2023 2nd International Conference on Innovations and Development of Information Technologies and Robotics (IDITR) | 979-8-3503-4557-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/IDITR57726.2023.10145987

# Deep Learning for Semantic Segmentation of Football Match Image

1st Yutian Wu
*Key Laboratory of Knowledge Automation for Industrial Processes of Ministry of Education, School of Automation and Electrical Engineering, University of Science and Technology Beijing,*
Beijing, China
wuyutian@ustb.edu.cn

2nd Wuqi Zhao
*Department of physical education, Beijing International Studies University,*
Beijing, China
zhaowuqi@bisu.edu.cn

3rd Chen Huang
*School of Automation and Electrical Engineering, University of Science and Technology Beijing,*
Beijing, China
U202141531@xs.ustb.edu.cn

4th Yaming Xi
*School of Automation and Electrical Engineering, University of Science and Technology Beijing,*
Beijing, China
U202142566@xs.ustb.edu.cn

5th Qing Li
*Key Laboratory of Knowledge Automation for Industrial Processes of Ministry of Education, School of Automation and Electrical Engineering, University of Science and Technology Beijing,*
Beijing, China
liqing@ies.ustb.edu.cn

6th Heng Wang
*Key Laboratory of Knowledge Automation for Industrial Processes of Ministry of Education, School of Automation and Electrical Engineering, University of Science and Technology Beijing,*
Beijing, China
hengwang@ustb.edu.cn

***Abstract*—As one of the most popular sports, football has been a subject to growth and advancements in technology. The combination of football and artificial intelligence is expected to be used for intelligent football analysis. Image semantic segmentation is an important basis for image analysis and understanding. This paper proposes a deep learning-based image segmentation model for pixel-level classification of the video recordings frames of football matches. Every pixel of football video frame is classified into one of the 10 classes, e.g., players, ball, goal bar and several background scenes. In this paper, we first test a variety of CNN architectures and pre-trained models and select the MobileNet-UNet architecture as our baseline. We note the severe unbalanced data distribution in football scene segmentation. To solve this problem, the weighted multi-class cross-entropy loss is adopted in training of MobileNet-UNet to redistribute the weights of classification loss, focusing on smaller foreground object classes and improving segmentation accuracy. We also propose to use image transformations and a random mixture sampling technique for training data augmentation to reduce model overfitting. The model is trained and validated in the well-annotated Football Semantic Segmentation Open Dataset. The proposed best model achieves 0.96 frequency weighted IoU and 0.90 mean IoU segmentation accuracy on validation set.**

***Index Terms*—semantic segmentation, football, deep neural network, UNet, data augmentation, weighted cross-entropy**

## I. Introduction

Since this century, artificial intelligence has penetrated into every aspect of daily life. Combining football and artificial intelligence technology is expected to be used for more advanced intelligent football analysis. At present, scholars are working on the in-depth combination of latest technology and football, such as detection [1], tracking [2] and post estimation of football and players, intelligent analysis and judgement of player behaviour, 3D reconstruction [3], identification of key events and key moments of football match [4].

Intelligent analysis of football matches first relies on various sensors, mainly cameras, to obtain data describing football scenes. In recent years, the development of computer vision has made it possible to model the biological structure of brain with deep neural networks to analyse and process camera images with high precision and speed. Compared with traditional image processing methods, deep neural networks [5] contain a larger number of parameters, which can solve more complex and high-level tasks, and can be automatically trained through labelled data, which greatly reduces the workload of tuning huge parameters. Moreover, with the development of computing resources and developing tools, a neural network model can often be easily constructed and trained in a few days or even hours.


Authorized licensed use limited to: National Taiwan Univ of Science and Technology. Downloaded on April 19,2024 at 02:37:48 UTC from IEEE Xplore. Restrictions apply.

Image semantic segmentation is the foundation of computer vision and image understanding. Semantic segmentation aims to identify and segment the image at the pixel level, to predict the class and precise location information of objects. The output of semantic segmentation can further help object detection, tracking, image analysis and understanding. Image semantic segmentation has been widely used in autonomous driving [6] and medical image processing [7]. With advances in deep learning, the accuracy of semantic segmentation has been continuously improved [8]–[10].

Previous studies have proposed to segment a small set of objects in football match, such as players or lines [11], referees [12].This paper proposes a deep learning model for semantic segmentation of the whole scene, gives a more complete object and scene classification for football image, including the identification of players, ball, goal bar and several background scenes. We also propose image augmentation techniques as well as model optimization ways to improve segmentation accuracy. To be specific, we first test various deep learning models in football match image and select the best performing model MobileNet-UNet [13] as our baseline model. Data often play a decisive role for the training of deep learning model. We use the recently released Football Semantic Segmentation Open Dataset (FSSOD) [14] as our train and validate dataset, which has a total of 100 frame. However, the training data are too less to train a deep learning model, which easily leads to overfitting and poor accuracy. Therefore, we propose four data augmentation techniques to expand the training data and randomly mix the augmentation techniques to further increase data diversity. Besides, in football match scene, there is an issue of unbalanced distribution of categories. Ball always occupies a small portion of pixels on the entire image, while pixels of grass fields, advertising boards, etc., occupy large proportions. The commonly used multi-class cross-entropy loss function gives equal attention to each class, and thus is prone to poor accuracy for classes with less pixels. To solve this problem, we adopt weighted multi-category cross-entropy loss to redistribute the weight of classification loss, to let the network focuses on foreground object categories, and further improve segmentation accuracy. We test and train the proposed model on FSSOD. Experimental results show that the proposed model can effectively perform semantic segmentation after training by a small amount and unbalanced data of football match image. The proposed model is introduced in section II. Section III describes the FSSOD dataset and data augmentation method. Section IV analyses the experiment results. Finally, we make conclusion in section V.

## II. Deep Neural Networks for Image Segmentation

### A. Network structure

Semantic segmentation is to classify each pixel of an image into different kinds of objects or backgrounds. The original deep learning methods [15] convert image segmentation into a patch classification problem, using CNN with fully connected layers as the end of the network to extract surrounding pixels of the target pixel to predict its classification result. However, the continuous downsampling of CNN lost spatial information, and the fully connected layer restricts the input to be fixed size and brings heavy computation. FCN [8] is a milestone work which proposed to use the deconvolution layer to learn to recover the spatial resolution. The fully connected layer is replaced by $1 \times 1$ convolution, so that the model can process input of any size. Also combining fine and coarse layers by using skip layers allows the model to make local predictions with concern of global semantic information.

Compared with FCN, UNet [7] adopts the Encoder-Decoder structure and adds convolution operations after deconvolution. High-resolution features from the encoder layers are fused with the upsampled high-semantic features. Successive convolutional layers in decoder can then learn to assemble more accurate outputs based on the fused feature.

UNet projects the discriminative features learnt at different stages of the encoder onto the pixel space. Usually, we adopt transfer learning, to reuse pre-trained image classification network [13], [16] as encoder. This allows model exploits the rich features learned from large-scale classification datasets to help image segmentation. Since the amount of football segmentation data are relatively small, the pre-trained network we choose is supposed to have smaller parameters to avoid overfitting. In Section 4, we tested and compared several popular models and chose the MobileNet, which has the smaller number of parameters but the best performance, as the encoder of UNet.

### B. Loss Function

The most commonly used loss function in image segmentation is the pixel-wise cross-entropy(CE) loss. Cross-entropy measures the difference between two probability distributions. It is also widely used as image classification loss.

The cross-entropy loss checks each pixel individually, comparing the prediction vector with the label vector $\hat{y}$. Football match scene segmentation is a multi-class prediction. Assume that we have a total of $K$ classes, the output of MobileNet-UNet is a $K$-dimensional vector. We first use the softmax activation function to normalize the output into probability distribution $y$ , then calculate the cross-entropy loss.

$$L_{CE} = -[ylog\hat{y} + (1-y)log(1-\hat{y})] \tag{1}$$

Cross-entropy loss evaluates the class prediction for each pixel separately and then averages the loss across all pixels, to let the network equally learn for each pixel in the image. In football match images, there is a problem of unbalanced class distribution. For example, as shown in Fig. 1, ball has very few pixels in image. We quantified the average portion of pixels of each class in the FSSOD dataset, shown in Fig. 2. On average, the pixels of ball and goal bar only account for 0.11%, 1.45% of the entire image, while person classes, including players and referee account for 17.2% in total. Ground, advertisement, and audience classes account for a large proportion, which are 40.35%, 15.63%, and 25.27% respectively. The imbalanced data distribution causes the network loss and gradient in


Authorized licensed use limited to: National Taiwan Univ of Science and Technology. Downloaded on April 19,2024 at 02:37:48 UTC from IEEE Xplore. Restrictions apply.


Fig. 1. Example of data on FSSOD. Left column: football match images, right column: corresponding label images, bottom row: the legend of label image. Best viewed in colors.

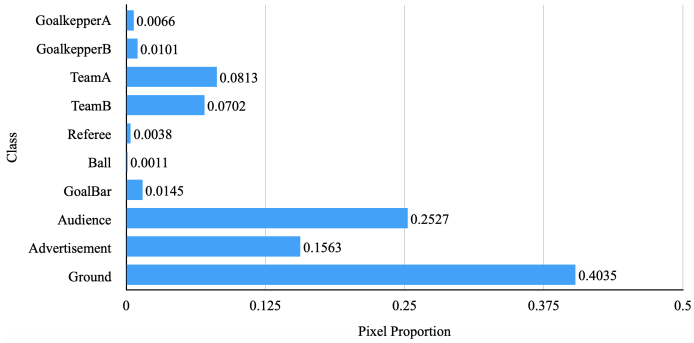



Fig. 2. The average proportion of pixels per class in FSSOD.

back propagation stage dominated by the classes with a large number of pixels, may lead to the poor accuracy in classes with less pixels.

To alleviate imbalanced learning, we add a class-wise weight parameter to cross-entropy loss, which denotes as weighted cross-entropy (WCE) loss, to give more punishment when misclassification occurs in the classes with less data. $\beta$ is a K-dimensional weight vector.

$$L_{WCE} = -[\beta y log\hat{y} + (1-y)log(1-\hat{y})] \tag{2}$$

## III. Football Segmentation Dataset and Data Augmentation Method

The source data of FSSOD was collected from the UEFA Super Cup match between Real Madrid and Manchester United in 2017. FSSOD totally has 100 frames each with fine pixel-level annotation. Every frame is filled with football match scene, while blurred frame or outliers is been replaced. There are 10 classes defined in FSSOD, including the goal-keepers, players of team A and B, referee, ball, goal bar, audience, advertisement, ground. To train and test the proposed deep learning model, we randomly divide FSSOD into $5:1$ and get 83 frames for training and 17 frames for validation.

83 frames are too less for a segmentation model training, it is easy to cause overfitting which means the model may memorize the training data and show poor performance in unseen validate data. In medical image segmentation, researchers exploit image augmentation to expend data. Image transformations is equally applied to both images and labels during training to create data diversity. In this paper, we

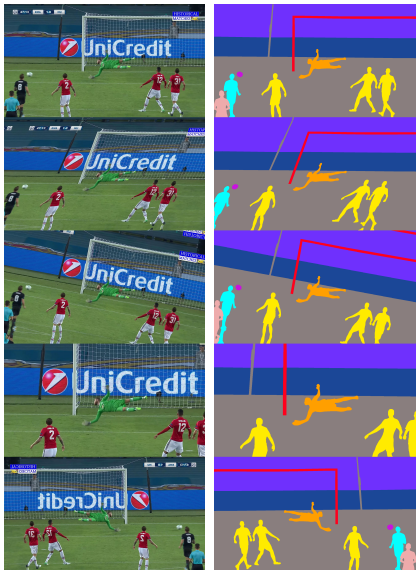


Fig. 3. Examples of image transformation in proposed data augmentation, Top row: original image and label, subsequent rows from top to down: image transformation of shear, rotating, scaling (zoom in/out), flipping.

apply the data augmentation to football match segmentation task to increase training data. Shooting football match has the characteristics of multi-angle, multi-rotation, zoom in and out, thus, we apply $4$ kinds of similar image transformations, including flip, scale, shear, rotate. The example of proposed image transformation is shown in the Fig. 3. We also propose a mixture sampling strategies to randomly select $[0, 4]$ kinds of image transformation method for each augmentation, which greatly increase the amount and diversity of training data.

## IV. Experiments

### A. Implementation Details

The proposed model is built using the Tensorflow Keras deep learning framework and trained by GPU. We use Adam optimizer with learning rate of $0.001$, train the model for $60$ epochs and record the model weight with best accuracy on validation set. The image size of FSSOD is $(1080, 1920)$. In order to improve the processing speed, we resize the image to $(416, 608)$ for training and validation. Class-wised weight parameters in WCE loss are set to $[8, 4, 2, 2, 8, 16, 4, 1, 1, 1]$. The parameters of image transformation in data augmentation are set as follows: the shear angle is randomly selected from $(-5, 5)$, the rotate angle is randomly selected from $(-5, 5)$, the scaling rate is randomly selected from$(1, 1.5)$, and the image is flipped with a probability of $0.5$.

To evaluate the performance of the proposed model, we compare the predicted segmentation result and the ground truth. The metrics we used are mean Intersection-Over-Union(MIoU) and frequently weight IoU(FWIoU), which are two of the most commonly used metrics in semantic segmentation task. IoU is defined as the area of overlap between the predicted segmentation and the ground truth divided by the area of union between the predicted segmentation and the



Authorized licensed use limited to: National Taiwan Univ of Science and Technology. Downloaded on April 19,2024 at 02:37:48 UTC from IEEE Xplore. Restrictions apply.

TABLE I
RESULTS ON FSSOD WITH FREQUENCY WEIGHTED IOU AND MEAN IOU

| Model | | | *Param* | *FWIoU* | *MIoU* |
|---|---|---|---|---|---|
| FCN | | | 69M | 0.9226 | 0.7784 |
| PSPNet | | | 3M | 0.8396 | 0.5846 |
| SegNet | | | 3M | 0.9051 | 0.6970 |
| VGGUnet | | | 12M | 0.9502 | 0.8364 |
| MobileNetUnet | | | 6M | 0.9536 | 0.8428 |
| | Aug. | WCE. | | | |
| MobileNetUnet | ✓ | | 6M | 0.9584 | 0.8793 |
| MobileNetUnet | ✓ | ✓ | 6M | 0.9615 | 0.9011 |

TABLE II
ABLATION STUDY RESULTS OF IMAGE AUGMENTATION ON FSSOD

| Flip | Scale | Shear | Rotate | Random | *FWIoU* | *MIoU* |
|---|---|---|---|---|---|---|
| | | | | | 0.9536 | 0.8428 |
| ✓ | | | | - | 0.9563 | 0.8750 |
| | ✓ | | | - | 0.9555 | 0.8656 |
| | | ✓ | | - | 0.9550 | 0.8617 |
| | | | ✓ | - | 0.9572 | 0.8735 |
| ✓ | ✓ | ✓ | ✓ | | 0.9584 | 0.8768 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 0.9584 | 0.8793 |

ground truth. MIoU is to evaluate the multi-class segmentation performance, which is calculated by averaging the IoU of each class. Assuming there are $K$ classes, $P_{ij}$ indicates the number of pixels labeled as class $i$ and predicted as class $j$.

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \quad (3)$$

FWIoU further extends the MIoU metrics by weighting the IoU of each class by its frequency of occurrence.

$$FWIoU = \frac{1}{\sum_{i=0}^{k}\sum_{j=0}^{k}p_{ij}}\sum_{i=0}^{k}\frac{p_{ii}\times\sum_{j=0}^{k}p_{ij}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}} \quad (4)$$

### *B. Results*

We test the performance of various deep learning semantic segmentation models in FSSOD for football scene segmentation. Table I shows that the segmentation accuracy of MobileNet-UNet is 95.36% FWIoU and 84.28% MIoU, which is the highest among all popular models. Moreover, the weight parameters of MobileNet-UNet are small, which indicates a low computational resource cost to create the model. Thus, we choose MobileNet-UNet as our baseline model.

After adding image enhancement, the segmentation accuracy is increased by 0.48% FWIoU and 3.65% MIoU. We further change the normal cross-entrophy loss to weighted cross-entrophy loss, the accuracy is increased by 0.31% FWIoU and 2.18% MIoU, which is the best model among all tested models. Fig. 4 shows the model inference results on unseen validate set of FSSOD. The baseline MobileNet-UNet can gives better classification of object details compared with SegNet and FCN. After adding WCE loss and image augmentation in training stage , we obtain a well-trained model which can perfectly classify most of pixels. Besides, because of the adopt of WCE loss, our proposed model can provide refined classification for small object class with less pixels, such as balls and goal bars.

We test the ablation experiments of adding four kinds of image transformation as image augmentations. Table II shows that after adding any kind of flip, scale, shear, and rotate, the segmentation accuracy is improved compared to the baseline. When sequentially implementing four kinds of transformations together, the result improves by 0.48% FWIoU and 3.4% MIoU, which indicates that the four transformations can well cooperate. We further perform random and mixture sampling of four transformations. The more randomness is added, more diverse training data is created. Result shows the accuracy gets the biggest boost of 0.48% FWIoU and 3.65% MIoU.

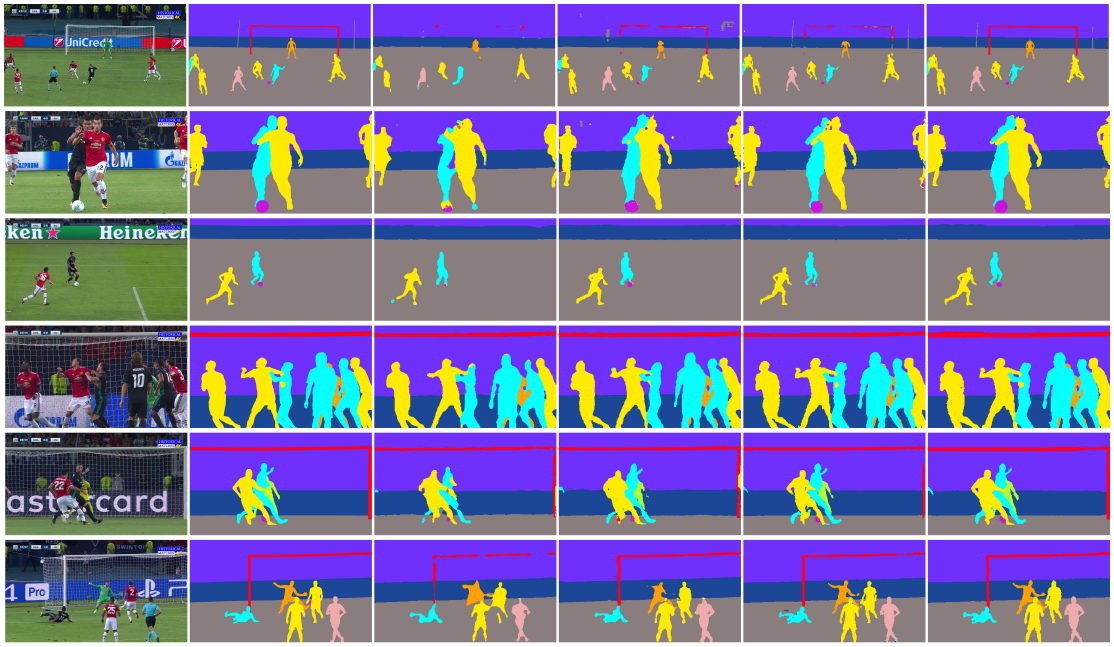
We also conduct ablation experiments to evaluate the model improvement of updating the loss from cross-entropy loss to weighted cross-entropy loss, and record the classification accuracy improvement by class in Table III. It can be seen that when using only cross-entropy loss, FMIoU has a relative higher score of 95.84% but MIoU only has 87.93%. The difference between FMIoU and MIoU is 7.91%. From the class-wise results, we can infer that the large difference between FMIoU and MIoU is caused by the inconsistent model classification ability for different classes. The model can better classify the classes with higher frequency of occurrence, such as advertisement, audience, and ground, all of which are above 95% IoU. However, the accuracy for small classes is poor, the IoU of ball is 77.1%, and the IoU of Referee is 84.72%, etc. After using WCE to enhance the loss of small classes, FMIoU and MIoU are increased respectively, and their gap becomes smaller. Meanwhile, small data classes get a bigger boost, the IoU of ball is increased to 81.95%, and the IoU of referee is increased to 92.53%. This shows that the weighted cross-entropy loss effectively improves the performance of football semantic segmentation with imbalanced class distributions.

## V. CONCLUSION

This paper proposed a deep learning model which can perform pixel-level semantic segmentation of football match image. We compared a variety of semantic segmentation models and built MobileNet-UNet as a baseline model. To solve the problem of insufficient football data and unbalanced data distribution, we proposed random sampling data augmentation techniques and weighted multi-category cross-entropy loss to improve the learning of model. Without affecting the inference speed, the accuracy of the model is greatly improved.

Currently, football image segmentation performance is still limited by the amount of training data. In the future research, our algorithm can also be considered as a semi-automatic labeling tools to provide more training data and further contribute to the construction of robust large-scale football match image segmentation models.


Authorized licensed use limited to: National Taiwan Univ of Science and Technology. Downloaded on April 19,2024 at 02:37:48 UTC from IEEE Xplore. Restrictions apply.



(a) Input image (b) Ground truth (c) SegNet (d) FCN (e) MobileNet-UNet (f) Ours

Fig. 4. Comparison of model inference results on unseen validate set of FSSOD. From left to right: football match image, ground truth label, prediction of other models and our proposed model.

TABLE III
COMPARISON OF CROSS-ENTROPY LOSS AND WEIGHTED CROSS-ENTROPY LOSS ON FSSOD WITH CLASS-WISE SEGMENTATION RESULT

| **Loss** | ***FWIoU*** | ***MIoU*** | **IoU per category** | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | ***GK.A*** | ***GK.B*** | ***TeamA*** | ***TeamB*** | ***Referee*** | ***Ball*** | ***GoalBar*** | ***Aud.*** | ***Adv.*** | ***Ground*** |
| CE | 0.9584 | 0.8793 | 0.7648 | 0.8458 | 0.9130 | 0.9078 | 0.8472 | 0.7710 | 0.8428 | 0.9674 | 0.9524 | 0.9804 |
| WCE | 0.9615 | 0.9011 | 0.8136 | 0.8696 | 0.9165 | 0.9171 | 0.9253 | 0.8195 | 0.8449 | 0.9692 | 0.9526 | 0.9824 |

## ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (62173029).

## REFERENCES

[1] J. Komorowski, G. Kurzejamski, and G. Sarwas, "Footandball: Integrated player and ball detector," *arXiv preprint arXiv:1912.05445*, 2019.

[2] S. Hurault, C. Ballester, and G. Haro, "Self-supervised small soccer player detection and tracking," in *Proceedings of the 3rd international workshop on multimedia content analysis in sports*, 2020, pp. 9–18.

[3] K. Rematas, I. Kemelmacher-Shlizerman, B. Curless, and S. Seitz, "Soccer on your tabletop," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4738–4747.

[4] A. Deliege, A. Cioppa, S. Giancola, M. J. Seikavandi, J. V. Dueholm, K. Nasrollahi, B. Ghanem, T. B. Moeslund, and M. Van Droogenbroeck, "Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4508–4519.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[6] X. Pan, Y. Wu, and H. Ogai, "Automatic training data generation method for pixel-level road lane segmentation," in *International Conference on Genetic and Evolutionary Computing*. Springer, 2019, pp. 473–481.

[7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[9] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[11] A. Cioppa, A. Deliege, and M. Van Droogenbroeck, "A bottom-up approach based on semantics for the interpretation of the main camera stream in soccer games," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1765–1774.

[12] J. R. Nunez, J. Facon, and A. de Souza Brito, "Soccer video segmentation: referee and player detection," in *2008 15th International Conference on Systems, Signals and Image Processing*. IEEE, 2008, pp. 279–282.

[13] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[14] A. AI, S. Roomy, M. M. Islam, A. B. S. Nayem, A. M. Tonmoy, and S. M. S. Islam, "Football (semantic segmentation)," 2022. [Online]. Available: https://www.kaggle.com/dsv/4572474

[15] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, "Learning hierarchical features for scene labeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1915–1929, 2012.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.



Authorized licensed use limited to: National Taiwan Univ of Science and Technology. Downloaded on April 19,2024 at 02:37:48 UTC from IEEE Xplore. Restrictions apply.