Supplementary Materials
SF-Net: Single-Frame Supervision for Temporal Action Localization

Fan Ma[1], Linchao Zhu[1], Yi Yang[1], Shengxin Zha[2], Gourab Kundu[2], Matt Feiszli[2], and Zheng Shou[2]

[1] University of Technology Sydney, Australia
[2] Facebook AI

1 Single-frame Annotation

We invite annotators with different backgrounds to label single-frames for all actions intances. Before annotating each dataset, four annotators have watched a few video examples containing different actions to be familiar with action classes. They are asked to annotate one single frame for each target action instance while watching the video by our designed annotation tool. Specifically, they are required to pause the video when they identify an action instance and choose the action class that the paused frame belongs to. Once they have chosen the action class, they need to continue watching the video and record the frames for the next target action instances. After watching the whole video, the annotator should press the generation button and the annotation tool will then automatically produce the timestamps and action classes of all operated frames for the given video. Compared to the annotation process in the weakly-supervised setting, this results into almost no extra time cost since the timestamps are automatically generated. The single-frame annotation process is much faster than annotating the temporal boundary of each action in which the annotator often watches the video many times to define the start and end timestamp of a given action.

1.1 Annotation guideline

Different people may have different understandings of what constitutes a given action. To reduce the ambiguity, we prepare a detailed annotation guideline, which includes both clear action definitions as well as positive/negative examples with detailed clarifications for each action. For each action, we give (1) textual action definition for single-frame annotation, (2) positive single-frame annotations, and (3) segmented action instances for annotator to be familiar with.

1.2 Annotation tool

Our annotation tool supports automatically recording timestamp for annotating single-frame. This makes the annotation process faster when annotators notice an

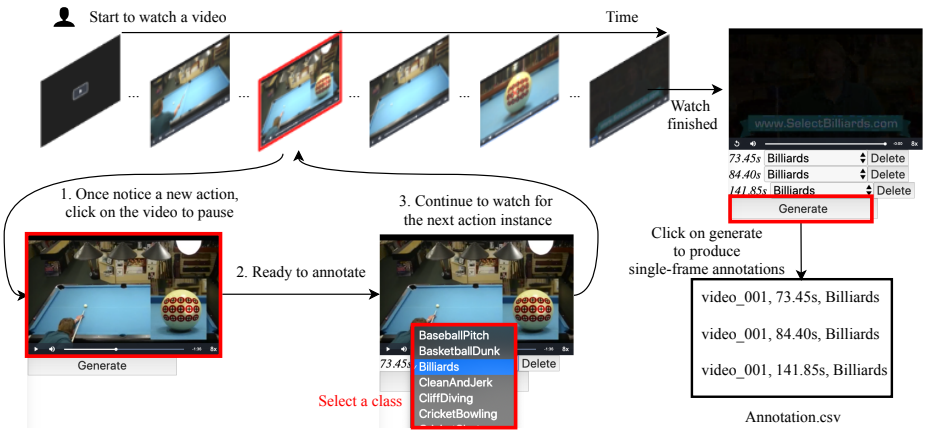

Fig. 1. Interface for annotating a single frame. First step is to pause the video when annotators notice an action while watching the video. The second step is to select the target class for the paused frame. After annotating an action instance, the annotator can click the video to keep watching the video for the next action instance. Note that the time is automatically generated by the annotation tool. After watching a whole video, the annotator can press the generate button to save all records into a csv file.

action and ready to label the paused frame. The interface of our annotation tool is presented in Figure 1. After watching a whole video, the annotator can press the generate button, the annotation results will be automatically saved into a csv file. When annotators think they made a wrong annotation, they can delete it at any time while watching the video. We have shown the one annotation example in the supplementary file. We have uploaded a video in the supplementary file to show how to annotate single-frame while watching the video.

1.3 Quality control

We make two efforts to improve the annotation quality. First of all, each video is labeled by four annotators, and the annotated single-frames of a view are randomly selected during experiments to reduce annotation bias. Secondly, we train annotators before annotating videos and make sure that they can notice target actions while watching the video.

2 Action Frame Mining

The action frame mining strategy is described in Algorithm 1. We treat the labeled action frame as the anchor frame and expand frames around it. We use a threshold ξ and the label consistency with neighbors to decide whether to add the unlabeled frame or not. The expanded frames are annotated with the same

Algorithm 1 Action Frame Mining

Input: video classification activation $C \in \mathcal{R}^{T \times N_c+1}$, labeled action frame at time t belonging to action class y, expand radius $r = 5$, threshold $\xi = 0.9$.
Output: expanded frames set $\mathcal{S}$
gather classification score $C(t)$ for the anchor frame
$\mathcal{S} \leftarrow \{(t, y)\}$
function EXPAND(s) ;
    for $j \leftarrow 1; j \leq r$ **do**
        $\hat{y}_{past} \leftarrow \text{argmin } C(t + (j-1)s)$
        $\hat{y}_{current} \leftarrow \text{argmin } C(t + js)$
        $\hat{y}_{future} \leftarrow \text{argmin } C(t + (j+1)s)$
        if $\hat{y}_{past} == \hat{y}_{current} == \hat{y}_{future}$ and $C(t + js)_y \geq \xi C(t)_y$ **then**
            $\mathcal{S} \leftarrow (t + js, y)$
        end if
        $j \leftarrow j + s$
    end for
end function
EXPAND(-1)
EXPAND(1)
Return $\mathcal{S}$

label as the anchor frame. As shown in Algorithm 1, we expand the frames at $t-1$ to the anchor frame. We first gather the classification score of three frames around $t-1$ frame. We then calculate the prediction classes for these three frames. If they all have the same predicted class and the classification score for the current frame at class y is above a threshold, we choose to put the current frame into the training frame set S. For all experiments in the current paper, we set $\xi = 0.9$ for fair comparison.

3 Evaluate Classification & Localization Independently

We evaluate our single-frame supervised model and weakly-supervised model in terms of classification and localization independently. We adopt mean average precision (mAP) in [8] to evaluate the video-level classification performance and AP at different IoU thresholds to evaluate the class-agnostic localization quality regardless of the action class. We report the video-level classification mAP in Table 1, showing only marginal gain as expected. This is because THUMOS14 only contains one or two action classes in a single video which makes the video be easily classified into the target action category. We also evaluates boundary detection AP regardless of the label in Table 1, showing large gain after adding single-frame supervision.

Table 1. Classification accuracy and class-agnostic localization AP on THUMOS14.

	Classification	Class-agnostic localization		
	mAP	AP@IoU=0.3	AP@IoU=0.5	AP@IoU=0.7
Ours w/o single-frame	97.8	42.1	18.1	5.5
Ours w/ single-frame	**98.5**	**58.8**	**32.4**	**9.4**

Table 2. The background η analysis on THUMOS14. AVG is the average mAP at IoU 0.1 to 0.7.

η	mAP@hit	mAP@IoU					
		0.1	0.3	0.5	0.6	0.7	AVG
0.0	44.4±0.56	58.6±0.55	41.1±0.80	20.2±0.69	12.9±0.58	7.3±0.10	31.7±0.47
1.0	57.7±0.41	68.3±0.37	51.1±0.57	28.2±0.52	17.7±0.09	9.4±0.31	39.3±0.13
3.0	60.6±1.36	71.0±1.21	53.8±0.71	29.3±1.14	18.9±0.88	9.4±0.43	41.1±0.80
5.0	60.6±0.85	70.6±0.92	53.7±1.21	29.1±0.39	19.1±1.31	10.2±0.84	41.1±0.78
7.0	60.9±0.56	70.7±0.08	54.3±1.18	29.5±0.13	19.0±0.50	10.1±0.27	41.3±0.44
9.0	60.2±1.12	70.3±0.83	53.4±0.8	29.6±0.58	18.8±0.99	10.1±0.37	41.0±0.60

Table 3. The loss coefficients analysis on THUMOS14. AVG is the average mAP at IoU 0.1 to 0.7.

parameter	mAP@hit	Segment mAP@IoU					
		0.1	0.3	0.5	0.6	0.7	AVG
$\alpha = 0.2$	61.9±0.34	71.6±0.73	54.2±1.31	29.3±0.47	18.4±0.62	9.7±0.35	41.3±0.56
$\alpha = 0.5$	61.9±0.68	71.8±0.36	54.4±0.68	30.2±0.41	19.3±0.92	10.2±1.14	41.9±0.47
$\alpha = 0.8$	60.7±0.95	71.0±0.40	53.8±0.64	29.4±0.26	19.0±0.23	10.0±0.25	41.2±0.22
$\beta = 0.2$	60.6±1.55	70.5±1.21	53.2±1.09	29.4±0.64	18.8±0.71	9.7±0.33	41.0±0.67
$\beta = 0.5$	60.2±0.69	70.5±0.55	53.7±0.71	29.4±0.16	18.8±0.47	10.0±0.34	41.1±0.42
$\beta = 0.8$	60.8±1.05	70.6±0.50	53.8±1.47	29.6±0.34	18.9±0.36	10.0±0.37	41.2±0.55

4 Sensitivity Analysis

4.1 Background Ratio

Table 2 shows the results with respect to different background ratios η on THUMOS14. The mean and standard deviation of segment and frame metrics are reported. We ran each experiment three times. The single-frame annotation for each video is randomly sampled from annotations by four annotators. From the table 2, we find that our proposed SF-Net boosts the segment and frame evaluation metrics on THUMOS14 dataset with background mining. The model becomes stable when the η is set in range from 3 to 9.

Supervision	Method	mAP @IoU			
		0.5	0.7	0.9	AVG
Full	CDC [5]	45.3	-	-	23.8
Full	SSN [9]	41.3	30.4	13.2	28.3
Weak	UntrimmedNet [7]	7.4	3.9	1.2	3.6
Weak	AutoLoc [6]	27.3	17.5	6.8	16.0
Weak	W-TALC [2]	37.0	14.6	-	18.0
Weak	Liu *et al.* [3]	36.8	-	-	**22.4**
Weak	3C-Net [4]	**37.2**	**23.7**	**9.2**	21.7
Single-frame	SF-Net (**Ours**)	37.8	24.6	10.3	22.8

Table 4. Segment localization results on ActivityNet1.2 validation set. The AVG indicates the average mAP from IoU 0.5 to 0.95.

4.2 Loss coefficients

We also conduct experiments to analyze the hyper-parameters of each loss item on the THUMOS14 in Table 3. The mean and standard deviation of segment and frame metrics are reported. We ran each experiment three times. The single-frame annotation for each video is randomly sampled from annotations by four annotators. The default values of α and β are 1. We change one hyper-parameter and fix the other one. From the Table 3, we observe that our model is not sensitive to the hyper-parameters.

5 ActivityNet Simulation Experiment

We conduct experiments on ActiviytNet1.2 by randomly sampling single-frame annotations from ground truth temporal boundaries. Table 4 presents the results on ActivityNet1.2 validation set. In this experiment, the annotations are generated by randomly sampling single frame from ground truth segments. We follow the standard evaluation protocal [1] by reporting the mean mAP scores at different thresholds (0.5:0.05:0.95). On the large scale dataset, our proposed method can still obtain a performance gain with single frame supervision.

6 Qualitative Results

We present the qualitative results on BEOID dataset in Figure 2. The first example has two action instances: *scan card* and *open door*. Our model localizes every action instance and classifies each action instance into the correct category. The temporal boundary for each instance is also close to the ground-truth annotation despite that we do not have any temporal boundary information during training. For the second example, there are three different actions and total four action instances. Our SF-Net has detected all the positive instances in the videos. The drawback is that the number of detected segments for each action class is

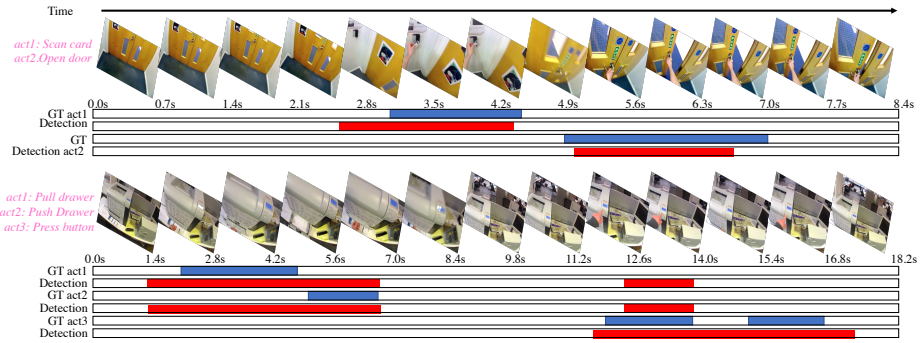

Fig. 2. Qualitative Results on BEOID dataset. GT denotes the ground truth and the action segment is marked with blue. Our proposed method detects all the action instances in the videos.

greater than the number of ground truth segments. To better distinguish actions of different classes, the model should encode the fine-grained action information from the target action area instead of the 1D feature directly extracted from the whole frame. We will consider this in the future work.

References

1. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015)
2. Ding, L., Xu, C.: Weakly-supervised action segmentation with iterative soft boundary assignment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6508–6516 (2018)
3. Liu, D., Jiang, T., Wang, Y.: Completeness modeling and context separation for weakly supervised temporal action localization. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
4. Narayan, S., Cholakkal, H., Khan, F.S., Shao, L.: 3c-net: Category count and center loss for weakly-supervised action localization. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019)
5. Shou, Z., Chan, J., Zareian, A., Miyazawa, K., Chang, S.F.: Cdc: Convolutional-deconvolutional networks for precise temporal action localization in untrimmed videos. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
6. Shou, Z., Gao, H., Zhang, L., Miyazawa, K., Chang, S.F.: Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 154–171 (2018)
7. Wang, L., Xiong, Y., Lin, D., Van Gool, L.: Untrimmednets for weakly supervised action recognition and detection. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 4325–4334 (2017)
8. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Val Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: ECCV (2016)
9. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., Lin, D.: Temporal action detection with structured segment networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2914–2923 (2017)