Supplementary Material: Predicting Action Tubes

Gurkirt Singh, Suman Saha, and Fabio Cuzzolin

Oxford Brookes University, UK
gurkirt.singh-2015@brookes.ac.uk

In this supplementary material, we first present implementation details in Section 1. Next we show extra plots at different detection threshold, where we include TPnet_{453} as well, as mentioned in the main paper.

1 Implementation details

All models are trained with batch size of 16 on two 1080Ti GPUs (11GB VRAM each) for AMTnet and TPnet. We used Pytorch library to implement all of the models (AMTnet, and TPnet) by following the implementation of Singh *et al.*[5] closely from their GitHub repository [1].

Input Data Preparations. The number of input frames to CNNs is more than 1 in our case e.g. 2 in AMTnet. Optical-flow stream uses a stack of 5 frames as input for each input frame in the sequence. In our case, sequence length is more than (in case AMTnet and TPnet equal to 2), so 2 stacks of 5 frames are used as input for flow stream in case of TPnet for flow stream. Number of input frames for RGB stream is equal to sequence length e.g. 2 for TPnet. Each optical flow image is 3 channel image, where first two channels are flow in x and y direction respectively and the third channel is magnitude (square-root of the sum of squares) of flow in both directions. We computed dense optical flow between each pair of successive video frames using the algorithms of [1].

SSD relies on **data augmentation** to boost the performance, we extended the data augmentation [3] function provided by Singh *et al.*[5] in their GitHub repository[1] to accept sequence of frames . The same data augmentation was applied to all the frames in one input sequence as used in [5, 2] .

VGG weight initialisation. Weights of VGG network (base network) are initialised with weights from a pre-trained ImageNet model [4][2] for appearance- and flow-based SSD networks for both training the initial SSD model, similar to [5]. Finally, weights of AMTnet and TPnet are initialised using trained SSD model on J-HMDB-21 dataset for both flow and appearance streams.

[1] https://github.com/gurkirt/realtime-action-detection

[2] https://gist.github.com/weiliu89/2ed6e13bfd5b57cf81d6

2 Early Label Prediction Performance (Accuracy).

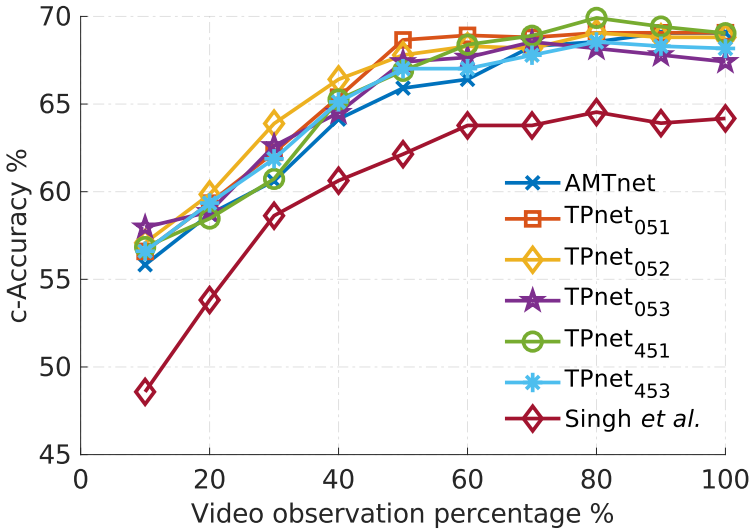

Fig. 1. *Early label prediction results (video-level label prediction accuracy) on J-HMDB-21 dataset in sub-figure. TPnet$_{abc}$ represents our TPnet where $a = \Delta_p$, $b = \Delta_f$ and $c = n$.*

3 Online Action Detection Performance (mAP).

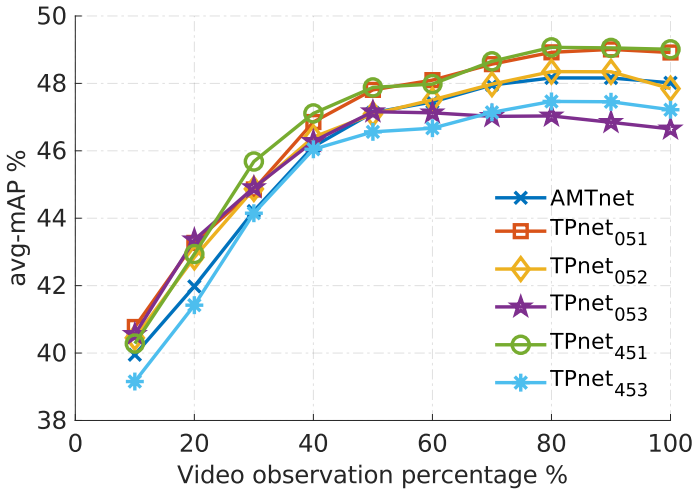

Fig. 2. *Online action detection performance (avg-mAP) on J-HMDB-21 dataset. TPnet$_{abc}$ represents our TPnet where $a = \Delta_p$, $b = \Delta_f$ and $c = n$.*

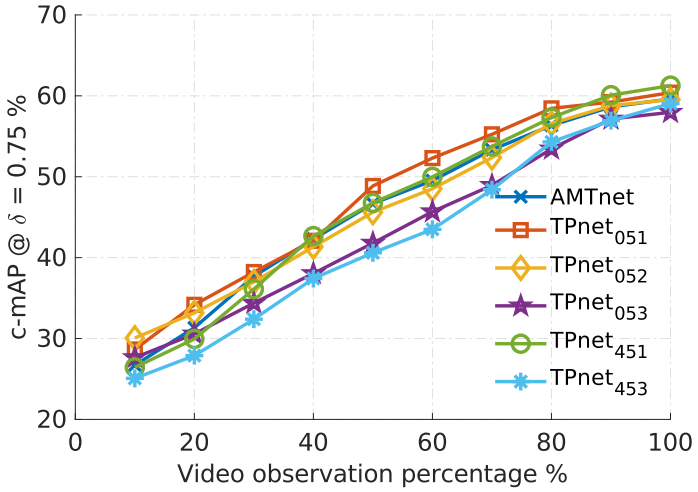

Fig. 3. *Online action detection performance (mAP $\delta = 0.75$) on J-HMDB-21 dataset. TPnet$_{abc}$ represents our TPnet where $a = \Delta_p$, $b = \Delta_f$ and $c = n$*

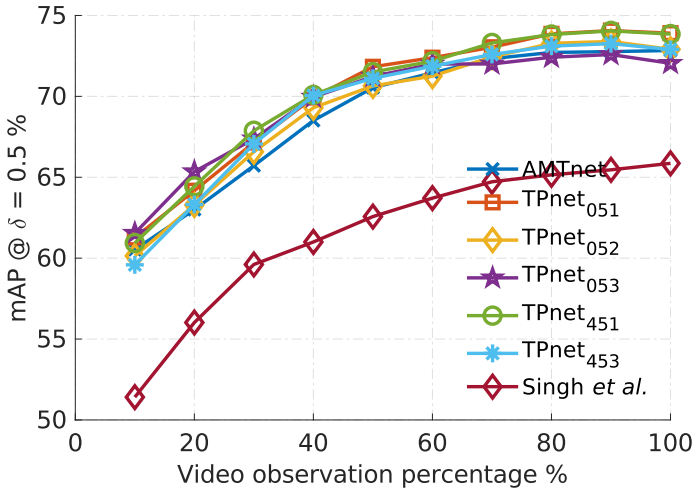

Fig. 4. *Online action detection performance (mAP $\delta = 0.5$) on J-HMDB-21 dataset. TPnet$_{abc}$ represents our TPnet where $a = \Delta_p$, $b = \Delta_f$ and $c = n$*

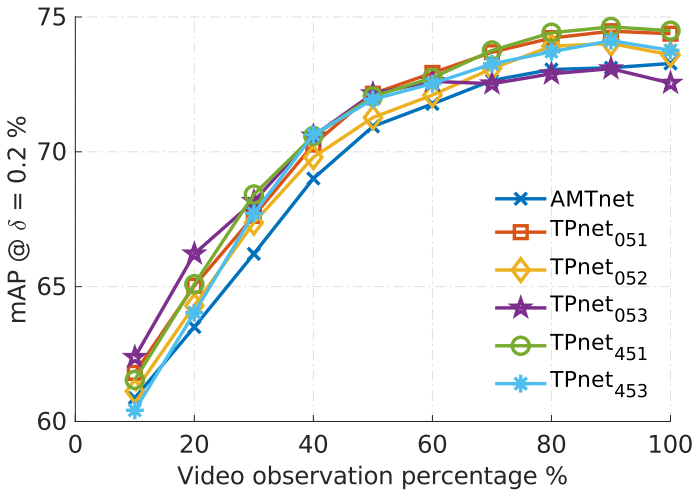

Fig. 5. *Online action detection performance (mAP $\delta = 0.2$) on J-HMDB-21 dataset. TPnet$_{abc}$ represents our TPnet where $a = \Delta_p$, $b = \Delta_f$ and $c = n$*

4 Action Tube Completion Performance (c-mAP).

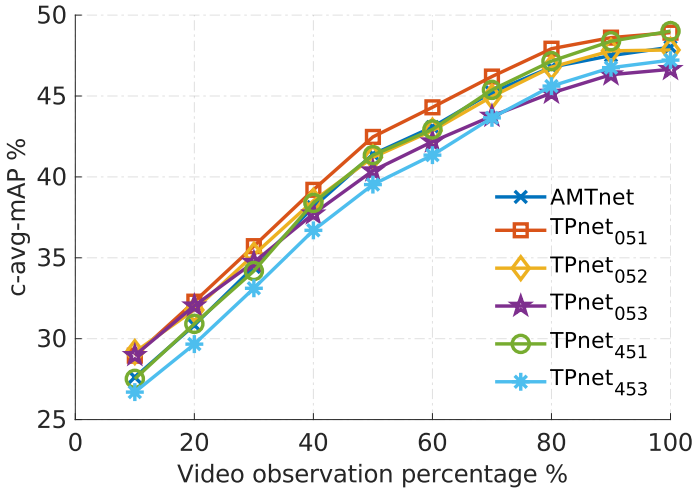

Fig. 6. *Action tube completion performance (c-avg-mAP) on J-HMDB-21 dataset. TPnet$_{abc}$ represents our TPnet where $a = \Delta_p$, $b = \Delta_f$ and $c = n$.*

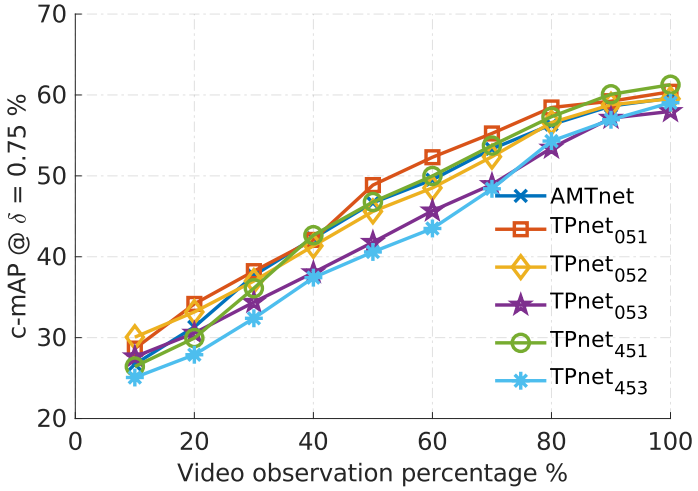

Fig. 7. *Action tube completion performance (c-mAP $\delta = 0.75$) on J-HMDB-21 dataset. TPnet$_{abc}$ represents our TPnet where $a = \Delta_p$, $b = \Delta_f$ and $c = n$*

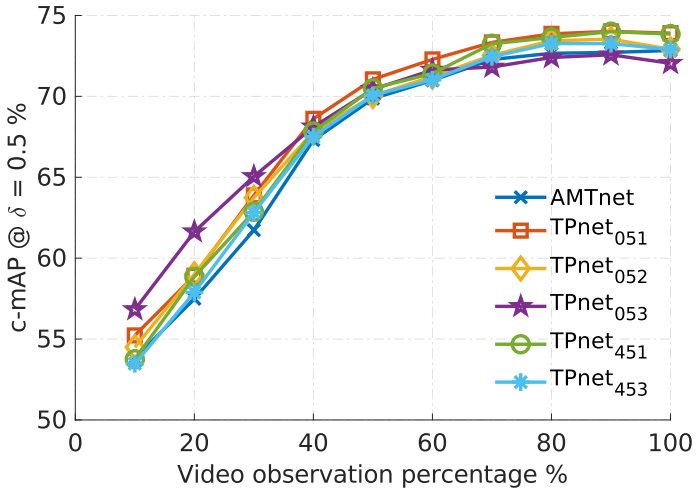

Fig. 8. *Action tube completion performance (c-mAP $\delta = 0.5$) on J-HMDB-21 dataset. TPnet$_{abc}$ represents our TPnet where $a = \Delta_p$, $b = \Delta_f$ and $c = n$*

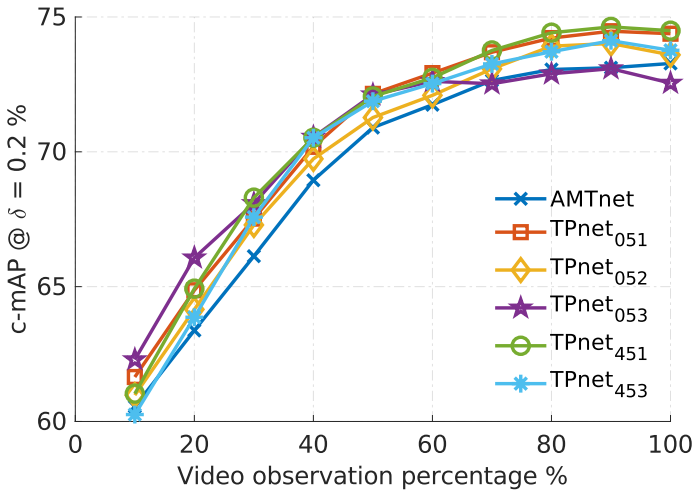

Fig. 9. *Action tube completion performance (c-mAP $\delta = 0.2$) on J-HMDB-21 dataset. TPnet$_{abc}$ represents our TPnet where $a = \Delta_p$, $b = \Delta_f$ and $c = n$*

5 Future Action Tube Prediction Performance (p-mAP).

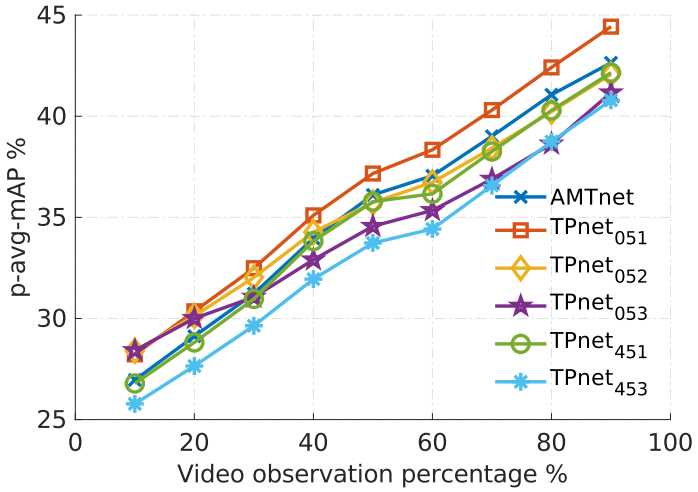

Fig. 10. *Future action tube prediction performance (p-avg-mAP) on J-HMDB-21 dataset. TPnet$_{abc}$ represents our TPnet where $a = \Delta_p$, $b = \Delta_f$ and $c = n$.*

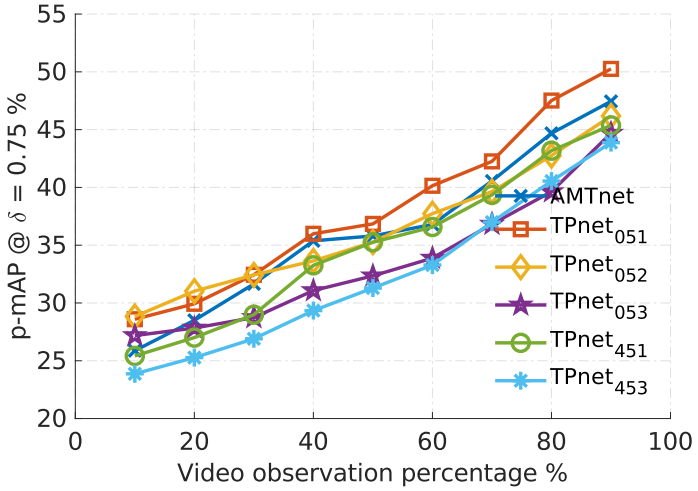

Fig. 11. *Future action tube prediction performance (p-mAP $\delta = 0.75$) on J-HMDB-21 dataset. TPnet$_{abc}$ represents our TPnet where $a = \Delta_p$, $b = \Delta_f$ and $c = n$*

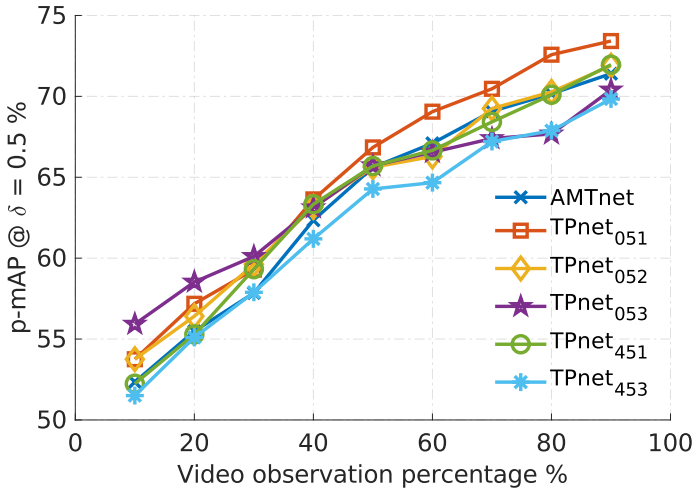

Fig. 12. *Future action tube prediction performance (p-mAP $\delta = 0.5$) on J-HMDB-21 dataset. TPnet$_{abc}$ represents our TPnet where $a = \Delta_p$, $b = \Delta_f$ and $c = n$*

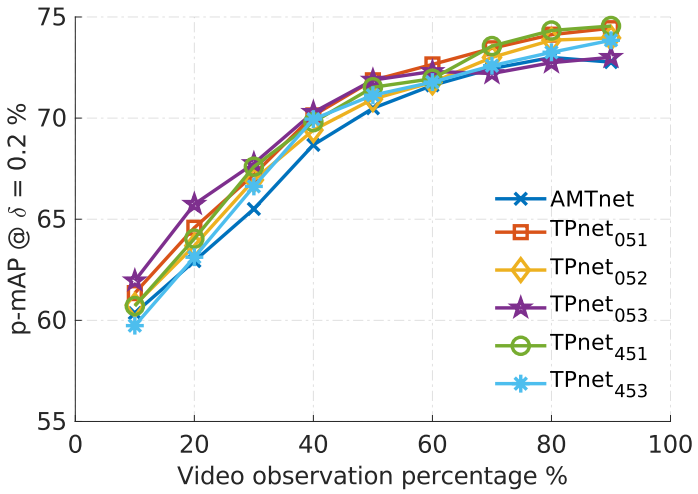

Fig. 13. *Future action tube prediction performance (p-mAP $\delta = 0.2$) on J-HMDB-21 dataset. TPnet$_{abc}$ represents our TPnet where $a = \Delta_p$, $b = \Delta_f$ and $c = n$*

References

1. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping (2004)
2. Kalogeiton, V., Weinzaepfel, P., Ferrari, V., Schmid, C.: Action tubelet detector for spatio-temporal action localization. In: IEEE Int. Conf. on Computer Vision (2017)
3. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: SSD: Single shot multibox detector. arXiv preprint arXiv:1512.02325 (2015)
4. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: Looking wider to see better. arXiv preprint arXiv:1506.04579 (2015)
5. Singh, G., Saha, S., Sapienza, M., Torr, P., Cuzzolin, F.: Online real-time multiple spatiotemporal action localisation and prediction. In: IEEE Int. Conf. on Computer Vision (2017)