Supplementary Material

Table 1. Comparisons of the annotation time cost under two modes. FM indicates the new developed frame mode, and VM represents the conventional video mode.

Task	samples	FM	VM
Assemble Bed	6	6:55	23:30
Boil Noodles	5	3:50	18:15
Lubricate A Lock	2	1:23	5:29
Make French Fries	6	5:57	20:24
Change Mobile Phone Battery	2	2:23	7:35
Replace A Bulb	2	1:30	6:40
Plant A Tree	2	1:45	6:37
Total	25	23:43	88:30

1. Annotation Time Cost Analysis

In section 3, we have introduced a toolbox for annotating COIN dataset. The toolbox has two modes: frame mode and video mode. The frame mode is new developed for efficient annotation, while the video mode is frequently used in previous works [1]. We have evaluated the annotation time on a small set of COIN, which contains 25 videos of 7 tasks. Table 1 shows the comparison of annotation time under two different modes. We observe that the annotation time under the frame mode is only 26.8% of that under the video mode, which shows the advantages of our developed toolbox.

2. Browse Times Analysis

In order to justify that the selected tasks meet the need of website viewers, we display the number of browse times across 180 tasks in Figure 1. We searched "How to" + name of 180 tasks, e.g., "How to Assemble Sofa", on YouTube respectively. Then we summed up the browse times of the videos appearing in the first pages (about 20 videos) to get the final results. "Make French Fries" is the most-viewed task, which has been browsed 1.7×10^8 times. And the browse times per task are 2.3×10^7 on average. These results demonstrate the selected tasks of our COIN dataset satisfy the need of website viewers, and also reveal the practical value of instructional video analysis.

3. Visualization Results

In section 5.1, we have visualized the step localization results of different methods and ground-truth annotations. Figure 2 shows more examples of task "install bicycle rack" and "make paper windmill". When applying our task-consistency method, we can discard the steps which do not belong to the correct task, e.g., "jack up the car" in the task "install the bicycle rack" and "crop the paper" in the task "make paper windmill". For more visualization results, please see the uploaded video.

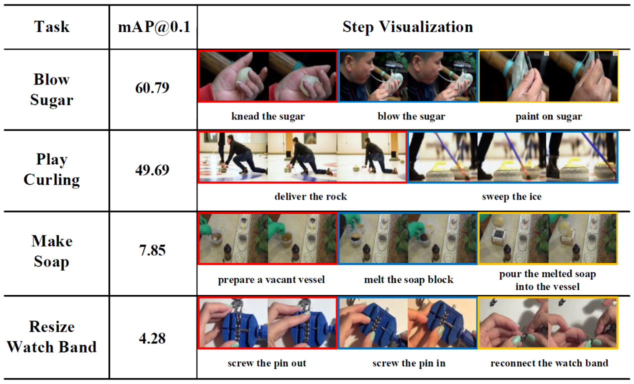

Figure 3. Comparisons of the step localization accuracy (%) of different tasks. We report the results obtained by SSN+TC$_{-Fusion}$ with $\alpha = 0.1$.

4. Step Localization Results of Different Tasks

In section 5.3, we have compared the performance across different domains. Figure 3 shows some examples from 4 different tasks as "blow sugar", "play curling", "make soap" and "resize watch band". They belong to the domain "sports", "leisure & performance", "gadgets" and "science and craft", which are the two of the easiest domains and the two of the hardest domains. For "blow sugar" and "play curling", different steps vary a lot in appearance, thus it is easier to localize them in videos. For "make soap" and "resize watch band", various steps tend to occur in similar scenes, hence the mAP accuracy of these tasks are inferior.

References

[1] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *ICCV*, pages 706–715, 2017. 1

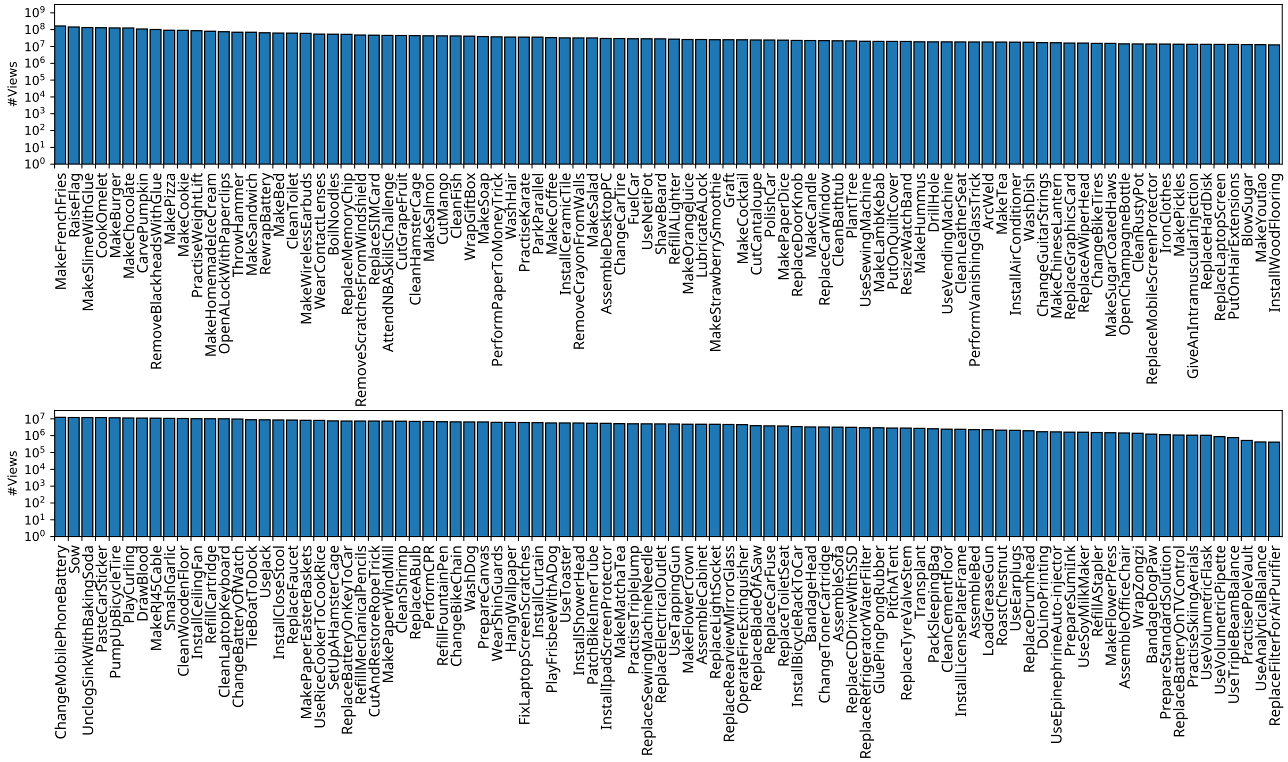

Figure 1. The browse time distributions of the selected 180 tasks on YouTube.

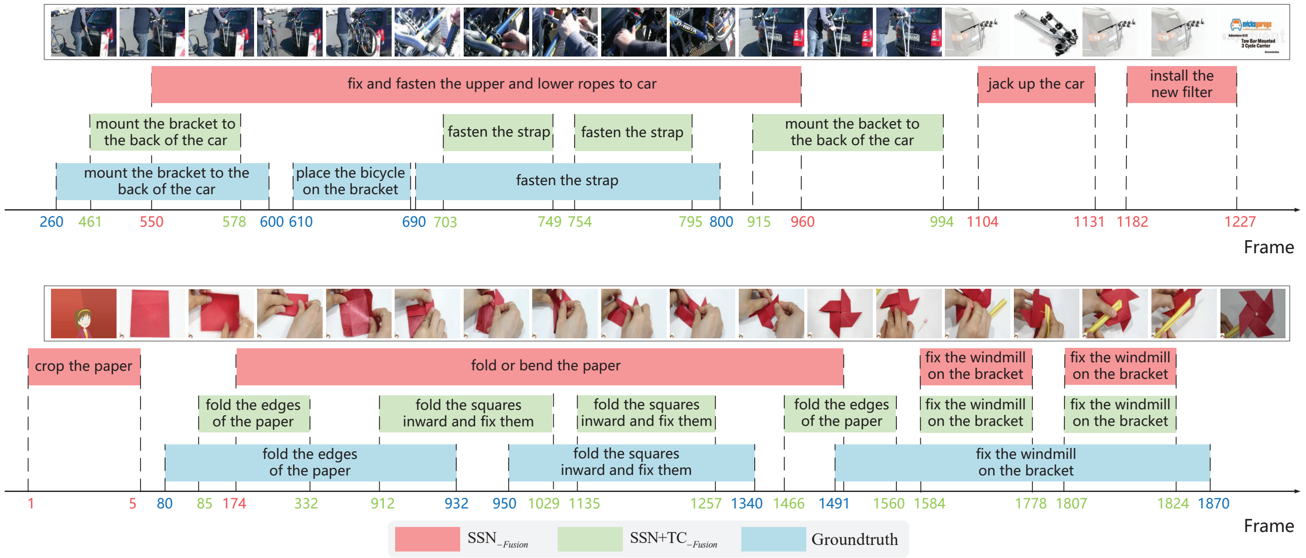
fix and fasten the upper and lower ropes to car
jack up the car
install the new filter
mount the bracket to the back of the car
fasten the strap
fasten the strap
mount the bracket to the back of the car
mount the bracket to the back of the car
place the bicycle on the bracket
fasten the strap
260
461
550
578
600
610
690
703
749
754
795
800
915
960
994
1104
1131
1182
1227
Frame
crop the paper
fold or bend the paper
fix the windmill on the bracket
fix the windmill on the bracket
fold the edges of the paper
fold the squares inward and fix them
fold the squares inward and fix them
fold the edges of the paper
fix the windmill on the bracket
fix the windmill on the bracket
fold the edges of the paper
fold the squares inward and fix them
fix the windmill on the bracket
1
5
80
85
174
332
912
932
950
1029
1135
1257
1340
1466
1491
1560
1584
1778
1807
1824
1870
Frame
SSN$_{-Fusion}$
SSN+TC$_{-Fusion}$
Groundtruth

Figure 2. Visualization of step localization results. The videos are associated with the task "install the bicycle rack" and "make paper windmill".