Supplementary Material for Actions as Moving Points

Yixuan Li*, Zixu Wang*, Limin Wang[0000−0002−3674−7718], and Gangshan Wu

State Key Laboratory for Novel Software Technology, Nanjing University, China
{liyixxxuan,zixuwang1997}@gmail.com, {lmwang, gswu}@nju.edu.cn

1 Study on Hyper-parameters

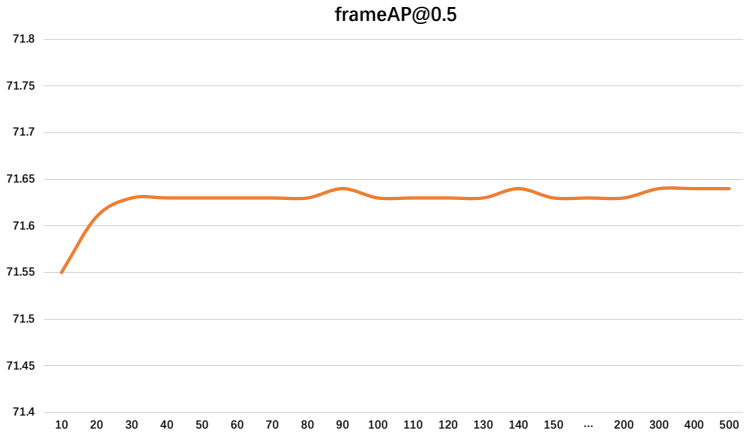

Fig. 1. Study on N. FrameAP@0.5 result on UCF101-24 [8] with tubelet length K=5 and only RGB input.

N in Center Branch. During inference, Center Branch keeps top N instances from all categories after max pooling operation, which is indicated in paper's Section 3.1. We follow CenterNet [11], which is an anchor-free object detector and set N as 100. As shown in Figure 1, we can see that the detection result is robust to N and changes slightly after 40.

a and b in Loss Function. Paper's Equation (9) is MOC's training objective consisting of three branches' loss. As shown in Figure 2, we have a linear search on a and b with tubelet length K=5 and only RGB input. We can see that a=1, b=0.1 performs best.

* Yixuan Li and Zixu Wang contribute equally to this work. This work is supported by Tencent AI Lab.

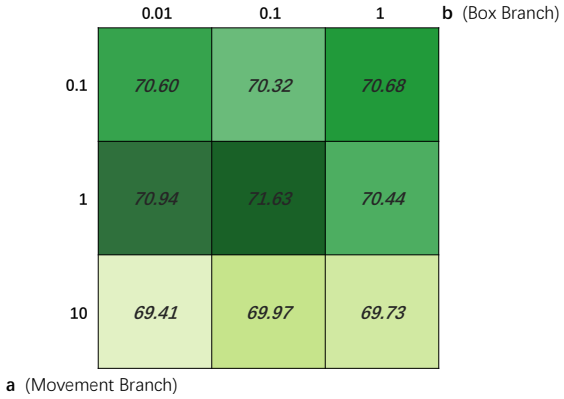

Fig. 2. Study on a and b. FrameAP@0.5 result on UCF101-24 [8] with tubelet length K=5 and only RGB input.

2 Error Analysis

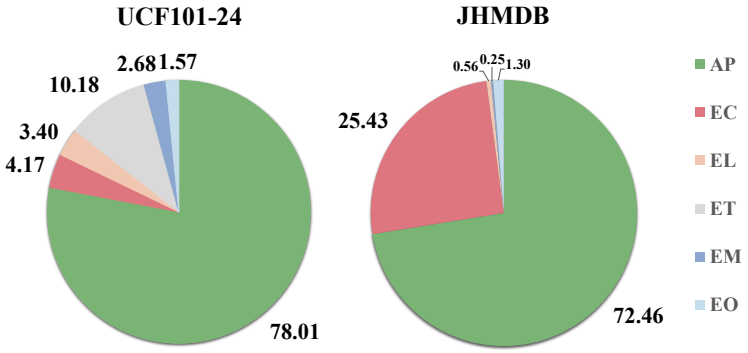

Fig. 3. Error analysis on UCF101-24 [8] and JHMDB [5] (only split 1). We report the detection error results according to five categories: (1) classification error E_C, (2) localization error E_L, (3) time error E_T, (4) missed detection E_M, and (5) other error E_O. The green part represents the correct detection. With tubelet length K = 7 and two-stream fusion.

In this section, following [6], we conduct an error analysis on the frame mAP to better explore our proposed MOC-detector. In particular, we investigate five kinds of tubelet detection error: (1) classification error E_C: the detection IoU is greater than 0.5 with the ground-truth box of another action class. (2) localization error E_L: the detection class is correct in a frame but the bounding box IoU with ground truth is less than 0.5. (3) time error E_T: the detection in the untrimmed video covers the frame that doesn't belong to the temporal extent of the current action instance. (4) missed detection error E_M: cannot detect out a

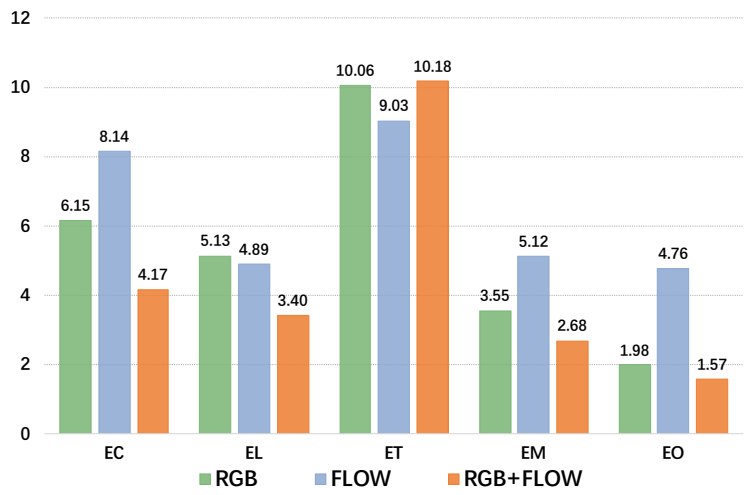

Fig. 4. Error Analysis with Two-stream Fusion. We report the detection error results according to five categories by changing input: (1) classification error E_C, (2) localization error E_L, (3) time error E_T, (4) missed detection E_M, and (5) other error E_O. With tubelet length K = 7 and two-stream fusion on UCF101-24 [8].

ground truth box. (5) other error E_O: the detection appears in a frame without the class and has IoU less than 0.5 with the ground truth bounding box of other classes.

We present error analysis on the untrimmed dataset UCF101-24 [8] and the trimmed dataset JHMDB [5] (only split 1) with tubelet length $K = 7$ and two-stream fusion. As shown in Figure 3, we find the major error is E_T, time error (10.18%), for the untrimmed dataset UCF101-24 [8] and E_C, classification error (25.43%), for the trimmed dataset JHMDB [5]. Although our MOC-detector has achieved state-of-art on both datasets, we will try to extend this framework to model longer temporal information to improve classification accuracy and model action boundary in the temporal dimension to eliminate time error.

We also visualize error analysis with two-stream fusion on UCF101-24 [8] and the results are reported in Figure 4. Note that we set tubelet length K as 7. First, spatial stream performs obviously better than the temporal stream for classification error and missed detection, owing to its richer information. Second, two-stream fusion improves the performance except for time error, which shows that two-stream fusion harms temporal localization.

3 More exploration on Box Branch

Previously, we tried to add temporal information into the bbox estimation by stacking features across time as input, which is as same as Movement Branch. As shown in Table 1, the performance drops after adding temporal information. It indicates that a single frame is sufficient for the bbox detection.

Table 1. Exploration study on the Box Branch design with only RGB as input and $K = 5$. Note that union means stacking feature together to add temporal information into the bbox estimation and separate (MOC) estimates bbox separately for each frame.

Method	F-mAP@0.5 (%)	Video-mAP (%)			
		@0.2	@0.5	@0.75	0.5:0.95
union	70.41	76.54	49.14	26.61	**26.14**
separate(MOC)	**71.63**	**77.74**	**49.55**	**27.04**	26.09

4 More Results on JHMDB

Table 2. Comparison with Gu et al. [3] and Sun et al. [9] on JHMDB [5] (3 splits) with tubelet length K=7 and two stream fusion. Ours (MOC) † is pretrained on ImageNet [2] , Ours (MOC)†† is pretrained on COCO [7] and Ours (MOC)††† is pretrained on UCF101-24 [8] for action detection.

Method	GFLOPs	JHMDB				
		Frame-mAP@0.5 (%)	Video-mAP (%)			
			@0.2	@0.5	@0.75	0.5:0.95
Ours (MOC)†	29.4	68.0	76.2	75.4	68.5	54.0
Ours (MOC)††	29.4	70.8	77.3	77.2	71.7	59.1
Ours (MOC)†††	29.4	74.0	**80.7**	**80.5**	**75.0**	**60.2**
Gu *et al.* 2018 [3] (I3D)	>91.0	73.3	-	78.6	-	-
Sun *et al.* 2018 [9] (S3D-G)	>65.5	**77.9**	-	80.1	-	-

Our MOC is a one stage tubelet detector with 2D backbone. We compare it with two-stage detectors with 3D backbone [3,9] in paper's Section 4.3, which perform comparably with us on UCF101-24 [8] while better than ours on JHMDB [5].

JHMDB [5] is really small and sensitive to the pre-train model. For fair comparison with 2D backbone methods in paper's Section 4.3, we just provide results with ImageNet [2] pretrain and COCO [7] pretrain. But Gu et al [3] and Sun et al. [9] both pretrain 3D backbone on Kinetics [1], which is a large-scale video classification dataset and always boosts task results especially on small datasets. We pretrain our MOC on UCF101-24 [8] for action detection in Table 2, which outperforms Gu et al. [3] for all metrics with saving more than 3 times computation cost and performs comparably with Sun et al. [9] with saving more than 2 times computation cost. Note that Gu et al. [3] and Sun et al. [9] do not provide implementation code, so we just roughly estimate the backbone computation for each frame's detection result, whose input is 20 frames with resolution of 320*400. For Gu et al. [3], we calculate ResNet50 (conv4) [4] for action localization and I3D (Mixed_4e) [1] for classification. For Sun et al. [9]

(Base Model), we calculate ResNet50 (conv4) [4] for action localization and S3D-G [10] for classification. For our MOC, we calculate the whole computation cost for each frame detection result. For fair comparison, we only use RGB as input to estimate GFLOPs for all methods.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
3. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6047–6056 (2018)
4. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
5. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: Proceedings of the IEEE international conference on computer vision. pp. 3192–3199 (2013)
6. Kalogeiton, V., Weinzaepfel, P., Ferrari, V., Schmid, C.: Action tubelet detector for spatio-temporal action localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4405–4413 (2017)
7. 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: European conference on computer vision. pp. 740–755. Springer (2014)
8. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
9. Sun, C., Shrivastava, A., Vondrick, C., Murphy, K., Sukthankar, R., Schmid, C.: Actor-centric relation network. In: ECCV. pp. 335–351 (2018)
10. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 305–321 (2018)
11. Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)