Supplementary Materials

A . Additional Details

In this section, we provide some additional details of our loss functions and visualise object tracking with scene graphs for all three datasets.

A .1. Loss Functions

Our problem is therefore multi-objective, as we aim at not only recognising the label of the activity taking place but also finding its boundary (start and end time).

Given the ground truth $y_s = \{y_{s_1}, \ldots, y_{s_i}, \ldots, y_{s_N}\}$ and predicted $\hat{y}_s = \{\hat{y}_{s_1}, \ldots, \hat{y}_{s_i}, \ldots, \hat{y}_{s_N}\}$ activity labels, and the ground truth $y_{br} = \{y_{br_1}, \ldots, y_{br_i}, \ldots, y_{br_N}\}$ and predicted $\hat{y}_{br} = \{\hat{y}_{br_1}, \ldots, \hat{y}_{br_i}, \ldots, \hat{y}_{br_N}\}$ temporal extent labels, our overall loss function is formulated as:

$$\mathcal{L} = \lambda \cdot \mathcal{L}_{Act} + \mathcal{L}_{Br}. \tag{1}$$

The second component denotes the binary cross entropy (BCE) loss $\mathcal{L}_{Br} = \{l_1, ..., l_i, ..., l_N\}^T$, where

$$l_i = -w_i \Big[y_{br_i} \cdot \log \hat{y}_{br_i} + (1 - y_{br_i}) \cdot \log(1 - \hat{y}_{br_i}) \Big],$$

driving the recognition of the activity's temporal extent in class-agnostic manner.

$\mathcal{L}_{Act}$ denotes, instead, the *BCEWithLogitsLoss* [**?**], defined as $\mathcal{L}_{Act} = \{l_{1,c}, \ldots, l_{i,c}, \ldots, l_{N,c}\}^T$, where

$$\begin{aligned} l_{i,c} = -w_{i,c} \Big[& p_c \cdot y_{s_i,c} \cdot \log \sigma(\hat{y}_{s_i,c}) \\ & + (1 - y_{s_i,c} \cdot \log(1 - \sigma(\hat{y}_{s_i,c})) \Big]. \end{aligned} \tag{2}$$

Here i is the index of the sample in the batch, c is the class number, p_c is the weight of a positive sample for class c, and σ is the sigmoid function. This loss component is used to predict the activity labels $\{y_{s_i}\}$, and combines classical binary cross entropy with a sigmoid layer[1]. This combination is proven to be numerically stronger (as it leverages the log-sum-exp trick) than using a plain sigmoid separately followed by a BCE loss.

Finally, λ is a weight term, which we set to the number of selected anchor proposals (128 in our case).

A .2. Objects Detection, Tracking, and Scene Graphs

In addition to Fig. 2 in the paper, we also provide a pictorial illustration of agent tubes in an example video segment from all three datasets, showing both a sample frame, with overlaid the detection bounding boxes, and a bird's-eye view of the agent tubes and a local scene graph for the snippet it belongs to.

ActivityNet-1.3 is one of the largest temporal action localisation datasets which includes activities performed by both individual and multiple objects (agents). Therefore, we illustrate the example of both classes single agent and multi-agent activities in Fig. 1 and 2, respectively.

Thumos-14 covers different types of sports actions performed by single or mostly multiple agents. Two sample actions; baseball pitch and soccer penalty are visualised in Fig. 3. In the figure, it can be seen that our model construct the seen graph of almost all agents despite the low-resolution videos.

ROAD. The activities in the road dataset are mostly performed by a large number of agents, however, there are a few cases where the number of detected agents is less in number as shown in Fig. 4 (above). On the other hand, in Fig. 4 (below) we also show an example of a night video where the activity is performed by a large number of agents moving very promptly.

B . Additional Experiments

B .1. Qualitative Results

We provide additional qualitative results of our method on all three datasets in Fig. 5. The figure shows two samples per dataset, and portrays a series of local scenes (snippets), skipping some for visualisation purposes, with superimposed the ground truth (in green) and the prediction of our model (in red).

[1] https://pytorch.org/docs/stable/generated/torch.nn.BCEWithLogitsLoss.html

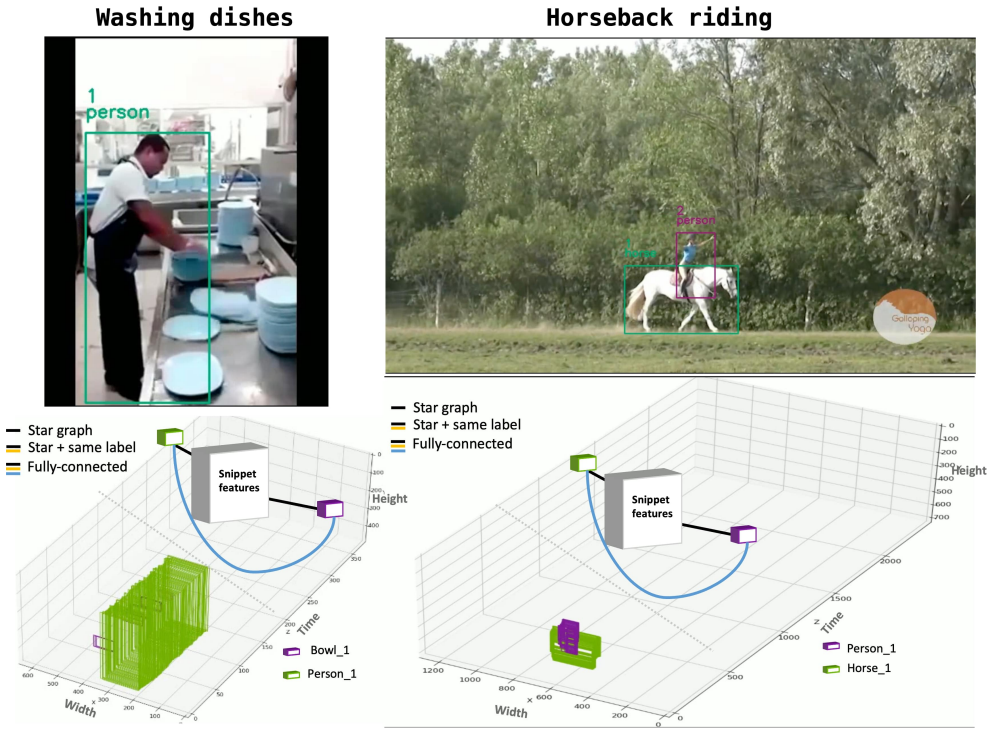

Figure 1. Visualisation of our agent detection and tracking stage using a bird's-eye view of the spatiotemporal volume corresponding to a video segment of *ActivityNet-1.3* dataset. In this figure, we present examples of activities performed by a single actor.

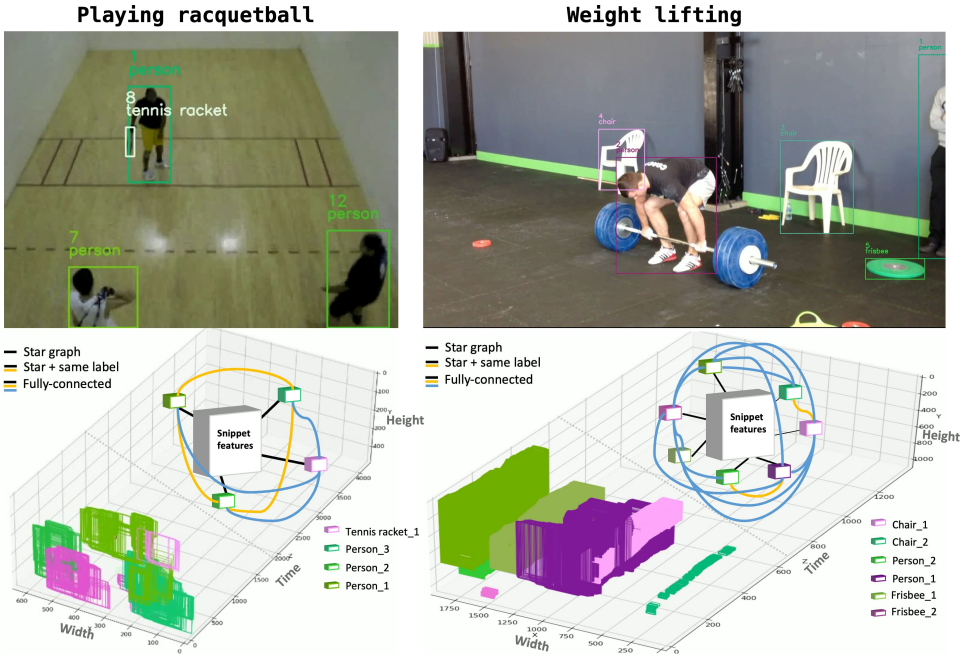

Figure 2. Visualisation of our agent detection and tracking stage using a bird's-eye view of the spatiotemporal volume corresponding to a video segment of *ActivityNet-1.3* dataset. In this figure, we show examples of activities performed by multiple actors (agent tubes).

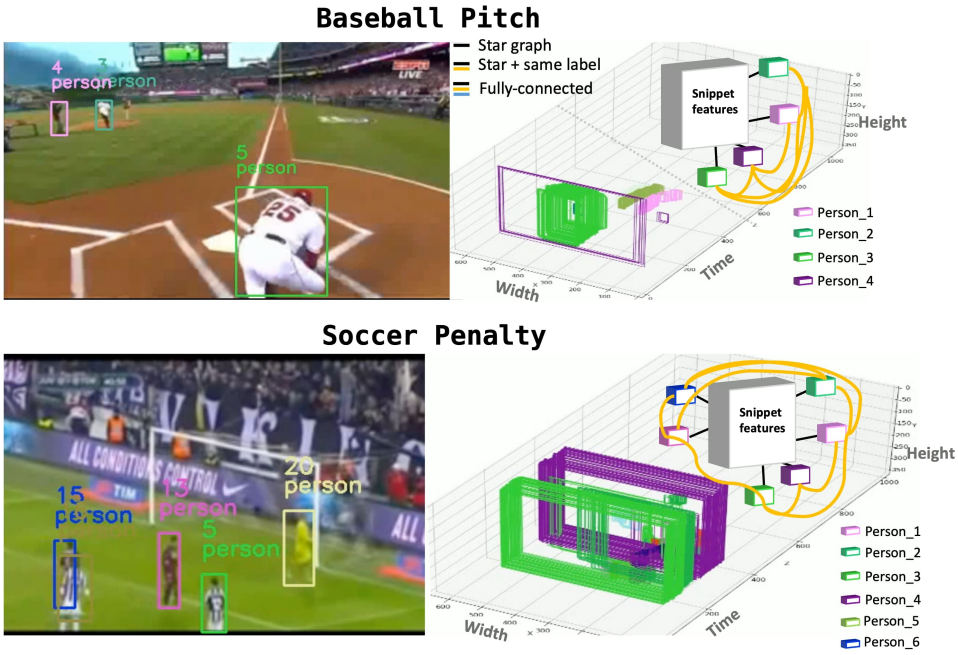

Figure 3. Visualisation of our agent detection and tracking stage using a bird's-eye view of the spatiotemporal volume corresponding to a video segment of *Thumos-14* dataset. Both of the examples show the activities performed by multiple agents (recognising sports activities).

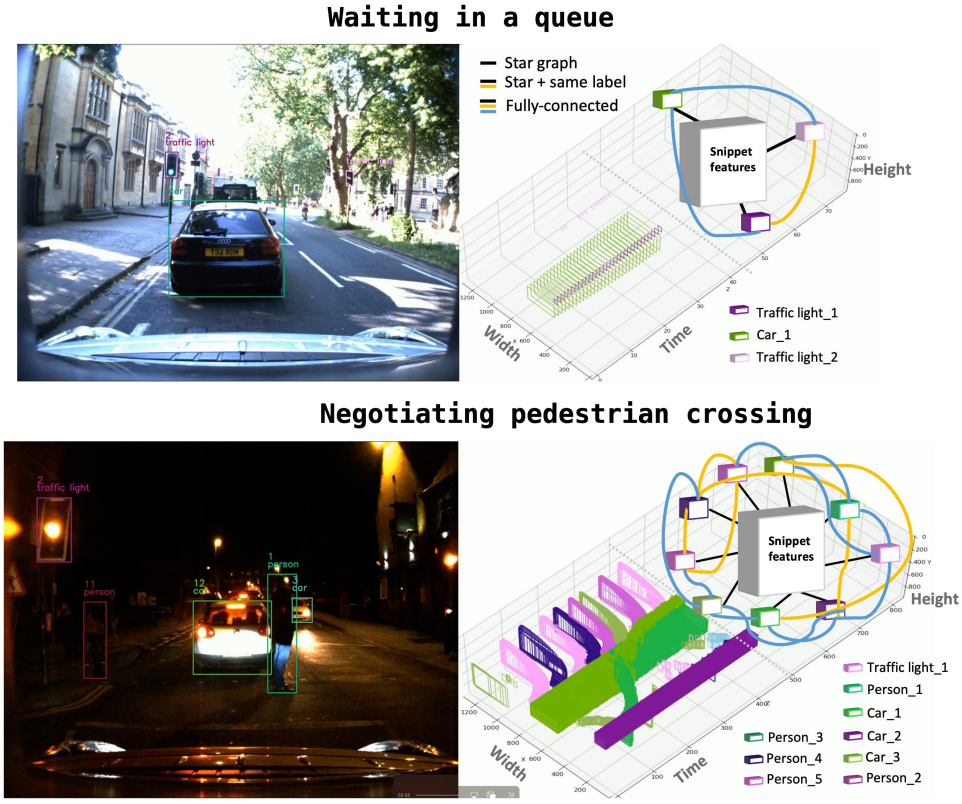

Figure 4. Visualisation of our agent detection and tracking stage using a bird's-eye view of the spatiotemporal volume corresponding to a video segment of the *ROAD* dataset. The upper section shows an example of activity performed by only three agents. Below is the example of a night video where the activity is performed by multiple agents.

Figure 5. The qualitative results of our proposed method for all three datasets. The green rectangles covering the snippets (local scenes) are the ground truth while the yellow boxes show the prediction of our model.