Journal of Information Systems Engineering and Management
2025, 10(45s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

SoccerSAM: Leveraging Segment Anything Model for Unlabeled Semantic Segmentation of Football Match Scenes

Chen Zhang[1] , Wan Ahmad Munsif Wan Pa[1*], Nur Shakila Mazalan[1]

[1]*Faculty of Education, Universiti Kebangsaan Malaysia, Bangi, Selangor Malaysia, 43600, Malaysia*

ARTICLE INFO

Received: 25 Dec 2024

Revised: 15 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

Introduction: This paper addresses a novel and challenging task: how to leverage the emerging Segment Anything Model (SAM), which demonstrates impressive zero-shot instance segmentation capabilities, to train a compact semantic segmentation model for football scenes (student) without requiring any labeled data.

Objectives: This presents significant challenges due to SAM's inability to provide semantic labels and the considerable capacity gap between SAM and the student model. To solve this, we introduce a novel framework, SoccerSAM, which incorporates a Semantic Bridging Module (SBM) to bridge this gap.

Methods: The SBM assigns semantic class probabilities to SAM-generated instance masks, producing dense semantic logits for training. To further enhance the model's boundary accuracy, we introduce a Boundary-Aware Consistency Loss that aligns the predicted edges with SAM's high-quality boundary information. Additionally, we propose a Logit-Level Consistency Loss to enforce alignment between the student's predictions and the pseudo-labels generated by SBM.

Results:Extensive experiments on the FSSOD dataset show that our SoccerSAM outperforms previous methods, achieving significant improvements in both mIoU and FWIoU, with a remarkable boost in performance, especially in small object segmentation, while maintaining a lightweight architecture.

Conclusions: In this paper, we introduced SoccerSAM, a novel framework for football scene segmentation that leverages the Segment Anything Model (SAM) and the Semantic Bridging Module (SBM) to generate soft semantic supervision without requiring labeled data. Extensive experiments on the FSSOD dataset demonstrate that SoccerSAM outperforms existing methods in terms of both mIoU and FWIoU, achieving significant improvements, especially in small object segmentation. Our approach provides a lightweight and effective solution for real-time football scene segmentation, and future work will explore extending the framework to handle more complex scenes and incorporate additional domain-specific knowledge.

Keywords: Soccer, Semantic Segmentation, Segment Anything Model, Knowledge Transfer

INTRODUCTION

Over the past decade, sports analytics has seen significant growth, fueled by advancements in artificial intelligence and computational tools [9]. In addition, the market was valued at over 1 billion dollars, and projections suggest it will grow by 500% in the next 5–10 years [7][3]. As a result, sports analytics is poised to play an even more crucial role in the sports industry moving forward. Several companies are already providing analytics services to sports clubs, aiming to enhance performance and improve rankings, which in turn generates increased revenue from ticket sales, advertising, and merchandise.

With the development of AI [5][16] and the high cost of manual annotation, most sports analytics products have shifted towards automated systems using AI and computer vision. One key aspect of these systems is segmentation [4][10][18],which plays an essential role in accurately analyzing the game. Segmentation is critical because it allows for the identification and precise localization of key objects, such as players and the ball, at the pixel level. This enables

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

more reliable tracking of player movements, assessment of team coverage, and analysis of key actions such as passes, all of which are integral to understanding game dynamics.

Previous studies have proposed segmenting specific elements in football match images, such as players and field lines [2], or referees [14]. In addition, [19] extended this idea to whole-scene segmentation, covering a broader range of objects including players, goalposts, and the ball, thereby enabling richer contextual understanding of the game. However, these approaches heavily rely on large-scale annotated datasets, which are time-consuming and expensive to obtain, and incur significant computational costs during training. This reliance on fully supervised learning limits their scalability and applicability, especially in dynamic or real-time scenarios where rapid domain adaptation is essential. Recently, there has been significant progress in the development of foundational models [8][13][17]. Large visual models (LVMs), like the Segment Anything Model (SAM) [8], which is trained on extensive datasets (over 1 billion masks across 11 million images), have garnered considerable attention. SAM's outstanding zeroshot instance segmentation capabilities on unseen datasets and tasks make it an ideal foundational model for a wide range of segmentation applications [12][20].

In this work, we address a new challenge in sports video understanding: how to exploit the powerful instance segmentation ability of the Segment Anything Model (SAM) to achieve semantic segmentation of football match scenes—without relying on any labeled data. Although SAM demonstrates impressive generalization and zero-shot performance, directly applying it to dense pixel-level segmentation tasks in structured sports scenes presents two key challenges: (1) SAM is designed for generic instance segmentation and lacks the ability to assign semantic categories; (2) there exists a notable discrepancy in model capacity and task granularity between SAM and lightweight segmentation networks typically used for real-time sports analytics.

To this end, we introduce SoccerSAM, a novel framework designed to achieve semantic segmentation of football match scenes without relying on labeled data. SoccerSAM leverages the strong generalization ability of the Segment Anything Model (SAM) and introduces a Semantic Bridging Module (SBM) to mediate the gap between instance-level segmentation and dense class-aware semantic understanding. Our SoccerSAM enjoys two key technical contributions. Specifically, we first propose the Semantic Bridging Module (SBM), which converts the category-agnostic masks generated by SAM into class-aware pseudo labels through visual-semantic alignment. This module enables us to generate reliable semantic supervision without any human annotations, allowing the student segmentation network to learn from SAM's structural priors while capturing class semantics. Second, we introduce a Boundary-Aware Consistency Loss, which enforces spatial alignment between the student's predictions and SAM's high-quality instance boundaries. This design significantly enhances the model's ability to distinguish closely packed objects—such as overlapping players or a player interacting with the ball—by guiding the network to refine boundaries at the pixel level. Together, these two innovations enable SoccerSAM to train a compact, task-specific segmentation network for football analysis using only unlabeled match imagery, making it highly scalable and practical for real-world applications where annotation is costly or unavailable.

In summary, our contributions are as follows: (I) Our work serves as the first attempt to perform semantic segmentation of football match scenes by transferring knowledge from SAM in a completely label-free setting. (II) We propose the SoccerSAM framework, which introduces a Semantic Bridging Module (SBM) to convert SAM's category-agnostic masks into semantic pseudo labels. By combining instance masks and semantic priors, SBM produces ensemble logits that guide the student model with high-quality supervision. We further incorporate a boundary-aware consistency loss to refine spatial predictions. (III) We demonstrate the effectiveness of SoccerSAM on football match segmentation tasks, showing that our framework achieves competitive performance without relying on any annotated data.

METHODOLOGY

2.1 Overview

Given an input football image $x \in \mathbb{R}^{H \times W \times 3}$, we first apply the frozen Segment Anything Model (SAM) to extract a set of high-quality instance masks $\{\mathcal{M}_i\}_{i=1}^{N}$ and their corresponding boundary map $\mathcal{B}_{SAM}$. These outputs capture rich structural cues but lack semantic class information. We then introduce a Semantic Bridging Module (SBM) to assign

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

class probabilities to each mask, generating a dense pseudo logit map as supervision. Finally, a compact student segmentation model is trained under the guidance of these logits and SAM's boundary priors. The overall framework is shown in Figure 1.

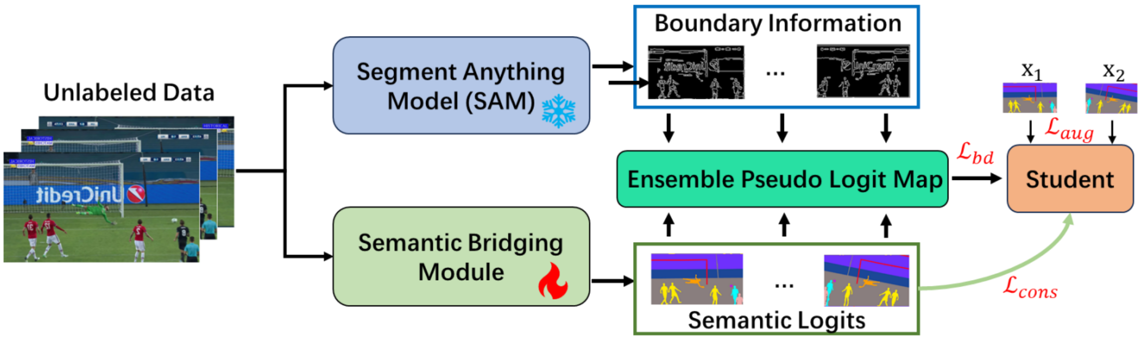

Figure 1. Illustration of SoccerSAM framework. The framework consists of two branches: the SAM branch that generates instance masks and boundary information, and the SBM branch that produces semantic logits. These outputs are fused to form the Ensemble Pseudo Logit Map, which serves as soft supervision for training the student segmentation model.

2.2 Semantic Bridging Module (SBM)

While SAM provides high-quality instance masks, it lacks semantic awareness and cannot produce class-specific predictions. To overcome this, we introduce a lightweight SemanticBridgingModule(SBM) that generates densesemanticlogits by combining SAM's instance masks with pretrained semantic priors. Given the instance masks $\{\mathcal{M}_i\}_{i=1}^{N}$ generated by the frozen SAM model and the image feature map $F \in \mathbb{R}^{H\times W\times D}$ extracted from a frozen backbone, SBM computes a region-wise semantic distribution for each mask. Specifically, for each $\mathcal{M}_i$, SBM performs masked average pooling to obtain a region embedding v_i:

$$v_i = \frac{1}{|\mathcal{M}_i|}\sum_{p\in\mathcal{M}_i} F\,(p) \qquad (1)$$

This embedding v_i is then passed through a lightweight classifier to produce the semantic logits of the mask, l.

Each l_i is broadcast to all pixels within its corresponding mask $\mathcal{M}_i$, resulting in per-mask semantic maps. To construct a dense semantic supervision signal, we aggregate all per-mask logits via pixel-wise averaging across overlapping masks:

$$\hat{y}(p) = \frac{1}{|\mathcal{J}p|}\sum i \in \mathcal{J}_p\, l_i,\ \text{where}\mathcal{J}_p = i \mid p \in \mathcal{M}_i \qquad (2)$$

The resulting $\hat{y} \in \mathbb{R}^{H\times W\times C}$ is referred to as the ensemble pseudo logit map, which serves as soft supervision for training the student segmentation model.

2.3 Boundary-Aware Consistency

While the ensemble pseudo logit map $\hat{y}$ provides semantic supervision, it may be less precise near object boundaries. To refine boundary localization, we introduce a Boundary − Aware Consistency Loss that aligns the structural edges of the student prediction with those implicitly embedded in SBM's output.

Specifically, we compute binary boundary maps from both the student's predicted logits P_S and the SBM-generated pseudo logits $\hat{y}$ using a gradient operator (e.g., Sobel) followed by thresholding. Denote these as $\mathcal{B}_S$ and $\mathcal{B}_{SBM}$, respectively.

We then define the boundary consistency loss as:

$$\mathcal{L}_{bd} = \frac{1}{|\mathcal{B}_{SBM}|}\sum_{p\in\mathcal{B}_{SBM}}|\mathcal{B}_S(p) - 1| \qquad (3)$$

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This loss encourages the student model to predict strong edges at the same locations as those derived from SBM's structure-aware logits, enhancing boundary precision without relying on ground-truth annotations.

2.4 Logit-Level Consistency

To further encourage semantic agreement, we introduce a two-fold logit-level consistency strategy that regularizes the student model from both external and internal perspectives.

$1)\mathrm{SBM}-\mathrm{to}-\mathrm{Student\ Consistency.}$

We use the dense pseudo logit map $\hat{y}$ produced by SBM to supervise the student's prediction P_S via pixel-wise soft supervision. We apply a pixel-wise Cross-Entropy loss between the two normalized distributions:

$$\mathcal{L}_{\mathrm{cons}} = \mathcal{L}_{\mathrm{CE}}(\hat{y}, P_S) \qquad (4)$$

This objective helps align the semantic distributions at each pixel and regularizes the student's output in ambiguous or uncertain regions.

$2)\mathrm{StudentSelf}-\mathrm{ConsistencyunderAugmentation.}$

To improve the student's robustness to input variations, we introduce a self-consistency constraint under data augmentation. Let $\mathcal{T}$ be a stochastic augmentation (e.g., flip, crop), and $\mathcal{T}^{-1}$ its spatial inverse. We enforce the outputs of x and its augmented view $x' = \mathcal{T}(x)$ to be consistent:

$$P_S^{(1)} = \mathrm{Student}(x_1),\ P_S^{(2)} = \mathrm{Student}(x_2) \qquad (5)$$

We enforce prediction consistency by minimizing the pixel-wise Cross-Entropy loss:

$$\mathcal{L}_{\mathrm{aug}} = \mathcal{L}_{\mathrm{CE}}\left(P_S^{(1)}, P_S^{(2)}\right) \qquad (6)$$

This loss encourages the student model to produce stable semantic predictions under stochastic image augmentations such as random crop, color jitter and flipping.

2.5 Training Objective

The student segmentation model is optimized by jointly minimizing three complementary losses: semantic alignment with SBM, boundary-level consistency, and augmentation-based prediction stability. The overall training objective is formulated as:

$$\mathcal{L}_{\mathrm{total}} = \lambda_{\mathrm{sbm}}\mathcal{L}_{\mathrm{sbm}} + \lambda_{\mathrm{bd}}\mathcal{L}_{\mathrm{bd}} + \lambda_{\mathrm{aug}}\mathcal{L}_{\mathrm{aug}} \qquad (7)$$

We empirically set $\lambda_{\mathrm{sbm}} = 0.3$, $\lambda_{\mathrm{bd}} = 1.0$, and $\lambda_{\mathrm{aug}} = 0.4$ based on hyperparameter analysis.

EXPERIMENTS

3.1 Datasets and Implementation Details

Datasets.The unlabeled training data and test data used in our experiments are both derived from the FSSOD[15] dataset, which consists of high-resolution football match images with pixel-level semantic annotations.

Implementation Details. We train the proposed SoccerSAM with PyTorch in 2 NVIDIA A6000 GPUs. We keep SAM frozen during and utilize it solely for providing boundary information and instance masks. The image size is resize to (416, 608) for training and validation. Our Semantic Bridging Module (SBM) adopts a lightweight architecture similar to MobileNet-UNet [6]. We employ two AdamW optimizers to update the parameters of the student segmentation network and the Semantic Bridging Module (SBM), respectively. The initial learning rate for both optimizers is set to 1×10^{-5}, with $\epsilon = 1 \times 10^{-8}$ and a weight decay of 5×10^{-4}. The metrics we used are MIoU and Frequency Weighted IoU (FWIoU), which are the popular metrics for segmentation task.

3.2 Results

Table 1 : Comparison of semantic segmentation methods on the FSSOD dataset. We report the performance using mean IoU (mIoU) and frequency-weighted IoU (FWIoU).

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Method	mIoU (%)	FWIoU (%)
FCN [11]	77.84	92.26
PSPNet [21]	58.46	83.96
SegNet [1]	69.7	90.51
VGG-UNet	83.64	95.02
MobileNet-UNet[6]	84.28	95.36
SoccerSAM (Ours)	**92.33**	**97.2**

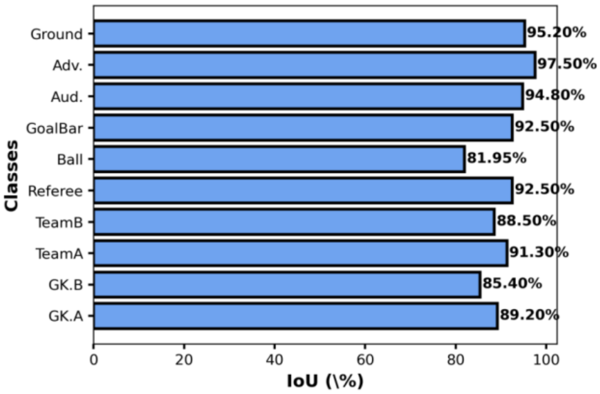

Figure 2. Class-wise IoU for SoccerSAM (Ours) on FSSOD Dataset.

We evaluate the proposed SoccerSAM framework on the FSSOD dataset and compare it with several widely-used semantic segmentation models, including FCN [11] , SegNet, PSPNet [21], VGG-UNet and MobileNet-UNet. Table 1 reports the performance in terms of both mean Intersection over Union (mIoU) and Frequency Weighted IoU (FWIoU). SoccerSAM achieves the best overall performance, significantly outperforming other methods. In addition, we evaluate the proposed SoccerSAM framework on the FSSOD dataset, as shown in Figure 2.

3.3 Ablation Study

We conduct an ablation study to assess the contribution of each component in the SoccerSAM framework. As shown in Table 2, removing the Semantic Bridging Module (SBM) loss leads to a decrease in both mIoU and FWIoU, demonstrating the importance of semantic supervision. Similarly, excluding the Boundary-Aware Consistency Loss results in a noticeable drop in FWIoU, indicating the benefit of boundary refinement for precise object segmentation. The removal of the Augmentation-Based Consistency Loss also leads to a slight reduction in performance, emphasizing the role of data augmentation in improving robustness. Finally, when all three losses are used, SoccerSAM achieves the best performance with a significant improvement in both mIoU (92.33%) and FWIoU (97.20%).

Table 2 : Ablation study results for SoccerSAM on the FSSOD dataset, evaluating the contribution of different loss components.

Method	mIoU (%)	FWIoU (%)
SoccerSAM w/o $\mathcal{L}_{\text{sbm}}$	88.37	93.54
SoccerSAM w/o $\mathcal{L}_{\text{bd}}$	87.42	94.8
SoccerSAM w/o $\mathcal{L}_{\text{aug}}$	89.2	95.4
SoccerSAM	**92.33**	**97.2**

3.4 Visualization

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We visualize the segmentation results of SoccerSAM on the FSSOD dataset to qualitatively assess its performance. Figure 3 shows the segmentation predictions for several football scenes. Our method not only successfully detects the players, ball, and goal bars but also captures the boundaries more accurately, especially in challenging scenes with densely packed objects.

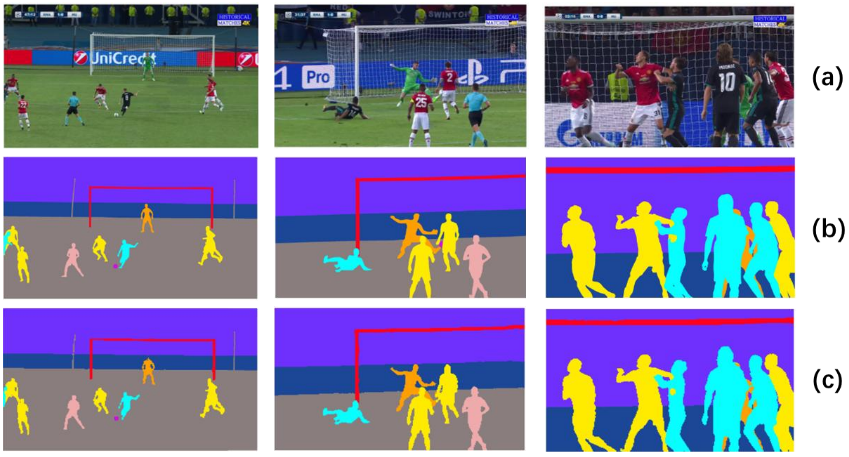

Figure 3. Example visualization results. (a) Input football match image, (b) Ground truth, (c) SoccerSAM.

CONCLUSION

In this paper, we introduced SoccerSAM, a novel framework for football scene segmentation that leverages the Segment Anything Model (SAM) and the Semantic Bridging Module (SBM) to generate soft semantic supervision without requiring labeled data. Extensive experiments on the FSSOD dataset demonstrate that SoccerSAM outperforms existing methods in terms of both mIoU and FWIoU, achieving significant improvements, especially in small object segmentation. Our approach provides a lightweight and effective solution for real-time football scene segmentation, and future work will explore extending the framework to handle more complex scenes and incorporate additional domain-specific knowledge.

REFRENCES

[1] Badrinarayanan, V., Handa, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv:1505.07293 (2015)

[2] Cioppa, A., Deliege, A., Van Droogenbroeck, M.: 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. pp. 1765–1774 (2018)

[3] Gough, C.: Market size of the sports analytics industry worldwide in 2020 and 2028 (2021 https://wwwstatistacom/statistics/1185536/sports-analytics-market size/#:~:text=The%20global%20sports%20analytics%20market,billion%20US%20dollars%20by%202028)

[4] Guo, Y., Liu, Y., Georgiou, T., Lew, M.S.: A review of semantic segmentation using deep neural networks. International journal of multimedia information retrieval 7, 87–93 (2018)

[5] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)

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

[7] Intelligence, M.: Sports analytics market – Growth, trends,COVID-19 impact, and forecasts (2022 - 2027) (2022 https://wwwmordorintelligencecom/industry-reports/sports-analytics-market)

[8] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L.,Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything.In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)

[9] Lab, D.S.A.: Why sports analytics (2019 https://dtaicskuleuvenbe/sports/)

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[10] Liu, R., Luo, T., Huang, S., Wu, Y., Jiang, Z., Zhang, H.: Crossmatch:Cross-view matching for semi-supervised remote sensing image segmentation. IEEE Transactions on Geoscience and Remote Sensing (2024)

[11] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3431–3440 (2015)

[12] Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications 15(1), 654 (2024)

[13] MH Nguyen, D., Nguyen, H., Diep, N., Pham, T.N., Cao, T., Nguyen, B.,Swoboda, P., Ho, N., Albarqouni, S., Xie, P., et al.: Lvm-med: Learning 10 xx.large-scale self-supervised vision models for medical imaging via secondorder graph matching. Advances in Neural Information Processing Systems 36, 27922–27950 (2023)

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

[15] Prime, S.R.: Football (semantic segmentation). https://www.kaggle.com/datasets/sadhliroomyprime/football-semantic-segmentation(2023), accessed: April 23, 2025

[16] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural informationprocessing systems 28 (2015)

[17] Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M.,Kiela, D.: Flava: A foundational language and vision alignment model. In:Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15638–15650 (2022)

[18] Strudel, R., Garcia, R., Laptev, I., Schmid, C.: Segmenter: Transformer for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7262–7272 (2021)

[19] Wu, Y., Zhao, W., Huang, C., Xi, Y., Li, Q., Wang, H.: Deep learning for semantic segmentation of football match image. In: 2023 2nd International Conference on Innovations and Development of Information Technologies and Robotics (IDITR). pp. 7–11. IEEE (2023)

[20] Zhang, Y., Shen, Z., Jiao, R.: Segment anything model for medical image segmentation: Current applications and future directions. Computers in Biology and Medicine p. 108238 (2024)

[21] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network.In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2881–2890 (2017)

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.