Discovery of Shared Semantic Spaces for Multi-Scene Video Query and Summarization: Supplementary Material

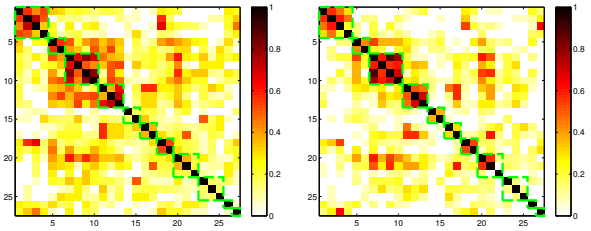
Xun Xu, Timothy M.Hospedales, and Shaogang Gong

I. Scene Alignment

In this section, we give further insight into scene alignment by analysing its quantitative impact on scene clustering and presenting some examples of matched scenes.

A. Quantitative Analysis of Scene Alignment

We demonstrate the impact of scene alignment by comparing the pairwise scene relatedness before and after alignment. Heat maps to visualise the relatedness (main manuscript Eq. (9)) between scenes are shown in Fig. 1. It is evident that after alignment, the affinity matrix becomes cleaner by increasing intra-cluster relatedness and reducing inter-cluster relatedness.



(a) Pairwise scene relatedness before alignment (b) Pairwise scene relatedness after alignment

Fig. 1: Pariwise scene relatedness before (a) and after (b) alignment. Scenes are in the order of Fig.7 in the manuscript. Green boxes indicate scene clusters.

We further investigated how the alignment affects scene clusters 3 and 7. We show the pairwise scene relatedness between 6 scenes across the two clusters in Table I, II and III.

The authors are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, UK.(e-mail: {xun.xu, t.hospedales, s.gong}@qmul.ac.uk)

TABLE I: Pairwise scene relatedness before alignment.

Pre-Alignment	Cluster 3				Cluster 7	
	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Scene 6
Scene 1	1.00	0.60	0.73	0.63	0.20	0.40
Scene 2	0.60	1.00	0.50	0.43	0.13	0.17
Scene 3	0.73	0.50	1.00	0.80	0.17	0.23
Scene 4	0.63	0.43	0.80	1.00	0.23	0.30
Scene 5	0.20	0.13	0.17	0.23	1.00	0.47
Scene 6	0.40	0.17	0.23	0.30	0.47	1.00

TABLE II: Pairwise scene relatedness after alignment.

Post-Alignment	Cluster 3				Cluster 7	
	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Scene 6
Scene 1	1.00	0.60	0.90	0.77	0.20	0.23
Scene 2	0.60	1.00	0.57	0.60	0.13	0.00
Scene 3	0.90	0.57	1.00	0.70	0.10	0.20
Scene 4	0.77	0.60	0.70	1.00	0.10	0.13
Scene 5	0.20	0.13	0.10	0.10	1.00	0.53
Scene 6	0.23	0.00	0.20	0.13	0.53	1.00

TABLE III: Relatedness difference between pre-alignment and post-alignment.

Difference	Cluster 3				Cluster 7	
	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5	Scene 6
Scene 1	0.00	0.00	0.17	0.13	0.00	-0.17
Scene 2	0.00	0.00	0.07	0.17	0.00	-0.17
Scene 3	0.17	0.07	0.00	-0.10	-0.07	-0.03
Scene 4	0.13	0.17	-0.10	0.00	-0.13	-0.17
Scene 5	0.00	0.00	-0.07	-0.13	0.00	0.07
Scene 6	-0.17	-0.17	-0.03	-0.17	0.07	0.00

It is evident from Table III that scene relatedness within clusters is mostly increased after alignment while relatedness across clusters is decreased. This initial scene alignment process enables the scene clustering in a later stage of the overall model design to be more meaningful. Examples of scene alignment within Scene Clusters 3, 7 and across clusters are shown in Fig. 2, 3 and 4.

II. Data Annotation

To investigate the consistency of our activity annotation ontology, we invited 8 independent annotators to annotate separately the data in scene clusters 3 and 7. These annotators are unaware of the modelling methodology, but were instructed to follow a annotation scheme shown in Fig. 5. Each scene was annotated 3 times in total by multiple annotators.

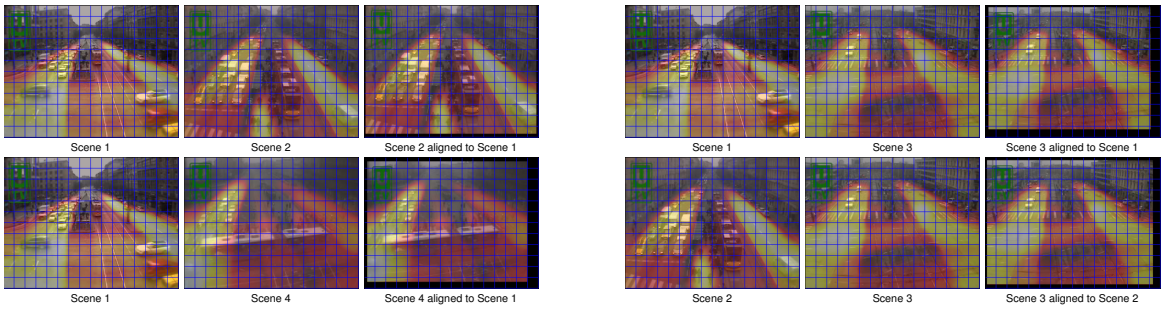


Fig. 2: Example scene alignment pairs within scene cluster 3. The overlapped heat map is the spatial frequency of visual words.

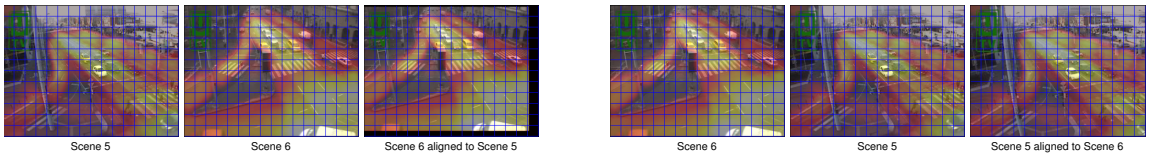


Fig. 3: Example scene alignment pairs within scene cluster 7. The overlapped heat map is the spatial frequency of visual words.

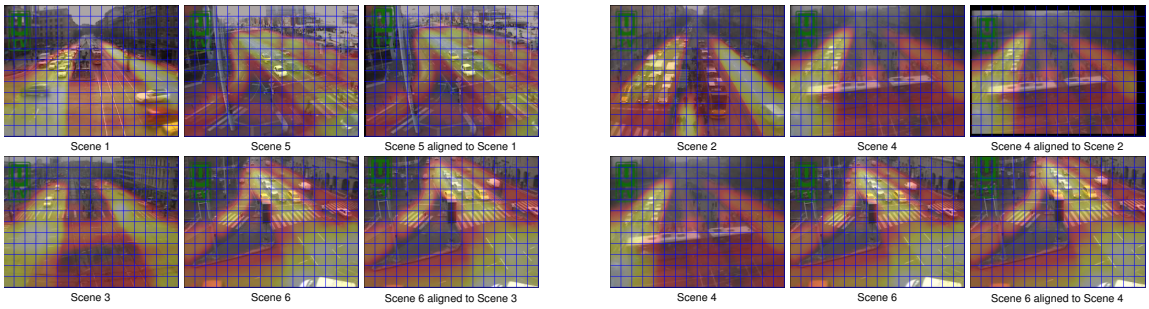
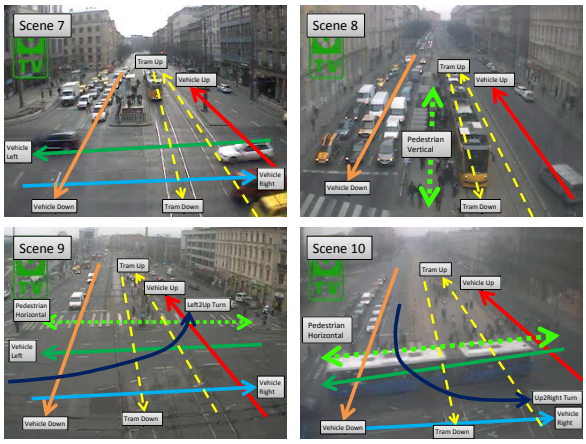


Fig. 4: Example scene alignment pairs across scene cluster 3 and 7. The overlapped heat map is the spatial frequency of visual words.

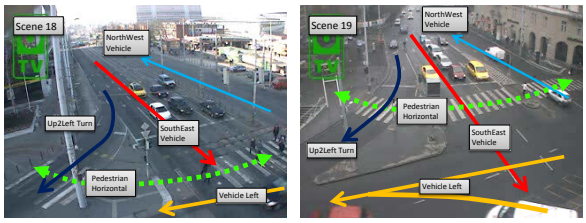
To quantitatively analyse the repeatability of the annotation scheme, we compared our reference annotation with the new independent annotations (additional annotations). Considering the reference annotation as a two dimension matrix $\{M^R(i,j)\}_{i=1\cdots672,j=1\cdots N_{act}}$ where i is the index for clip (we have 6 scenes each with 112 clips. So $112 \times 6 = 672$ clips in total) and j is the index for unique activities (defined in Table I in the main manuscript). Recall that we have three annotation schemes, original scheme with 19 unique activities, merge scheme 1 with 13 unique activities and merge scheme 2 with 10 unique activities. The additional annotations can be considered as a three dimensional matrix $\{M^A(i,j,k)\}_{i=1\cdots672,j=1\cdots N_{act},k=1\cdots N_{usr}}$ where k indicates the user who annotated the ith clip. We present quantitative measurements of annotation repeatability broken down by cluster and by activity: (1) We compute the Hamming distance between additional annotation and reference annotation for clip i and user k as:

$$d(i,k) = \frac{\sum_{j=1}^{N_{act}} \mathbb{1}(M^R(i,j) = M^A(i,j,k))}{N_{act}} \quad (1)$$

where $\mathbb{1}$ is the indicator function; (2) We calculate the agreement between the additional and reference annotations for each activity as how many binary activity annotations are consistent throughout all clips. The agreement



(a) Annotation scheme for Cluster 3



(b) Annotation scheme for Cluster 7

Fig. 5: The annotation guide provided for scenes in clusters 3 and 7.

for jth activity is defined as:

$$ag(j) = \frac{\sum_{k=1}^{N_{usr}} \sum_{i=1}^{672} \mathbb{1}(M^R(i,j) = M^A(i,j,k))}{N_{usr} \times 672} \quad (2)$$

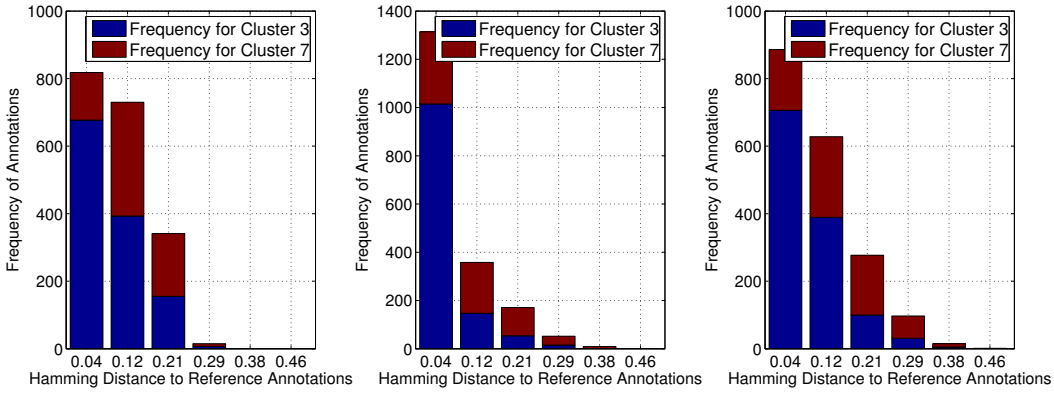
Fig. 6(a)-(c) show the differences between annotations by additional eight different annotators and the original reference annotation, measured by the Hamming distance (Eq. (1)). These are from three annotation schemes across all 6 scenes, including both within scene cluster 3 and within scene cluster 7. It is evident that the new additional annotations are fairly consistent with the original (reference) annotation, as most clips have less than 0.2 hamming distance from the reference. That is, all the additional annotations of all the activities in all the video clips have more than 80% in agreement on with reference annotation. We further show the breakdown in activity agreement for the three (fine to coarse) annotation schemes (see Table I in the main manuscript) in Fig.7. It can be seen that some activities' annotations by the new different annotators have relatively lower agreement (around 50%) with the reference annotation. Most of the activities with larger discrepancy in annotation are vehicle activities with either sparse or dense options. This may be due to that the sparsity of moving vehicles can be interpreted differently by different human annotators even if a quantitative criterion is given. By merging some dense and sparse tags, also left/right, up/down, southeast/northwest tags, the annotation agreement is increased notably to 65%. It is further noted that the annotation of pedestrian horizontal activity also has lower agreement because pedestrians are usually very small and move relatively slowly in the scene, making annotations hard to be consistent. The problem of annotation consistency and interpretation could be an interesting topic to further investigate in future work. A summary of average activity annotation agreement across three annotation schemes are given in Table IV. It is evident that the average agreement increases from fine to coarse annotation in Fig. 7(a)-(c). This supports our analysis.

TABLE IV: The average activity agreement.

	All 6 Scenes	Cluster 3	Cluster 7
Org. Annot.	90.0	91.9	85.8
Merge Sch. 1	91.0	92.7	87.0
Merge Sch. 2	91.9	93.5	88.2

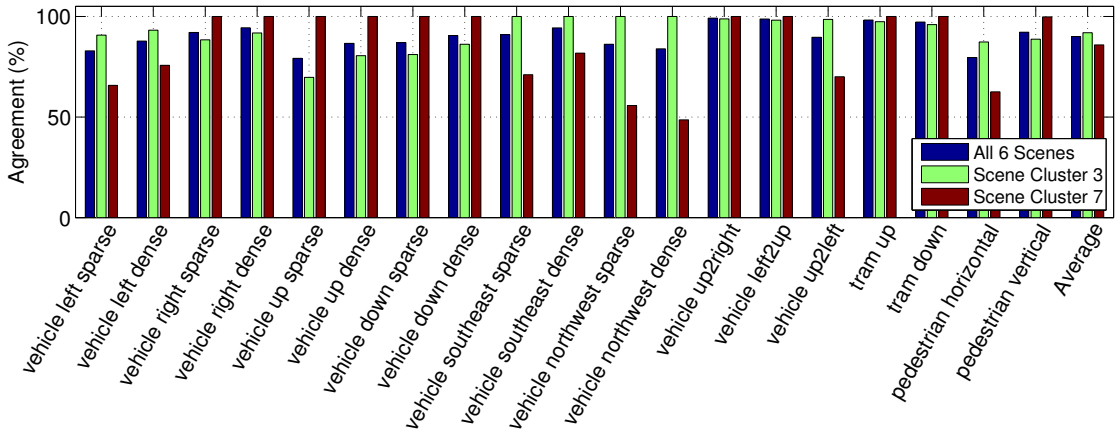
III. Multi-Scene Summarization

Finally, we illustrate multi-scene summarization qualitatively, using a smaller range of video clips to make visualisation manageable. We select randomly 32 clips of video from each scene in scene cluster 3, with 128 candidate video clips overall. Then we run the *Multi-Scene Kcenter* method on these 128 clips, setting the summary length 32. All of the candidate video clips are illustrated in Fig. 8 (a-d), while the resulting multi-scene summary is given in Fig. 8 (e). The behaviour label for each clip is displayed at the top of each image. It is evident from the motion overlays of each clip that the original videos (Fig. 8(a-d)) have extensive redundancy (many similar overlays), whereas the summary video (Fig. 8(e)) is much more salient and concise with a small set of clips exhibiting visually distinctive behaviours. More specifically, behaviour type 2, 6, 11, 15, 19, 20, 21, 22, 24, 25, 26, 27, 28, 29, 30, 31 exist in the candidate clips and all of these unique behaviours are discovered in the summary clips.

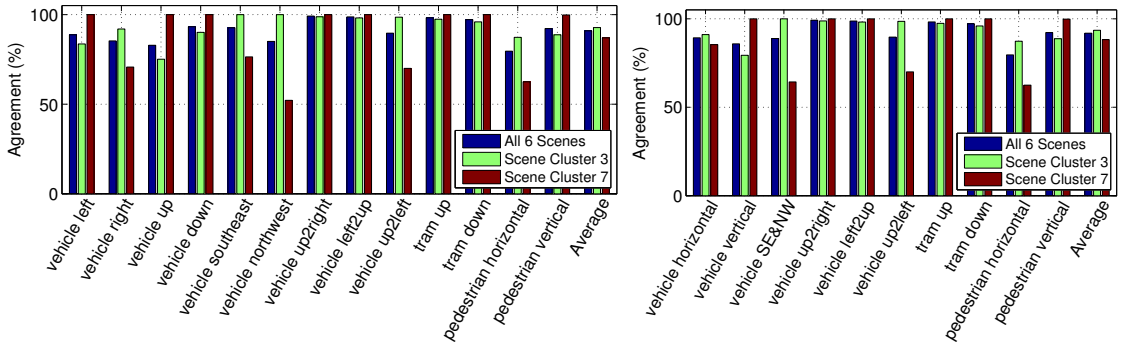


(a) Original Scheme (b) Merge Scheme 1 (c) Merge Scheme 2

Fig. 6: Histogram of Hamming distances between the additional and reference annotations.

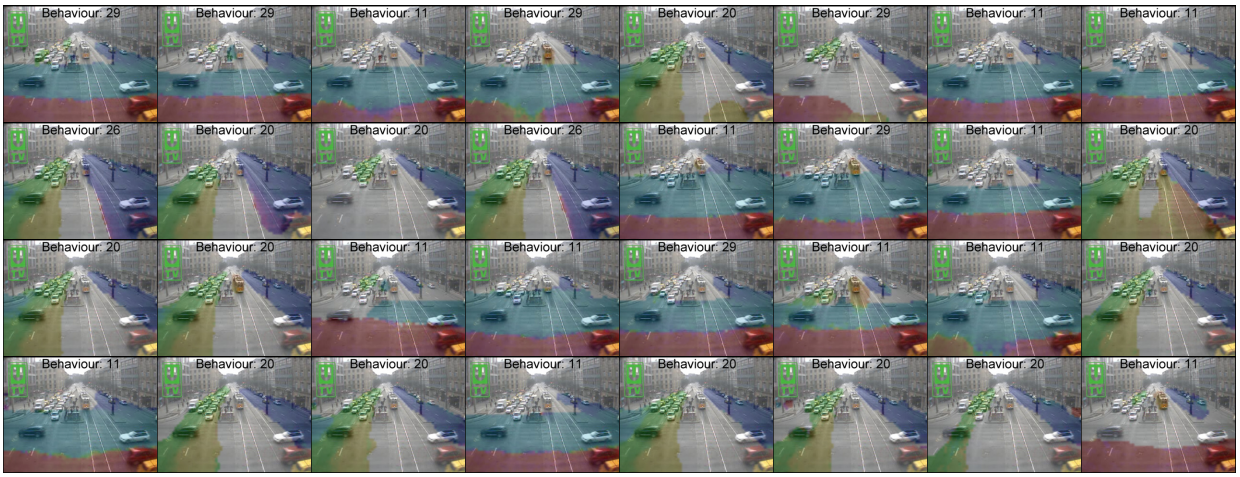


(a) Fine-grained 19-activity annotation

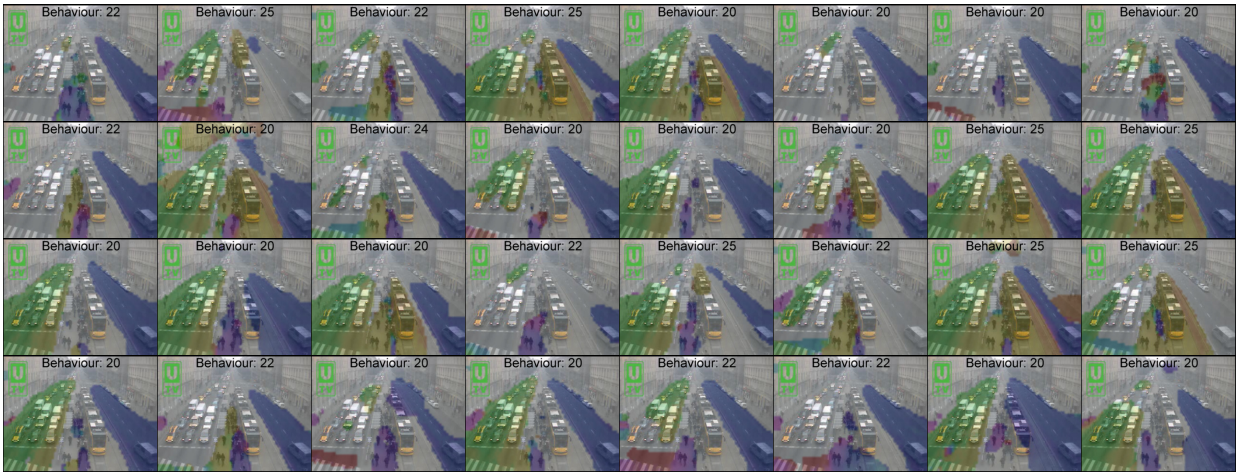


(b) 13 activity annotation (c) Coarse 10 activity annotation

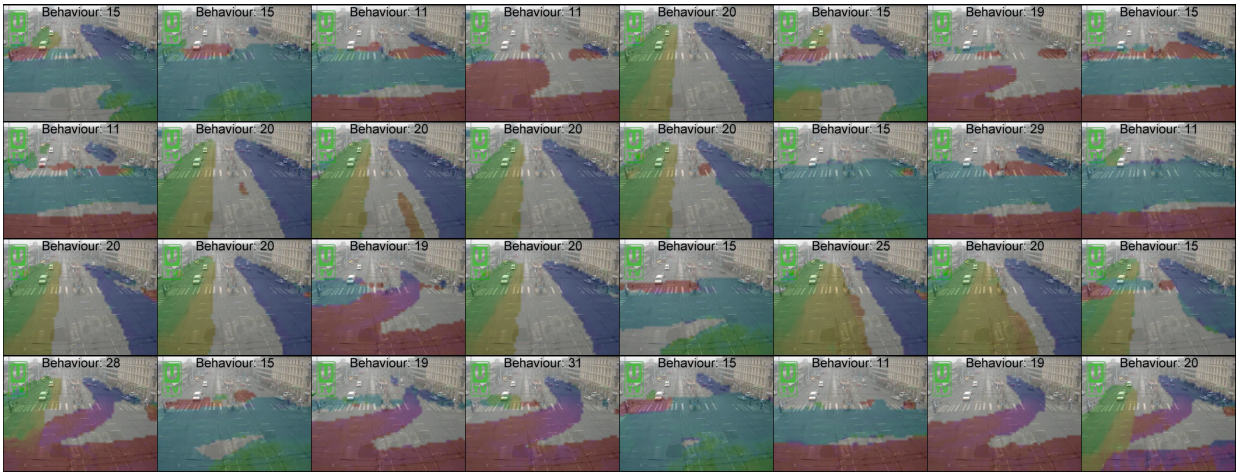
Fig. 7: Activity annotation consistency broken down by activity type. Agreement between reference and additional annotations.



(a) 32 video candidate video clips randomly selected from scene 7 cluster 3

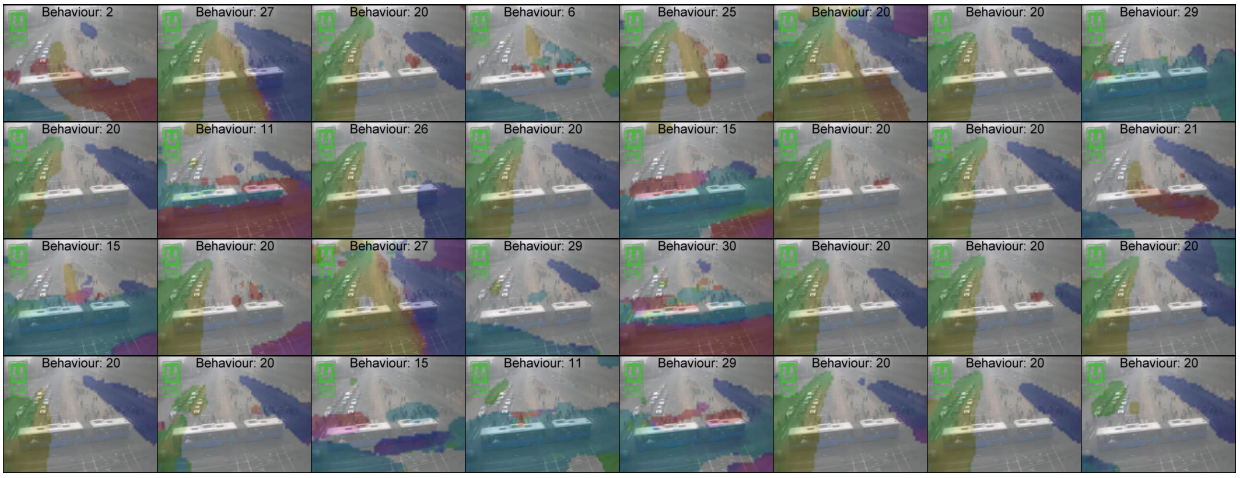


(b) 32 video candidate video clips randomly selected from scene 8 cluster 3

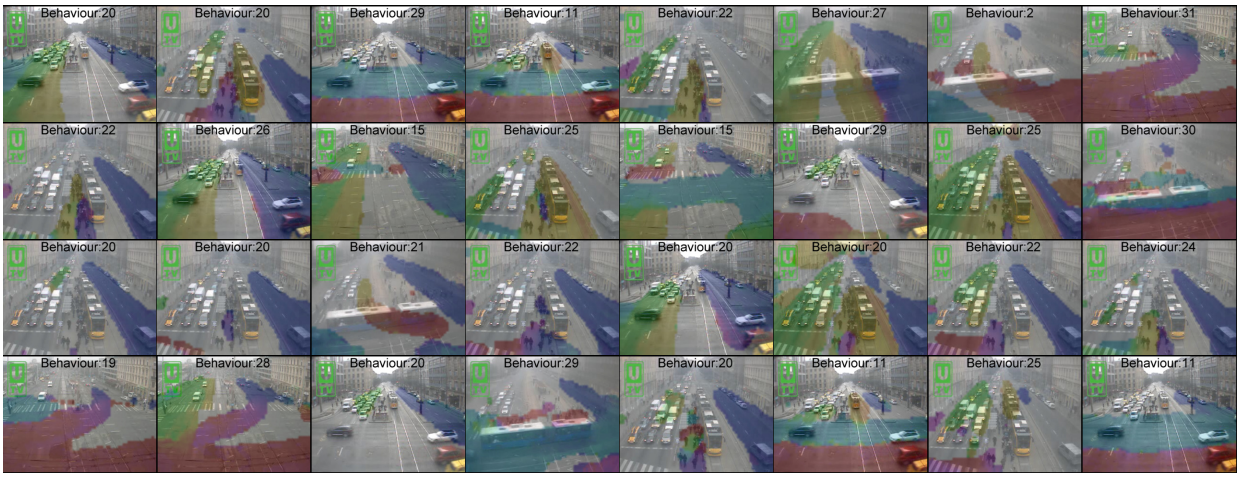


(c) 32 video candidate video clips randomly selected from scene 9 cluster 3

Fig. 8: Histogram of agreement between additional and reference annotations



(d) 32 video candidate video clips randomly selected from scene 10 cluster 3



(e) Multi-scene summary clips.

Fig. 8: (a)-(d)A total of 128 video clips are randomly selected as candidate video clips cropped from scene cluster 3. The behaviour category is marked in the top of each frame. (e) The 32 summary clips selected from 128 candidate video clips by our framework.