Title: Contact Detection in American Football: A Comprehensive Approach

G076 (s2435255, s2340198, s2442472)

Abstract

This research focused on detecting player contact in American football using SVM as a baseline model, and compared it with deep learning models ResNet-50, ResNet-152, and EfficientNet-B0. Although the models had better results than SVM, overfitting occurred. To address this, image augmentation was applied with ResNet-50, which gave the best accuracy. Additionally, the study examined the performance of state-of-the-art object detection models (Faster RCNN, Mask2Former, YOLOX) on cropped versus larger images. YOLOX consistently outperformed the others, demonstrating its potential to be used for development of real-time contact detection system in American football.

1. Introduction

American Football is a highly physical sport, characterized by frequent collisions and tackles between players. This makes the players prone to various injuries, some of which can be severe and have long-lasting consequences. Therefore, being able to detect contacts in real- time can be really helpful in determining which player is prone to injury and potentially changing the strategies involved in the game. The main challenge with this task is that the players are constantly engaging in physical contact, making it exceedingly challenging to accurately detect and track the interactions among all players within any given video frame. The aim of this paper is to address this challenge of contact detection in American Football.

While existing literature addresses problems related to injury, collision, and contact detection in various domains, the specific task of detecting contacts in American Football remains under-explored. Current publications are limited to utilizing a Support Vector Machine (SVM) model to detect head injuries in the American Football, action recognition using deep learning in Australian football and other publications included player detection using deep learning. However, none of these approaches fully address the unique challenges of detecting contacts in American Football, where multiple players often engage in close physical proximity.

To bridge this gap, the state-of-the-art object detection or segmentation models from the mmdetection framework were used on our dataset of American Football. From this it was observed that these models sometimes struggled to accurately detect players in close proximity or merged multiple players into one. To overcome this limitation, cropping the images and re-evaluating the detection models on the cropped images was considered thus gave rise to our second research question i.e. to check performance of these state-of-the-art models on a large image and small cropped version of same image.

For the primary task of contact detection, due to lack of computational resources instead of object detection and then segregation to classes, the given csv files in the dataset [7] were used to preprocess and were segregated into two classes as mentioned in detail in Data set section. Then the SVM was used as our baseline model for contact detection based on [11] and further compare its performance with that of deep learning models, such as ResNet50 and Resnet 152 as discussed in [5]. Additionally we also evaluated performance of EfficientNet [10]. We present a thorough analysis of the training and validation accuracy and loss curves for each model, as well as other relevant metrics (as discussed in the Experimental Results section). Our research aims to provide a comprehensive evaluation of various approaches for contact detection in American Football, ultimately contributing to the development of a real-time contact detection system that can help mitigate injury risks and inform game strategies.

2. Data set and task

The dataset for American contact detection was taken from the Kaggle competition nflplayercontactdetection [7]. This collection consists of 720 videos in total which includes the Sideline, Endzone, and All29 views. We considered only the Sideline and Endzone videos because the All29 view is not timestamped. Train_player_tracking.csv, Train_labels.csv, and Train_baseline_helmets.csv are additional CSV files included with the dataset. Information on player tracking, including location, speed, and acceleration, etc., is contained in the train_player_tracking.csv file. While the train_labels.csv file includes information about contact or no contact labels, the baseline_helmets.csv file contains data about the bounding boxes of the helmets.

The videos in this dataset have a frame rate of 59.94 frames/sec, and the game starts 5 seconds into the video. Contact information is stored from the 5-second mark and every 0.1 seconds afterward until the end of the video.

We then created frames from these videos and extracted frames at intervals of one second beginning at the five-second point in order to prepare the data.

Then we extended the bounding boxes to 256 X 256 by increasing the size of the helmet bounding box information for each player in each frame. The pictures are given names in the following format: Game_play_view_step_player1Number_player2Number (for example, 58168_003392_Endzone_0_A23_A32) using the train_labels.csv and train_player_tracking.csv files. Based on the titles, images are divided into two distinct folders contact and no contact.

We randomly choosed 10,000 images from the no-contact class and about 8,000 images from the contact class because there was an class imbalance, with the majority of the images falling to the no-contact class. After which, we cropped the images to 150 x 150 pixels to reduce the number of features to train so as to avoid overfitting.

This study's goal is to identify player contact and decide whether or not that contact will harm the player. In order to solve this particular issue, the researchers intend to review previous work on object detection, apply the algorithms from this work, and extend them.

3. Methodology

For the task of Player Contact Detection in American Football we first thought of cropping images using the pretrained object detection models and to determine parameters to determine contact and no-contact and then run another model. But midway we came across a problem where when players were in close proximity the model was detecting two persons as one or not detecting one of them. So due to lack of computation prowess to retrain the model for our dataset we decided to crop images directly using the given csv files where bounding box of helmet were given. Later, we determined classes and ran Image Classification models to predict contact. Here, SVM was used as a baseline model and we compared it with resnet50, resnet 152 and efficientnetB0. For this we trained data based on made classes of contact or no-contact. For our second task of determine if missing persons in orginal images are detected in smaller cropped version of same image we used pre-trained models such as Faster R-CNN R50, Mask2Former, and YOLOX X.

We took the video dataset from Kaggle for our task. Then we applied the preprocessing steps as described in the previous section. The frames generated were then passed to three state of the art object detection models Faster R-CNN R50, Mask2Former, and YOLOX X. This is used for the second research question of seeing whether these models are able to better detect the players in a cropped image compared to the full image. The frames were also cropped based on the bounding boxes of helmet given in the dataset. The bounding boxes of the helmet were increased so that the person and its immediate surroundings become visible.

For our first research question these cropped images were passed to SVM model which we considered as our basline model and to three different deep learning models ResNet50, ResNet152, and EfficientNet. Each of the models predicted the outcome of contact or no contact. The accuracy of each of them were seen and compared which is described in detail in the experimental section. The SVM, ResNet and Efficient models are explained below with the reason as to why we chose them.

The paper by Wu et al. [11] uses SVM to detect head impacts in American Football which being used in a similar setting gave us an idea to use the SVM model as the baseline for our task of prediction player contact.

SVM is a model known to be used for the task of image classification. The feature vectors are used by the SVM technique to train a binary classification model. The model's objective is to identify a hyperplane that divides the two image classes. The hyperplane is chosen so that the distance between the hyperplane and the closest data points from each class (called support vectors) is maximized. This is known as the maximum margin approach.

After the SVM model has been trained, it can be used for binary classification of two images like in our task. The input image is again represented as a feature vector, and the SVM model then predicts which class the image belongs to based on its position relative to the hyperplane.

After training, the SVM model may be used to divide new images into the two groups. Using the input image's position in relation to the hyperplane and its representation as a feature vector once more, the SVM model determines which class the image belongs to.

For the task of binary classification SVMs can perform well but are only useful if the number of data samples is less as SVMs can be computationally expensive and may not scale well to large datasets or high-dimensional feature spaces. In addition, the choice of the feature extraction technique and the parameters of the SVM model can greatly impact its performance.

Hence we tried various deep learning models explained below to detect contact which are currently state of the art models for the task of image classification.

ResNet [6] is based on the concept of residual learning, which uses skip connections to facilitate the easier transmission of information through a network. Because of this, ResNet is able to have many deeper layers than earlier networks without encountering the vanishing gradient problem, in which gradients get smaller as they pass through more layers. In computer vision applications including picture classification, object recognition, and semantic segmentation, ResNet has been extensively employed. The paper [5] uses ResNet50, ResNet152 for action recognition in Australian football. We thought of using this for our task of contact detection in American football. The ResNet152 used in this paper performed reasonable well and gave an accuracy of 77.45%. We decided to use these models as they were successful in the research paper mentioned which had a similar surrounding environment of a football field.

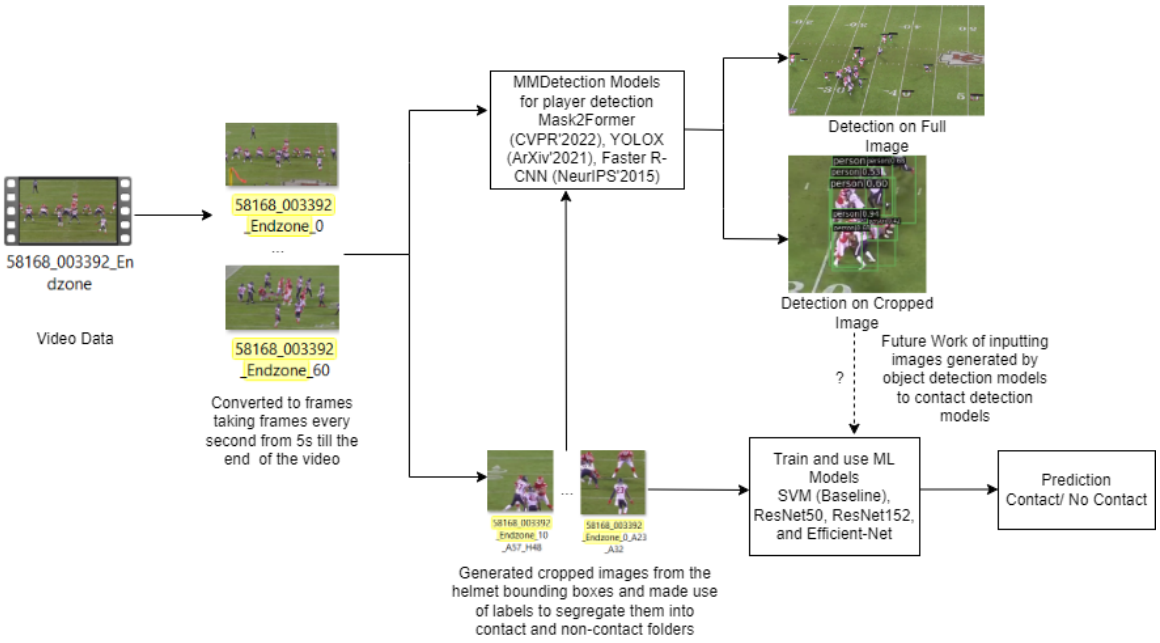

Figure 1. Model architecture

EfficientNet [10] achieves state-of-the-art performance on picture classification problems while also being computationally efficient.

EfficientNet's design is built on a compound scaling technique that combines the three scaling factors of depth, width, and resolution. Resolution is the size of the input image, depth is the number of neural network layers, width is the number of channels in each layer.

In order to improve performance on various tasks while limiting computational cost, EfficientNet scales up or down the network design using a combination of these scaling factors. Neural architecture search (NAS), a method for automating the design of neural networks, is used to optimise the architecture of EfficientNet.

The use of a new block termed the "swish" activation function, which has been proved to perform better than conventional activation functions like ReLU, is one of the main advances in EfficientNet. The use of a "squeeze-and-excitation" (SE) block, which rebalances the channel-wise feature replies to help the network focus on essential features, is another novelty.

It is mentioned in [10] that it outperforms ResNet50 and ResNet152 and thus we chose this model in addition to tht ResNet models which we got inspiration from [5].

For the second research question, we wanted to check whether the state-of-the-art models perform better on the cropped images rather than the whole playfield image to detect a person in the American Football game. The models used are explained in brief below.

Faster R-CNN R50[8] is a two-stage object detection algorithm that uses a Region Proposal Network (RPN) to generate object proposals and a Fast R-CNN network to classify objects and refine the proposals. The backbone network of Faster R-CNN R50 is ResNet-50, which has 50 convolutional layers. Faster R-CNN R50 is known for its accuracy and is a popular choice for object detection tasks that require high precision.

Mask2Former[3] is a transformer-based model that uses a self-attention mechanism to encode image features and generate object proposals. It is a one-stage object detector, meaning that it directly predicts object bounding boxes and class probabilities without generating object proposals. Mask2Former is known for its efficiency and speed, making it a good choice for real-time applications or when processing large volumes of images.

YOLOX X[4] is a variant of the YOLO family of models, which are also one-stage object detectors. YOLOX X uses a modified version of the CSPDarkNet network and has a multi-scale prediction head, allowing it to detect objects at different scales. YOLOX X is known for its speed and high recall rate, making it a good choice for detecting objects in complex scenes.

4. Experiments

As our objective was to detect contact between two players in American football game, this requires object detection as a prerequisite. In order to determine the most suitable object detection model for our task, we have decided to compare three states of the art pre-trained models on COCO dataset: R-CNN R50, Mask2Former, and YOLOX X. But due to the lack of resources we decided to come up with new approach for Contact detection. And, we came up with new question as mentioned in introduction i.e. to check if the above state-of-the-art models detect missing objects (i.e. persons in our case) in the original image are detected in cropped images.

4.1. Contact Detection

For this task we first started our experiments with the SVM model which we took as our baseline model. Due to our nature of dataset SVM was pretty slow, so we could only test in case of two scenarios. The first scenario is where we kept the default parameters of the sklearn library of python where it was able to get train accuracy of 49.51% and test accuracy of 49.19%. For the other scenario we ran it for 200 iterations and with a gamma value of 0.1 which was able to give a train accuracy of 48.03% and a test accuracy of 47%. It could have done better by fine tuning some parameters but clearly it is not a good model for our task of predicting contact between players.

We thus tried three state-of-the-art models ResNet50, ResNet152, and EfficientNet and trained our dataset with these models. We ran and plotted accuracy and loss graphs for each of the models and compared them.

First the experiments were run using the ResNet50 model. In the beginning we trained our dataset for 100 epochs and observed that the model overfitted very soon. Thus, we ran the experiments with a learning rate of 0.01 and dropout of 0.5. The model still overfitted and which can be noticed in the Figure 2. It had the best validation accuracy of 72.17%. Hence, we finally tried with augmenting the images and ran computations with the same hyperparameters using the ResNet50 model. Only 50 epochs was ran as this was quite slow and we observed that we were able to reduce the overfitting problem. This had an improved best validation accuracy of 74.89%. The loss and accuracy curves for the non-augmented and augmented graph with l2 learning rate of 0.01 and dropout of 0.5 can be seen in the Figure 2 and 3.

Further experiments were run on ResNet152 and EfficientNet without augmentation due to time constraints. It was found that ResNet152 only had slight improvement over the non-augmented ResNet50. The non-augmented EfficientNet was the best among all the three models which had the best validation accuracy of 74.49%. Their accuracy and loss curves can be seen in figures 4 and 5 respectively. The summary of the accuracy of all the models can be seen in Table 1.

Thus, we can concluded that augmenting the images helps in solving the overfitting problem and EfficientNet is the better model among the three models for the task of detecting contact in American football.

Model	Best Validation Accuracy
SVM	49.19%
SVM (gamma = 0.1)	47%
ResNet50	72.17%
ResNet50 (Augmented)	74.89%
ResNet152	73.27%
EfficientNet	74.49%

Table 1. Accuracy scores of the models for the task of contact detection.

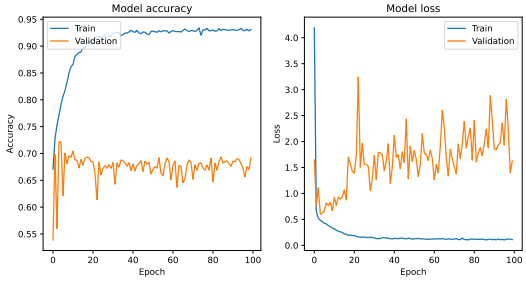

Figure 2. Accuracy and Loss curves for ResNet50 without Augmentation

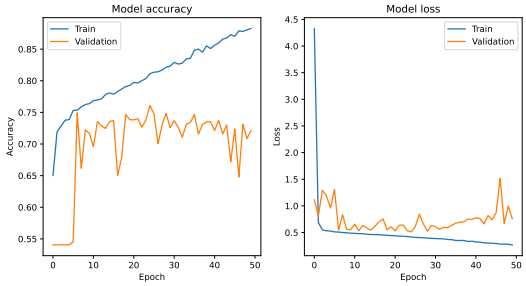

Figure 3. Accuracy and Loss curves for ResNet50 with Augmentation

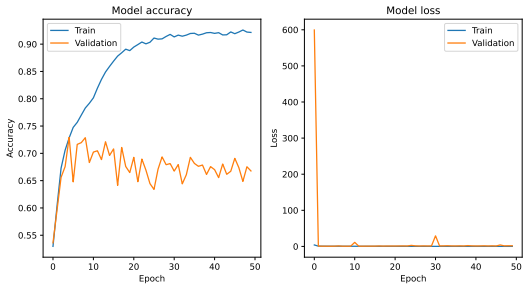

Figure 4. Accuracy and Loss curves for ResNet152

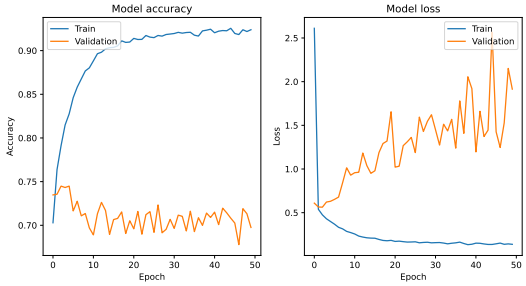

Figure 5. Accuracy and Loss curves for EfficientNetB0

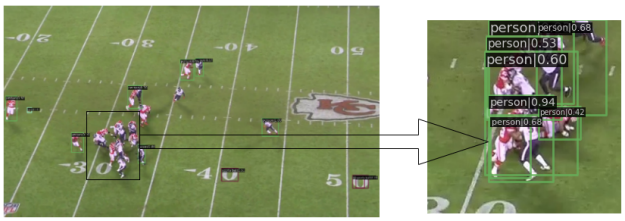

Figure 6. Faster RCNN Results of Detecting Players with full image v/s cropped image

Model	Recall	Precision	F1-score
Fast RCNN (original)	0.91	0.83	0.86
Fast RCNN (Cropped)	0.99	0.85	0.91
Mask2Former (original)	0.99	0.84	0.91
Mask2Former (Cropped)	0.91	0.89	0.90
Yolo (original)	1.00	0.85	0.92
Yolo (Cropped)	0.99	0.87	0.92

Table 2. Accuracy scores of the three models on Original and cropped images

4.2. Evaluating Object Detection Models

We evaluated the models R-CNN R50, Mask2Former, and YOLOX X on our American Football player dataset, we check their performance in detecting persons within image. This helped us determine which model is most capable of detecting and localizing players in our NFL contact detection task.

To evaluate the performance of our object detection algorithm in detecting contact between two players in NFL, we used two different settings. The first setting included the complete photo, while the second setting included cropped images of the 15 images that were provided, resulting in a total of 47 cropped images. The results for object detection were then recorded for both settings. By comparing the results from the two settings. These are shown below:

To quantitatively evaluate the performance of each object detection model, we calculated the precision and recall values for each image in both settings. These values were then averaged for each model in each setting, providing a more detailed assessment of the performance of each model.

The resulting precision and recall values provided a clear comparison of the performance of each model in each setting. By analyzing this, we were able to identify the strengths and weaknesses of each model and make informed decisions regarding the most suitable model for our NFL contact detection task.

Along with this we calculated the F1 score metric that combines both precision and recall into a single score, providing a more comprehensive evaluation of the performance of an object detection model. The F1 score ranges between 0 and 1, with a higher score indicating better performance.

From the table 2 above it can be seen that YOLOX X performed consistently well in both settings, achieving the highest F1 score of 0.9201 without cropped images and 0.9259 with cropped images. This suggests that YOLOX X is a reliable model for object detection in the NFL contact detection task, regardless of the scale and resolution of the input images.

Also, Faster R-CNN R50 and Mask2Former performed differently in the two settings. Faster R-CNN R50 achieved a higher F1 score in the cropped image setting, while Mask2Former achieved a higher F1 score in the setting without cropped images. This suggests that the performance of these models is sensitive to the image scale and resolution.

4.3. Using other pre-trained models

In addition to above tasks, we also looked for various pretrained models whichcould be relevant to our take. One of which was a state-of-the-art violence detection model [1]. We ran this model on our NFL Dataset of videos and observed the live output, but we encountered a significant number of false positives. Despite the promising results reported in the research paper, our experiment highlights the need to thoroughly evaluate and fine-tune deep learning models for specific use cases. While a model may perform well on a general dataset, it may not be suitable for a specific application i.e. sports without appropriate adjustments. Our experience also highlights the importance of thoroughly evaluating deep learning models for false positives and other types of errors. False positives can be particularly problematic in scenarios such as contact detection in American Football, where accuracy is crucial for player safety and performance analysis.Moving forward, we will continue to explore various machine learning models and techniques for contact detection in American Football and thoroughly evaluate their performance to ensure accurate and reliable results.

5. Related Work

In this section we discuss the relevant literature for the problem of contact detection in American football while highlighting the different approaches and methods used in the area.

Sports contact has frequently been detected using sensor-based methods, especially when accelerometers and gyroscopes are employed. A wearable device with an accelerometer and gyroscope was used to identify head impacts in football players in a study by [11], proving the viability of using such devices for impact detection in contact sports.

Sports contact recognition has also benefited from the use of video analysis. A real-time system for detecting actions in rugby using video analysis and computer vision techniques was suggested in a study by [5]. To detect contacts, this system used optical flow, feature extraction, and machine learning algorithms. [9] investigated the application of

convolutional neural networks (CNNs) for head contact detection in soccer matches in another research.

Numerous object detection tasks have used CNNs and other deep learning methods. A residual learning framework called ResNet was introduced by [6] and used in sports video analysis. A deep learning model built on 3D CNNs was used in a research by [2]. Similar to this, [9] used video footage to apply a CNN-based algorithm to identify head impacts in soccer players. In another research, [5] combined player tracking and head impact identification to suggest a two-stage deep learning model for detecting actions in Australian football.

By combining sensor data with video analysis, contact recognition can be accomplished in its entirety. In their study, [11] suggested a SVM system for head impact detection in American football that combined the benefits of both video analysis and sensor data.

The literature on sensor-based methods, video analysis, computer vision techniques, and deep learning techniques for real-time contact detection in American football has shown their efficacy in detecting and tracking contacts. These techniques can be combined to create a comprehensive strategy to contact detection in American football, which can aid in injury management and injury prevention. Future studies should concentrate on combining different methods and creating real-time tools that can be used in American football matches.

6. Conclusions

In conclusion, we evaluate the potential different approaches, such as deep learning models, object recognition algorithms, to try to predict player contacts during games. From the above experimental results it can be observed that without augmentation the Efficient net [10] performed the best with 74.49% accuracy. It can also be observed that resnet 50 [6] performed better with Image augmentation. This proves that Image augmentation can help improve performance of other models as well as it solves the overfitting problem. We couldn't do the experiments for other models with different image augmentation, image processing and other hyper parameter tuning due to lack of computational resources and time constraint.

For our second task, it can be observed that overall object detection models had a considerable improvement in F1-score in detecting objects in cropped images when compared to orginal image. And, the performance slightly decreased in segmentation model in detecting objects in cropped images when compared to orginal image. And, the results indicate that YOLOX X [4] is the most reliable model for object detection in the NFL [7], achieving the highest F1 score in both settings. However, the performance of the other models may also be optimized by adjusting their parameters or fine-tuning their training on the specific task.

In future we would like to generate the cropped images using the state-of-the-art object detection models by retraining them on our whole dataset and based on the segmentation or detection boxes made we would determine heuristics to determine if there could've been contact in between persons and then use this results to train a model which would classify contact or no-contact. This way the system can work on real-time without any external parameters. This can be possible with better computational resources.

Other than that, the problem with our current models was that they were overfitting, we would like to use different image augmentation techniques, image processing techniques and hyperparameters to try which would be the best combination of parameters for our given task.

Future research should focus on refining the integration of various techniques and developing practical real-time systems for use in American football games as mentioned in Future Work. This will ultimately contributing to injury prevention, better management of player contacts, and improvements in the sport as a whole.

References

[1] Aldahoul, Nouar, Karim, Hezerul Abdul, Datta, Rishav, Gupta, Shreyash, Agrawal, Kashish, and Albunni, Ahmad. Convolutional neural network-long short term memory based iot node for violence detection. Institute of Electrical and Electronics Engineers Inc., 9 2021. ISBN 9781665428996. doi: 10.1109/IICAIET51634.2021.9573691.

[2] Chen, Yixin, Dwivedi, Sai Kumar, Black, Michael J., and Tzionas, Dimitrios. Detecting human-object contact in images. 3 2023. URL http://arxiv.org/abs/2303.03373.

[3] Cheng, Bowen, Misra, Ishan, Schwing, Alexander G., Kirillov, Alexander, and Girdhar, Rohit. Masked-attention mask transformer for universal image segmentation. 12 2021. URL http://arxiv.org/abs/2112.01527.

[4] Ge, Zheng, Liu, Songtao, Wang, Feng, Li, Zeming, and Sun, Jian. Yolox: Exceeding yolo series in 2021 v100 batch 1 latency (ms) yolox-l yolov5-l yolox-darknet53 yolov5-darknet53 efficientdet5 coco ap (number of parameters (m) figure 1: Speed-accuracy trade-off of accurate models (top) and size-accuracy curve of lite models on mobile devices (bottom) for yolox and other state-of-the-art object detectors. URL https://github.com/ultralytics/yolov3.

[5] Groen, Derek, de Mulatier, Clélia, Paszynski, Maciej, Krzhizhanovskaya, Valeria V., Dongarra, Jack J., and Sloot, Peter M. A. (eds.). *Computational Science – ICCS 2022*, volume 13352. Springer International Publishing, 2022. ISBN 978-3-031-08756-1. doi: 10.1007/978-3-031-08757-8. URL https://link.springer.com/10.1007/978-3-031-08757-8.

[6] He, Kaiming, Zhang, Xiangyu, Ren, Shaoqing, and Sun, Jian. Deep residual learning for image recognition. 12 2015. URL http://arxiv.org/abs/1512.03385.

[7] Kaggle. Player contact detection. 2022. URL https://www.kaggle.com/competitions/nfl-player-contact-detection/overview.

[8] Ren, Shaoqing, He, Kaiming, Girshick, Ross, and Sun, Jian. Faster r-cnn: Towards real-time object detection with region proposal networks. URL https://github.com/.

[9] Rezaei, Ahmad and Wu, Lyndia C. Automated soccer head impact exposure tracking using video and deep learning. *Scientific Reports*, 12, 12 2022. ISSN 20452322. doi: 10.1038/s41598-022-13220-2.

[10] Tan, Mingxing and Le, Quoc V. Efficientnet: Rethinking model scaling for convolutional neural networks.

[11] Wu, Lyndia C., Kuo, Calvin, Loza, Jesus, Kurt, Mehmet, Laksari, Kaveh, Yanez, Livia Z., Senif, Daniel, Anderson, Scott C., Miller, Logan E., Urban, Jillian E., Stitzel, Joel D., and Camarillo, David B. Detection of american football head impacts using biomechanical features and support vector machine classification. *Scientific Reports*, 8, 12 2018. ISSN 20452322. doi: 10.1038/s41598-017-17864-3.