model has not been trained on any images from our dataset. Zero-shot models make predictions by selecting the



Supervision	Model	Precision	Recall	mAP50	mAP50-95
Zero-Shot Object Detection	YOLOWorld	0.15	0.27	0.1	0.088
Supervised Object detection	YOLOv8	0.87	0.81	0.86	0.74
Supervised Instance Segmentation	YOLOv8seg	0.86	0.81	0.85	0.73

**Table 4.** Species Localization Baselines Results.

most likely answer from a predefined set of categories. Linear probing, on the other hand, is a resource-efficient approach where we employ a neural network to obtain latent representations (embeddings) of the images, and then tune small machine learning models to classify these embeddings. Linear probing requires less training data and computational resources than supervised learning, but it relies on the assumption that the embeddings adequately represent the images for the specific task. As we will observe in a later section, this assumption does not hold true for CLIP embeddings in the object counting task, highlighting the limitations of this method. Supervised learning, on the other hand, refers to the generic training of a model using labeled data.

In a supplementary experiment, we explore the impact of training data composition on model performance. Specifically, we conduct an experiment where only unpacked fruits and vegetables are used for model training, while the entire dataset, including both unpacked and packed images, is utilized for testing. This experimental design allows us to evaluate the models' ability to generalize to more complex recognition.

To perform a comprehensive analysis, we conduct separate evaluations on subsets of varying complexity. This includes different levels of label hierarchy granularity, such as species and varieties, and different packaging types: not packed, packed in plastic bags, and a combination of both. It is worth noting that given a balanced dataset with respect to packaging, the evaluation on all samples yields results that are close to the average of packed and unpacked items.

Additionally, we take advantage of having four different viewpoints for each layout by evaluating a multi-view pipeline that averages the predictions from individual images. However, due to the lengthy inference time, the results for LLaVA in the multi-view setting are not included.

**Result Discussions.** The results presented in Table 3 are categorized according to the type of supervision used, and the best results in the zero-shot group and overall are highlighted in bold. This distinction is made to emphasize the varying data requirements for each method. Zero-shot pipelines do not require any training data, while supervised learning with non-packed-only data is relatively easier to collect manually or obtain from existing datasets. On the other hand, supervised learning and linear probing utilize the entire dataset.

The reported results in Table 3 demonstrate the F1-score metrics for each method. Overall, as expected, approaches that leverage more data tend to yield better performance. Interestingly, zero-shot methods still provide reasonably accurate results. Furthermore, the results clearly indicate that species classification is an easier task compared to varieties classification, and objects packed in plastic bags are generally easier to distinguish than unpacked items. The benefit of utilizing multiple viewpoints is evident, as it consistently improves the performance across all methods.

Among our observations, BioCLIP demonstrates improved performance on common fruits and vegetables among zero-shot methods, but only when they are not packed. However, its performance declines significantly on packed samples, and it lacks representation capabilities for linear probing. It is important to mention that finding an effective textual representation for BioCLIP was challenging, and incorrect combinations of common names and scientific taxons led to worse outcomes.

LLaVA, on the other hand, utilizes a more extensive and sophisticated language model built on top of CLIP visual encoder, offering a more intuitive natural language interface and applicability beyond classification tasks. LLaVA also enhances performance on species recognition. However, it is more resource-intensive and time-consuming. Additionally, its performance on varieties recognition is lower, which we attribute to the use of common species in the training of CLIP and LLaMA<sup>49</sup>, whereas varieties recognition requires more domain-specific knowledge. This observation suggests that introducing rarer fruits and vegetables could potentially lead to a significant decrease in accuracy.

The supervised pipeline, as expected, sets a high bar for other methods, effectively solving both species and varieties classification with good scores. Notably, in certain cases, linear probing achieves equivalent or even higher results compared to the supervised approach. Furthermore, when excluding packed samples from the training subset, the performance on unpacked test images improves slightly compared to the full dataset, but the results on packed samples deteriorate drastically.

**Localization Results.** Table 4 presents the results of our object localization experiments. For the zero-shot setting, we employ the YOLOWorld model, a state-of-the-art on-device object detection system pre-trained on a diverse set of classes. In contrast, for the supervised experiment, we utilize the YOLOv8 model for object detection and YOLOv8seg for instance segmentation. The metrics exhibited in the table differ significantly, providing insight into the performance expectations for each approach. The zero-shot method, without fine-tuning, performs poorly even on common species. It struggles to accurately localize and classify objects, particularly those that are packed. This highlights the limitations of zero-shot object localization, especially when dealing with specific classes or packed items. On the other hand, the supervised training of YOLOv8 and YOLOv8seg yields excellent results, even on challenging samples. This confirms that our dataset is sufficiently large and clear to enable effective supervised training for object localization tasks.

#Train Samples	Packed	Not Packed
10	0.014	0.017
100	0.12	0.27
250	0.46	0.65
500	0.74	0.84
All	0.94	0.96

**Table 5.** Observations on Scaling Laws (F1-score).

**Ablation Results.** *Object Counting.* In addition to the primary classification task, we also evaluate the models' ability to perform object counting, specifically focusing on fruits and vegetables. Notably, we do not utilize specialized object counting models, instead opting for more general-purpose models that can simultaneously tackle the main classification task.

The results for the LLaVA model reveal a mean average percentage error (MAPE) of 70%, which is a relatively modest performance.

Furthermore, we investigate the feasibility of linear probing the CLIP model on a simplified object counting task, where objects are categorized into four bins: one object, two to three objects, four to six objects, and seven or more objects. The results yield an F1-score of 0.78 for non-packed objects and 0.63 for packed objects.

An alternative approach to object counting involves post-processing the detection results, which are reported in the preceding subsection.

*Packaging Recognition.* The presence of packed samples is a prominent feature of our dataset, prompting an investigation into how models recognize packaging. Zero-shot LLaVA achieves an 89% accuracy rate in understanding that objects are packed, while supervised tuning of ConvNext yields an expected 99.96% accuracy. As anticipated, there is minimal confusion with samples like salads, which may have plastic wrapping but are not entirely packed.

*Scaling Laws.* To substantiate the necessity of a large dataset for effective training, Table 5 presents the results of an experiment where we train the ConvNext model for varieties classification using smaller subsets of our dataset. This experimental setup allows us to assess the impact of dataset size on the model's performance and demonstrates the value of having a substantial amount of labeled data for achieving better generalization and overall accuracy.

*The Influence of the Background Type.* We investigate the influence of background type on model accuracy by evaluating zero-shot CLIP on non-packed species classification. The F1-score for images with simple, uniform backgrounds is 0.583, while for complex backgrounds, it slightly decreases to 0.573. For the more challenging task of varieties classification, the F1-scores are 0.376 and 0.347, respectively. Our findings confirm that background context plays a crucial role in prediction quality, with a more significant impact on more complex tasks. These results highlight the importance of considering background information in model development and evaluation.

**Limitations.** Despite its notable strengths, the PackEat dataset exhibits certain limitations. Firstly, the dataset predominantly focuses on fruits and vegetables, thus limiting its scope in terms of product diversity. Future expansions that encompass a broader range of retail products could enhance its applicability and versatility. Additionally, while the dataset exhibits a relatively balanced distribution across varieties, there remain instances of outliers that may introduce biases or challenges in model training. Furthermore, the dataset's image backgrounds, although varied, may not entirely reflect natural shop or storage environments. Another limitation lies in the temporal scope of the dataset, as it was collected during a single month, neglecting seasonal variations in product appearances throughout the year. However, it provides a good representation of the target vegetable and fruit species, as the photos were captured in multiple stores and locations under varying sensing conditions, including different cameras and lighting setups. These factors contribute to the dataset's reliability. We assume that the visual appearance of the vegetables and fruits remains generally consistent across seasons, as some are grown in greenhouses and are therefore less affected by seasonal variations. Lighting conditions, however, are a critical factor that can influence a neural network model's performance. While the dataset already includes diverse lighting scenarios, additional augmentation techniques could be applied to further enhance variability, improving the model's robustness to real-world conditions<sup>50</sup>.

The background complexity labels are currently subjective, reflecting the annotators' general assessment. The variety of background patterns is intended to help neural network models adapt to diverse and challenging conditions, so even backgrounds that are unusual for grocery store environments are included to enhance dataset diversity. In the future, background complexity could be assessed using general or specialized computer vision techniques for homogeneity evaluation. For instance, texture uniformity could be measured using Local Binary Patterns (LBP) or entropy-based measures, while gradient-based approaches such as edge density or Sobel filters could serve as fundamental tools for quantifying complexity. Additionally, community feedback could be incorporated to improve dataset completeness and suitability for future versions, ensuring better adaptability to various background conditions in real-world applications.

Notwithstanding these limitations, PackEat offers significant value by facilitating the development and evaluation of multitask models in computer vision and retail automation. Its contribution lies in providing a specialized dataset tailored for the retail domain, thereby advancing visual recognition capabilities in real-world retail environments. The dataset is expected to serve as a valuable resource for the research community, stimulating innovative solutions and fostering advancements in the field.

### Code availability

The code is available through the link: <https://www.kaggle.com/datasets/sergeynesteruk/packed-fruits-and-vegetables-recognition-benchmark/data>.

### Data availability

The dataset is available through the link: <https://zenodo.org/records/16901177> with the record number 16901177 on Zenodo. It includes images in 'jpg' format. The dataset is split into 'train' and 'test' folders. The segmentation images and annotations for them are stored in folders with extension 'segmentation'. For the segmentation task, annotation is stored in 'txt' format to reduce the required memory. We also provided two files in 'CSV' format describing taxonomy for each variety of the collected data ('taxonomy.csv') and information for variety classification ('variety\_classification.csv').

Received: 17 January 2025; Accepted: 10 November 2025;

Published online: 28 November 2025

### References

- Begley, S., Marohn, E., Mikha, S. & Rettaliata, A. Digital disruption at the grocery store. *McKinsey & Company: London, UK* 1–8 (2020).
- Wei, Y., Tran, S., Xu, S., Kang, B. & Springer, M. Deep learning for retail product recognition: Challenges and techniques. *Computational intelligence and neuroscience* **2020** (2020).
- Tang, Y. M., Chau, K. Y., Lau, Y.-y & Zheng, Z. Data-intensive inventory forecasting with artificial intelligence models for cross-border e-commerce service automation. *Applied Sciences* **13**, 3051 (2023).
- Fahad, L. G. *et al.* Fruits and vegetables freshness categorization using deep learning. *Computers, Materials & Continua* **71** (2022).
- Sultana, N., Jahan, M. & Uddin, M. S. An extensive dataset for successful recognition of fresh and rotten fruits. *Data in Brief* **44**, 108552 (2022).
- Štulec, I., Petljak, K. & Kukor, A. The role of store layout and visual merchandising in food retailing (2021).
- Chen, C., Anjum, S. & Gurari, D. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19098–19107 (2022).
- Guimarães, V., Nascimento, J., Viana, P. & Carvalhal, P. A review of recent advances and challenges in grocery label detection and recognition. *Applied Sciences* **13**, 2871 (2023).
- Ninja, D. Visualization tools for fruits & vegetable detection dataset. <https://datasetninja.com/fruits-and-vegetable-detection> Visited on 2024-05-12 (2024).
- Selvam, P. & Koilraj, J. A. S. A deep learning framework for grocery product detection and recognition. *Food Analytical Methods* **15**, 3498–3522 (2022).
- Nesteruk, S., Illarionova, S., Elina, T., Bezzateev, S. & Burnaev, E. Packeat: Packed fruits and vegetables visual classification and segmentation benchmark <https://doi.org/10.5281/zenodo.16901178> (2025).
- Wudhikarn, R., Charoenkwan, P. & Malang, K. Deep learning in barcode recognition: A systematic literature review. *IEEE Access* **10**, 8049–8072 (2022).
- Seghezzi, A., Mangiaracina, R. & Tumino, A. E-grocery logistics: exploring the gap between research and practice. *The International Journal of Logistics Management* **34**, 1675–1699 (2023).
- Saihui Hou, Y. F. & Wang, Z. Vegfru: A domain-specific dataset for fine-grained visual categorization. In *IEEE International Conference on Computer Vision* (2017).
- Klasson, M., Zhang, C. & Kjellström, H. Using variational multi-view learning for classification of grocery items. *Patterns* **1** (2020).
- Marin, J. *et al.* Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**, 187–203 (2021).
- Olóndriz, D. A., Puigdevall, P. P. & Palau, A. S. Foodi-ml: a large multi-language dataset of food, drinks and groceries images and descriptions. *arXiv preprint arXiv:2110.02035* (2021).
- Chen, X., Zhu, Y., Zhou, H., Diao, L. & Wang, D. ChineseFoodnet: A large-scale image dataset for chinese food recognition. *arXiv preprint arXiv:1705.02743* (2017).
- Nesteruk, S. *et al.* Pseudoaugment: Enabling smart checkout adoption for new classes without human annotation. *IEEE Access* **11**, 76869–76882, <https://doi.org/10.1109/ACCESS.2023.3296854> (2023).
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695 (2022).
- Zhu, J.-Y., Park, T., Isola, P. & Efros, A. A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232 (2017).
- Muresan, H. & Oltean, M. Fruit recognition from images using deep learning. *Acta Universitatis Sapientiae, Informatica* **10**, 26–42 (2018).
- Liu, X. *et al.* The recognition of apple fruits in plastic bags based on block classification. *Precision agriculture* **19**, 735–749 (2018).
- Rojas-Aranda, J. L., Nunez-Varela, J. I., Cuevas-Tello, J. C. & Rangel-Ramirez, G. Fruit classification for retail stores using deep learning. In *Pattern Recognition: 12th Mexican Conference, MCPR 2020, Morelia, Mexico, June 24–27, 2020, Proceedings* 12, 3–13 (Springer, 2020).
- Bogomasov, K. & Conrad, S. Efficient fruit and vegetable classification and counting for retail applications using deep learning. In *Proceedings of the 5th International Conference on Advances in Artificial Intelligence*, 65–71 (2021).
- Hemming, J., Ruizendaal, J., Hofstee, J. W. & Van Henten, E. J. Fruit detectability analysis for different camera positions in sweet-pepper. *Sensors* **14**, 6032–6044 (2014).
- Nesteruk, S. *et al.* Xtremeaugment: Getting more from your data through combination of image collection and image augmentation. *IEEE Access* **10**, 24010–24028 (2022).
- Lowe, D. G. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, vol. 2, 1150–1157 (Ieee, 1999).
- Navneet, D. Histograms of oriented gradients for human detection. In *International Conference on Computer Vision & Pattern Recognition, 2005*, vol. 2, 886–893 (2005).
- Franco, A., Maltoni, D. & Papi, S. Grocery product detection and recognition. *Expert Systems with Applications* **81**, 163–176 (2017).

31. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
32. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
33. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826 (2016).
34. Liu, Z. *et al.* A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 11976–11986 (2022).
35. Koch, G., Zemel, R., Salakhutdinov, R. *et al.* Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, vol. 2 (Lille, 2015).
36. Wang, C., Huang, C., Zhu, X. & Zhao, L. One-shot retail product identification based on improved siamese neural networks. *Circuits, Systems, and Signal Processing* **41**, 6098–6112 (2022).
37. Horng, S.-J. & Huang, P.-S. Building unmanned store identification systems using yolov4 and siamese network. *Applied Sciences* **12**, 3826 (2022).
38. Nguyen, P. A. & Tran, S. T. Tracking customers in crowded retail scenes with siamese tracker. In *2020 RIVF international conference on computing and communication technologies (RIVF)*, 1–6 (IEEE, 2020).
39. Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PMLR, 2021).
40. Lee, U. *et al.* An automated, clip-type, small internet of things camera-based tomato flower and fruit monitoring and harvest prediction system. *Sensors* **22**, 2456 (2022).
41. Liu, H., Li, C., Li, Y., & Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 26296–26306) (2024).
42. Rees, J. A. & Cranston, K. Automated assembly of a reference taxonomy for phylogenetic data synthesis. *Biodiversity Data Journal* (2017).
43. Hinchliff, C. E. *et al.* Synthesis of phylogeny and taxonomy into a comprehensive tree of life. *Proceedings of the National Academy of Sciences* **112**, 12764–12769 (2015).
44. Patterson, D., Mozzherin, D., Shorthouse, D. P. & Thessen, A. Challenges with using names to link digital biodiversity information. *Biodiversity data journal* (2016).
45. Bánki, O. *et al.* Catalogue of life checklist (annual checklist 2023). *Cat Life* (2023).
46. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).
47. Liu, R. *et al.* Zero-1-to-3: Zero-shot one image to 3d object 2303.11328 (2023).
48. Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
49. Touvron, H. *et al.* Llama: Open and efficient foundation language models. *ArXiv abs/2302.13971*, <https://api.semanticscholar.org/CorpusID:257219404> (2023).
50. Nesteruk, S., Illarionova, S. & Somov, A. Image dataset augmentation a survey and taxonomy. *Measurements and Instrumentation for Machine Vision* 110–136 (2024).

## Acknowledgements

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4F0002 and the agreement with Skoltech No139-10-2025-033.

## Author contributions

Conceptualization, S.N.; methodology, S.N. and S.I.; software, S.N.; validation, S.I. and S.N.; formal analysis, S.I. and S.N.; investigation, all authors; resources, E.B.; data curation, S.I.; writing—original draft preparation, all authors; writing—review and editing, all authors; visualization, S.N.; supervision, E.B.; funding acquisition, E.B. All authors reviewed the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to S.I.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025