Dance with Flow: Two-in-One Stream Action Detection Supplementary Material

Jiaojiao Zhao and Cees G. M. Snoek
University of Amsterdam
j.zhao3, cgmsnoek@uva.nl

1. Extra Experiments and Results

All the results in this supplementary material are from a single-frame network trained on *UCF101-24*. We report mAP at the high IoU threshold of 0.5:0.95. BroxFlow is applied here.

Influence of Fusion Performance of different fusion methods are seen in Table 1. Besides mean fusion, we use another two fusion methods to fuse RGB-stream and flow-stream following [1]. To conduct boost fusion, we perform L-1 normalization on the detection boxes scores after fusion and then retain any flow detection boxes for which an associated appearance based box was not found. The other way is retaining the union $\{b_i^a\} \cup \{b_j^f\}$ of the two sets of RGB-stream $\{b_i^a\}$ and flow-stream $\{b_j^f\}$ detection boxes, respectively. All of the three fusion methods applied to two stream help to improve results of single stream. Mean fusion is the best way to fuse RGB- and flow- stream. Our two-in-one stream beats all of them with only half runtime and the number of parameters.

Quantitative Results per Action We report deteciton results per action from *UCF101-24* in Table 2. For some challenging cases such as basketball dunk, cricket bowling, pole vault and volleyball spiking, which have crowded and cluttered backgrounds, our two-in-one stream achieves better results than other methods. For instance two-in-one stream outperforms two-stream by 6% and RGB-stream by 9% for pole vault. For the cases where multiple instances may occur, such as fencing, ice dancing and salsa spin, our two-in-one stream also boosts the accuracy. Notably, fusing RGB and optical flow, improves results in most cases except for skiing. Both two-stream and two-in-one stream perform worse than the RGB-stream for skiing. The flow images are very noisy and even make results worse. Overall, the proposed two-in-one stream outperforms alternatives for 16 out of 24 action classes.

Qualitative Results per Action Some successful detected results of challenging cases using our two-in-one stream are visualized in Figure 1 and Figure 2. The green boxes represent the ground truth boxes. The yellow boxes with the labels are detected boxes with the classification scores.

Basketball dunk is difficult as there are many interfering actors. An RGB-stream cannot detect any actions for the scenes shown in the Figure 1. For cliff diving, when the actor reaches the surface of the water, the action is mistaken as surfing due to the sea context captured by the RGB-stream. The RGB-stream model may pay more attention to backgrounds. However, our two-in-one stream using flow condition to modulate RGB features, focuses more on actions and improves the results. For pole vault, it is easily mistaken as cliff diving when the actor falls down from up using the RGB stream. Our two-in-one strem performs better. In Figure 2, we also show some multi-instance cases such as ice dancing and salsa spin. In these cases only one ground-truth box is given for each image. However, it is reasonable that our two-in-one stream is capable to detect multiple instances. Thus, the results are actually better than the AP values of these cases shown in Table 2. It is worth to be mentioned that the detection boxes of our two-in-one stream have high overlap with the ground-truth boxes.

Failure Cases We show three kinds of failure cases in Figure 3. It is difficult to define whether the frames at the bound of an action are action or not. In the first row, these frames follow an action tennis swing. The model still takes them as tennis swing. Without considering ground-truth, we think it is reasonable. In the second row, when the actor appears blurry, our model still gives correct detection in the last three frames. However, there are no actions in the ground-truth for these frames. In the third row, the model successfully locates the actors, but assigns the wrong action label. The real action is pole vault, which has a similar run-up with floor gymnastics in the beginning of the action.

References

[1] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip HS Torr, and Fabio Cuzzolin. Online real-time multiple spatiotemporal action localisation and prediction. In *ICCV*, 2017. 1

	Accuracy	**Efficiency**	
		sec/frame	# param. (M)
flow-stream	11.60	0.04	26.82
RGB-stream	18.49	0.04	26.82
two-stream (boost-fusion)	18.97	0.09	53.64
two-stream (union-set)	19.42	0.09	53.64
two-stream (mean-fusion)	19.79	0.09	53.64
two-in-one stream	**21.51**	0.04	26.93

Table 1: **Influence of Fusion** Performance ($mAP@IoU$ = 0.5:0.95, runtime and # param.) comparison on *UCF101-24*. Three different fusion methods are used in two-stream method. Mean fusion achieves better results than the two others. Our two-in-one stream outperforms all of them.

video-mAP(%)	mAP	Basketball	BasketballDunk	Biking	CliffDiving	CricketBowling	Diving	Fencing	FloorGymnastics
flow-stream	11.60	0.03	0.06	9.09	2.57	0.11	1.40	33.21	31.79
RGB-stream	18.49	0.00	0.27	23.25	8.15	0.38	5.90	46.90	45.30
two-stream	19.79	0.01	0.12	**24.81**	**11.04**	0.20	4.95	45.03	43.45
two-in-one stream	**21.51**	**0.17**	**1.16**	23.13	9.13	**0.89**	**7.30**	**53.70**	**54.31**
		GolfSwing	HorseRiding	IceDancing	LongJump	PoleVault	RopeClimbing	SalsaSpin	SkateBoarding
flow-stream		4.64	35.69	15.63	14.75	2.81	23.69	0.34	28.17
RGB-stream		10.45	42.35	17.11	20.70	4.10	31.00	0.75	41.10
two-stream		12.38	40.77	16.87	**25.53**	7.94	36.46	0.44	44.35
two-in-one stream		**14.37**	**45.30**	**19.00**	23.44	**13.70**	**46.40**	**1.70**	**46.40**
		Skiing	Skijet	SoccerJuggling	Surfing	TennisSwing	TrampolineJumping	VolleyballSpiking	WalkingWithDog
flow-stream		6.08	2.84	44.29	4.13	0.20	5.14	0.00	11.71
RGB-stream		**37.03**	26.66	25.41	18.12	0.15	5.72	0.00	26.85
two-stream		34.25	**28.84**	**45.98**	16.02	0.17	**6.60**	0.00	**28.61**
two-in-one stream		35.57	28.11	41.02	**18.56**	**0.46**	5.65	**0.06**	26.80

Table 2: Per action class video $AP@IoU$ = 0.5:0.95 on *UCF101-24*. Our two-in-one stream achieves better results for 16 action classes. Especially for diving, fencing, floor gymnastics, horse riding, ice dancing, pole vault and rope climbing, there are obvious improvements. For fencing, ice dancing and salsa spin, where multiple instances may occur, two-in-one stream also boosts the accuracy. For skiing, both two-stream and two-in-one stream are lower than RGB-stream in accuracy. But our two-in-one stream is still better than two stream in this case.

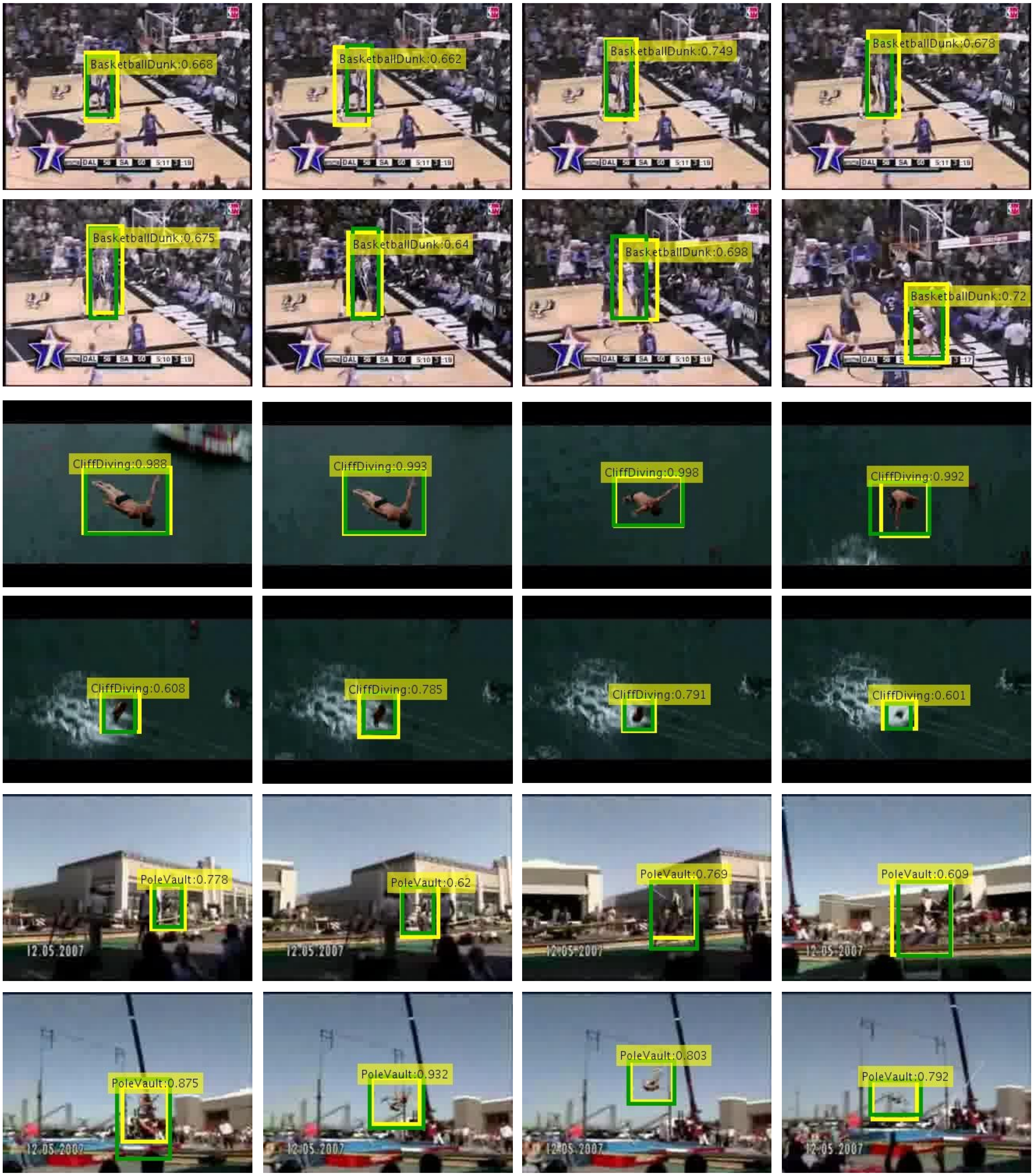

Figure 1: Examples of successful detected results of some challenging cases using our two-in-one stream. The green boxes represent ground-truth boxes. The yellow boxes with labels mean detection boxes with classification scores. Our two-in-one stream performs well at high IoU thresholds.

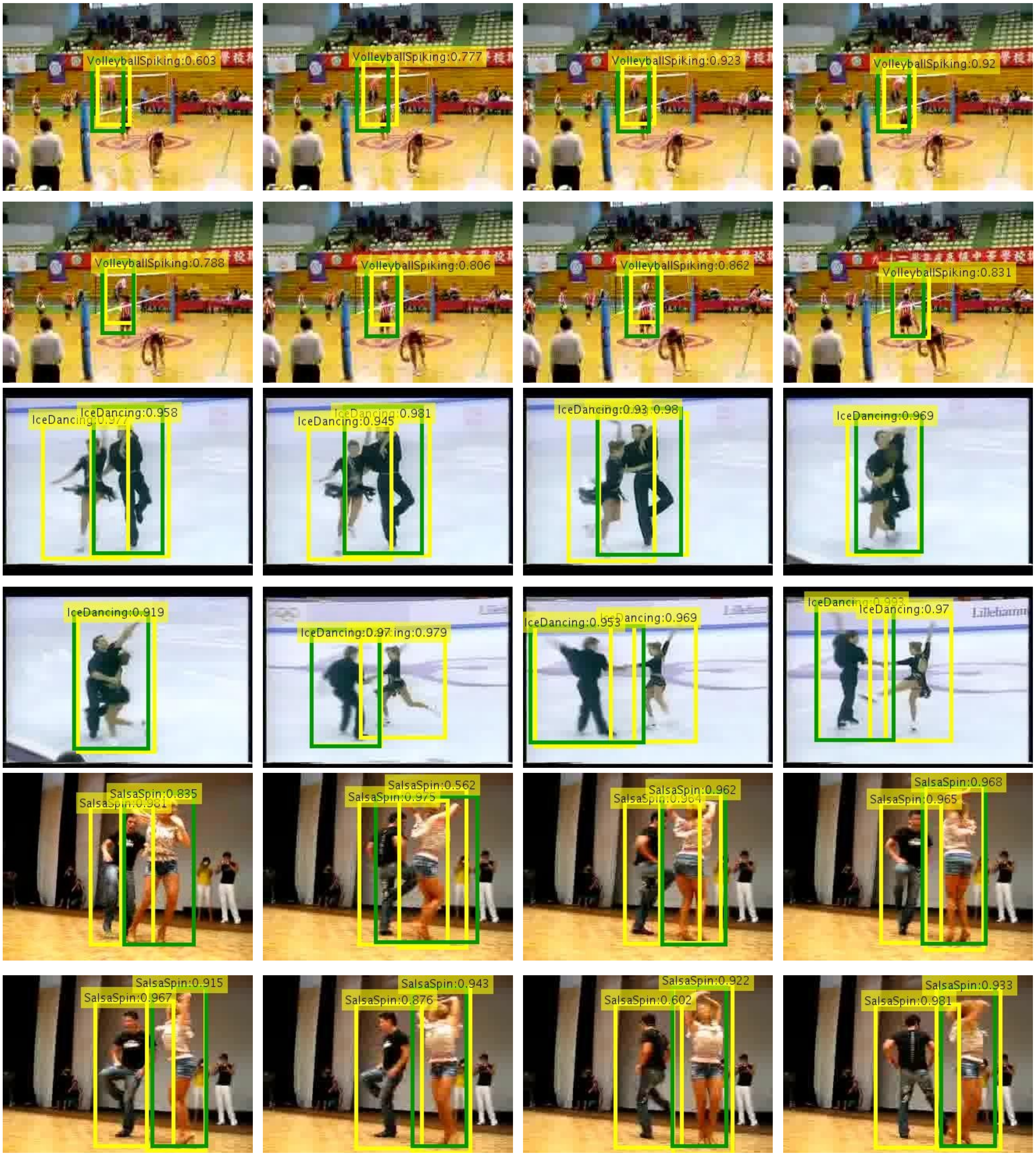

Figure 2: Examples of successful detected results. Our two-in-one stream is able to detect multiple instances for ice dancing and salsa spin. It is reasonable even only one actor is labeled in the ground-truth.

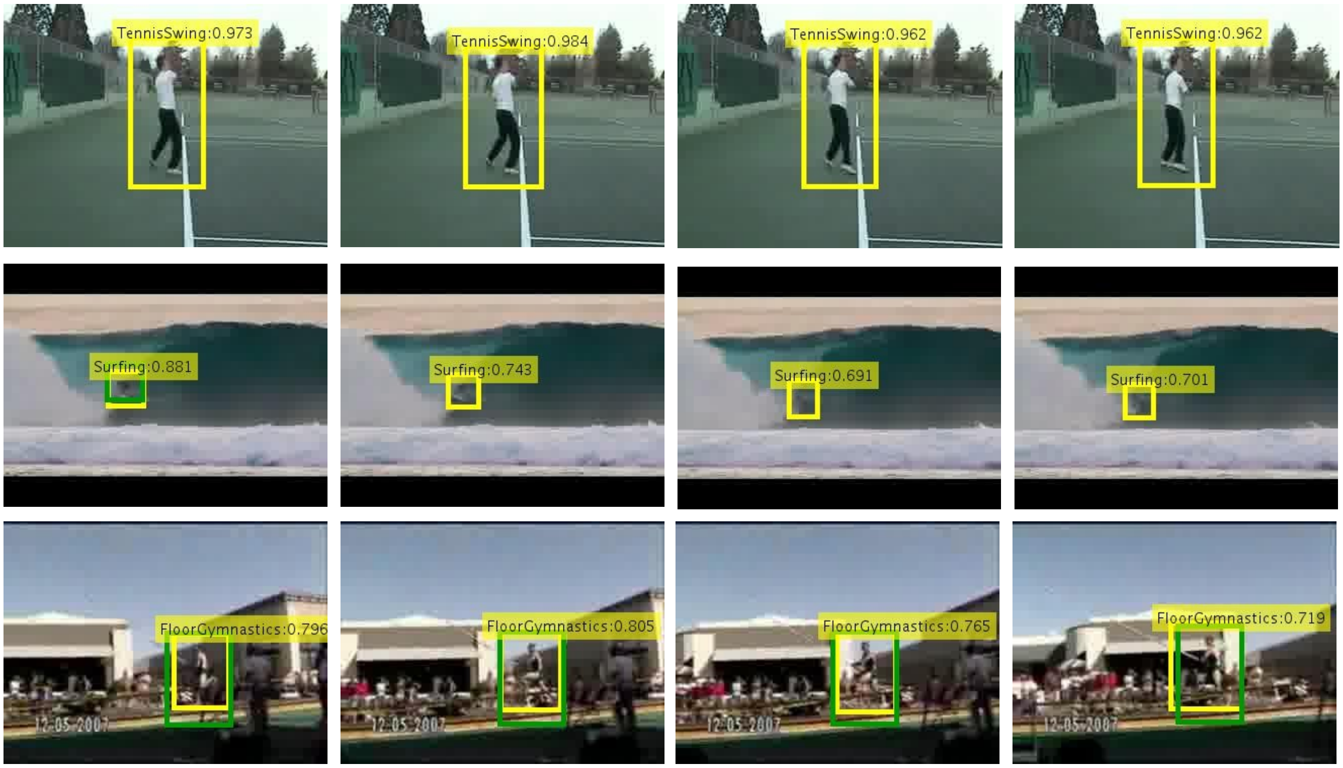

Figure 3: Failure cases. It is reasonable to assign tennis swing to these actions in first row even there is no action for these frames in the ground-truth. In the second row, our model still gives correct detections for the blurry actions even no actions are labeled in the ground-truth. In the last row, our model successfully locates the actors. But pole vault is mistaken as floor gymnastics as the two actions have the similar run-up in the beginning.