Master Of Science in Computer Engineering

- Intelligent Systems -

Vision-based spatiotemporal analysis of football matches

Francesco De Lucchini
Salvatore La Porta

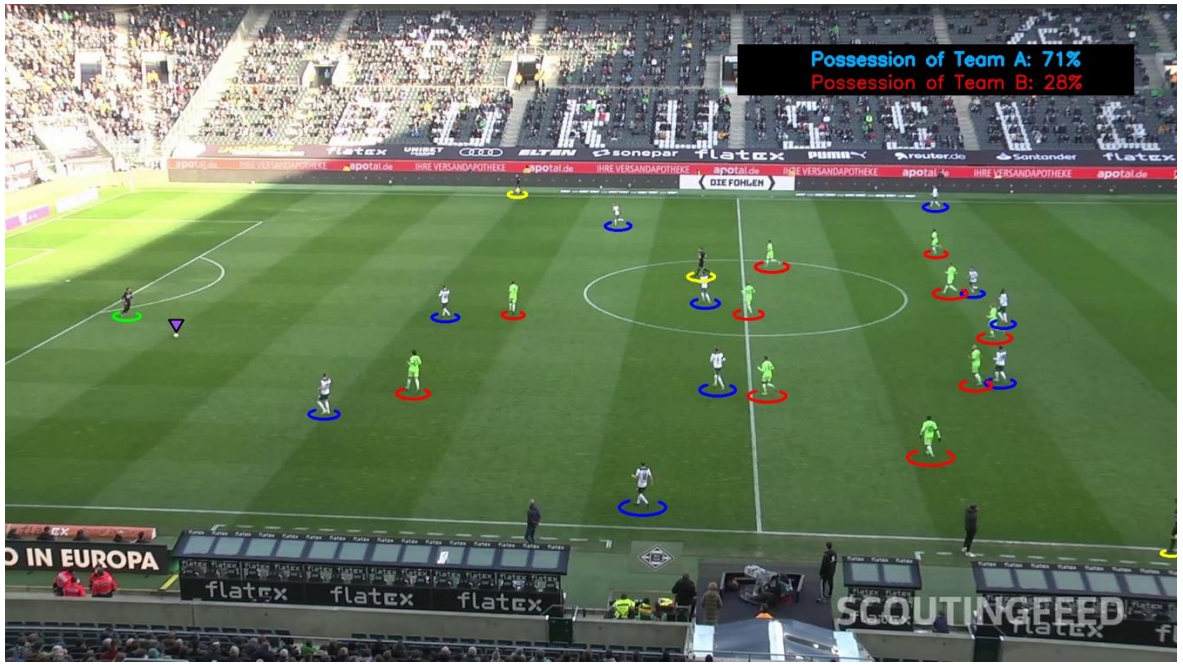
A.Y. 2024 / 2025

Problem introduction

We want to develop a **football tracking software**:

- Given a (panoramic) video feed of a football match, the software must be able to recognize the **referee**, the **goalkeepers**, the **players** and the **ball**

⇒



State of the art

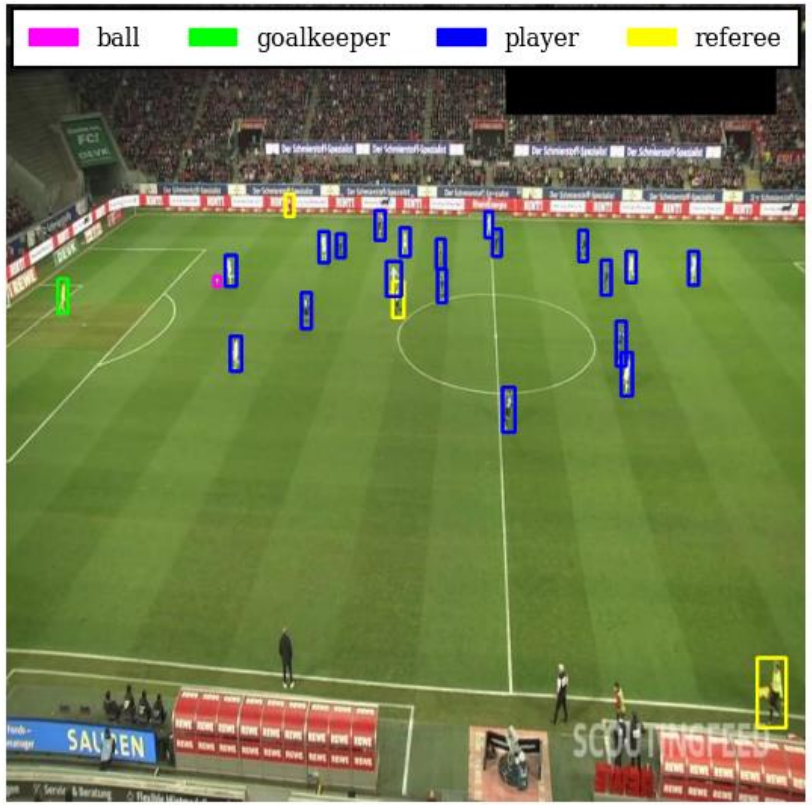
- Object detection models are generally classified into two main categories:

  - **Two-stage detectors** (e.g., Faster R-CNN): first generate region proposals, then classify and refine bounding boxes
    - ➢ Usually more accurate but slower
  - **One-stage detectors** (e.g., The YOLO family): perform detection in a single forward pass, directly predicting class probabilities and bounding boxes
    - ➢ More popular, good compromise between accuracy and speed

Dataset description

- The dataset (https://universe.roboflow.com/roboflow-jvuqo/football-players-detection-3zvbc/dataset/14) contains 372 images, each with a resolution of 640×640 pixels, annotated with **four classes**:

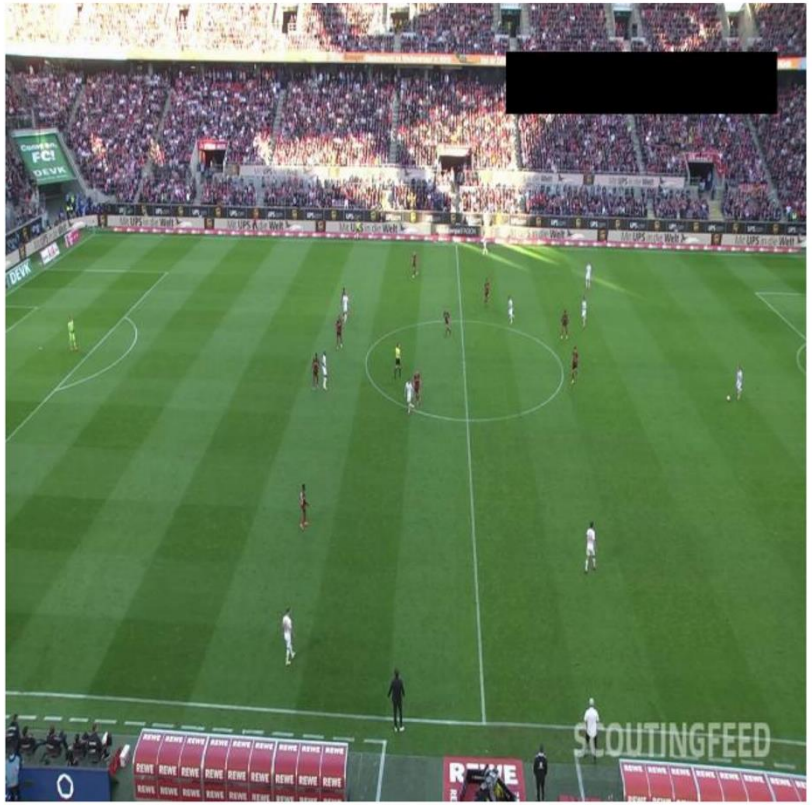
	Ball	Goalkeeper	Player	Referee
Mean occurrences per image	0.88	0.77	19.96	2.32



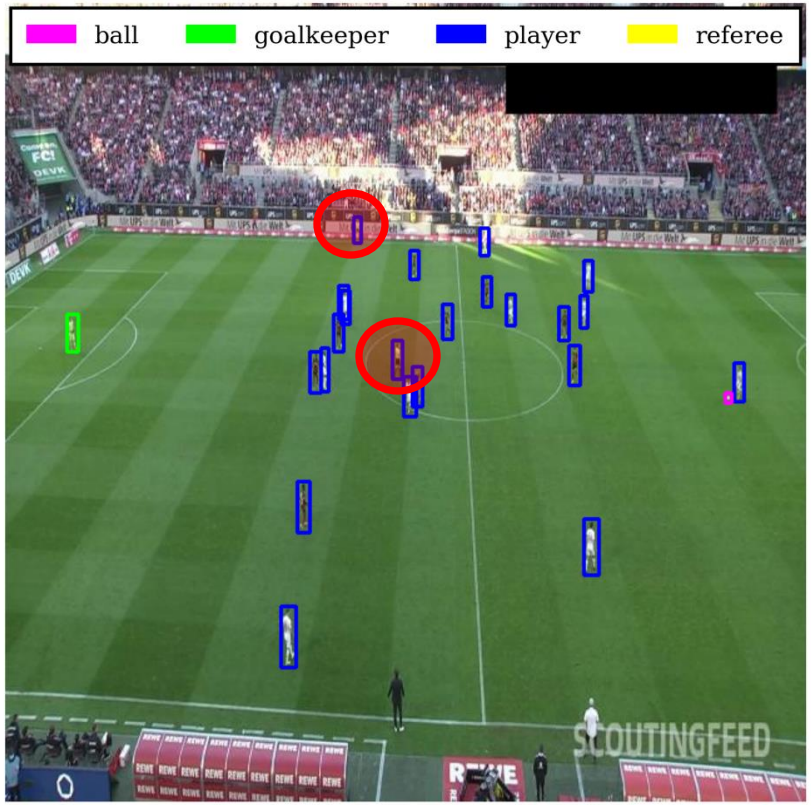
Consistency checks

- Are there any images with more than 20 (moving) players?
- Are there any labels with either 0 or more than 3 referees?

Original image



Annotated image (ground truth)

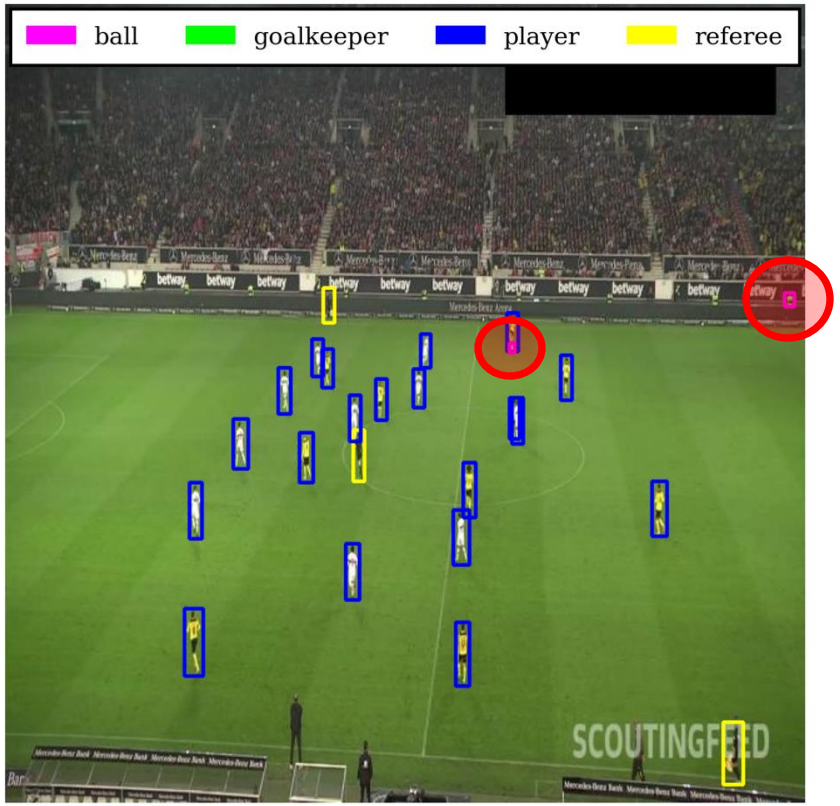


Consistency checks

- Are there any labels with more than 2 goalkeepers?
- Are there any images with more than one ball?

Original image

Annotated image (ground truth)



Refined dataset

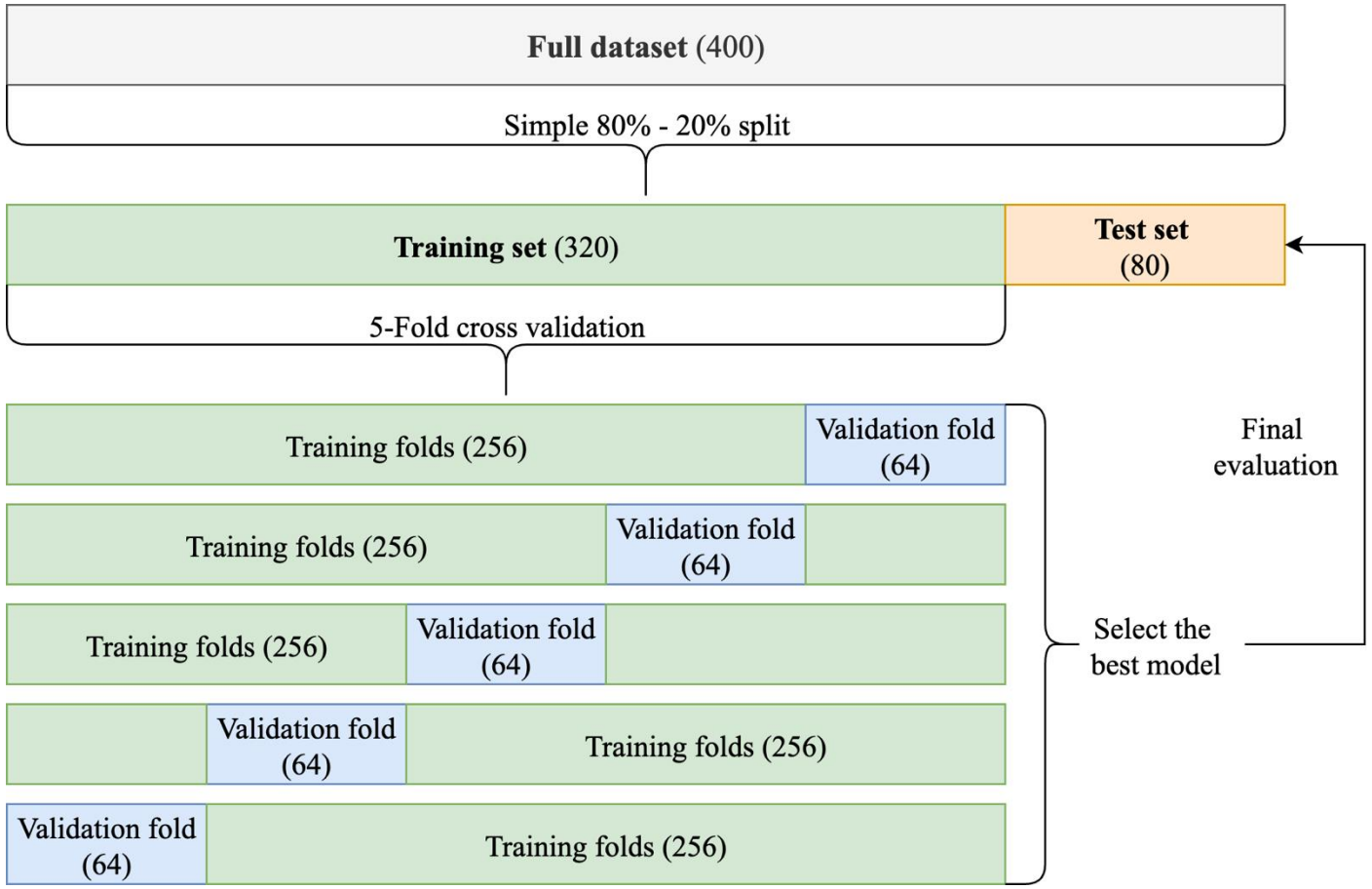
- The **original dataset is of relatively low quality**

⟹ We manually reviewed all 372 images it contained and, when necessary, we **re-annotated** them using “Label Studio”

Label Studio

- Additionally, we **expanded** the dataset by introducing 28 new images, bringing the total to **400**

Data splitting

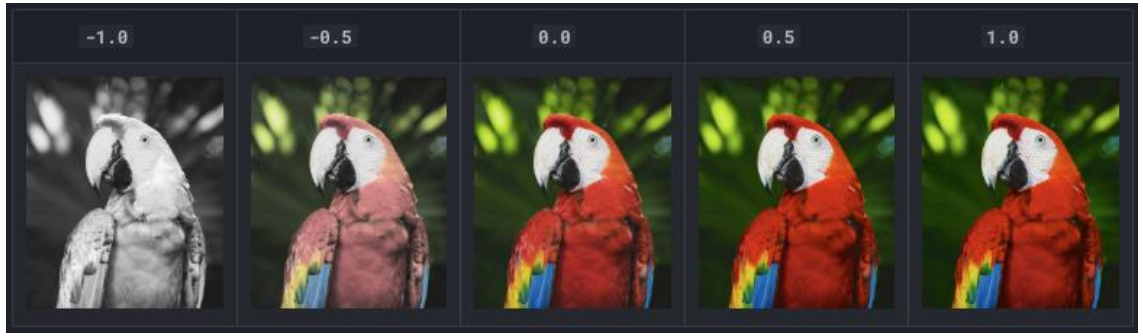


Ideally, with a small dataset, one should adopt **nested cross validation** (or similar techniques)

Data augmentation

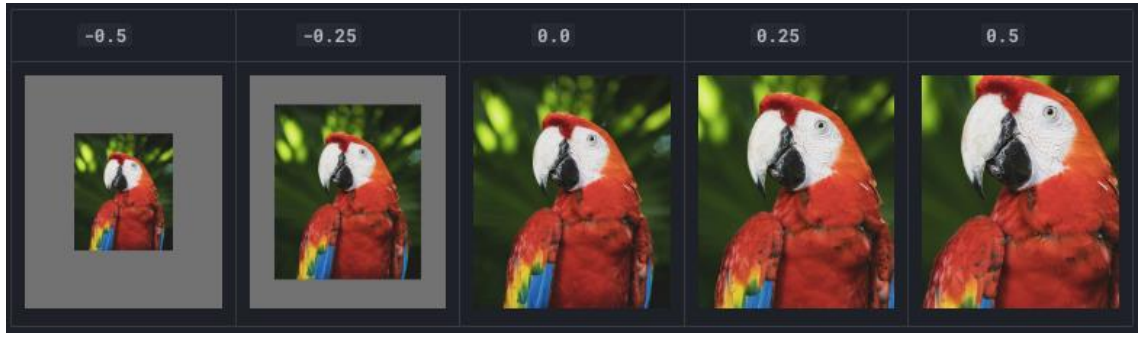
- Ultralytics YOLO's training framework automatically applies several data augmentation techniques:

Saturation

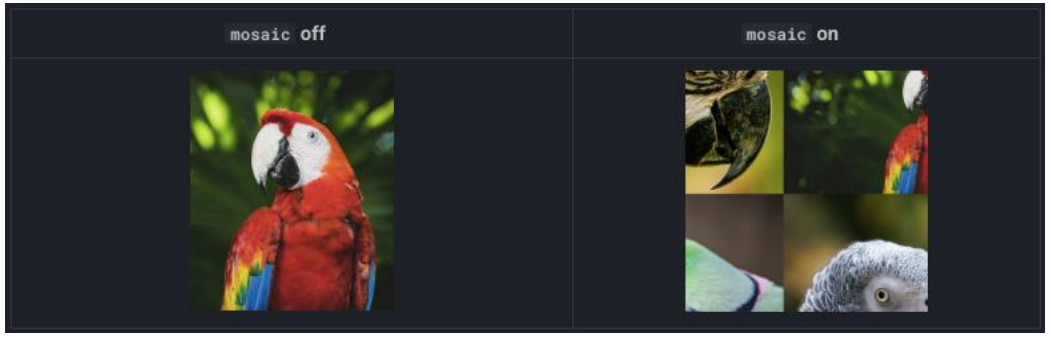


Horizontal flip

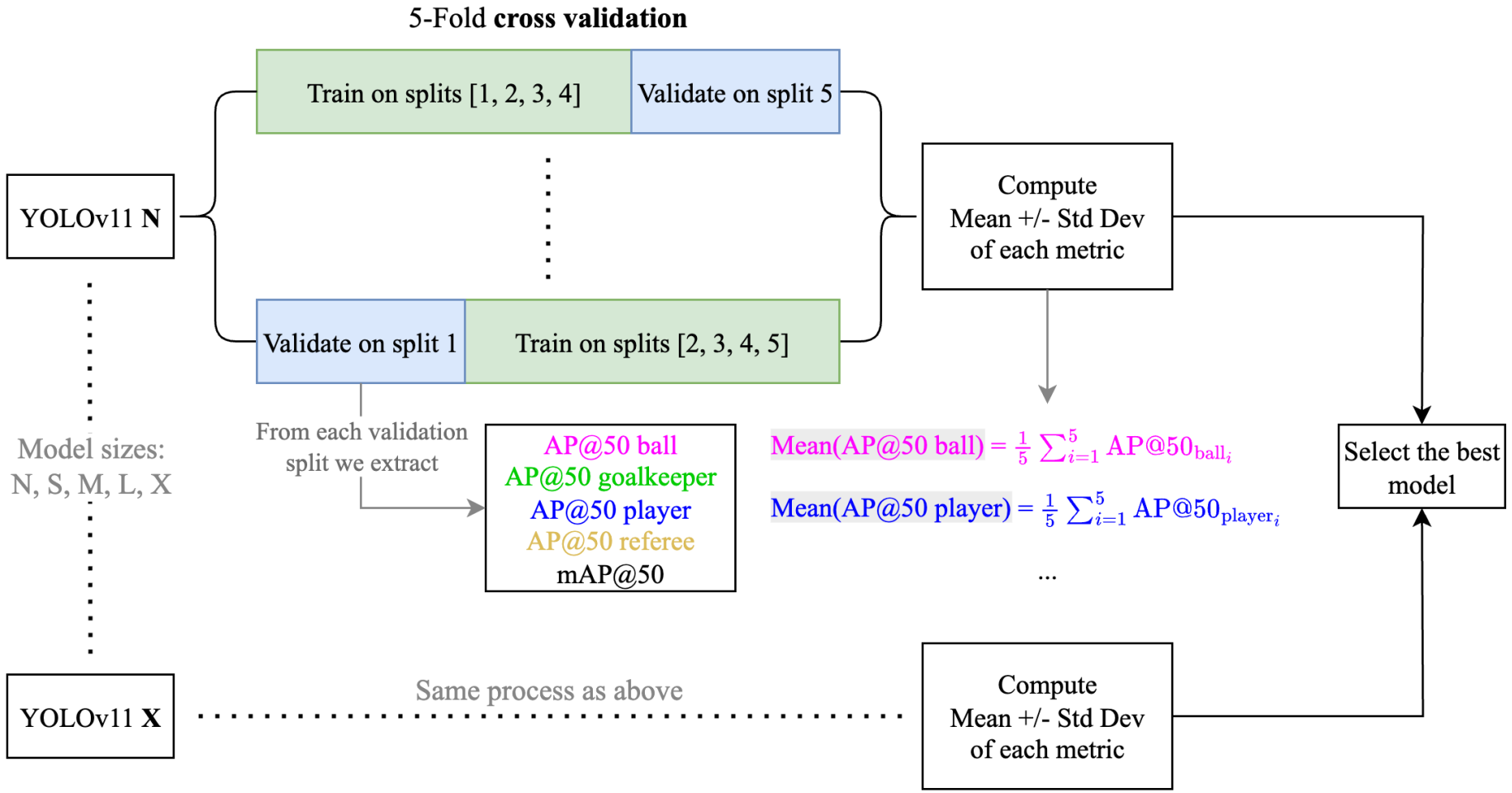
Scale



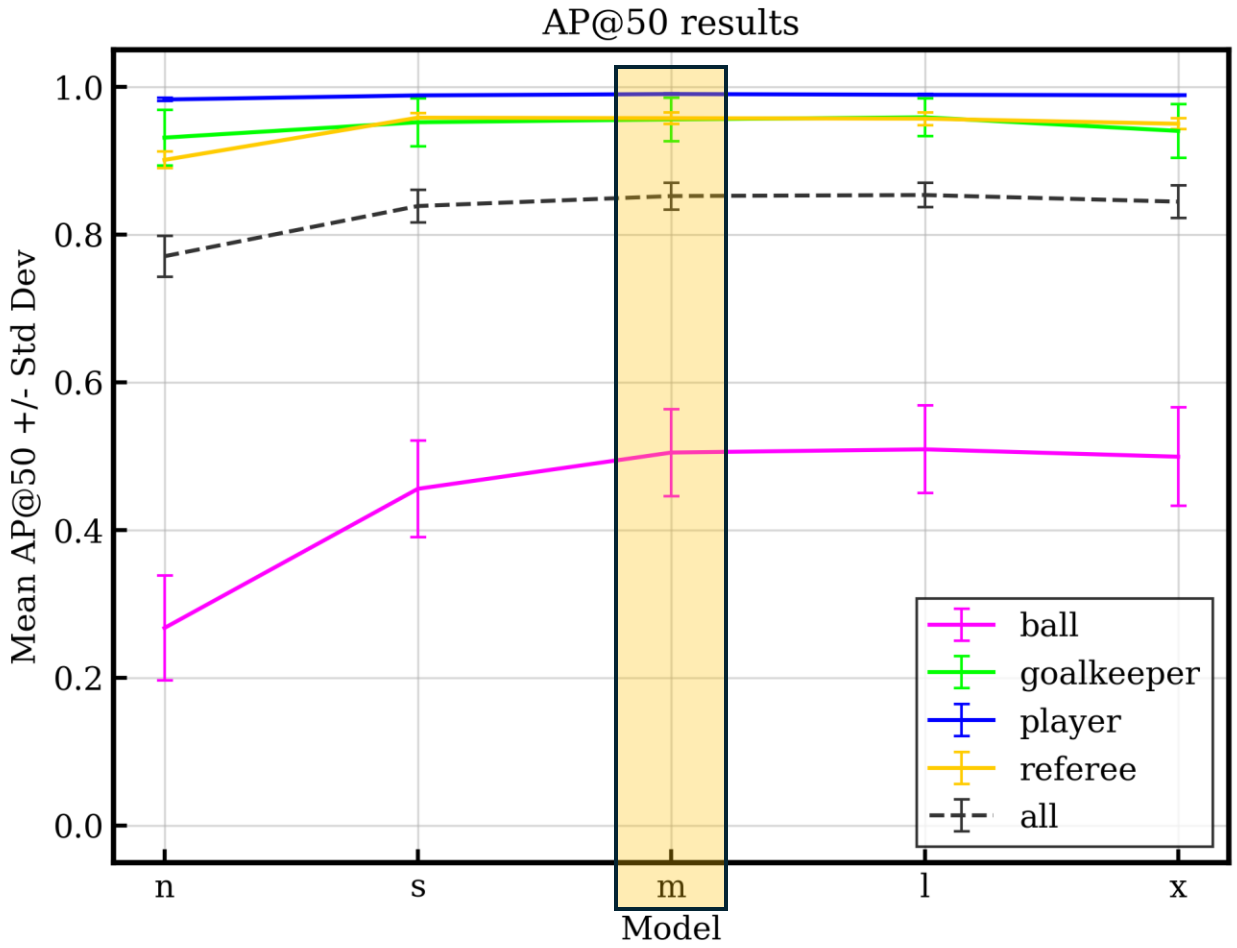
Mosaic



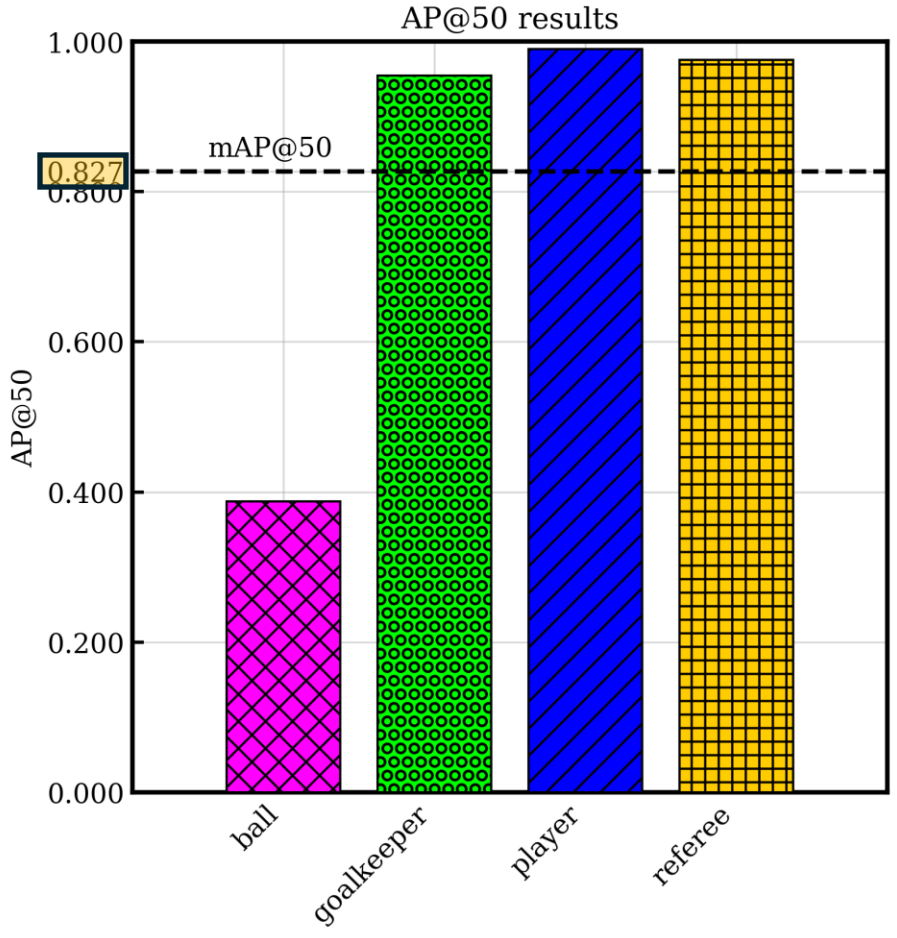
Training strategy

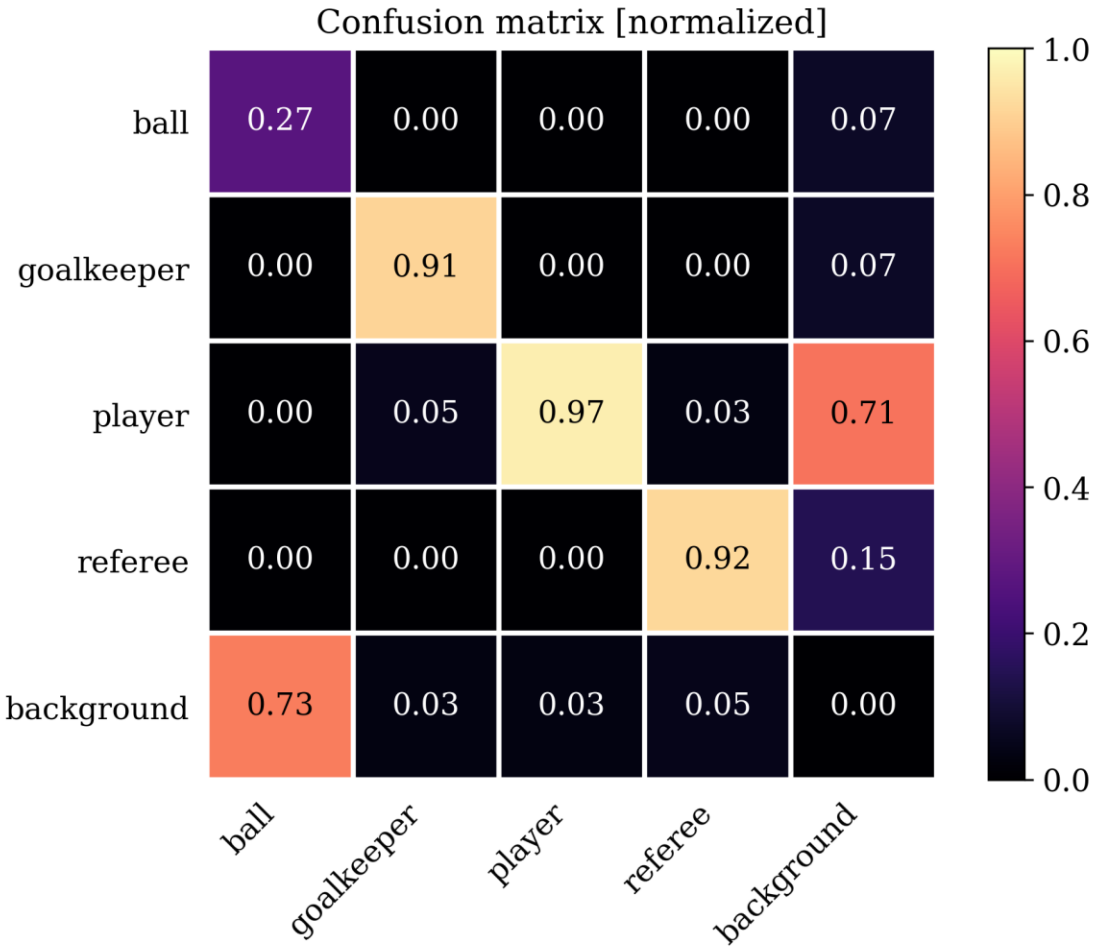


Selection of the best model



Final evaluation of the selected model





Final evaluation of the selected model

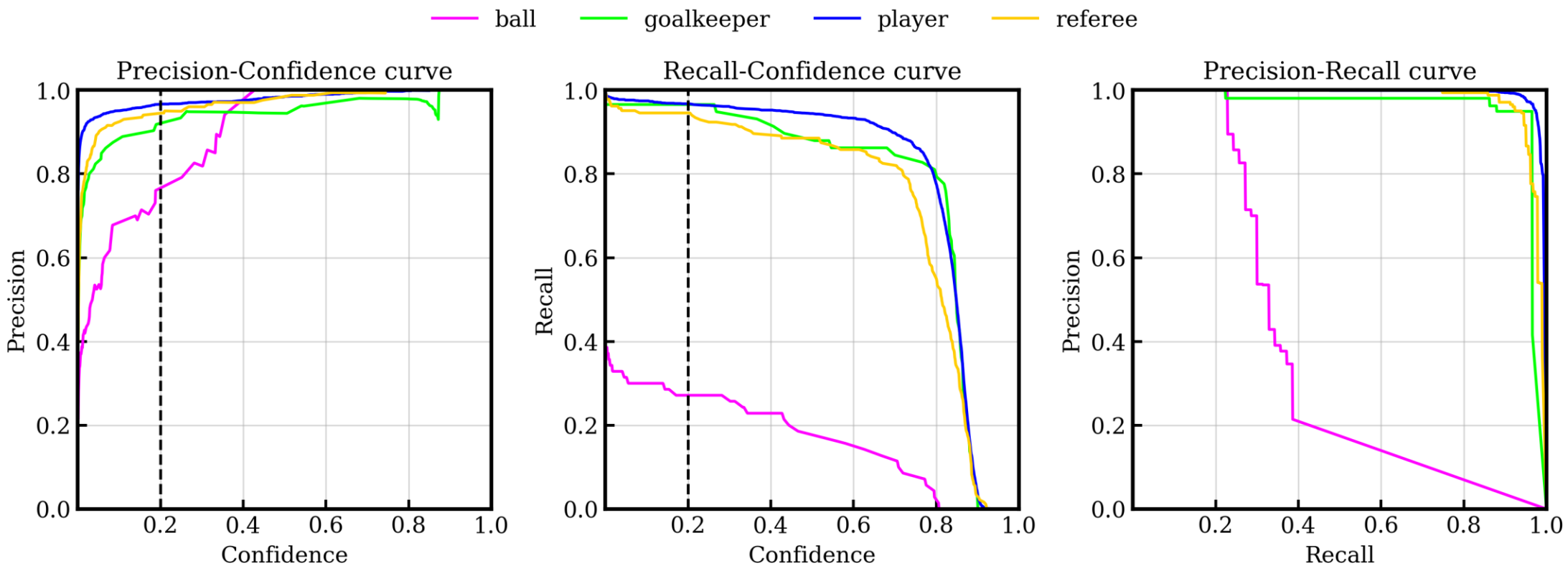
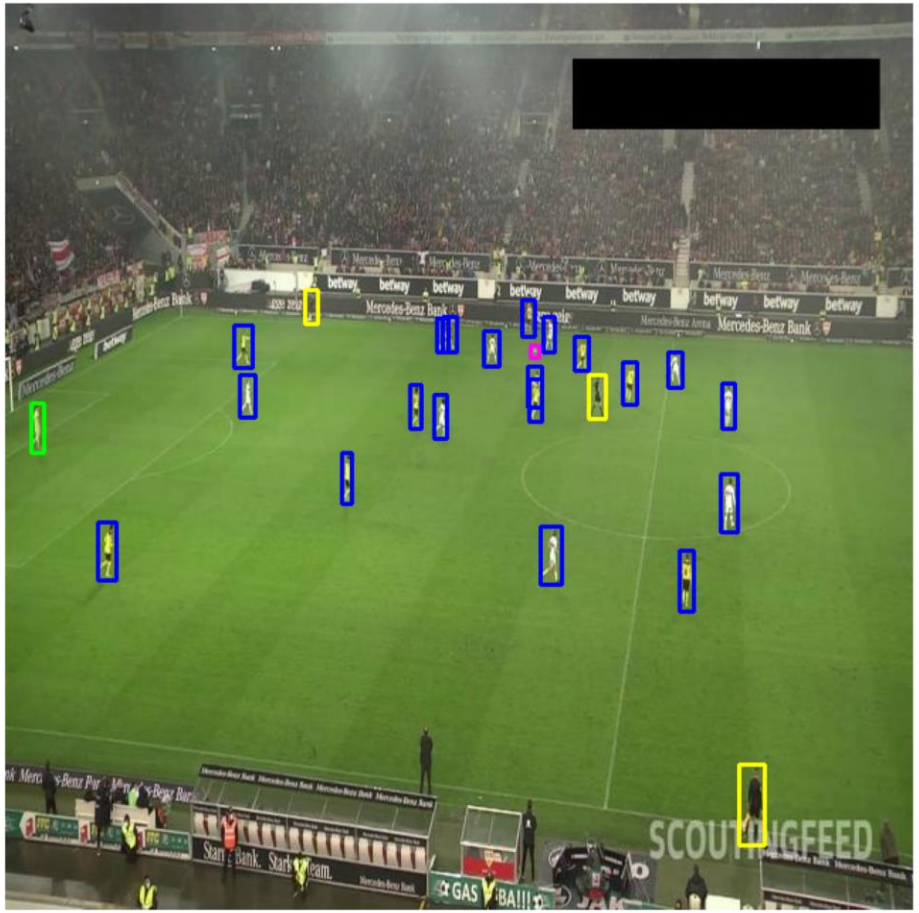
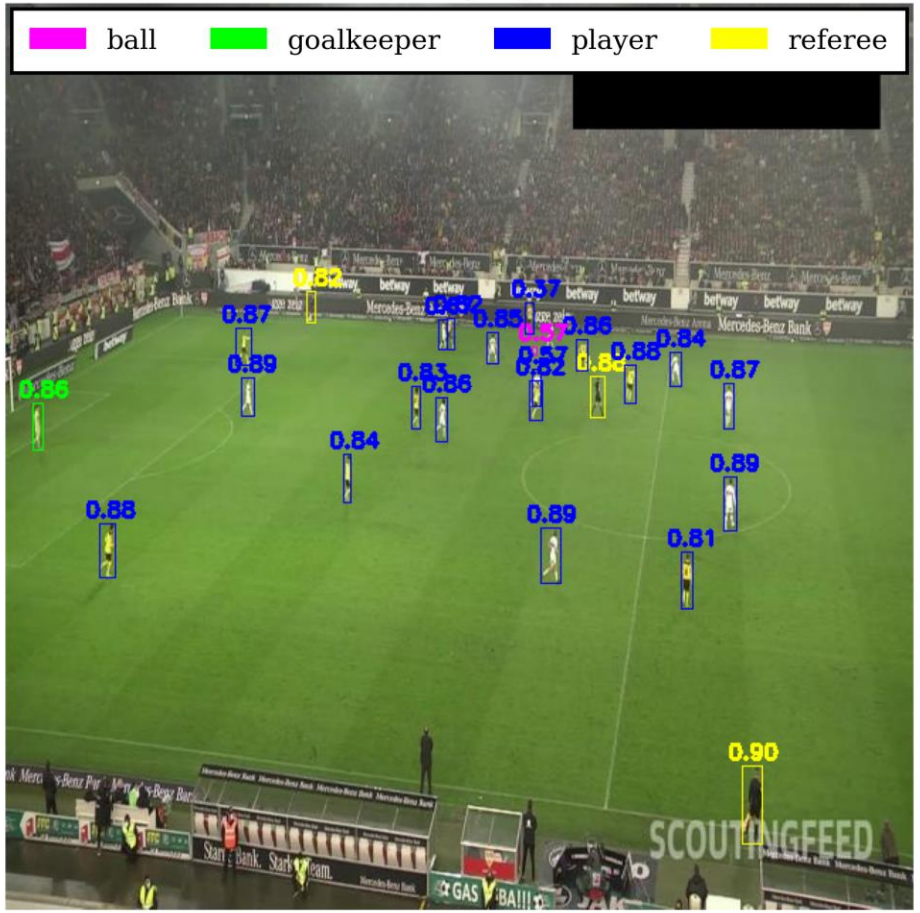


Image prediction demo

Annotated image (ground truth)



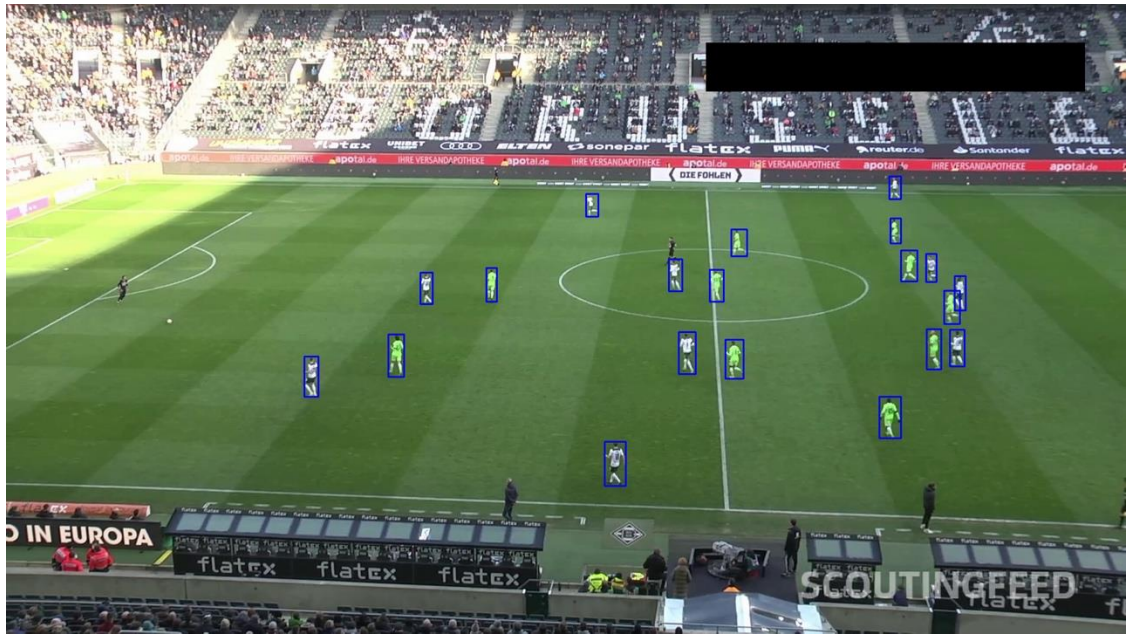
Annotated image (predicted)



Team detection

Since there are no team labels, we must use an **unsupervised approach**:

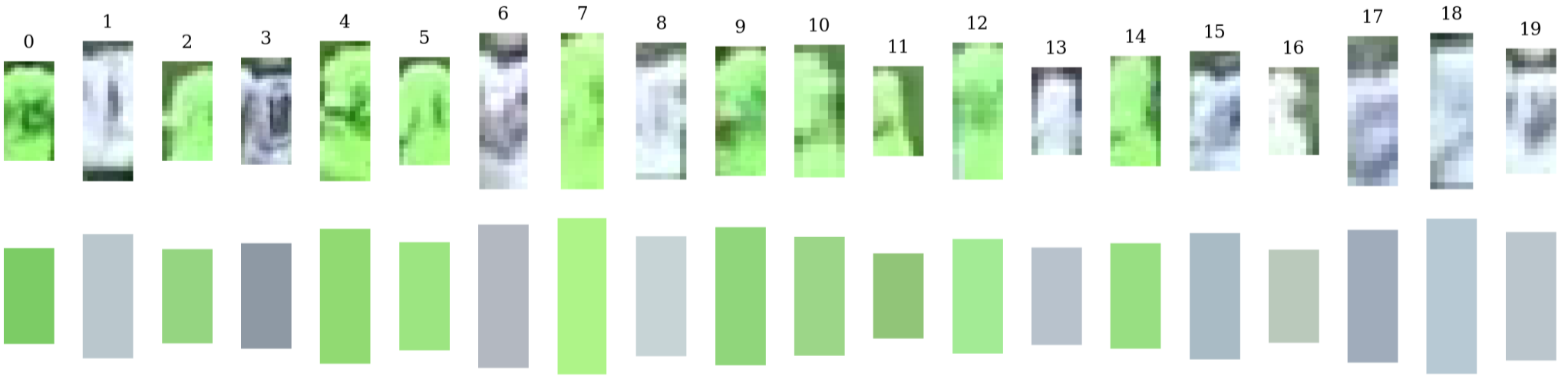
1. Isolate the players



Team detection

Since there are no team labels, we must use an **unsupervised approach**:

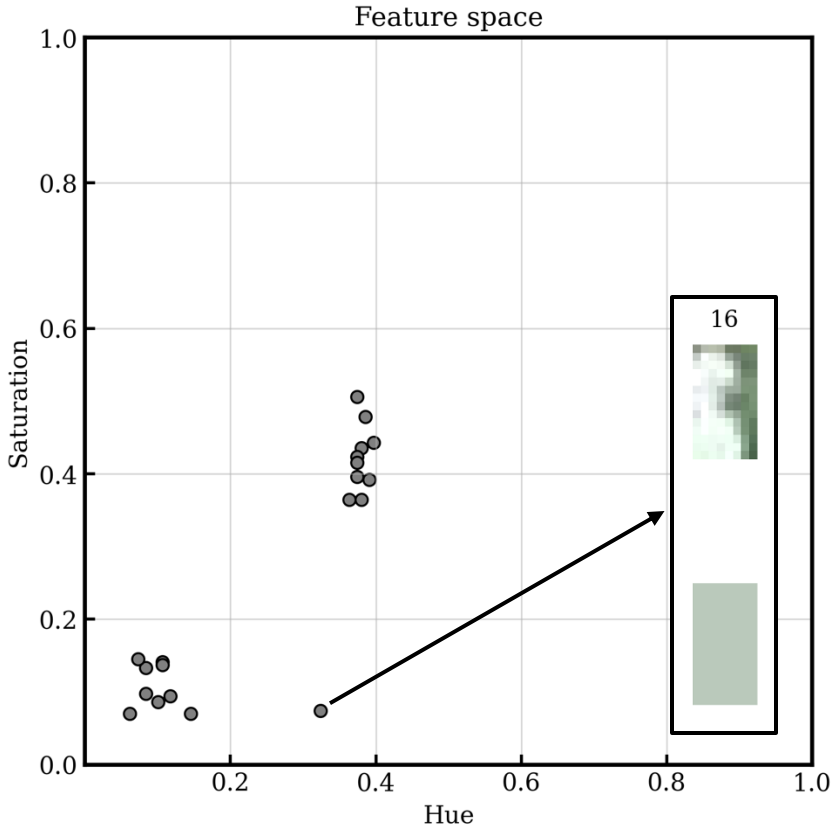
2. Take the upper third of the bounding box (i.e., the player's shirt) and calculate its mean color



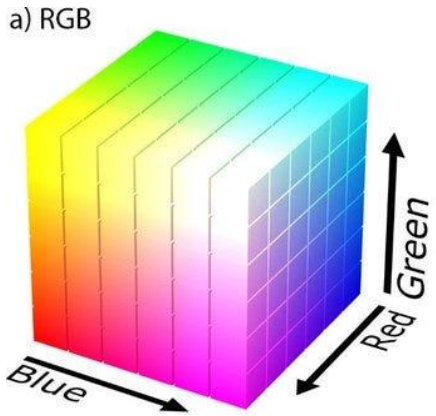
Team detection

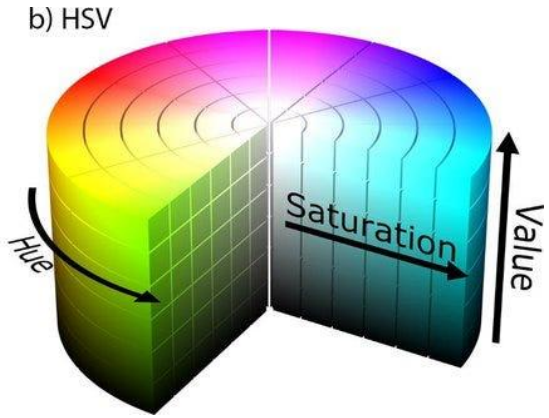
Since there are no team labels, we must use an **unsupervised approach**:

3. Convert the mean colors from RGB to **HSV** space for better clustering



We don't care about the **V**alue (brightness) since it does not define color

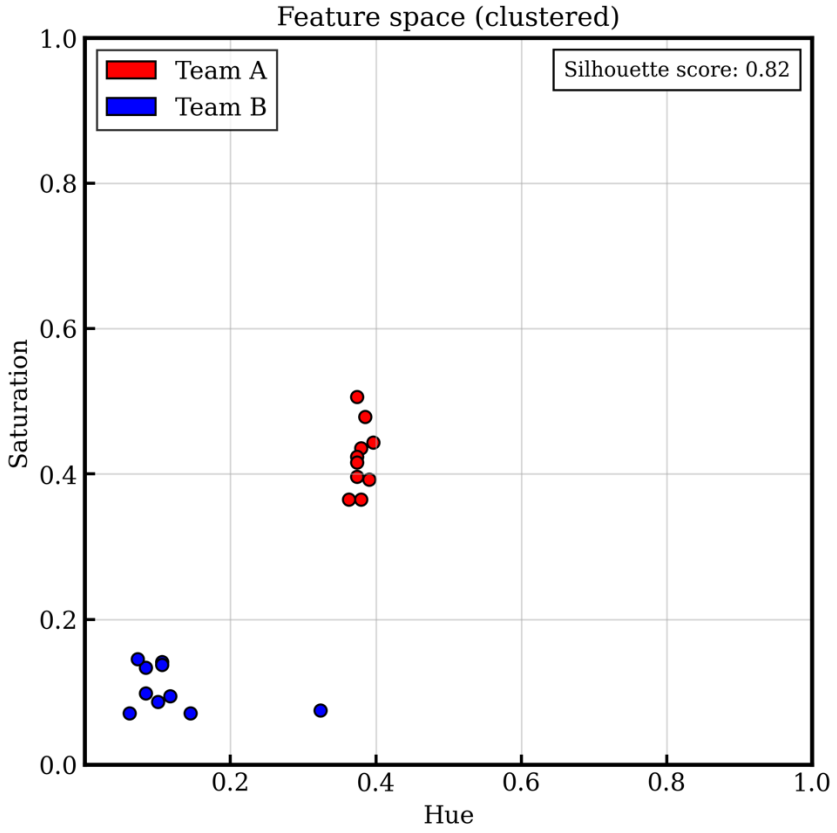




Team detection

Since there are no team labels, we must use an **unsupervised approach**:

4. Apply any clustering/separation algorithm (we used **K-Means** with K=2)



5. Annotate the image with the right colors