Object Detection in Heritage Archives using a Human-in-Loop Concept

KASTURI, Surya, SHENFIELD, Alex <http://orcid.org/0000-0002-2931-8077>, ROAST, Chris <http://orcid.org/0000-0002-6931-6252>, LE PAGE, Danny and BROOME, Alice

Published version

KASTURI, Surya, SHENFIELD, Alex, ROAST, Chris, LE PAGE, Danny and BROOME, Alice (2024). Object Detection in Heritage Archives using a Human-in-Loop Concept. In: NAIK, Nitin, JENKINS, Paul, GRACE, Paul, YANG, Longzhi and PRAJAPAT, Shaligram, (eds.) Advances in Computational Intelligence Systems. Contributions Presented at The 22nd UK Workshop on Computational Intelligence (UKCI 2023), September 6-8, 2023, Birmingham, UK. Advances in Intelligent Systems and Computing (1453). Cham, Springer, 170-181.

Object Detection in Heritage Archives using a Human-in-Loop Concept

Surya Kasturi[1], Alex Shenfield[1], Chris Roast[1], Danny Le Page[2], and Alice Broome[2]

[1] Sheffield Hallam University, Sheffield, UK,
[2] British Online Archives, Leeds, UK

Abstract. The use of object detection has become common within the area of computer vision and has been considered essential for a numerous applications. Currently, the field of object detection has undergone significant development and can be broadly classified into two categories: traditional machine learning methods that employ diverse computer vision techniques, and deep learning methods. This paper proposes a methodology that incorporates the human-in-loop feedback concept to enhance the deep learning object detection capabilities of pre-trained models. These Deep Learning models were developed using a custom humanities and social science dataset that was obtained from the British Online Archives collections database.

Keywords: Object Detection, Human-in-Loop, Deep Learning

1 Introduction

Machine learning (ML) is a widely known concept that has gained significant interest in various domains, such as computer vision, pattern recognition, and data retrieval. ML allows computers to learn from data without explicit programming, improving themselves through experience. ML algorithms analyze historical data, identify patterns, and establish mathematical relationships between inputs and outputs. This technique relies on large training databases and computational power. While ML is fascinating, artificial intelligence (AI) is an even more advanced and intriguing technology. AI involves computer systems simulating human cognitive processes, including learning and problem-solving.

Human involvement plays a crucial role in every step of the machine learning (ML) pipeline, starting from data preparation to result inference. Before constructing a model, data scientists dedicate substantial time to data preprocessing [13]. This involves tasks such as data extraction, integration, and cleaning. The data is then categorized and divided into separate training and test sets. Throughout the entire development process of training and testing the ML model, human participation is evident. The following sections of this paper explore existing knowledge on human involvement across different phases of ML development. Additionally, we present a methodology and corresponding results

that showcase improvements in object detection techniques applied to archival documents.

British Online Archives faces challenges due to the time-consuming publication process, limiting the volume and richness of curated collections and metadata. Humanities researchers rely on curated collections around a topic of interest to inspire and facilitate their work, but time constraints result in limited metadata provided. Researchers in humanities require consideration of both written and graphical content, but searching graphical content remains challenging compared to textual content. This complexity hinders systematic search and analysis of graphical material. In order to expedite the process of curating and publishing archives while also generating detailed and easily searchable metadata, we propose a machine learning pipeline as to produce comprehensive metadata about the elements within the collection. This extensive metadata, which describes various aspects of the curated collection, is automatically generated. This automation allows editors to concentrate on validating, organizing, and refining the contents of the collection. Once the collection is published, users can access the metadata, which provides detailed information. This enhanced accessibility enables users to systematically search for graphical content using both keywords and free-text queries, improving their overall experience. The object detection is crucial component of our research and a part of this research has been proposed in this paper. This research is a part of KTP (Knowledge Transfer Partnership) project which is funded by UKRI through Innovate UK.

2 Prior Work

Utilising pre-existing knowledge into the learning framework is a viable strategy for addressing data sparsity, as it obviates the need for the learner to derive the knowledge solely from the available data [3]. Humans possess extensive prior knowledge as specialised agents. The developer has the potential to facilitate machine learning through the incorporation of human wisdom and knowledge, which can aid in addressing the issue of sparse data, particularly in domains where there is insufficient training data [24]. To address these challenges a new concept named as Human-in-Loop (HIL) has been proposed. This approach primarily focuses on involving human expertise into the modelling procedure [7].

A conventional machine learning algorithm generally comprises of three components [21]: data pre-processing, data modelling, and process optimisation via developer modifications to enhance the performance of the model. In the typical process of model development, human intervention is required during the data pre-processing stage to transform unstructured data into structured label data. This practise has been identified by some researchers as an application of the Human-in-the-Loop (HIL) concept [1]. Usually, the efficiency of deep learning is dependent upon the quality of the data. To obtain effective performance in a novel task, a substantial quantity of accurately labelled data is required. The process of annotating extensive sets of data necessitates significant effort and time investment. This can pose a challenge for tasks that require multiple it-

erations and cannot accommodate the associated costs and delays. In contrast to data annotation, iterative data labelling places greater emphasis on user experience enabling users to engage in the data annotation process directly. So, here the objectives has been divided into two primary areas: first, improving the learning system through iterative labeling and second involves giving importance to engaging and communicating with users. This means actively involving users in the learning process, gathering their feedback, and incorporating their insights to enhance the system's performance.

Yu et al. [23] employed a labelling scheme that was partially automated, utilising deep learning techniques with human-in-the-loop to reduce the need for manual labour in the annotation process. This represents the fundamental model of uncomplicated iterative annotation. Several domains within the realm of Artificial Intelligence, including Natural Language Processing (NLP) and Computer Vision (CV), employ diverse methodologies that utilise human intelligence for the purposes of training and inferring experimental outcomes. Research related for both NLP and CV covers a range of techniques that combine human and machine intelligence. The utilisation of heuristic methods has considered the varied nature of human creativity in order to attain outcomes of superior quality.

The utilisation of Deep Learning techniques, specifically neural network-based methods, has become the leading approach for executing various computer vision tasks, as evidenced by recent studies [20]. In order to enhance the efficiency of stated techniques, human feedback has been incorporated into the deep learning framework to improve the system's overall intelligence in addressing difficult scenarios that are beyond the model's capacity to handle. Object detection, which is considered to be a fundamental and challenging problem in the field of computer vision, has drawn substantial interest in recent times [4]. Yao et al. [22] highlight that the repeated cycles of queries can incur significant costs and consume substantial time, rendering it impractical to engage in interactions with end-users. They proposed an interactive architecture for object detection that enables users to rectify a limited number of annotations suggested by a model for an unannotated image or test dataset with the highest predicted annotation cost. Madono et al. [12] proposed a proficient framework for object detection that involves human-in-the-loop. The framework is comprised of bi-directional deep SORT [19] and annotation-free segment identification (AFSID). The responsibility of humans within this architecture pertains to the verification of object candidates that cannot be automatically detected by bi-directional deep SORT. Subsequently, the model should be trained on the supplementary objects that have been annotated by individuals.

Numerous researchers have been dedicating their efforts towards enhancing the performance of object detection models. These models can be classified into two categories: one-stage object detectors and two-stage object detectors. One-stage object detection models execute classification and regression operations on closely spaced anchor boxes, without generating a sparsely populated Region of Interest (RoI) set. The YOLO algorithm[14], represents an initial foray into the direct detection of objects on a feature map with high density. The utilisation

of multi-scale features has been proposed by SSD [11] as a means of detecting objects with varying scales. Later, RetinaNet [10] introduced the use of focal loss as a solution to tackle the issue of imbalanced classes in the context of dense object detection.

Currently, two-stage detectors exhibit superior performance in terms of detection accuracy. The detectors employ a two-stage approach wherein the initial stage generates sparse region proposals, followed by a subsequent stage that performs regression and classification on the proposed regions. The RCNN model [5] employed computer vision techniques such as Selective Search [18] and Edge Boxes [25] at a low level to produce proposals. Subsequently, a CNN was utilised to extract features for the purpose of training an SVM classifier and bounding box regressor. Fast R-CNN [4] then proposed a method of feature extraction for individual proposals on a feature map that is shared, through spatial pyramid pooling. Later, building on this, Faster R-CNN [15] incorporated the region proposal process within the deep ConvNet architecture, resulting in a detector that can be trained end-to-end.

The authors of R-FCN [2] introduced a region-based fully convolutional network as a means of producing features that are sensitive to regions for the purpose of detection which traditional methods lacked. By directly producing region-sensitive features using a fully convolutional network, R-FCN achieves faster inference times and better localization accuracy. FPN (Featured Pyramid Network)[9] an architectural approach that employs top-down processing and lateral connections to produce a feature pyramid suitable for detecting objects at multiple scales. FPN preserves both semantic information and spatial details, improving object detection across various scales. This approach has become widely adopted and has advanced the accuracy and robustness of object detection models. The EfficientDet model [16] utilises a compound scaling technique to simultaneously increase the dimensions of depth, width, and resolution for the backbone, BiFPN, and box/class prediction networks. The compound scaling technique used in EfficientDet enhances the model's capacity, improves feature representation, and allows for more precise object detection across different scales, contributing to its success in the field of object detection.

The current research emphasises on the development of a pipeline that is defined by ease of use and robustness. Even though involving humans in model inference incurs additional costs [22], we believe that human in loop techniques such as interactive machine learning will actually provide significant improvements in the process where there is a scarcity of data for training the model.

3 Object Detection

Our implementation of Human-in-Loop for object detection in archival documents involves six fundamental steps:

1. Dataset collection and annotation using the Label Studio [8] tool
2. Object detection model training using a transfer learning approach (which also entails selecting the appropriate model)

3. Inference on validation data
4. Modification or correction of the model's inference outcomes (using customised Label Studio user interface)
5. Retraining the model with new learning parameters after collecting a few newly annotated samples
6. Evaluation of the results in a held out test set

3.1 Dataset for base model

The models have undergone training on a dataset comprising 146 images containing of 180 objects and 3 classes. The valid dataset during training has 35 images containing 53 objects. The test dataset, on the other hand, consists of 37 images containing 54 objects and the same classes as the other datasets. Some dataset samples are illustrated in Figure 2.

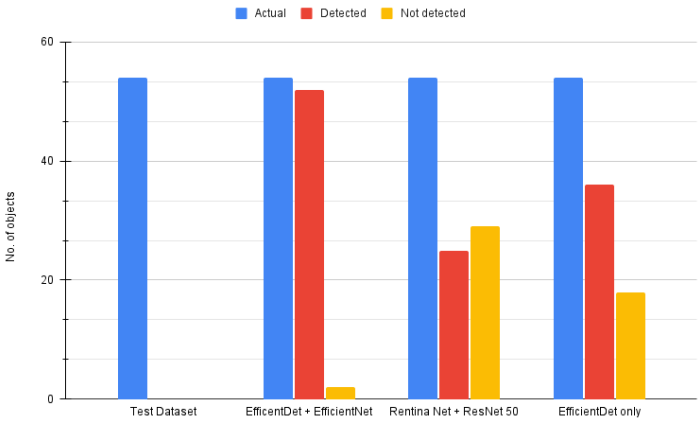


Fig. 1: Comparisons of various base models performance

3.2 Model Configuration

In this study our baseline object detection model is a two-stage fine-tuned EfficientDet architecture with a second stage EfficientNet classfier. Initial evaluation work indicates that this combination outperforms a single stage EfficientDet model and a two stage model based on RetinaNet and ResNet50. The comparison of different base models is presented in Figure 1. It is evident that the two-stage model, namely EfficientDet + EfficientNet, outperforms the other two models. The two stage EfficientDet + EfficientNet model configuration is then further tuned using the Human-in-the-Loop (HIL) implementation discussed in the Section 4 to improve the overall performance of the system.

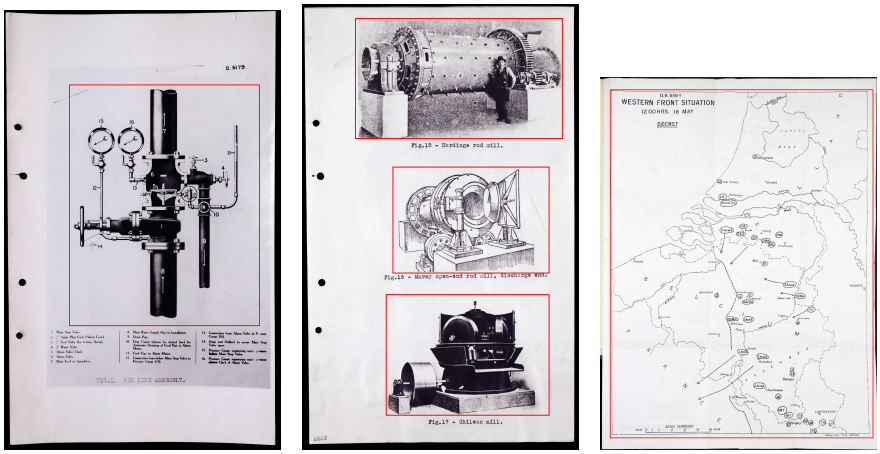


Fig. 2: Examples with bounding boxes

4 Implementation of Human-in-Loop

The essential elements of the Human-in-Loop framework entail the development of a user interface to facilitate user inputs and the establishment of a pipeline to enable automatic model retraining in response to human feedback. The subsequent sections will elaborate on the utilisation of Label Studio [8] as a user-facing interface for the purpose of rectifying or altering the outcomes generated by the model.

4.1 The Interface

The Human-in-Loop system requires an interface component that must exhibit simplicity in order to ensure ease of use for all users. The dataset employed in Section 3.1 was curated through the utilisation of Label Studio, a tool that enables the importing of extensive image datasets from cloud storage platforms like S3. All the data utilised in our study was obtained from the British Online Archives. The visual representation depicted in Figure 3 provides an overview of the interface design intended for the user's perspective. This shows how users are able to access the predictions generated by the model and provide feedback to the pipeline.

4.2 The Pipeline

The effective implementation of machine learning pipeline integration constitutes another significant element of human-in-the-loop. The incorporation of this integration enhances the model's ability to acquire knowledge from user feedback. During the initial stage of the pipeline, the data undergoes pre-processing, which involves the creation of annotated data and the removal of abnormal data. The

Fig. 3: Screenshot of the Label Studio Interface

implementation of augmentation and normalisation techniques on the training dataset is utilised to enhance the quality of the training process. The transfer learning [17] methodology is employed in order to create a baseline model for the implementation of our Human-in-Loop (HIL) process. Non-Maximum Suppression (NMS) [6] is employed during post-processing to eliminate redundant bounding boxes and facilitate the selection of optimal bounding boxes. A distinct test dataset was generated, which was not exposed to the model during the HIL training phase, in order to assess its efficiency during evaluation. Finally the prediction of model on the test dataset available to the user on the Label Studio platform. Users then have the opportunity to review the predicted images and make necessary adjustments to the bounding boxes. The adjusted images data will then be collected and fed back into the model, thereby enhancing its performance.

The parameters utilised for our object detection model are as follows:

1. Image size : 256 x 256
2. Learning rate : 0.005 (for initial training) & 0.00005 (For re-training based on user feedback)
3. Batch size: 13
4. IOU threshold: 0.45
5. Prediction confidence = 0.50

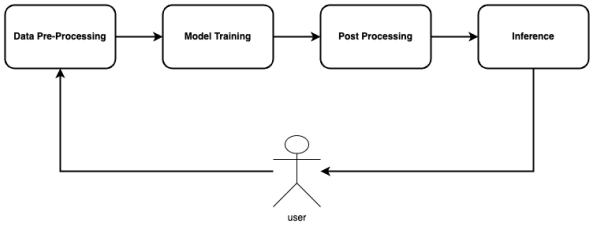


Fig. 4: Human-in-loop process

4.3 Evaluation Method

Following the integration of interface and pipeline, an essential aspect of the human-in-the-loop process is result evaluation. In this section, we present our evaluation method. Initially, we refer to the results obtained from the first inference as HIL-0%, indicating that no human feedback was involved in generating these results. Subsequently, we introduce HIL-10%, HIL-15%, and HIL-20%, which signify that users have corrected 10%, 15%, and 20% of poorly performing predicted bounding boxes in the test dataset, respectively.

The evaluation employs two distinct test datasets, namely test-1 and test-2. One of these datasets will be utilised for the purpose of rectifying the predictions, while the other dataset will be exclusively utilised for evaluating the model's performance across varying levels of HIL. Upon the user's modification of the bounding boxes on one of the test dataset, the corresponding corrected images will replace some random images in the initial training dataset which then becomes a new training dataset. A small learning rate of 0.00005 (in our case) is employed to retrain the model using the new training dataset. To retain previously learned information while incorporating the user-provided data, we utilize transfer-learning techniques that load the pre-trained weights of the model from HIL-0%. This approach allows us to make gradual adjustments to the model's weights, ensuring the assimilation of the new data without compromising the existing knowledge. Subsequently, we will conduct a comparison of the mean Intersection over Union (mIOU), mean Average Precision (mAP) scores, Precision and Recall specifically at IOU values of 0.5 and 0.75, for each of the distinct stages of models involved in this procedure. The comprehensive examination and outcomes can be found in the following section 5

5 Results

The diverse outcomes of the model's performance at different percentages of HIL (Human-in-the-Loop) corrections are evident from the provided metrics in table 1. An example object detection at different HIL levels are shown in Figure 6

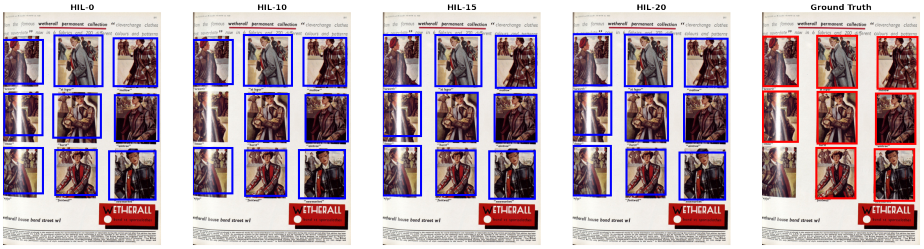
The table 1 represents the model performance at various stages of the HIL process. The results show that the model's performance improves as the percentage of images corrected by the user increases. This is because a newer training

Table 1: Table of evaluation results

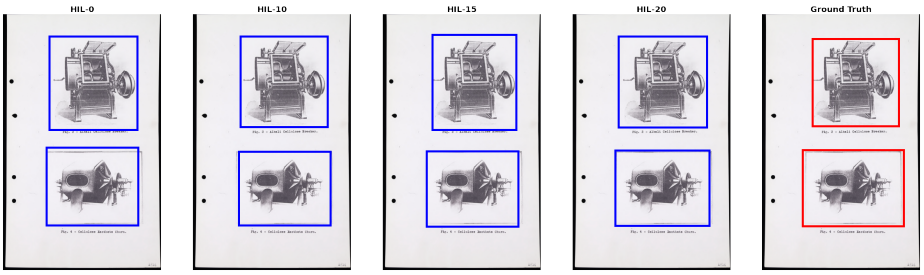
		mAP		Precision		Recall		FNs		FPs	
HIL	mIOU	@0.5	@0.75	@0.5	@0.75	@0.5	@0.75	@0.5	@0.75	@0.5	@0.75
0%	0.789	0.966	0.920	0.925	0.807	0.832	0.726	161	263	64	166
10%	0.805	0.966	0.933	0.927	0.817	0.856	0.754	138	236	64	162
15%	0.801	0.963	0.929	0.920	0.815	0.855	0.757	139	233	71	165
20%	**0.810**	**0.967**	**0.935**	**0.933**	**0.831**	**0.87**	**0.776**	**124**	**215**	**60**	**151**

dataset allows the model to learn more about the different types of objects that it is likely to encounter. Specifically, the mIOU and mAP scores for HIL-20 are higher than the scores for HIL-0, HIL-10, and HIL-15. This suggests that using 20% of the corrected images might provide a better results. However, it is worth noting that HIL-15 had a higher number of false positives and false negatives compared to the other models except HIL-0. This increase in false positives and fasle negatives could be attributed to various factors, including human errors during the correction process, imbalanced distribution of objects in the dataset, or the complexity and small size of the objects leading the model to predict bounding boxes for non-existent objects or missing some objects. Overall, the results show promise, indicating that using a higher percentage of corrected images (such as 20%) for training might yield better performance for this dataset. Nevertheless, it is essential to continue evaluating the model on different datasets to assess its adaptability and performance across various object types. Figure 7 presents an overview of precision-recall curves at various stages of HIL. The key observation is that as the HIL percentage increases, there is a less pronounced decrease in precision at the initial increase in recall which concludes that HIL can be used to improve the accuracy of object detection models, without sacrificing too much precision.

Fig. 5: Human Error (Missed annotation highlighted in yellow circle)

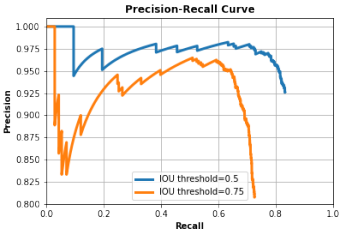


(a)

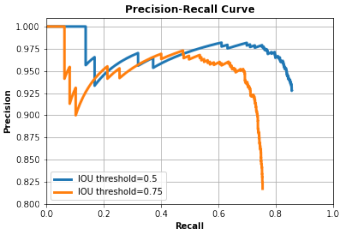


(b)

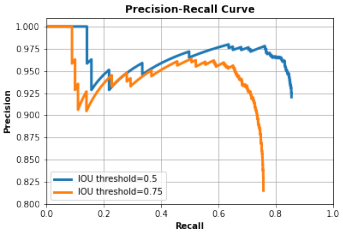
Fig. 6: Test dataset HIL progress



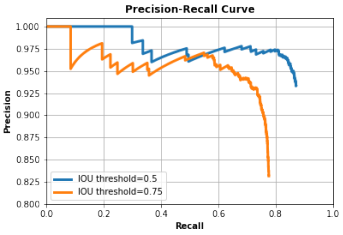
(a) HIL-0%



(b) HIL-10%



(c) HIL-15%



(d) HIL-20%

Fig. 7: Precision-Recall curves at different % of HIL

6 Conclusion

The findings indicate that the inclusion of HIL corrections at a moderate level (approximately 10-20%) can improve the performance of the model in tasks related to object detection. Furthermore, based on the findings in section 5 HIL helps the model to improve the localisation of the objects. Nevertheless, augmenting the dependence on human corrections beyond a particular threshold could potentially give rise to incongruities and impede the precision of the model. Striking a balance between automated predictions and human corrections is crucial for achieving optimal performance in these tasks.

One potential avenue for further investigation and analysis to determining the optimal threshold for incorporating human-in-the-loop (HIL) corrections, which can yield the most substantial enhancements in performance for tasks related to object detection. Additionally, examine diverse methodologies or computational procedures for integrating human-in-the-loop (HIL) corrections in an efficient manner. Analyse the effects of various correction mechanisms, including active learning, reinforcement learning, and selective correction sampling, on improving the accuracy and efficiency of the model. Further investigation in the field of natural language processing (NLP), specifically focusing on machine translation, presents promising opportunities for significant advancements. An area worth investigating is the possibility of utilising human-in-the-loop (HIL) corrections as a means of improving the calibre of machine translation results.

References

1. Chai, C., Li, G.: Human-in-the-loop techniques in machine learning. IEEE Data Eng. Bull. 43(3), 37–52 (2020)
2. Dai, J., Li, Y., He, K., Sun, J.: R-fcn: Object detection via region-based fully convolutional networks. Advances in neural information processing systems 29 (2016)
3. Diligenti, M., Roychowdhury, S., Gori, M.: Integrating prior knowledge into deep learning. In: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 920–923 (2017)
4. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 1440–1448 (2015)
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 580–587 (2014)
6. Hosang, J., Benenson, R., Schiele, B.: A convnet for non-maximum suppression. In: Pattern Recognition: 38th German Conference, GCPR 2016, Hannover, Germany, September 12-15, 2016, Proceedings 38. pp. 192–204. Springer (2016)
7. Kumar, V., Smith-Renner, A., Findlater, L., Seppi, K., Boyd-Graber, J.: Why didn't you listen to me? comparing user control of human-in-the-loop topic models. arXiv preprint arXiv:1905.09864 (2019)
8. Label Studio contributors: Label Studio. `https://labelstud.io/` (2021), [Online; accessed September 2021]
9. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017)

10. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
11. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C., Berg, A.: Ssd: Single shot multibox detector,‖ in european conference on computer vision (eccv) (2016)
12. Madono, K., Nakano, T., Kobayashi, T., Ogawa, T.: Efficient human-in-the-loop object detection using bi-directional deep sort and annotation-free segment identification. In: 2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 1226–1233. IEEE (2020)
13. Obaid, H.S., Dheyab, S.A., Sabry, S.S.: The impact of data pre-processing techniques and dimensionality reduction on the accuracy of machine learning. In: 2019 9th annual information technology, electromechanical engineering and microelectronics conference (iemecon). pp. 279–283. IEEE (2019)
14. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
15. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems 28 (2015)
16. Tan, M., Pang, R., Le, Q.V.: Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10781–10790 (2020)
17. Torrey, L., Shavlik, J.: Transfer learning. In: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, pp. 242–264. IGI global (2010)
18. Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. International journal of computer vision 104, 154–171 (2013)
19. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: 2017 IEEE international conference on image processing (ICIP). pp. 3645–3649. IEEE (2017)
20. Wu, X., Xu, B., Zheng, Y., Ye, H., Yang, J., He, L.: Fast video crowd counting with a temporal aware network. Neurocomputing 403, 13–20 (2020)
21. Xin, D., Ma, L., Liu, J., Macke, S., Song, S., Parameswaran, A.: Accelerating human-in-the-loop machine learning: Challenges and opportunities. In: Proceedings of the second workshop on data management for end-to-end machine learning. pp. 1–4 (2018)
22. Yao, A., Gall, J., Leistner, C., Van Gool, L.: Interactive object detection. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 3242–3249. IEEE (2012)
23. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)
24. Zhang, R., Torabi, F., Guan, L., Ballard, D.H., Stone, P.: Leveraging human guidance for deep reinforcement learning tasks. arXiv preprint arXiv:1909.09906 (2019)
25. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 391–405. Springer (2014)