Self-supervised animal detection in constrained environment

Fayaz Rahman[1,2]
fayazrahman4u@pg.cusat.ac.in
C. B. Dev Narayan[1,2]
devcb@pg.cusat.ac.in
Mohib Ullah[2]
mohib.ullah@ntnu.no
Muhammad Mudassar Yamin[2]
muhammad.m.yamin@ntnu.no
Øyvind Nordbø[3]
oyvind.nordbo@norsvin.no
Christopher Coello[3]
christopher.coello@norsvin.no
Ali Shariq Imran[2]
ali.imran@ntnu.no
Santhosh Kumar G.[1]
san@cusat.ac.in
Madhu S. Nair[1]
msn@cusat.ac.in
Faouzi Alaya Cheikh[3]
faouzi.cheikh@ntnu.no

[1] Artificial Intelligence & Computer Vision Lab
Department of Computer Science
Cochin University of Science and Technology
Kochi - 682022, India.

[2] Intelligent Systems and Analytics (ISA) Research Group, Department of Computer Science, Norwegian University of Science and Technology
815 Gjøvik
Norway.

[3] Norsvin SA
Storhamargata 44
2317 Hamar
Norway.

Abstract

Animal detection is a critical component of behavioural research in animal sciences. Improvement in detection accuracy along with model efficiency is essential for good results in animal phenotyping. Collecting data on a farm is easy, but labelling the data is tedious. Therefore, most current methods rely on limited training data, which is a restrictive factor for achieving optimal results. Recently, self-supervised learning methods leveraging unlabelled data have been shown to produce improved accuracy on extensive detection benchmarks. In this work, we proposed a self-supervised animal detection pipeline for animals in a constrained environment. The proposed pipeline employs a modified version of Barlow Twins in its pre-training stage. It is tested on a specially created dataset with labelled images along with unlabelled images for self-supervised learning. With our proposed pipeline, we achieved a boost in *mAP* across various thresholds without compromising the efficiency of the detection models and having a high FPS video output even on non-industrial-grade GPUs, making it suitable for online tracking applications. The link to our code is publically available at https://github.com/FayazRahman/barlow-effdet.

© 2023. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

1 Introduction

Animals rely on body posture and sound to express themselves. Therefore, Animal-Computer Interaction studies are making a significant effort to develop technology for automatically recognising animal behaviour and body postures [16]. The ability to process videos of small animals and automatically score their behaviour [2] is crucial for several common tasks in the life sciences, such as measuring the locomotive activity of an animal, defining its position in an arena, quantifying its interactions with an object, or assessing its engagement in defensive behaviours like freezing [15]. To improve farm products while complying with animal welfare regulations, breeding companies aim to leverage vision-based solutions to monitor animal living and conceive novel animal traits that can enhance breeding programs.

Object detection is a critical component of animal behaviour analysis frameworks by means of tracking [3, 7, 17]. Recently, there has been huge progress in object detection models, giving high accuracy results such as YOLO [18] and FasterRCNN [20], but many of these methods trade-off model efficiency for better results. However, model efficiency is essential for behaviour analysis, which requires fast results. Most of the earlier work in this area mainly relies on a fully supervised learning paradigm. For example, PigPose [10] proposed keypoint detection in its framework for animal pose estimation and tracking. Similarly, Zhang et al. [28] introduced a hierarchical pig detection and correlation-based tracking.

Compared to fully supervised learning, there has been a growing interest in semi-supervised and self-supervised object detection algorithms that leverage unlabelled data to improve detection performance [6, 7, 23, 25, 26]. Inspired by self-learning, we developed a pipeline that increased detection accuracy without compromising model efficiency. Primarily, we incorporated a modified version of the Barlow Twin's methodology [27] to pre-train the smaller EfficientDet detectors - variants D0 and D1 [22], on unlabelled data. In a nutshell, the contributions of our work are two-fold:

- A self-supervised training methodology that improves animal detection accuracy in a constrained environment.
- A well-curated, unique dataset collected from a pig farm.

The rest of the paper is organized in the following order. In section 2, the description of the proposed method is given. The dataset details, performance metrics, and implementation details are given in section 3. Section 4 lists the quantitative and qualitative results. The discussion and final remarks are given in section 5 that concludes the paper.

2 Methodology

Our proposed model consists of two training phases. The first phase involves pre-training the object detector using unlabelled data, and the second phase involves training it using labelled data. The graphical depiction of our proposed model and training methodology is shown in (Fig. 1). The model consists of two sister branches, as in a Siamese neural network [24]. In each branch, the first component is the object detection model, followed by a projector network. The projector network produces feature embeddings that are used to optimize the weights of the backbone object detector. In the following subsections, each component is briefly explained.

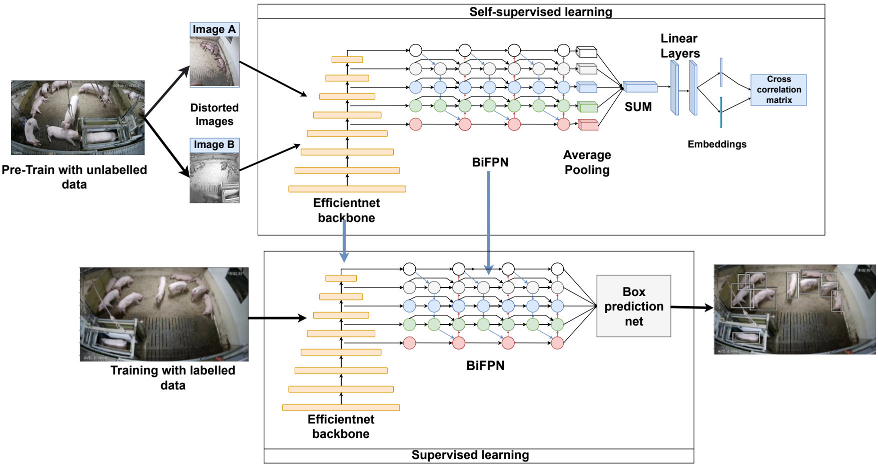

Figure 1: Self-learning architecture of our detection model. Pre-training with our modified Barlow Twins method is depicted at the top. The resulting detector model, then trained with labelled data, is depicted at the bottom.

2.1 Object detector architecture

We used EfficientDet-D0 as our base model for animal detection [22]; however, the methodology can be tuned for any generic object detector. It consists of an ImageNet pre-trained EfficientNet-B0 convolution neural network as its backbone network [21] and makes use of a feature pyramid network, namely BiFPN [21], which is equipped with cross-scale connections similar to PANet [14]. BiFPN applies repeated feature fusion of the bidirectional connections using Fast Normalized Fusion given by (Eq. 1):

$$O = \sum_i \frac{w_i}{\varepsilon + \sum_j w_j} . I_i \tag{1}$$

where w_i is a learnable weight, I_i is the input feature vector and $\varepsilon = 10^{-4}$ is a constant to avoid numerical instability. The fast normalized fusion is then applied on features from levels 3 to 7 of the backbone network to obtain five output features [22]. The extracted features are fed into a class and box prediction network to predict boxes and their corresponding confidence scores. The final boxes are filtered by application of Soft-NMS [4]. Soft-NMS chooses the box with the highest confidence score, and the confidence scores of boxes overlapping with the chosen box are decreased by (Eq. 2):

$$s_i = s_i e^{\frac{-IoU(M,b_i)^2}{\sigma}}, \forall b_i \notin D \tag{2}$$

where M is the bounding box with the maximum confidence score, D is the set of selected final boxes, s_i is the score of the box b_i, and σ is a hyperparameter to control the decay of scores. $IoU(M,b_i)$ refers to the ratio of the area of intersection to the area of union between the boxes M and b_i.

2.2 Self-supervised pre-training of detector

Contrary to direct training using labelled data as in a fully supervised learning paradigm, we pre-train the object detection model on unlabelled data using a modified Barlow twins strategy [27]. We apply the method on a condensed representation of the output of the Feature Pyramid Network of the EfficientDet network instead of the detector's backbone. This method aims to learn features invariant to distortions applied to the sample. In other words, it aims to find a representation that conserves information about the sample while being least informative about the specific distortions applied.

2.2.1 Pre-training architecture:

The self-supervised training involves passing two different variants of the input image to two identical networks. The distorted version of images is generated by applying random augmentation. The networks consist of:

1. EfficientDet network without the final classification and box prediction layers (Section. 2.1).

2. Projection Network: In the projection network, we apply global average pooling on the multi-scale representations obtained from the Bi-FPN layer of the detector and sum the results. This is followed by two linear layers with a batch normalization layer and ReLU activation in between. Thus, we obtain the embeddings from the projection network, which is fed into the loss function.

3. Barlow Twins loss function: This loss function helps in bringing the cross-correlation matrix between the embeddings from the projection network closer to the identity matrix.

2.2.2 Loss function:

The loss function used in pre-training calculates the cross-correlation matrix between the embeddings from the network for each of the different distortions and tries to bring it closer to the identity matrix. The loss function is given below (Eq. 3):

$$L_{bt} = \sum_i (1 - C_{ii})^2 + \lambda \sum_i \sum_{j \neq i} {C_{ij}}^2 \tag{3}$$

In Eq. 3, λ is a constant used to trade off the importance of the two loss terms. C is the cross-correlation matrix, which will be the identity matrix in the ideal case. The cross-correlation matrix between embeddings A and B, where b indexes samples and i,j indexes the embeddings given by z is given by (Eq. 4):

$$C_{ij} = \frac{\sum_b z^A_{b,i} z^B_{b,j}}{\sqrt{\sum_b (z^A_{b,i})^2} \sqrt{\sum_b (z^B_{b,j})^2}} \tag{4}$$

3 Experiments

3.1 Dataset description

The data is recorded in pig form. In the testing station, the animals are grouped, with eleven similar-sized animals in each pen, during the growth period from 30-120 kgs. Videos were recorded 24/7 in top-down view by LOREX (4K Ultra HD IP NVR) and ELOTEC (4MP Bullet, IP67) cameras with a resolution of 1920 × 1080 under different lighting conditions. Some frames of these video sequences are extracted and manually annotated by the Darwin V7 labs [1] and coco annotator [5]. These annotations contain bounding boxes, segmentation masks, and key points for each animal instance in the frame. The dataset used in this work contains 1674 images as training data with their ground truths. The ground truths are bounding boxes stored in JSON files created by manual animal boundary tracing by experts using COCO API. The dataset was split into 1339 training data points and 335 testing data points in an 80-20 split.

3.2 Evaluation metrics

The results are evaluated by calculating the Average Precision across various IoU thresholds. The IoU threshold is a measure to determine the minimum required overlap between predicted and ground truth bounding boxes for considering them as a true positive match. The metric mAP (mean Average Precision) [13] is given in Eq. 5 and Eq. 6

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{5}$$

$$AP = \frac{TP}{TP+FP} \tag{6}$$

The average detection precision is the number of true positives among the total positive detections. A detection is considered True Positive if it has an IoU above a specified IoU threshold (ϕ). The mAP is calculated across the IoU thresholds of $\phi = 0.5, 0.75, 0.5-0.95$.

3.3 Implementation Details

We conducted all of our experiments using the open-source library Pytorch [11] and PyTorch Lightning [9] on a machine with 16GB NVIDIA GeForce RTX 3080 Laptop GPU. As a preprocessing step, all images of our respective datasets (training, testing and validation) are resized to 512×512 and normalized to the $[0, 1]$ range.

3.3.1 Pre-training:

For pre-training using Barlow Twins, we used the augmentations in BYOL [8]. Each image needs to be augmented to produce two views consisting of random cropping, resizing, horizontal flipping, colour jittering, grayscale, Gaussian blurring, and solarization. Cropping and resizing are consistently applied, while the last five are applied randomly, with some probability. This probability is different for the two distorted views in the last two transformations (blurring and solarization). The hyperparameters used in the pre-training stage are listed in Table 1.

Table 1: Implementation details of pre-training

Hyperparameter	Value
λ for Barlow twins loss	$5e-3$
Embeddings size	4096
Batch size	16
Learning rate	$1e-4$
Epochs	50
Optimizer	AdamW

3.3.2 Training:

We apply a horizontal flip with a probability of 0.5 for training using labelled data. The backbone is initialized as a model pre-trained on ImageNet. The hyperparameters used for the same are listed in Table 2.

Table 2: Implementation details of training

Hyperparameter	Value
Batch size	24
Learning rate	$3e-3$
Epochs	24
Optimizer	AdamW

4 Results

The quantitative results are given in table 3. Our proposed model is compared with some of the standard detection methods, and the evaluations are done on our unique dataset mentioned in section 3. Some of the qualitative results of the model can be seen in Fig. 2. The quantitative results and the corresponding number of parameters and FLOPS are presented in Table 3.In the table, "Proposed" refers to the application of our method. It is apparent that the proposed method is able to push the accuracy of the model, especially in the higher IOU thresholds, placing it on par with newer models such as YoloV4-Tiny [12] and a much larger model like FasterRCNN [19] while being highly efficient at 5.8B FLOPS in the case of EfficientDet-D1, proving that the efficiency of our model is superior, making it suitable for applications such as tracking and behaviour analysis of animals.

Table 3: Quantitative evaluation of detection results: The proposed model (values noted in bold) is evaluated by using mAP, which is calculated across the IoU thresholds of $\phi = 0.5, 0.75, 0.5-0.95$.

Detector	mAP			Params	FLOPS
	0.5	0.75	0.5 : 0.96		
EfficientDet D0	0.953	0.688	0.592	3.8M	2.4B
EfficientDet D1	0.957	0.796	0.662	6.6M	5.8B
FasterRCNN + ResNet50 FPN	0.967	0.849	0.698	41.3M	13.3B
YoloV4 Tiny	0.962	0.794	0.664	6.06M	10.1B
Proposed EfficentDet-D0	**0.952**	**0.707**	**0.599**	**3.8M**	**2.4B**
Proposed EfficentDet-D1	**0.956**	**0.817**	**0.665**	**6.6M**	**5.8B**

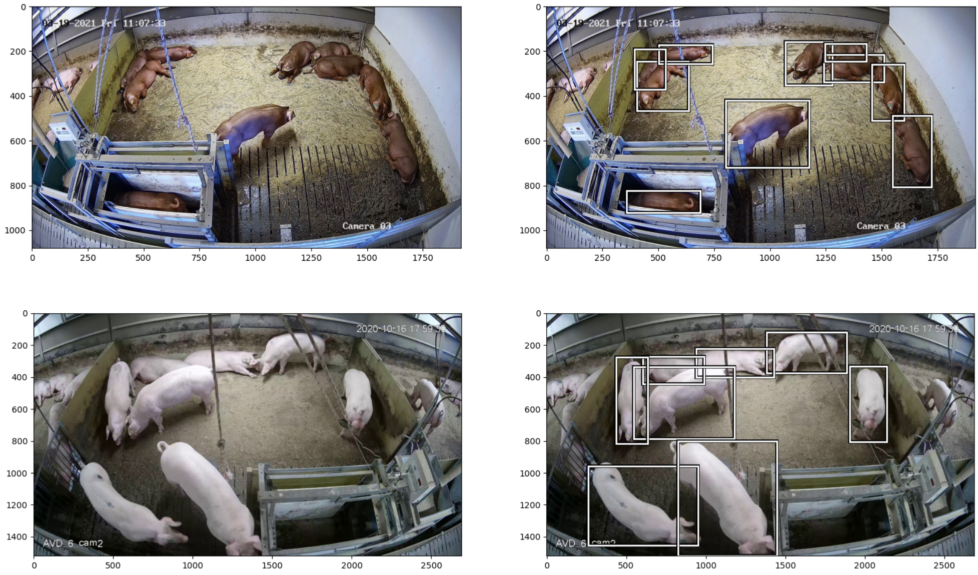

Figure 2: Reference images and their corresponding predicted bounding boxes.

4.1 Discussion

The potential reasons for the high accuracy and efficiency can be summarized in the following:

- The object detector is able to learn the variance in the unlabelled data and the underlying visual features through self-learning.
- The detector is able to learn valuable representations from the labelled data.
- The high efficiency can be attributed to the detector having fewer parameters.

5 Conclusion

We propose a methodology for the self-supervised detection of animals in a constrained environment. We pre-train base object detectors with modified Barlow Twins loss function and use the unlabelled data for inference. A custom-built dataset is used to train and evaluate the model. We achieved improvements in mAP in higher IoU ranges with the proposed strategy. The resulting model is highly efficient in terms of FLOPS. In the future, we plan to apply our method to improve the tracking of animals.

Acknowledgement

We thank the Research Council of Norway for funding this study through INID, the INTPART project no. 309857 and the BIONÆR program, project numbers 282252 and 321409. We also thank Rune Sagevik Norsvin SA for the image acquisition.

References

[1] Darwin V7 Labs. https://darwin.v7labs.com/.

[2] Hina Afridi, Mohib Ullah, Øyvind Nordbø, Faouzi Alaya Cheikh, and Anne Guro Larsgard. Optimized deep-learning-based method for cattle udder traits classification. *Mathematics*, 10 (17):3097, 2022.

[3] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. IEEE, sep 2016. doi: 10.1109/icip.2016.7533003. URL https://doi.org/10.1109%2Ficip.2016.7533003.

[4] Navaneeth Bodla, Bharat Singh, Rama Chellappa, and Larry S Davis. Soft-nms–improving object detection with one line of code. In *Proceedings of the IEEE international conference on computer vision*, pages 5561–5569, 2017.

[5] Justin Brooks. COCO Annotator. https://github.com/jsbroks/coco-annotator/, 2019.

[6] Binghui Chen, Pengyu Li, Xiang Chen, Biao Wang, Lei Zhang, and Xian-Sheng Hua. Dense learning based semi-supervised object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4815–4824, 2022.

[7] Dev Narayan et al. Tracking-by-self detection: A self-supervised framework for multiple animal tracking. In *International Conference on Artificial Intelligence Applications and Innovations*, pages 561–572. Springer, 2023.

[8] Jean-Bastien Grill et al. Bootstrap your own latent: A new approach to self-supervised learning. *CoRR*, abs/2006.07733, 2020. URL https://arxiv.org/abs/2006.07733.

[9] Jirka Borovec et al. Lightning-ai/lightning-bolts: Minor patch release, December 2022. URL https://doi.org/10.5281/zenodo.7447212.

[10] Kresovic et al. Pigpose: A realtime framework for farm animal pose estimation and tracking. In *Artificial Intelligence Applications and Innovations*, pages 204–215. Springer, 2022.

[11] Paszke et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[12] Zicong Jiang, Liquan Zhao, Shuaiyang Li, and Yanfei Jia. Real-time object detection method based on improved yolov4-tiny. *arXiv preprint arXiv:2011.04244*, 2020.

[13] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

[14] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8759–8768, 2018.

[15] Zachary T Pennington, Zhe Dong, Yu Feng, Lauren M Vetere, Lucia Page-Harley, Tristan Shuman, and Denise J Cai. eztrack: An open-source video analysis pipeline for the investigation of animal behavior. *Scientific reports*, 9(1):1–11, 2019.

[16] Patricia Pons, Javier Jaen, and Alejandro Catala. Assessing machine learning classifiers for the detection of animals' behavior using depth-based tracking. *Expert Systems with Applications*, 86: 235–246, 2017. ISSN 0957-4174.

[17] Akif Quddus Khan, Salman Khan, Mohib Ullah, and Faouzi Alaya Cheikh. A bottom-up approach for pig skeleton extraction using rgb data. In *International Conference on Image and Signal Processing, ICISP 2020*, pages 54–61. Springer, 2020.

[18] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE transactions on pattern analysis and machine intelligence*, 39(6):1137–1149, 2016.

[21] Mingxing Tan and Quoc V Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*, 2019.

[22] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[23] Peng Tang, Chetan Ramaiah, Yan Wang, Ran Xu, and Caiming Xiong. Proposal learning for semi-supervised object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2291–2301, January 2021.

[24] Mohib Ullah, Habib Ullah, and Faouzi Alaya Cheikh. Single shot appearance model (ssam) for multi-target tracking. *Electronic Imaging*, 2019(7):466–1, 2019.

[25] Mohib Ullah, Zolbayar Shagdar, Habib Ullah, and Faouzi Alaya Cheikh. Semi-supervised principal neighbourhood aggregation model for sar image classification. In *2022 16th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 211–217. IEEE, 2022.

[26] Mengde Xu, Zheng Zhang, Han Hu, Jianfeng Wang, Lijuan Wang, Fangyun Wei, Xiang Bai, and Zicheng Liu. End-to-end semi-supervised object detection with soft teacher. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3060–3069, October 2021.

[27] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.

[28] Lei Zhang, Helen Gray, Xujiong Ye, Lisa Collins, and Nigel Allinson. Automatic individual pig detection and tracking in pig farms. *Sensors*, 19(5):1188, 2019.