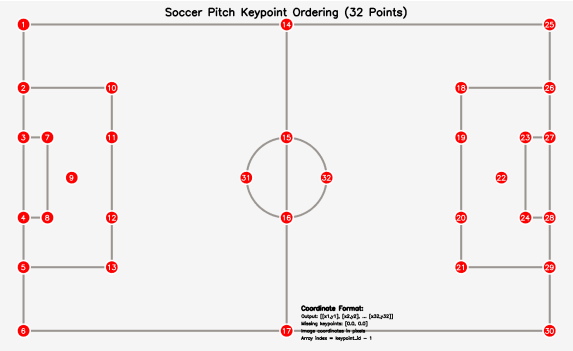
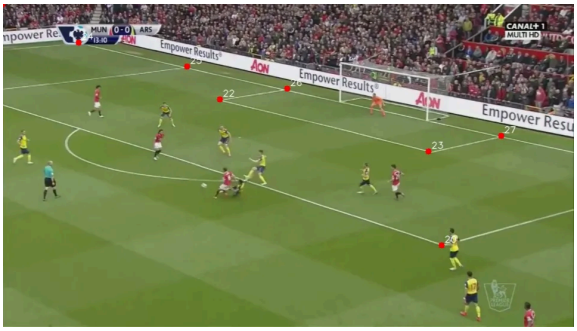
New Keypoint Scoring Mechanism

Introduction

What are Keypoints?

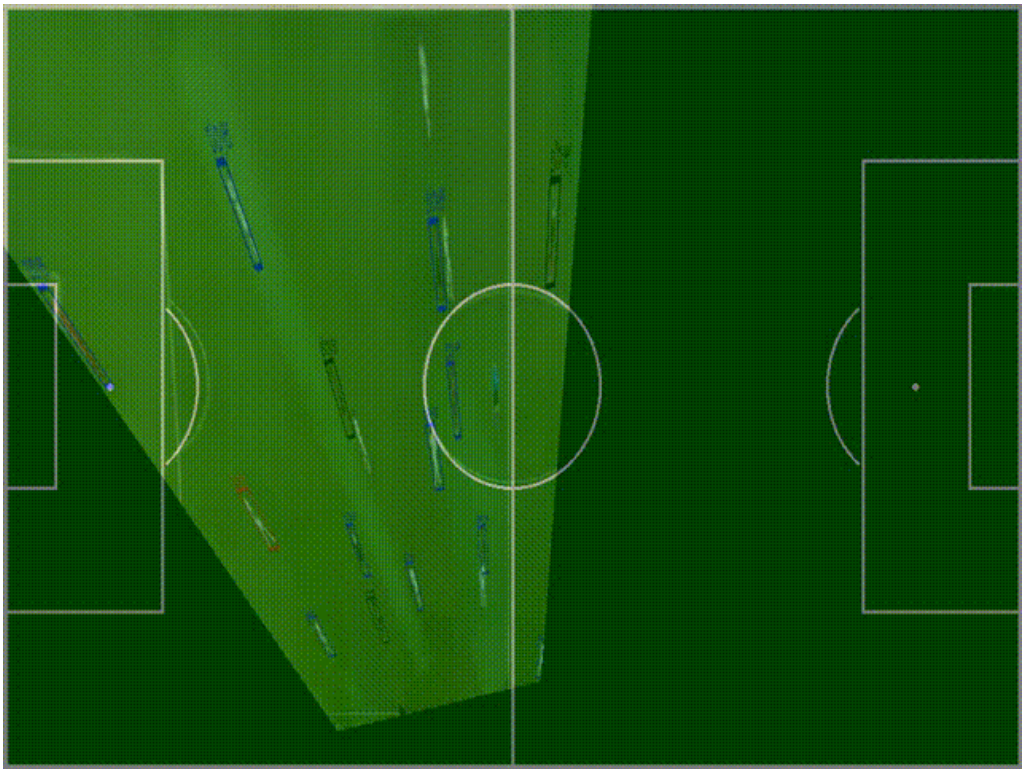
A ***keypoint*** is simply a coordinate `(x, y)` in the image that you can identify (e.g. a goal post, the center circle, the pitch corner, etc) and it serves as a **reference point** that **anchors** pixels in the image to known positions in the real world.



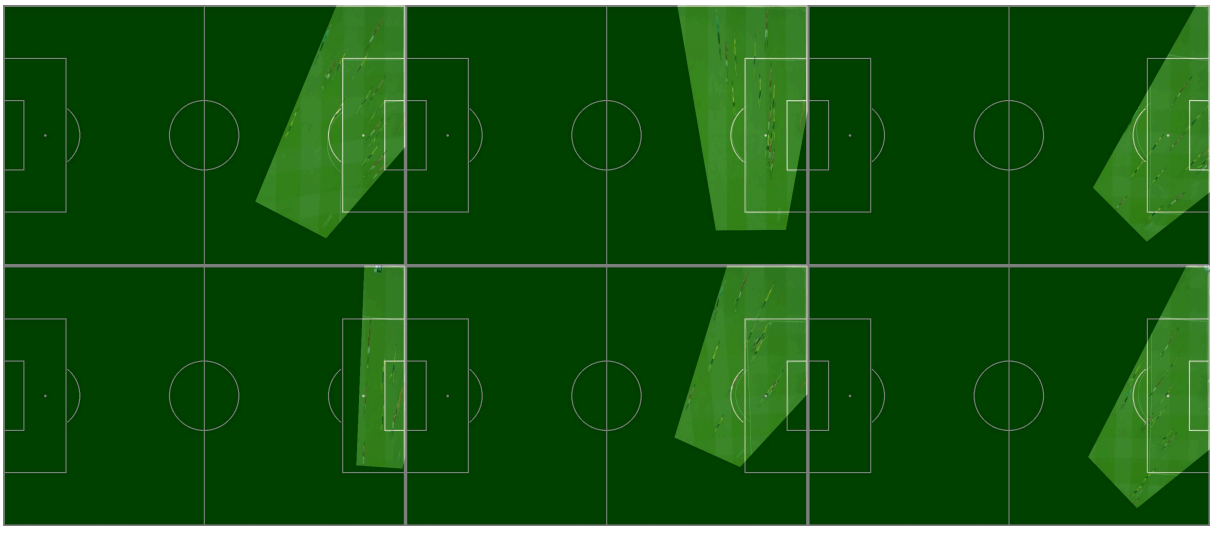


We have selected 32 keypoints for the football challenge type (above left), a few of which also appear in the frame of a football match (right).

Why Keypoints?



If a sufficient number of keypoints (>4) are identified in a frame, the pixels in that frame can be reliably projected onto a global map, making it easier to locate and track the motion of objects of interest (i.e. players, goalies, ball, etc) throughout the game.



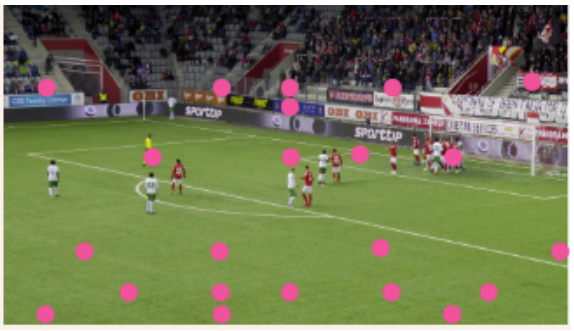
Lack of Label Problem (No Ground Truth Keypoints)

> Most algorithms are typically trained in a fully or semi-supervised fashion, which relies on expensive training datasets of human-labelled images which are time-prohibitive to produce. Additional manual annotations are relied

> upon to evaluate these algorithms during training and after. Alternatively, a human-in-the-loop is employed to assess quality, but this is also slow and potentially unreliable as it depends on the visual inspection of a few selected samples instead of a systematic approach based on quantitative metrics.

Pseudo Ground Truth Keypoints

Previously, we solved the lack of label problem with a sophisticated VLM-based approach which could generate Pseudo Ground-Truth labels (that would then be validated using another VLM-as-Judge). VLM's were shown to be very good at generating pseudo-GT annotations for bounding boxes around objects. However, they did not prove as effective when attempting to generate pseudo-GT keypoints!

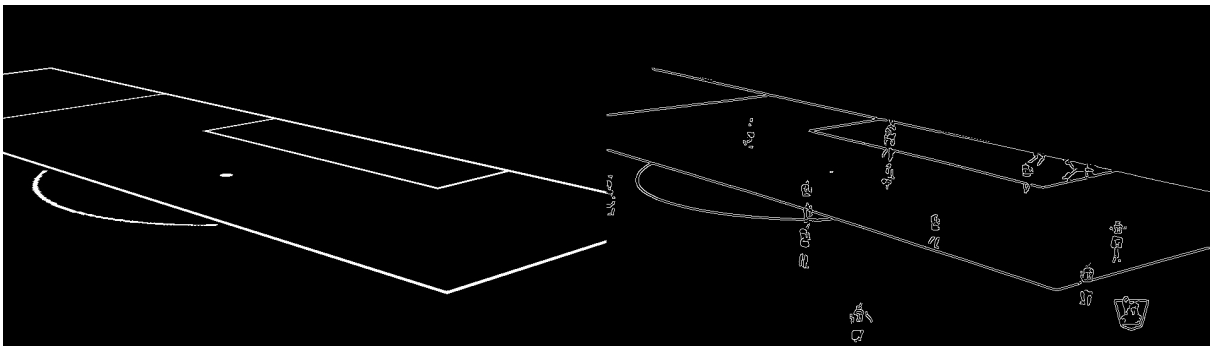


They would either generate keypoint coordinates which were anchored to players, as opposed to keypoints (left) or just a random scatter of meaningless coordinates (right). Even MOLMO, a new VLM trained to predict keypoints, performed surprisingly bad at this task

Label-Free Keypoint Validation

Without pseudo-GT annotations, we began to take a step back and ask if we actually NEED keypoint annotations to be able to validate miner keypoint predictions. How else can we judge the accuracy of a miner's keypoints if we do not have any GT keypoints to compare them against?

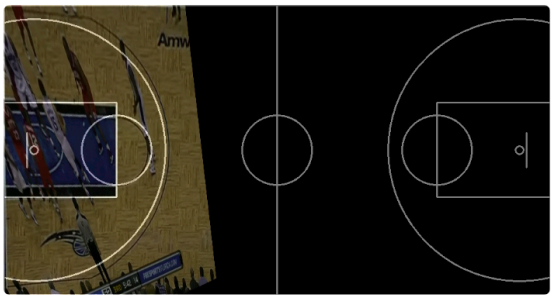
During experimentations trying to solve this validation task, a unique approach was arrived at that could measure the quality of a miner's keypoint predictions without requiring a single GT keypoint at at all!



Instead, the miner's keypoints are used to project the global template into the perspective of the the video frame. The more accurate the keypoints are, the more aligned the projected template lines would be with the pixels shown in the frame. With this in mind, the lines of the projected template (left) can be compared directly to the lines of the pitch in the frame (right).

Generalisable

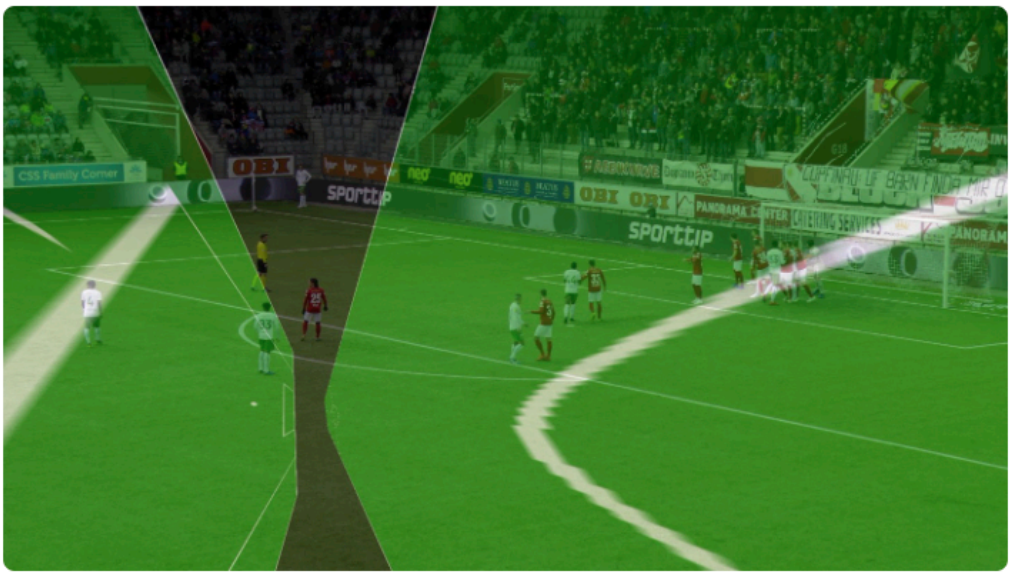
This method is agnostic to the challenge type and can be applied to any video that contains distinct reference markings on the ground (e.g. basketball, cricket, etc)



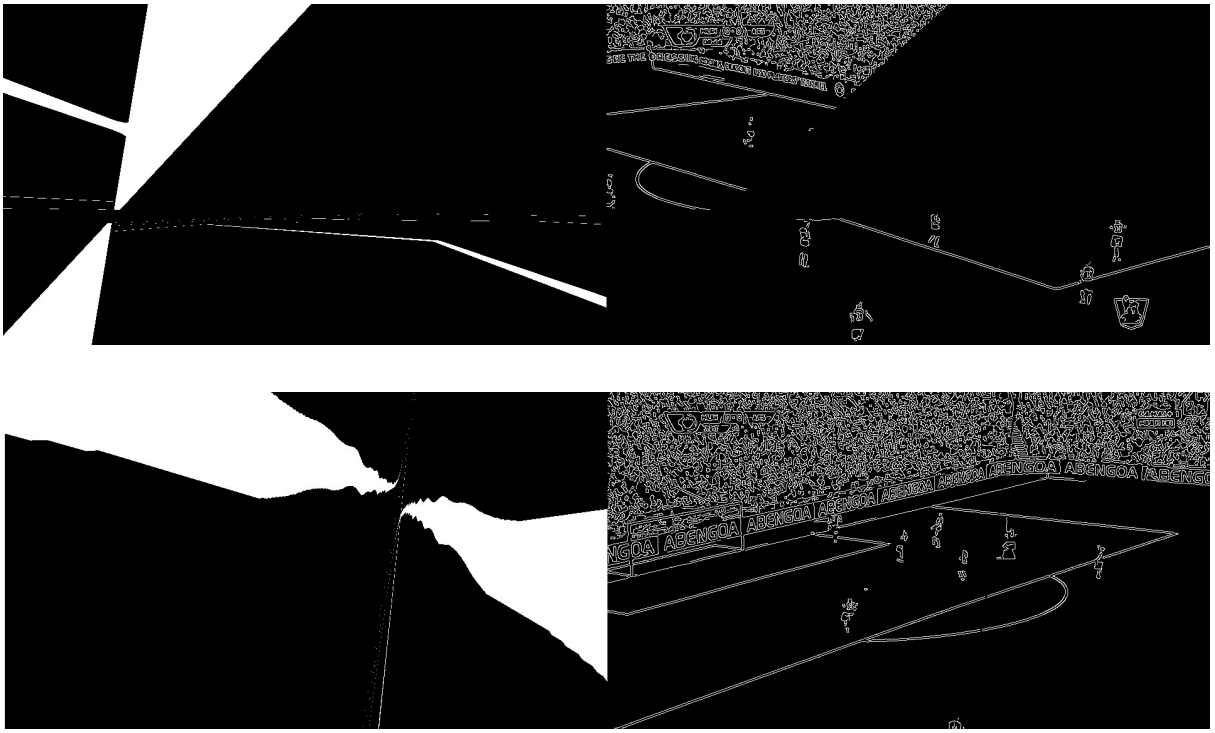
Scalable

The method bypasses the need to generate pseudo-GT labels too, so there is no heavy computation involved and it can be quickly applied to all miners and all frames in the challenge video (as opposed to only applying it to a subset).

Robust against Exploits



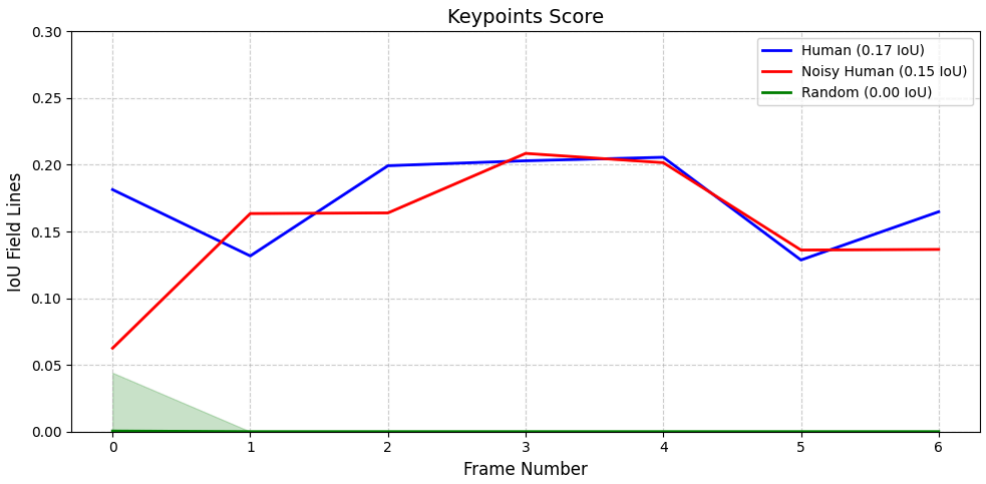
It is not easy to use this metric to reverse-engineer the keypoints (which will be unique to each frame) since there are so many possible coordinates that each keypoint could be. While a brute-force permutation is possible to search for keypoints that result in the best scoring transform, it would be so slow and inefficient, that it would be easier to use an intelligent search algorithm for the keypoints or train a model to predict them (randomly placed keypoints almost always produce concave transforms because it is so rare to find a combination of points which result in a sensible projection - see below).



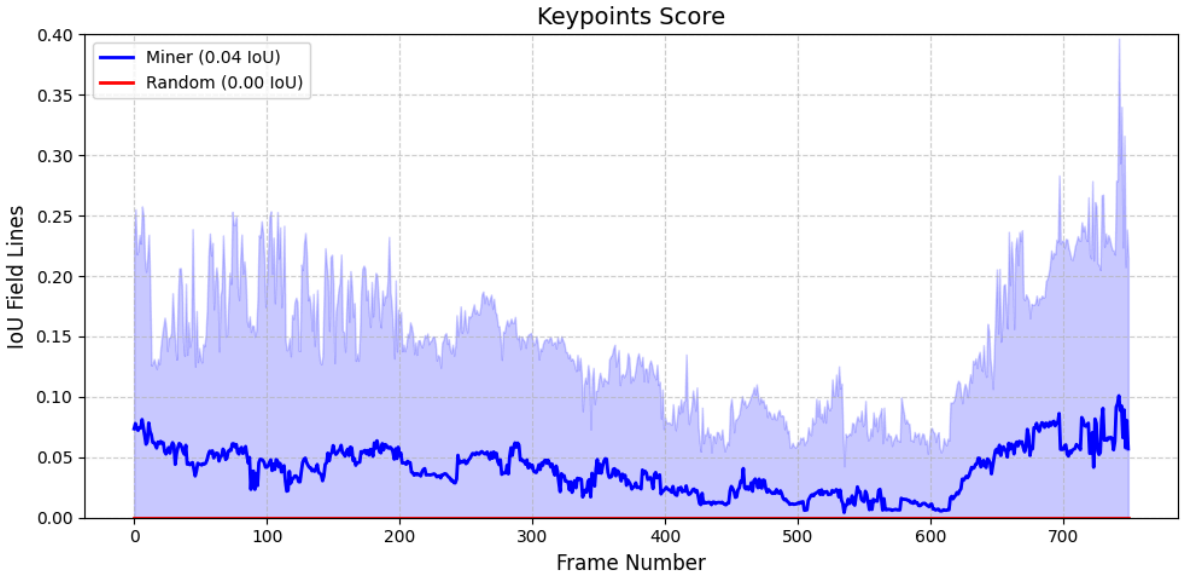
However, another exploit involves finding a projection that makes the lines become so distorted that they end up covering an extremely large area of the image, thus resulting in a higher overlap score (since the quality of the keypoints are ultimately measured via the overlap (IoU) of the lines). But these exploits are cut off because the transformed template is validated to ensure it is concave (ie not twisted or wildly distorted) - as we know the optimal projection will be in the relatively smaller convex subset (and only exploits lie in the concave set).

Benchmarking the Scoring mechanism

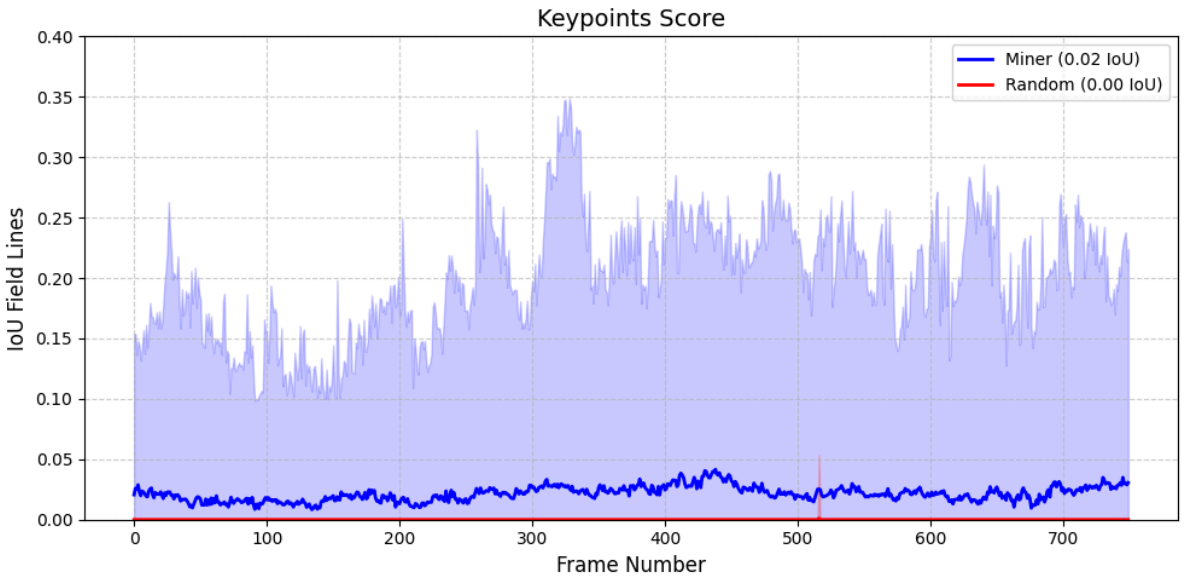
We evaluated the metric by comparing human annotated keypoints to the same human annotations with some jitter added to the coordinates (making them less accurate) and also to randomly generated keypoints (an open-source dataset was found which derived keypoints from human annotated soccernet lines). We expect to see the human annotations score consistently higher than the noisy human annotations and the random keypoints. This was indeed the case



We also had a lot of in-house data for keypoint predictions generated by miners. Although the quality of the keypoints predicted by miners is still much lower than the quality of human annotated keypoints, they should perform better than randomly generated keypoints on average.



Miner G (above) was tested using 10 challenge videos (each with 750 frames) and miner T (below) was tested using 50 challenge videos. Both miners did indeed perform better than random, again confirming the reliability of the scoring mechanism. In fact, miner G produced a higher average score than miner T which was also is inline with the qualitative analysis of the predictions (upon inspection, miner G appeared to have far more stable keypoints than miner T).



The Scorevision baseline miner was also tested on the same 50 challenges and its results are shown below (compared against miner T):