We thank all reviewers for the insightful comments and suggestions. We appreciate the positive comments on 1) the technically sound method that achieves close accuracy to the state-of-the-art methods in an extremely fast speed, 2) very thorough experimental evaluation, 3) very clearly written and thoroughly convincing of this paper, and 4) moderate algorithmic novelty with detailed attention to comparing different methods. We believe the remaining issues can be fully addressed. We will revise this paper according to reviewer 1's constructive suggestions. We respectfully disagree reviewer 2's misunderstanding of our contribution and framework.

For reviewer_1

C1 (abbr. for Comment): why non-consecutive sampling method works

We very appreciate this question. Our experiment indeed shows that non-consecutive sampling method obtains comparable performance with consecutive sampling method. We speculate that this sampling method can well model the global information of the entire video, which results in promising performance. Next we will explore more segment sampling methods to further model the overall video.

C2: real-time action recognition and online action recognition

Thanks sincerely for the question. Online action recognition has been extensively studied in the past few years. Our method mainly focuses on accelerating the video process speed. We do not apply the proposed method to solve the online action recognition. However, our method is able to conduct online recognition as video frames are received. We will add discussion about utilizing our method for online action recognition in the camera-ready paper.

For reviewer_2

C1 and C2: what makes the proposed system faster than the existing 3D CNNs.

Thanks for your question. We describe the balance between speed and accuracy in detail in the Section 4.4. Besides, as Table 5 shows, our framework uses around 20% sampling number of traditional methods (S clips vs all clips) to make the final prediction, which results in a faster inference speed. S is the number of clips adopted by our framework, which is set to 3 in the experiments. More importantly, our network is trained with S clips' aggregated scores. Thus, feeding S clips into the network is suitable to capture the feature of video. As a result, our method can obtain a higher accuracy at a faster speed than the existing 3D CNNs.

C3 and Q1 (abbr. for Question): the main contribution only is RGB input.

Thanks sincerely for the question. However, the reviewer may have some misunderstandings about our contributions. I hope to make the contributions clearer as below.

First, RGB input is definitely not our contribution. The popular two-stream action recognition architecture exploits two CNNs to model RGB and optical flow respectively. However, Two-stream CNNs cannot process videos at real-time since the optical flow is computationally expensive. Owing to the contributions of the proposed framework, our method can achieve higher accuracy at a faster speed than the two-stream architectures even only with the input of RGB frames.

The main contribution of this paper is proposing a Temporal Convolutional 3D Network recognition framework which learns video representations in a hierarchical multi-granularity manner. The framework contains three important components：1) high-level video sampling; 2)

residual 3D-CNN; and 3) temporal encoding technique. More details:

1) High-level video sampling. The traditional methods always use the dense sampling method. Differently, we explore to sample the video segments with a small part of frames, which avoid extracting most video frames.

2) Residual 3D-CNN. We feed the sampling video segments into the residual 3D-CNN to extract the fine-grained spatial and temporal information. It is much faster than optical flow based methods.

3) Temporal encoding technique. After sampling multiple segments, the temporal encoding technique can calculate the multi-clip aggregated loss rather than a single one. Thus, the temporal encoding technique drives the network to learn the video-level information.

The combination of the three components achieves significantly better performance than state-of-the-art real-time methods by over 5.4% in terms of accuracy and 2 times faster in terms of inference speed. To make the paper more solid, we will release the codes.

Q2: the illustration of Figure 2

Thanks sincerely for the question. A detailed description of Figure 2 can be found at the beginning of Section 3.

Figure 2 shows the pipeline of the proposed network. Given a video to be recognized, we firstly divide it into S (e.g., 3) parts. Then several frames are selected from each part to make up a clip. Next, S clips represented the entire video are fed into the residual 3D-CNN. Furthermore, feature maps and category scores of S clips are fused by an aggregation function to yield the video-level score. Finally, the proposed method calculates the loss value from the video-level score, which drives the residual 3D-CNN to learn the video-level feature.

Q3: the meaning of Q in the Eq. (1)

We appreciate this question. We explicitly define the Q in line 2 on page 4. Q denotes one of the four pooling methods described in Section 3.4. In table 3, we also summarize the results of different Q functions.

Q4: the purpose to elaborate four pooling methods

We appreciate this question. Based on Eq. (1), the pooling method (also known as the aggregation function) is a key factor to model the overall features of videos. We elaborate pooling methods to find the best aggregation function to achieve the promising accuracy. In addition, investigating pooling methods is one of the common-use evaluations in the video-based action recognition.

Q5: typos and grammatical errors

We will improve the writing and polish the paper with the assistance of a native speaker.