We thank reviewers for the valuable comments. Please see below for our responses to specific comments.

R1: About the termination network. Only using a switch network is not reliable as switching to another tracker does not necessarily improve the tracking. Thus we tend to keep using a certain tracker as much as possible. Our termination network makes the decision module avoid oscillating between the two trackers especially when they have similar accuracy. In the training phrase, the tracker that actually performs well could suffer from an improper termination due to the probability indicating its tracking performance. However, our termination scheme enforces the agent to explore more states that would not have been selected. We also observe that the probability is typically close to either 0 or 1 at the later stage of training, which means that improper termination hardly occurs at that stage.

R1: Extend to more trackers? Yes, our framework can be easily extended to more trackers. For instance, the results of using 3 trackers including ACT, FCT and SiamFC are shown in the second row of the table below. Considering both accuracy and efficiency, we use two trackers in the proposed DTNet.

Method	OTB2015		TrackingNet		UAV-123		LaSOT		VOT18			Speed (fps)
	AUC	Prec.	AUC	Prec.	AUC	Prec.	AUC	Prec.	Accuracy	Robustness	EAO	
DTNet (FCT+SiamFC)	0.660	0.891	0.610	0.583	0.533	0.731	0.360	0.341	0.518	0.277	0.300	36
DTNet (ACT+FCT+SiamFC)	0.665	0.893	0.621	0.585	0.539	0.726	0.364	0.342	0.521	0.287	0.298	23
DTNet (ATOM+SiamRPN++)	0.701	0.916	0.737	0.698	0.649	0.831	0.516	0.579	0.604	0.197	0.418	27
SiamRPN++	0.696	0.914	0.733	0.694	0.613	0.807	0.496	0.569	0.600	0.234	0.414	35
ATOM	0.661	0.867	0.703	0.648	0.642	0.825	0.515	0.576	0.590	0.204	0.401	30
DiMP	0.660	0.859	0.723	0.666	0.643	0.821	0.532	0.581	0.594	0.182	0.402	57
Manually designed rule-based	0.544	0.716	0.416	0.402	0.453	0.618	0.283	0.302	0.470	0.682	0.158	19

R1: Tracking affected by previous frames? No, the tracking delivered by the DTNet is not affected by the tracking results of previous frames. The benefit is that the tracker is not influenced by inaccurate tracking on previous frames.

R2: Comparison with latest SOTA trackers. As suggested, we have compared our method with the latest SOTA trackers such as SiamRPN++, ATOM and DiMP. Specifically, we replace the baseline trackers FCT and SiamFC with SiamRPN++ and ATOM, and perform our decision module to make them compete with each other. The table above shows that our DTNet still improves both SiamRPN++ and ATOM in all datasets. This is because in our method, two baseline trackers could work alternatively to conduct tracking within different scenes that they are adept in. Such results indicate that combining different trackers based on an intelligent switching scheme is superior over a single tracker even if it is the SOTA which already integrates the advantages of both template and detection based trackers.

R2: Evaluation on other datasets. As shown in the table above, we have evaluated our DTNet with different baseline trackers on other datasets suggested by the reviewers including OTB2015, TrackingNet, UAV-123, LaSOT and VOT18. It can be seen that our method achieves consistent improvement over various datasets benefitting from the proposed decision module which could select different types of trackers for handling different scenes.

R3: Motivation. We agree that combining different kinds of clues is expected to make the tracker stronger. In fact, what we mean here is that such a fusion manner might not be the best choice. In this work, instead of fusing different types of trackers into one, we advocate an intelligent switching strategy to make them coexist and compete with each other for different scenes. To the best of our knowledge, this has never been explored before. The results in Table 2 of the paper show that the proposed strategy can utilize the advantages of different types of trackers and produce significant gains. Moreover, even with the two similar fusion-based trackers such as SiamRPN++ and ATOM, our method still makes improvement as shown in the table above. This also shows the potential of our method in more general cases.

R2: About the FCT tracker. Please note that the FCT tracker is not the primary contribution of our work. Our main contribution lies in the novel decision module which automatically selects a tracker to handle different scenes as recognized by R1 and R4. Yes, FCT is extended from MDNet. Compared to MDNet, the proposed FCT uses pixel-level classification and regression which does not require the expensive proposal generation. We prefer such a proposal-free tracker as it does not affect the efficiency of the whole ensemble framework much.

R3: Compare with manually designed rule-based decision module. We have included the manually designed rule-based decision module for comparison. It is implemented by picking a particular tracker based on the confidence score of tracking subject to the thresholds set manually. The results are given in the bottom row of the table above. Apparently, our automated decision module significantly outperforms such a handcrafted one which relies on handcrafted thresholds for tracker selection. Besides, our method is more efficient as it only performs each tracker once in the decision-making process while the handcrafted module has to carry out both trackers and use their output confidence scores for decision.

R4: About update scheme of the decision module. The actor of the update scheme defined by Eq. 9 outputs a probability to indicate how possible the tracker should be terminated. The critic defined by Eq.8 is then updated subject to this termination probability when evaluating the value of a state-option pair. And the actor still updates parameters in proportion to the action-value gradient of the critic.

R4: Inconsistent symbols in Eq. 7 and line 173. Thx. We have made corrections to avoid using inconsistent symbols.