

 Future Processing

# Ball tracking and detection in soccer videos

*Comparison of VLMs and traditional pipelines*

**Maciej Szymkowski, Ph.D., Eng.**

*Paris, 08 April 2026  
PyTorch Conference Europe 2026*

[www.future-processing.com](http://www.future-processing.com)




# Agenda

---

- Database used in the experiments
- AI-based specialized pipeline: ball detection with YOLOv11
- Vision-Language Models: efficient solution for object detection?
- Conclusions and future work

# Database used in the experiments

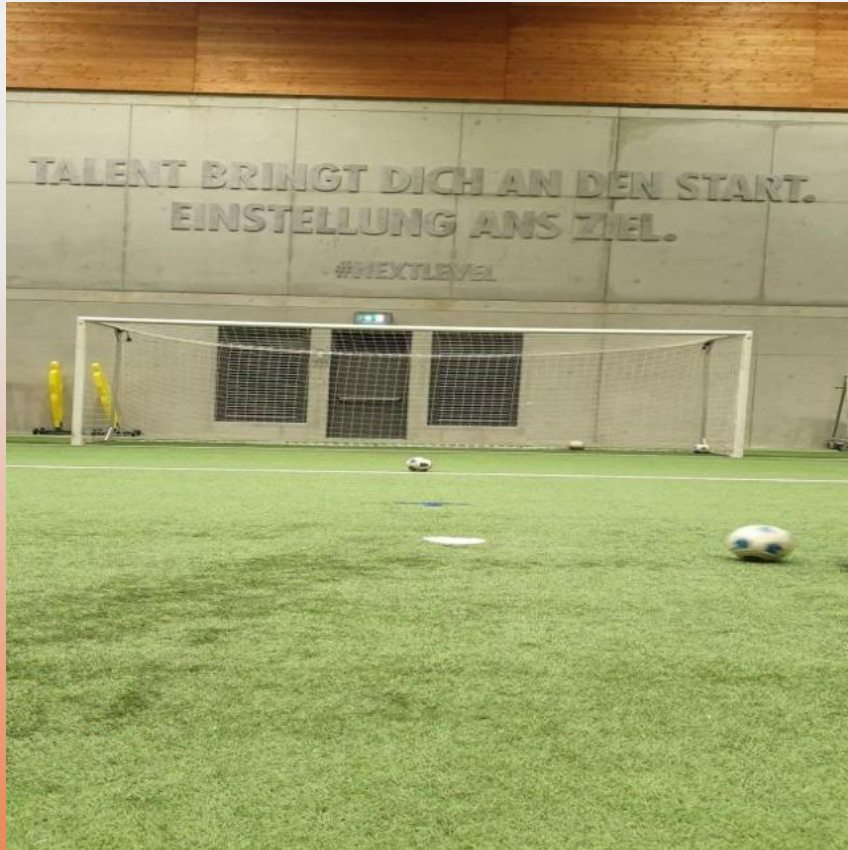


# Database

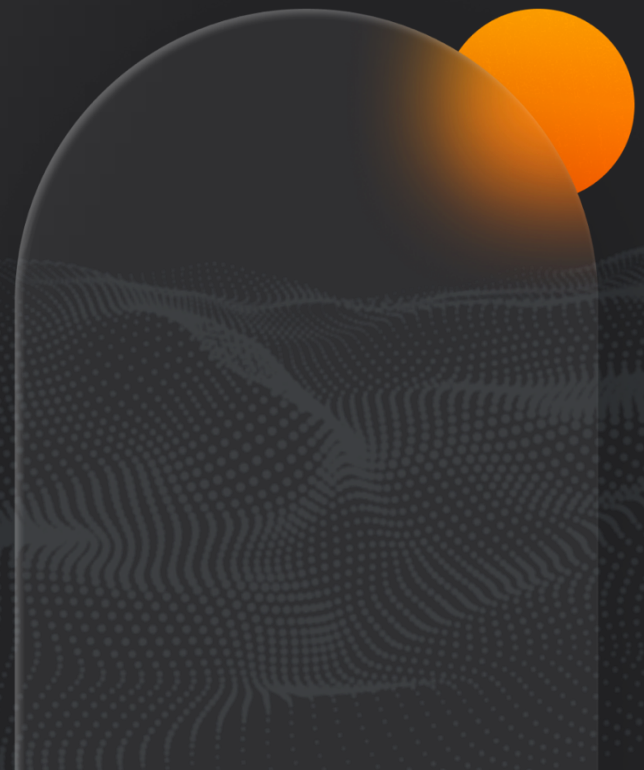
---

- For all performed experiments, we used the dataset called “ball\_detection Computer Vision Model” that is available on Roboflow platform (link: [https://universe.roboflow.com/tezeract/ball\\_detection-e6g8x](https://universe.roboflow.com/tezeract/ball_detection-e6g8x))
- Images are showing a ball in diversified scenarios (there are sometimes more than one ball in the picture) – e.g., during training (when someone heads the ball or taking a free kick).
- All samples have their ground truth versions – it means that in all the cases position of the ball(s) has been marked by hand.
- The dataset consists of *three* subsets:
  - **Training** – 11848 images
  - **Testing** - 44 images
  - **Validation** – 41 images

# Database (*examples*)



# Ball detection with YOLOv11

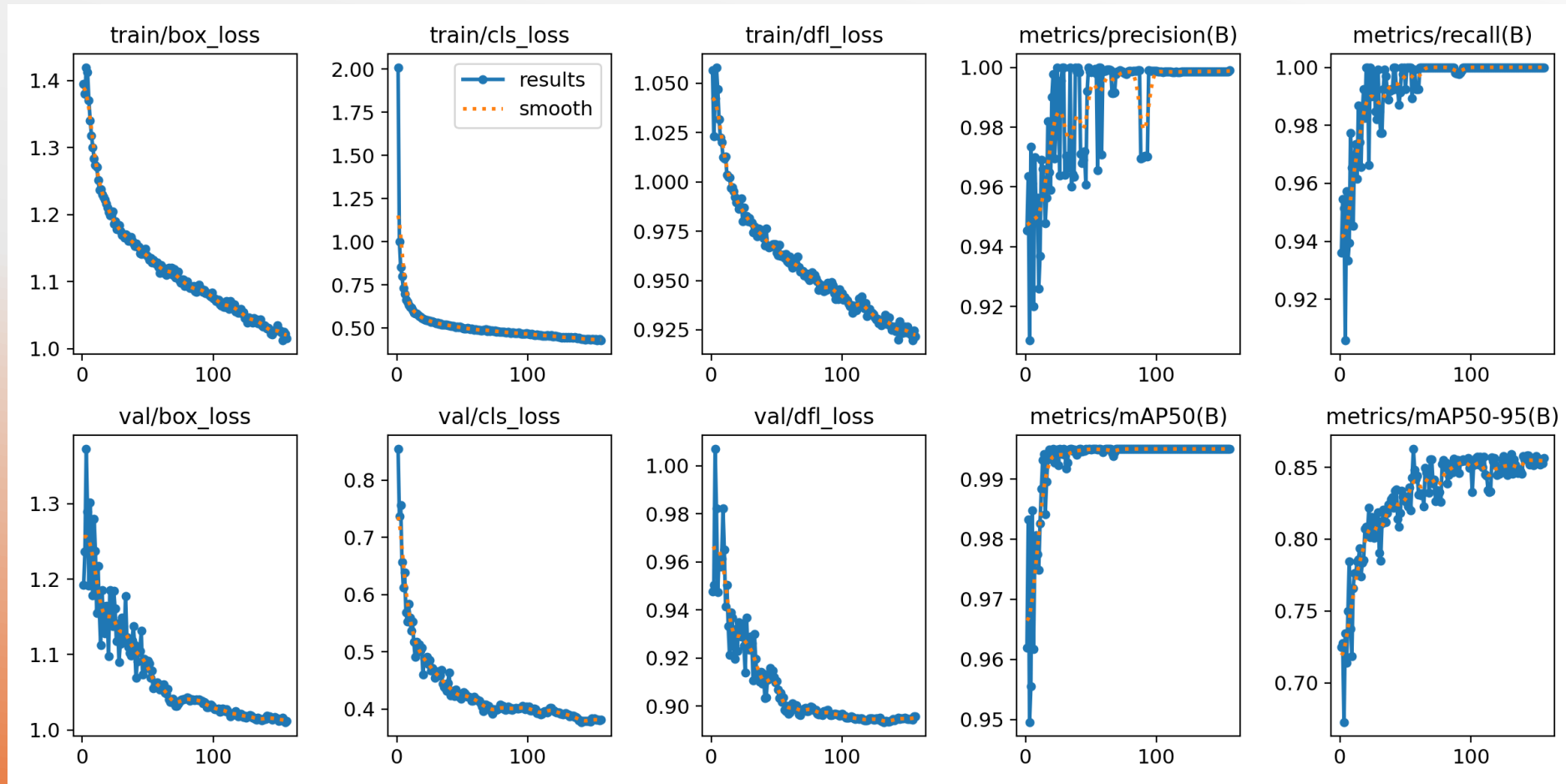


# Ball detection with YOLOv11

---

- In the first part of the experiments, we fine-tuned YOLOv11 model (with subset called “training” – mentioned previously). It was selected due to its high efficiency and accuracy in object detection and recognition tasks.
- It also must be pointed out that fine-tuning of this model is not resource-heavy (it does not require huge computing power, we used gaming PC with single Nvidia RTX 4090 graphics card). Inference of the model can also be done within low resources environment (it is possible to run the model even with single-board computers like Raspberry Pi).
- Fine-tuning process was set on 250 epochs but was finished after 155 (early stopping when the changes are small took place).

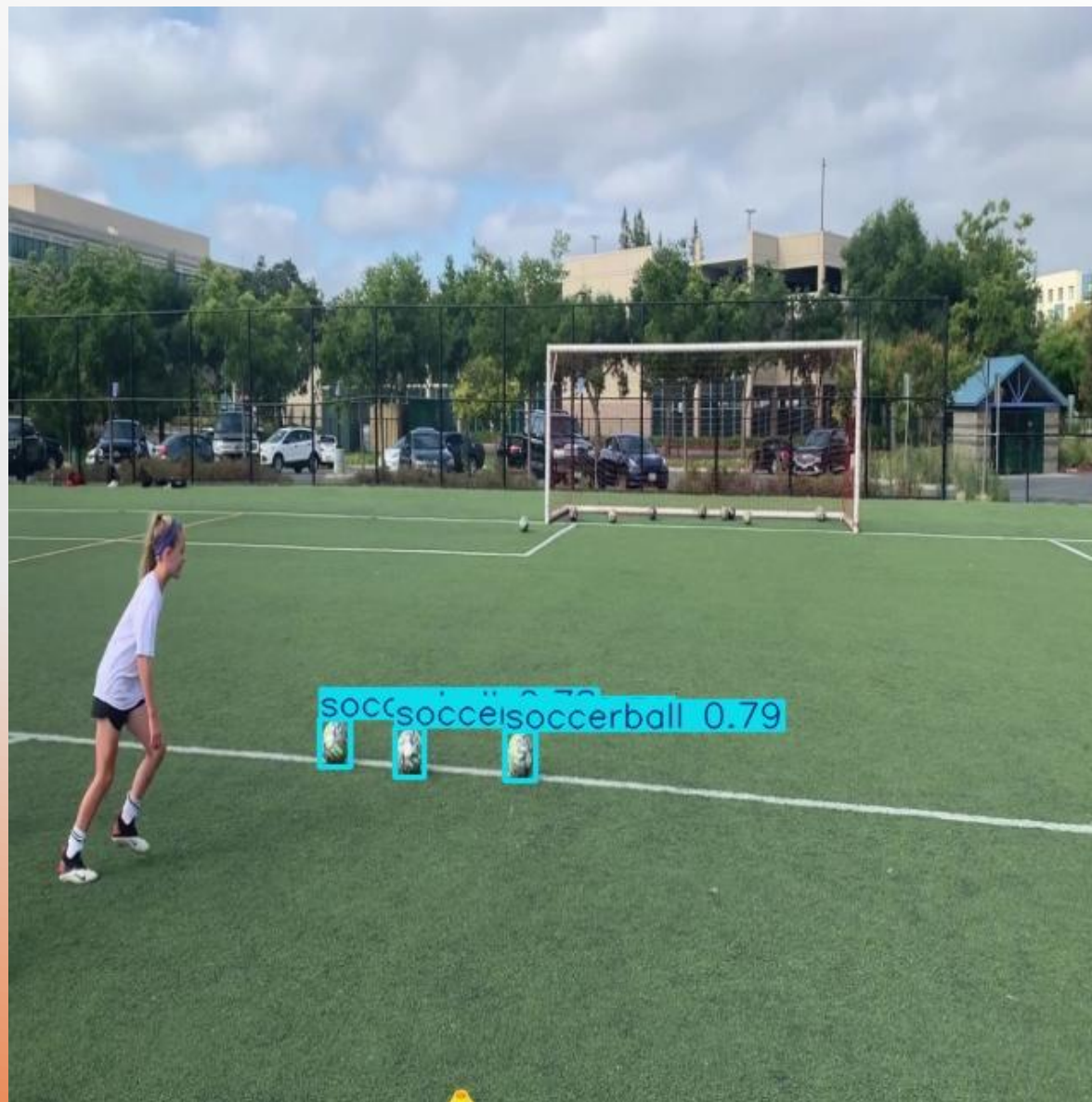
# Results of fine-tuning process



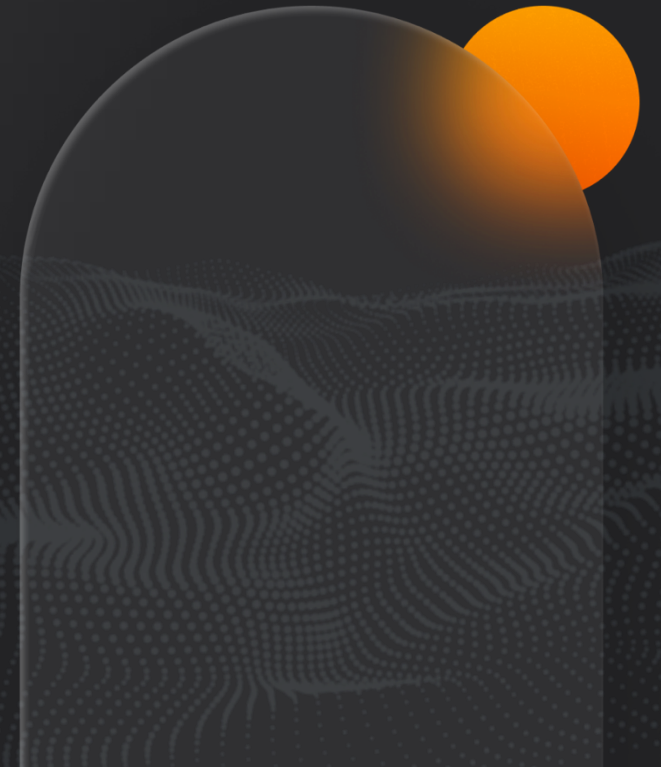
# Results on validation set



# But...



Vision-Language  
Models: efficient  
solution for object  
detection?



# Vision-Language Models (VLMs)

---

- Nowadays, one of the trends in Computer Vision is to use Vision-Language Models (VLMs) to detect specific objects or structures in the image (also in medical samples).
- VLMs can provide satisfactory results (even within complex environments) but their usage can have some bottlenecks:
  - Resources – high VRAM demand
  - Inference time – can be much longer than in the case of specialized pipeline
  - SBC – it is hard to run huge models on these devices
- In that part of the experiments, we gave a chance to three VLMs: Gemini 3, Qwen-3 and InternVL-3.5. Two of them were prepared as on-premise models.

# Prompt used in all models

---

*You are a specialist in analysis of soccer images. Your task is to mark all balls in the image - there can be more than one ball. As the result of your work, return information how many balls are in the picture, their positions and your certainty about the detection result (for each of them).*

# Gemini 3 (thinking)



frame4fr.jpg	11	1. [650, 283, 693, 309] (Line 1)	100%
		2. [656, 351, 700, 379] (Line 2)	100%
		3. [658, 452, 705, 482] (Line 3)	90%
		4-11. Balls lined up on goal line	

# Qwen-3-8B (thinking)



There are 3 balls visible in the image.

1. **Ball 1**: Located in the foreground, slightly to the left of the center of the image. It is also on the white line.
2. **Ball 2**: Located in the middle of the foreground, to the right of Ball 1. It is also on the white line.
3. **Ball 3**: Located further back, to the right of Ball 2, also on the white line. I am certain about its detection.

There is also a fourth ball visible in the image, which is much further in the background, near the goal.

Therefore, I am certain about the detection of 3 balls.

**Final Answer:**

There are 3 balls in the picture. Their positions are approximately:

- Ball 1: [x1, y1] to [x2, y2] (in the foreground, left of center)
- Ball 2: [x1, y1] to [x2, y2] (in the foreground, center)
- Ball 3: [x1, y1] to [x2, y2] (in the foreground, right of center)

My certainty for each detection is high (certain).

# InternVL-3.5-8B

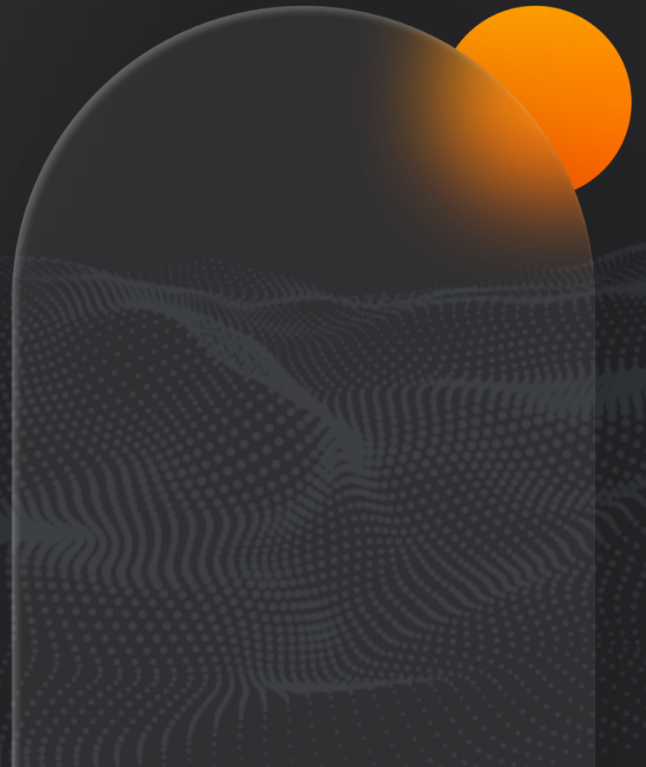


1. **Ball 1**
  - **Position**: Located slightly to the left, closer to the edge of the field. It appears to be positioned near the center line.
  - **Certainty**: Highly certain (95%) of the detection. The ball is clearly visible and distinct from the background.
2. **Ball 2**
  - **Position**: Positioned slightly to the right of Ball 1, closer to the center of the field. It is located near the center line.
  - **Certainty**: Highly certain (95%) of the detection. The ball is clearly visible and distinct from the background.
3. **Ball 3**
  - **Position**: Located further to the right of Ball 2, closer to the goal. It is positioned just to the left of the goal line.
  - **Certainty**: Highly certain (95%) of the detection. The ball is clearly visible and distinct from the background.

In summary, there are three soccer balls in the image:

1. Ball 1: Bottom left area, slightly left of center.
2. Ball 2: Centered horizontally, middle of the field.
3. Ball 3: Just left of the goal line, centered vertically.

# Conclusions and Future Work

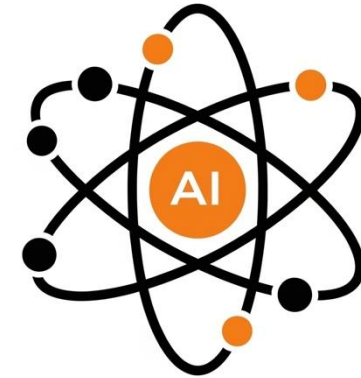


# Conclusions and Future Work

---

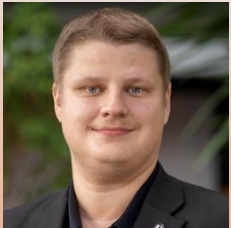
- In fact, all tested approaches (both YOLOv11-based and VLMs) worked perfectly in the images where ball was clearly visible and was close to the camera. All of them were certain about the detection.
- YOLOv11, Qwen-3, and InternVL-3.5 had huge troubles if the balls were smaller (were placed in longer distance from the camera) – in neither of the tested cases, these balls were detected.
- The best results in the case of accuracy were obtained with Gemini 3. It was able to observe the ball in both simple and complex environments. Moreover, the analysis of a single image was done in no more than 2 seconds – it was the second-best results (better was YOLOv11 – processing time was around of 30 ms per image).
- In the next step we would like to continue the experiments with specialized models – so that they are able to observe even smaller balls. We would like to also check different sort of models – like RF-DETR and newer YOLOv26.

# Questions?



**AI Practice**

# Thank you for your attention!



**Maciej Szymkowski, Ph.D., Eng.**  
mszymkowski@future-processing.com

