

# Multi-Modal Soccer Scene Analysis with Masked Pre-Training

## Supplementary Material

Marc Peral<sup>1,2</sup> Guillem Capellera<sup>1,2</sup> Luis Ferraz<sup>2</sup> Antonio Rubio<sup>2</sup> Antonio Agudo<sup>1</sup>  
<sup>1</sup>Institut de Robòtica i Informàtica Industrial, CSIC-UPC <sup>2</sup>Kognia Sports Intelligence

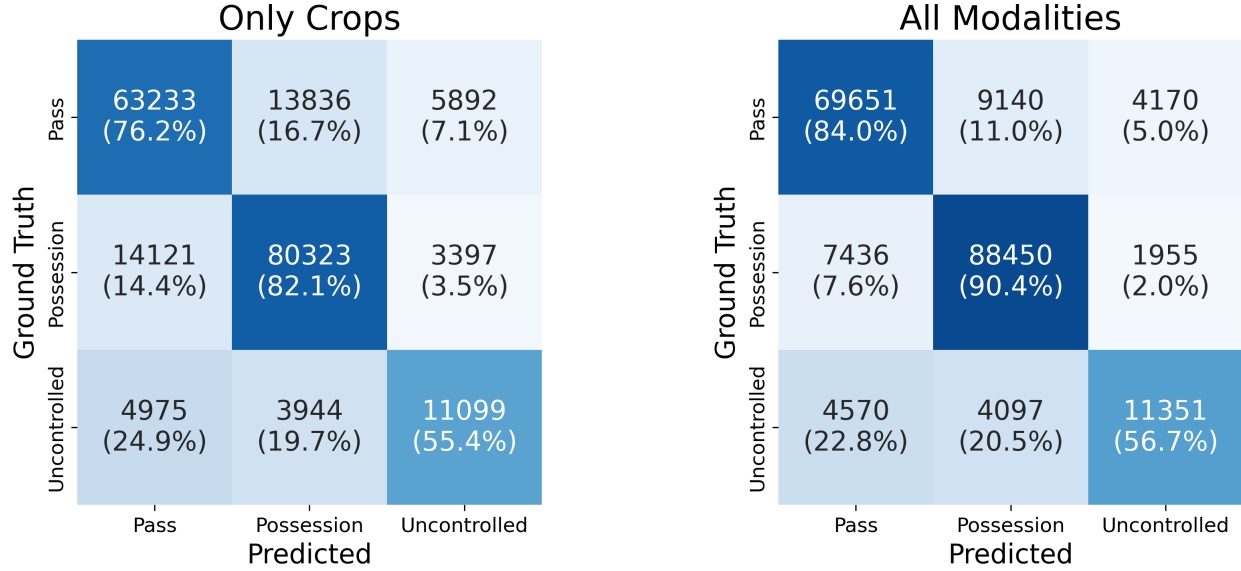


Figure 1. **Confusion matrices for ball state classification.** Left matrix shows results for the setting using only crops, while the right matrix corresponds to our full model with all three input modalities.

### 1. Ablation on Multi-modality

Table 2 reports Average Displacement Error (ADE) and Max Error (ME), both in meters, for ball trajectory, and State Accuracy (SA) and Possessor Accuracy (PA), in percentages, for different input modality combinations. As expected, trajectories dominate ball position inference since player types and crops carry no positional information, though they still provide complementary cues when combined with trajectories. One might argue that the reasonable results obtained in the classification tasks without trajectories could be due to dataset biases (e.g., frequency-based guessing), which would make these tasks trivial.

For possessor identification, however, the dataset contains many sequences where different players hold possession, leading to a large number of classes and a highly distributed label space. Combined with the model’s equivariance property, ensuring predictions do not depend on the order of players, this rules out the possibility of such trivial

biases.

For ball state classification, we already provide class distributions in the main paper, but to further address this concern we include confusion matrices in Fig. 1. The left matrix shows results for the setting using only crops (first line of Table 2), while the one matrix corresponds to our full model with all three input modalities (last line of Table 2).