Supplemental Material

1 Introduction

In this document we present additional results and analysis of our framework, including implementation details as well as quantitative and qualitative comparisons. We also provide a short video in **eccv2020_2107.mp4** to demonstrate the qualitative results on the Imagenet VID 2015 dataset [6]. For viewing purposes, we show the video at 15 fps although the original video is captured at 30 fps.

2 Implementation Details

We implement our framework with PyTorch. The models are trained on TITAN X GPUs, and the inference speed is reported on an Intel Xeon E5 CPU. When training the single-image detectors, we observe that adding class weights for the imbalance issue helps improve the accuracy on several classes but decreases the overall mAP. Hence, we randomly sample training images from the VID [6], DET [6], and COCO [2] datasets without setting particular class weights. The hyper-parameter μ for the heuristic keyframe policy is set to normalize the tracking score s. For correlation-based trackers like KCF [1], s is usually within the range of $(0, 1)$. Therefore, we set $\mu = 1$ in our experiments. Detailed architecture of our LSTM module is shown in Fig. 1, where all internal layers have 32 channels.

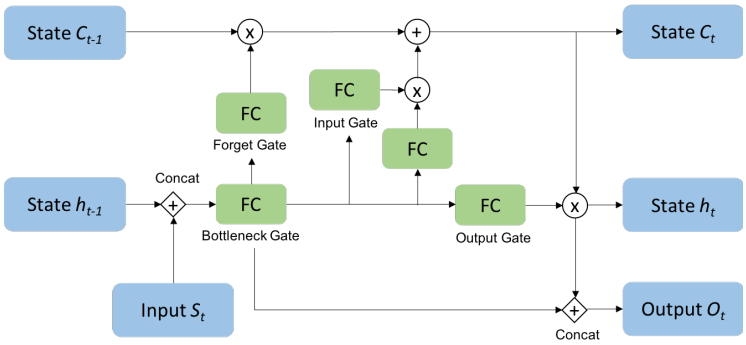


Fig. 1: Detailed architecture of our Bottleneck-LSTM module.

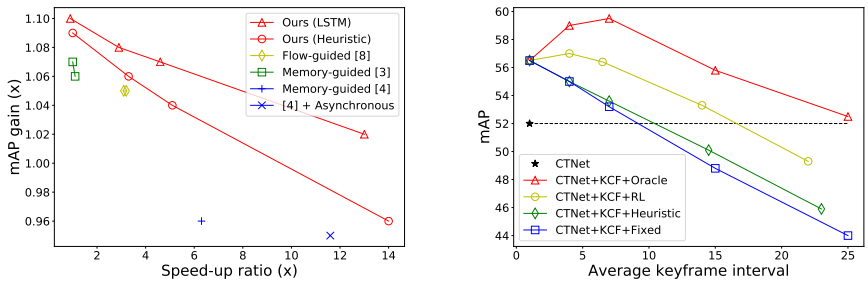
To pre-train the RL policy, we assign a ground truth action in different cases given the oracle keyframes. If no keyframe should be triggered within the current

interval D_t, then $a_t = 2$ (fix interval) is encouraged. If an oracle keyframe is within D_t but not within $D_t/2$, then ground truth action is $a_t = 1$ (shorten interval). Otherwise, we encourage the policy to take $a_t = 0$ (detect instantly).

3 Quantitative Comparisons

3.1 Using a Different Single-image Detector

In Fig. 2 we show the validation results of our framework with CenterNet [7] + KCF. We visualize the comparison with previous approaches [8,3,4] in (a), and the speed-accuracy tradeoff of different keyframe policies in (b). Similar to using YOLOv3 [5] as the single-image detector, our framework with CenterNet generally has a better tradeoff between mAP gain and speed-up ratio. Interestingly, the oracle keyframe policy achieves higher mAP when applying the detection model sparsely. It indicates that simply tracking the objects can produce higher accuracy than applying detection in certain situations. The observation exemplifies the importance of keyframe selection.

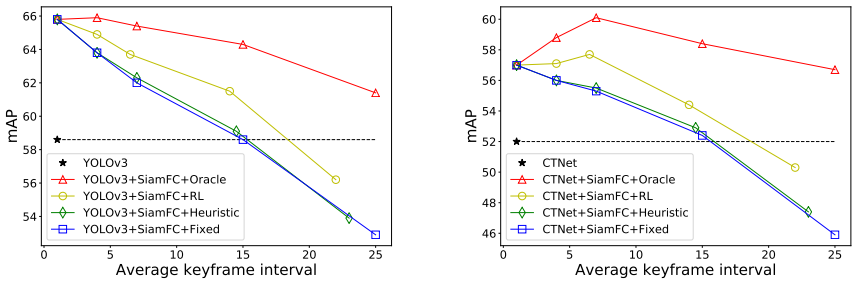


(a) (b)

Fig. 2: Quantitative comparisons on the Imagenet VID validation set. In (a) we plot the relative mAP gain versus speed-up ratio compared to the single-image baselines. The performance of different keyframe policies are shown in (b).

3.2 Using a Different Object Tracker

We also plot the tradeoff curves using SiamFC trackers in Fig. 3. With the deep trackers, the accuracy drops slower as the keyframe interval increases. However, the accuracy gain by the heuristic scheduler is barely noticeable since the score s provided by CNN filtering does not directly correspond to the tracking quality. Our RL policy still produces a significant performance gain by leveraging additional features such as detection confidence and box size variation.



(a) (b)

Fig. 3: Quantitative comparisons of different keyframe schedulers using (a) YOLOv3 + SiamFC and (b) CenterNet + SiamFC. All models are run with LSTM-based temporal aggregation.

3.3 Analysis of Different Classes

The detailed accuracy of the 30 VID classes are listed in Table 1. Our method with $D_{base} = 1$ produces the highest mAP on most classes. For the object classes with slower motions, *e.g.*bears, lizards, and watercrafts, a larger detection interval results in higher accuracy. The classes like antelopes, birds, and lions are often misclassified as other classes by the detection model. Our temporal aggregation tends to filter out the correct classification on these classes, leading to a lower mAP.

Table 1: Detailed accuracy of the 30 classes in the VID dataset. We report the mAP scores of CenterNet as the single-image baseline, and KCF trackers + LSTM-based temporal aggregation for our methods.

	Air plane	Antelope	Bear	Bicycle	Bird	Bus	Car	Cattle	Dog	Domestic cat	Elephant	Fox	Giant panda	Hamster	Horse
Single-image	77.4	**78.7**	55.4	60.1	**63.5**	56.1	41.8	37.3	35.2	38.7	58.1	60.3	62.7	79.3	60.4
Ours ($D_{base} = 1$)	**81.0**	69.5	68.0	**67.1**	54.1	**69.6**	**51.0**	**38.3**	**39.8**	**50.2**	**63.1**	**65.9**	70.1	**89.5**	**65.1**
Ours ($D_{base} = 7$)	79.2	70.3	**69.8**	63.4	56.9	66.9	49.1	37.9	39.4	49.5	62.5	62.6	**71.2**	88.7	63.3

	Lion	Lizard	Monkey	Motorcycle	Rabbit	Red panda	Sheep	Snake	Squirrel	Tiger	Train	Turtle	Watercraft	Whale	Zebra
Single-image	**14.0**	59.6	23.3	**70.2**	**39.7**	**36.9**	25.6	27.9	23.1	68.2	67.9	54.8	50.9	45.3	83.8
Ours ($D_{base} = 1$)	6.3	54.6	**34.7**	68.7	37.8	33.9	31.0	**42.6**	21.6	**82.6**	**71.2**	**55.5**	57.7	**57.6**	**87.0**
Ours ($D_{base} = 7$)	13.0	**63.3**	31.3	63.6	38.5	35.7	**32.4**	36.4	**24.7**	73.2	70.4	50.9	**61.3**	51.3	83.8

4 Qualitative Results

In Fig. 4-8 we show additional qualitative comparisons. Each detected object is labeled by a colored box with its class name, detection confidence, and tracking score on the top. The keyframes are marked with red outlines. From top to bottom, we show the results of a) single-image baseline, b) temporal aggregation at $D_{base} = 1$, c) fixed keyframe interval at $D_{base} = 15$, and d) adaptive keyframe scheduling at $D_{base} = 15$. All of our models adopt CenterNet detectors, KCF trackers, LSTM-based temporal aggregation, and RL keyframe scheduler.

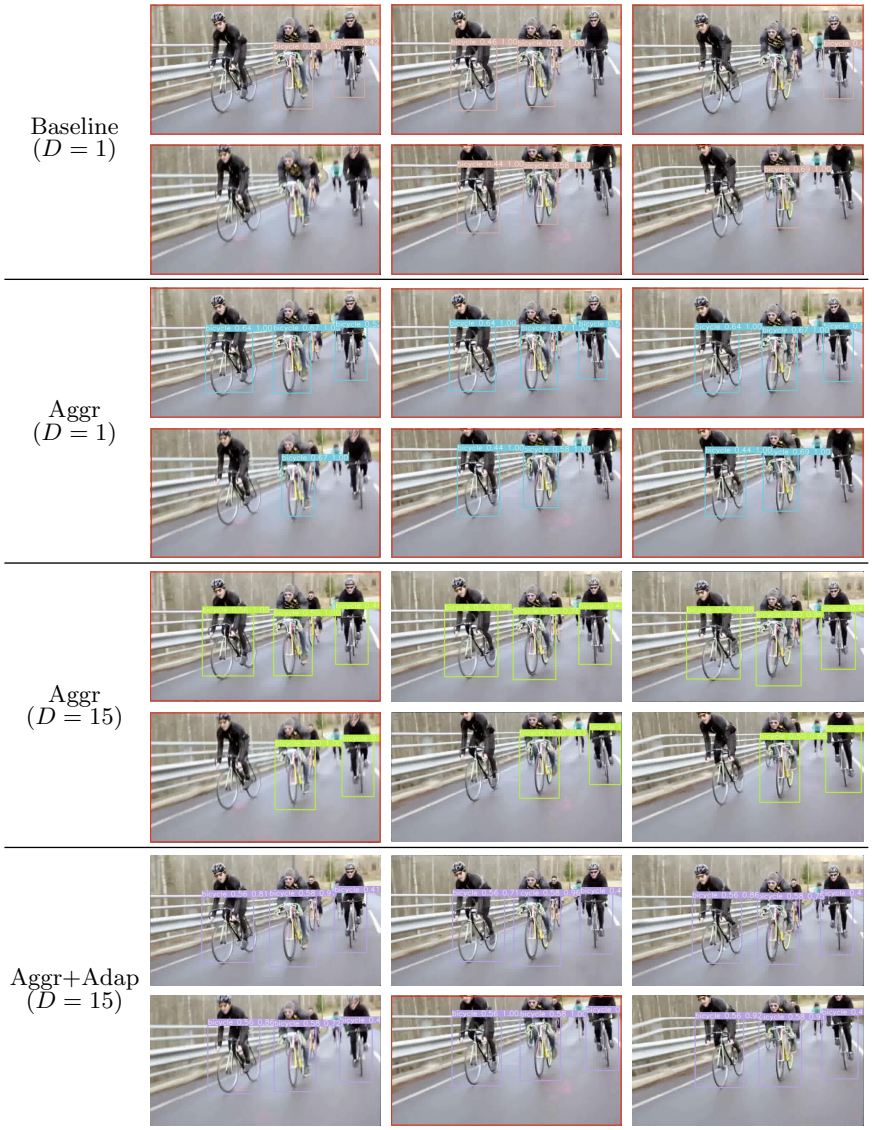


Fig. 4: Qualitative comparisons of our methods with single-image baseline.

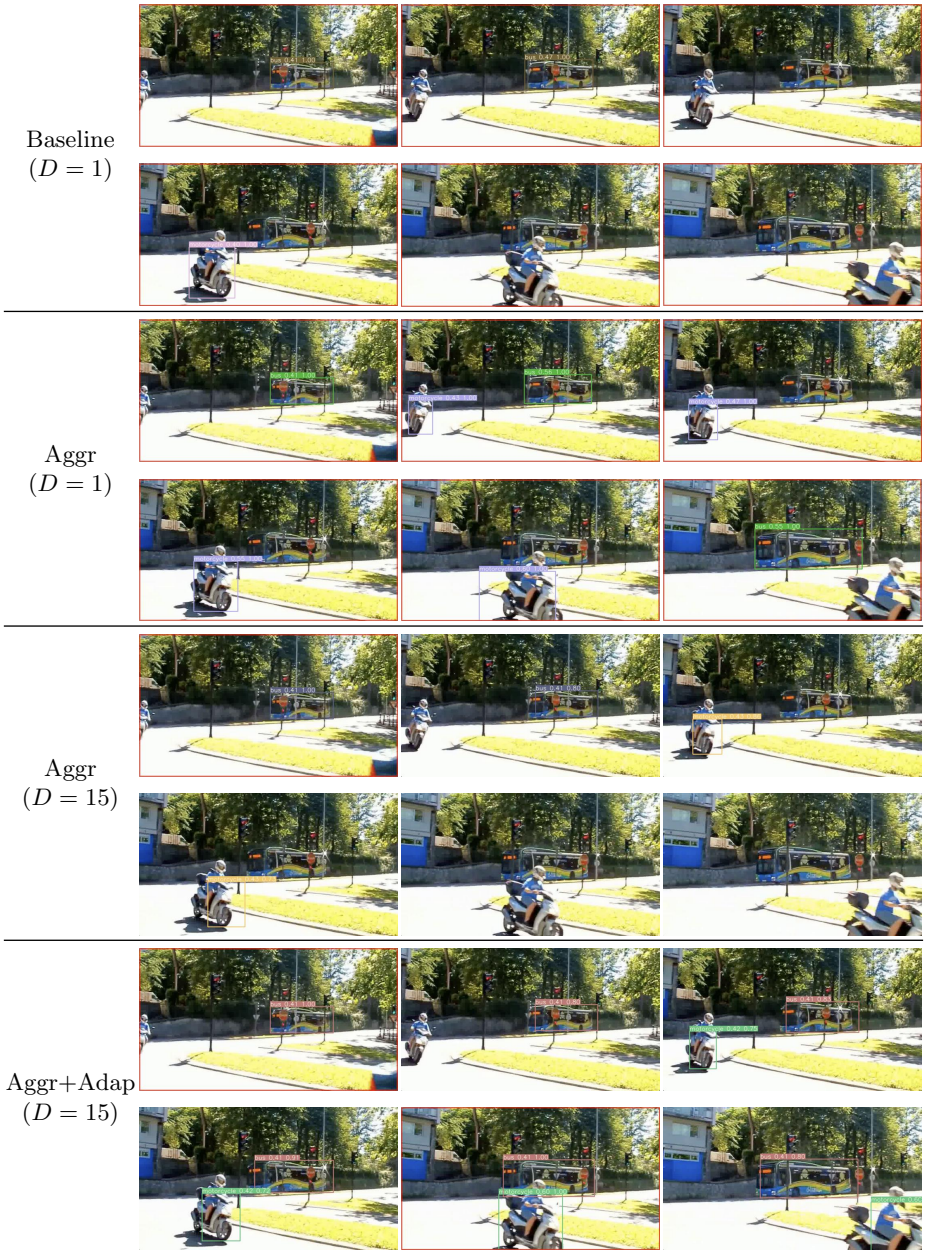


Fig. 5: Qualitative comparisons of our methods with single-image baseline. Our temporal aggregation module produces a more consistent prediction on the fast moving motorcycle. The adaptive keyframe scheduler further improves the results of the partially occluded bus when applying detection sparsely.

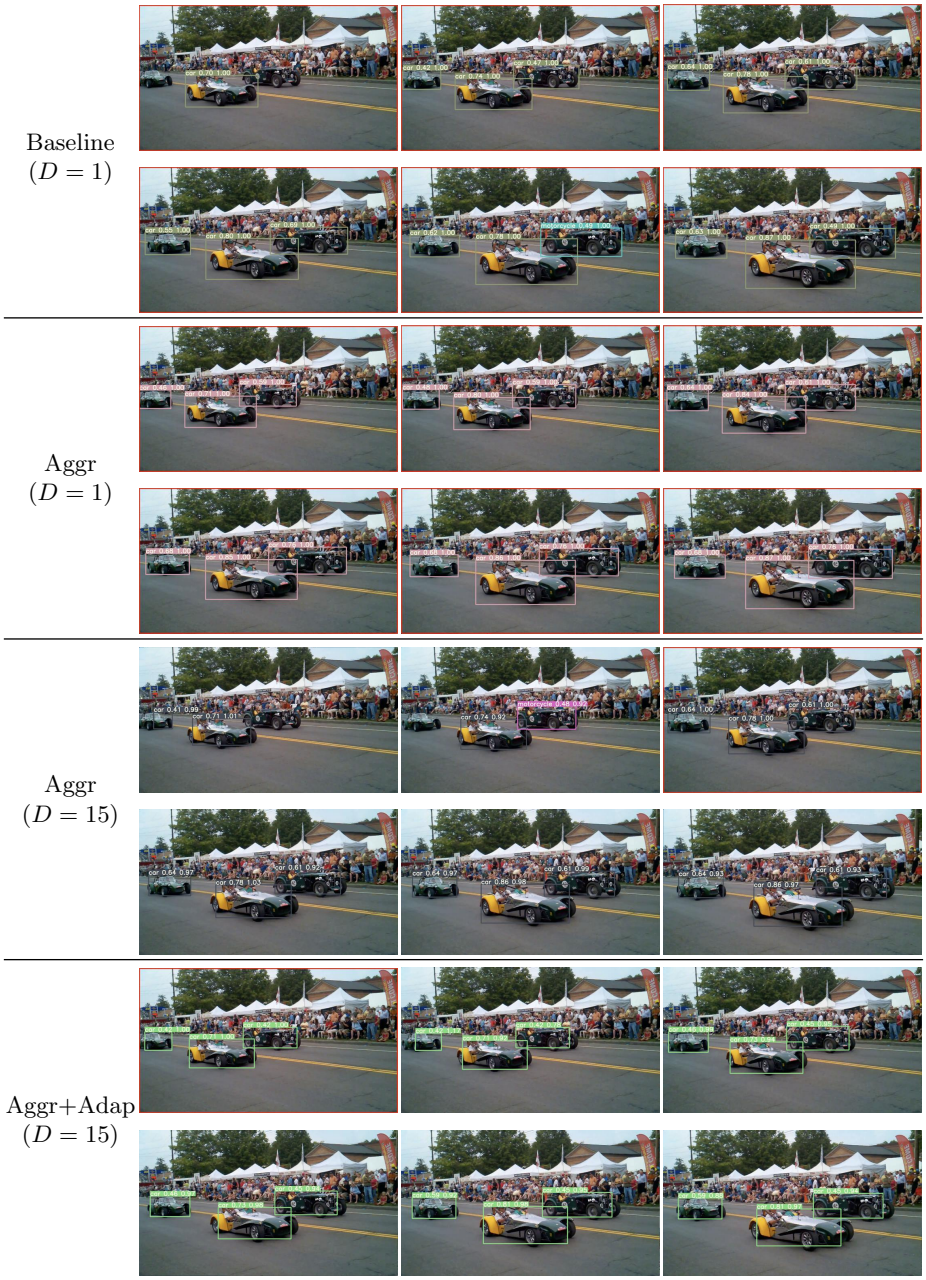


Fig. 6: Qualitative comparisons of our methods with single-image baseline. Our temporal aggregation module and adaptive keyframe scheduler produce more consistent and accurate predictions on the fast moving cars.

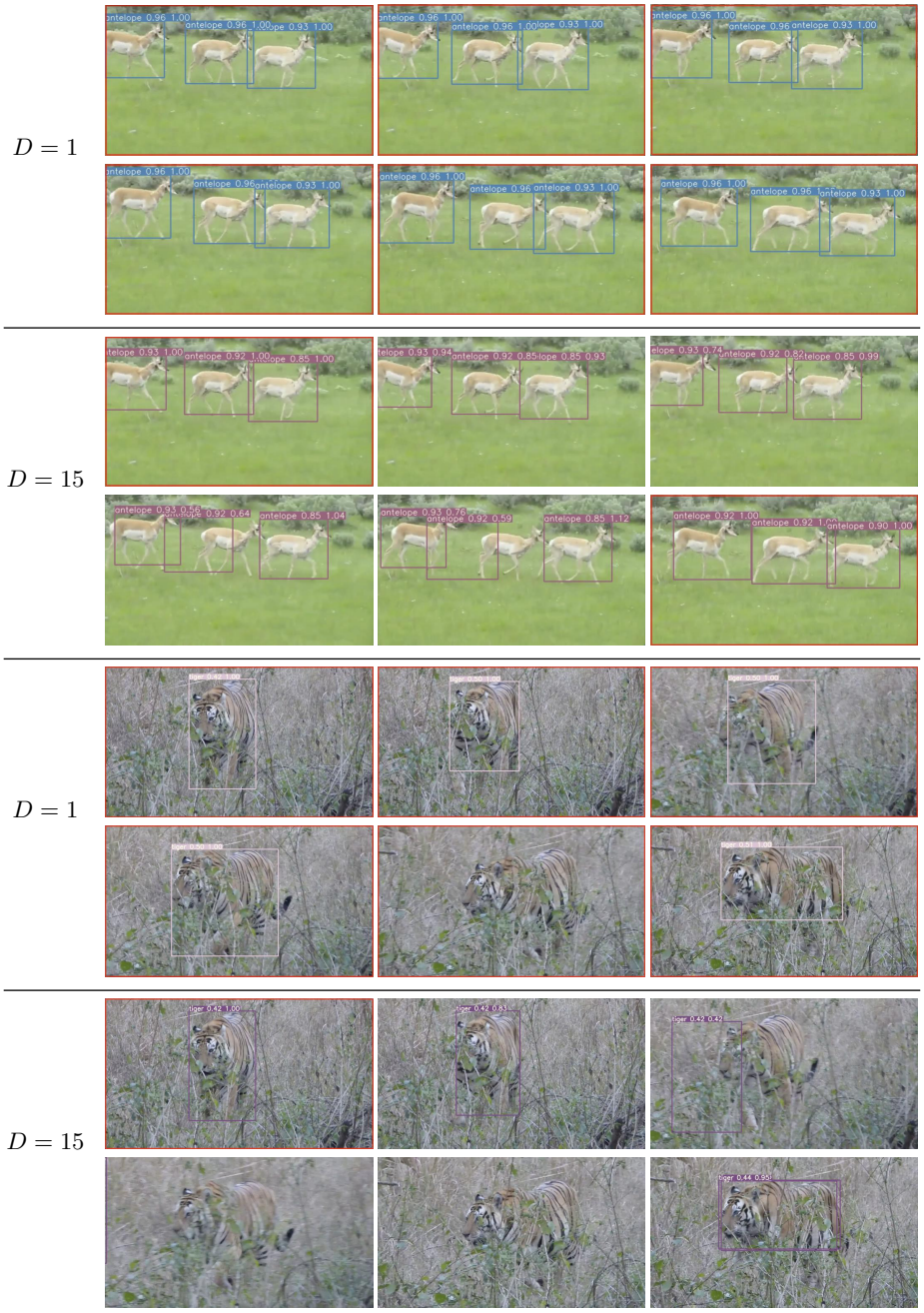
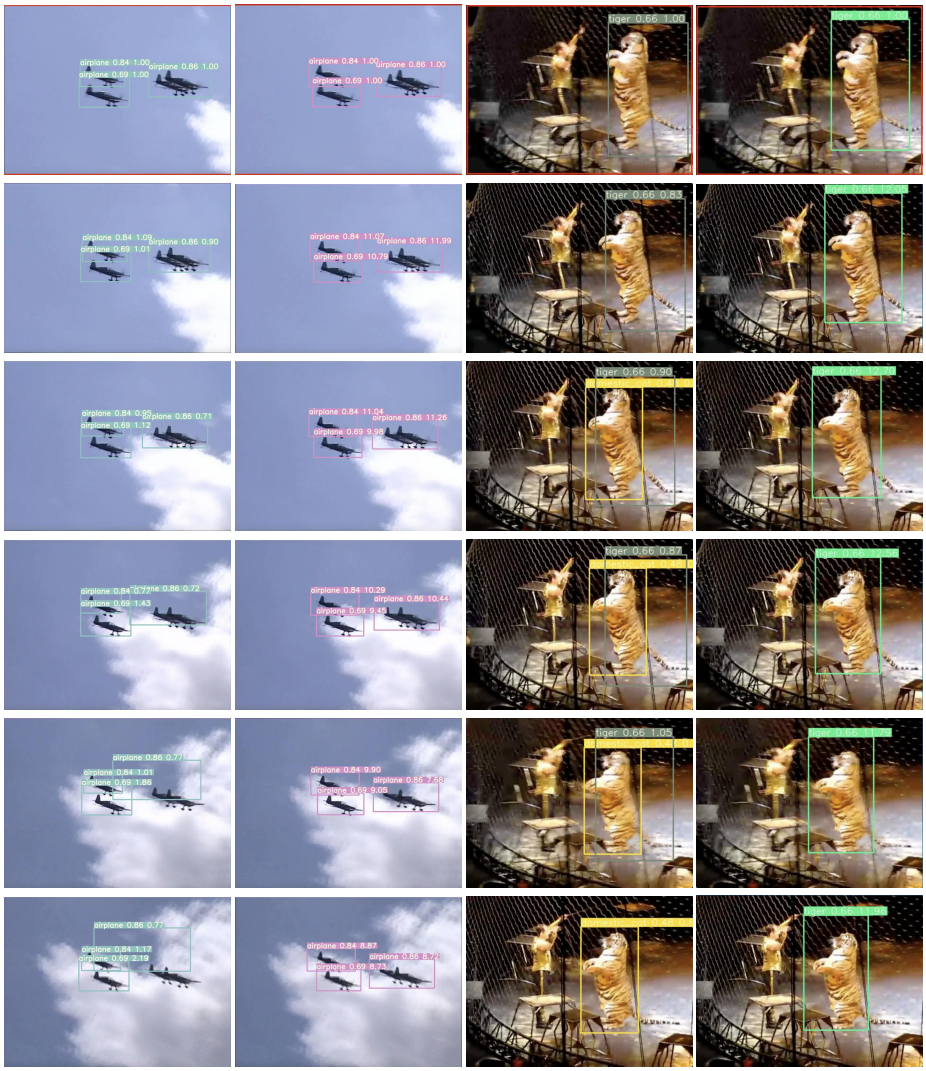


Fig. 7: Comparisons of different base interval. We show the results of CenterNet + KCF + LSTM-based aggregation with fixed keyframe intervals. They exemplify the need for frequent detection during rapid object movement or deformation.



KCF SiamFC KCF SiamFC

Fig. 8: Comparisons of different object trackers. We show the results of CenterNet + LSTM-based aggregation with a fixed keyframe interval $D_{base} = 7$. SiamFC trackers perform better than KCF in the videos where occlusion and rapid motion are present.

References

1. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.: High-speed tracking with kernelized correlation filters. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(3), 583–596 (2014) 1
2. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Proceedings of the European Conference on Computer Vision. pp. 740–755 (2014) 1
3. Liu, M., Zhu, M.: Mobile video object detection with temporally-aware feature maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5686–5695 (2018) 2
4. Liu, M., Zhu, M., White, M., Li, Y., Kalenichenko, D.: Looking fast and slow: Memory-guided mobile video object detection. arXiv preprint arXiv:1903.10172 (2019) 2
5. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018) 2
6. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y 1
7. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. In: arXiv preprint arXiv:1904.07850 (2019) 2
8. Zhu, X., Dai, J., Zhu, X., Wei, Y., Yuan, L.: Towards high performance video object detection for mobiles. arXiv preprint arXiv:1804.05830 (2018) 2