ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA241066

Frisbees and Dogs: Domain Adaptation for Object Detection with Limited Labels in Rugby Data

Will Connors [a,*], **Ellen Rushe**[b] **and Anthony Ventresque**[a]

[a]SFI Lero @ School of Computer Science and Statistics, Trinity College Dublin, Ireland
[b]School of Computing, Dublin City University, Ireland
ORCID (Ellen Rushe): https://orcid.org/0000-0001-5869-5333, ORCID (Anthony Ventresque): https://orcid.org/0000-0003-2064-1238

Abstract. Object detection often struggles when applied to low-resource, domain-specific datasets. This challenge is exacerbated when dealing with sports-related data such as rugby, where fast-paced gameplay and tackles result in frequent instances of motion blur and occlusion, representing a substantial domain-shift from widely available pre-trained models. Given the high cost of manual labelling, we seek to determine whether we can minimise the number examples needed for fine-tuning by identifying implausible label classifications made by pre-trained object detection models. We do this using a coarse-grained labelling approach in the absence of detailed ground truth bounding boxes, allowing us to determine whether a label is implausible within the context of a rugby pitch. This is done to maximize the information provided by each example used for fine-tuning with the goal of minimizing the number of examples needed. Our results show that using pool-based, single-step uncertainty sampling to select examples from a subset of frames with implausible labels improves the model performance. More specifically, we show that fine-tuning on frames with the lowest confidence scores first can lead to greater performance after roughly 30 examples.

1 Introduction

Modern-day visual perception algorithms rely heavily on training datasets consisting of thousands of images such as MS COCO [12] and PASCAL VOC [7]. MS COCO, for example, has 91 common classes covering a broad range of objects with 82 of them having over 5,000 labelled instances. Although models trained on these datasets perform well when classifying in-distribution data, they still require fine-tuning when we seek to generalise to domain-specific instances [14]. Tracking tackles and player movements is a key concern for rugby players and coaches with there being a distinct lack of algorithms to detect tackles in rugby union [4], therefore developing a robust, efficient object detection algorithm will be beneficial to future work in tracking rugby tackles. Object detection in sporting scenarios is somewhat unusual in that it has a limited number of objects that are relevant to actual gameplay in addition to those that are essentially irrelevant to sporting analysis (e.g. spectators). There are also unconventional object classes that must be considered such as the horizontal and vertical line markings on pitches which can affect the type of gameplay that is likely to occur.

Though out-of-the-box object detectors will likely detect some instances of “seen" classes, the domain specific nature of rugby data leads to some challenges. Some of these are common to general object detection, such as occlusion. Others are due to shifting domains where particular variations of certain classes are uncommon in more “generalised" settings. For example, it is likely that the positions of players’ bodies will differ from what might be seen in day-to-day scenarios. Objects might also be moving at unusual speeds such as a ball travelling at a high speed in a professional rugby environment. A high degree of motion blur is therefore to be expected. Additionally, in a given frame, there is often a significant level of class imbalance. For example, there will be several players in a frame at any given time but only a single ball object. Labelling a single frame will therefore not yield the same amount of information for each class. It is also worth considering the sources of many larger datasets. The most likely characteristics of sporting equipment will vary from location to location depending on the most popular local sports (for instance the “baseball bat" label in MS-COCO), leading to a certain degree of bias based on geographical location (i.e. where the data was collected or labelled). The geographical location of data collection can also bias data in more indirect ways such as the environmental characteristics and common weather conditions. In this paper, we aim to evaluate off-the-shelf object detection specifically for the rugby domain in the absence of large amounts of labelled data and attempt to use the information from these models in order to maximally utilise available training examples for each class individually.

The remainder of the paper is structured as follows. First, we discuss works in the area of tackle detection in rugby and the need for effective object detection. Next we detail our method of evaluating pre-trained models using “cheap" coarse-grained labels. We then go on to detail our experimental procedure for fine-tuning using confidence ordering. Following that, we discuss the results obtained using different variations of confidence ordering and finish with some concluding remarks.

2 Related Work

In order to motivate the work on domain adaptation, it is useful to first look at the contexts where these models are used. For example,

* Corresponding Author. Email: connorsw@tcd.ie.
This research was supported by Science Foundation Ireland grant 13/RC/2094 P2 to Lero - the Science Foundation Ireland Research Centre for Software (www.lero.ie).

there have been several works that attempt to analyse the quality of rugby gameplay, in particular with the aim of providing feedback that could minimize the risk of injury to players.

For example, Daly et al. [6] attempt to detect unsafe tackles for a proof-of-concept mobile training platform measuring orientation using inertial measurement units captured with a sensor worn on the inside of a player's jersey. The use of sensors for measurement, however, is invasive, depending on the placement.

Non-invasive alternatives have therefore been proposed in a number of works. Martin et al. [15] used YOLO (version 4) [1] to detect both the ball and players in videos, followed by OpenPose pose estimations to determine the kinematic measurements of players and a Kalman filter to track the movement of the ball. High risk tackles were determined by tracking the head centre of the tackler and the ball-carrier. The authors trained YOLO with a few hundred images however the minimum number of ground truth labelled datapoints required by this type of fine-tuning was not discussed. Similarly, Nonaka et al. [16] created an automated system to detect high-risk tackles directly from video footage. The authors compared a number of variations of ResNet models for tackle frame selection. Following frame selection they then compare a set of object detection models (YOLO version 3 [17], DETR [3] and RetinaNet [11]) by fine-tuning them to detect the bounding boxes of tackles. Pose estimation was then applied to the players involved in the tackle using CenterTrack [21]. Finally, Naive Bayes was used to detect whether or not a tackle was high risk. Overall they found that a ResNet with (2+1)D convolutions [19] with RetinaNet and CenterTrack performed best.

One noteworthy observation from the works above is that object detection plays an important role in identifying players and their positions. Correctly identifying players also affects downstream pose estimation which is also frequently performed in these works. Fine-tuning is common in the literature due to the domain-shift between the more "general" data used for pre-training and the target sports-related scenes. It is often also necessary to account for the presence of novel objects in the target domain. The minimum number of labelled examples required to perform fine-tuning is not explored in the works described, though this is an important consideration as labelling game footage is a laborious task, requiring both time and expertise. Minimizing the number of examples needed to perform domain adaptation, as in many other applications, is desirable.

3 Methodology

In this section, we will describe our methodology for evaluating the "off-the-shelf" performance of a widely used object detection model, YOLOv5, when detailed ground truth labels are not yet available and without taking steps to shift the domain of the model; Next we describe how analysis of these initial predictions can be used in order to identify the most salient points on which to fine-tune.

3.1 Using Coarse-Grained Labels to Measure Performance

A key difficulty of evaluating pre-trained object detection models on real-world data is that, without ground truth bounding boxes and labels, measuring performance is challenging. As a result, it is difficult to quantify the degree of fine-tuning that is necessary to perform and therefore the number of examples to label in order to effectively transfer to the target domain. We approach this problem by using a coarse-grained labelling approach. To this end, we introduce the concept of *plausibility*. Here, given predictions made with a pre-trained model using its original class set, we simply ask a human to choose the labels that are most plausible *given the context* out of all unique labels.

This is a "cheap" evaluation measure as it does not require bounding boxes to be drawn or for individual frames to be labelled, but gives us a rough estimate of the performance of a given pre-trained object detector using a small amount of contextual information. Intuitively, if a large number of implausible labels are predicted, it is likely that the model is particularly ill-adapted to the target domain. We also hypothesise that the frames classified as implausible objects likely contain novel objects or previously unseen variations of known objects.

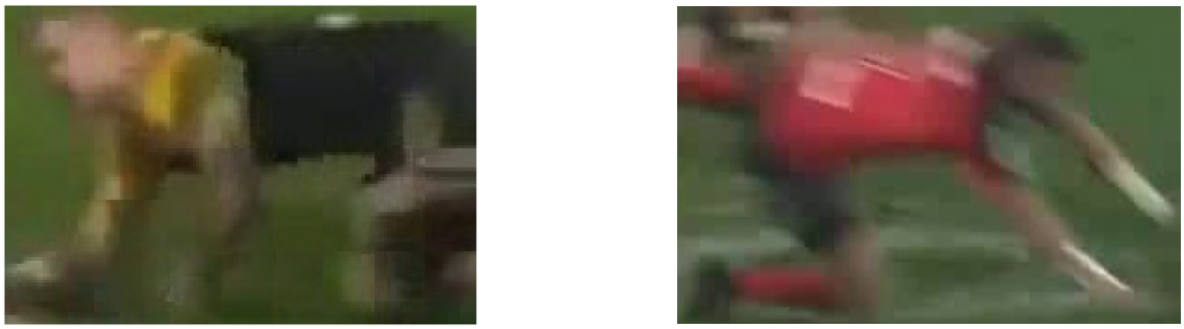
Pre-trained models also typically contain some measure of confidence. In this work we additionally sought to establish the degree to which the confidence of labels was associated with their plausibility in the target domain. Finding an association between the proportion of implausible labels and lower confidence scores could give us an indication of those examples in the unlabelled dataset that present the greatest challenge to the model and this, in turn, can be leveraged to more carefully select examples to label for the purposes of domain adaptation.

3.2 Fine-tuning with Single-Step Pool-based Uncertainty Sampling

So far, we have described a potential means to understand whether the initial predictions of an "off-the-shelf" model can provide us with information on the difference between the source and target domain. We hypothesise that, provided that we can show that there is sufficient separation between the confidence scores of plausible and implausible labels, these scores can be leveraged in an active learning framework to minimize the number of labelled examples needed to fine-tune a pre-trained object detector. Intuitively, plausible labels with high confidence are likely to be correct and therefore do not provide additional information on novel or domain-specific variations of a given class, making examples with lesser scores more informative.

This strategy is a form of pool-based uncertainty sampling [9, 10] using *least confidence* [18], where labels are queried from an oracle ordered by a measure of confidence or uncertainty. In our case, however, given that there are several different objects within each frame, this measure of confidence must be aggregated to form a single score. We note that the confidence scores for each image may not be normally distributed, therefore we evaluate three variations of aggregation to assess which provides the most benefit in fine-tuning. We score each frame based on the minimum, mean and median confidence of all objects in a single frame. Furthermore, we also reduce the pool of candidate examples by specifically selecting frames with implausible labels.

To enhance the applicability of this strategy, we also consider the labelling load on the oracle by applying active learning using a single batch. Active learning methods often employ an iterative strategy, where labels are obtained from the oracle sequentially. This requires the oracle to perform labelling during the training process, an often unrealistic expectation. It is far more realistic for a small number of carefully selected examples to be given to a domain expert at once to label which can be used for fine-tuning in a single step (i.e. single shot) [5, 20].



(a) (b)

Figure 1: Examples of players labelled as “dog" rather than “person".

4 Experimental Setup

We use an injury risk assessment dataset which was curated by a registered rugby analysis database [15]. It consists of 109 tackle segments which have been sourced from both the Rugby World Cup in 2019 and Super Rugby [15]. A tackle segment was defined using a tackle classification framework [8] and each tackle segment was manually clipped by a rugby video analysis expert [15].

The resolution of video frames varies from pixel dimensions 854×480 to 1920×1080, the number of frames per second varies from 20 to 30 [15]. The data is extracted directly from the broadcast view and varies between multiple camera angles, some being further away from the players than others. When the boundary box for a given player was less than 14% of the pixel height for the overall frame or there was an excess of occlusion, the authors determined that the players would lack the detail necessary for detection therefore these segments were not included [15]. The number of frames within each tackle segment varied from a minimum of nine frames to a maximum of 153 frames. The final dataset contains 3,560 total frames drawn from the 109 distinct videos. This entire dataset was used for our initial analysis of the efficacy of our pre-trained model on a new domain.

The following details the process of sampling training and testing datasets from these 3,560 frames. Note that the training and test sets were split using video IDs, rather than splitting on randomly sampled frames to ensure a more realistic evaluation of the model's generalisation performance.

After running an out-of-the-box YOLO model on the data, a single expert human rater simply marked each of the object classes detected as either “plausible” or “implausible”. The plausibility of the detected labels was then used in creating the following training and test sets:

- Fifty frames were used in the training data which were randomly selected from a sample of 132 frames containing an implausible object that was manually inspected and relabelled. These selected examples were then given fine-grained labels, i.e. each object in the frame was given a correct bounding box and label. As discussed in the Methodology Section, the idea here is that frames containing implausible labels are likely to be more informative to training. This relabelled training data was then broken up into five segments, each with an incrementally larger sample size. The purpose of these sets was to evaluate the number of labelled examples required to gain improved performance given single step active learning approach. The first segment contained 20% of the labelled training set, the second contained 40% and so on, increasing by 20%, until a segment with all corrected training data was reached. We emphasise once again that unlike standard uncertainty sampling [10], we do not sequentially relabel data during the training process as this requires an expert to label after each training update, which is a less realistic scenario as this is a more time-consuming task. We therefore present the expert with all examples to be labelled at one.

- The test set consisted of 100 different frames. 50 frames were selected at random from a sample of 546 that contained an implausible object. To compose the second half of the test set, another 50 frames were selected that contained a high-confidence object (object confidence greater than 85%). The intuition here is that we should evaluate the performance of examples that previously contained low confidence objects along with that of those that contained objects detected with high confidence. This enables us to evaluate whether, in the process of fine-tuning, we have degraded the detection performance for examples that were likely to previously have been detected well.

For the purpose of adapting the object detection model to the rugby domain, the training and testing datasets required manual labelling by a domain expert. The labels were reduced to the following classes: *person*, *line* and *sports ball*. Coco-Annotator [2] was used to manually label each object or person's bounding box. The *line* label suffered a significant level of occlusion, therefore each non-occluded line segment was given its own bounding box. Additionally, the set of frames were also checked to ensure there were no duplicates.

We use the PyTorch implementation of YOLO v5 [1] with a batch size of 1, 100 training epochs and YOLO v5's default training parameters. A model was fine-tuned for each of the five training data segments, with the first training batch containing 10 frames, increasing by a further 10 frames per batch until the final training set, which contained all 50 frames.

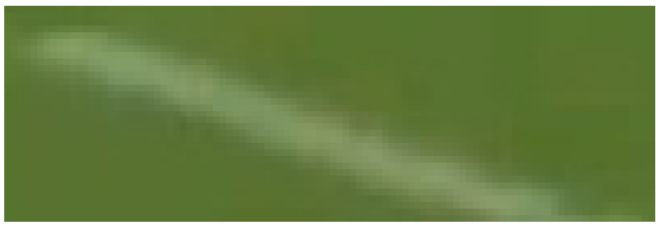
5 Results

In this section we will first discuss the domain shift between the pre-trained YOLO model used and the target domain without fine-tuning using the analysis strategy outlines in Section 3.1. We then go on

[1] https://github.com/ultralytics/yolov5

to discuss the performance improvement gained by providing incrementally larger training segments obtained using the sampling technique outlined in Section 3.2.

(a) A rugby ball



(b) A pitch marker

Figure 2: Example of objects labelled as "frisbee".

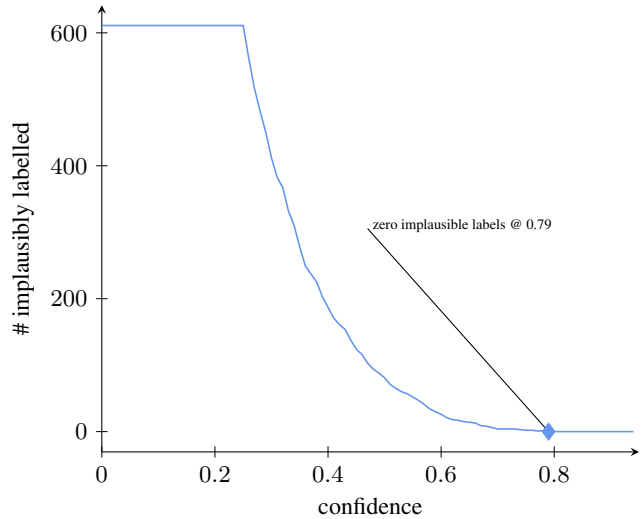


Figure 3: Number of implausible labels at different confidence thresholds.

5.1 Establishing Difference in Domain

We first evaluate the initial predictions on the entire dataset using plausibility in the absence of detailed ground-truth labels. Looking at the initial classifications, a clear difference in domain is visible. Figures 1a and 1b show a person who has been detected as the *dog* class, despite the presence of a *person* class in the dataset used for pre-training. This is likely due to the horizontal, four-limb stance being far more prevalent in the *dog* class training images than in those of the *person* class. While a vertical, two-limb stance is far more likely for humans in everyday scenarios, the high-contact nature of a sport such as rugby significantly increases the likelihood of falls which are naturally underrepresented in non-domain-specific datasets.

Figure 2a shows another image of a rugby ball classified by the "off-the-shelf" model as *frisbee* despite being an obviously out-of-place object to the human eye given the context of a rugby game. In the absence of this type of "common sense", the object detector has no way of determining this information without additional input. We can also see in Figure 2b that novel concepts, such as the pitch markings are classified incorrectly.

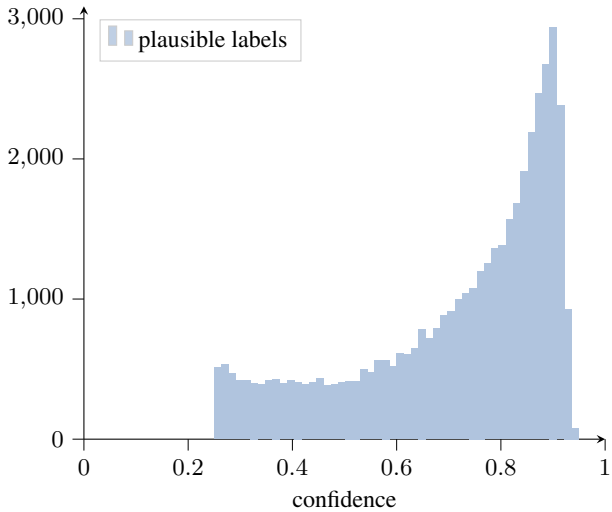
To quantify the number of implausible labels, coarse-grained labelling is first performed, with the only plausible labels from the source pre-training object detection task determined to be *person* and *sports ball*. Figure 3 plots the number of implausible labels against increasing confidence scores. We can see that there are no objects with confidence greater than 0.79 with an implausible label.

In fact, Figure 4a, shows that the distribution of the confidence scores of objects with plausible labels is heavily skewed towards higher values. The scores of objects with implausible labels shown in Figure 4b, on the other hand, are skewed in the opposite direction. This indicates that the model is reasonably well calibrated. The number of implausible labels therefore becomes a surprisingly good proxy of the performance on a target dataset, without the need for extensive detailed labelling. Given that the confidence scores also appear to be associated with unknown concepts, learning from frames with low confidence classifications should provide more information to the model than higher confidence frames. This confirms that implausible labels are unlikely to exhibit high confidence scores, as hypothesised in Section 3.1.

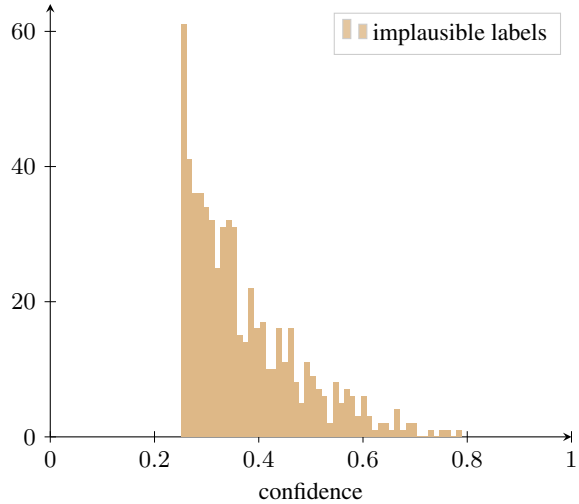
5.2 Fine-tuning with Confidence Ordering

As discussed in Section 3.2, given the association between low confidence and implausible labels, we hypothesise that frames with lower confidence classification are the most informative datapoints for learning. We tested this hypothesis by fine-tuning the YOLO model on frames ordered by confidence. As also described in Section 3.2, given that there are multiple instances of each class in each frame, we must aggregate the confidence score for each frame. We chose three methods: minimum, mean and median frame confidence. The mean Average Precision (mAP) over Intersection Over Union (IOU) thresholds 0.5 to 0.95 (mAP50-95) for all classes is shown in Figure 5a ordered using these three aggregation methods along with a model trained in the "standard" way, with examples for each subset of data randomly sampled from the total pool of 50 selected implausible samples (i.e. with random ordering).

Though standard random sampling provides the highest mAP score when fine-tuning using between 10 and 20 frames, as the number of frames increases, there is a significant improvement when ordering samples by both the mean and median confidence method's mAP score. After 30 frames, all three ordering methods begin to outperform standard random sampling with both mean and minimum ordering proving to be marginally better than median ordering. Though it is tempting to conclude that mean ordering is the optimal strategy, this conclusion is slightly misguided as the majority of the examples within the overall sample were within the *person* class, with *sports ball* and *line* accounting for just a small proportion of the sample at just 5% and 23% respectively. We therefore analyse each class individually.



(a) Confidence scores for *plausible* objects.



(b) Confidence scores for *implausible* objects.

Figure 4: Histograms comparing distribution of of YOLOv5 confidence scores for *plausible* (4a) and *implausible* (4b) objects (Figure 4b differs slightly from Figure 3 as the latter is the actual count of implausible labels at *particular* confidence thresholds as opposed to the count over a small ranges of thresholds in a histogram).

In the case of class *line*, after 20 frames, median confidence ordering outperforms standard sampling, with frames ordered by mean and minimum confidence below both. After training on 40 frames, a substantial increase in performance is achieved using minimum ordering. Conversely, mean ordering performs comparably to standard sampling while median ordering results in the poorest performance.

We can see in Figure 5c that the standard random sampling achieved the highest detection score after 10 frames with an mAP score just above 0.1. However, mean and median confidence ordering have the most substantial increase in mAP, both outperforming standard sampling after 30 and 40 frames. The performance of the minimum ordering strategy does not improve as significantly after 30 and 40 frames, performing just below median and mean, though still above standard sampling.

Our model struggled to improve performance on the *sports ball* object class. This may be due to the limited sample size across all ordering techniques, with the model only beginning to detect any objects after 30 frames. This is shown in Figure 5d. Minimum confidence ordering proved the worst approach for fine tuning, returning an mAP score of 0 after 40 frames. Both median and mean ordering outperform standard sampling, substantially increasing the mAP score after 40 frames. A final mAP score of 0.06 is the lowest recorded score of all three labels. This is likely due to the small sample size, the level of occlusion for this class, and the higher likelihood of motion blur.

Though no single confidence ordering technique consistently proved optimal for fine-tuning, in the case of all three classes (*line*, *sports ball* and *person*), one or more of the three confidence ordering methods evaluated out-performed standard random sampling after 30 frames. It is noteworthy, however, that standard random sampling appears optimal when fine-tuning on under 30 frames. This is, perhaps, unsurprising given the likely under-representation of minority classes in each subset of data.

5.3 The Endemic Challenge of Shifting Domains

As noted in Section 2, object detection often proceeds pose estimation in tackle analysis. Given the improved detection of players shown here, we completed the additional step of applying a common pose estimation model on the crops of the detected players using the common pose estimation model, Mediapipe [13]. However, although more detections were made by the fine-tuned model proposed in this paper, there were a higher proportion of instances of failed pose estimation for the players detected by the fine-tuned model relative to the set of players detected with the “off-the-shelf” model – despite the fine-tuned model identifying the players more effectively. This suggests that the domain shift presented by these, likely more domain-specific, examples also effects downstream pose estimation models. To investigate the potential causes of this, we took a random sample of 30 instances where pose estimation failed for both models and analysed their characteristics.

From the fine-tuned model sample, from 30 detections we found 28 instances where a person was recognisable to a human annotator, 13 instances where the whole body was captured within the crop, and another 14 instances where the back of the person was turned away. The out-of-the-box model had 21 instances where the person was recognisable to a human, 8 instances of a full body captured within the crop and 11 instances where the person’s back was turned. It should be noted that, in large part, the reason for the number of “human-recognisable" instances being lower in the “off-the-shelf” model was due to some detections containing more than just one person within the crop.

the “off-the-shelf” model had multiple instances where there were 2 or more objects within the same frame, therefore, were determined to be not human-recognisable

We noted that in most cases, the players pose was visible, though, once again, the positions may not be those common to the training data used for pose estimation in a similar way the training data used to train common object detection models. Future work will

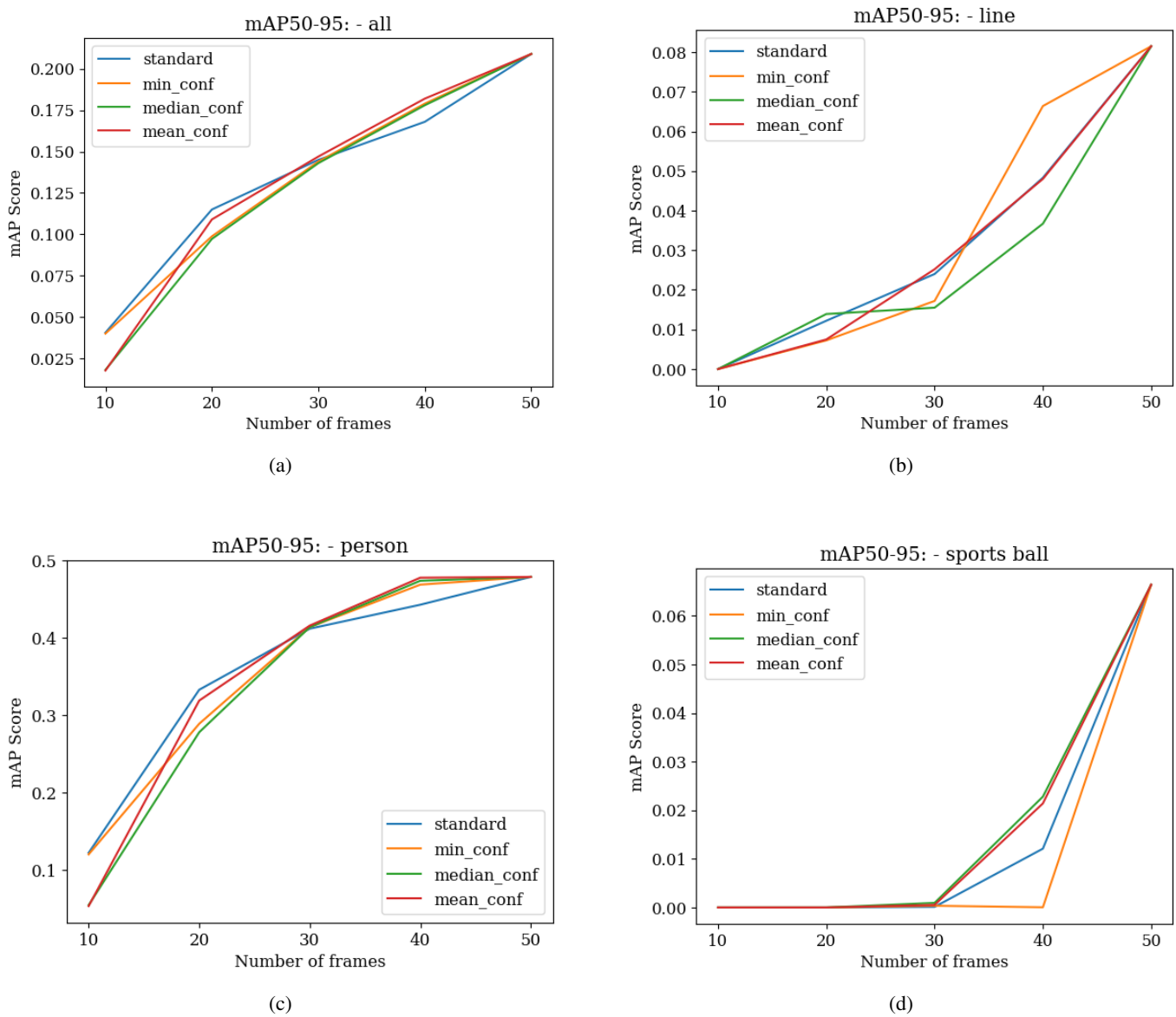


Figure 5: mAP over IOU thresholds of 0.5 to 0.95 for all classes (5a), the *line* class (5b), the *person* class (5c), and the *sports ball* class 5d.

look to evaluate pose estimation models to determine whether these pipelines can be made more adaptable to low-resource, domain-specific environments.

Table 1: Human analysis of 30 object crops from both models

Scenario	Custom Model	OTB Model
Human Recognizable	28	21
Fully Visible Body	13	8
Back Turned	14	11

6 Conclusion

In this paper, we have presented a means of evaluating the "off-the-shelf" performance of object detection algorithms on domain shifted data through the use of coarse-grained plausibility labels, using the example domain of rugby. We have shown that the predictions of implausible labels appear to skew towards lower confidence values. We show that by fine-tuning using pool-based, single-step uncertainty sampling on frames with objects or people classified with implausible labels, we can obtain significant performance gains in just a few examples. Additionally, we see that training on frames with low confidence objects first appears to lead to greater performance after only around 30 examples for all object and person classes that we attempted to classify. Furthermore, we perform additional preliminary analysis that suggests that even supposedly "generalised" pose esti-

mation models can perform poorly even when the people are clearly visible to a human annotator within a crop of an image. This suggests that domain adaptation may too be necessary for even these more generalised models.

Acknowledgements

We warmly thank Thomas Laurent and Ruth Holmes for their feedback on earlier drafts of this paper.

References

[1] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

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

[3] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[4] R. M. Chambers, T. J. Gabbett, R. Gupta, C. Josman, R. Bown, P. Stridgeon, and M. H. Cole. Automatic detection of one-on-one tackles and ruck events using microtechnology in rugby union. *Journal of science and medicine in sport*, 22(7):827–832, 2019.

[5] G. Contardo, L. Denoyer, and T. Artières. A meta-learning approach to one-step active learning. *arXiv preprint arXiv:1706.08334*, 2017.

[6] E. Daly, P. Esser, A. Griffin, D. Costello, J. Servis, D. Gallagher, and L. Ryan. Development of a novel coaching platform to improve tackle technique in youth rugby players: A proof of concept. *Sensors*, 22(9): 3315, 2022.

[7] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010.

[8] S. D. Hollander, C. Ponce, M. Lambert, B. Jones, and S. Hendricks. Tackle and ruck technical proficiency in rugby union and rugby league: A systematic scoping review. *International Journal of Sports Science & Coaching*, 16(2):421–434, 2021. doi: 10.1177/1747954120976943. URL https://doi.org/10.1177/1747954120976943.

[9] D. Lewis and W. Gale. A sequential algorithmfor training text classifiers. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 3–12, 1994.

[10] D. D. Lewis. A sequential algorithm for training text classifiers: Corrigendum and additional data. In *Acm Sigir Forum*, volume 29, pages 13–19. ACM New York, NY, USA, 1995.

[11] T. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. arxiv. *arXiv preprint arXiv:1708.02002*, 2017.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[13] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. Yong, J. Lee, W.-T. Chang, W. Hua, M. Georg, and M. Grundmann. Mediapipe: A framework for perceiving and processing reality. In *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR) 2019*, 2019. URL https://mixedreality.cs.cornell.edu/s/NewTitle_May1_MediaPipe_CVPR_CV4ARVR_Workshop_2019.pdf.

[14] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[15] Z. Martin, S. Hendricks, and A. Patel. Automated tackle injury risk assessment in contact-based sports-a rugby union example. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4594–4603, 2021.

[16] N. Nonaka, R. Fujihira, M. Nishio, H. Murakami, T. Tajima, M. Yamada, A. Maeda, and J. Seita. End-to-end high-risk tackle detection system for rugby. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3550–3559, 2022.

[17] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[18] B. Settles. Active learning literature survey. *Computer Sciences Department Technical Report #1648*, 2009.

[19] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[20] Y. Yang and M. Loog. Single shot active learning using pseudo annotators. *Pattern Recognition*, 89:22–31, 2019.

[21] X. Zhou, V. Koltun, and P. Krähenbühl. Tracking objects as points. *ECCV*, 2020.