International Journal of Computer Vision (2023) 131:1163–1181
https://doi.org/10.1007/s11263-023-01748-3

Dynamic Curriculum Learning for Great Ape Detection in the Wild

Xinyu Yang[1] · **Tilo Burghardt**[1] · **Majid Mirmehdi**[1]

Received: 29 April 2022 / Accepted: 2 January 2023 / Published online: 16 January 2023

Abstract
We propose a novel end-to-end curriculum learning approach for sparsely labelled animal datasets leveraging large volumes of unlabelled data to improve supervised species detectors. We exemplify the method in detail on the task of finding great apes in camera trap footage taken in challenging real-world jungle environments. In contrast to previous semi-supervised methods, our approach adjusts learning parameters dynamically over time and gradually improves detection quality by steering training towards virtuous self-reinforcement. To achieve this, we propose integrating pseudo-labelling with curriculum learning policies and show how learning collapse can be avoided. We discuss theoretical arguments, ablations, and significant performance improvements against various state-of-the-art systems when evaluating on the Extended PanAfrican Dataset holding approx. 1.8M frames. We also demonstrate our method can outperform supervised baselines with significant margins on sparse label versions of other animal datasets such as Bees and Snapshot Serengeti. We note that performance advantages are strongest for smaller labelled ratios common in ecological applications. Finally, we show that our approach achieves competitive benchmarks for generic object detection in MS-COCO and PASCAL-VOC indicating wider applicability of the dynamic learning concepts introduced. We publish all relevant source code, network weights, and data access details for full reproducibility.

Keywords Semi-supervised learning · Curriculum learning · Great ape conservation · Species detection · Wildlife detection · MS-COCO · PASCAL-VOC

1 Introduction

Motivation-Automated visual monitoring of animals filmed in their natural habitats is gaining significant traction, boosted recently by a plethora of deep learning methods and applications (Tabak et al., 2019; Norouzzadeh et al., 2021; Tuia et al., 2022). However, developing and advancing relevant computer vision tools remains challenging due to several factors. Animals in their natural environments are often hard to detect, obscured by dynamic backgrounds, varying illumination conditions, occlusions, camouflage effects, and more. Deploying network models trained on prevalent image and video databases, such as ImageNet (Deng et al., 2009), MS-COCO (Lin et al., 2014), Kinetics (Carreira and Zisserman, 2017), are often insufficient on their own, even after taking advantage of the potentials of transfer learning. To further exacerbate the difficulty of deploying machine learning methods to their fullest extent in the domain, there is still a distinct lack of large-scale, annotated training datasets for particular species despite evolving general frameworks (Beery et al., 2019). Whilst crowd sourcing annotations can help, low labelling rates relative to archive sizes remain the norm in the field. For great apes in particular, several recent works have attempted to address some of the above mentioned challenges (Yang et al., 2019; Schofield et al., 2019; Sakib & Burghardt, 2021; Bain et al., 2021). However, these works still either only pretrain on datasets from other domains or rely on relatively small datasets for supervised training due to the complexities associated with obtaining annotations. Thus, while these methods have advanced the cause somewhat regarding great ape detection in jungle settings, they have also emphasised the urge for bet-

Communicated by Helge Rhodin.

✉ Xinyu Yang
xinyu.yang@bristol.ac.uk

Tilo Burghardt
tilo@cs.bris.ac.uk

Majid Mirmehdi
majid@cs.bris.ac.uk

[1] Department of Computer Science, University of Bristol, Bristol, UK

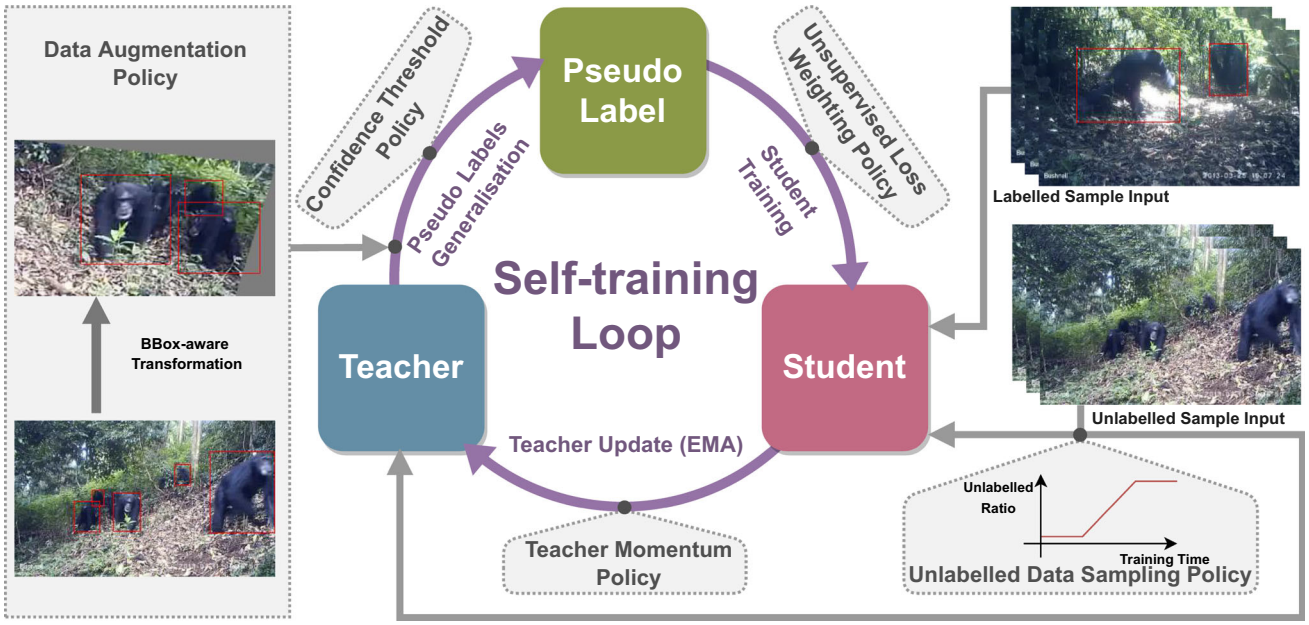


Fig. 1 Conceptual overview. We utilise a student-teacher paradigm for learning where the teacher produces pseudo-labels for the student to learn from while being updated by an exponentially moving average (EMA) of the student model. We apply five dynamic policies to this learning loop that we show can lead to effective (i.e. virtuous) self-training cycles: *unlabelled data sampling policy* to control the unlabelled sample input, *confidence threshold policy* to filter unreliable pseudo-labels, *data augmentation policy* to diversify the unlabelled training data, *unsupervised loss weighting policy* to balance unsupervised and supervised losses, and *teacher momentum policy* to adjust the update speed of the teacher model

ter use of the vast archives of completely unlabelled camera trap footage.

Paper Concept-In response, this paper introduces a novel curriculum learning approach that intertwines traditional supervised detector training with unlabelled data utilisation. The approach demonstrates by proof-of-concept that, exemplified for great apes, large unlabelled camera trap archives can indeed be exploited to enrich and empower real-world animal detector construction without any further labelling efforts. We leverage lessons learned from recent self-supervised (Grill et al., 2020; Caron et al., 2021; Chen and He, 2021) and semi-supervised (Sohn et al., 2020b, a; Xu et al., 2021) methods on feature representation learning and image classification to propose an end-to-end student-teacher based detection pipeline that integrates self-training (via pseudo-labels) and dynamic training polices into one cyclical curriculum learning design. Our model learns from unlabelled data in the curriculum by generating high quality pseudo-labels on the fly. In turn, these virtual annotations of otherwise unlabelled samples are exploited by the student whose update influences the teacher and a next round of pseudo-label generation. This cyclical self-training idea can be illustrated conceptually as a learning loop shown in Fig. 1. We will demonstrate that carefully fine-tuned curriculum learning policies in this loop can blend labelled and unlabelled sample input in a way that leads to virtuous training cycles (as opposed to vicious training cycles) which increasingly and consistently improve model performance. Critically, we show that dynamic learning adjustments can be controlled stably by policies and can improve performance over static learning. Intuitively, the approach expands model coverage of the vast space of animal appearance in particular, slowly from the labelled sample base, guided and channelled by the policies. We show that this approach can significantly improve great ape detection benchmarks, as well as other benchmarks including Bees and Snapshot Serengeti. We also demonstrate that the method is applicable beyond the targeted animal domain and achieves competitive or state-of-the-art results on the MS-COCO and PASCAL-VOC object detection challenges without a need for dataset-specific hyperparameter fine-tuning.

Contributions-Overall, the contributions of this paper can be summarised as, (i) a novel end-to-end *dynamic* detection framework for semi-supervised curriculum learning designed to improve species detectors built from sparsely labelled datasets, (ii) a dynamic policy system with stable hyper-parameters for temporal control over changing learning properties in semi-supervised detector training promoting self-reinforcing virtuous training loops, (iii) extensive experiments and ablations on a large scale real-world great ape camera trap dataset - we report improvements to the state-of-the-art for the semi-supervised great ape detection task

evaluated on the Extended PanAfrican Dataset, (iv) we offer new semi-supervised detection benchmarks on sparse labelling versions of two other animal datasets - Bees and Snapshot Serengeti, contributing towards handling annotation shortage in the animal domain, and finally (v) we also provide competitive and state-of-the-art semi-supervised object detection results for the MS-COCO and PASCAL-VOC datasets, demonstrating broader applicability.

2 Related Work

In this section, we consider works related to the key topics of interest with focus on the state-of-the-art.

Semi-supervised Learning (SSL)-SSL exploits the potential of unlabelled data to facilitate model learning with limited amounts of annotated data (Rebuffi et al., 2020). Training computer vision models such as objection detection or action recognition networks, relies on the availability of annotated datasets which can be costly to generate. This has motivated the development of semi-supervised methods (Jeong et al., 2019; Berthelot et al., 2019; Zhai et al., 2019; Sohn et al., 2020a, b; Zhang et al., 2021; Xu et al., 2021; Tang et al., 2021).

One dominant SSL approach is consistency regularisation where the model is regularised to generate consistent predictions on data with different augmentations (Jeong et al., 2019; Berthelot et al., 2019; Zhai et al., 2019). Another approach is based on generating pseudo-labels for unlabelled data and updating the model by training on a mix of unlabelled data with pseudo-labels and labelled data with manually-annotated labels (Sohn et al., 2020a, b; Zhang et al., 2021; Xu et al., 2021; Tang et al., 2021; Liu et al., 2021). What type of pseudo-labelling to use is critical to the success of SSL in particular scenarios. FixMatch (Sohn et al., 2020a) applied a high confidence threshold for mining pseudo-labels and then these sharpened and strongly-augmented pseudo-labels were utilised for model training. STAC (Sohn et al., 2020b) extended FixMatch from image classification to objection detection by introducing self-training and augmentation-driven consistency regularisation. More recently, Xu et al. (2021) introduced the soft teacher mechanism to alleviate the issue of unreliable pseudo-labels generated by the teacher in SSL object detection. Liu et al. (2021) jointly train a student and a teacher in a mutually-beneficial manner by applying a class-balance loss to down-weight overly confident pseudo-label impact. In the light of the success of these methods, our approach follows the pseudo-labelling concept, but addresses the model learning challenges differently.

Object detection-This area of computer vision has advanced in leaps and bounds since the very start of the modern era of deep learning. Some notable early works are: (i) single-stage detection frameworks, such as (Redmon et al., 2016; Liu et al., 2016; Lin et al., 2017b; Tian et al., 2019), which perform object classification and bounding box regression directly, without using pre-generated region proposals. They are typically applied over a dense sampling of possible object locations to estimate the class probabilities and bounding box coordinates directly. (ii) In contrast, two-stage detection frameworks, such as (Ren et al., 2015; He et al., 2017; Lin et al., 2017a) utilise a region proposal network to generate class-agnostic regions of interest (ROIs) and only then perform ROI bounding box regression and object classification. More recently, DEtection with TRansformers (DETR) (Carion et al., 2020) built the first end-to-end detection pipeline by viewing object detection as a direct set-prediction problem. DETR eliminated the need for anchor-based target assignment pre-processing and non-maximum suppression (NMS) post-processing, prevalent in commonly used object detectors. It combined CNNs for feature extraction and transformers for feature interpretation to directly translate object queries to class and bounding boxes by leveraging cross attention (Vaswani et al., 2017) on image features. However, the vanilla DETR suffers from slow convergence and hence longer training time than detectors based on YOLO, SSD and Faster-RCNN. The Deformable DETR (Zhu et al., 2020) proposed a deformable attention module that only attend to a small set of prominent key elements to replace the attention in DETR. This improvement led to faster convergence and a better performance. We select this variant as our model for the various detection components of our proposed curriculum learning framework.

Curriculum Learning (CL)-The CL training approach (Bengio et al., 2009; Wang et al., 2022b) has had significant impact on the design of computer vision algorithms, such as (Karras et al., 2018; Wang et al., 2018; Huang et al., 2020; Wang et al., 2022a; Zhang et al., 2021). Wang et al. (2018), for instance, use average precision of each sample to re-rank the data from easy to hard and train the object detector in an easy-to-hard fashion applied to the pre-ranked order of data. Wang et al. (2022a) propose a pseudo-labelled auto-curriculum learning framework that engages reinforcement learning to learn a series of dynamic thresholds for the pseudo-labels for semi-supervised key-point localisation. FixMatch (Sohn et al., 2020a), on the other hand, applied a constant threshold to select unlabelled samples for training, which fails to address the learning difficulties at different time steps. Thus, it can allow poor quality samples to get through. FlexMatch (Zhang et al., 2021) improved on FixMatch by dynamically adjusting the threshold at each time step to filter unlabelled samples and pseudo-labels.

Both FixMatch and FlexMatch applied a pre-trained pseudo-label generator which does not get updated during the semi-supervised learning stage, thus, failing to consider the evolution of the pseudo-label generator as the learning progressing. To address this issue, we propose student-teacher

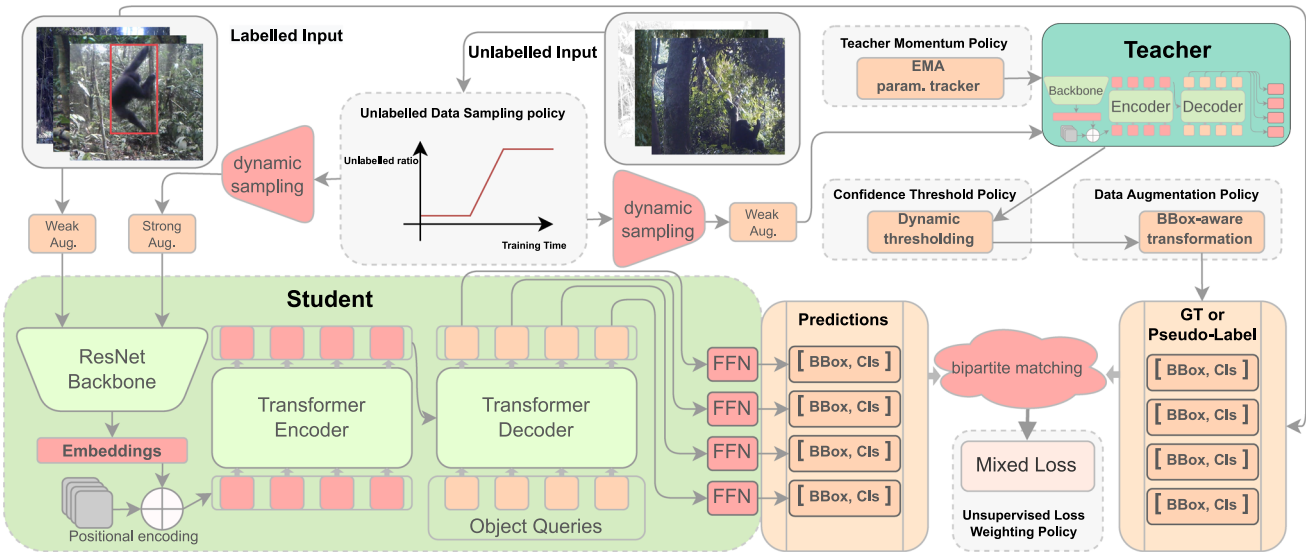


Fig. 2 Detailed end-to-end self-training great ape detection pipeline. We utilise the Deformable DETR (Zhu et al., 2020) framework with a ResNet backbone as detector architecture. The student network (light green) uses this architecture as well as the teacher network (dark green). All labelled data along with dynamically sampled and policy-controlled unlabelled data are mixed during training. The teacher performs pseudo-label generation with purely unlabelled input on the fly. The pseudo-labels are filtered with an adaptive threshold and then augmented via a bounding box-aware transformation. The teacher network is updated by student model via a dynamic momentum coefficient. The final loss is the sum of supervised and unsupervised detection losses balanced by a policy-controlled dynamic weight. We carefully designed the policies of the system to achieve an effective (a.k.a. virtuous) self-reinforcing training cycle (Color figure online)

learning paradigms inspired by the recent advances in self-supervised learning methods (Grill et al., 2020; Chen and He, 2021; Caron et al., 2021) that evolve the teacher component dynamically guided by a set of curriculum learning policies and controls. We will now describe our approach in detail.

3 Proposed Method

We introduce an end-to-end curriculum learning pipeline for effective semi-supervised Great Ape detection in camera trap footage. Our framework follows a student-teacher training scheme, as illustrated in detail in Fig. 2 and operates as follows: in each learning iteration, we train a student model built around a Deformable DETR detector (Zhu et al., 2020) by a mix of labelled and unlabelled videos where the unlabelled videos are sampled by our curriculum sampling policy π. The teacher performs pseudo-label generation with unlabelled input. Pseudo-labels are then refined by a dynamic threshold ς_t and transformed by augmentation policy $\mathcal{A}$. Together, both the pseudo-labels and manually-annotated labels are fed into the student network for learning. The student network is then updated by the gradient from the overall loss which is balanced by unsupervised loss weight α_t. Finally, the teacher network is updated by the exponential moving average (EMA) of the student parameters via a dynamic momentum coefficient m_t. This completes one iteration of the learning loop leading to an updated teacher and student model. Our target will be to design the mentioned policies and an appropriate loss in a fashion that virtuous, that is effective, learning can be practically achieved.

3.1 Problem Definition

Let us consider that frames are sampled from the video at a frequency of ω for both labelled and unlabelled videos. The teacher is trained to generate the pseudo-labels for unlabelled frames only, while the student is trained to fit the pseudo-labels with the unlabelled input frames, as well as the ground-truth labels with the labelled input frames. Thus, the overall loss for the student is defined as the weighted sum of supervised and unsupervised losses:

$$\mathcal{L}_{all} = \mathcal{L} + \alpha \mathcal{L}', \tag{1}$$

where $\mathcal{L}$ and $\mathcal{L}'$ denote the supervised loss of labelled samples and unsupervised loss of unlabelled samples respectively, and α represents the balancing weight.

Further consider we have a labelled sample set D_X with M labelled samples X_i and their corresponding class and box labels (C_i, B_i) and an unlabelled sample set D_U with N unlabelled samples U_j (regardless of the sampling approach) used for training. Also, let $\mathfrak{S}(\theta)$ be the student model param-

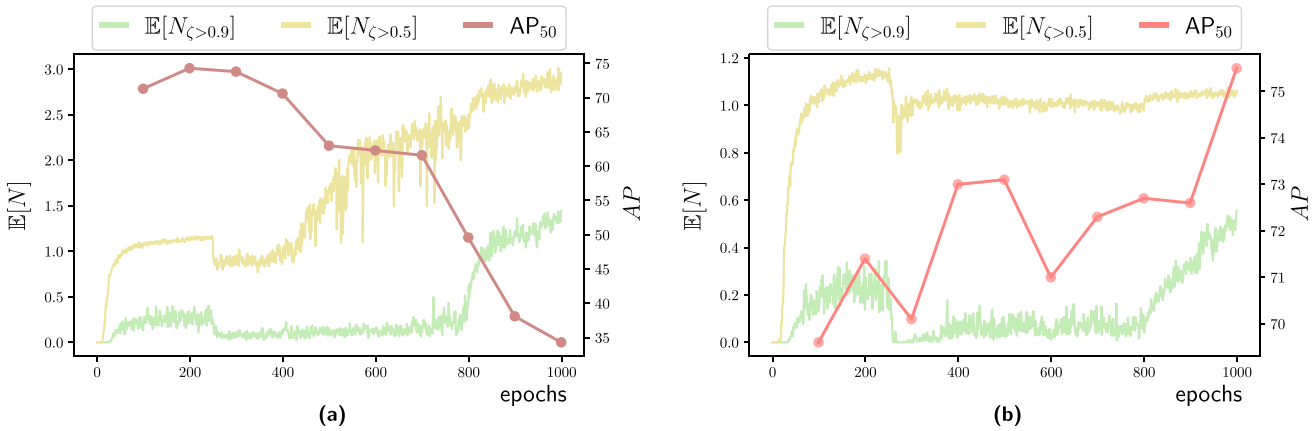


Fig. 3 Bipolar behavioural dynamics of learning via self-training loops. Two representative cases illustrating the bipolar dynamics of training success when using the proposed architecture: *vicious* collapse in (**a**) and *virtuous* effective learning in (**b**). The scenarios differ only in policy parameterisation. (For reproducibility of behaviour in Fig. 3, the exact parameters used were: (a) π: constant policy, constant $\varsigma = 0.1$ and constant $\alpha = 0.1$; (b) π: linear increase policy, linear increase ς $0.3 \rightarrow 0.5$, constant $\alpha = 0.5$. Both were trained for 1000 epochs with the first 250 epochs for warmup, lr decreases at 800^{th} epoch). In both plots, the right ordinate indicates the AP_{50} on the validation set whilst the left ordinate represents the average number of samples with confident score $\zeta > 0.9$ or $\zeta > 0.5$

eterised by θ, and $\mathfrak{T}(\theta')$ be the teacher model parameterised by θ'. Eq. (1) can then be expanded to:

$$\begin{aligned}\mathcal{L}_{all}(\theta) &= \mathcal{L} + \alpha\mathcal{L}' \\ &= \frac{1}{M}\sum_{X_i \in D_X} L_\theta\left(X_i\right) + \alpha\frac{1}{N}\sum_{U_j \in D_U} L'_\theta\left(U_j\right) ,\end{aligned} \tag{2}$$

where $L_\theta\left(X_i\right)$ is the loss for labelled sample X_i,

$$\begin{aligned}L_\theta\left(X_i\right) &= \text{Loss}\left(\mathfrak{S}\left(X_i,\theta\right),[C_i,B_i]\right) \\ &= L_{reg}\left(\mathfrak{S}\left(X_i,\theta\right),B_i\right) + L_{ce}\left(\mathfrak{S}\left(X_i,\theta\right),C_i\right) ,\end{aligned} \tag{3}$$

and $L'_\theta\left(U_j\right)$ is the loss for the unlabelled sample U_j,

$$\begin{aligned}L'_\theta\left(U_j\right) &= \text{Loss}\left(\mathfrak{S}\left(U_j,\theta\right),\mathfrak{T}\left(U_j,\theta'\right)\right) , \\ &= L_{reg}\left(\mathfrak{S}\left(U_j,\theta\right),\mathfrak{T}\left(U_j,\theta'\right)\right) \\ &\quad + L_{ce}\left(\mathfrak{S}\left(U_j,\theta\right),\mathfrak{T}\left(U_j,\theta'\right)\right) ,\end{aligned} \tag{4}$$

where L_{reg} represents the bounding box regression loss and L_{ce} represents the classification loss.

We follow common practice in self-supervised learning methods, such as (Caron et al., 2021; Grill et al., 2020), so that the teacher is updated by the EMA of the student,

$$\theta'_t \leftarrow m\theta'_{t-1} + (1-m)\theta_t . \tag{5}$$

Our objective is to find a set of student parameters θ^* that minimises the expected overall loss $\mathcal{L}_{all}(\theta)$, such that

$$\theta^* = \arg\min_\theta \mathcal{L}_{all}(\theta) . \tag{6}$$

3.2 Self-reinforcing Training Loop

The evolution of the student and teacher network is conceptually a cyclic relationship. On the one hand, the performance of the student detector depends on the quality of the pseudo-labels, which in turn relies on the teacher, and on the other hand, the teacher is updated according to student status. Thus, there is an intricate interdependence between the student, the teacher, and the pseudo-labels forming a self-training loop which is controlled by the learning policies.

In practice, we observe a bipolarisation phenomenon for the training of models with different settings where they gradually become more confident of their predictions, but show two drastically different performance trajectories. As illustrated on sample training runs shown in Fig. 3, whilst a gradual increase of confidence indicators $\mathbb{E}[N_{\zeta>0.9}]$ and $\mathbb{E}[N_{\zeta>0.5}]$, which represent the average number of predicted objects whose confident scores ζ are over 0.9 or 0.5 respectively, can be observed; the validation performance can be erratic, either collapsing or improving effectively. In the example illustrated in Fig. 3a, a decrease of AP_{50} on the validation set was observed after a few hundred training epochs. Initially, one may assume the model is simply over-fitting at this stage in learning. However, as shown in the second example in Fig. 3b for a different parameterisation, a long-term increase in AP_{50} can be observed, which suggests the model can learn from the training set well into training cycles. We hypothesise that the bipolar collapse or success of learning with regard to generalisation is critically linked to the self-reinforcement property of the training loop parameterisation and policies.

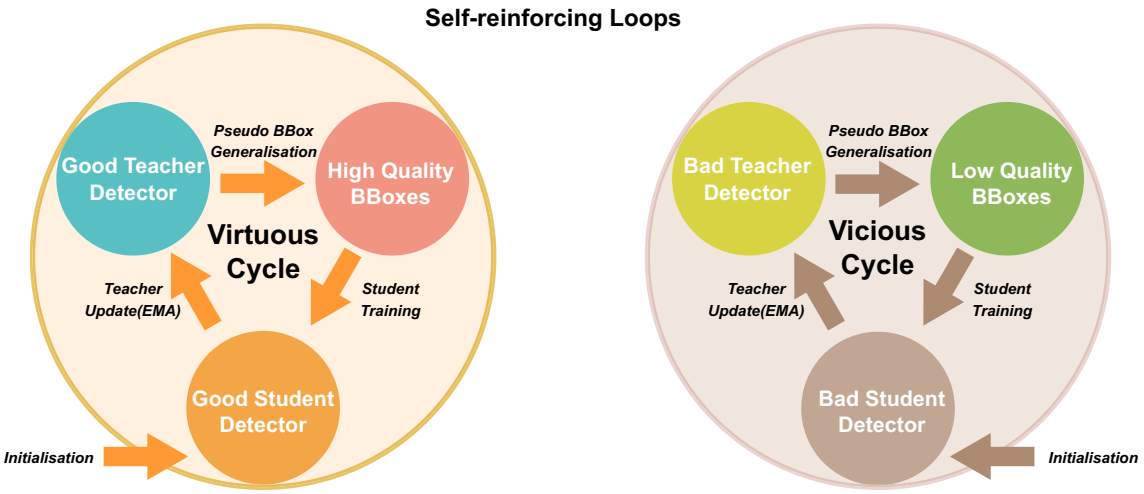


Fig. 4 Processes within self-reinforcing Loops. Illustration of four key processes (arrows) involved in training loops. Note that a destabilised *Virtuous Cycle* where low quality pseudo labels or highly inaccurate student or teacher networks are produced turns into a *Vicious Cycle* and vice versa. Thus, effective parameterisations and policies for the key processes are required to to promote stable learning

We categorise bipolarisation as two different types of learning cycles, effective *Virtuous Cycles* and collapsing *Vicious Cycles* as illustrated in Fig. 4. In the virtuous cycle state, the teacher model generates pseudo-labels of sufficient quality as to contribute to the training of the student model, allowing both models to improve continually. In contrast, the vicious cycle sees the teacher generate insufficiently low quality pseudo-labels that degrade the training of the student model, thus both models degenerate continually.

Fig. 4 depicts four key processes in the learning loop (shown as arrows), which are crucially influencing the trajectory of learning: (i) *Initialisation*: initialising the student model before the self-training phase; (ii) *Teacher Update*: updating the teacher network according to student status; (iii) *Pseudo-label Generalisation*: generating pseudo-labels by teacher; (iv) *Student Training*: using pseudo-labels to update student. Our goal is to find suitable controls that guide the above processes and can maintain the development of a virtuous self-training loop and, for robustness, also transition from a vicious to a virtuous setting.

For initialisation, we confirmed experimentally that the proposed system operates in a stable manner with fixed, standard backbone initialisations across all tested datasets. In particular, we use the self-supervised ImageNet pre-trained ResNet weights from SWAV (Caron et al., 2020) for our detection backbone. In addition, to show that general training stability can also be maintained in a supervised initialisation scenario, we test supervised ImageNet pre-trained ResNet (He et al., 2016) weights too. Note that any such fixed initialisation is essential as random initialisation triggers vicious training cycles, however, the fix is not sensitive to target dataset properties as transfer between scenarios still produces stable learning (see Sect 7).

For the other three processes above, we propose appropriate 'policies' that guide learning within the confounds of effective 'virtuous' learning cycles—these are described next.

3.3 Student Training

Student network training is guided by two policies that allow the student model to exploit the unlabelled sample data and their pseudo-labels effectively.

Unlabelled Data Sampling Policy controls the number of unlabelled samples to use in the self-training loop at different time steps. It can be expressed with an additional Bayesian prior on Eq.(2), i.e.

$$\begin{aligned}\mathcal{L}_\pi(\theta) &= \hat{\mathbb{E}}\left[L_\theta\right] \\ &= \frac{1}{M}\sum_{X_i \in D_X} L_\theta\left(X_i\right) + \alpha \frac{1}{N\mathbb{E}[\pi]} \sum_{U_j \in D_U} L'_\theta\left(U_j\right)\pi(U_j),\end{aligned} \tag{7}$$

where $\pi(U_j)$ is the probability for using the unsupervised loss of U_j in the self-training stage. For simplicity, we substitute $\alpha \frac{\pi(U_j)}{N\mathbb{E}[\pi]}$ with $p(U_j)$ in Eq. (7) to obtain:

$$\begin{aligned}\mathcal{L}_\pi(\theta) &= \frac{1}{M}\sum_{i=1}^{M} L_\theta\left(X_i\right) + \sum_{j=1}^{N} L'_\theta\left(U_j\right) p(U_j) \\ &= \frac{1}{M}\sum_{i=1}^{M} L_\theta\left(X_i\right) + \sum_{j=1}^{N} \left(L'_\theta\left(U_j\right) p(U_j) - \hat{\mathbb{E}}[L'_\theta] p(U_j)\right.\end{aligned}$$

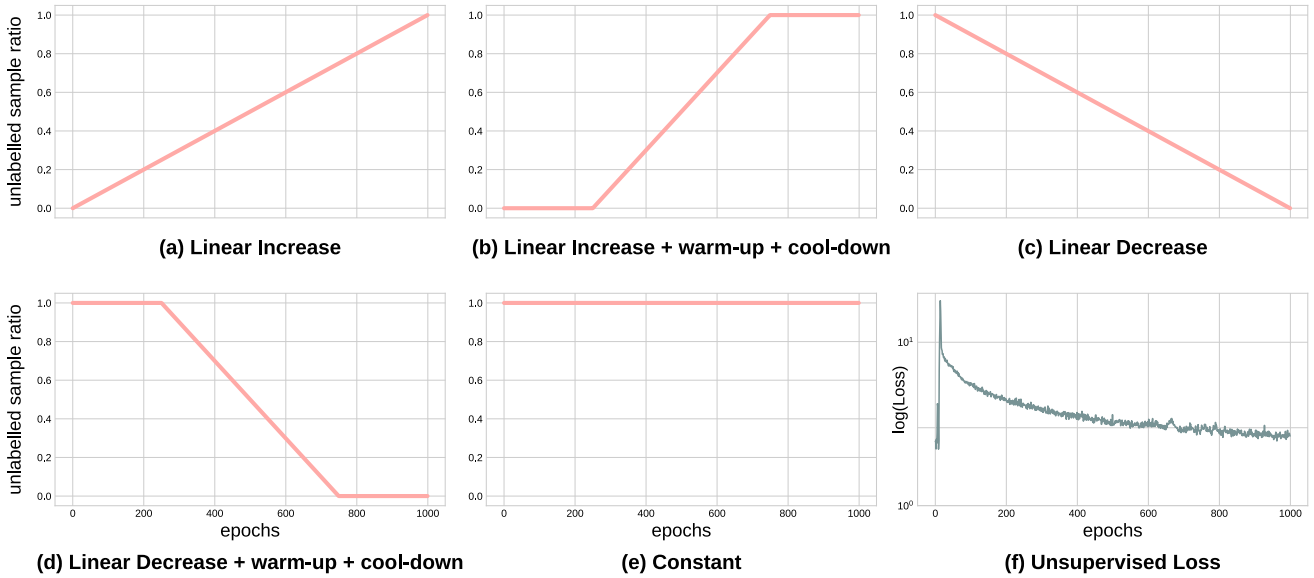


Fig. 5 Unlabelled data sampling policies & Unsupervised Loss Dynamics. (**a**) models a linear increase of the unlabelled data ratio from 0 to 1 over epochs, (**b**) combineswarm-up and cool-down phases with linear increase, (**c**) linear decrease of the unlabelled data ratio, (**d**) combines linear decrease with warm-up and cool-down phases, (**e**) keeps a constant ratio, and (**f**) shows the unsupervised loss L'_θ observed in training

$$
\begin{aligned}
&- \hat{\mathbb{E}}[p]L'_\theta\left(U_j\right) + \hat{\mathbb{E}}[L'_\theta]\hat{\mathbb{E}}[p]\Big) + N\hat{\mathbb{E}}[L'_\theta]\hat{\mathbb{E}}[p] \\
&= \frac{1}{M}\sum_{i=1}^{M} L_\theta\left(X_i\right) + N\hat{\mathbb{E}}[L'_\theta]\hat{\mathbb{E}}[p] \\
&\quad + \sum_{j=1}^{N}(L'_\theta\left(U_j\right) - \hat{\mathbb{E}}[L'_\theta])(p(U_j) - \hat{\mathbb{E}}[p]) \\
&= \frac{1}{M}\sum_{i=1}^{M} L_\theta\left(X_i\right) + N\hat{\mathbb{E}}[L'_\theta]\hat{\mathbb{E}}[p] + N\hat{\mathrm{Cov}}[L'_\theta, p]\,.
\end{aligned}
\tag{8}
$$

Based on the definition of $\mathcal{L}_{all}(\theta)$ in Eq. (2), Eq. (8) can be simplified to:

$$
\mathcal{L}_\pi(\theta) = \mathcal{L}_{all}(\theta) + N\hat{\mathrm{Cov}}[L'_\theta, p] \tag{9}
$$

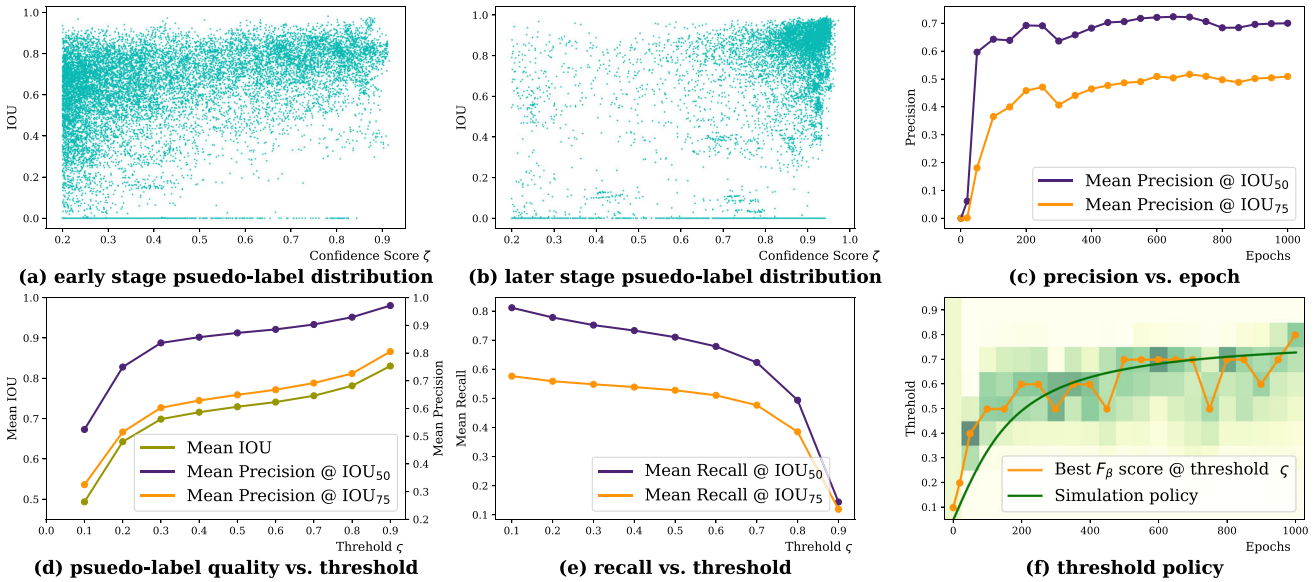
The goal is to search for the best unlabelled data sampling policy π^* that can yield the lowest possible loss for Eq. (9), such that:

$$
\begin{aligned}
\pi^* &= \arg\min_\pi \mathcal{L}_\pi(\theta) \\
&= \arg\min_\pi \mathcal{L}_{all}(\theta) + N\hat{\mathrm{Cov}}[L'_\theta, p] \\
&= \arg\min_\pi \hat{\mathrm{Cov}}[L'_\theta, p]
\end{aligned}
\tag{10}
$$

Equation (10) suggests that if L'_θ and p are negatively correlated then we can arrive at an effective policy π^*. Given that p is positively correlated with π, since $\frac{\alpha}{N\mathbb{E}[\pi]}$ is positive, an effective unlabelled data sampling policy π^* should be negatively correlated to L'_θ. The model gets updated for each iteration, thus one may assume naïvely that $L'_\theta(U_{j+1}) < L'_\theta(U_j)$, because $L'_\theta(U_{j+1})$ is generated after backpropagation of $L'_\theta(U_j)$. In practice, during training, we also observed such a decrease of $\mathbb{E}[L'_\theta]$ as shown in Fig. 5(f).

In summary, considering π^* and L'_θ are negatively correlated, and L'_θ is indeed decreasing over time, we can conclude that π can consequently be obtained via cyclical curriculum learning. Practically, this may be carried out via a gradual increase of unlabelled sample input in the ways shown in Fig. 5a or b, where the latter includes warm-up and cool-down periods. Conceptually, these policies expand learning slowly but steadily towards the unexplored data domain in order to allow for a gradual expansion of high quality model expertise and prevent erratic learning collapse. For comparison and to emphasise the importance of this policy choice, we later also experimentally examine other policies depicted in Figs. 5c, d, and e.

Unsupervised Loss Weighting Policy is tasked with balancing the weighting between the supervised and unsupervised losses. The performance of the student detector depends on the quality of the pseudo-labels. Fig. 6a and b depict pseudo-label distributions captured at an early stage and a late stage of the training, respectively, plotted against the confidence score ζ. Label confidence and IOU quality clearly increase over training at these snapshot points. The associated @IOU$_{50}$ and @IOU$_{75}$ precision curves in Fig. 6c illustrate that the average quality of the pseudo-labels increases over time. We



(a) early stage psuedo-label distribution **(b) later stage psuedo-label distribution** **(c) precision vs. epoch**

(d) psuedo-label quality vs. threshold **(e) recall vs. threshold** **(f) threshold policy**

Fig. 6 Pseudo-label analysis. We use 70k pseudo-labels generated from the teacher network to conduct this analysis. Note that unlabelled samples do not use ground truth in training, but we use it for this analysis. (**a**) and (**b**) show the distributions of each pseudo-label's ground truth IOU against the confidence score ζ visualised at the early stage (100^{th} epoch snapshot) and later stage (800^{th} epoch snapshot) of training, respectively. (**c**) the pseudo-label mean precision at IOU_{50} and IOU_{75} without applying a threshold over the training epochs. (**d**) pseudo-label quality against the threshold ς represented by IOU_{50} and IOU_{75} precision averaged across epochs, and IoU of real ground truth labels averaged on all epochs. (**e**) mean recall at IOU_{50} and IOU_{75} for different ς values, (**f**) heatmap indicating the normalised F_β score for ς_t at different epochs - light to dark colours for low to high scores, orange shows the best F_β score at each epoch, and the dark green plot represents our confidence threshold policy as the arctan function that approximates the best F_β scores (Color figure online)

note that recent works (Sohn et al., 2020b; Xu et al., 2021; Tang et al., 2021) on this topic only applied a fixed weighting to all pseudo-labels throughout the training. Yet, given this observed gradual change in pseudo-label quality, there is an opportunity to design an adaptive weighting policy that applies smaller unsupervised loss weights for less reliable pseudo-labels in the early training stages and larger weights for more reliable pseudo-labels generated in the later training stages.

To implement this, we use a curriculum learning approach for the unsupervised loss weighting parameter α, which is made subject to an adaptive weighting policy. Theoretically, more optimal policies would keep track of the bounding-box pseudo-label qualities. However, in practice, this is hard to do on the fly due to the unavailability of the ground truth and extensive computational needs. We thus opt for a simple linear increase of α as a first approximation.

3.4 Pseudo-label Generation

We use two policies to generate reliable pseudo-labels from the teacher's output to promote a virtuous cycle for training. *Confidence Threshold Policy* allows us to examine the reliability of pseudo-labels at different threshold values of ς, ranging from 0.1 to 0.9, and averaged across epochs. Our aim is to select an optimised value in order to discard unreliable pseudo-labels most effectively.

Three metrics are applied to assess the quality of the pseudo-labels: IOU, Precision @IOU_{50} and Precision @IOU_{75}. The plots in Fig. 6d for all measures show that they increase as ς does. Trivially, the higher the value of ς, the higher the probability of obtaining more reliable pseudo-labels. So for highest quality one could select $\varsigma = 0.9$. However, as a consequence the recall rate is significantly suppressed, with both mean @IOU_{50} and @IOU_{75} recalls of course negatively correlated to ς (see Fig. 6e). For example, when $\varsigma = 0.9$, mean precision reaches approx. 95%, while the mean recall drops to approx. 15%.

To address this issue, our dynamic confidence threshold policy increases the quality of pseudo-labels by controlling false negatives explicitly and thereby balancing precision and recall. We apply the F_β score which is the weighted harmonic mean of precision and recall, with $\beta = 0.5$ to allow the F_β score assign more weight to the precision than the recall on the basis that the false positives have more negative impact than the false negatives in our pipeline. Compared to anchor-based detectors which assign missed bounding boxes as negatives (non-object class), such as Ren et al. (2015); He et al. (2017); Lin et al. (2017b), the DETR family of methods do not. Their bipartite matching stage only matches the predictions with the ground truth, thus any missed bounding

Fig. 7 Augmentation strategy. Visualisation of colour augmentation and geometric augmentation examples used in the experiments. Augmentations are selected such that the results reflect the variance found across different camera and acquisition settings commonly seen in the dataset. (best viewed under zoom)

boxes are ignored and there are no penalties for this in training. Further, bounding box-aware crops are applied in the augmentation stage, thus false negative areas could be wiped out in the image. For example, see the chimp on the right side of the first image in the last column in Fig. 7, which if undetected by the teacher, it will disappear after a geometric transformation, as shown in the last image in the last column.

After fixing β, which may be changed for different application scenarios, the goal of our confidence threshold policy is to search for the threshold ς at time step t that can maximise F_β, i.e.

$$\varsigma_t^* = \arg\max_{\varsigma_t} F_\beta(\mathcal{P}_t, \mathcal{R}_t|\varsigma_t), \tag{11}$$

where $\mathcal{P}_t$ and $\mathcal{R}_t$ represent the precision and recall rates at time step t, determined by threshold ς_t. The heatmap in Fig. 6(f) shows the normalised F_β score for each time step (darker colour means higher value), with the best F_β at threshold ς_t shown in the orange plot. We use the approximate fit of the arctan function (dark green plot) to represent our confidence threshold policy.

Data Augmentation Policy ensures consistent augmentation of unlabelled data under the pseudo-labels produced by the teacher.

This is an indispensable element in semi-supervised and self-supervised methods. Self-supervised methods, such as DINO (Caron et al., 2021) and BYOL (Grill et al., 2020) minimise different views of data generated by data augmentation. Recently, semi-supervised methods such as FixMatch (Sohn et al., 2020a) and STAC (Sohn et al., 2020b), use augmentation-driven consistency regularisation for classification and detection.

Following STAC, we explore different variants of transformations on the Extended PanAfrican Dataset as our augmentation policy $\mathcal{A}$. We apply transformation operations in sequence as follows: first, we randomly apply bounding-box-aware crop and resize on the image, and then we apply a randomly-selected geometric transformation, followed by a random transformation on the colour statistics of the image (see code for all details).

Finally, for strong augmentation $\mathcal{A}_s$, we apply random erase (Zhong et al., 2020) or cutout (DeVries and Taylor, 2017) at multiple random locations of the whole image. For a weak augmentation $\mathcal{A}_w$, we just decrease the intensity for each transformation. Some examples are shown in Fig. 7 to illustrate the augmentation process.

3.5 Teacher Update

Teacher Momentum Policy controls the update speed of the teacher model and is encapsulated in the momentum coefficient m. In Fig. 6c, we can see the pseudo-label precision rate increases steeply in the early training stages, but slowly in the later training stages. Given that learning happens faster in the early stage, it motivates us to design a dynamic momentum policy which takes this fact onboard and stabilises teacher updates. Eq. (5) suggests that a lower momentum coefficient allows faster updates of the teacher model. To match the learning speed of the model at different time steps, we use a lower momentum coefficient m at the early stages and gradu-

Algorithm 1 Semi-supervised Training of Great Ape Detector using all Policies

Require: D_X, D_U ▷ labelled and unlabelled data
Require: $\mathfrak{S}_\theta$, $\mathfrak{T}_{\theta'}$ ▷ student and teacher models
Require: Π ▷ curriculum learning strategy
$\mathfrak{S}_\theta \leftarrow$ initialisation $\mathfrak{T}_{\theta'} \leftarrow$ initialisation ▷ initialise student and teacher models
do
  $\pi_t, \varsigma_t, \alpha_t, m_t, \mathcal{A} \leftarrow \Pi(\mathrm{t})$ ▷ instantiate five policies at each time step
  $X_i, [C_i, B_i] \leftarrow mini\text{-}batch(D_X)$ ▷ sample labelled mini-batch data
  $U_j \leftarrow mini\text{-}batch(D_U)$ ▷ sample unlabelled mini-batch data
  $U_j \leftarrow \pi_t(U_j)$ ▷ apply unlabelled data sampling policy
  $\mathcal{A}_s, \mathcal{A}_w \leftarrow \mathcal{A}$ ▷ sample strong and weak augumentations
  $[C_j, B_j], \zeta_j \leftarrow \mathfrak{D}_{\theta'}(\mathcal{A}_w(U_j))$ ▷ generate pseudo-labels by teacher
  $\widehat{U_j}, [\widehat{C_j}, \widehat{B_j}] \leftarrow \mathbb{1}(U_j, [C_j, B_j] \mid \zeta_j > \varsigma_t)$ ▷ apply confidence threshold policy
  $\mathcal{L} \leftarrow Loss\big(\mathfrak{D}_\theta(\mathcal{A}_w(X_i)), \mathcal{A}_w([C_i, B_i])\big)$ ▷ get supervised loss with ground truth
  $\mathcal{L}' \leftarrow Loss\big(\mathfrak{D}_\theta(\mathcal{A}_s(\widehat{U_j})), \mathcal{A}_s([\widehat{C_j}, \widehat{B_j}])\big)$ ▷ get unsupervised loss with pseudo-labels
  $\mathcal{L}_{all} \leftarrow \mathcal{L} + \alpha_t \mathcal{L}'$ ▷ apply unsupervised loss weighting policy
  $\Delta\theta \leftarrow -\nabla_{\mathcal{L}_{all}}\theta$ ▷ backpropagate the overall loss
  $\theta \leftarrow \theta + \Delta\theta$ ▷ undate student networks by gradient
  $\theta' \leftarrow m_t\theta' + (1 - m_t)\theta$ ▷ undate teacher with teacher momentum policy
  $t \leftarrow t + 1$ ▷ next time step and repeat
until $\mathcal{L}_{all}$ converge
end

ally increase it with time. In practice, we use a cosine increase of m in our pipeline which has also been explored in DINO (Caron et al., 2021). In experiments, we find that this dynamic momentum policy leads to consistently better performance than a constant one (see Table 2).

3.6 Combined Policy Application

All five policies are implemented in unison as a dynamic curriculum learning strategy Π for our wildlife detection pipeline, comprising the unlabelled data sampling policy $\pi_t = \Pi_\pi(t)$, the unsupervised loss weighting policy $\alpha_t = \Pi_\alpha(t)$, the confidence threshold policy $\varsigma_t = \Pi_\varsigma(t)$, the data augmentation policy $\mathcal{A} = \Pi_\mathcal{A}(t)$ and the teacher momentum policy $m_t = \Pi_m(t)$. Algorithm 1 illustrates this curriculum learning strategy $\Pi = \{\pi_t, \alpha_t, \varsigma_t, \mathcal{A}, m_t\}$ in its complete form.

4 Experiments

Datasets-We test our method on the Extended PanAfrican Dataset from the PanAf programme (Max-Planck-Institute, 2022) which contains camera-trap footage captured in natural Great Ape habitats in central Africa. There are two major species of Great Apes in the dataset, gorillas and chimpanzees. The archive footage contains around 20K videos adding up to around 600 hours. We use a subset of 5219 videos, with 500 videos (totalling over 180K frames) manually annotated with per frame great ape location bounding boxes, species and further categories (Yang et al., 2019; Sakib & Burghardt, 2021). This labelled data is split into `trainset`, `valset`, `testset` at a ratio of 80%, 5%, 15% respectively. All labels and metadata are fully published (Yang et al., 2019) and source videos may be obtained as detailed in the Acknowledgements.

Following standard evaluation protocols as used in (Sohn et al., 2020b; Xu et al., 2021; Zhou et al., 2021; Liu et al., 2021), we utilise the Extended PanAfrican Dataset for system training and benchmarking under two general paradigms:

1. *Partially Labelled Data (PLD)*. In this setting, either 10%, 20%, or 50% of the annotated `trainset` data are sampled as labelled training data, and the complete remainder of all data is used as unlabelled data. For each quantity, we create 3 different data folds and report the performance on `testset` with mean average precision (mAP) as the evaluation metric.
2. *Fully Labelled Data (FLD)*. In this setting, the whole annotated `trainset` is utilised as the labelled training data and only the remaining ∼5K unlabelled videos, totalling ∼1.8M frames, are used as additional unlabelled data.

In addition, we investigate two other animal datasets under sparse labelling settings - Bees[1] and Snapshot Serengeti (Swanson et al., 2015) - to explore system effectiveness under sparse labelling regimes further across the domain of animal visuals. We also present results on the MS-COCO (Lin et al., 2014) and PASCAL-VOC (Everingham et al., 2010) datasets to explore wider applicability of the introduced concepts to mainstream object detection.

[1] Available at https://lila.science/datasets/boxes-on-bees-and-pollen.

Table 1 Results and detailed comparative evaluation on the extended panAfrican dataset

Method	Labelled ratio	Setting	mAP	mAP_{50}	mAP_{75}
Supervised baseline	10%	PLD	32.17 ± 0.70	75.57 ± 1.24	21.40 ± 2.45
STAC (Sohn et al., 2020b)	10%	PLD	38.04 ± 3.88	73.31 ± 7.01	35.34 ± 2.31
SoftTeacher* (Xu et al., 2021)	10%	PLD	39.37 ± 7.97	63.03 ± 11.42	44.50 ± 9.72
Ubteacher* (Liu et al., 2021)	10%	PLD	44.03 ± 0.26	76.69 ± 2.15	47.25 ± 1.21
Ours	10%	PLD	**45.96** ± 2.97	**78.10** ± 6.14	**47.67** ± 3.12
Supervised baseline	20%	PLD	46.93 ± 1.30	86.47 ± 0.74	46.00 ± 2.42
STAC	20%	PLD	51.35 ± 2.39	83.71 ± 2.24	56.58 ± 4.12
SoftTeacher*	20%	PLD	50.87 ± 2.99	79.57 ± 6.29	58.67 ± 2.89
UbTeacher*	20%	PLD	55.78 ± 0.45	88.07 ± 1.88	63.02 ± 0.67
Ours	20%	PLD	**59.01** ± 1.57	**89.23** ± 0.98	**66.95** ± 2.45
Supervised baseline	50%	PLD	59.50 ± 1.40	92.37 ± 0.92	65.47 ± 2.04
STAC	50%	PLD	59.93 ± 1.21	92.35 ± 0.65	67.40 ± 2.10
SoftTeacher*	50%	PLD	60.47 ± 3.58	86.93 ± 4.23	69.63 ± 3.35
UbTeacher*	50%	PLD	61.66 ± 1.73	91.79 ± 1.45	**72.71** ± 1.42
Ours	50%	PLD	**63.39** ± 1.34	**92.96** ± 0.68	70.00 ± 3.45
Supervised baseline	100%	FLD	65.53	95.28	74.52
STAC	100%	FLD	46.98	80.76	50.61
SoftTeacher*	100%	FLD	**70.70**	94.90	**81.90**
UbTeacher*	100%	FLD	66.45	94.13	79.35
Ours	100%	FLD	67.64	**95.87**	76.81

Mean and standard deviation on test set portion evaluated over 3 data folds for 10%, 20% and 50% Labelled Ratio are reported. Supervised baseline refers to the same model trained on the labelled data only. Other state-of-the-art methods are re-evaluated on the dataset based on their publicly available codebase. We evaluate the methods with PLD and FLD settings which represent the Partially Labelled Data and Fully Labelled Data paradigms. PLD evaluation in particular was performed at scale using the Labelled Ratio portion of 500 labelled videos (i.e. ∼180k annotated frames) as labelled input and adding remaining videos plus ∼5000 additional unlabelled videos (i.e. ∼1.8M frames) of the same domain for unlabelled input. Note that ∗ indicates the code's data loader was changed for this dataset
The bold and underlined values represent the best and second-best results

Implementation Details-We use a Deformable DETR architecture with a ResNet-50 backbone as our default detection model (see Fig. 2) for evaluating the effectiveness of our method. The transformer decoder and encoder are randomly initialised and the ImageNet pre-trained ResNet-50 weights from SWAV (Caron et al., 2020) are used as initial parameters for our backbone. The student model is trained with the AdamW optimizer (Loshchilov and Hutter, 2018) with a weight decay of 0.0004 and a batch size of 64, distributed over 4 GPUs. We follow Caron et al. (2021) using a linear scale rule of $lr = 0.0005 \times batchsize/64$ and apply a slightly lower learning rate of $0.1 \times lr$ for the backbone.

We use randomly sampled frames for each video at each epoch with frequency $\omega = 10$, and the frames are rescaled so that the smaller axis of the frame is in range [320, 480]. The PLD model is trained for 1000 epochs with the first quarter as the warm-up phase and the last quarter as the cool-down phase (Fig. 5a), and lr decreases to $5e-5$ at the 800^{th} epoch. The momentum m for updating the teacher follows a cosine schedule from 0.998 to 0.9998. Since the amount of training data for the partially labelled data setting and the fully labelled data setting is quite different, training parameters vary slightly from that for FLD.[2]

Comparative Evaluation-We first evaluate our method for the PLD and FLD settings against a supervised baseline and state-of-the-art works STAC (Sohn et al., 2020b), SoftTeacher (Xu et al., 2021) and UbTeacher (Liu et al., 2021) at various ratios of labelled data. Table 1 summarises the results.

Our proposed method shows significant performance improvements under almost all test settings. For example, in the mAP column, we outperform the supervised baseline by 13.79%, 12.08%, 3.89%, STAC by 7.92%, 7.66%, 3.46%, SoftTeacher by 6.59%, 8.14%, 2.92% and UbTeacher by 1.93%, 3.23%, 1.73% when 10%, 20%, 50% of labelled data are provided, respectively. We find that our method works

[2] For FLD, we use total epochs=1100 with the first 500 as warmup, the last 100 as cooldown and lr decreases at the 1000th epoch. The unlabelled ratio is bounded at 10 in minibatch. Linear increase α is 0.3→1 and arctan increase is ς 0.3→0.6.

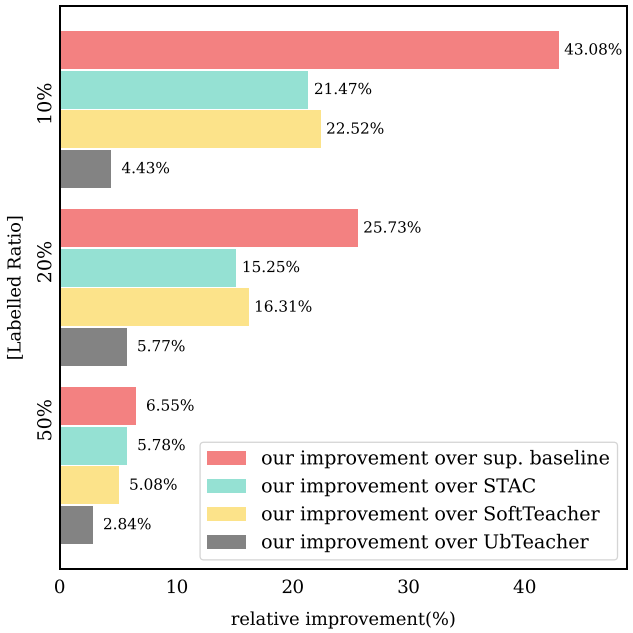


Fig. 8 Relative improvement comparisons. Relative improvement of mAP for our method over the supervised baseline, STAC, SoftTeacher, and UbTeacher across various PLD settings. We find our method shows particularly strong performance in lower annotation ratio regimes typical for many wildlife data settings

better than others particularly when the provided labelled data is small as illustrated in Fig. 8. We note again that such a setting is particularly common in wildlife applications where camera trap archives are large and accurate annotation ratios are very small. We can see that competitor methods also show sizeable improvements over the supervised baseline for smaller splits, indicating unsurprisingly that extra unlabelled data has particularly high value when very little labelling is available in the first place. However, note that the performance gap between the proposed method and other approaches is also particularly large in exactly this setting, confirming the specific applicability of our enhanced dynamics for curriculum learning in low labelling ratio settings. Qualitative results across all methods are exemplified and discussed in Fig. 9. This is complemented by visualisations of some failure cases in Fig. 10.

5 Ablation Studies

In this section, we evaluate our key contributions by examining the importance of each policy. We use the `fold1` split of the 50% PLD setting as the base for the conducted ablations.
Unlabelled Data Sample Policy-The motivation of our unlabelled data sample policy (Eq. 10) is to make sure that if the unlabelled samples loss L'_θ and unlabelled data sampling policy π are negatively correlated, the minimum possible loss can be achieved. Based on this hypothesis, we designed five experiments for five different possible characteristics for policy π: (i) linear increase of the number of unlabelled samples (Curriculum learning) used in training over the training process (depicted in Fig. 5a), (ii) linear increase but with warm-up and cool-down phases at the beginning and at the end respectively (Fig. 5b), (iii) start with all unlabelled samples but linearly decrease the number of unlabelled samples throughout training (Fig. 5c), (iv) The opposite of (ii) (Fig. 5d), (v) using all unlabelled samples constantly throughout training (Fig. 5e).

The results in Table 2(a) show that a gradual increase of the number of unlabelled samples during the self-training phase can gain around 1.16% mAP compared with constantly using all the unlabelled samples. The best performance however is achieved by introducing a warm-up and a cool-down phase at 64.3% mAP. This ablation experiment demonstrates that both the choice of 'phasing in' unlabelled data underpinned by our theoretical discussion in Sect. 3.3 have a positive measurable effect on learning performance.
Unsupervised Loss Weight Policy-The results in Table 2(b) demonstrate the effects of the unsupervised loss weight policy. We find that setting the unsupervised loss weight α is a challenge since both a large loss weight and a small loss weight can harm the performance. A 9.30% mAP drop occurs when $\alpha = 2.0$ compared to $\alpha = 0.5$. We argue that constantly applying a large α would harm the training at the beginning because α would assign a large weight for the loss produced by the unreliable pseudo-labels in the early stages which would mislead the model, causing it to get caught up in vicious training cycles. In contrast, applying our dynamic weighting approach, the performance reaches 64.73%, which is 0.75% better than a constant $\alpha = 0.5$, 2.55% better than $\alpha = 0.1$ and about 10% better than $\alpha = 2$. As discussed in Sect. 3.3, while our linear increase performs best, it is a naïve approach since the best global policy is difficult and computationally costly to find across the space of monotonously growing functions. Further exploration of this policy is subject to future work.
Confidence Threshold Policy- Table 2(c) displays the effects of different approaches for the confidence threshold policy. Both low and high thresholds cause significant performance degradation, at both low and high thresholds, e.g. $\varsigma = 0.05$ and $\varsigma = 0.9$, respectively, with lower thresholds being worse. This suggests that false positive pseudo-labels (appearing at low thresholds) have more negative impact than the false negative pseudo-labels (that appear at higher thresholds). As noted in Sect. 3.4, this motivated us to use a new weighted metric F_β to assess the best choice of threshold for this policy. Applying the arctan increasing ς approach, we see a significant increase in performance.

For comparison, we conducted a linear increase approach from 0.1 to 0.6, which takes similar strides, although arctan

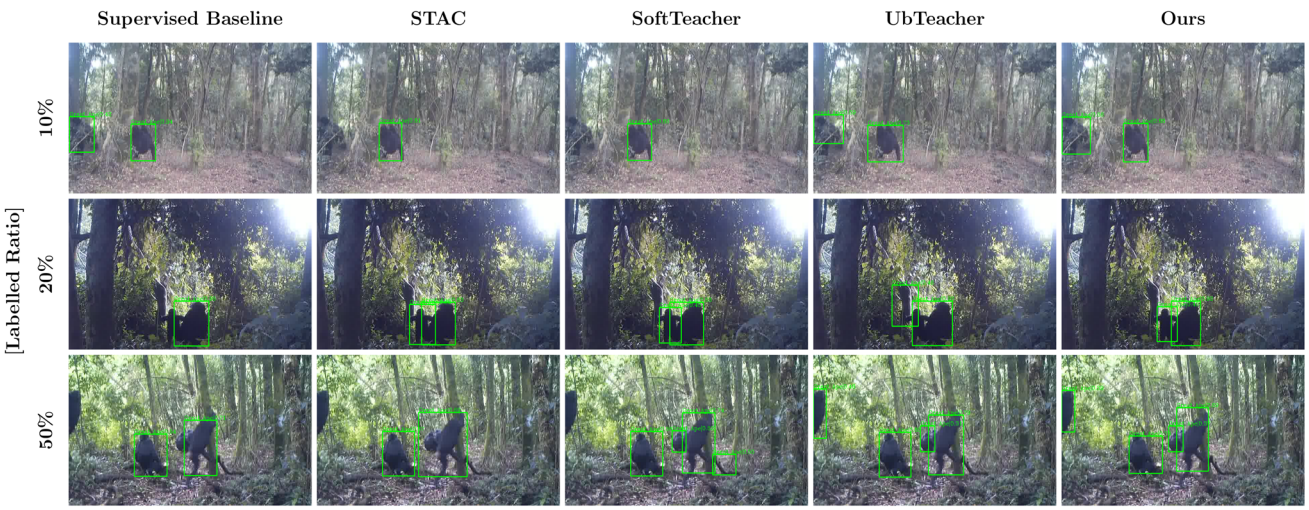


Fig. 9 Qualitative detection examples. We compare our method with other state-of-the-art approaches tested and the supervised baseline under the PLD setting with Labelled Ratios of 10%, 20%, 50%. Note examples where our proposed method reliably detects partly occluded apes and ignores tree structures which distract some of the other models (best viewed under zoom)

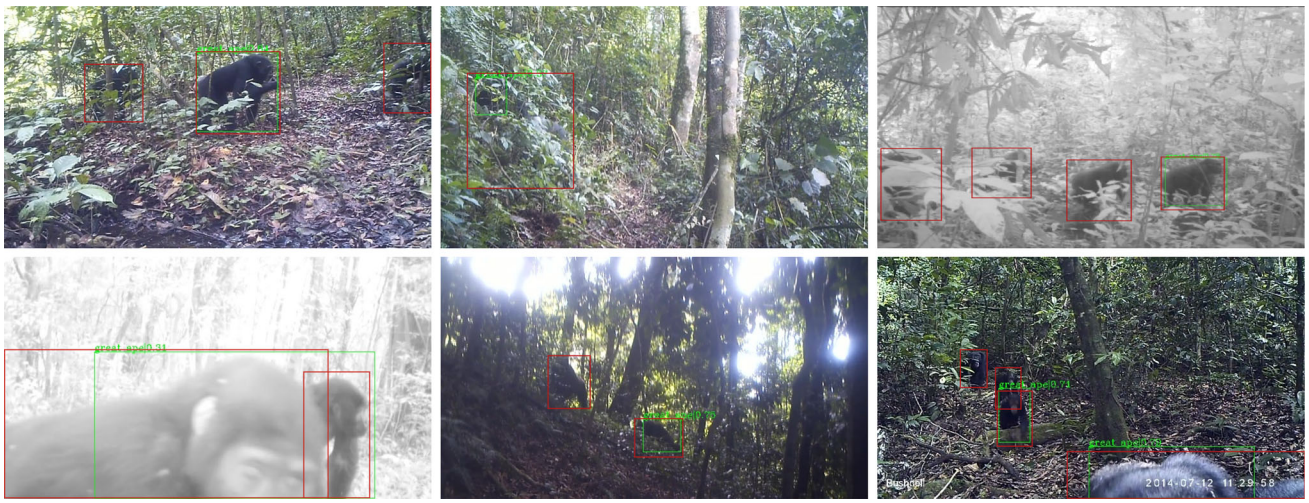


Fig. 10 Examples of failure cases. Visualised are failure cases under the 10% PLD setting. Ground-truth labels are annotated in red, and our detection results are shown in green. Note that partial occlusions form one hard-to-learn aspect given sparse label availability for training (best viewed under zoom) (Color figure online)

increases more aggressively in the early stages and achieves a slightly better outcome.

Augmentation Policy-A simple ablation is performed on the proposed augmentation policy, i.e. $\mathcal{A}_w$ and $\mathcal{A}_s$ augmentations versus no augmentation. Results show a significant improvement by 9.32% as seen in Table 2(d).

Teacher Momentum Policy-We compare a static momentum coefficient approach with a dynamic momentum approach for m which is used to update the teacher network. As shown in Table 2(e), both settings have a similar expected value but the dynamic momentum policy improves the performance significantly by 4.66%.

Initialisation-The initial status of the student model is crucial since it can affect training direction from the start towards an effective virtuous or catastrophic vicious cycle of learning. We see in Table 2(f) that a random initialisation of the model can lead to such catastrophic failure in training, while a SWAV-based self-supervised initialisation outperforms a supervised one.

6 Experiments on the MS-COCO Dataset

Our method primarily addresses the problem of handling sparsity of labelled data in animal biometrics whenever large

Table 2 Ablation studies of all learning policies

(a) Sample policy π	mAP	*(b)* Loss weight policy α	mAP	*(c)* Conf. threshold policy ς	mAP
Linear increase	61.62	Constant 0.1	62.18	Constant 0.05	51.73
Linear increase$^{\dagger *}$	**64.73**	Constant 0.5	63.98	Constant 0.6	62.62
Linear decrease	53.47	Constant 1.0	63.21	Constant 0.9	58.61
Linear decrease†	59.99	Constant 2.0	54.68	Linear 0.3 $\rightarrow$ 0.5	62.38
Constant	60.46	Linear 0.1 $\rightarrow$ 1*	**64.73**	Linear 0.1 $\rightarrow$ 0.6	64.10
		Linear 0.1 $\rightarrow$ 2	61.17	arctan 0.1 $\rightarrow$ 0.6*	**64.73**

(d) Augmentation policy $\mathcal{A}$	mAP	*(e)* Momentum Policy m	mAP	*(f)* Initialisation	mAP
No augmentation	55.41	Constant 0.999	60.07	Random init.	0.17
Augmentation with $\mathcal{A}_w$, $\mathcal{A}_s^*$	**64.73**	Constant 0.998	59.19	SWAV Init.*	**64.73**
		Constant 0.9998	57.60	supervised init.	61.55
		cos 0.998 $\rightarrow$ 0.9998*	**6**4.73		

The effectiveness of the introduced policies are verified via ablation. All studies are conducted for the `fold1` split at 50% Labelled Ratio in the PLD setting. The symbol * represents the default settings which we use in the system and the † symbol denotes the presence of warm-up and cool-down phases. Only one policy is varied for each study to isolate the effect. 'Constant' represents 'no curriculum' learning where the related variable or unlabelled sample pool stays constant

The bold and underlined values represent the best and second-best results

Table 3 Comparison with state-of-the-art methods on MS-COCO val2017 with PLD setting

Method	Venue	1% PLD	5% PLD	10% PLD
STAC (Sohn et al., 2020b)	Arxiv'20	13.97±0.35	24.38±0.12	28.64±0.21
Instant-teaching (Zhou et al., 2021)	CVPR'21	18.05±0.15	26.75±0.05	30.40±0.05
Humble-teacher (Tang et al., 2021)	CVPR'21	16.96±0.38	27.70±0.15	31.61±0.28
Unbiased-teacher (Liu et al., 2021)	ICLR'21	20.75±0.12	28.27±0.11	31.50±0.10
Soft-teacher (Xu et al., 2021)	ICCV'21	20.46±0.39	**30.74**±0.08	34.04±0.14
DETReg (Bar et al., 2022)	CVPR'22	14.58±0.30	24.80±0.20	29.12±0.20
MUM (Kim et al., 2022)	CVPR'22	**21.88**±0.12	28.52±0.09	31.87±0.30
Our supervised baseline	–	11.31±0.30	21.33±0.20	26.34±0.10
Ours	–	17.36±0.22	29.84±0.21	**35.08**±0.34

The $\text{mAP}_{50:95}$ standard COCO evaluation metrics on the COCO validation set are reported by models trained on 1, 5, 10% Labelled Ratio under PLD settings. The results are the average of 5 experiments with different random seeds. Our supervised baseline refers to our model without the unlabelled branch, leaving a Deformable DETR setup with ResNet-50 backbone identically initialised to our full method for fair comparison. Note that our full method demonstrates competitive or superior performance in comparison, indicating that concepts introduced here for wildlife detection are still applicable to general object detection

The bold and underlined values represent the best and second-best results

unlabelled data is available. Yet, it is nevertheless both conceptually and practically applicable to mainstream object detection. The concept of slowly expanding detection capabilities of a model in a policy-controlled way to learn highly complex and variable object appearance is indeed not limited to animal detection. In order to experimentally support any claim of wider applicability, we next evaluated our proposed method on the popular MS-COCO dataset under a low data regime (PLD) and with extra unlabelled data (FLD). For a fair comparison, we followed the evaluation approach used by STAC (Sohn et al., 2020b) using their splits between labelled and unlabelled data for PLD settings. We trained our model with Labelled Ratios of 1%, 5%, and 10% evaluated on the standard COCO `val2017` with the $\text{mAP}_{50:95}$ metrics. For the FLD option, we trained our model using the fully labelled COCO `train2017`, plus additional unlabelled COCO `unlabeled2017` following the same procedure described in (Sohn et al., 2020b; Yang et al., 2021; Liu et al., 2021; Tang et al., 2021; Zhou et al., 2021; Xu et al., 2021; Kim et al., 2022). As shown in Tables 3 and 4, we achieve leading state-of-the-art results for a 10% PLD Labelled Ratio and the FLD setting. At other ratios, our

Table 4 Comparison with state-of-the-art methods on MS-COCO val2017 with FLD Setting

Method	Venue	mAP
STAC (Sohn et al., 2020b)	Arxiv'20	39.21
ISMT (Yang et al., 2021)	CVPR'21	39.64
Unbiased-teacher (Liu et al., 2021)	ICLR'21	41.30
Humble-teacher (Tang et al., 2021)	CVPR'21	42.37
Instant-teaching (Zhou et al., 2021)	CVPR'21	40.20
Soft-teacher (Xu et al., 2021)	ICCV'21	44.50
MUM (Kim et al., 2022)	CVPR'22	42.11
Ours	–	**45.30**

The $mAP_{50:95}$ standard COCO evaluation metrics on the COCO validation set are reported by models trained on all the labelled `train2017` set plus additional unlabelled `unlabeled2017`. Our method dominates SOTA methods
The bold and underlined values represent the best and second-best results

benchmarks remain competitive: for 5% PLD our method trails only 0.90% below the best result by SoftTeacher, and for 1% PLD it scores 4.52% below the SOTA MUM model. This demonstrates that the introduced concepts of dynamic control in curriculum learning are certainly applicable to a wider domain of general object detection. We find that our curriculum learning method is less sensitive to hyper-parameters. In practice, the hyper-parameter configurations for COCO dataset[3] are inherited from the hyper-parameters that are fine-tuned on the PanAfrica dataset. They can indeed outperform the state-of-the-art under certain configuration after searching. Further research will be required to stipulate in how far truly dataset-optimal hyper-parameterisation of dynamic training regimes such as the one presented is computationally feasible. For practical purposes, it is important to note that hyper-parameter transfer does not lead to learning collapse or vastly degraded performance as will be shown again in our experiments outlined in the next section.

7 Experiments on the PASCAL VOC Dataset

In order to understand applicability to mainstream object detection further, we utilise another popular object detection benchmark to evaluate our model. We follow the standard FLD evaluation process on the PASCAL VOC dataset (Everingham et al. (2010)), as in (Sohn et al., 2020b; Liu et al., 2021; Zhou et al., 2021), with the performance of our model reported on `VOC07-test`, trained using `VOC07-trainval` as the labelled training set, and `VOC12-trainval` or `VOC12-trainval + COCO20cls`[4].

As shown in Table 5, we explore two different policy-parameter settings in the experiments, (i) without policy-parameter searching[5] (with † notation in the Table), and (ii) with pseudo-label analysis and policy-parameter searching. For VOC12, Row 13 shows the leading 57.02% mAP and the second best 81.89 % mAP_{50}, and for VOC12 + COCO20cls, Row 13 offers the second best 58.28% mAP and competitive 81.82% mAP_{50} among state-of-the-art methods. It also achieves a 14.89% and 16.15% gain in mAP over the supervised baselines, respectively, by simply adopting the configuration from the MS-COCO experiments. This further supports the argument that policy and parameter transfer does not lead to learning collapse or vast performance degradation.

When we systematically analyse the pseudo-labels and perform policy-hyper-parameter finetuning[6], the performance of our method can be boosted, achieving state-of-the-art 57.65% mAP for VOC12 and 82.34% mAP_{50} for VOC12 + COCO20cls (row 14 in Table 5).

Our experimental results on PASCAL VOC suggest i) the proposed method can have applications beyond animal detection, ii) it does not need heuristic tuning for hyper-parameters, since merely adopting the COCO ones to PASCAL VOC can lead to virtuous training cycles and achieve competitive results.

8 Experiments on the Bees Dataset

The Bees dataset[7] contains approximately 5K images of bees captured in hives. The bees and pollen that appear in each image are annotated with Bounding boxes and most of the data includes crowded scenes where bees are densely located.

In this experiment, we consider only PLD settings as we do not have extra bee data. We randomly spilt 80% of the whole data as a training set and use the rest as a testing set. For the training set, we construct three different PLD labelled ratios sampled with five different random seeds, where the labels are randomly masked so that the proportion of labelled data is 5%, 10%, and 20%, respectively. We evaluate the supervised baseline (using the same baseline approach as for MS-COCO) and our proposed model over five data folds for 5%, 10%, and 20% labelled ratios and

[3] To handle the large size of MS-COCO, the training epochs are adjusted to range from 50 to 100 depending on the labelled ratio so that the total training iteration is fixed to 180k, while keeping the other hyper-parameters of the policy the same.

[4] This set is tailored from MS-COCO dataset, which keeps the same 20 categories as PASCAL VOC as the unlabelled training set.

[5] We heuristically adopt the policy-parameter fine-tuned for MS-COCO.

[6] For reproducibility of this experiment, the exact parameters used were: π Linear Increase with warm-up and cool-down phases; linear increase α $0.1 \rightarrow 1$; arctan increase ς $0.2 \rightarrow 0.5$.

[7] Available at https://lila.science/datasets/boxes-on-bees-and-pollen.

Table 5 Comparison on PASCAL VOC dataset

No.	Method	Venue	VOC12		VOC12+COCO20cls	
			mAP_{50}	mAP	mAP_{50}	mAP
1.	Supervised	–	72.63	42.13	72.63	42.13
2.	STAC (Sohn et al., 2020b)	Arxiv'20	77.45 (+4.82)	44.64 (+2.51)	79.08 (+6.45)	46.01 (+3.88)
3.	ISMT (Yang et al., 2021)	CVPR'21	77.23 (+4.60)	46.23 (+4.10)	77.75 (+5.12)	49.59 (+7.46)
4.	Instant-teaching (Zhou et al., 2021)	CVPR'21	79.20 (+6.57)	50.00 (+7.87)	79.00 (+6.37)	50.80 (+8.67)
5.	Humble-teacher (Tang et al., 2021)	CVPR'21	80.94 (+8.31)	53.04 (+10.91)	81.29 (+8.66)	54.41 (+12.28)
6.	Unbiased-teacher (Liu et al., 2021)	CVPR'21	77.37 (+4.74)	48.69 (+6.56)	78.82 (+8.19)	50.34 (+8.21)
7.	Unbiased-teacher-v2 (Liu et al., 2022)	CVPR'22	81.29 (+8.66)	56.87 (+14.74)	82.04 (+9.41)	58.08 (+15.95)
8.	MUM (Kim et al., 2022)	CVPR'22	78.94 (+6.31)	50.22 (+8.09)	80.45 (+7.82)	52.31 (+10.18)
9.	Labelmatch (Chen et al., 2022a)	CVPR'22	**85.48** (+12.85)	55.11 (+12.98)	–	–
10.	DSL (Chen et al., 2022b)	CVPR'22	80.70 (+8.07)	<u>56.80</u> (+14.67)	<u>82.10</u> (+9.47)	**59.80** (+17.67)
11.	ACRST (Zhang et al., 2022)	AAAI'22	81.11 (+8.48)	54.30 (+12.17)	–	–
12.	Dense-teacher (Zhou et al., 2022)	ECCV'22	79.89 (+7.26)	55.87 (+13.74)	81.23 (+8.60)	57.52 (+15.39)
13.	Ours†		81.89 (+9.26)	57.02 (+14.89)	81.82 (+9.19)	58.28 (+16.15)
14.	Ours		<u>82.09</u> (+9.46)	**57.65** (+15.52)	**82.34** (+9.71)	<u>58.85</u> (+16.72)

In experiment, `VOC2007-trainval` is used as the labelled set and `VOC2012-trainval` used as the unlabelled set for all the models. The results are reported based on the evaluation on `VOC2007-test`. † represents the hyperparameters that are directly inherited from COCO without further finetuning efforts.
The bold and underlined values represent the best and second-best results

Table 6 Experimental results on bees dataset

Method	Labelled ratio	mAP	mAP_{50}	mAP_{75}
Supervised baseline	5%	26.15±1.47	65.24±2.55	14.96±1.83
Ours	5%	**32.81**±1.40 (+6.66)	**73.19**±2.93 (+7.95)	**22.12**±0.89 (+7.16)
Supervised baseline	10%	35.40±1.15	75.82±1.31	27.16±1.91
Ours	10%	**40.47**±0.25 (+5.07)	**79.15**±0.60 (+3.33)	**35.83**±1.05 (+8.67)
Supervised baseline	20%	42.24±1.33	82.34±1.65	37.99±2.12
Ours	20%	**45.17**±1.02 (+2.93)	**83.79**±0.89 (+1.45)	**44.09**±0.85 (+6.10)

Mean and standard deviation on test set portion evaluated over 5 data folds for 5%, 10%, 20% labelled ratio are reported. Supervised baseline refers to the same model trained on the labelled data only. Note that we adopt the policy hyper-parameters optimised for the PanAfrica dataset on these experiments.

Table 7 Experimental results on snapshot serengeti dataset

Labelled	fold	Supervised baseline			Ours		
		mAP	mAP_{50}	mAP_{75}	mAP	mAP_{50}	mAP_{75}
5%	0	52.13	74.94	56.45	55.40	78.48	59.20
	1	53.31	75.75	57.34	56.11	78.16	60.18
	2	52.05	74.98	56.12	55.38	78.17	59.61
	avg.	52.50	75.22	56.64	55.63	78.27	59.66
10%	0	56.24	78.00	60.83	59.56	80.92	63.86
	1	55.96	78.46	60.95	59.14	80.75	63.47
	2	55.89	78.87	60.81	58.95	80.61	63.60
	avg.	56.03	78.44	60.86	59.22	80.76	64.33
20%	0	58.98	81.19	64.26	60.97	82.53	66.04
	1	59.21	80.96	64.17	61.39	82.78	66.32
	2	59.74	81.51	64.81	62.12	82.90	67.26
	avg.	59.31	81.22	64.41	61.49	82.74	66.54

Three folds and their average results on test set evaluated are reported for 5%, 10%, 20% labelled ratio. Supervised baseline refers to our model without any aspect of the unlabelled branch. Note that we adopt the policy hyper-parameters optimised for the PanAfrica dataset on these experiments

report the mean and standard deviation of mAP. In Table 6, we demonstrate a substantial performance boost by applying our policy-guided semi-supervised learning, especially under lower data regimes, with 6.66% gain over the baseline in the 5% PLD setting.

9 Experiments on the Snapshot Serengeti Dataset

We finally conducted experiments on sparsely labelled versions of the Snapshot Serengeti dataset[8] (Swanson et al. (2015)) in which overall around 78K images (out of 7.1M) are labelled with instance-level bounding boxes that allow us to test our proposed method. We conducted our experiments under PLD settings where the model was trained with 5% and 10% of the labelled data out of 78K labelled images.

As shown in Table 7, substantial boosts of mAP, mAP_{50}, mAP_{75} can be observed under limited label regimes when comparing the supervised baseline (composed as before for the MS-COCO and Bees datasets) to our full system. Moreover, our method can reach similar or better performance than the supervised baseline while using only half of the labelled data.

To provide some further context to the wider literature on this dataset, we note that using only 20% of the labels our method's performance at mAP_{50} of 82.7 comes close to published results for fully supervised training with 100% of the Snapshot Serengeti labels using Mask-RCNN (Ibraheam et al., 2021) at mAP_{50} of 85.7 and significantly outperforms full label training with Faster-RCNN Ibraheam et al. (2021) at mAP_{50} of 73.2 or Context-RCNN (Beery et al., 2020) at mAP_{50} of 55.9.

Finally, we note that the multi-dataset learning approach of the MegaDetectorV5a* (Beery et al., 2019) leading to mAP_{50} of 90.65 on this dataset prevents fair apple-to-apple comparison with our method. However, the multi-dataset training regime is clearly highly effective in utilising label information across animal species boundaries. Future work into benchmarking our presented work in such multi-dataset training settings seems a promising avenue to improve results for species-specific and species-agnostic detectors further.

10 Conclusion

In this paper, we introduced an end-to-end dynamic curriculum learning framework for semi-supervised detection in sparely labelled datasets unlocking information in the unlabelled data portions. We demonstrated that bipolarity in the behaviour of cyclical student-teacher training regimes can lead to either effective virtuous or collapsing vicious training loops. We discussed the importance of expanding model coverage of new data slowly and in a controlled way to keep expanding detector and label quality without collapse. To achieve this, we proposed five policies to guide the dynamics of training and promote steady, simultaneous improvements to the student detector, the teacher detector, and the quality of the pseudo-labels. We showed that the described approach is effective in significantly advancing the state-of-the-art in great ape detection performance when evaluated under various settings on the large Extended PanAfrican Dataset. Our method is also shown to be beneficial to sparse labelling versions of other datasets without specialising hyper-parameterisations or policies. We have demonstrated this for the Bees and Snapshot Serengeti datasets in the animal domain. Finally, we showed that evaluation on general object detection tasks in MS-COCO and PASCAL-VOC achieves competitive or superior performance over existing state-of-the-art methods.

We conclude that the work holds the promise for dynamic curriculum learning controlled by training policies to be applied effectively to sparsely labelled wildlife data and thereby help unlock the full wealth of information so far widely sealed in steadily growing unlabelled camera trap archives.

Acknowledgements We would like to thank the entire team of the Pan African Programme: 'The Cultured Chimpanzee' Max-Planck-Institute (2022) and its collaborators for allowing the use of their data for this project. Please contact the copyright holder Pan African Programme at http://panafrican.eva.mpg.de to obtain the source videos from the dataset. Particularly, we thank: H Kuehl, C Boesch, M Arandjelovic, and P Dieguez. We would also like to thank: K Zuberbuehler, K Corogenes, E Normand, V Vergnes, A Meier, J Lapuente, D Dowd, S Jones, V Leinert, EWessling, H Eshuis, K Langergraber, S Angedakin, S Marrocoli, K Dierks, T C Hicks, J Hart, K Lee, and M Murai. Thanks also to the team at https://www.chimpandsee.org. The work that allowed for the collection of the dataset was funded by the Max Planck Society, Max Planck Society Innovation Fund, and Heinz L. Krekeler. In this respect we would also like to thank: Foundation Ministre de la Recherche Scientifique, and Ministre des Eaux et Forłts in Cote d'Ivoire; Institut Congolais pour la Conservation de la Nature and Ministre de la Recherch Scientifique in DR Congo; Forestry Development Authority in Liberia; Direction des Eaux, Forłts Chasses et de la Conservation des Sols, Senegal; and Uganda National Council for Science and Technology, Uganda Wildlife Authority, National Forestry Authority in Uganda.

[8] Available at https://lila.science/datasets/snapshot-serengeti.

References

Bain, M., Nagrani, A., Schofield, D., Berdugo, S., Bessa, J., Owen, J., Hockings, K. J., Matsuzawa, T., Hayashi, M., Biro, D., Carvalho, S., & Zisserman, A. (2021). Automated audiovisual behavior recognition in wild primates. *Science Advances*. https://doi.org/10.1126/sciadv.abi4883

Bar, A., Wang, X., Kantorov, V., Reed, CJ., Herzig, R., Chechik, G., Rohrbach, A., Darrell, T., Globerson, A. (2022). Detreg: Unsupervised pretraining with region priors for object detection. In: CVPR.

Beery, S., Morris, D., & Yang, S. (2019). Efficient pipeline for camera trap image review. CoRR abs/1907.06772, http://arxiv.org/abs/1907.06772, 1907.06772.

Beery, S., Wu, G., Rathod, V., Votel, R., & Huang, J. (2020). Context r-cnn: Long term temporal context for per-camera object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 13075–13085.

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning, pp 41–48.

Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A.,& Raffel, CA. (2019). Mixmatch: A holistic approach to semi-supervised learning. NIPS 32.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). *End-to-end object detection with transformers* (pp. 213–229). Springer, London: ECCV.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *NIPS, 33*, 9912–9924.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski. P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. In: ICCV, pp 9650–9660.

Carreira, J.,& Zisserman, A. (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR, pp 6299–6308.

Chen, B., Chen, W., Yang, S., Xuan, Y., Song, J., Xie, D., Pu, S., Song, M.,& Zhuang, Y. (2022a). Label matching semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14381–14390.

Chen, B., Li, P., Chen, X., Wang, B., Zhang, L., & Hua, XS. (2022b). Dense learning based semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4815–4824.

Chen, X., & He, K. (2021). Exploring simple siamese representation learning. In: CVPR, pp 15750–15758.

Deng, J., Dong, W., Socher, R., Li, LJ., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In: CVPR, IEEE, pp 248–255.

DeVrie,s T.,& Taylor, GW. (2017). Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552.

Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010). The pascal visual object classes (voc) challenge. *International Journal of Computer Vision, 88*(2), 303–338.

Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., et al. (2020). Bootstrap your own latent-a new approach to self-supervised learning. *NIPS, 33*, 21271–21284.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In: CVPR, pp 770–778.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In: ICCV, pp 2961–2969.

Huang, Y., Wang, Y., Tai, Y., Liu, X., Shen, P., Li, S., Li, J., &Huang, F. (2020). Curricularface: adaptive curriculum learning loss for deep face recognition. In: CVPR, pp 5901–5910.

Ibraheam, M., Li, K. F., Gebali, F., & Sielecki, L. E. (2021). A performance comparison and enhancement of animal species detection in images with various r-cnn models. *AI, 2*(4), 552–577.

Jeong, J., Lee, S., Kim, J., & Kwak, N. (2019). Consistency-based semi-supervised learning for object detection. NIPS 32.

Karras, T., Aila, T., Laine, S., & Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation. In: ICLR.

Kim, J., Jang, J., Seo, S., Jeong, J., Na, J.,& Kwak, N. (2022). Mum: Mix image tiles and unmix feature tiles for semi-supervised object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 14512–14521.

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. *ECCV* (pp. 740–755). London: Springer.

Lin, TY., Dollár, P., Girshick, R., He, K., Hariharan, B.,& Belongie, S. (2017a). Feature pyramid networks for object detection. In: CVPR, pp 2117–2125.

Lin, TY., Goyal, P., Girshick, R., He, K.,& Dollár, P. (2017b). Focal loss for dense object detection. In: ICCV, pp 2980–2988.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016). Ssd: Single shot multibox detector. *ECCV* (pp. 21–37). London: Springer.

Liu, YC., Ma, CY., He, Z., Kuo, CW., Chen, K., Zhang, P., Wu, B., Kira, Z.,& Vajda, P. (2021). Unbiased teacher for semi-supervised object detection. In: ICLR.

Liu, YC., Ma, CY., & Kira, Z. (2022). Unbiased teacher v2: Semi-supervised object detection for anchor-free and anchor-based detectors. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 9819–9828.

Loshchilov, I., & Hutter, F. (2018). Fixing weight decay regularization in adam.

Max-Planck-Institute (2022) Pan african programme: The cultured chimpanzee. http://panafrican.eva.mpg.de/index.php.

Norouzzadeh, M. S., Morris, D., Beery, S., Joshi, N., Jojic, N., & Clune, J. (2021). A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution, 12*(1), 150–161.

Rebuffi, SA., Ehrhardt, S., Han, K., Vedaldi, A., & Zisserman, A. (2020). Semi-supervised learning with scarce annotations. In: CVPRW.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In: CVPR, pp 779–788.

Ren, S., He, K., Girshick, R., &Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. NIPS 28.

Sakib, F.,& Burghardt, T. (2021). Visual recognition of great ape behaviours in the wild. In: IEEE/IAPR International Conference on Pattern Recognition (ICPR) Workshop on Visual Observation and Analysis of Vertebrate And Insect Behavior (VAIB).

Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., & Carvalho, S. (2019). Chimpanzee face recognition from videos in the wild using deep learning. *Science Advances, 5*(9), eaaw0736.

Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A., & Li, C. L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NIPS, 33*, 596–608.

Sohn, K., Zhang, Z., Li, CL., Zhang, H., Lee, CY., & Pfister, T. (2020b). A simple semi-supervised learning framework for object detection. arXiv preprint arXiv:2005.04757.

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., & Packer, C. (2015). Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific Data, 2*(1), 1–14.

Tabak, M. A., Norouzzadeh, M. S., Wolfson, D. W., Sweeney, S. J., VerCauteren, K. C., Snow, N. P., Halseth, J. M., Di Salvo, P. A.,

Lewis, J. S., White, M. D., et al. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution, 10*(4), 585–590.

Tang, Y., Chen, W., Luo, Y.,& Zhang, Y. (2021). Humble teachers teach better students for semi-supervised object detection. In: CVPR, pp 3132–3141.

Tian, Z., Shen, C., Chen, H., & He, T. (2019). Fcos: Fully convolutional one-stage object detection. In: ICCV, pp 9627–9636.

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., van Langevelde, F., Burghardt, T., et al. (2022). Perspectives in machine learning for wildlife conservation. *Nature Communications, 13*(1), 1–15.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, AN., Kaiser ,Ł., & Polosukhin, I. (2017). Attention is all you need. NIPS 30.

Wang, C., Jin, S., Guan, Y., Liu, W., Qian, C., Luo, P.,& Ouyang, W. (2022a). Pseudo-labeled auto-curriculum learning for semi-supervised keypoint localization. In: ICLR.

Wang, J., Wang, X.,& Liu, W. (2018). Weakly-and semi-supervised faster r-cnn with curriculum learning. In: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, pp 2416–2421.

Wang, X., Chen, Y., & Zhu, W. (2022). A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(9), 4555–4576. https://doi.org/10.1109/TPAMI.2021.3069908

Xu, M., Zhang, Z., Hu, H., Wang, J., Wang, L., Wei, F., Bai, X., & Liu, Z. (2021). End-to-end semi-supervised object detection with soft teacher. In: ICCV, pp 3060–3069.

Yang, Q., Wei, X., Wang, B., Hua, XS.,& Zhang, L. (2021). Interactive self-training with mean teachers for semi-supervised object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5941–5950.

Yang, X., Mirmehdi, M., & Burghardt, T. (2019). Great ape detection in challenging jungle camera trap footage via attention-based spatial and temporal feature blending. In: ICCVW.

Zhai, X., Oliver, A., Kolesnikov, A., & Beyer, L. (2019). S4l: Self-supervised semi-supervised learning. In: ICCV, pp 1476–1485.

Zhang, B., Wang, Y., Hou, W., Wu, H., Wang, J., Okumura, M., &Shinozaki, T. (2021). Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. NIPS 34.

Zhang, F., Pan, T., & Wang, B. (2022). Semi-supervised object detection with adaptive class-rebalancing self-training. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*, 3252–3261.

Zhong, Z., Zheng, L., Kang, G., Li, S., & Yang, Y. (2020). Random erasing data augmentation. *AAAI, 34*, 13001–13008.

Zhou, H., Ge, Z., Liu, S., Mao, W., Li, Z., Yu, H., & Sun, J. (2022). Dense teacher: Dense pseudo-labels for semi-supervised object detection. In: ECCV.

Zhou, Q., Yu, C., Wang, Z., Qian, Q.,& Li. H. (2021). Instant-teaching: An end-to-end semi-supervised object detection framework. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4081–4090.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., Dai, J. (2020). Deformable detr: Deformable transformers for end-to-end object detection. In: ICLR.