Appendix

Appendix A: Runtime Performance

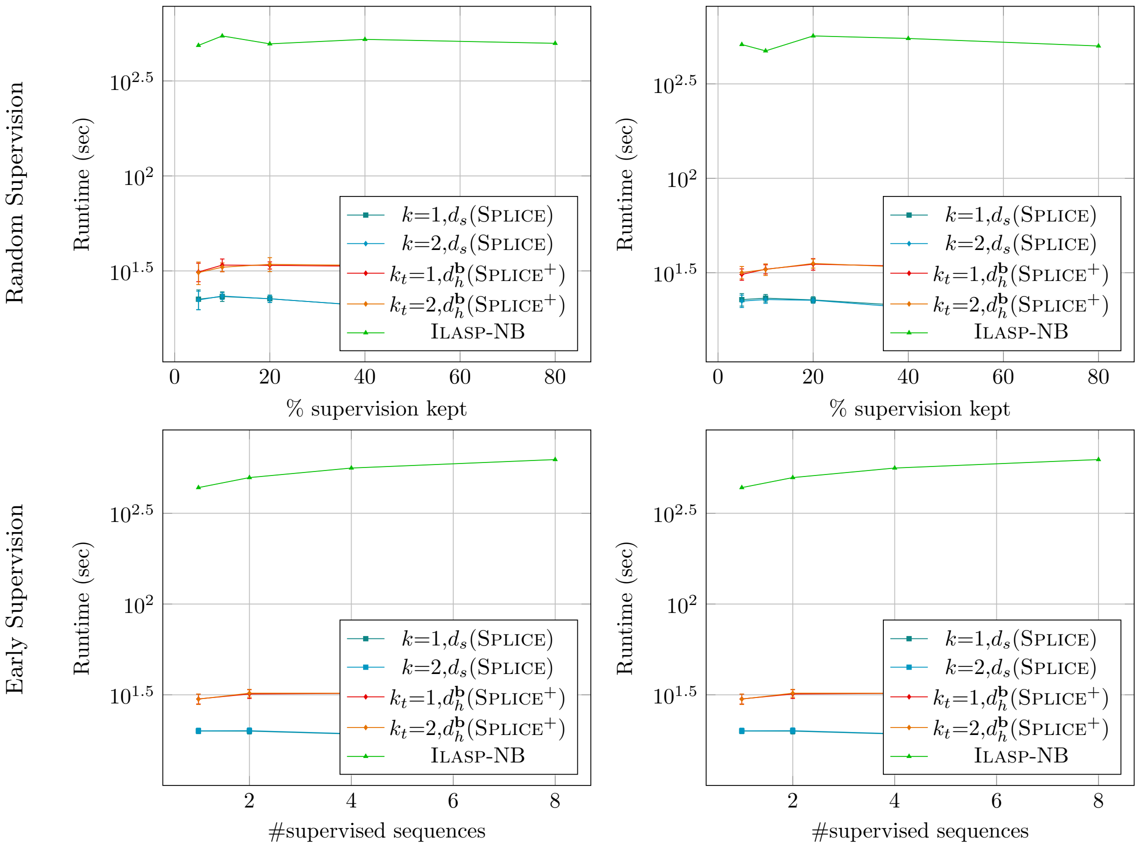

Figure 1: Runtime of supervision completion on `meet` (left) and `move` (right) as supervision increases. The runtime is macro-averaged over all samples.

Figure 1 presents the runtime performance of Splice$^+$ against Splice and ILASP-NB in log scale. As discussed in the experimental evaluation, Splice$^+$ is slower than Splice due to the tree updates and feature subset selection. In addition, note that ILASP-NB is much slower than Splice$^+$, since it is a batch learning algorithm that needs to operate over all data at once and may perform many iterations over the data to converge.

Figure 2 depicts the runtime cost for the maritime monitoring dataset. The computational penalty is is typically below 25 seconds. The variation of runtime cost between different supervision levels, due to feature selection, is much more apparent in the random supervision setting of the `rendezvous` CE, where we observe a 6 seconds increase between 5% and 20% supervision, due to frequent feature selection. Due to the very high runtime cost of ILASP-NB, we have omitted the results from the figures.

Finally, Figure 3 presents the runtime performance on the fleet management dataset in log scale. The runtime of Splice and Splice$^+$ is comparable to activity recognition and maritime monitoring, while ILASP-NB is faster than in the activity recognition still

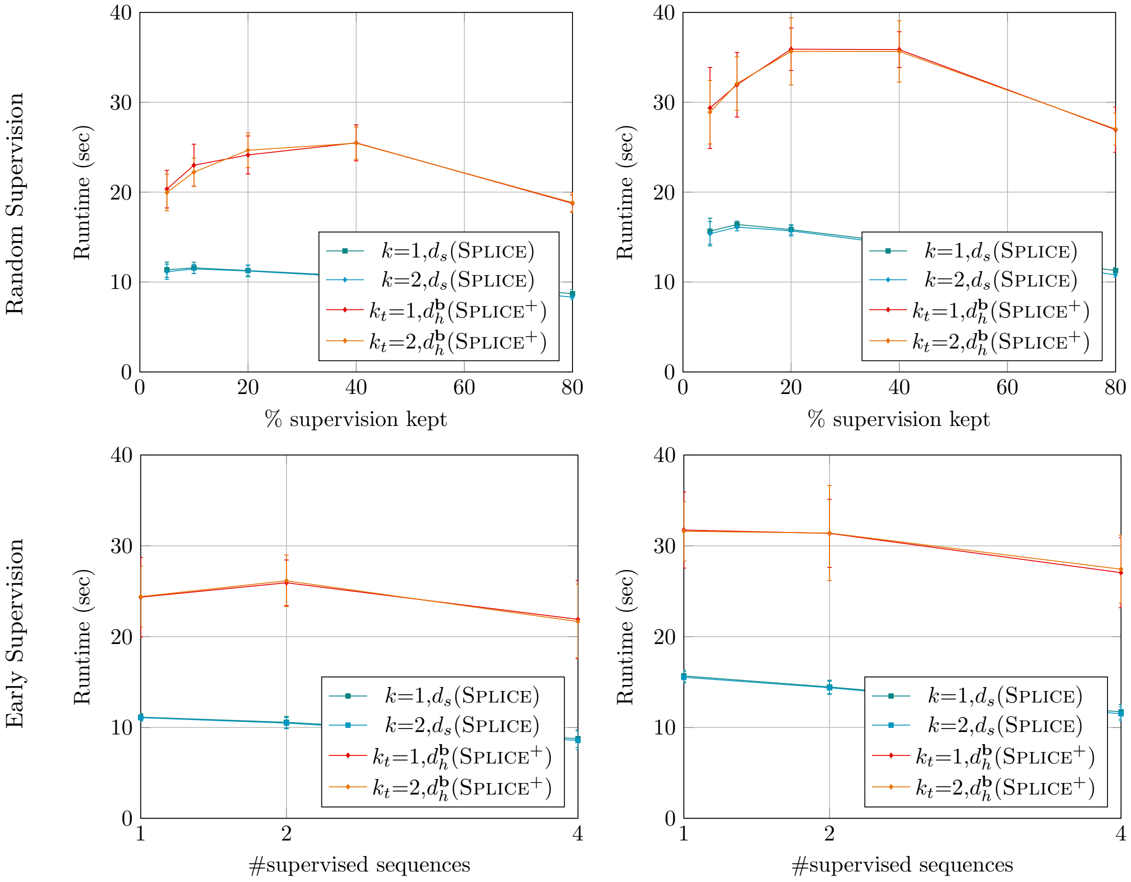

Figure 2: Runtime of supervision completion on `pilotOps` (left) and `rendezvous` (right) as supervision increases. The runtime is macro-averaged over all samples.

remains 3 times slower than Splice$^+$.

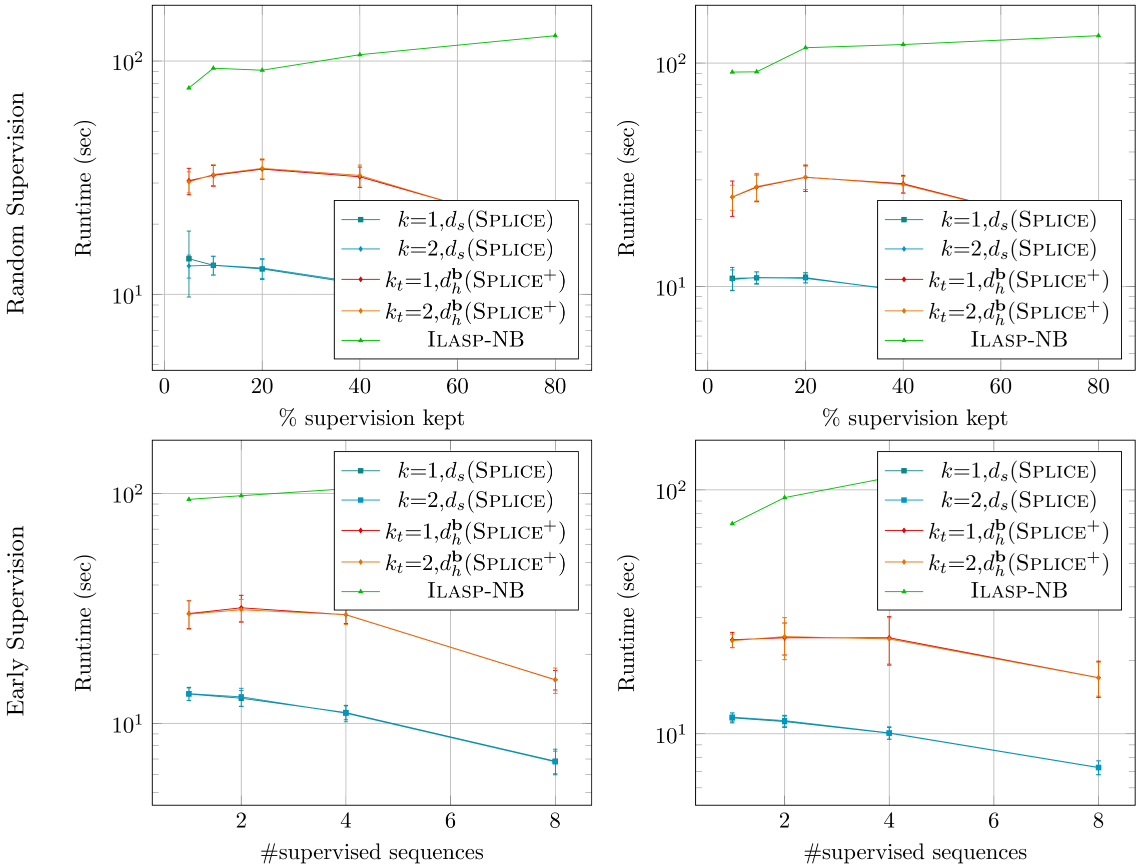

Figure 3: Runtime of supervision completion on `pilotOps` (left) and `rendezvous` (right) as supervision increases. The runtime is macro-averaged over all samples.

Appendix B: Batch Sizes

CE	Batch size	Number of supervised sequences			
		1	2	4	8
`meet`	10	0.44/0.69	0.59/0.78	0.73/0.78	0.78/0.93
	25	0.43/0.69	0.57/0.74	0.72/0.78	0.78/0.93
	50	0.42/0.69	0.51/0.77	0.67/0.77	0.77/0.93
	100	0.42/0.69	0.56/0.76	0.75/0.80	0.77/0.93
`move`	10	0.66/0.73	0.73/0.75	0.71/0.79	0.84/0.94
	25	0.66/0.73	0.74/0.74	0.72/0.79	0.84/0.94
	50	0.66/0.73	0.74/0.78	0.74/0.81	0.84/0.94
	100	0.66/0.73	0.73/0.75	0.73/0.80	0.84/0.94

Table 1: F_1-score as batch size increases for `meet` and `move` CEs: Splice/Splice$^+$.

Table 1 presents the change in performance as the batch size increases on the activity recognition dataset. The F_1-score of Splice tends to fluctuate more than that of Splice$^+$as the batch size increases. For instance, in the `meet` CE, when 2 or 4 supervised sequences are provided, the F_1-score of Splice varies from 0.01 to 0.08, while Splice$^+$ varies from 0.01 to 0.04. Corresponding changes are also noticeable when 1 or 8 supervised sequences are given to Splice, while Splice$^+$ does not vary at all in these cases. Such variations also appear at a much smaller scale in `move`. These results suggest that Splice$^+$ seems to be more robust to different batch sizes than its predecessor.

Appendix C: Ablation Study

CE	Distance	Random Supervision		Early Supervision	
		5%	10%	1	2
`meet`	d_s	0.62	0.70	0.56	0.69
	$d_s^{\mathbf{b}}$	0.59	0.70	0.64	0.71
	m	0.65	0.75	0.64	0.70
	d_h	0.63	0.76	0.65	0.73
	$d_h^{\mathbf{b}}$	**0.67**	**0.77**	**0.70**	**0.76**
`move`	d_s	0.57	0.64	0.67	0.71
	$d_s^{\mathbf{b}}$	**0.58**	0.67	0.69	0.71
	m	0.56	0.68	0.60	0.54
	d_h	0.57	**0.69**	0.70	0.73
	$d_h^{\mathbf{b}}$	**0.58**	**0.69**	**0.73**	**0.75**

Table 2: Comparison of Splice$^+$ on `meet` and `move` using the simple structural distance (d_s) and the hybrid distance ($d_h^{\mathbf{b}}$).

In order to further examine the contribution of each of the proposed improvements over

Splice, in Table 2 we present an ablation study on the activity recognition dataset, to compare the different components of Splice$^+$ against each other. Since, in both scenarios Splice$^+$ performs better using k_t=1 instead of k_t=2 we only use k_t=1 here. The first important observation is that the structural distance (d_s) using temporal kNN performs very well, which indicates the importance of the temporal connectivity. Mass-based dissimilarity alone ($\tilde{m}$) performs well enough in `meet` but rather poorly in `move`, especially in the early supervision setting. However, when combined with the structural distance (d_h) it always performs better than the structural distance alone. Another interesting observation is that while feature selection on the structural distance ($d_s^{\mathbf{b}}$) does not always achieve better results than the structural distance alone (d_s), it yields a synergistic effect when combined with the mass-based dissimilarity ($d_h^{\mathbf{b}}$). For instance, in the random supervision setting, using $d_h^{\mathbf{b}}$ instead of d_h increases F$_1$-score from 0.63 to 0.67 for `meet` (corresponding to 242 errors on average), while on the early supervision setting, it increases F$_1$-score from 0.65 to 0.7 for `meet` and 0.7 to 0.73 for `move` (corresponding to 162 and 136 errors respectively). Therefore, although not by a large margin, the proposed hybrid measure achieves the best performance.

CE	Distance	Random Supervision		Early Supervision	
		5%	10%	1	2
`rendezvous`	d_s	0.59	0.70	0.69	0.79
	$d_s^{\mathbf{b}}$	0.59	0.70	0.69	0.79
	m	0.53	0.65	0.53	0.53
	d_h	**0.62**	**0.75**	**0.74**	**0.81**
	$d_h^{\mathbf{b}}$	**0.62**	**0.75**	**0.74**	**0.81**
`pilotOps`	d_s	0.47	0.63	0.78	0.90
	$d_s^{\mathbf{b}}$	0.47	0.63	0.78	0.90
	m	**0.56**	**0.69**	0.94	0.94
	d_h	**0.56**	**0.69**	**0.95**	**0.96**
	$d_h^{\mathbf{b}}$	**0.56**	**0.69**	**0.95**	**0.96**

Table 3: Comparison of Splice$^+$ on `pilotOps` and `rendezvous` using the simple structural distance (d_s) and the hybrid distance ($d_h^{\mathbf{b}}$).

Similar to the task of human activity recognition, we analyse in the maritime monitoring dataset, the contribution of each of the proposed improvements over Splice. In Table 3 we compare the different components of Splice$^+$ against each other. Again we present results for k_t=1, as it yields the best performance. Note that the structural distance (d_s) alone, using temporal kNN performs very well, which indicates the importance of temporal connectivity. Mass-based dissimilarity alone ($\tilde{m}$) seems to perform well in the `pilotOps` CE, but yields very poor performance in the `rendezvous` CE. Similar to activity recognition, when combined with the structural distance (d_h) leads to a small, but consistent improvement of the performance. In contrast to activity recognition, feature selection on the structural distance ($d_s^{\mathbf{b}}$) does not seem to improve the performance further. This is due to the synthetic supervision of the maritime dataset, which leads to noise-free labels and features. Moreover, there are three irrelevant features in the dataset but only one of them appears frequently in the positive examples of the CEs, which renders the measurements of d_s quite similar to $d_s^{\mathbf{b}}$. Feature selection in this context is not expected to add value.

Finally, note that the results in the fleet management dataset do not provide any insight in the discussion, since there are no irrelevant or noisy features in the dataset, leading to very similar distance measurements. As a general conclusion from the ablation studies that we performed in all datasets, the use of SPLICE$^+$ with all of its features, seems to guarantee the best performance, irrespective of the supervision setting.