Journal of Computational Social Science
https://doi.org/10.1007/s42001-024-00266-7

RESEARCH ARTICLE

Efficient annotation reduction with active learning for computer vision-based Retail Product Recognition

Niels Griffioen[1] · Nevena Rankovic[1] · Federico Zamberlan[2] · Monisha Punith[3]

Received: 27 September 2023 / Accepted: 27 February 2024

Abstract

The retail industry encounters huge obstacles with computer vision (CV) technology due to frequent model retraining with changing products and time-consuming, costly data annotation. Previous research in this field has been primarily focused on optimizing model performance rather than minimizing annotation effort. Therefore, the main idea of this paper is to evaluate active learning as a method to minimize annotation effort in the retail industry. The MVTEC Densely Segmented Supermarket dataset is used to evaluate various active learning methods such as the Least Confident, Entropy and Cost-Effective Active Learning (CEAL) along with Mask R-CNN model. The results demonstrate that annotating only 20.83−24.34% of the data achieves 95% of the full dataset's performance. When training, out-of-sample data share similar characteristics, the Least Confident and CEAL methods reduce annotation requirements by 7.7−15.7% while maintaining 95% and 97% of the full dataset's performance. However, the Entropy method under-performs compared to the random selection baseline. Ultimately, none of the methods show a clear advantage when the data characteristics differ between training and out-of-sample data. Finally, the proposed active learning methods on an industry-specific retail dataset remarkably propels the development of highly efficient and cost-effective CV solutions meticulously tailored for the retail industry.

Keywords Active learning · Annotation effort · Retail image analysis · Instance segmentation

Introduction

In the retail sector, CV techniques can be used for product recognition without requiring alterations to the product or packaging. However, frequent changes in products and packaging require regular retraining of models with new data, which is expensive to collect and annotate [1, 2]. This challenge is further compounded by complex CV methods such as instance segmentation, which rely on extensive

Extended author information available on the last page of the article

Published online: 12 April 2024

datasets with pixel-level annotations [3]. In the retail industry, the expensive and time-consuming process of creating such a dataset, may have to be repeated regularly due to frequent changes in the products. In this paper, active learning is evaluated as a method to reduce the annotation effort required to train a instance-segmentation CV model in the retail industry. It thereby aims to fulfill the need for efficiency in (re-)training CV models within the retail sector by evaluating active learning as a mechanism to mitigate annotation burdens. The presented methodology contributes a solution to the challenges encountered by the retail sector in the adoption of CV technology. Moreover, it furthers the comprehension of automated product segmentation using CV-based methods through the lens of an industry-specific challenge.

Product recognition within a retail context entails identifying items on shelves or during checkout, achieved either manually or through (semi-)automated processes [2]. Automated recognition offers economic advantages due to its reliability and time-efficiency. However, prevailing (semi-)automatic solutions often require modifications to products or packaging, such as barcodes or RFID tags [1]. In contrast, CV based methods solely require a camera, providing a cost-effective alternative [4]. Consequently, the integration of CV technology holds promising implications for retailers. Research within this domain has predominantly concentrated on establishing the feasibility of CV technology within the retail context. For example, studies by [5, 6] examine the application of CV methods in retail, with a focus on optimizing model performance. However, the exploration of annotation effort reduction remains unexplored within their investigations. Partial insight into this topic is offered by [7] wherein minimal data prerequisites are investigated, establishing that object detection requires over 100 instances per product for robust performance. Nevertheless, methods aimed at curtailing these minimal data requirements are unexplored in their study.

The scientific relevance lies in the evaluation of active learning as a method to address the challenge of minimum data requirements within the retail sector. Active learning strategically selects data points from an unlabeled dataset to optimize information gain, facilitating the prioritization of the annotation process by human annotators [3]. The retail industry's necessity for frequent retraining, make it a fitting candidate for the integration of active learning. Based on the current understanding and available research in this field, there are no studies that examine the use of active learning in the retail industry. The scientific merit of addressing this research gap comes from the application of active learning in a domain-specific context. Through the evaluation of active learning within the context of a domain-specific issue, this paper contributes to an improved understanding of the combination of automatic segmentation and active learning within the field of CV. From a societal perspective, the strategic selection of data-points for annotation reallocates annotator effort from tasks with minimal impact to those yielding more substantial outcomes. This reallocation serves to optimize human labor utilization, thereby resulting in cost reduction that enhances the feasibility of employing CV systems for product identification.

The Mask R-CNN model was chosen for its strong performance in instance segmentation, as evidenced in the COCO 2016 challenge, and its suitability for real-time applications in self-checkout systems, as highlighted by [5]. This study

evaluates three active learning strategies: Least Confident, Label Entropy, and Cost-Effective Active Learning (CEAL), to determine their effectiveness in reducing annotation effort in a retail setting. This paper aims to assess the impact of active learning on minimizing manual annotation work for retail product recognition.

The paper's contributions include:

1. An evaluation of the role of active learning methods in reducing human annotation effort for retail product recognition and quantifying the extent of efficiency improvement in the annotation process.
2. An analysis of Mask R-CNN's performance when trained on a selected subset of informative images, focusing on its impact on mean Average Precision (mAP) in a retail environment.
3. A comparative study of active learning strategies-Least Confident, Label Entropy, and CEAL-against random selection (RANDOM), specifically examining their effectiveness in minimizing manual annotation effort in retail settings.

The rest of the paper is organized as follows: Literature review provides an overview of related literature in the automated product recognition and active learning domain. Research methodology and experimental setup describes the methodology employed in this research. Results presents the obtained results which are further discussed in Sect. Discussion. Finally, concluding remarks are given in Sect. Conclusion.

Literature review

This Section offers a review of related academic literature. In Sect. Automated product recognition in a retail environment, automated product recognition in the retail context is examined, and Sect. Active learning explores active learning as a means to minimize effort in annotation tasks.

Automated product recognition in a retail environment

Currently, the primary technologies for retail product recognition are barcodes and RFID tags [1]. These technologies require alterations to the products by adding a RFID tag or barcode. In contrast, CV methods only require a camera to gather data, presenting an affordable option for product recognition [4].

Feature extraction is an crucial step in the many CV method. [4] categorize feature extraction methods into five main groups. The first four groups encompass key point, gradient, pattern, and color-based methods. These methods involve extracting specific features and transforming them into feature vectors. However, [1] note that hand-crafted features often fail to capture all the necessary information for accurate product classification. As a result, the fifth group, deep learning-based feature extraction, has gained prominence. Among these, Convolutional Neural Networks (CNNs) [8] are the most prevalent. For data with multiple objects per image, such as those found in the retail industry, efficient extraction of regions is crucial for

CNNs. [9] categorizes models into two main groups based on their region extraction approach. The first group are the one-stage models that regress from various positions in the image to determine object spatial positions. The second group are two-stage models that propose potential object locations in the first stage and then classify each region in the subsequent stage. For two-stage models, numerous techniques have been suggested for selecting regions of interest. For instance, R-CNN introduced selective search as an unsupervised region proposal method [10]. However, selective search can be computationally intensive since it involves CNN calculations for each region. To mitigate this, faster R-CNN integrated a region proposal network that shares convolutional features with the detection network to reduce redundant computations [11].

Mask R-CNN extends the Faster R-CNN framework by adding a mask head to the faster R-CNN framework [12]. This enables Mask R-CNN to perform instance segmentation. Instance segmentation combines object detection and semantic segmentation by generating separate pixel-masks for each object in an image [13]. This approach provides both a detailed understanding of the spatial layout and a pixel-wise comprehension of the image's objects [14, 15]. In comparison to semantic segmentation, it retains the pixel-wise object understanding and adds the capability to distinctly detect multiple objects from the same class [15]. In a study by [5], various single-stage and two-stage architectures such as SSD, YOLO v2, Faster R-CNN, and Mask R-CNN were compared on a retail dataset. Mask R-CNN was evaluated as a suitable model for smart retail product detection systems, due to it's combination of performance and average inference time [5].

[7] assessed the smallest amount of data needed for satisfactory performance in a practical retail setting. They studied three CNN architectures (Inception, Resnet, and Mobilenet). For image classification, they found that 6 to 20 instances were needed for 90% accuracy, and 26 to 51 instances for 95% accuracy. Object detection, however, demanded more data; around 42 instances per class for 90% accuracy and over 100 instances for 95% accuracy. Although [7] established these minimal data requirements, they did not explore methods to lower or alleviate these demands.

There's a growing focus on real-time shelf commodity detection to enhance customer service. Study by [16] highlight the emergence of visual-based detection methods, notably deep learning-based target detection, for its efficiency in interpreting image features and advancing shelf identification. Comparative analyses between standard target detection datasets and those specific to shelf goods have been conducted, examining features, benefits, and applications. Innovative approaches for constructing robust datasets and refining them through data enhancement techniques, especially for imperfect datasets, are discussed. Further, advancements in package identification, including large-scale, small target, and partially occluded object recognition, are reviewed.

Additionally, promising results The new method, proposed by [17] integrating class distribution-aware adaptive classification margins and cluster-based embedding, has been evaluated for classifying fruits and vegetables with similar features. Traditional deep convolutional neural networks (DCNNs) struggle with such similar items, but this approach uses ResNet50-generated features, projecting them into an enhanced feature space to improve class distinction and compactness. Various

adaptive margins have been tested, with vector projection assessing similarity for better intra-class compactness. This method, also addressing imbalanced dataset distribution, has shown significant clustering and classification improvements, offering real-world applicability with minimal added complexity.

In a study by [18], they tackled the challenge of needing minimal data for effective object detection by introducing a novel model architecture. Their experiments revealed that this new architecture outperforms Faster R-CNN on smaller datasets. However, the best achieved detection performance on the tested dataset was an mAP of 91.3. This result still falls behind other methods, like the Mask R-CNN model which reached an mAP of 99.27 in the work of [7]. Moreover, instance segmentation is more intricate than object detection, as it demands pixel-level ground-truth information [3]. Gathering and annotating a dataset of sufficient size is both time-consuming and expensive. Active learning has emerged as a potential approach to alleviate the data annotation burden for instance segmentation [3].

Active learning

Active learning (AL) operates by selecting unlabeled data points that offer the most valuable information. Typically, it follows an iterative approach where the usefulness of unlabeled data points is measured, and a batch of images is chosen for annotation. These chosen images are then labeled by human annotators, followed by model re-training. This ongoing process is depicted in Fig. 1. By focusing on informative data points, active learning decreases the workload for human annotators, zeroing in on the most crucial data for annotation. The ultimate aim is to preserve most of the model's performance while greatly reducing the need for human annotations. A common measure of data-point informativeness is the uncertainty of model predictions. The assumption is that more uncertain predictions yield greater information gain when their true labels are included in model training [19].

In [20], three processes are described for model-human annotation interactions. Membership query synthesis generates labels for newly generated, unlabeled

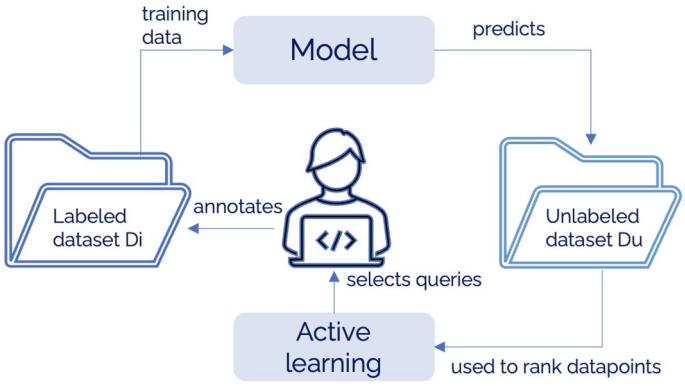


Fig. 1 Iterative active learning approach

instances, potentially posing challenges in human annotation due to arbitrary queries. Stream-based selective sampling involves sampling an instance from the unlabeled data and deciding whether to query it or discard it. However, this approach sends only one query at a time. Pool-based active learning, on the other hand, is suitable for scenarios with large unlabeled datasets, selecting multiple queries from a pool of unlabeled data for simultaneous annotation [20].

In the study by [21], they introduce Cost-Effective Active Learning (CEAL) as an approach for image classification within a pool-based active learning setup. Their method involves selecting images based not only on their informativeness for human annotation but also on the high confidence of their predictions. The model uses pseudo-labels from these highly confident images in the next training iteration to further reduce annotation needs. Another method, proposed by [22], called Batch-BALD, calculates the mutual information gain for an entire batch of sampled data-points. This approach addresses the limitation of ranking single data-points, where similar information might be grouped together, potentially missing additional information. Calculating the information gain for a batch of data-points is expected to yield better performance [22].

On one hand, the research proposed by [23] addresses the challenge of high-cost data annotation in specialized fields by exploring deep active learning (DeepAL). It offers a first comprehensive survey of DeepAL, presenting a formal classification, systematic overview, application insights, and discusses current issues and future directions in this emerging field. On the other hand, the study by [24] suggests the task-aware variational adversarial AL (TA-VAAL), which alters the task-agnostic VAAL, which took into account the distribution of data for both labeled and unlabeled pools. This is achieved by converting task learning loss prediction to ranking loss prediction and by embedding normalized ranking loss information on VAAL using a ranking conditional generative adversarial network. The suggested TA-VAAL outperforms state-of-the-arts in terms of semantic segmentation and its task-agnostic and task-aware AL features, as well as classifications with balanced or imbalanced labels on a variety of benchmark datasets. Particularly, it has been shown by [25] that combining the notions of uncertainty and diversity, easily scales to batch sizes (100K-1 M) several orders of magnitude larger than used in previous studies and provides significant improvements in AL model training efficiency compared to recent baselines.

Summarizing the existing literature, In previous studies the main focus has been on improving the performance of CV models. The Mask R-CNN model stands out for its efficient compute time during inference, strong performance, and its ability to understand objects at a pixel level. Despite its advantages, the challenge of annotating data remains, as creating a sufficiently large annotated dataset is expensive. Previous research in the retail sector has attempted to overcome this challenge by designing new model architectures that perform well with fewer samples. However, these models still fall short of the performance achieved by the Mask R-CNN method. This creates a gap in the literature regarding how to minimize annotation requirements for an existing retail dataset. This research aims to fill this gap by exploring active learning methods, which involve selecting the most informative data points for human annotation. While active learning has been successfully

used to reduce annotation efforts in other contexts, its effectiveness on retail-specific datasets has not been thoroughly investigated. Therefore, this study aims to examine the potential of active learning to decrease human annotation efforts in the retail industry.

Research methodology and experimental setup

This Section breaks down the research methodology employed. In Sect. Description of the data set, an introductory overview of the dataset is provided. Moving forward, Sect. Methodology elaborates on the instance segmentation method and active learning techniques that were applied. Within Sect. Experimental design, we delve into a detailed explanation of the experimental framework. Finally, in Sect. Evaluation method a detailed explanation of the evaluation method is provided.

Description of the data set

The MVTec D2s: Densely Segmented Supermarket Dataset (MVTec D2s) simulates the industrial setting of an automated checkout system in a retail store [13]. Common benchmark segmentation datasets such as COCO [26] and Cityscapes [27] frequently do not address industrial constraints such as insufficient diversity and scarcity of labeled training data, thereby restraining the applicability within an industrial setting [13]. The MVtec D2s dataset simulates a real-world, industrial setting, thereby addressing the real-world application and challenges found in the retail industry. The dataset contain 21,000 high-resolution color images featuring grocery products categorized into 60 classes. The grocery products are arranges on a turntable, and then photographed in ten rotations and three different light intensities.

The dataset is split into a train, validation and test set which is presented in Table 1. The train and validation set have pixel-level annotations available, however the test-set annotations are not made public. In addition, the statistics summarized in Table 1 show a disparity in the distribution of objects per image and the occurrence of occlusion and clutter across the various datasets. These disparities are visually illustrated in Fig. 2. These disparities have intentionally been introduced into the dataset to simulate industrial demands, such as the necessity for reduced labeling efforts, streamlined dataset procurement and effortless incorporation of new product classes. In order to mitigate the absence of annotations for the test-set and evaluate the robustness of the active learning methods in an industrial setting, the

Table 1 MVTec D2s: Densely Segmented Supermarket Dataset

Split	Train	Validation	Test
# of images	4380	3600	13,020
# objects/image	1.58	4.35	3.83
# scenes w. occlusion	10	84	299
# scenes w. clutter	0	18	68

Fig. 2 Samples from the MVTec D2s dataset. Row **a** contains samples from the original training set, row **b** contains samples from the validation and test set

experimental approach is assessed on two new subsets of the dataset, designed to replicate distinct scenarios.

Scenario 1: In scenario 1, the training data emulates what might be procured at a retail checkout, using real customer transactions as a model. Since the training data emulates being collected from real customer transactions, the characteristics of the training data should closely match the characteristics of the out-of-sample data that is expected from real customer transactions. To optimize hyper-parameters, the 3600 validation images are randomly redistributed into new training, validation, and test sets. Post hyper-parameter optimization on the validation set, the training and validation sets are merged and hyper-parameters are fixed for all further stages of the process.

Scenario 2: The second scenario replicates the disparities that are present in the original dataset. The validation and test set remain consistent with those in scenario 1. For the training data, the 4380 images in the original training set are used, and 10.000 artifically generate images created by [13]. The original training data contains an average of 1.58 products per image, and does not contain occlusions or clutter. The out-of-sample data contains on average 4.35 products per image, and incorporates occlusion or clutter in a some of it's images. The artificially generated images consist of a random selection of one to fifteen products drawn from the original dataset, which are randomly placed within an image. This design attempts to emulate the industrial prerequisites as mentioned by [13]. It thereby replicates a scenario where data-collection practices meet industrial demands for minimized labeling effort and streamlined data acquisition. An instance of this scenario arises when the training data is amassed independently from the data utilized for inference, for instance in a warehouse. This divergence introduces dissonance between the attributes of the training data and the data upon which the model is required to perform inference during the checkout process.

Methodology

The model selected for automated product recognition is Mask R-CNN. This model is selected because of it's demonstrated performance during the COCO 2016 challenge [12]. This decision is reinforced by [5], who, in a retail dataset study, compared Mask R-CNN against SSD, Faster R-CNN, and YOLO, finding it to be an optimal choice for real-time processing at a self-checkout system due to its combination of performance and efficiency. Mask R-CNN, an extension of Faster R-CNN object detection model, makes instance segmentation possible through the addition of a parallel fully convolutional network (FCN) branch [12].

The Mask R-CNN model can be interpreted as a sequence of distinct steps that collectively produces robust instance segmentation capabilities. The first step consists of a convolutional network that forms the backbone of the model. It transforms the raw input image into a feature map that contains the essential feature representations. This feature representation is further used in the subsequent step which is the Region Proposal Network (RPN). This step scans the feature map to identify potential object-containing regions. The RPN places anchor centers across the feature map, and creates a diverse range of anchor boxes characterized by varying scales and aspect rations. A binary classifier then determines whether an anchor box contains an object and a bounding-box regressor predicts the offset between each anchor box and the true bounding box. Overlapping predictions are resoleved through non-maximum suppression, yielding the achor boxes with the highest classification scores which form the regions of interest (ROIs).

Following this step, the Region of Interest Alignment (RoiAlign) operation uses ROIs generated by the RPN and overlays them onto the feature map. Each ROI is divided into a number of discrete bins, and from each bin a number of feature values are randomly sampled and processed through a bottleneck model, to create a fixed-size feature map for each region. Finally, each fixed-size feature map is processed through two separate branches. The first branch employs fully connected layers to map the feature maps to object class identities and precise bounding box coordinates. Simultaneously, the second branch employs a fully convolutional network to generate segmentation masks for each region. A visual depiction of these steps is presented in Fig. 3.

Active learning methods are compared against a passive baseline (RANDOM), which uniformly selects images. This method is also called passive learning because it relies on the assumption that there each image has an equal information gain. Therefore selection of any combination of images, would not perform better than any other combination of images. In contrast, active learning methods rely on the assumption that there is a variance in information gain between images. Under this assumption selecting the correct combination of images would lead to a better performance. Three active learning metrics are chosen for comparison against the baseline.

The first approach, denoted as Least Confident, relies on the class probability of predicted objects. This method has performed well in various domains such as the information extraction tasks, or conditional random fields [28–30]. This metric

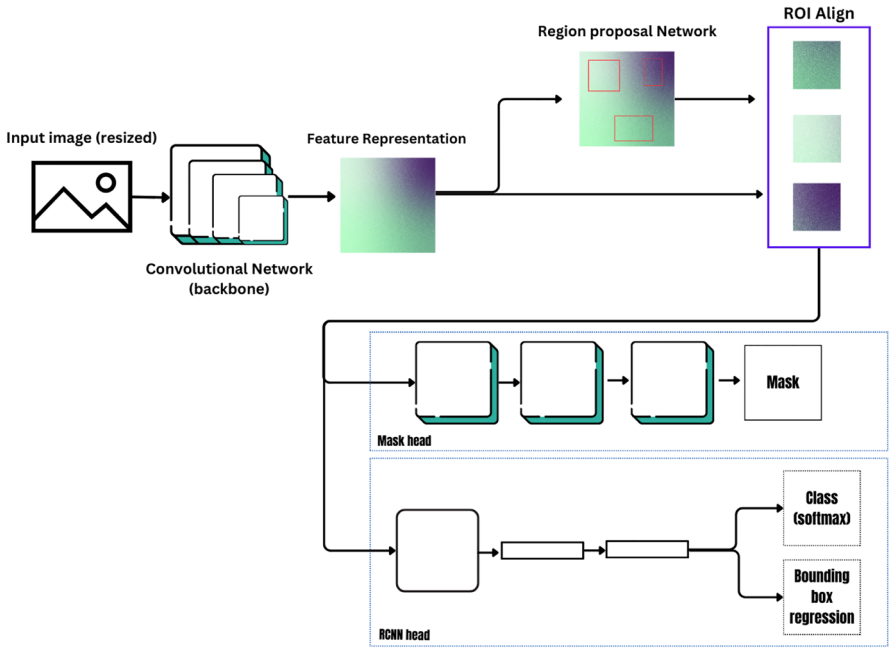


Fig. 3 Flowchart of the Mask R-CNN method

measures the uncertainty of an instance, based on the most probable predicted label for that instance using equation 1 and 2.

$$I_{LC} = \arg\max_{x} (1 - P(\hat{y}|x)), \tag{1}$$

$$\hat{y} = \arg\max_{y}(P(y_i|x)). \tag{2}$$

where $\hat{y}$ is the most probable label for instance x, and $P(\hat{y}|x)$ is the predicted probability of x being the most probable label $\hat{y}$. The information gain I_{LC} of the entire image is calculated by taking the maximum uncertainty among all instances x.

The Least Confident method uses only most probable label in it's calculation of uncertainty for each instance. The second approach, denoted as Entropy is an extension that measures the uncertainty of an instance based on all possible class labels for that instance [19]. The information for an image using the Entropy metric is done by equation 3 [19].

$$I_{EN} = \arg\max_{x} \left(- \sum_{i} P(\hat{y}_i|x) log(P(\hat{y}_i|x)) \right), \tag{3}$$

where $P(\hat{y}_i|x)$ is the probability of instance x belonging to the ith class. The information gain I_{EN} of the entire image is calculated by taking the maximum uncertainty among all instances x.

The third approach is known as Cost-Effective Active Learning (CEAL) [21]. This method is chosen due to its cost-effective nature, as demonstrated in the study by [21], and its compatibility with various other active learning methods. The efficacy of CEAL lies in its utilization of two types of sample selection. The first type involves using another active learning method to select samples that are then annotated by human annotators. The second type selects images for which the model exhibits high confidence in its predictions. These images are assigned pseudo-labels based on their model predictions, thereby avoiding the need for human annotation. The assignment of pseudo-labels is governed by equation 4 [21].

$$\begin{aligned} j^* &= argmaxp(y_i = j|x_i; W), \\ y_i &= \begin{cases} j^*, & en_i \quad lt;\delta, \\ 0 & \text{otherwise}, \end{cases} \end{aligned} \tag{4}$$

where en_i represents the label entropy for a given instance i, and j* indicates the most likely category for that instance. The threshold parameter δ is calculated using Eq. 5 [21].

$$\delta = \begin{cases} \delta_0, & t = 0, \\ \delta_0 - dr * t, & t > 0, \end{cases} \tag{5}$$

where δ_0 represents the initial threshold, which is set as a hyper-parameter, dr is another hyper-parameter governing the decay rate of the threshold, and t signifies the iteration number.

Experimental design

The experimental framework is structured into two distinct stages. The initial stage consists of hyper-parameter optimization using the complete training set. The resulting performance is denoted as the benchmark performance. This metric represents the upper bound that the iterative procedure in the further stage can approach. The second stage is an iterative process that is separately conducted for the RANDOM

baseline approach and each of the active learning methodologies as outlined in algorithm 1.

Algorithm 1 Iterative active learning algorithm

Require: Training samples D_{train}, Test samples D_{test}, sample selection size K, maximum iteration number T
Ensure: 1D array containing mAP performance metrics for each iteration W

Initialize empty 1D array W
Assign K random samples from D_{train} to D_l
Assign all other samples from D_{train} to D_u, so that $D_l \cap D_u = \emptyset$ and $D_l \cup D_u = D_{train}$
while not reached maximum iteration T **do**
    Initialize Mask R-CNN on D_l
    Evaluate mAP on D_{test} and append result to W
    Predict samples in D_u
    Rank instances in D_u using RANDOM, Least Confident, Entropy or CEAL method.
    From D_u move K top ranked samples to D_l with ground-truth annotations.
end while
return W

A visual representation of the algorithm is illustrated in Fig. 4. The anticipated behavior of the algorithm is that as the set of labeled images D_l expands, the performance of the Mask R-CNN model will gradually converge towards the benchmark level. The active learning methods are expected to systematically select the most informative images from D_u during each iteration, thereby approaching the benchmark performance with a reduced number of iterations compared to the baseline RANDOM method. Stated differently, with an equivalent quantity of annotated images in the set D_l, the performance of the Mask R-CNN model,

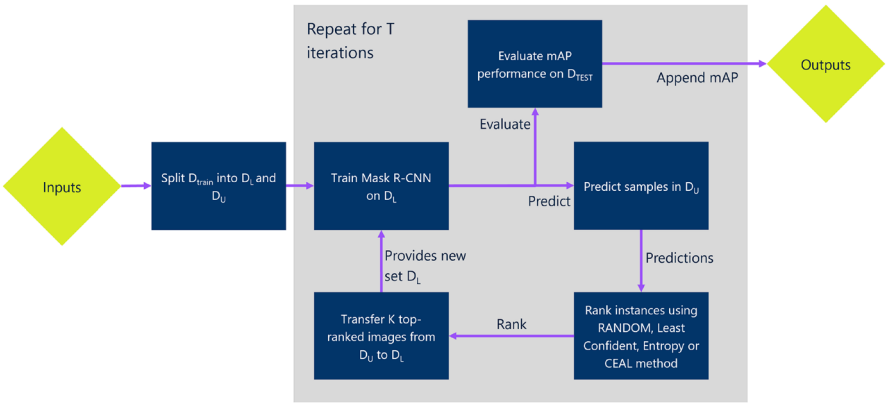


Fig. 4 Diagram depicting the iterative experimental setup utilized in the study

when employing any of the active learning approaches, is anticipated to exhibit closer proximity to the benchmark performance than when compared to the baseline RANDOM method.

During the hyper-parameter optimization stage, in scenario 1 the hyper-parameters are initially optimized using the validation set. Because both the training and validation are representative of customer transaction in a checkout scenario, the model is retrained on the combined set to report the final performance as a benchmark. Contrary to scenario 1, in scenario 2 the the training and validation sets are not combined when reporting the final performance. This separation is deliberate, as this scenario aims to emulate real-world industrial requirements as detailed in Sect. 3.1. Specifically, the validation set in scenario 2 does not meet the typical standards of industrial data collection, making it a test case for the model's ability to adapt to varying industrial conditions.

For the iterative process detailed in algorithm 1, training samples consists of the combined training and validation set in scenario 1, whereas in scenario 2 it consists only of the original training set. For both scenarios, once the hyper-parameters are optimized in the hyper-parameter stage, they remain fixed. There are no further adjustments to these parameters in the subsequent iterative stage of the methodology.

Evaluation method

The evaluation of models employs the mean average precision (mAP) metric, a widely used measure in instance segmentation and major competitions like the 2007 PASCAL VOC challenge and COCO segmentation challenge [26, 31]. To compute mAP, the Intersection over Union (IoU) metric is calculated pixel by pixel, as shown in equation 6.

$$IoU = \frac{target \cap prediction}{target \cup prediction}, \tag{6}$$

where $target \cap prediction$ counts pixels at the intersection of the prediction mask and ground truth, while $target \cup prediction$ counts pixels in their union [32].

Predictions are categorized using an IoU threshold. A prediction with an IoU score above the threshold is categorized as True Positive (TP), while those below are categorized as False Positive (FP). False Negatives (FN) indicate ground truths not predicted by the model. The average precision (AP) is determined by the area under the precision-recall curve for each class at a specific IoU threshold. Recall and precision are computed using equation 7.

$$\begin{aligned} Recall &= \frac{TP}{TP + FN}, \\ Precision &= \frac{TP}{TP + FP}. \end{aligned} \tag{7}$$

Finally, the AP is averaged across classes and thresholds within the [0.5,0.95] range with 0.05 increments to compute the mAP.

Results

In this section, we present the results obtained in this research. Section Benchmark Performance shows the benchmark performance. Sections Least Confident, Entropy, and Cost effective active learning show the performance achieved by the Least Confident, Entropy, and CEAL active learning methods.

Benchmark performance

This Section shows the performance of the Mask R-CNN model across both scenarios, utilizing the complete dataset. This analysis sets the benchmark performance, representing the highest level that the iterative process will approach as more images are annotated and moved from D_u to D_l.

For the benchmark, optimization of hyper-parameters occurred on the validation set. The ResNet architecture was chosen for its performance, for instance during the COCO 2015 detection and segmentation challenges [33]. Among the ResNet50 and ResNet101 variants, the latter demonstrated superior performance. The optimal epochs were 50 for scenario 1 and 20 for scenario 2. Among learning rates, 0.001 outperformed 0.00025. Performance was further improved by introducing a learning rate decay step for the final 25% of iterations, using the decay rate of 0.1 (which reduces the learning rate by a factor of 10). For the number of ROI heads,the values were 128, 256, and 512, with the latter achieving higher performance. A summary of the top 5 combinations for scenario 1 and scenario 2 can be referenced in Appendix A, Tables 3 and 4.

The performance on the out-of-sample data are presented in Table 2. The first row showcases the mAP across an intersection over IOU threshold range of [0.5, 0.95] with increments of 0.05. In scenario 1, an mAP of 96.72 was attained, while only 78.69 mAP was achieved in scenario 2. These values will form the benchmark performance against which the iterative process of the active learning methods is compared. The disparity between the two scenarios indicates that the Mask R-CNN model performs better when training data closely mirrors out-of-sample data in terms of inherent characteristics. The augmented data in scenario 2 inadequately

Table 2 Out-of-sample performance of the Mask R-CNN model for scenario 1 and 2

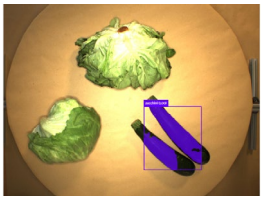
Metric	Scenario 1	Scenario 2
mAP [0.5,0.95]	96.72	78.30
mAP [0.5]	99.81	86.14
mAP [0.75]	99.64	83.33
mAP medium objects	81.26	10.00
mAP large objects	96.87	78.64

addressed disparities in object count per image and the presence of occlusion or clutter.

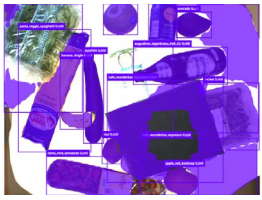
Additional commonly used metrics in object detection and classification, such as Precision, Recall, and F1-score for both scenarios, are available in Tables 5 and 6 in Appendix B. These supplementary metrics provide further evidence of the performance discrepancy between the two scenarios. For a detailed breakdown, the confusion matrices for scenario 1 and 2 are provided in Figs. 12 and 13 in Appendix C. These matrices affirm the observation that the model in scenario 1 accurately predicts the majority of instances, while the model in scenario 2 exhibits incorrect predictions for several product classes. Notably, in scenario 2, a significant number of "adelholzener alpenquelle classic" instances are inaccurately predicted as "adelholzener alpenquelle naturell". The most common prediction errors are failing to predict a product when present (false negative), erroneously predicting a product that is absent (false positive), and imprecise pixel masks. A selection of these typical prediction errors is illustrated in Fig. 5.

To assess the disparities in performance between the two scenarios, a quantitative examination is conducted. An imprecise mask predictions may be a plausible contributing factor to the divergence observed between the scenarios. This evaluation involves assessing the model at reduced IOU thresholds, wherein the assessment method considers less accurate mask predictions as valid true positive predictions. The evaluation is conducted individually at IOU thresholds of 0.5 and 0.75. The corresponding results are shown in rows 2 and 3 of Table 2. In scenario 1, the difference between the thresholds is marginal. Conversely, a larger difference exists between the thresholds in scenario 2. This dissimilarity implies that masks in scenario 2 exhibit comparatively lower accuracy compared to those in scenario 1. This observation lends support to the conjecture that the discernible performance variation between the two scenarios can be attributed, at least in part, to the imprecision in mask predictions within scenario 2.

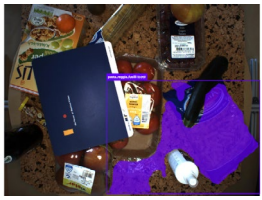
Another contributing factor may be the different product sizes. As per the COCO evaluation method, object sizes are classified into small (area $< 32^2$ px), medium



(a) False Negative



(b) False Positive



(c) Inaccurate Mask

Fig. 5 Samples of the common prediction errors. Sub-figure **a** contains false negatives, sub-figure **b** contains false positives for the caffee wunderbar class and sub-figure **c** contains an inaccurate mask for pasta reggia fusilli

(32^2 px < area < 96^2 px), and large (area > 96^2 px) categories [26]. The the same product class can fit in different categories based on it's orientation. The MVTEC D2S dataset encompasses medium and large object sizes. Seperate performance metrics corresponding to these categories are reflected in rows 4 and 5 of Table 2. In scenario 2, the discrepancy between medium and large object sizes is more pronounced than that observed in scenario 1. Within scenario 2, the model achieves an mAP of 10 for medium-sized objects, in contrast to 78.64 mAP for large-sized objects. This disparity implies that a sizable portion of the larger error rate in scenario 2 can be ascribed to relatively poorer performance on medium-sized objects. Conversely, this phenomenon is comparatively less pronounced in scenario 1.

Lastly, we consider performance disparities accross product classes as a potential contributing factor for the difference in performance. An overview of the AP for each product class can be found in Appendix D. In scenario 1, the least performing product classes have respective AP values of 87.37 and 89.27, while the most successful class has a rounded AP score of 100. The divergence in scenario 2 is larger, with the lowest AP recorded at 40.87 and the highest at 99. This contrast implies that a portion of the performance distinction between the two scenarios comes from variations in predictive prowess among distinct product classes.

Least confident

In scenario 1, the initial division splits the dataset into 50 images for D_l and 2830 images for D_u. This division is based on [21]'s investigation, achieving a 97% performance with around 35% of the data. We set the sample selection size K at 50 and the maximum iteration count T at 20. This configuration aims to leverage approximately 35% of the complete training set by the conclusion of the 20th iteration.

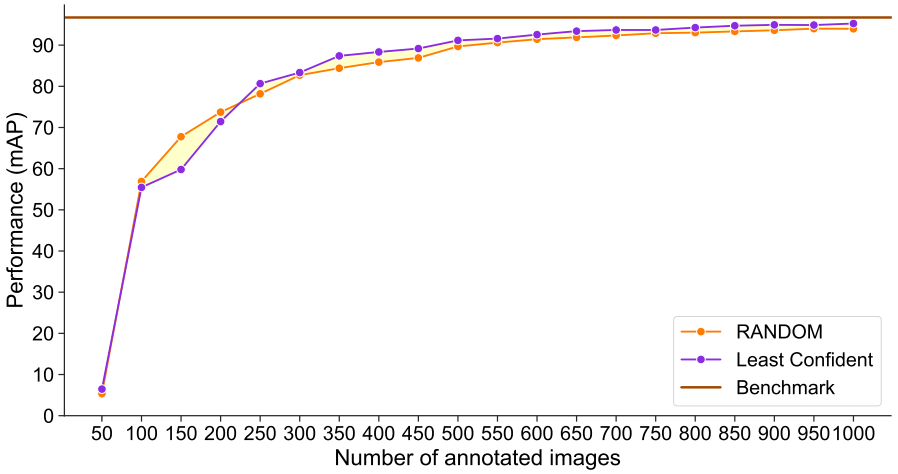


Fig. 6 Mask R-CNN model performance in active learning cycle: scenario 1 with Least Confident and RANDOM baseline

Figure 6 displays the evolving mean Average Precision (mAP) across iterations for scenario 1. The red line signifies the model's 97.72 mAP performance with the complete dataset, discussed in Sect. Benchmark performance. Initially, the baseline RANDOM approach outperforms the Least Confident active learning method in the initial iterations. However, after the fourth iteration, the Least Confident method consistently achieves superior performance. The concave trajectory of the lines in the Figure signifies diminishing returns as additional images are annotated. At the 20^{th} iteration, the Least Confident approach records a 95.25 mAP, whereas RANDOM achieves a 93.97 mAP. Relative to the benchmark, the Least Confident method attains 95% of the benchmark's performance by the 12^{th} iteration, utilizing only 20.83% of the data. In contrast, the RANDOM method requires 13 iterations to reach the same milestone, signifying a 50-image (7.69%) reduction in annotation needs at this stage. As the process advances, the divergence in annotation efforts becomes more pronounced. For instance, to achieve a 97% performance akin to the benchmark's full dataset achievement, the Least Confident method accomplishes this by the 16th iteration, while RANDOM requires 19 iterations. Consequently, the Least Confident approach reduces annotation necessity by 150 images (15.79%).

In scenario 2, the training dataset's substantial size necessitates setting the selection size at 250 to correspond to the same 35% data utilization as in scenario 1. Owing to the increased computational complexity of the larger training set, the maximum iteration count T in this context is constrained by the server's computational capacity. The server can run for up to 36 h concurrently, allowing for 17 iterations within this time frame. Figure 7 illustrates the mAP progression across each iteration. Scenario 2 showcases no performance difference between the Least Confident method and the RANDOM baseline. While the Least Confident active learning method outperforms RANDOM at 1750, 2000, and 4250 images,

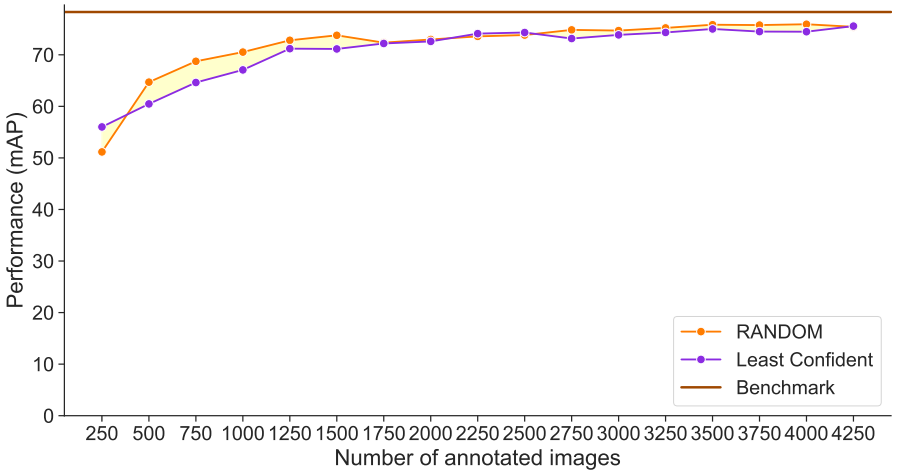


Fig. 7 Mask R-CNN model performance in active learning cycle: scenario 2 with Least Confident and RANDOM baseline

it underperforms the latter in all other iterations. Consequently, in this setting, the Least Confident method struggles to pinpoint the most informative unlabeled samples. The line's concave shape also suggests a diminishing trend in the returns from annotation efforts within scenario 2. The red line represents the benchmark performance of 78.30 mAP derived from the complete dataset. To attain 95% of the benchmark's performance, both methods employ 24.34% of the data.

Entropy

Commencing this section, Fig. 8 showcases the performance observed in scenario 1. For most iterations, the Entropy method falls short of the RANDOM method, implying that the ranking approach through Entropy fares worse than random selection. By the 20th iteration, RANDOM attains a mAP of 93.96, while Entropy achieves 92.35. The outcomes for scenario 2 are depicted in Fig. 9, with a comparable pattern where Entropy lags behind the RANDOM baseline in most iterations. Both the RANDOM baseline and Entropy active learning method reach 95% of the benchmark performance by the 11th iteration, utilizing 19.12% of the data.

Cost effective active learning

The CEAL active learning method introduces three new hyper-parameters: the threshold value, threshold decay, and the selection methods for identifying the most informative images. In previous work, sensitivity analysis by [21] evaluated the threshold and threshold decay values, revealing minimal influence on model performance in their datasets. Nonetheless, their study utilized a CNN model rather than the Mask R-CNN model. For the Mask R-CNN model, we evaluated numerous combinations

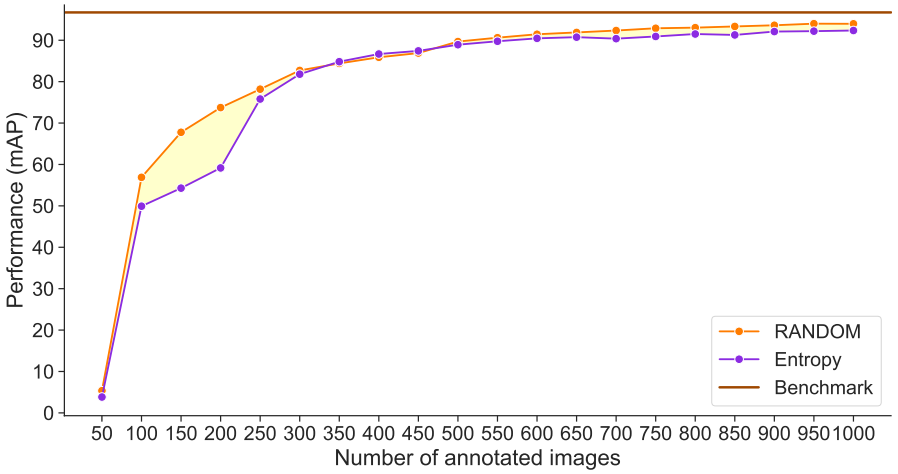


Fig. 8 Mask R-CNN model performance in active learning cycle: scenario 1 with entropy and RANDOM baseline

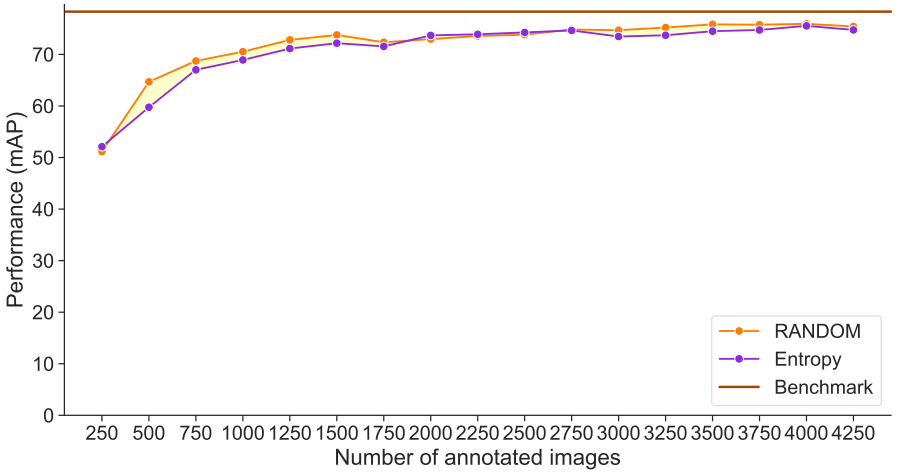


Fig. 9 Mask R-CNN model performance in active learning cycle: scenario 2 with entropy and RANDOM baseline

of these new hyper-parameters on subset 1, keeping other hyper-parameters constant. Tables 9 and 10 in Appendix E detail the outcomes of these assessments. Both scenarios exhibited optimal performance with the Least Confident uncertainty selection method, employing a threshold value of 0.05 and a threshold decay value of 0.0033.

Figure 10 depicts the mAP for scenario 1 across each iteration, while scenario 2 results are presented in Fig. 11. In scenario 1, the CEAL method, utilizing the optimal hyper-parameter combination, demonstrated better performance compared to the RANDOM method starting from the 13^{th} iteration onward. However, prior to

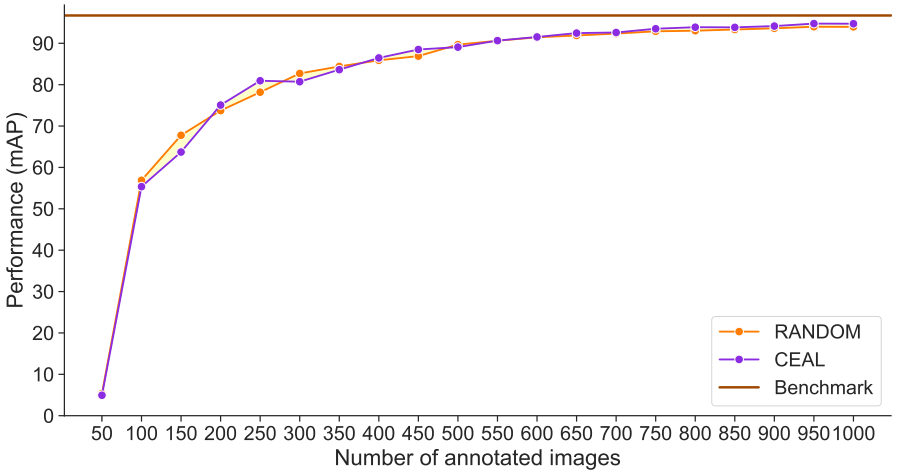


Fig. 10 Mask R-CNN model performance in active learning cycle: scenario 1 with CEAL and RANDOM baseline

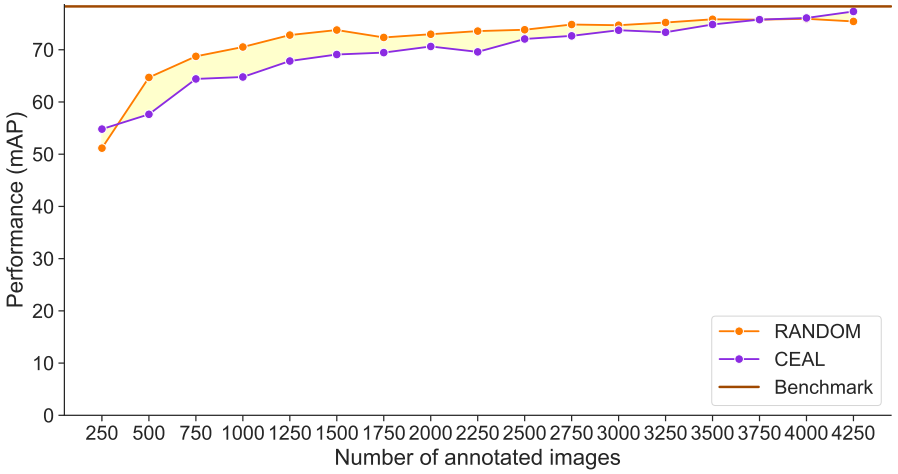


Fig. 11 Mask R-CNN model performance in active learning cycle: scenario 2 with CEAL and RANDOM baseline

this point, no decisive performance difference between the two methods emerged. To achieve 97% of the benchmark performance, the CEAL method required 16 iterations, while the RANDOM method needed 19 iterations. Therefore, the CEAL method reduced annotation needs by 150 images (15.79%) while achieving equivalent performance. In scenario 2, the CEAL method surpassed the RANDOM method only from the 15^{th} iteration onward, with the RANDOM method generally exhibiting better performance before this threshold.

Discussion

This paper aims to address the challenge of resource-intensive annotations in the retail industry. To assess this objective, active learning methods are evaluated as a means to streamline the human annotation effort for CV-based retail product recognition. Firstly, the results highlight a diminishing trend in returns as more images are annotated, underscoring the importance of finding a balance between annotated data and model performance. Secondly, the study reveals differing performance outcomes for active learning methods based on method-scenario combinations, indicating that their effectiveness could be influenced by specific data or model characteristics.

The study's findings reveal the Mask R-CNN model's effective performance with smaller subsets containing only the most informative images. The model achieved 95% of the benchmark performance using 20.83% to 24.34% of the data. Additionally, the study highlights the diminishing returns associated with annotating additional batches of images. While the Least Confident method outperformed the random baseline in the first scenario, it failed to do so in the second scenario. The Entropy-based active learning approach did not surpass the random baseline. In scenario 1, the CEAL method excelled over the RANDOM baseline from the

13th iteration onwards. However, in scenario 2, the CEAL method outperformed the RANDOM baseline only from the 15th iteration onwards, with the RANDOM baseline generally displaying better performance prior to that.

When comparing our findings to existing research, the benchmark performance in scenario 1 aligns with the outcomes observed by [5]. In their study, they achieved top-1 classification performance above 90% for 21 out of 23 classes, suggesting that the Mask R-CNN model is well-suited for smart retail product detection. Our results in scenario 1 follow a similar pattern, with only 2 out of 60 product classes having an mAP below 90. However, scenario 2 introduces a significant distinction to [5]'s conclusion. The benchmark results for scenario 2 indicate that 40 out of 60 product classes achieve an mAP below 90. This disparity underscores that [5]'s conclusion might hold true primarily when the Mask R-CNN model is trained on data closely resembling the inference conditions. Accounting for the industrial demand for simplified data collection and annotation, it becomes evident that the Mask R-CNN model might fall short of meeting the required performance levels.

In a previous study by [21], the combination of a CNN architecture with the CEAL method achieved 95% of the full dataset's performance using 34% to 36% of the data. In our study, employing the Mask R-CNN model, we achieved 95% of the benchmark performance using 20.83% to 24.34% of the data. This suggests that the Mask R-CNN model performs relatively well even with a smaller data subset. However, the reduction in annotation effort observed in our study is more modest compared to [21]. They reported reductions ranging from 15% to 36%, while our results indicate reductions in the range of 7.69% to 15.79% for scenario 1, and no definitive reductions for scenario 2.

The insights gained from this study contribute to a better understanding of active learning methods' applicability in the retail sector. While active learning offers the potential to reduce annotation needs, the results are influenced by the attributes of the data and model. The study reinforces the effectiveness of the Mask R-CNN model for automated product recognition, yet underscores the importance of data alignment between training and inference. This suggests that the model may not fully meet industrial needs for streamlined data collection and annotation. For retail businesses, the practical takeaway is a need to reassess data-collection and annotation strategies, prioritizing data that closely resembles real-world inference scenarios. The observed diminishing returns emphasize the need to strike a balance between annotation effort and model performance to optimize resource allocation.

Future direction

The findings presented within this study have several notable limitations. Firstly, the chosen active learning methods encompass only a subset of the available strategies, suggesting a potential avenue for further exploration to achieve a more comprehensive grasp of active learning within a retail context. Secondly, while the methods employed draw on information derived from the object detection branch, they overlook the potential insights offered by the mask branch. Incorporating mask branch information into the active learning process bears potential to amplify the efficacy of these methods. Thirdly, while the

dataset utilized herein endeavors to emulate a self-checkout scenario, the observed performance might diverge in application to real-world data. Lastly, this study assumes equal annotation effort across all annotated images, neglecting potential variations in annotation complexity that may arise across distinct images.

Further studies can address the limitations and strengthen the validity of these results by extending the scope to a broader spectrum of active learning methodologies, thereby encompassing a more diverse array of potential approaches to determine information gain. Moreover, a promising research direction involves investigating the inclusion of mask-derived uncertainty measures within the active learning framework to fully exploit the capabilities inherent to the Mask R-CNN model architecture. Additionally, to substantiate the outcomes' robustness and applicability within real-world contexts, validation using real-world data would bolster the robustness of the findings. Furthermore, recognizing the potential variance in annotation complexity among images, future endeavors should devise mechanisms to incorporate such intricacies into the active learning process such as resource requirements, annotation cost, ambiguity, image complexity and computational costs. Such research collectively contributes to increasing the comprehension of active learning's efficacy within the retail industry. The resulting insights will hold practical value for retail stores, offering guidance for the integration of active learning methodologies to streamline the image annotation process effectively.

Conclusion

This research addresses a challenge within the retail sector stemming from the dynamic nature of products and packaging. This challenge necessitates frequent retraining of CV models with new data. The acquisition and annotation of such data is time-intensive and therefore costly. Existing research within the retail domain has predominantly centered on establishing the viability of CV technology within retail environments, often not considering strategies aimed at mitigating the annotation overhead for established models. Within the academic framework, this research attempts to bridge this gap by evaluating the performance of the Mask R-CNN model in combination with active learning techniques such as Least Confident, Entropy, and CEAL in mitigating annotation requirements. Thus, the study makes a contribution to the refinement of more streamlined and cost-efficient CV solutions, tailored to the requisites of the retail industry.

The implications drawn from this paper provide insights into the dynamic between automatic product segmentation and active learning within the realm of CV, particularly as applied to the retail industry. From a societal perspective, the implications in this study show the capacity to strategically redistribute annotation efforts, redirecting resources from relatively inconsequential annotation tasks to those of more import. This shift enhances the overall efficiency of CV model training. For retail enterprises, the implications underscore the need for a reevaluation of prevailing data collection and annotation practices. The diminishing return in performance from each successive batch of images emphasizes the need to strike a balance between annotated data volume and model efficacy. The incorporation of active

learning methodologies introduces an avenue for substantial reduction in annotation prerequisites. However, the efficacy of active learning methods is contingent upon the characteristics of the training data or the active learning method used.

Appendix A (see Tables 3, 4)

Table 3 The 5 best performing combinations of hyper-parameters tested on the validation set for scenario 1

Backbone	Epochs	Learning rate	Decay step	ROI heads	mAP
Resnet101	50	0.001	0.75	512	96.6
Resnet101	50	0.001	0.75	128	96.07
Resnet101	35	0.001	0.75	128	95.96
Resnet101	10	0.001	0.75	512	95.61
Resnet101	25	0.001	0.75	128	95.61

Table 4 The 5 best performing combinations of hyper-parameters tested on the validation set for scenario 2

Backbone	Epochs	Learning rate	Decay step	ROI heads	mAP
Resnet101	20	0.001	0.75	512	81.24
Resnet101	35	0.001	0.75	512	80.53
Resnet101	20	0.001	0.75	128	80.42
Resnet101	50	0.001	0.75	512	80.25
Resnet101	15	0.001	0.75	512	80.16

Appendix B (see Tables 5, 6; Figs. 12, 13)

Table 5 Precision, Recall, F1-score and Support for subset 1

Metric	Precision	Recall	f1-score	Support
Micro avg	0.99	1	1	3169
Macro avg	0.99	1	1	3169
Weighted avg	0.99	1	1	3169

Table 6 Precision, recall, F1-score and support for subset 2

Metric	Precision	Recall	f1-score	Support
Micro avg	0.75	0.89	0.81	3169
Macro avg	0.8	0.9	0.82	3169
Weighted avg	0.83	0.89	0.84	3169

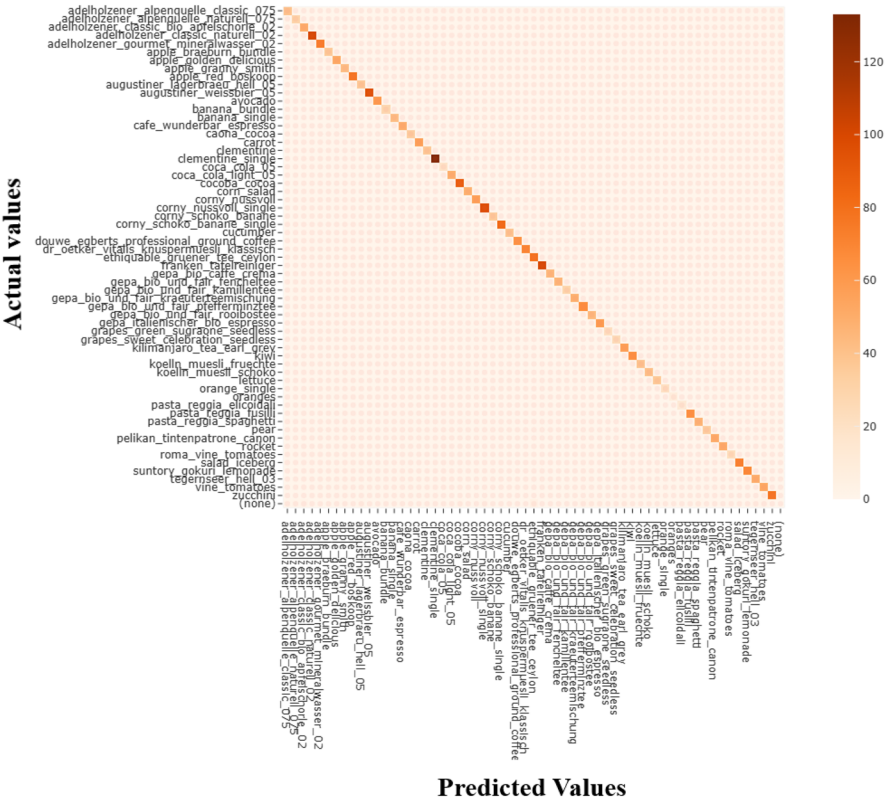


Fig. 12 Confusion matrix for each product class in subset 1

Appendix C

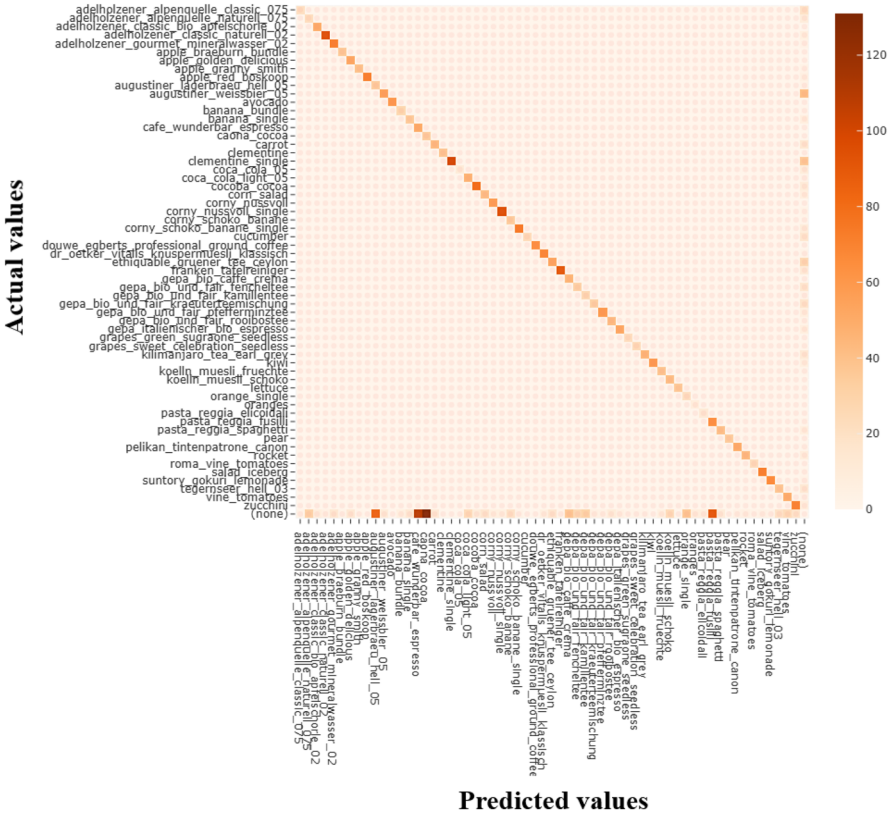


Fig. 13 Confusion matrix for each product Class in subset 2

Appendix D (see Tables 7, 8)

Table 7 Per class AP for scenario 1

Product class	AP	Product class	AP
Adelholzener alpenquelle classic 075	99.41	Banana single	89.27
Adelholzener alpenquelle naturell 075	99.37	Clementine	98.75
Adelholzener classic bio apfelschorle 02	99.61	Clementine single	96.05
Adelholzener classic naturell 02	96.42	Grapes green sugraone seedless	98.97
Adelholzener gourmet mineralwasser 02	97.33	Grapes sweet celebration seedless	98.63
Augustiner lagerbraeu hell 05	97.69	Kiwi	99.19
Augustiner weissbier 05	99.57	Orange single	99.30
Coca cola 05	100.00	Oranges	100.00
Coca cola light 05	100.00	Pear	96.59
Suntory gokuri lemonade	99.46	Pasta reggia elicoidali	99.11
Tegernseer hell 03	98.93	Pasta reggia fusilli	97.51
Corny nussvoll	99.45	Pasta reggia spaghetti	95.34
Corny nussvoll single	92.47	Franken tafelreiniger	95.00
Corny schoko banane	96.42	Pelikan tintenpatrone canon	98.53
Corny schoko banane single	90.96	Ethiquable gruener tee ceylon	92.18
Dr oetker vitalis knuspermuesli klassisch	95.17	Gepa bio und fair fencheltee	98.68
Koelln muesli fruechte	99.54	Gepa bio und fair kamillentee	97.51
Koelln muesli schoko	97.77	Gepa bio und fair kraeuterteemischung	96.18
Caona cocoa	93.47	Gepa bio und fair pfefferminztee	98.93
Cocoba cocoa	90.84	Gepa bio und fair rooibostee	99.50
Cafe wunderbar espresso	98.38	Kilimanjaro tea earl grey	91.13
Douwe egberts professional ground coffee	98.59	Cucumber	91.82
Gepa bio caffe crema	100.00	Carrot	87.37
Gepa italienischer bio espresso	96.31	Corn salad	94.94
Apple braeburn bundle	100.00	Lettuce	95.22
Apple golden delicious	98.55	Vine tomatoes	90.06
Apple granny smith	97.59	Roma vine tomatoes	91.01
Apple red boskoop	100.00	Rocket	98.78
Avocado	99.70	Salad iceberg	99.75
Banana bundle	93.97	Zucchini	90.67

Table 8 Per class AP for scenario 2

Product class	AP	Product class	AP
Adelholzener alpenquelle classic 075	52.62	Banana single	67.42
Adelholzener alpenquelle naturell 075	45.07	Clementine	94.60
Adelholzener classic bio apfelschorle 02	95.16	Clementine single	69.83
Adelholzener classic naturell 02	86.31	Grapes green sugraone seedless	91.34
Adelholzener gourmet mineralwasser 02	85.62	Grapes sweet celebration seedless	86.51
Augustiner lagerbraeu hell 05	53.49	Kiwi	87.67
Augustiner weissbier 05	52.27	Orange single	40.87
Coca cola 05	51.72	Oranges	96.24
Coca cola light 05	71.39	Pear	91.49
Suntory gokuri lemonade	91.07	Pasta reggia elicoidali	97.28
Tegernseer hell 03	67.79	Pasta reggia fusilli	87.30
Corny nussvoll	97.39	Pasta reggia spaghetti	78.60
Corny nussvoll single	86.97	Franken tafelreiniger	82.86
Corny schoko banane	79.01	Pelikan tintenpatrone canon	91.10
Corny schoko banane single	77.70	Ethiquable gruener tee ceylon	50.34
Dr oetker vitalis knuspermuesli klassisch	85.59	Gepa bio und fair fencheltee	55.57
Koelln muesli fruechte	92.75	Gepa bio und fair kamillentee	73.08
Koelln muesli schoko	92.76	Gepa bio und fair kraeuterteemischung	56.96
Caona cocoa	85.85	Gepa bio und fair pfefferminztee	78.53
Cocoba cocoa	65.57	Gepa bio und fair rooibostee	92.30
Cafe wunderbar espresso	93.79	Kilimanjaro tea earl grey	54.51
Douwe egberts professional ground coffee	93.89	Cucumber	53.10
Gepa bio caffe crema	98.33	Carrot	60.67
Gepa italienischer bio espresso	77.06	Corn salad	69.67
Apple braeburn bundle	95.97	Lettuce	94.71
Apple golden delicious	98.62	Vine tomatoes	65.79
Apple granny smith	88.18	Roma vine tomatoes	70.85
Apple red boskoop	91.69	Rocket	75.80
Avocado	99.00	Salad iceberg	89.78
Banana bundle	73.58	Zucchini	66.74

Appendix E (see Tables 9, 10)

Table 9 Hyper-parameter optimization on subset 1 for new parameters introduced by the CEAL method. The new hyper-parameters are: Uncertainty selection method, Threshold Value, Threshold Decay Value

Training iteration / percentage of total samples annotated	1	3	5	7	9	11	13	15	17	19
Method - Uncertainty selection - Threshold - Decay Value	1.74%	5.21%	8.68%	12.15%	15.63%	19.10%	22.57%	26.04%	29.51%	32.99%
Ceal - Least Confident - Threshold: 0.05 - Decay: 0.0033	7.07	63.35	83.57	88.26	92.05	93.30	93.96	94.34	94.67	95.24
Ceal - Least Confident - Threshold: 0.05 - Decay: 0.0015	2.21	63.55	84.33	88.05	90.02	91.78	93.54	94.18	94.78	94.81
Ceal - Least Confident - Threshold: 0.05 - Decay: 0.0025	7.10	66.92	84.58	89.24	90.68	92.62	93.62	93.91	94.09	94.63
Ceal - Least Confident - Threshold: 0.10 - Decay: 0.0033	4.24	62.30	81.47	86.40	88.39	89.96	92.72	93.14	93.68	94.27
Ceal - Least Confident - Threshold: 0.10 - Decay: 0.0025	5.49	62.11	79.29	84.03	87.38	89.62	91.53	92.56	92.61	94.26
Ceal - Entropy - Threshold: 0.05 - Decay: 0.0033	1.72	61.71	76.97	83.35	87.33	88.73	90.27	91.58	92.06	92.69
Ceal - Entropy - Threshold: 0.05 - Decay: 0.0025	4.72	65.25	78.79	82.49	86.82	87.44	87.70	89.43	91.36	92.67
Ceal - Entropy - Threshold: 0.05 - Decay: 0.0015	7.89	64.41	83.05	83.65	84.99	89.50	91.62	88.86	92.05	91.95
Ceal - Entropy - Threshold: 0.10 - Decay: 0.0033	5.19	54.31	78.40	82.69	86.47	87.24	88.86	90.38	90.24	90.27
Ceal - Entropy - Threshold: 0.15 - Decay: 0.0033	3.53	55.89	74.24	78.91	83.92	82.47	84.46	86.18	87.21	87.09
Ceal - Entropy - Threshold: 0.20 - Decay: 0.0033	7.39	36.58	66.86	75.36	76.84	77.87	82.70	84.08	86.65	86.61

Table 10 Hyper-parameter optimization on subset 1 for new parameters introduced by the CEAL method. The new hyper-parameters are: Uncertainty selection method, Threshold Value, Threshold Decay Value

Training iteration / percentage of total samples annotated	1	3	5	7	9	11	13	15	16
Method - Uncertainty selection - Threshold - Decay Value	1.74%	5.22%	8.69%	12.17%	15.65%	19.12%	22.60%	26.08%	27.82%
CEAL - Least Confident - Threshold: 0.05 - Decay: 0.0033	57.13	67.02	72.95	73.69	74.56	75.26	75.43	75.61	78.11
CEAL - Least Confident - Threshold: 0.05 - Decay: 0.0015	55.24	65.81	68.44	72.45	72.17	75.04	75.77	75.46	76.32
CEAL - Least Confident - Threshold: 0.05 - Decay: 0.0033	58.17	55.69	66.06	66.98	69.47	71.25	73.01	75.20	76.05
CEAL - Least Confident - Threshold: 0.05 - Decay: 0.0025	55.85	56.85	63.13	68.78	70.49	71.99	71.49	73.23	73.98

Funding The authors received no external funding.

Data availability Availability of data and materials: The MVTec D2S: Densely Segmented Supermarket Dataset is publicly available and could be downloaded through the paper by Follmann P. (2018) et al. https://doi.org/10.1007/978-3-030-01249-6_35...

Code availability All authors have read and agreed to the published version of the manuscript.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication All authors have read and agreed to the published version of the manuscript.

References

1. Wei, Y., Tran, S., Xu, S., Kang, B., & Springer, M. (2020). Deep learning for retail product recognition: Challenges and techniques. *Computational Intelligence and Neuroscience*. 2020: 1–23. https://doi.org/10.1155/2020/8875910
2. Wei, X.-S., Cui, Q., Yang, L., Wang, P., Liu, L., & Yang, J. (2022). RPC: a large-scale and fine-grained retail product checkout dataset. *Science China Information Sciences*. https://doi.org/10.1007/s11432-022-3513-y
3. Kovashka, A., Russakovsky, O., Fei-Fei, L., & Grauman, K. (2016). Crowdsourcing in computer vision. Foundations and Trends. *Computer Graphics and Vision, 10*(3), 177–243. https://doi.org/10.1561/0600000071
4. Santra, B., & Mukherjee, D. P. (2019). A comprehensive survey on computer vision based approaches for automatic identification of products in retail store. *Image and Vision Computing., 86*, 45–63. https://doi.org/10.1016/j.imavis.2019.03.005
5. Hsia, C.-H., Chang, T.-H.W., Chiang, C.-Y., & Chan, H.-T. (2022). Mask r-CNN with new data augmentation features for smart detection of retail products. *Applied Sciences, 12*(6), 2902. https://doi.org/10.3390/app12062902
6. Bartl, V., Spanhel, J., & Herout, A. (2022). PersonGONE: Image inpainting for automated checkout solution. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3115–3123. https://doi.org/10.1109/cvprw56347.2022.00351
7. Fuchs, K., Grundmann, T., & Fleisch, E. (2019). Towards identification of packaged products via computer vision. In: Proceedings of the 9th International Conference on the Internet of Things. https://doi.org/10.1145/3365871.3365899
8. Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. https://doi.org/10.1109/5.726791

9. Zhao, Z.-Q., Zheng, P., Xu, S.-T., & Wu, X. (2019). Object detection with deep learning: A review. *IEEE Transactions on Neural Networks and Learning Systems, 30*(11), 3212–3232. https://doi.org/10.1109/tnnls.2018.2876865
10. Uijlings, J. R. R., Sande, K. E. A., Gevers, T., & Smeulders, A. W. M. (2013). Selective search for object recognition. *International Journal of Computer Vision, 104*(2), 154–171. https://doi.org/10.1007/s11263-013-0620-5
11. Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster r-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(6), 1137–1149. https://doi.org/10.1109/tpami.2016.2577031
12. He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask r-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2961–2969. https://doi.org/10.1109/iccv.2017.322
13. Follmann, P., Böttger, T., Härtinger, P., König, R., & Ulrich, M. (2018). MVTec d2s: Densely segmented supermarket dataset. In: Computer Vision – ECCV 2018. pp. 581–597. https://doi.org/10.1007/978-3-030-01249-6_35
14. Shelhamer, E., Long, J., & Darrell, T. (2017). Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 39*(4), 640–651. https://doi.org/10.1109/tpami.2016.2572683
15. Bai, M., & Urtasun, R. (2017). Deep watershed transform for instance segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5221–5229. https://doi.org/10.1109/cvpr.2017.305
16. Chen, S., Liu, D., Pu, Y., & Zhong, Y. (2022). Advances in deep learning-based image recognition of product packaging. *Image and Vision Computing., 128*, 104571.
17. Hameed, K., Chai, D., & Rassau, A. (2021). Class distribution-aware adaptive margins and cluster embedding for classification of fruit and vegetables at supermarket self-checkouts. *Neurocomputing, 461*, 292–309.
18. Karlinsky, L., Shtok, J., Tzur, Y., & Tzadok, A. (2017). Fine-grained recognition of thousands of object categories with single-example training. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2017.109
19. Budd, S., Robinson, E. C., & Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis, 71*: 102062.
20. Settles, B. (2012). *Active Learning*. https://doi.org/10.1007/978-3-031-01560-1
21. Wang, K., Zhang, D., Li, Y., Zhang, R., & Lin, L. (2017). Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology, 27*(12), 2591–2600. https://doi.org/10.1109/tcsvt.2016.2589879
22. Kirsch, A., Van Amersfoort, J., Gal, Y. (2019). Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in Neural Information Processing Systems*, *32*
23. Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., & Wang, X. (2021). A survey of deep active learning. *ACM computing surveys (CSUR), 54*(9), 1–40.
24. Kim, K., Park, D., Kim, K.I., & Chun, S.Y. (2021). Task-aware variational adversarial active learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8166–8175
25. Citovsky, G., DeSalvo, G., Gentile, C., Karydas, L., Rajagopalan, A., Rostamizadeh, A., & Kumar, S. (2021). Batch active learning at scale. *Advances in Neural Information Processing Systems, 34*, 11933–11944.
26. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C.L. (2014). Microsoft COCO: Common objects in context. In: Computer Vision – ECCV 2014. pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
27. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., & Schiele, B. (2016). The cityscapes dataset for semantic urban scene understanding. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/cvpr.2016.350
28. Culotta, A., & McCallum, A. (jan 2005). Reducing labeling effort for structured prediction tasks. Technical Report. https://doi.org/10.21236/ada440382
29. Settles, B., & Craven, M. (2008). An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing–EMNLP '08. https://doi.org/10.3115/1613715.1613855

30. Lafferty, J.D., McCallum, A., & Pereira, F.C.N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the Eighteenth International Conference on Machine Learning. ICML '01, pp. 282–289. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
31. Everingham, M., Eslami, S. M. A., Gool, L. V., Williams, C. K. I., Winn, J., & Zisserman, A. (2014). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision, 111*(1), 98–136. https://doi.org/10.1007/s11263-014-0733-5
32. Hu, T., Deng, Y., Deng, Y., & Ge, A. (2021). Fully convolutional network variations and method on small dataset. In: 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE). pp. 40–46. https://doi.org/10.1109/iccece51280.2021.9342059
33. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. https://doi.org/10.1109/cvpr.2016.90

Authors and Affiliations

Niels Griffioen[1] · Nevena Rankovic[1] · Federico Zamberlan[2] · Monisha Punith[3]

✉ Nevena Rankovic
n.rankovic@uvt.nl

Niels Griffioen
njcgriffioen@gmail.com

Federico Zamberlan
f.zamberlan@uvt.nl

Monisha Punith
Monisha.Punith@uantwerpen.be

[1] Department of Cognitive Science and Artificial Intelligence, Tilburg University, Waraandelan, 5037 AB Tilburg, North-Brabant, Netherlands

[2] Departamento de Física, Universidad de Buenos Aires, Intendente Güiraldes 2160, C1428EGA Buenos Aires, Argentina

[3] Department of economics, University of Antwerp, Prinsstraat 13, 2000 Antwerp, Flanders, Belgium