Relevant Instance Segmentation in American Football Practice Images to Aid Risky Tackle Detection

Nasik Muhammad Nafi
Dept. of Computer Science
Kansas State University
Manhattan, KS, USA
nnafi@ksu.edu

Ashley Rediger
Dept. of Computer Science
Kansas State University
Manhattan, KS, USA
ashleyrediger@ksu.edu

Scott Dietrich
Dept. of Sport and Exercise Science
Barry University
Miami Shores, FL, USA
sdietrich@barry.edu

William Hsu
Dept. of Computer Science
Kansas State University
Manhattan, KS, USA
bhsu@ksu.edu

***Abstract*—This paper addresses the problem of *relevant region* segmentation, as a pretext task for a defined multi-object scene classification task, with a specialized application to risky tackle detection from American football practice videos. The downstream task of classifying each frame from such a video as depicting a risky tackle or not depends on the interaction between the tackle-performing player and the target dummy. In both automated and manual approaches, if these two objects can not be differentiated from other objects as part of the analysis, false positive and false negative scene misclassification errors may result, to the detriment of both precision and recall. While player detection appears to be a simple human detection task, specific poses and occlusion due to the dummy make the instance segmentation task particularly challenging in the case of American football practice videos. In this paper, we present a new annotated dataset of tackle practice images and for the first time demonstrate instance segmentation in American football practice images leveraging the new dataset. Further, we show that the Cascade Mask R-CNN based segmentation approach is more suitable for the problem than another popular segmentation model, simple Mask R-CNNs, by characterizing the inherent difficulty of the task and comparing experimental results.**

***Index Terms*—american football, sports safety, risky tackle, player detection, mask segmentation, cascade r-cnn**

I. Introduction

Among American football players, concussion is one of the most frequent injuries. According to the Centers for Disease Control (CDC) data, American football has the highest share of sports-related concussions (SRC) [1] [2]. Further, other types of head injuries such as hemorrhage and edema are highly prevalent in American football. Research also shows that frequent incidents of head impacts can result in long-term neuropsychiatric and cognitive disorders, especially when a young player is involved in American football [3]. Therefore, coaches and athletic trainers emphasize learning proper tackle technique as a concussion management protocol from an early age [1] [4].

To minimize player-to-player head impact, during practice sessions, a football practice dummy, commonly known as a blocking dummy, is used to simulate a real player. Coaches may give real-time feedback on risky and safe tackles to players. However, more comprehensive feedback is provided based on the recorded session of players tackling blocking dummies. Athletic trainers or coaches generally analyze the videos offline and come up with their judgments and suggestions to the players [5] [6] [7]. The manual categorization of risky and safe tackles from practice session videos requires a significant amount of time and effort. Both manual and automated analysis of the player and the blocking dummy can be tedious, difficult, and demanding due to the background noise present in the forms of other players, varying backgrounds, service carts, etc.

Fig. 1: Sample images from our newly created dataset. Images are extracted from collected tackle practice videos.

Figure 1 presents some representative frame samples from a few tackle practice videos. The issue of background noise is clearly evident from the sample frames. Existing work that deals with American football practice videos [8] also suggests removing the unnecessary spatial regions to improve the detection accuracy of risky tackles from American football. However, such methods directly leverage a pre-trained object detection model that already consists of a `human` class to detect the player bounding box, and enlarge the bounding box to include the dummy. Further, [8] uses a simple rule-based approach to identify the player in case of multiple persons present in the image. However, our analysis reveals that direct use of such a pre-trained human detection model along with rule-based pruning of bounding boxes can easily be fooled in the case of a more cluttered environment and may select a

Fig. 2: Prior work uses spatial segmentation of the region of interest. The middle figure shows that such segmentation while removes the unnecessary elements far from the player, but retains unnecessary elements such as other players. This complicates the problem of scene understanding. Our proposed mask segmentation, on the other hand, keeps only the player and the dummy masks while removing the background.

wrong person as the tackle-performing player.

We also observe that extended spatial region selection based on only player detection often discards parts of the dummy. Also, simple spatial region extraction still retains the background noise within that spatial region (See Figure 2). We argue that as instance-level segmentation generates masks of the targeted objects (player and dummy in this case), this can significantly enhance the removal of unwanted elements while retaining the original shape of the object. Instance segmentation in individual frames will facilitate the pre-processing of the American football practice videos. Such pre-processed masked videos will, in turn, make the risky tackle detection task easier both for human judges and modern neural network-based approaches.

In this work, we introduce the new task of relevant instance segmentation in American football practice images containing a blocking dummy. The task is to segment the particular single instance of the target objects - the actual player and the dummy. We created a dataset of American football tackle practice images and manually annotated the masks for the player and dummy. We analyze the entity-specific aspects of the tackle practice scenes and the attributes of the existing state-of-the-art segmentation models to identify the effective model for the task. We observe that generic models may suffer from false positives because most of the scenes contain negative objects very similar to the positive ones (player vs. other human subject) often with a high degree of overlap. We therefore propose to leverage Cascade Mask R-CNN [9], an object detection mechanism consisting of multiple detectors, to refine predictions. Finally, we conduct a comparative analysis based on average precision at different thresholds of intersection over union (IoU) and recommend models for future segmentation tasks in the context of American football practice.

II. Related Work

Object or instance segmentation is considered a generic problem of computer vision [10] [11]. Proposal-based approaches for instance segmentation have been widely used in the last few years across many domains [12]. Most such systems first identify a number of object proposals and then attempt to classify the objects that reside within the proposals [11] [13] [9]. The efficacy of these approaches often relies on the number of good proposals predicted by the initial Region Proposal Network. Recently, transformer-based approaches have gained much popularity [14] [15] [16]; however, they require many samples to train to a state-of-the-art (SOTA)-competitive levels [10].

For sports-related segmentation tasks, domain-specific knowledge about the particular sport has generally been leveraged to improve upon the standard state-of-the-art. [17] uses a proposal-free approach for simultaneous ball segmentation and multi-player pose estimation in team sports. Several approaches have been proposed for improved player detection [18]. However, different sports require distinct treatment and refinement to enable superior performance for that sports-specific task.

III. Instance Segmentation: Player and Dummy

In this work, our objective is to precisely extract the mask segmentation of the player and the dummy, which are crucial to the task of risky tackle detection. While human detection and segmentation is a well-studied area, player segmentation in different sports scenarios remains challenging due to the lack of enough representative samples for each particular sports activity and the presence of human subjects other than the player. Therefore, we first present a new annotated dataset for American football practice images. We then train the Cascade Mask R-CNN object segmentation model to generate segmentation masks for the two classes: player and dummy.

A. A New Dataset: Data Collection and Annotation

Our dataset consists of 700 images extracted from over 100 tackle practice videos. We collected tackle practice videos from different practice fields in the United States. While recording these videos, a standard guideline was used to set up cameras. In all videos, one player performs the tackle and a dummy is placed in the same direction in which the player runs. The videos are recorded in different formats: MOV, MOD, MKV, and MP4. The frame rate for all videos is 30 frames per second (fps). Finally, we randomly extract tackle frames from the videos to ensure high diversity from a wide

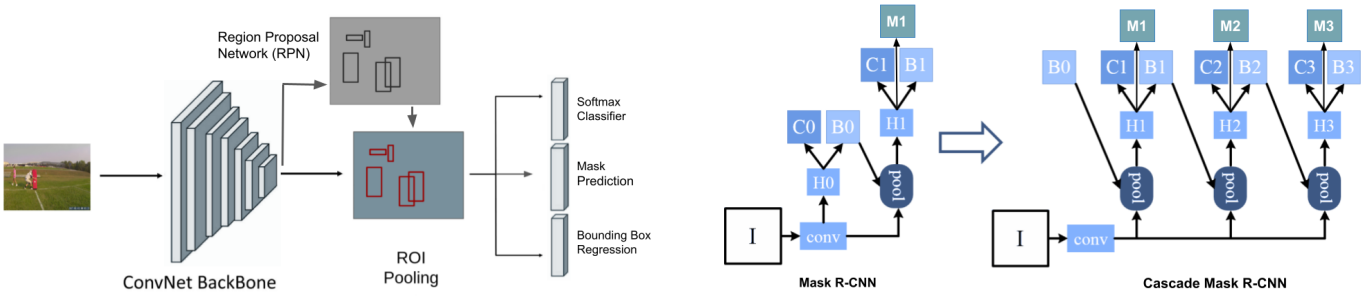

Fig. 3: **(Left)** Overall pipeline of a generic region proposal-based segmentation model. In the first phase, the ConvNet extracts a map of features, and then a Region Proposal Network (RPN) outputs potential object regions. Finally, the ROI pooling passes the features corresponding to the proposals and predicts the object class, bounding box, and mask. **(Right)** Difference between Cascade Mask R-CNN and standard Mask R-CNN models. The Cascade Mask R-CNN iteratively refines the prediction with different detector **H** and uses the prediction from the previous stage.

variety of real-world backgrounds and multi-object contexts. In some images, the tackle-performing player may not be present; instead, another person holding the dummy, or the coach, may be present. This makes the task harder as the model needs to distinguish between the tackle-performing player from the other persons. The extracted images are in the *jpg* format.

We annotate all images in the dataset by creating the segmentation masks for the dummy and the tackle-performing player if either is present in the frame. We use the make-sense.ai platform to annotate images [19]. This platform enables the generation of XML or JSON files for image annotations. We obtain the ground truth annotation by combining markup by two individual annotators.

B. Cascade Mask R-CNN for Object Segmentation

We propose to use Cascade Mask R-CNN [20] for the mask segmentation because this multi-stage variant of the R-CNN is more robust against false positives and our task is highly vulnerable to false positives. To understand the issue in depth, let's consider an image frame where there are multiple additional human subjects along with the tackle-performing player. Because of the similarity in features shared by all human subjects (as the player is indeed a human), any model is likely to predict other human subjects also as the player. However, any such prediction is a false positive that we need to avoid. The player is mainly characterized by posture. The same thing may happen if additional dummies are present in the frame (see the first image in Fig. 1 where there are a few dummy-like stuff lying on the ground).

Figure 3 shows the basic architecture of the generic proposal-based object segmentation pipeline used in Faster R-CNN or Mask R-CNN. The images are first fed into a deep Convolutional Neural Network (CNN) backbone for feature extraction. The CNN backbone can be a pre-trained network like VGG, ResNet, or any other architecture. Then, based on those spatial features Region Proposal Network (RPN) predicts multiple possible object regions with bounding box coordinates. Finally, Region of Interest (ROI) pooling takes the cropped feature maps corresponding to the proposed regions to the next stage where every proposal is classified against each class from the category list. Along with the class prediction, the models predict the mask and the bounding boxes. Thus the loss function is a combination of the three losses:

$$\mathbf{L} = \mathbf{L}_{cls} + \mathbf{L}_{box} + \mathbf{L}_{mask} \tag{1}$$

Cascade R-CNN uses a cascade of detectors as shown in Figure 3 (right) with increasing levels of precision where detectors deeper into the cascade are more selective against close false positives. One can view the multiple stages of detectors as multi-layer filtering with increasingly strict conditions. In each stage, the model defines a specific false positive target rate. While in the first stage, the model is trained with a reasonable false positive rate, in the subsequent stages the target false positive rate is reduced further. Only region proposals that cross the current stage's threshold are passed to the next stage. This helps the model to focus on hard examples in the downstream stages and discard easy negative proposals early.

Based on the discussion above, we argue that in a cascade architecture, easy negative proposals like objects other than the player and the dummy will be discarded in the early stages. While other human subjects may be considered as potential objects in the early layer they will be removed in the subsequent stages especially when the player's bounding box overlaps with the other human subject. A similar argument also holds for the dummy. Thus, a cascade architecture that employs multiple stages of the detector will reduce the false positives and improve performance for our task which is inherently prone to false positives.

IV. Experiments and Results

A. Experimental Setup and Implementation Details

We perform an image-level 70%-30% train-test split of the dataset. Thus, we have 332 player instances and 616 dummy instances in the train split while there are 156 player instances and 286 dummy instances in the test split. We created separate JSON files for train and test annotations.

Fig. 4: Examples of mask segmentation results. The first and third images from the left represent the original images while the other two images are the output of the mask segmentation of the corresponding images.

TABLE I: Hyperparameters for model training and testing

hyperparameters	values
batch size	4
optimizer	SGD
learning rate	0.00025
anchor aspect ratio	[0.5, 1.0, 2.0]
FPN used	yes
anchor sizes	[32, 64, 128, 256, 512]
environments per worker	64
total number of epochs	2000
ROI head batch size per image	128
ROI head score threshold	0.5

To show the validity of our hypothesis, we perform instance segmentation using Cascade Mask R-CNN and compare the experimental results with the popular state-of-the-art proposal-based object segmentation model Mask R-CNN. In particular, to learn American football-specific representation we fine-tune the models with Resnet-50 backbone [11] [9] [21]. Additionally, to analyze the proposal or object detection performance, along with the segmentation models we use Faster-RCNN model [13] with Resnet-50 backbone to report only bounding box prediction as Faster R-CNN does not have mask prediction.

We implement the Cascade-RCNN [9], Mask-RCNN [11], and Faster-RCNN [13] models using Detectron2 library [22] which is a PyTorch-based modular object detection and segmentation library. We initialize the models with the corresponding pre-trained weights obtained via training on the ImageNet dataset. ImageNet has a wide variety of classes totaling 1024. However, in our case, there are only two classes - *dummy* and *player*. Thus, we restructure the final layer to accommodate just two classes and train the models.

We train all the models for 3000 epochs, however, observe that the models tend to overfit after 2000 epochs. Thus, we report results for the models trained for 2000 epochs. Stochastic Gradient Descent (SGD) with momentum is used as the optimizer with a base learning rate of 0.00025. We do not opt for any learning rate decay as the chosen learning rate is already low. Table. I shows a comprehensive list of hyperparameters used in our experiments.

B. Results and Discussion

Figure 4 presents the segmentation results obtained from our trained model as opposed to the original frames. This clearly shows the benefit of mask segmentation as the position and interaction between the player and the dummy is more pronounced compared to the raw frames.

TABLE II: Comparison of the State-of-the-Art object segmentation models on our annotated dataset

Approach	mAP	AP75	AP50	APl	AR
Mask R-CNN	69.501	83.883	**93.981**	72.414	75.5
Cascade Mask R-CNN	**70.835**	**85.607**	93.250	**73.584**	**75.6**

TABLE III: Comparison of per category/class average precision for the segmentation task

Classes	Mask R-CNN	Cascade Mask R-CNN
Dummy	76.540	**77.740**
Player	62.463	**63.931**

We report evaluation results of Cascade Mask R-CNN and Mask R-CNN object segmentation model based on the COCO evaluation protocol. Table II presents the mean Average Precision (mAP) at 50% to 95% IoU, AP at 75% IoU (AP75), AP at 50% IoU (AP50), AP for large objects (APl), and Average Recall (AR) for the two segmentation method. Our reported scores are an average of three independent trials. Cascade R-CNN achieves higher mAP, AP75, APl, and AR, and Mask R-CNN performs better at lower IoU as evidenced by the value of AP50. Comparatively low performance at lower IoU while achieving high mAP denotes that Cascade Mask R-CNN performs better at higher IoUs. That means Cascade Mask R-CNN produces high-quality masks that attain higher overlap with the ground truth. Table III shows that Cascade Mask R-CNN improves the mAP of each category or class.

Further, we present bounding box detection results for three different models in Table IV. Cascade Mask R-CNN outperforms all other models in terms of both classes. This represents the validity of our proposed notion. As this variant of Cascade R-CNN uses ResNet-50 as the backbone which is computationally efficient and achieves better performance, we suggest Cascade Mask R-CNN for practical use.

TABLE IV: Comparison of per category/class average precision for the object detection task

Classes	Faster R-CNN	Mask R-CNN	Cascade Mask R-CNN
Dummy	74.222	75.040	**77.451**
Player	74.062	74.970	**79.689**

V. Conclusion

In this paper, we present an annotated dataset for instance segmentation of players and dummies in American football practice video frames. We leverage state-of-the-art object detection and segmentation models to extract masks of the instances. We demonstrate Cascade Mask R-CNN performs better than the other approaches. Limitations of our current work include comparatively lower performance on the player class. The use of an automated instance segmentation approach deployed in real-time video frames can greatly facilitate the coaches' ability to decide faster. Further, extracted masks can also be used in different downstream tasks such as classifying risky tackles from the videos using deep learning models.

References

[1] M. E. Kelley, J. M. Kane, M. A. Espeland, L. E. Miller, A. K. Powers, J. D. Stitzel, and J. E. Urban, "Head impact exposure measured in a single youth football team during practice drills," *Journal of Neurosurgery: Pediatrics*, vol. 20, no. 5, pp. 489–497, 2017.

[2] J. Cournoyer and B. L. Tripp, "Concussion knowledge in high school football players," *Journal of athletic training*, vol. 49, no. 5, pp. 654–658, 2014.

[3] M. Alosco, A. Kasimis, J. Stamm, A. Chua, C. Baugh, D. Daneshvar, C. Robbins, M. Mariani, J. Hayden, S. Conneely *et al.*, "Age of first exposure to american football and long-term neuropsychiatric and cognitive outcomes," *Translational psychiatry*, vol. 7, no. 9, pp. e1236–e1236, 2017.

[4] Z. Y. Kerr, S. Yeargin, T. C. Valovich McLeod, V. C. Nittoli, J. Mensch, T. Dodge, R. Hayden, and T. P. Dompier, "Comprehensive coach education and practice contact restriction guidelines result in lower injury rates in youth american football," *Orthopaedic journal of sports medicine*, vol. 3, no. 7, p. 2325967115594578, 2015.

[5] D. J. Lessley, R. W. Kent, J. R. Funk, C. P. Sherwood, J. M. Cormier, J. R. Crandall, K. B. Arbogast, and B. S. Myers, "Video analysis of reported concussion events in the national football league during the 2015-2016 and 2016-2017 seasons," *The American journal of sports medicine*, vol. 46, no. 14, pp. 3502–3510, 2018.

[6] G. J. Tierney, C. Kuo, L. Wu, D. Weaving, and D. Camarillo, "Analysis of head acceleration events in collegiate-level american football: A combination of qualitative video analysis and in-vivo head kinematic measurement," *Journal of Biomechanics*, vol. 110, p. 109969, 2020.

[7] M. Kelley, J. Urban, D. Jones, A. Powers, C. T. Whitlow, J. Maldjian, and J. Stitzel, "Football concussion case series using biomechanical and video analysis," *Neurology*, vol. 91, no. 23 Supplement 1, pp. S2–S2, 2018.

[8] N. M. Nafi, S. Dietrich, and W. Hsu, "Risky tackle detection from american football practice videos using 3d convolutional networks," in *Machine Learning and Data Mining in Pattern Recognition, 18th International Conference on Machine Learning and Data Mining, MLDM 2022, New York, USA, July 16-21, 2022 Proceedings*, P. Perner, Ed. ibai publishing, 2022, pp. 105–119.

[9] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6154–6162.

[10] X. Li, H. Ding, W. Zhang, H. Yuan, J. Pang, G. Cheng, K. Chen, Z. Liu, and C. C. Loy, "Transformer-based visual segmentation: A survey," *arXiv preprint arXiv:2304.09854*, 2023.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[12] P. Bharati and A. Pramanik, "Deep learning techniques—r-cnn to mask r-cnn: a survey," *Computational Intelligence in Pattern Recognition: Proceedings of CIPR 2019*, pp. 657–668, 2020.

[13] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *arXiv preprint: 1506.01497*, 2015.

[14] J. Jain, J. Li, M. T. Chiu, A. Hassani, N. Orlov, and H. Shi, "Oneformer: One transformer to rule universal image segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 2989–2998.

[15] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask dino: Towards a unified transformer-based framework for object detection and segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3041–3050.

[16] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[17] S. A. Ghasemzadeh, G. Van Zandycke, M. Istasse, N. Sayez, A. Moshtaghpour, and C. De Vleeschouwer, "Deepsportlab: a unified framework for ball detection, player instance segmentation and pose estimation in team sports scenes," *arXiv preprint arXiv:2112.00627*, 2021.

[18] K. Vats, P. Walters, M. Fani, D. A. Clausi, and J. S. Zelek, "Player tracking and identification in ice hockey," *Expert Systems with Applications*, p. 119250, 2022.

[19] P. Skalski, "Make Sense," https://github.com/SkalskiP/make-sense/, 2019.

[20] Z. Cai and N. Vasconcelos, "Cascade r-cnn: High quality object detection and instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1483–1498, 2021.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.