Unsupervised Annotation and Detection of Novel Objects Using Known Objectness

Harsh Singh Jadon*, Jagdish Deshmukh*, Kamakshya Prasad Nayak[a],
Kamalakar Vijay Thakare[b] and Debi Prosad Dogra†[c]
Indian Institute of Technology Bhubaneswar, Odisha, PIN 752050, India

Keywords: Novel Object Detection, Unsupervised Learning, Weakly Annotated Dataset, Clustering.

Abstract: The paper proposes a new approach to detecting and annotating novel objects in images that are not precisely part of a training dataset. The ability to detect novel objects is essential in computer vision, enabling machines to recognise objects that have not been seen before. Current models often fail to detect novel objects as they rely on predefined categories in the training data. Our approach overcomes this limitation by leveraging a large and diverse dataset of objects obtained through web scraping. We extract features using a backbone network and perform clustering to remove redundant data. The resulting dataset is used to retrain the object detection models to obtain results. The method provides deep insights into the effect of clustering and data redundancy removal on performance. Overall, the work contributes to the field of object detection by providing a new approach for detecting novel objects. The method has the potential to be applied to a variety of real-world CV applications.

1 INTRODUCTION

Region proposals are vital building blocks in object detection tasks. An improved pre-processing of proposals usually leads to significant improvements in various applications, such as object recognition (Li et al., 2022b; Li et al., 2022a; Yang et al., 2022; Cheng et al., 2022; Hou et al., 2022), video tracking (Qin et al., 2022; Tang and Ling, 2022; Blatter et al., 2023), object discovery (Hénaff et al., 2022; Wang et al., 2022; Bao et al., 2022) and segmentation (Park et al., 2022; Xu et al., 2022). Due to such advancement, object detection using region proposals has attracted tremendous attention. Such tasks aim to generate region proposals and predict the labels with acceptable threshold. However, one major limitation of current object detection models is their inability to detect novel objects that are not present in the training data. This limitation hinders the ability of the models to recognize unknown objects, which may be a critical requirement for various real-world applications.

Several attempts (LaBonte et al., 2023; Xu et al., 2020; Wu et al., 2021; Zhu et al., 2021; Fan et al., 2020) have been made in recent years to detect novel objects. Kuo et al. (Kuo et al., 2015) have developed DeepBox that generates bottom-up proposals and reranks them using CNNs, given any RGB image. Kang et al. (Kang et al., 2019) have designed a novel few-shot detection model that 1) learns generalized meta-features and 2) automatically re-weights the features for novel class detection by producing class-specific activating coefficients from a few support samples. However, unlike the learning-free counterparts (Bao et al., 2022; Hénaff et al., 2022; Kuo et al., 2015), these methods tend to over-fit annotated categories and struggle with novel objects.

To overcome this limitation, this paper proposes a new approach to detect novel objects in images that are not precisely part of a training dataset. Our approach leverages a large and diverse dataset of objects obtained through web scraping. It enables the detection of novel objects with higher accuracy as compared to recent methods. We extract features using well-known backbone networks and perform clustering to remove redundant data. The resulting dataset is then used to retrain the object detection model to obtain final results. Existing models rely on predefined categories and context-free learning in the training

[a] https://orcid.org/0000-0002-0356-8377
[b] https://orcid.org/0000-0003-4587-4126
[c] https://orcid.org/0000-0002-3904-732X
*These authors contributed equally to this work
†Corresponding author

Jadon, H., Deshmukh, J., Nayak, K., Thakare, K. and Dograψ, D.
Unsupervised Annotation and Detection of Novel Objects Using Known Objectness.
DOI: 10.5220/0012412400003660
Paper published under CC license (CC BY-NC-ND 4.0)
In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 2: VISAPP*, pages 694-701
ISBN: 978-989-758-679-8; ISSN: 2184-4321

data and they often fail to detect novel objects. Our approach has been evaluated on several benchmarks, demonstrating its effectiveness in detecting novel objects with high accuracy. Moreover, we provide insights into the effect of clustering and data redundancy removal on the performance of the proposed approach. The paper offers the following technical contributions:

- We propose a systematic, easy-to-follow approach to accumulate relevant data samples for novel object discovery. With sufficient related keywords, the proposed approach can generate a large pool of auto-annotated images that can be used for training object detectors.
- We comprehensively evaluate pre-trained object detection models on a newly generated dataset. This evaluation encompasses various aspects such as threshold tuning, accuracy on novel classes, and the effect of the number of images on the model's accuracy.

The rest of the paper is organized as follows. Section 2 provides a review of related work in the field of object detection and detection of novel objects. Section 3 describes the proposed approach in detail, including data collection, feature extraction, clustering, and retraining. Section 4 presents the experimental setup and results. Section 5 discusses the proposed approach's results and limitations. Finally, Section 6 concludes the paper and discusses future work.

2 RELATED WORK

2.1 Object Detection

Traditionally, object detection algorithms relied heavily on handcrafted features. However, the advent of RCNNs, proposed by Girshick et al. (Girshick, 2015) has significantly boosted the progress of object detection. Subsequently, Faster-RCNNs and Fast-RCNNs have been introduced, which have improved the RCNN models by jointly training a detector and a bounding box regressor within the same network configuration. Based on Faster RCNNs (Ren et al., 2015), the authors have proposed Feature Pyramid Networks (FPN). The FPN architecture uses a top-down approach with lateral connections to generate high-level semantics at all scales. FPN has exhibited significant progress in detecting objects at varying scales. Another improvement in object detection is Mask RCNN proposed by He et al. (He et al., 2017), which incorporates pixel-level masks into Faster-RCNN. However, majority of the object detection algorithms demand fully annotated data for an object class they aim to detect. With the remarkable progress in object detection, the focus of recent object proposal research has transitioned from object discovery to detection. Object discovery proposals aim to propose all objects in an image, whereas detection proposals are designed to propose only the labelled categories for downstream classification.

2.2 Class Agnostic Object Detection

Class-agnostic object detection is a subfield of computer vision that aims to detect objects without relying on pre-defined classes. It focuses on detecting and localizing objects in an image regardless of their category. Class-agnostic object detection has numerous applications, including object tracking, scene understanding, and robotics. Recent advancements in deep learning models, such as one-stage detectors has significantly improved the accuracy and speed of class-agnostic object detection. However, this field still has several challenges, including object occlusion, scale variations, and background clutter. Previously, class-agnostic object detection has been tackled using traditional methods like Selective Search (Uijlings et al., 2013), EdgeBox (Zitnick and Dollar, 2014), DeepMask (O. Pinheiro et al., 2015), and MCG (Pont-Tuset et al., 2017). However, recent advancements such as the Object Localisation Network (Kim et al., 2021) and Multiscale Attention Vision Transformer with Late Fusion (Maaz et al., 2022) have emerged in this field. While existing approaches have made significant advancements, they still have some limitations that need to be addressed such as:

Limited Generalization to Novel Objects. Pre-trained object detection models often struggle to generalize well to novel objects not present in their training dataset. When faced with unseen object categories, these models may exhibit reduced accuracy and reliability, hindering their performance in real-world scenarios where object variability is high.

Dependency on Annotated Training Data. Most state-of-the-art object detection models rely heavily on large-scale annotated training data. The manual annotation process is time-consuming, expensive, and subject to human error, limiting the scalability and accessibility of these approaches.

3 METHOD

The pipeline of the proposed approach is summarized in Fig. 1. It consists of three stages: Accumulating novel object dataset using an image search engine,

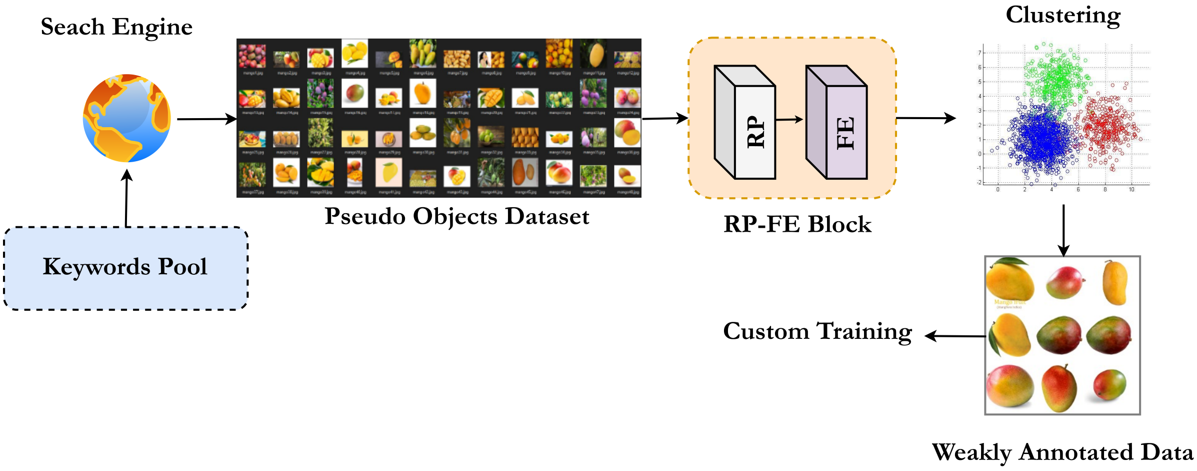

Figure 1: The architecture of the proposed framework. It has three important stages: (i) Extraction of relevant object dataset using Web Scrapper tool leveraging a large set of keyword pools. (ii) RP-FE block: Extract region proposals using class-agnostic object detectors and feed them to the backbone network for feature extraction. (iii) Utilize clustering to remove redundant image samples and use the newly generated dataset to custom training of the existing object detection model.

RP-FE block, and refinement of most relevant data using clustering followed by custom training of detection model with new data. In the following sections, we have explained each stage in detail.

3.1 Dataset Generation

The critical component of a typical object detection method is obtaining the labelled training data that contains examples of well-known object classes. However, there is an absence of labelled data accessible for training regarding new things. It is challenging to build supervised learning techniques for identifying new objects due to the scarcity of data. Moreover, generating labelled data is a tedious task.

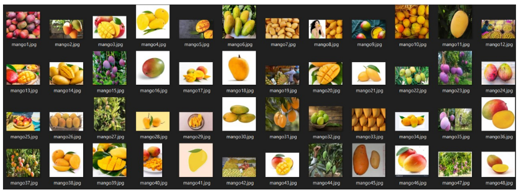

Figure 2: A depiction of a few image samples retrieved using the Web Scraping. The keywords are, *mango*, *ripped mango*, *mango with white background*, and so on.

To mitigate the aforementioned problem, we have employed a Web Scraper. This tool directly searches the World Wide Web based on a search keyword using the Hypertext Transfer Protocol and extracts image files from embedded HTML of the web page. We have employed a fully automated tool that retrieves non-copyright images from websites based on a particular keyword. Thus, we can obtain *novel* objects data as it is directly retrieved from the internet without supervision. Fig. 2 depicts a collection of image samples for the object *mango*.

Let the set of n keywords (similar) for an unknown or novel object be given in (1). For example, if $\mathbf{K}$ is the set of keywords related to *mango*, then k_1 = ripped mango, k_2 = green mango, k_3 = mango with knife, etc. We have employed various keywords for individual objects based on natural observations, brainstorming sessions, and suggestions from search engines. For instance, *mango with leaf*, *tennis ball with racket*, etc., have been used to search images on web.

$$\mathbf{K} = \{k_1, k_2, ..., k_n\} \tag{1}$$

We also assume $\mathbf{C} = \{C_1, C_2, ..., C_n\}$ be the set of object images retrieved from the web using the keyword set $\mathbf{K}$. The dataset $\mathbf{C}$ is an extensive collection of non-copyrighted images related to that particular object. The images of the dataset are then processed through an existing object detection model as discussed in the next section.

3.2 RP-FE Block

The approach uses a class-agnostic object detection model to detect all potential objects present in all the images of the dataset. By employing a class-agnostic model, we are able to capture a wide range of objects without being limited to a predefined class of object. The detection process generates region-of-interest proposals (RP) that encompass the spatial locations of these objects. Subsequently, we extract discriminative features (FE) from these object proposals, enabling us to capture meaningful information specific to each object.

Region Pooling (RP) Extraction. Let M be the class-agnostic object detection model with pre-trained weights $\mathbf{W_M}$ as given in (2), where p_M represents the parameters.

$$M = (\mathbf{W_M}, p_M) \tag{2}$$

Also, we assume that $F = (\mathbf{W_F}, p_F)$ is a backbone network that extracts features from the given image. We first employ a class-agnostic object detector M and obtain the set of bounding boxes $\mathbf{B} = \{B_1, B_2, \ldots, B_n\}$, where B_i is the set of coordinates of objects present in C_i. Since $\mathbf{C}$ is a collection of novel object images, we need a refinement strategy to filter the most relevant data that accurately represent K. To accomplish this, we have employed well-known clustering techniques and obtained the curated data.

Feature Extraction (FE). First, we extract features by employing a set of backbone networks, $\mathbf{F} = \{F_1, F_2, F_3\}$. This extraction is shown in (3), where D_1 features are extracted from the image set C_1 using $i-th$ network with pre-trained weights W_F.

$$D_1 = F_i(B_1, \mathbf{W_{Fi}}) \tag{3}$$

Here, F_1,F_2 and F_3 are popular ResNet-50, InceptionV3, and EfficientNet. In the later stage, $D = \{D_1, D_2, \ldots, D_n\}$ is the set of extracted features fed to the clustering algorithm. The clustering step is important in eliminating redundant data and organizing the extracted features into distinct groups or clusters. This process helps us to identify and retain the novel object's most relevant and representative data samples. Since the unlabelled images are obtained through keyword searching, clustering enables us to group similar features, thereby facilitating the selection of the most meaningful and discriminative representations for further analysis. To capture the most discriminative features with minimal intra-class variations, we have selected the smallest cluster for experimentation.

3.3 Custom Training

Once we obtain the most relevant data through clustering, we create a new weakly labelled dataset. The dataset is the foundation for the custom training of an existing object detection model. The objective is to fine-tune an existing trained model using the weakly generated dataset, focusing on the targeted class of objects. During the class-agnostic detection and ROI pooling stages, we not only detect objects but also preserve their spatial locations within the image. This spatial information is important for accurate annotation preparation and training. By maintaining the spatial context of each object, we ensure that the model learns to recognize and localize the targeted class effectively.

Training. The annotation process involves labelling the weakly generated dataset and providing class-specific annotations for the targeted objects. These annotations serve as ground truth labels for training the pre-trained object detection model and refining its ability to detect and classify the desired class of objects accurately. To learn the parameters of the model, we have incorporated IoU loss ($\mathcal{L}_{\text{IoU}}$), which is given in (4), where G_T is a set of four coordinates as ground truth labels of the objects and P is set of coordinates predicted by the model.

$$\mathcal{L}_{\text{IoU}} = -\ln \frac{(G_T \cap P)}{(G_T \cup P)} \tag{4}$$

This IoU loss helps the model to learn intrinsic features resulting better localization results.

Evaluation. We have comprehensively evaluated the pre-trained object detection models using a newly available dataset. As the newly acquired dataset is utilized for training purposes, we have derived the test set from this newly constructed data. To assess the performance of the models, we have designed and conducted four distinct experiments. Firstly, we have performed threshold tuning, systematically varying the threshold values to investigate their impact on the models' detection accuracy. Secondly, we have evaluated the models' accuracy on novel classes, which are not included in the original training set. This has allowed us to measure the ability to generalize to unseen object categories. Thirdly, after fine-tuning, we have quantified the accuracy improvement achieved by the pre-trained models on these novel classes. This analysis has provided us insights into the transferability of the learned representations. Finally, we have explored the effect of varying the number of training images on the models' performance, investigating how the scale of the dataset influenced their detection capabilities.

4 EXPERIMENTS AND RESULTS

In this section, we present experimental results for the following four categories: (i) Threshold tuning, (ii) accuracy on novel classes, iii) Accuracy on existing classes, and iv) effect of number of training samples.

4.1 Implementation Details

We have utilised two class agnostic objection detection models: OLN (Kim et al., 2021) and MAVL (Maaz et al., 2022). OLN (Kim et al.,

2021) is a two-stage object proposer similar to Faster R-CNN (Ren et al., 2015). It consists of a fully convolutional FCN and a region-based ROI stage followed by locality predictions. Meanwhile, MAVL (Maaz et al., 2022) has been trained on LMDet (Peng et al., 2020), a large-scale object detection dataset with 478,000 images and 10.5 millions of object instances. For the feature extraction phase of the pipeline, we have considered three feature extraction models: ResNet50, InceptionV3 and EfficientNetB0. These models have been selected due to their excellent performance on image classification. We have used the implementations available with ScikitLearn's for KMeans and DBSCAN clustering approaches. In addition, we have trained and evaluated YOLOv7, currently the best model in the YOLO family.

Evaluation Metrics. We have used **recall (R)**, **precision (P)**, and **mAP** (mean Average Precision) metrics for evaluation. Recall measures the ability to identify all relevant instances. Precision evaluates the accuracy of the detected instances. And mAP assesses the overall detection performance by considering the average precision across all classes or categories. These metrics help evaluate the model's performance, offering insights into its object detection capabilities and accuracy.

4.2 Threshold Tuning

Threshold tuning is very important in object detection as it allows for adjusting the detection sensitivity. By selecting an appropriate threshold, the trade-off between false positives and false negatives can be controlled, thereby influencing the overall precision and recall of the system. Optimal threshold tuning ensures that the model balances accurately, detecting objects and minimizing erroneous detections, leading to improved performance and more reliable results. We have conducted several experiments using different thresholds on OLN (Kim et al., 2021) and MAVL (Maaz et al., 2022) models. Tab. 1 represents the values of different accuracy metrics on varying threshold values.

4.3 Experiments on Pre-Trained Classes

Even though the COCO dataset includes the sports ball class and YOLOv7 has been trained on it, the model often identifies the tennis balls as instances of oranges instead, resulting in a reduced precision for the testing partition. The prediction for the tennis ball has been shown in Fig. 5. We have utilized the OLN, ResNet50, and K-means clustering with the search query "tennis ball" to create a tennis ball dataset with the instances of tennis balls labelled as sports balls. We created two partitions, one for training and the other for testing. Tab 2 and 3 present the prediction results related to improvement in accuracy when the model is retrained with the proposed approach. As we can see, the accuracy gain is higher for *sports ball* as the number of samples for the tennis ball is low in the COCO dataset and often gets predicted as orange. After retraining the model with a new dataset, the model successfully detected tennis balls. However, it is not

Table 1: Accuracy metric values at various thresholds for the OLN (Kim et al., 2021) and MAVL (Maaz et al., 2022) model.

Threshold Tuning					
Stage-I		**Accuracy Metrics**			
Model	Threshold	P	R	mAP@.5	mAP@.5:.95
OLN	0.5	0.819	0.964	0.793	0.664
OLN	0.6	0.906	0.96	0.902	0.74
OLN	0.66	0.96	0.96	0.941	0.78
OLN	0.7	0.96	0.938	0.937	0.755
OLN	0.75	0.963	0.88	0.921	0.728
MAVL	0.7	0.41	0.81	0.55	0.389
MAVL	0.75	0.44	0.809	0.6	0.44
MAVL	0.8	0.49	0.806	0.613	0.471
MAVL	0.85	0.49	0.756	0.606	0.436
MAVL	0.9	0.49	0.72	0.577	0.38

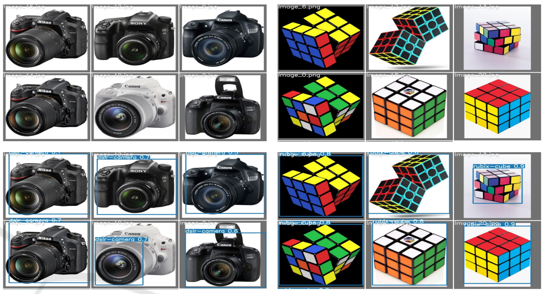

Figure 3: Object detection results using YOLOv7 on novel category *DSLR camera* and *Rubic cube*. The top row is the YOLOv7 prediction before the custom training and the bottom row presents the prediction of the retrained model with the proposed approach.

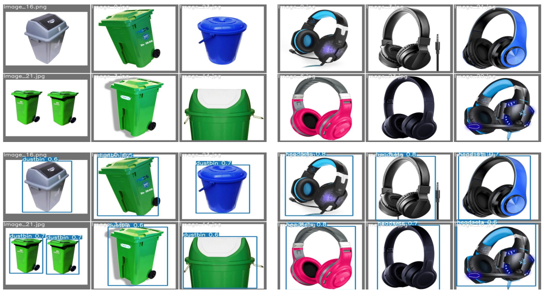

Figure 4: Object detection results using YOLOv7 on two novel categories e.g. *dustbin* and *Headphones*. The top row is the YOLOv7 prediction before custom training and the bottom is the prediction of the retrained model with the proposed approach.

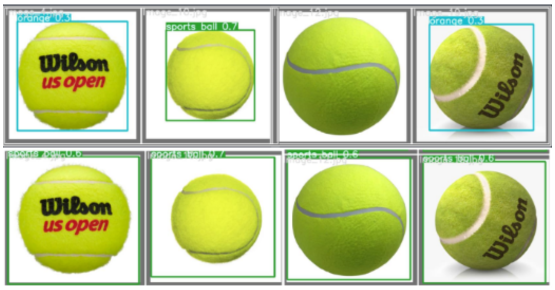

Figure 5: Top row: tennis ball is part of the COCO dataset, but YOLOv7 predicts its label incorrectly. Bottom row: Prediction results after retraining the new dataset generated by the proposed pipeline.

the same with *suitcase* as it is often correctly classified.

Table 2: Improvement in accuracy metrics after additional training on the "sports ball" class.

Sports Ball				
Accuracy Metrics	**P**	**R**	**mAP@.5**	**mAP@.5:.95**
Before Training	0.669	0.806	0.705	0.655
After Training	0.923 (+0.254)	0.854 (+0.048)	0.902 (+0.197)	0.775 (+0.12)

Table 3: Improvement in accuracy metrics after extra training on the "suitcase" class.

Suitcase				
Accuracy Metrics	**P**	**R**	**mAP@.5**	**mAP@.5:.95**
Before Training	0.917	0.834	0.902	0.724
After Training	0.941 (+0.024)	0.857 (+0.023)	0.919 (+0.017)	0.739 (+0.015)

4.4 Accuracy on Novel Classes

Evaluating the accuracy of the pre-trained object detection models on novel classes is essential to assess their generalization capabilities. We have trained and evaluated YOLOv7 with the following classes: Rubic cube, dustbin, headset headphones, DSLR camera, and tennis ball categories. We have reported the accuracy of YOLOv7 with different metrics on the Rubic cube in Table 4 and DSLR Camera in Table 5. Few resulting predictions are shown in Fig. 3 and 4.

4.5 Results Using Varying Number of Images

As we are scraping images from web without using any paid APIs, we can fetch a maximum of 400-500 images on a single run. This number can further vary based on the quality of the internet connection and the server load. Hence we decided to observe the effect of the number of fetched images on the Novel Class Training accuracy. The results presented in Tab. 6 show that higher mAP can be achieved with more number of samples.

Table 4: Accuracy metrics of YOLOv7 when trained on the dataset produced by the proposed pipeline using the search term "Rubix Cube", considering all possible combinations of the pipeline's three stages.

Search Query: Rubix Cube						
Task			*Accuracy Metrics*			
Class Agnostic OD	**Feature Extraction**	**Clustering**	**P**	**R**	**mAP@.5**	**mAP@.5:.95**
OLN	ResNet50	K-Means	0.96	0.96	0.941	0.78
OLN	InceptionV3	K-Means	0.842	0.95	0.869	0.74
OLN	EfficientNet	K-Means	0.889	0.96	0.922	0.788
MAVL	ResNet50	K-Means	0.45	0.731	0.512	0.39
MAVL	InceptionV3	K-Means	0.47	0.721	0.517	0.4
MAVL	EfficientNet	K-Means	0.49	0.806	0.613	0.471
OLN	ResNet50	DBSCAN	0.72	0.69	0.702	0.541
OLN	InceptionV3	DBSCAN	0.641	0.604	0.647	0.48
OLN	EfficientNet	DBSCAN	0.556	0.62	0.57	0.39
MAVL	ResNet50	DBSCAN	0.69	0.68	0.67	0.5
MAVL	InceptionV3	DBSCAN	0.56	0.54	0.547	0.372
MAVL	EfficientNet	DBSCAN	0.492	0.523	0.48	0.31

Table 5: Accuracy metrics of YOLOv7 when trained on the dataset produced by the proposed pipeline using the search term "DSLR Camera", considering all possible combinations of the pipeline's three stages.

Search Query: DSLR Camera						
Task			*Accuracy Metrics*			
Class Agnostic OD	**Feature Extraction**	**Clustering**	**P**	**R**	**mAP@.5**	**mAP@.5:.95**
OLN	ResNet50	K-Means	0.958	0.958	0.985	0.826
OLN	InceptionV3	K-Means	0.833	0.95	0.86	0.737
OLN	EfficientNet	K-Means	0.88	0.955	0.917	0.779
MAVL	ResNet50	K-Means	0.552	0.709	0.518	0.464
MAVL	InceptionV3	K-Means	0.556	0.711	0.518	0.498
MAVL	EfficientNet	K-Means	0.589	0.83	0.54	0.523
OLN	ResNet50	DBSCAN	0.57	0.713	0.532	0.499
OLN	InceptionV3	DBSCAN	0.548	0.719	0.51	0.49
OLN	EfficientNet	DBSCAN	0.536	0.808	0.504	0.482
MAVL	ResNet50	DBSCAN	0.432	0.607	0.41	0.343
MAVL	InceptionV3	DBSCAN	0.433	0.619	0.403	0.33
MAVL	EfficientNet	DBSCAN	0.459	0.64	0.429	0.37

Table 6: Effect of nnumber of scraped images on accuracy metrics after training.

Effect of no. of scraped images				
No of images\Metrics	**P**	**R**	**mAP@.5**	**mAP@.5:.95**
50	0.378	0.4	0.371	0.2
100	0.457	0.44	0.487	0.338
200	0.579	0.48	0.553	0.378
300	0.823	0.56	0.706	0.544
400	0.923	0.949	0.929	0.72
500	0.96	0.96	0.941	0.78

5 CONCLUSIONS

In conclusion, the research paper presents a comprehensive approach of unsupervised novel object detection. We have detected a few potential objects in the image using a class-agnostic object detection model and obtained ROI proposals. Extracting features from these proposals has facilitated the representation of each object in a discriminative manner. Through clustering, we have reduced redundancy and obtained a refined dataset generated from data obtained via keyword searching on the web. Custom training has been performed using the weakly generated dataset to enhance the object detection model further. Spatial information of the objects has been preserved during the -agnostic detection and ROI pooling stages, ensuring accurate annotation preparation. The model has been fine-tuned on weakly generated datasets, focusing on the targeted class, resulting in improved object detection capabilities.

Evaluation of the custom-trained model has demonstrated its effectiveness in detecting and localizing the targeted class of objects. The integration of clustering, weakly generated data, spatial preservation, and custom training has contributed to the overall success of the proposed approach. This research provides new insights into unsupervised novel object detection, addressing the challenges of limited labelled data for novel objects. The methodology presented in this paper offers a practical framework for detecting and localizing novel objects in various domains, paving the way for advancements in computer vision and object detection research. Future work can focus on extending this approach to real-time applications and exploring additional techniques to enhance the accuracy and efficiency of unsupervised novel object detection systems.

REFERENCES

Bao, Z., Tokmakov, P., Jabri, A., Wang, Y.-X., Gaidon, A., and Hebert, M. (2022). Discovering objects that can move. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11789–11798.

Blatter, P., Kanakis, M., Danelljan, M., and Van Gool, L. (2023). Efficient visual tracking with exemplar transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1571–1581.

Cheng, G., Wang, J., Li, K., Xie, X., Lang, C., Yao, Y., and Han, J. (2022). Anchor-free oriented proposal generator for object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11.

Fan, Q., Zhuo, W., Tang, C.-K., and Tai, Y.-W. (2020). Few-shot object detection with attention-rpn and multi-relation detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4013–4022.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. (2017). Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969.

Hénaff, O. J., Koppula, S., Shelhamer, E., Zoran, D., Jaegle, A., Zisserman, A., Carreira, J., and Arandjelović, R. (2022). Object discovery and representation networks. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII*, pages 123–143. Springer.

Hou, L., Lu, K., Xue, J., and Li, Y. (2022). Shape-adaptive selection and measurement for oriented object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 923–932.

Kang, B., Liu, Z., Wang, X., Yu, F., Feng, J., and Darrell, T. (2019). Few-shot object detection via feature reweighting.

Kim, D., Lin, T., Angelova, A., Kweon, I. S., and Kuo, W. (2021). Learning open-world object proposals without learning to classify. *CoRR*, abs/2108.06753.

Kuo, W., Hariharan, B., and Malik, J. (2015). Deepbox: Learning objectness with convolutional networks.

LaBonte, T., Song, Y., Wang, X., Vineet, V., and Joshi, N. (2023). Scaling novel object detection with weakly supervised detection transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 85–96.

Li, W., Chen, Y., Hu, K., and Zhu, J. (2022a). Oriented reppoints for aerial object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1829–1838.

Li, Y., Mao, H., Girshick, R., and He, K. (2022b). Exploring plain vision transformer backbones for object detection. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 280–296. Springer.

Maaz, M., Rasheed, H., Khan, S., Khan, F. S., Anwer, R. M., and Yang, M.-H. (2022). Class-agnostic object detection with multi-modal transformer. In *17th European Conference on Computer Vision (ECCV)*. Springer.

O. Pinheiro, P. O., Collobert, R., and Dollar, P. (2015). Learning to segment object candidates. In Cortes, C., Lawrence, N., Lee, D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Park, K., Woo, S., Oh, S. W., Kweon, I. S., and Lee, J.-Y. (2022). Per-clip video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1352–1361.

Peng, J., Bu, X., Sun, M., Zhang, Z., Tan, T., and Yan, J. (2020). Large-scale object detection in the wild from imbalanced multi-labels. *CoRR*, abs/2005.08455.

Pont-Tuset, J., Arbelaez, P., T.Barron, J., Marques, F., and Malik, J. (2017). Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(1):128–140.

Qin, H., Yu, C., Gao, C., and Sang, N. (2022). D2t: A framework for transferring detection to tracking. *Pattern Recognition*, 126:108544.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Tang, F. and Ling, Q. (2022). Ranking-based siamese visual tracking. In *Proceedings of the IEEE/CVF Conference* *on Computer Vision and Pattern Recognition*, pages 8741–8750.

Uijlings, J. R. R., van de Sande, K. E. A., Gevers, T., and Smeulders, A. W. M. (2013). Selective search for object recognition. *International Journal of Computer Vision*, 104(2):154–171.

Wang, Y., Shen, X., Hu, S. X., Yuan, Y., Crowley, J. L., and Vaufreydaz, D. (2022). Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553.

Wu, A., Han, Y., Zhu, L., and Yang, Y. (2021). Universal-prototype enhancing for few-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9567–9576.

Xu, Q., Fang, F., Gauthier, N., Li, L., and Lim, J.-H. (2020). Active image sampling on canonical views for novel object detection. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 2241–2245. IEEE.

Xu, X., Wang, J., Li, X., and Lu, Y. (2022). Reliable propagation-correction modulation for video object segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2946–2954.

Yang, J., Liu, S., Li, Z., Li, X., and Sun, J. (2022). Real-time object detection for streaming perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5385–5395.

Zhu, C., Chen, F., Ahmed, U., Shen, Z., and Savvides, M. (2021). Semantic relation reasoning for shot-stable few-shot object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8782–8791.

Zitnick, L. and Dollar, P. (2014). Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*.