Object Detection and Description for Tennis Matches

Andrea Grandi 196222 — Daniele Vellani 196186 — David Wuttke 197156

Computer Vision and Cognitive Systems

Abstract. The accurate detection and description of players and balls in tennis match images is crucial for detailed match analysis. This project addresses the challenge of achieving high-quality detection of tennis players and balls to facilitate comprehensive analysis of tennis matches, including tracking in video data. Our approach leverages the YOLO (You Only Look Once) model for object detection and the BLIP model for generating natural language descriptions that capture the spatial relationships between detected objects on the court.

To enhance our analysis, we incorporated a tracking component using TrackNet, allowing us to extend the detection capabilities to video data and analyze continuous movements of players and balls. This holistic approach ensures that both static and dynamic aspects of tennis matches are effectively captured.

Our results demonstrate that the combined use of YOLO and BLIP models achieves remarkable accuracy and speed in detecting players and balls. The YOLO model's state-of-the-art performance in object detection, coupled with BLIP's ability to generate detailed spatial descriptions, provides a robust solution for tennis match analysis. Additionally, the integration of TrackNet for tracking significantly enhances the system's applicability to real-time and recorded videos, offering a comprehensive tool for tennis analytics.

1 Introduction

Tennis is a sport experiencing a surge in popularity, with a constantly growing number of players and spectators [1][2]. As with other televised sports like football, technology is playing an increasingly important role in tennis analysis. Spectators are demanding more visual insights into the game, while players and coaches seek advanced analytics to improve their strategies. In particular, the detection and tracking of the tennis ball are crucial for understanding game dynamics [3][4].

This paper proposes an approach for analyzing tennis matches using deep learning techniques. We employ a YOLO [5] model to detect both tennis players and the ball on the court. Additionally, we leverage the BLIP large language model (LLM) [6] to generate natural language descriptions of the tennis scene, including the spatial relationships between tennis players. To capture the court's

geometry, we utilize a Tracknet convolutional neural network [7] and a Modified ResNet50 [8] for keypoint detection. Furthermore, we introduce a tracking component that extends our solution to video data. This allows us to analyze the continuous movement of players and the ball, providing a comprehensive understanding of the match's flow and strategic decisions. By combining these techniques, our system offers a powerful tool for enhancing tennis analysis.

2 Related Work

Projects that work on tennis players and ball detection have already been implemented in various scientific papers. Older papers such as the one by Pingali et al. attempt to recognize tennis balls using only computer vision methods [9]. Reno et al. show in their paper how successful the use of machine learning algorithms, such as CNNs, is for recognizing tennis balls [4]. Deepa et al., for example, even compared different machine learning algorithms for their performance in recognizing and predicting the trajectory of tennis balls [10].

The YOLO model is one state-of-the-art model for object detection that outperforms comparable models like DPM and R-CNN [11] [12] [13]. The use of YOLO for recognizing players in a sport is already much used. Buric et al. show in their paper an approach to use a YOLO model to detect players in handball [14]. In a different paper, Nady et al. show a way to train a YOLO model to detect players in several sports, like basketball and hockey [15].

The BLIP model was first published in 2022 by Li et al and can be used as a vision-language model for image-text retrieval, visual-question-asking, and image-captioning tasks. Our approach only uses the image-captioning ability of BLIP. Other papers already used BLIP for image-caption generation. The paper of Chiang et al. for example, fine-tuned the BLIP LLM on mobile screenshot captioning [16]. Other LLMs such as VisualGPT [17], or CLIP [18] are often used for image captioning. However, the usage of BLIP proved to be the easiest way to implement our project. BLIP2, the follow-up to BLIP, is also used in many papers [19] [20]. Since this model is more complex and BLIP is sufficient for our use case, we use BLIP.

The TrackNet paper by Shu et al. demonstrates an application for tracking tennis balls and players across different frames in a video. The system uses deep learning to handle the fast movements and occlusions common in tennis, providing accurate tracking results.

We found no related paper describing tennis players' spatial positions and relationships on the tennis court.

3 Used Data

We used three different data sets to train the CNN for tennis court key point detection, the YOLO model for tennis player and tennis ball detection and the BLIP LLM for image captioning.

3.1 Data Set for YOLO Tennis Player and Tennis Ball Detection

To train the YOLO model on tennis player and tennis ball detection, we used a dataset from RoboFlow [21] [22]. The dataset for players detection consists of 2445 images, and the dataset for the ball detection is made of 578 images from approximately the same perspective of the tennis court. Each image contains only two tennis players and a tennis ball. The corresponding label shows where the players and the ball are located in the image. An example image for the dataset is shown in Fig. 1.

Fig. 1: Example Image from the players and ball detection dataset.

3.2 Data Set for CNN Key Point Detection

The dataset used for training the CNN consists of 8841 images, separated into a train set (75%) and a validation set (25%). Each image of the data set has 14 annotated key points. The resolution of images is 1280x720. This dataset contains all court types (hard, clay, grass). The data set can be accessed via Github [23]. An example image for the dataset is shown in Fig. 2.

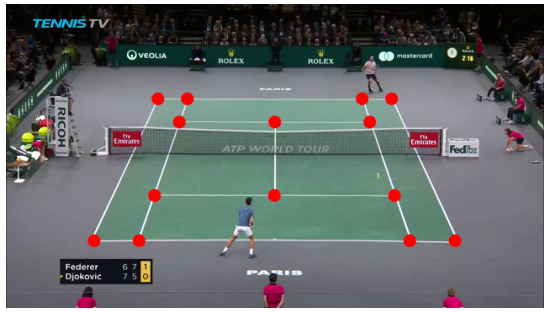

Fig. 2: Example Image from the tennis court key point detection data set.

3.3 Data Set for BLIP Image Captioning

The data set for fine-tuning BLIP includes 266 images with corresponding labels. The dataset is self-created and can be found under the following HuggingFace tag: DinoDave/SpatialRelationsTennis_masked. The images it contains are taken from various sources on the internet, such as YouTube, newspapers, or live streams. Each image shows only two tennis players, Player A and Player B, from approximately the same perspective. The dataset shows all common court types. The labels describe the spatial position of player A and player B. The players can either be at the net, at the baseline, at the T-zone, in the middle of the field, or outside the field. The images used have also been provided with bounding boxes by our own fine-tuned YOLO model. An example image for the dataset is shown in Fig. 3. Since all images in the data set have different sizes, they are resized to the same size in pre-processing.

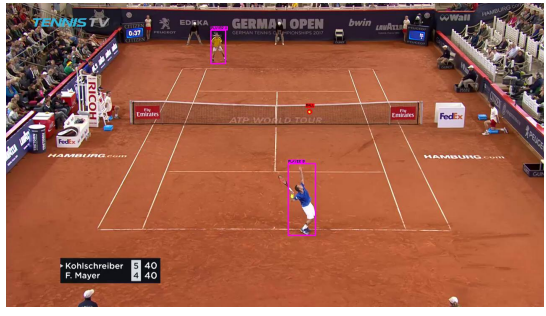

Fig. 3: Example Image from the DinoDave/SpatialRelationsTennis_masked dataset. Label: Player A is standing at the baseline and player B is standing at the baseline.

4 Methods

We developed a comprehensive system for tennis object detection and description, focusing on player and ball detection, spatial relationships, and tracking.

4.1 YOLO for Object Detection

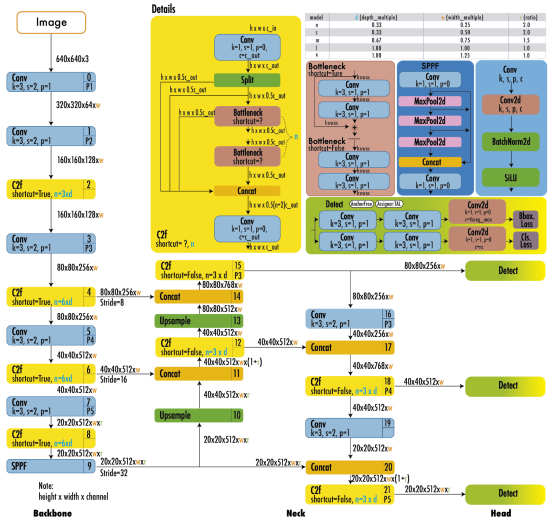

Fig. 4: Architecture of YOLOv8. The architecture uses a modified CSPDarknet53 backbone. The C2f module replaces the CSPLayer used in YOLOv5. A spatial pyramid pooling fast (SPPF) layer accelerates computation by pooling features into a fixed-size map. Each convolution has batch normalization and SiLU activation. The head is decoupled to process objectness, classification, and regression tasks independently.

Due to its exceptional speed and accuracy, YOLOv8 [24] is a powerful tool for detecting tennis players and the ball, ensuring that we can capture subjects effectively. This architecture also integrates anchor-free detection, reducing the complexity and computational load by eliminating the need for predefined anchor boxes. The model's performance is further optimized by the use of advanced training strategies such as mosaic data augmentation, which combines multiple training images to improve robustness and generalization. These technical innovations collectively contribute to superior performance in real-time object detection tasks, making it particularly effective for dynamic and fast-paced environments like tennis. We initially experimented with Faster R-CNN [25] and other versions of YOLO for object detection, in particular YOLOv5 [26], known for its simplicity and speed, making it a popular choice for real-time applications, and YOLOv7 [27]. However, YOLOv8 performed significantly better in

terms of both accuracy and processing speed. In Table 1, we present a detailed comparison between YOLOv8 and Faster R-CNN in terms of mAP@50 and GPU Latency.

Table 1: Comparison of YOLOv8 and Faster R-CNN performance.

Algorithm	mAP@50	GPU Latency (ms)
YOLOv8	0.62	1.3
Faster R-CNN	0.41	54

In Figure 5, the performance of various YOLO versions is analyzed based on COCO mAP and latency metrics. It is evident that YOLOv8 exhibits superior performance compared to YOLOv7, YOLOv6, and YOLOv5 across both metrics. This underscores YOLOv8's effectiveness in balancing accuracy and speed, making it a top choice for object detection applications.

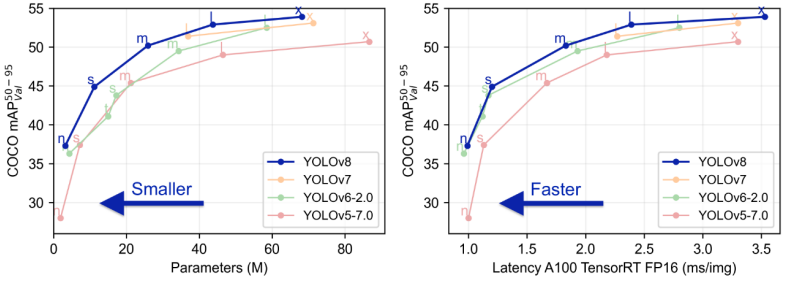

Fig. 5: Comparison of different YOLO versions in terms of COCO mAP and latency. The left plot shows the mAP against the number of parameters, and the right plot shows the mAP against the latency.

4.2 Image Processing and Perspective Distortion

Perspective Distortion: To address perspective differences between input images, we incorporated an image perspective distortion component (Homography), ensuring that all input images were aligned to the same perspective. This step helped maintain consistency across different images, allowing for more accurate detection.

Image Processing for Keypoints Detection: In our first attempt at recognizing the playing field, we used the Scale-Invariant Feature Transform (SIFT) [28] for keypoint detection. This procedure started with an image processing operator

that applied a morphological dilation operator to amplify edge and line characteristics. It helped in refining the detected features thus improving the overall quality and accuracy of the process of detecting keypoints through alignment. Then we employed Canny edge detector [29] to determine the edges in the image. The next step involved using Hough Transform [30] for recognizing court lines. Finally, SIFT was applied to detect keypoints filtered out from those that did not lie on lines which were detected by Hough Transform.

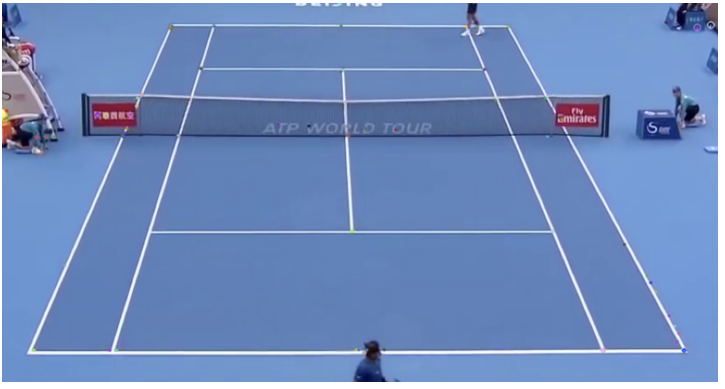

Fig. 6: Keypoints extraction using SIFT

This method aimed to ensure that the keypoints used for analysis were relevant to the structure of the court, enhancing the accuracy of our detection system. However, our ResNet50 and TrackNet provided superior results compared to our initial SIFT-based approach, as we will see in the next chapter.

4.3 Tracking and Keypoints Extraction

For tracking, we employed TrackNet, which is specifically designed for sports tracking applications.

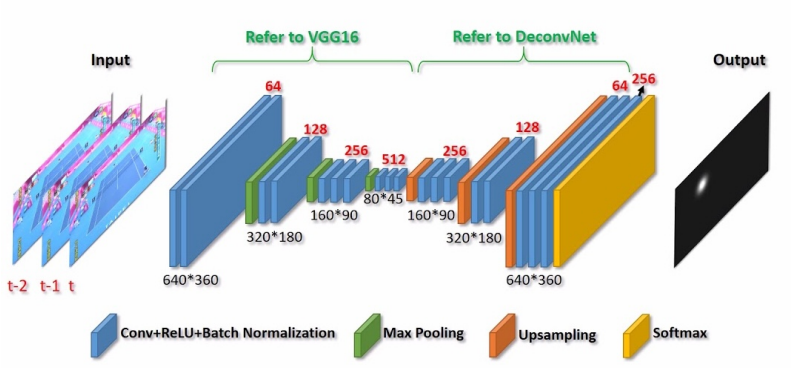

Fig. 7: Architecture of TrackNet. The network structure includes convolutional layers with ReLU and Batch Normalization, max pooling layers, upsampling layers, and a final softmax layer. The input consists of three consecutive frames, and the output is a heatmap indicating the position of the ball.

TrackNet's robustness in handling occlusions and fast movements ensures reliable tracking of both players and the ball throughout the match. This capability is crucial for maintaining accurate tracking data even when players momentarily block the view of the ball or each other, a common occurrence in tennis. In our project, we incorporated the tracking capabilities of this neural network to extend our analysis to video data. By leveraging TrackNet's tracking features, we were able to analyze match videos to provide a more comprehensive understanding of player movements and ball trajectories.

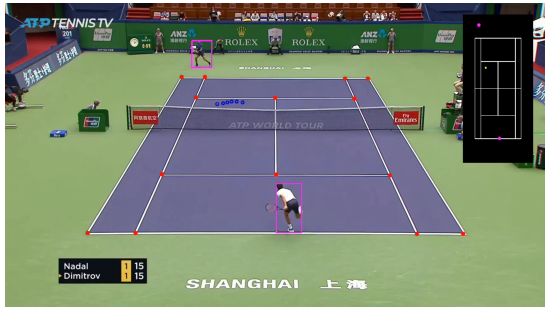

Fig. 8: Tracking frame example

In our work, we have employed both TrackNet and a modified ResNet50 architecture for keypoints detection. The modified ResNet50 includes channel and spatial attention mechanisms to enhance the network's focus on informative features and regions, followed by a linear layer to output the 14 pairs of keypoints.

CNN Design: For improve the keypoints detection capabilities, we have integrated channel and spatial attention mechanisms into a modified ResNet50 architecture. The channel attention mechanism refine the network's ability to focus on informative features while suppressing irrelevant ones. Specifically, it recalibrates feature maps by learning attention weights that emphasize important channels. It exploits inter-channel relationships to generate a channel attention map, which directs the network to important features in the input image. The process begins by squeezing the spatial dimensions of the feature map to aggregate spatial information using average-pooling and max-pooling operations, producing descriptors z_{avg} and z_{max} respectively. These descriptors are then fed into a shared network, typically a multi-layer perceptron (MLP) with one hidden layer, designed to compute the channel attention map A_c. To reduce parameter overhead, the size of the hidden activation F is set based on a reduction ratio r, where $F = \frac{C}{r}$, with C being the number of channels. The calculation of channel attention A_c is formulated as follows:

$$\mathbf{M_c}(\mathbf{F}) = \sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool}(\mathbf{F})\right) + \mathrm{MLP}\left(\mathrm{MaxPool}(\mathbf{F})\right)\right) \tag{1}$$

Or equivalently,

$$\mathbf{M_c}(\mathbf{F}) = \sigma\left(\mathbf{W_1}\left(\mathbf{W_0}\left(\mathbf{F}^c_{avg}\right)\right) + \mathbf{W_1}\left(\mathbf{W_0}\left(\mathbf{F}^c_{max}\right)\right)\right) \tag{2}$$

where $\mathbf{F}^c_{avg}$ and $\mathbf{F}^c_{max}$ denote the average-pooled and max-pooled features across channels, respectively.

Here:

- σ denotes the sigmoid function,
- $\mathbf{W_0}$ and $\mathbf{W_1}$ are weights applied to $\mathbf{F}^c_{avg}$ and $\mathbf{F}^c_{max}$ respectively,
- MLP refers to the multi-layer perceptron,
- AvgPool and MaxPool represent average pooling and max pooling operations on $\mathbf{F}$.

The channel attention map $\mathbf{M_c}(\mathbf{F})$ represents the attention weights, where higher values indicate channels that are more relevant for the task at hand, facilitating improved feature representation.

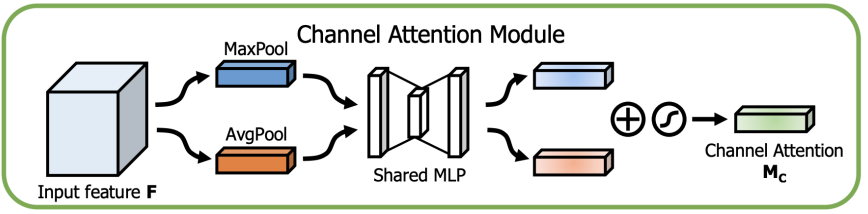

Fig. 9: Channel Attention Module

In addition to the channel attention mechanism, a spatial attention mechanism is employed to focus on informative regions within the feature maps. The spatial attention mechanism generates an attention map by applying average-pooling and max-pooling operations along the channel axis, followed by a convolutional layer to aggregate spatial information.

Finally, a linear layer is added to the end of the ResNet50 network to output the 14 pairs of keypoints.

4.4 Spatial Relationships

Understanding the player and ball's position in relation to each other is important for a complete analysis of tennis match images. This includes detecting the objects (players and ball) as well as understanding how they interact with each other and their positions within the scene. Ours is an approach which utilizes the keypoint extraction model, and BLIP (Bootstrapping Language-Image Pre-training), a natural language generation model in improving the system interpretability to such spatial dynamics.

BLIP Model for Descriptive Understanding: To make sense of this spatial data we incorporate the BLIP model. In reality it is created so that it can give rich context aware descriptions of scenes from visual perspective only.

Detailed Analysis of Spatial Relationships:

1. **Relative Positioning**: The system assesses the relative positions of players and the ball. For example, it can describe scenarios where a player is positioned at the baseline, approaching the net, or moving laterally to intercept the ball.
2. **Movement and Interaction**: BLIP's descriptions capture the dynamics of the match, detailing actions such as a player running to hit the ball, jumping for a smash, or reaching out for a volley.
3. **Proximity and Engagement Zones**: The system identifies zones of engagement, where the action is most intense. It can describe how close players are to each other and to the ball, and identify zones such as the baseline,

service box, and net area. This helps in understanding strategic positions and potential areas of advantage.

Integration and Outputs: The integration of BLIP results in a sophisticated output that combines visual data with natural language processing to offer comprehensive scene descriptions.

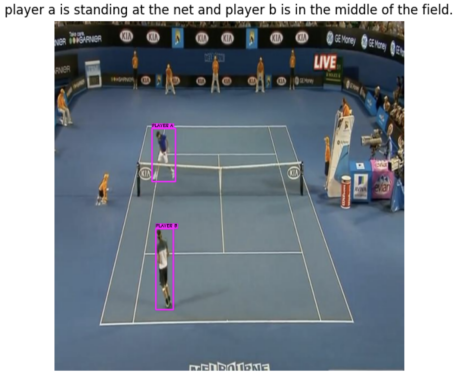

Fig. 10: Output Example

In summary, BLIP model generate descriptions that allows our system to understand and articulate the spatial relationships in tennis match images.

4.5 Text-to-Image Retrieval

Image retrieval consists in techniques used for finding some specific images among a dataset, by using keywords, titles or descriptions; in this case we needed to use full sentences because we wanted to search for particular situations in tennis matches. All the retrieval part was done by using the result of the BLIP model. The program will ask the user for a query, a situation to look for. Then, while the BLIP model is testing, it will also check if the query and the caption of the image are similar, using the sentence transformer. We used Sentence Trasformer (a.k.a. SBERT) [31] that is a Python module to compute embeddings or to calculate similarity scores using Cross-Encoder models. If the transformer find a similarity higher than 0.8, then it will take the index of that particular image and it will be saved in a list of indexes. As a reference value we used 0.8 because we needed to find the most similar sentences among all, but not too similar: actually if we set the value higher we will be looking for equal phrases, and we saw that the BLIP model doesn't always give the same description for images that can be described in the same way.

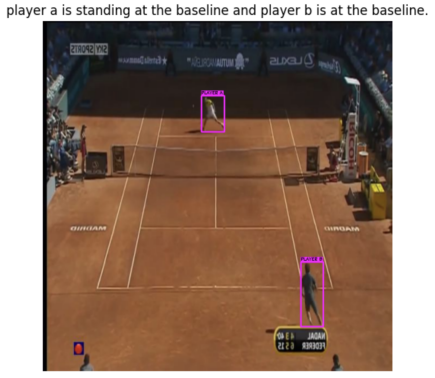

Fig. 11: Retrieval Output Example with Query: "Players are both staying on the baseline"

For example, if there are two player standing at the baseline the image can be described differently:

- '*player a is standing at the baseline and player b is standing at the baseline.*'
- '*both player a and b are standing at the baseline of the tennis field.*'

5 Experiments

To demonstrate the efficacy of our approach in solving the problem of tennis player and ball detection, keypoints detection and spatial relationships, we conducted several experiments. These experiments compared our method with previously published methods, evaluated the impact of various components through ablation studies, experimented with different hyperparameters and architectural choices, and utilized visualization techniques to gain insights into model performance. Common failure modes were also analyzed to understand the limitations of our approach. Below, we present detailed results and analyses from our experiments.

5.1 Performance of the YOLO Model for Tennis Player and Ball Detection

The model has shown excellent performance in detecting players and ball. Figure 12 shows the confusion matrix for players detections, illustrating high accuracy in distinguishing between different classes.

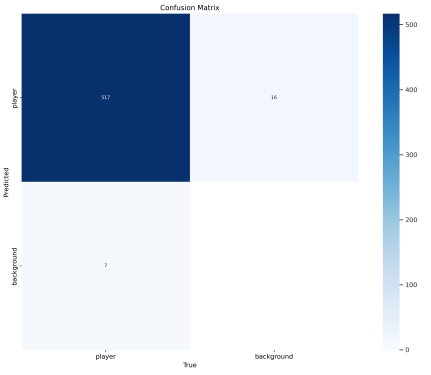

Fig. 12: Confusion matrix illustrating the performance of the YOLOv8 model in detecting players

To further evaluate the model's effectiveness, we also examined the F1-score, which provides a balanced measure of precision and recall. Figures 13 and 14 presents the F1-score, demonstrating that YOLOv8 maintains high precision and recall. The consistently high F1-scores underscore the model's robustness and reliability in player detection tasks.

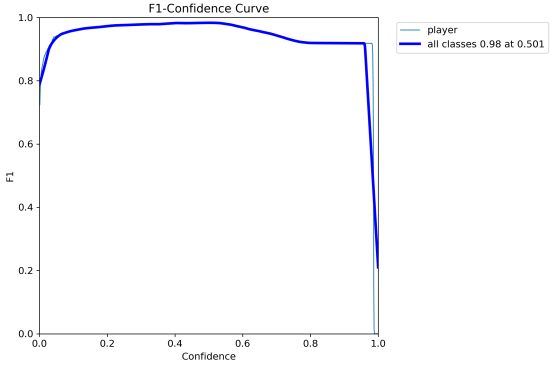

Fig. 13: F1-Score for Players Detections

Additionally, Figure 16 and Figure 15 displays sample results of the model's predictions on a batch of images. The detections are precise, showcasing the model's effectiveness including different player positions and ball positions. Importantly, we also trained YOLOv8 for tennis ball detection using a dedicated dataset. This approach optimized the model to accurately and reliably identify tennis balls during matches.

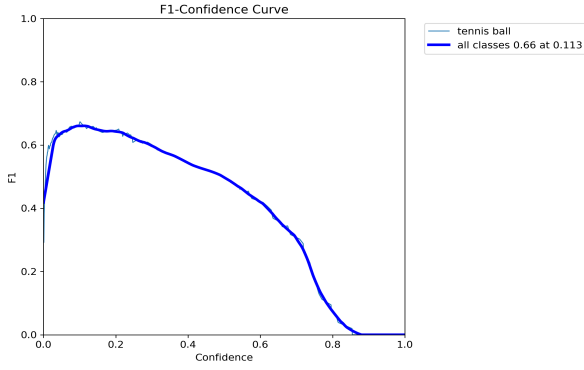

Fig. 14: F1-Score for Ball Predictions

Moreover, the processing speed of our model ensures it can be deployed in real-time tennis match analysis without significant latency. The balance between accuracy and speed is critical for live sports analysis, where timely and accurate data are essential.

Fig. 15: Examples of Players Detection

Fig. 16: Examples of Ball Detections

Overall, YOLOv8's performance, as evidenced by the confusion matrix, high F1-scores, and precise detections, makes it highly suitable for object detection applications in tennis match analysis, providing reliable and quick insights into players and ball positions.

5.2 Performance of ResNet50 for Keypoints Detection

(a) Keypoints with TrackNet (b) Keypoints with ResNet

Fig. 17: Comparison between TrackNet and Modified ResNet50 for Keypoints Detection

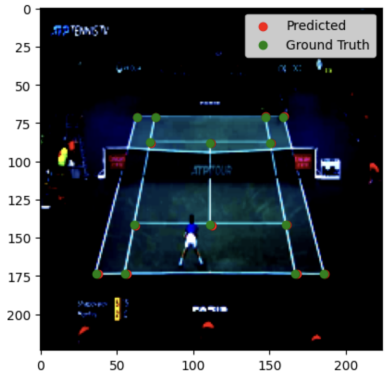

Fig. 18: Experimental Result

The modified ResNet50, enhanced with an attention channel, achieved comparable results to TrackNet in keypoint detection. The attention mechanism allowed the model to focus on the most relevant parts of the image, leading to more accurate keypoint localization. Figure 17 illustrates this comparison, showing that the ResNet50 with the attention block effectively identifies keypoints similarly to TrackNet. Moreover, with Figure 18 demonstrates the performance of the network by comparing the predicted keypoints with the ground truth. This improvement underscores the potential of integrating attention mechanisms for keypoints detection tasks.

5.3 Performance of the BLIP LLM for Image Captioning

To determine the performance of the fine-tuned BLIP LLM, we calculated the common performance measures. These are BLEU, METEOR, ROUGE, CIDEr and SPICE. We calculated the values of these individual measures using the test data set. The results are shown in table 2. The figure shows that the model performs best when trained with a learning rate of 3e-10.

Table 2: Performance of the BLIP LLM for different learning rates.

Learning Rate	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr	SPICE
3e-10	0.7201	0.6618	0.6012	0.5384	0.5832	0.7226	3.0161	0.3628
4e-10	0.6697	0.5808	0.4934	0.4166	0.5140	0.6671	1.7080	0.3191
5e-10	0.6577	0.5672	0.4799	0.4057	0.4929	0.6594	1.4662	0.3046

A BLEU-1 score of 0.7201 indicates a high unigram precision, meaning there is a strong overlap of individual words between the generated captions and the reference captions. A BLEU-2 score, which measures bigram precision, is also high at

66.18%, showing good overlap for pairs of consecutive words. With the BLEU-3 score of 60.12%, the model still maintains good overlap for sequences of three consecutive words. The BLEU-4 score of 53.84% reflects a strong performance in generating sequences of four consecutive words, indicating good fluency and context in the captions. A METEOR score of 58.32% is quite high and indicates good alignment between the generated captions and the reference captions, taking into account precision, recall, and synonyms. This score suggests that the captions not only contain the right words but are also semantically similar to the reference captions. A ROUGE-L score of 72.26% indicates a high degree of overlap in the longest common subsequence between the generated and reference captions. This suggests that the model is good at generating captions that are structurally similar to the reference captions. A CIDEr score of 3.0161 is very high and indicates strong consensus with multiple reference captions. This metric weights the importance of n-grams based on their occurrence across different references, and such a high score suggests that the generated captions are highly relevant and well-aligned with the references. A SPICE score of 36.28% is considered quite good. This metric evaluates the semantic content of the captions by comparing scene graphs, and a score in this range indicates that the generated captions capture the key objects, attributes, and relationships present in the reference captions.

The model performs very well across all metrics, with particularly high scores in BLEU, METEOR, and CIDEr. This suggests that the generated captions are both precise and contextually relevant, closely matching the reference captions in terms of both word overlap and semantic content.

6 Conclusions

In conclusion, our approach demonstrably achieves high-quality object detection and description of tennis matches. The YOLO model's accuracy, coupled with TrackNet's tracking capabilities, and our custom ResNet50, effectively captures both static and dynamic aspects of tennis matches. Furthermore, BLIP's ability to generate descriptive captions enriches the analysis with semantic information about the spatial relationships between objects. Given these results for the BLIP LLM, our model demonstrates strong performance in image captioning. It generates captions that are both accurate and semantically rich. To further improve, we might focus on enhancing the diversity and depth of the generated captions. To implement this, we need to continue working on the data set that was used for fine-tuning. With 266 images, this data set tends to be small. We can also improve the depiction of spatial relationships by enhancing the representation of the tennis court, also including ball speed and players relative speed movements.

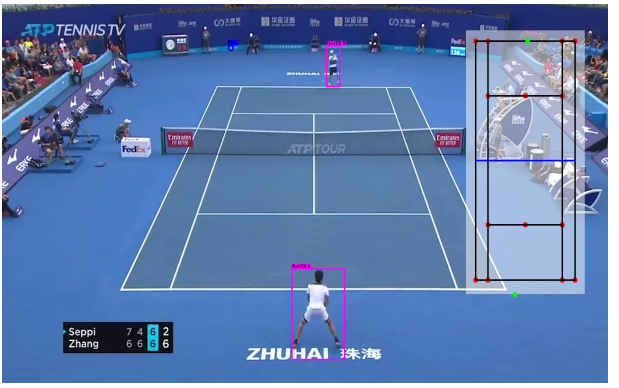

Fig. 19: On the right we can see the players relative positions inside the court

This would lead to a more precise analysis of object positions and movements. Moreover, enhancing the retrieval process would enable more efficient and accurate searches for tennis match images, thereby supporting better data acquisition and system performance.

References

1. WashingtonPost: Opinion — pickleball might be the 'in' sport now, but tennis is still strong - the washington post. https://www.washingtonpost.com/opinions/2023/09/22/pickleball-tennis-popularity/ (2024) (Accessed on 05/03/2024).
2. SueddeutscheZeitung: Sport im landkreis - tennis ist wieder im trend - landkreis münchen - sz.de. https://www.sueddeutsche.de/muenchen/landkreismuenchen/sport-im-landkreis-tennis-ist-wieder-im-trend-1.4209690 (2024) (Accessed on 05/05/2024).
3. Kamble, P.R., Keskar, A.G., Bhurchandi, K.M.: Ball tracking in sports: a survey. Artificial Intelligence Review **52** (2019) 1655–1705
4. Reno, V., Mosca, N., Marani, R., Nitti, M., D'Orazio, T., Stella, E.: Convolutional neural networks based ball detection in tennis games. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. (2018) 1758–1764
5. Varghese, R., M., S.: Yolov8: A novel object detection algorithm with enhanced performance and robustness. In: 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). (2024) 1–6
6. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S., eds.: Proceedings of the 39th International Conference on Machine Learning. Volume 162 of Proceedings of Machine Learning Research., PMLR (17–23 Jul 2022) 12888–12900
7. Huang, Y.C., Liao, I.N., Chen, C.H., İk, T.U., Peng, W.C.: Tracknet: A deep learning network for tracking high-speed and tiny objects in sports applications. In: 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). (2019) 1–8
8. Mo, C., Lin, Q., Pan, L., Wu, S.: Improved resnet-50 based low altitude, small rcs and slow speed target recognition. In: 2023 3rd International Conference on Electronic Information Engineering and Computer Science (EIECS). (2023) 698–704
9. Pingali, G., Opalach, A., Jean, Y.: Ball tracking and virtual replays for innovative tennis broadcasts. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000. Volume 4., IEEE (2000) 152–156
10. Deepa, R., Tamilselvan, E., Abrar, E., Sampath, S.: Comparison of yolo, ssd, faster rcnn for real time tennis ball tracking for action decision networks. In: 2019 International conference on advances in computing and communication engineering (ICACCE), IEEE (2019) 1–4
11. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (June 2016)
12. Zou, Z., Chen, K., Shi, Z., Guo, Y., Ye, J.: Object detection in 20 years: A survey. Proceedings of the IEEE **111**(3) (2023) 257–276
13. Vijayakumar, A., Vairavasundaram, S.: Yolo-based object detection models: A review and its applications. Multimedia Tools and Applications (2024) 1–40
14. Burić, M., Pobar, M., Ivašić-Kos, M.: Adapting yolo network for ball and player detection. In: Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods. Volume 1. (2019) 845–851
15. Nady, A., Hemayed, E.E.: Player identification in different sports. In: VISIGRAPP (5: VISAPP). (2021) 653–660

16. Chiang, C.Y., Chang, I.H., Liao, S.W.: Blip-adapter: Parameter-efficient transfer learning for mobile screenshot captioning. arXiv preprint arXiv:2309.14774 (2023)
17. Chen, J., Guo, H., Yi, K., Li, B., Elhoseiny, M.: Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2022) 18030–18040
18. Mokady, R., Hertz, A., Bermano, A.H.: Clipcap: Clip prefix for image captioning. arXiv preprint arXiv:2111.09734 (2021)
19. Zhu, D., Chen, J., Haydarov, K., Shen, X., Zhang, W., Elhoseiny, M.: Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. arXiv preprint arXiv:2303.06594 (2023)
20. Nguyen, T., Gadre, S.Y., Ilharco, G., Oh, S., Schmidt, L.: Improving multimodal datasets with image captioning. Advances in Neural Information Processing Systems **36** (2024)
21. Dhanwani, V.: tennis ball detection dataset. https://universe.roboflow.com/virendhanwani/tennis-ball-detection (feb 2023) visited on 2024-06-11.
22. University: Player detection for tcmr dataset. https://universe.roboflow.com/university-zjdmp/player-detection-for-tcmr (aug 2023) visited on 2024-06-21.
23. yastrebksv: tennis court detection dataset. https://github.com/yastrebksv/TennisCourtDetector (2024) visited on 2024-06-11.
24. Terven, J.R., Esparza, D.M.C., Romero-González, J.A.: A comprehensive review of yolo architectures in computer vision: From yolov1 to yolov8 and yolo-nas. Mach. Learn. Knowl. Extr. **5** (2023) 1680–1716
25. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
26. Karthi, M., Muthulakshmi, V., Priscilla, R., Praveen, P., Vanisri, K.: Evolution of yolo-v5 algorithm for object detection: Automated detection of library books and performace validation of dataset. In: 2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES). (2021) 1–6
27. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.M.: Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2023) 7464–7475
28. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision (2004)
29. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-8**(6) (1986) 679–698
30. Illingworth, J., Kittler, J.: A survey of the hough transform. Computer Vision, Graphics, and Image Processing **44**(1) (1988) 87–116
31. Wang, B., Kuo, C.C.J.: Sbert-wk: A sentence embedding method by dissecting bert-based word models. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28** (2020) 2146–2157