FineAction: A Fine-Grained Video Dataset for Temporal Action Localization

Yi Liu, Limin Wang, Member, IEEE, Yali Wang, Member, IEEE, Xiao Ma, and Yu Qiao, Senior Member, IEEE

Abstract—Temporal action localization (TAL) is an important and challenging problem in video understanding. However, most existing TAL benchmarks are built upon the coarse granularity of action classes, which exhibits two major limitations in this task. First, coarse-level actions can make the localization models overfit in high-level context information, and ignore the atomic action details in the video. Second, the coarse action classes often lead to the ambiguous annotations of temporal boundaries, which are inappropriate for temporal action localization. To tackle these problems, we develop a novel large-scale and fine-grained video dataset, coined as FineAction, for temporal action localization. In total, FineAction contains 103K temporal instances of 106 action categories, annotated in 17K untrimmed videos. Compared to the existing TAL datasets, our FineAction takes distinct characteristics of fine action classes with rich diversity, dense annotations of multiple instances, and co-occurring actions of different classes, which introduces new opportunities and challenges for temporal action localization.

To benchmark FineAction, we systematically investigate the performance of several popular temporal localization methods on it, and deeply analyze the influence of fine-grained instances in temporal action localization. As a minor contribution, we present a simple baseline approach for handling the fine-grained action detection, which achieves an mAP of 13.17% on our FineAction. We believe that FineAction can advance research of temporal action localization and beyond. The dataset is available at https://deeperaction.github.io/datasets/fineaction.

Index Terms—Temporal action localization, fine-grained, video understanding, deep learning.

I. INTRODUCTION

Video understanding [1], [2], [3], [4], [5], [6], [7], [8] is becoming an important problem in the computer vision community due to its wide applications in video content analysis, intelligent surveillance, and human-computer interaction. Temporal action localization (TAL) is a fundamental task in this research, which aims to localize the start and end time of each action instance in the untrimmed videos. The advance in this area has been mainly driven by deep learning models [9], [10], [11], [12], [13] with large-scale benchmarks such as THUMOS14 [14], ActivityNet [15], and HACS Segment [16].

However, the existing benchmarks have several limitations that may hinder the development of TAL methods. First, some datasets only contain sparse annotations of coarse-level action classes (e.g., ActivityNet and HACS Segment), where each action instance could occupy a large portion of the whole video. We argue that this fact fails to reflect the general and realistic requirement of TAL, such as localizing dense action instances in a video, determining more accurate temporal duration for fine-grained actions, handling co-occurring actions of different classes. As shown in Figure 1, for example, ActivityNet contains a single action instance of class Futsal, which covers a long duration in the video. We observe that it is inappropriate to localize such high-level event as there is no clear definition on temporal boundary of Futsal. Instead we can only accurately localize more fine-grained actions such as Pass soccer, Keep goal and Shot soccer due to its more accurate definition in time. In addition, detecting these fine-grained actions will pose more challenges to TAL methods with requirement of focusing on action itself rather than background context. Second, some action datasets (e.g., THUMOS14) contains small number of action classes with detailed annotation and simply focuses on a specific domain such as sport activities. The small number of training videos would increase the overfitting risk in model development. Meanwhile, it is unclear whether these methods that perform well on sports action temporal localization could be still effective for other daily life actions. Thus, it is expected...
to increase the scale and diversity of fine-grained TAL datasets to promote the future research.

Based on the discussion above, we think it is necessary to establish a new benchmark for TAL by proposing the following criteria. 1) We should focus on more fine-grained actions with clear temporal boundaries, which is more suitable for the task of TAL. Meanwhile, localizing such fine-grained actions can provide more detailed understanding of video content. 2) We should perform relatively dense annotations of fine-grained action classes, possibly assigning multiple labels to overlapped segments, as different fine-grained actions happen simultaneously. For example, *Keep goal* and *Shot soccer* are coupled in the soccer videos, as shown in Figure 1. These dense annotations will pose new challenges to TAL methods and are more similar to realistic applications such as sports analysis. 3) Instead of working on specific sports actions, this dataset should be large scale and cover various complex human activities in our daily life.

To achieve this goal, we introduce a large-scale and fine-grained video dataset, coined as *FineAction*, for the task of TAL. Different from the existing datasets, our FineAction is unique in aspects of fine-grained action classes, multi-label and dense annotations, relatively large-scale capacity, and rich action diversity. In total, FineAction contains 103K temporal instances of 106 action categories annotated in 17K untrimmed videos. The whole construction process consists of three critical stages. First, we define a new class taxonomy with three levels of category granularity. Our taxonomy excludes coarse-level action definition, and categorizes the highly-related actions into the same subordinate. As a result, this taxonomy can effectively organize action categories according to its fine-grained characteristics in a hierarchical manner. Second, in order to increase data scale as well as diversity, we progressively collect videos from both existing benchmarks and Internet. Third, we develop a customized annotation tool with temporal annotation guidance, so that annotators leverage image-text illustration to clearly understand fine-grained actions, and efficiently annotate temporal boundaries of action instances in a video. We also conduct annotation quality inspection with multiple rounds of professional checking. Finally, we systematically study a number of popular TAL methods on our new benchmark. Our experiments show that their performance on our FineAction is much worse than other benchmarks, demonstrating that our FineAction brings new challenges for the task of TAL, due to its fine-grained characteristics and detailed annotations.

In summary, FineAction benchmark contributes to the research of TAL in three different ways: 1) The definition of action categories is much more fine-grained with a distinct three-granularity taxonomy. 2) We provide effective annotation guidance and carefully control the annotation quality, trying to reduce the bias of annotations and the boundary uncertainty of action instances. 3) We propose an initial solution with multi-scale transformer neck to address the fine-grained action localization and the proposed baseline achieves the mAP of 13.17% on our FineAction. 4) We conduct in-depth studies on FineAction, which reveals the key challenges that arise in the localization of fine-grained instances. We expect that our dataset can facilitate the future research of fine-grained TAL.

II. RELATED WORKS

A. Action Datasets

Action recognition aims at classifying action in a trimmed video clip. Due to its potential value in realistic applications, it has been widely explored in the past decades, from handcrafted methods [17], [18] to deep learning models [3], [5], [19]. In particular, more challenging datasets were proposed subsequently, including UCF101 [4], Kinetics [19], ActivityNet [15], Moments in Time [20], and others [1], [14], [16], [21], [22]. With fast development of these large-scale benchmarks, deeply-learned video models exhibit high accuracy by learning spatial and temporal evolution of actions.

Unlike early datasets of action recognition, which mainly focus on action classification, more challenging datasets are proposed for other tasks such as temporal action detection and spatio-temporal action detection. Temporal action detection in untrimmed videos is crucial to inspection of massive videos uploaded on the Internet in real world. Several datasets with fine granularity of classes have been presented in the narrow domains. For example, THUMOS Challenge 2014 [14] includes 413 untrimmed videos on 20 sport actions, which was extended into Multi-THUMOS [21] with 65 action classes. FineGym V1.0 [23] provides temporal annotations with a three-level semantic hierarchy for gymnastic videos. Other datasets include MPII Cooking [24], [25] and EPIC-Kitchens [26] mainly focus on scene in the kitchen. Models trained on such domain-specific datasets may not generalize well to daily activities. Conversely, some datasets like ActivityNet-v1.3 [15] and HACS Segment [16] were designed to include more general, every-day activities. ActivityNet-v1.3 [15] includes 20K untrimmed videos with 23K temporal action annotations, and HACS Segment [16] contains 49k untrimmed videos with 122k temporal action annotations. But these datasets are lack of fine-grained annotations for daily activities.

Spatio-temporal action detection datasets, such as UCF Sports [27], UCF101-24 [4], J-HMDB [28], DALY [29],AVA [30] and AVA-Kinetics [31], typically evaluate spatio-temporal action detection for short videos with frame-level
Fig. 2. Complete organizational taxonomy behind our FineAction, which consists of 4 top-level categories, 14 middle-level categories, and 106 bottom-level categories.

action annotations. These benchmarks pay more attention to spatial information with frame-level detectors and clip-level detectors, which are limited to fully utilize temporal information. Recently, MultiSports [32] presented a spatio-temporal action detection dataset on sports actions. Alternatively, we mainly focus on temporal localization of detailed human action instances in this work.

B. Temporal Action Localization

Temporal action localization aims at localizing the start and end points of action clips from the entire untrimmed video with full observation. Among temporal action detection methods, anchor-based action localization methods [33], [34], [35], [36], [37], [38], [39] in a top-down pipeline, try to predict confident of action segments on default anchors and refine the action boundaries. For example, [38], [39] transforms Faster R-CNN [40] for temporal action localization. On the other hand, actionness-based action localization methods [10], [11], [41], [42] in a bottom-up pipeline, which predict the frame-level actionness score to discover actions boundaries. For example, [10] leverages actionness as guidance to obtain temporal intervals and then classifies these intervals by learning temporal structure and action completeness. Reference [42] generates proposals via learning starting and ending probability using a temporal convolutional network. Besides, there are methods based on reinforcement learning [43], gated recurrent unit [44], [45] and graph convolutional network [46], [47]. Recently [48] presents a simple and end-to-end
learnable framework for direct action proposal generation, by re-purposing a Transformer-alike architecture.

III. FineAction Dataset

Our goal is to build a large-scale and high-quality dataset with fine-grained action classes and dense annotations for temporal action localization. In this section, we present the action taxonomy, the collection and annotation process, the dataset statistics, and the dataset properties.

A. Dataset Preparation

First, we build up a pool of action categories from the existing benchmarks for video collection and annotation. Since ActivityNet [15] and Kinetics [19] contain a wide range of action categories from sports to daily activities, we choose the union set of their action categories as our pool to increase class diversity. Second, to select suitable action classes with clear temporal boundaries, we conduct filtering based on this pool. Note that, fine-grained classes refer to subordinate-level categories in the superior class, and their inter-class differences are subtle. Base on this principle, we carefully define fine-grained action classes with two following cases. (1) if an action can be decomposed into subordinate-level actions by Wikipedia and/or professional manuals, it will be chosen as coarse-grained action to be decomposed (e.g., layup is broken down into dribbling, passing ball, dunking). (2) if several classes are often confused in the research of action recognition, these classes are also picked as our fine-grained classes due to subtle inter-class differences (e.g., play trumpet vs play saxophone, ride a bike vs ride scooter).

1) Action Taxonomy: Based on these rules, we finally generate 106 action classes within a new taxonomy of three-level granularity. Figure 2 shows the full organizational taxonomy behind FineAction. This taxonomy consists of 4 top-level categories (Household Activities, Personal Care, Socializing, Relaxing and Sports, Exercise), 14 middle-level categories, and 106 bottom-level categories. It is worth noting that, our built taxonomy is different from the one in ActivityNet [15], and HACS Segments [16], due to its fine-grained characteristics. For example, Makeup is actually a middle-level action that consists of 7 bottom-level actions in our taxonomy, as shown in Figure 3 (a), while it is a bottom-level action in ActivityNet. In practice, thanks to our more specific action definition, our taxonomy is more reasonable and suitable for temporal action localization in videos.

2) Video Collection: Based on our action taxonomy, we progressively collect videos from both existing benchmarks and Internet. First, we manually select videos from the related categories of the existing video datasets, i.e., YouTube8M [49], Kinetics400 [19], and FCVID [50]. Note that, for the bottom-level action categories that do not appear in these datasets, we use the corresponding middle-level action categories for selection. Second, we annotate these videos (the procedure is described in the next section), and count the number of instances per category. If the number of instances is smaller than 200 in one bottom-level category, we use keyword to crawl YouTube videos to increase instances in this action. Finally, we remove video duplicates by computing the pairwise video similarity with deep features. As for reproducibility, we will release our dataset with raw videos & URLs, annotation file, and extracted features. But like ImageNet, our release is based on request, where researchers must submit the license form and agree to use this dataset for only academic purpose, before they can access to it.

B. Dataset Annotation

After data collection, we perform manual annotation of action instances for each video. Specifically, given a video, the trained annotator is asked to find the start and the end of each action instance as well as associating action tag with this instance. To boost annotation efficiency and consistency, we design an annotation tool as shown in Figure 3 (c). First, the annotator loads a video into video browsing area, for quickly previewing the entire video. Second, the annotator...
carefully checks the video frames in frame selection area, and determines the start and the end of each action instance. Third, the annotator operates various buttons (e.g., adding, deleting, and modifying labels) of operation menu area for detailed annotation. Finally, the annotator makes a double check for the annotation results on label display area.

We design a rigorous guidance and checking procedure to ensure the annotation quality. First, we develop a guide to give the specific definition of each action class, which can ensure the annotation consistency of temporal boundaries among different annotators. Specifically, we provide the start and end description of each action category with image-text definition. For sport activities, we use their definition and explanation from Wikipedia and/or professional sport manuals. For other daily activities, a team of professional researchers in TAL are asked to summarize the clear boundary definitions to avoid annotation bias. Figure 3 (b) shows an example of Paint Eye Shadow in our annotation guide. Second, we arrange another annotator team to check all the annotated videos in multiple rounds to increase annotation quality. Specifically, this team will calculate the annotation statistics of each category from different annotators, including the average frequency of the action, the average duration of the action, etc. If different annotators have large statistical variations for the same category, it means that this category tends to be controversial in the annotation process, due to cognitive bias between different annotators. In this case, our evaluation team will double-check all the boundary annotations of this category, by discussing these videos with professional TAL researchers. Finally, the professional researchers are asked to make the sampling inspection to further alleviate inaccurate and/or inappropriate annotations.

C. Dataset Statistics

We first compare our FineAction with the large-scale datasets in TAL, i.e., ActivityNet [15] and HACS Segment [16]. As shown in Table I, FineAction has the comparable data size but exhibits more fine-grained properties than ActivityNet and HACS Segment. For example, our FineAction has multi-label and dense annotations rather than one instance per video and the fine-grained actions often happen in a short period.

Hence, the average duration of temporal instance in our FineAction (7.1s) is much shorter than the one in ActivityNet (49.2s) and HACS Segment (33.2s). Moreover, we further make comparison in terms of instance-duration distribution in Table II. The distribution of our FineAction is different from that of ActivityNet and HACS Segments, i.e., FineAction contains much more short-duration action instances (66,890) than the one in ActivityNet (900) and HACS Segment (8,874) for 0-2s instances. On one hand, it requires much more effort to label such elaborate annotations. On the other hand, it brings much larger challenge for temporal action localization. Note that, the average number of action instances per video is 1.15/2.47/6.17 for ActivityNet/HACS Segment/FineAction. Hence, our FineAction is more densely annotated among these large-scale datasets. Finally, several fine-grained actions can happen simultaneously. Hence, 11.5% of temporal segments have multiple action labels with overlaps in our FineAction, while all the temporal segments are tagged with a single label in ActivityNet and HACS Segment. Hence, these multi-label annotations would greatly increase the challenge of our FineAction.

Then, we compare our FineAction with other fine-grained temporal datasets. As shown in Table I, most of these existing benchmarks focus on narrow domains such as sports (THUMOS14 [14] and FineGym V1.0 [23]) or kitchen activities (MPII Cooking [24] and EPIC-Kitchens [26]), while our FineAction contains multiple types of activities with rich diversity. In addition, the recent large-scale fine-grained datasets mainly work on other video tasks, e.g., action recognition (FineGym V1.0) and anticipation (EPIC-Kitchens). THUMOS14 is the most similar dataset to our FineAction on TAL. However, this dataset is quiet small with limited number of videos and action instances. These comparison demonstrates that our FineAction is unique with distinct properties of fine-grained action classes, multi-label and dense annotations, relatively large-scale capacity, and rich action diversity.

Finally, we show the number of instances for each action category in Figure 4. For each top-level category, we visualize

| Database          | Category | M-L | Video | Instance | Overlap | Duration | Action type | Main task          |
|-------------------|----------|-----|-------|----------|---------|----------|-------------|--------------------|
| MPII Cooking [24] | 65       | ✓   | 45    | 5,609    | 0.1%    | 11.1 m   | kitchens    | Action Classification |
| EPIC-Kitchens [26]| 4,025    | ✓   | 700   | 89,979   | 28.1%   | 3.1 s    | kitchens    | Action Classification |
| FineGym V1.0 [23] | 530      | ✓   | 303   | 32,697   | 0.0%    | 1.7 s    | sports      | Action Classification |
| THUMOS14 [14]     | 20       | x   | 413   | 6,316    | 17.5%   | 4.3 s    | sports      | Temporal Action Localization |
| ActivityNet [15]  | 200      | x   | 19,994| 23,064   | 0.0%    | 49.2 s   | daily events| Temporal Action Localization |
| HACS Segment [16] | 200      | x   | 49,485| 122,304  | 0.0%    | 33.2 s   | daily events| Temporal Action Localization |
| FineAction        | 106      | ✓   | 16,732| 103,324  | 11.5%   | 7.1 s    | daily events| Temporal Action Localization |

| Database          | 0-2 s | 2-6 s | 6-15 s | >15 s | Ins / Vid |
|-------------------|-------|-------|--------|-------|-----------|
| THUMOS14 [14]     | 2,028 | 2,753 | 1,437  | 99    | 15.29     |
| ActivityNet [15]  | 900   | 3,253 | 4,426  | 14,485| 1.15      |
| HACS Segment [16] | 8,874 | 29,644| 31,982 | 51,804| 2.47      |
| FineAction        | 66,890| 15,253| 10,523 | 10,586| 6.17      |
the instance distribution over the corresponding bottom-level categories. The number of instances exhibits the natural long-tailed distribution, which also arises new challenges and opportunities for temporal action detection.

D. Dataset Properties

1) Dataset Challenges: Compared with the existing benchmarks, our FineAction dataset shares the following distinct challenges. 1) **Fine-grained action categories with a long-tailed distribution.** It is difficult to recognize these detailed actions, if the models mainly focus on distinguishing background and action categories without learning subtle motion patterns. 2) **Densely annotated instances with short temporal duration.** It leads to the difficulties in precisely localizing temporal boundaries, since various short actions are densely distributed in the entire video. 3) **Temporal segments with multiple action labels in overlaps.** It is challenging to recognize different actions in a single segment, if the models are designed without concurrent action analysis and/or action relation learning. 4) **Diverse action classes with a wide range of semantics.** Without specific prior knowledge, it is hard to design TAL models to perform well on various fine-grained human actions.

2) **High Quality:** FineAction is progressively annotated in a careful and efficient procedure. Our annotation teams are rigidly supervised by professional researchers. With our customized annotation tool and guidance, action instances are correctly labeled with precise temporal boundaries. Besides, our researchers make multiple calibrations to remove annotation bias as much as possible.

3) **Richness and Diversity:** From the perspective of action definition, FineAction establishes a new taxonomy which contains fine-grained actions from sports to daily activities. Hence, it is a preferable benchmark to understand detailed human actions in the realistic scenarios. From the perspective of data collection, FineAction is collected from various data resources via rigid quality control. Such richness brings new opportunity to develop powerful deep learning models for TAL.

IV. Experiment

Our goal is to show challenges and values of FineAction in TAL research. Hence, we mainly focus on systematical investigation of widely-used SOTA approaches in this paper. Via extensive ablation studies and rigorous error analysis, we aim to sufficiently explore core contribution of our FineAction, and provide insightful findings that are beneficial for future research of TAL. Due to the quality of video features can seriously affect the performance of TAL, as our proposed TAL network takes video features as input. We firstly carried out the experiment of untrimmed video classification task by using different vision backbone.

A. Dataset Split

**FineAction benchmark.** To build a solid TAL benchmark, we manually split the videos into the training set (50%), validation set (25%), and testing set (25%). In total, FineAction contains 57,752 training instances from 8,440 videos, 24,236 validation instances from 4,174 videos and 21,336 testing instances from 4,118 videos. Unless otherwise mentioned, we report the results by training on the training set and testing on the validation set. We withhold the annotations of testing set in the public release and encourage to submit the testing results to our server for evaluation.
B. Untrimmed Video Classification

1) Evaluation Metrics: As for the untrimmed video classification task with multi-label, we choose Mean Average Precision (mAP) as the main evaluation metric.

2) Results: Based on our FineAction benchmark, it is natural to introduce a classification task on untrimmed videos. In this work, we apply the widely-used video models for task evaluation, including TSN [51], I3D [19], SlowFast [52], and Video Swin Transformer [53]. The results of each model can be seen in Table IV. First, the performance is getting better as the model becomes stronger, e.g., Video Swin Transformer achieves the best classification performance, with the powerful transformer architecture. Second, the overall performance is still weak for this untrimmed video classification task, due to multi-label and fine-grained action clips in FineAction.

3) Implementation Details: All these models are implemented by the MMAction2 [54] framework. For training, we initialize the network weights from the K400 [19] pretrain models and employ an AdamW [55] optimizer for 30 epochs using a cosine decay learning rate scheduler and 2.5 epochs of linear warm-up. A batch size of 64 is used. For inference, we use 2x down-sampling by 1D convolutional network with ReLU to extract the embedded feature $F^1$ as the input of transformer layer. Second, we introduce five transformer layers as the neck architecture, apply LayerNorm (LN) before every block, and add the residual connection after every block. Specifically, we use 2x down-sampling by 1D convolution between every two transformer layers to capture the multi-scale features from different temporal scales. This is given by:

$$ F^{ℓ+1} = \Downarrow (\text{Transformer}_ℓ(F^{ℓ})), \quad ℓ = 1 \ldots L - 1 $$

where $F^ℓ \in \mathbb{R}^{T \times D}$ and $\Downarrow$ is the 2x down-sampling by 1D convolution. Finally, we follow BMN [11] to design the head architecture with Temporal Evaluation Module and Proposal Evaluation Module. In the inference phase, we fuse the boundary probability and the confidence map from multi-scale predictions, for generating the final confidence scores and proposals. The whole architecture is shown in Table V.

3) Results: We report the performance of the state-of-the-art methods whose codes are publicly available, including evaluation metric, similar to ActivityNet dataset, where AN ranges from 0 to 100. For the temporal action localization task, mean Average Precision (mAP) is a conventional evaluation metric, where Average Precision (AP) is calculated for each action category. For our FineAction, mAP with IoU thresholds [0.50:0.05:0.95] is used to compute the average mAP.

2) Baseline Approach: To learn fine-grained action proposals, we introduce a simple but effective multi-scale transformer for FineAction localization [58]. First, given an untrimmed video $V$, we can extract a visual feature sequence $F \in \mathbb{R}^{T \times D}$. Before the multi-scale transformer layers, we use two 1D convolutional network with ReLU to extract the embedded feature $F^1$ as the input of transformer layer. Second, we introduce five transformer layers as the neck architecture, apply LayerNorm (LN) before every block, and add the residual connection after every block. Specifically, we use 2x down-sampling by 1D convolution between every two transformer layers to capture the multi-scale features from different temporal scales. This is given by:

$$ F^{ℓ+1} = \Downarrow (\text{Transformer}_ℓ(F^{ℓ})), \quad ℓ = 1 \ldots L - 1 $$

where $F^ℓ \in \mathbb{R}^{T \times D}$ and $\Downarrow$ is the 2x down-sampling by 1D convolution. Finally, we follow BMN [11] to design the head architecture with Temporal Evaluation Module and Proposal Evaluation Module. In the inference phase, we fuse the boundary probability and the confidence map from multi-scale predictions, for generating the final confidence scores and proposals. The whole architecture is shown in Table V.

3) Results: We report the performance of the state-of-the-art methods whose codes are publicly available, including...
BMN [11], DBG [13], G-TAD [12] and our baseline approach, on two main tasks for temporal action localization. To ensure a fair comparison, we keep the video features on the same scale for all the methods. Table III left shows the results on Action Proposal Generation, in terms of AR@AN with SNMS at IoU thresholds [0.50:0.05:0.95], where SNMS stands for Soft-NMS. Table III right shows the results of Action Detection, in terms of mAP at IoU thresholds [0.50:0.05:0.95]. To assign the global classification results to the proposals, we adopt top-1 video-level classification results generated by TSN [6], and we use confidence scores of BMN proposals for detection results retrieving. It clearly shows that, the performances of these existing methods on our FineAction for the two tasks are comparable to each other, and the modality between RGB and Flow has a certain complementarity. Moreover, our proposed baseline achieves the mAP of 13.17% and this further demonstrates the advantage of the multi-scale transformer neck on distinguishing fine-grained action. To further compare our FineAction with other benchmarks, we present the performance evaluation of BMN method on different benchmarks in Figure 5. The performance on our dataset is far lower than other benchmarks for both action proposal generation and action detection, demonstrating the challenges of our FineAction dataset.

4) Implementation Details: For feature encoding, we utilize two-stream architecture [3] to extract features for video frames. In our experiment, two separate I3D [19] models are pre-trained from consecutive RGB frames and optical Flow frames on Kinetics [19], respectively. I3D model takes non-overlapping snippets of 16 stacked RGB or optical Flow frames as input and extracts 2048-dimensional features for each stream. For the fusion of modality, we concatenate the features of RGB and Flow. By default, we rescale the feature sequence of input videos to length \( L = 100 \) by linear interpolation following [42]. We also investigate the effect of rescaled length in our ablation study. For hyper parameter settings, we have investigated different parameter settings of TAL methods. For three methods, L (window length) is 100, D (maximum duration length) is 50 and SNMS threshold is 0.65. For DBG, \( \epsilon \) of Gaussian function is 0.75. For G-TAD, \( \tau_1 \) and \( \tau_2 \) are 16 and 16.

D. Ablation Studies

1) Study on Cross-Dataset Evaluation: We conduct cross dataset evaluation for temporal action localization. Specifically, we choose BMN [11] as the basic architecture. Subsequently, we use ActivityNet to pretrain BMN, and use FineAction to fine-tune it. Alternatively, we also use FineAction to pretrain BMN, and use ActivityNet to fine-tune it. The results are shown in Table VI. As expected, the mAP performance drops for both cases, compared with learning from scratch. The main reason is that, our FineAction consists of fine-grained actions, while ActivityNet contains coarse-grained actions. They exhibit different data characteristics with domain shift. This brings opportunities and challenges for our FineAction benchmark.

2) Study on Sequence Length: We investigate the length of input feature sequence on the FineAction dataset. We use
BMN [11] as baseline for evaluation and the result is reported in Figure 6. As expected, increasing the sequence length contributes to a better localization performance when $L$ is from 100 to 200, and the performance tends to saturate when $L$ is from 200 to 250. We analyze that BMN can leverage longer sequence to aggregate more detailed context for localization. However, too long sequence may limit the batch size and has a side effect on network training. Hence, appropriately increasing sequence length can improve fine-grained temporal action localization.

3) Error Analysis: We analyze typical errors in FineAction by using BMN method with the protocol in [57]. As shown in Figure 7, Background Error and Localization Error are the two main sources of false positives. We can conclude that the BMN method often generates invalid proposals and incorrect boundaries in our FineAction, perhaps due to the fine-grained and multi-labeled action instances. To further demonstrate it, we visualize the typical failure cases in Figure 8. Note that, Background Error and Localization Error are the two main sources of false positives. We can conclude that many detected
Fig. 9. Localization performance AR@100 on different duration of instances. Our FineAction is more challenging than ActivityNet in different durations.

Table VII

| Categories          | AR@100 | Avg. Duration | Num |
|---------------------|--------|---------------|-----|
| straighten hair     | 1.6    | 1.17 s        | 1,934 |
| apply eyebrows      | 1.7    | 1.13 s        | 3,428 |
| dig hole            | 2.2    | 0.83 s        | 544  |
| freestyle relay     | 88.4   | 53.12 s       | 345  |
| breaststroke        | 94.5   | 53.59 s       | 211  |
| play the harp       | 96.2   | 69.21 s       | 268  |

Fig. 10. Visualization the video clips for the challenging category of apply eyebrows.

Fig. 11. Illustration of the proposed new metric. In this example, the IoU of the ground truth moment and two proposals (Proposal 1 and Proposal 2) are both 0.5.

Table VIII

Comparison of Different SOTA Methods on New Metric at tIoU [0.50:0.05:0.95]. Results are worse on the new metric, which can better evaluate the performance for shorter instances.

| Method   | Modality | AUC old | AUC new |
|----------|----------|---------|---------|
| BMN [11] | RGB      | 17.49   | 16.42   |
|          | Flow     | 18.94   | 17.94   |
|          | RGB+Flow | 19.19   | 18.19   |
| DBG [13] | RGB      | 15.48   | 14.33   |
|          | Flow     | 17.70   | 16.74   |
|          | RGB+Flow | 17.24   | 16.32   |
| G-TAD [12]| RGB     | 16.06   | 14.90   |
|          | Flow     | 17.09   | 15.97   |
|          | RGB+Flow | 17.65   | 16.54   |

4) Which Categories are More Challenging?: After carefully analyzing the typical errors, we delve into which categories are more challenging in our FineAction. First, we divide the instance into four levels according to its duration, followed by 0-2s, 2-5s, 5-10s, >10s, and present the average recall with 100 BMN proposals for each level. As shown in Figure 9, we make comparisons with ActivityNet, which shows our dataset is more challenging in different durations due to its fine-grained and multi-label annotations. Moreover, instances in the 0-2s level have the worst performance, even though this level covers the largest proportion of our FineAction dataset (Table II). It clearly illustrates the localization difficulty in short instances. Moreover, we further show the action categories with top-3 worst and top-3 best localization performance in Table VII. We can see that, the 3 worst categories (straighten hair, apply eyebrows and dig hole) are highly fine-grained with much shorter duration. In Figure 10, we visualize the video clips for one challenging category of apply eyebrows. Note that, both on the spacial and temporal, apply eyebrows is a fine-grained action with short duration and slight movement. Hence, it is quite challenging to detect these actions, even with a large number of training instances.

5) Rethink the Evaluation Metrics for Shorter Instances: As mentioned in Section IV-A, we think that the prevailing evaluation metric AR@AN IoU is unreliable for shorter instances. For the short ground truth, it is preferable to generate a short proposal with the similar scale, instead of producing a long proposal to cover the ground truth blindly. For example, as shown in Figure 11, both Proposal 1 and Proposal 2 meet the IoU=0.5 condition, i.e., these two proposals are regarded equally under the original AR@AN IoU. However, since the temporal boundaries of the Proposal 2 are further from the ground truth, we propose a new metric as follows, which penalizes more on Proposal 2:

\[ \text{IoU}_{\text{new}} = \text{IoU} \cdot \alpha_s \cdot \alpha_e , \]

where \( \alpha_s = 1 - \text{abs}(g^s_t - p^s_t) \) and \( \alpha_e = 1 - \text{abs}(g^e_t - p^e_t) \). Both \( p^s_t \) and \( p^e_t \) are normalized to 0-1 by dividing the whole video length. As shown in Table VIII, the new metric has lower performance than the old metric and is more suitable for
fine-grained annotations. It also shows that action localization for shorter instances still faces with great challenges.

V. CONCLUSION

In this paper, we have proposed a new large-scale and fine-grained video dataset, coined as FineAction, for temporal action localization. FineAction differs from the existing datasets in aspects of fine-grained action definitions, high-quality and dense annotations, co-occurring instances of multiple classes. We have empirically investigated several popular methods for temporal action localization on our FineAction and performed in-depth error analysis, which shows great challenges posed by our dataset for temporal action localization. As for the future research, we think that fine-grained module design might improve the performance over FineAction. We hope our FineAction can facilitate new advances in the field of temporal action localization.

VI. APPENDIX

We apply multi-round user studies to ensure clear definition of action boundary as following two steps:

1. A team of professional TAL researchers summarize boundary definitions of each action, by Wikipedia and/or professional manuals and design a rigorous guidance.

2. Our annotators are asked to label our datasets with this guide. Meanwhile, another annotator team and professional TAL researchers check the annotated videos in multiple rounds to increase annotation quality. As shown in Figure 12, we have listed some action examples with start and end definition from our rigorous guidance. The following is a detailed definition of these categories with Start and End Definition.

001 cut nails. **Start:** The moment when nail clipper touches your nails. **End:** The moment when you take out the nail clipper from your nails.
002 wash feet. Start: The moment when you just put your feet in the basin filled with water. End: The moment when your feet leave from the basin.

003 wash face. Start: The moment when you touch the cheek with water by your hands. End: The moment before you start wiping your face with a towel.

004 tie shoelaces. Start: The moment when your hands touch the shoelaces. End: The moment when your hands leave from shoelaces.

005 brush hair. Start: The moment when you divide the hair into several pieces and prepare to start braiding. End: The moment when the comb leaves from your hair.

006 cut hair. Start: The moment when your scissors touch your hair. End: The moment when the scissors leave from the hair.

007 dry hair. Start: The moment when you hold a towel, ready to wipe the hair or turn on the hair dryer to the hair. End: The moment when you take the towel off your hair or take the hair dryer off.

008 apply eyeliner. Start: The moment when you take out the tool and put it on your eyelid. End: The moment when you take off the tool from your eyelid.

009 apply nails. Start: The moment when nail polish tool touches your fingernails. End: The moment when the nail polish tool leaves from your fingernails.

010 bungee jumping. Start: The moment before you jump off from high platform. End: The moment when your body stops swinging.

011 long jump. Start: The moment when you start approach running. End: The moment when your feet fall in the bunker.

012 throw discus. Start: The moment when you hold the discus with your hand and start pre-tilting. End: The moment when the discus leaves from your hand.

013 javelin throw. Start: The moment when you start running with a javelin. End: The moment when the javelin leaves from your hand.

014 squat snatch. Start: The moment when you grab the barbell. End: The moment when the barbell is lifted over the head, the arms are straight, and the legs are straight.

015 arm wrestling. Start: The moment when the two elbows are on the table and in a straight line. End: The moment when one elbow totally touches the table.

016 archery. Start: The moment when you just complete the bow posture. End: The moment when arrows shot out.

017 hit baseball. Start: The moment when you bend your knees, hold your baseball bat with both hands, and prepare to swing the bat. End: The moment when you hit the baseball or fail to hit the baseball.

018 keep goal. Start: The moment when someone try to shoots the soccer. End: The moment when you catch soccer or fail.

019 pass soccer. Start: The moment when you kick soccer to your teammate. End: The moment when your teammate receives the soccer.

020 shot soccer. Start: The moment when you try to shoots the soccer and kick the ball. End: The moment when the soccer enters the goal or fails.

021 play ice hockey. Start: The moment when you hit ice hockey with a ice hockey stick. End: The moment when the ice hockey leaves away.

022 play bowling. Start: The moment when you insert your finger into the bowling hole. End: The moment when the bowling ball leaves away.

023 drop golf. Start: The moment when you straighten your knees, hold the golf club with both hands, aim the ball. End: The moment when you hit the golf ball.

024 play the violin. Start: The moment when you lighter ignites a cigarette. End: The moment when you throw out the cigarette.

026 ride a bike. Start: The moment when you get on the bike and leave your feet from the ground. End: The moment when your feet touch the ground and stop riding.

027 make a telephone call. Start: The moment when you put the phone near your ear. End: The moment when you hang up the phone.

028 chop wood. Start: The moment when you pick up the axe, aim at the wood. End: The moment when you take out the axe from the chopped wood.

REFERENCES

[1] H. Kuehne, H. Juang, E. Garrote, T. Poggio, and T. Serre, “HMDB: A large video database for human motion recognition,” in Proc. Int. Conf. Comput. Vis., Nov. 2011, pp. 2556–2563.

[2] Y. Shao, Y. Zhao, B. Dai, and D. Lin, “Intra and inter-action understanding via temporal action parsing,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 730–739.

[3] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in Proc. NIPS, 2014, pp. 1–9.

[4] K. Soomro, A. R. Zamir, and M. Shah, “UCF101: A dataset of 101 human actions classes from videos in the wild,” 2012, arXiv:1212.0402.

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 4489–4497.

[6] L. Wang et al., “Temporal segment networks for action recognition in videos,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 11, pp. 2740–2755, Nov. 2019.

[7] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, “Temporal pyramid network for action recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 591–600.

[8] L. Wang, Z. Tong, B. Ji, and G. Wu, “TDN: Temporal difference networks for efficient action recognition,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 1895–1904.

[9] Z. Shou, D. Wang, and S.-F. Chang, “Temporal action localization in untrimmed videos via multi-stage CNNs,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 1049–1058.

[10] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, “Temporal action detection with structured segment networks,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 2914–2923.

[11] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, “BMN: Boundary-matching network for temporal action proposal generation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 3889–3898.

[12] M. Xu, C. Zhao, D. S. Rojas, A. Thabet, and B. Ghanem, “G-TAD: Sub-graph localization for temporal action detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 10156–10165.

[13] C. Lin et al., “Fast learning of temporal action proposal via dense boundary generator,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 11499–11506.

[14] H. Idrees et al., “The Thumos challenge on action recognition for videos ‘in the wild,’” Comput. Vis. Image Understand., vol. 155, pp. 1–23, Feb. 2017.
[15] F. C. Heilbron, V.ESC. B. Ghanem, and J. C. Niebles, “ActivityNet: A large-scale video benchmark for human activity understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 961–970.

[16] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, “HACS: Human action clips and segments dataset for recognition and temporal localization,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 8668–8678.

[17] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, “Learning realistic human actions from movies,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2008, pp. 1–8.

[18] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, “Dense trajectories and motion boundary descriptors for action recognition,” Int. J. Comput. Vis., vol. 103, no. 1, pp. 60–79, May 2013.

[19] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new model and the kinetics dataset,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6299–6308.

[20] M. Monfort et al., “Moments in time dataset: One million videos for fine-grained action understanding,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2016, pp. 510–526.

[21] Y.-W. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and J. Deng, “Hollywood in homes: Crowdsourcing data collection for activity understanding,” in Proc. 25th ACM Int. Conf. Multimedia, Oct. 2017, pp. 988–996.

[22] Y. Liu, L. Ma, Y. Zhang, D. Deng, and R. Cao, “Rethinking the faster R-CNN architecture for temporal action localization,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1130–1139.

[23] H. Xu, A. Das, and K. Saenko, “R-3D: Region convolutional 3D network for temporal action detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Oct. 2017, pp. 5783–5792.

[24] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards real-time object detection with region proposal networks,” in Proc. Adv. Neural Inf. Process. Syst., 2015, pp. 91–99.

[25] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, “CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5734–5743.

[26] Y. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, “BSN: Boundary sensitive network for temporal action proposal generation,” in Proc. ECCV, 2018, pp. 3–19.

[27] S. Yeung, O. Russakovsky, G. Morl, and L. Fei-Fei, “End-to-end learning of action detection from frame glimpses in videos,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 2678–2687.

[28] S. Buch, V. ESC. C. Shen, B. Ghanem, and J. C. Niebles, “SST: Single-stream temporal action proposals,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 2911–2920.

[29] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, “Uncovering the temporal context for video question answering,” Int. J. Comput. Vis., vol. 124, no. 3, pp. 409–421, Sep. 2017.

[30] R. Zeng et al., “Graph convolutional networks for temporal action localization,” in Proc. ICCV, 2019, pp. 7094–7103.

[31] Y. Bai, Y. Wang, Y. Tong, Y. Yang, Q. Liu, and J. Liu, “Boundary content graph neural network for temporal action proposal generation,” in Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer, 2020, pp. 121–137.

[32] J. Tan, J. Tang, L. Wang, and G. Wu, “Relaxed transformer decoders for direct action proposal generation,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2021, pp. 13526–13535.

[33] S. Abu-El-Haija et al., “YouTube-8M: A large-scale video classification benchmark,” 2016, arXiv:1609.08675.

[34] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang, “Exploiting feature and class relationships in video categorization with regularized deep neural networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 2, pp. 352–364, 2018, doi: 10.1109/TPAMI.2017.2670560.

[35] C. Zhang, J. Wu, and Y. Li, “ActionFormer: Localizing moments of action detection, video understanding, and deep learning.”
**Limin Wang** (Member, IEEE) received the B.Sc. degree from Nanjing University, Nanjing, China, in 2011, and the Ph.D. degree from the Chinese University of Hong Kong, Hong Kong, in 2015. From 2015 to 2018, he was a Post-doctoral Researcher with the Computer Vision Laboratory, ETH Zurich. He is currently a Professor with the Department of Computer Science and Technology, Nanjing University. His research interests include computer vision and deep learning. He was the first runner-up at the ImageNet Large Scale Visual Recognition Challenge 2015 in scene recognition and the winner at the ActivityNet Large Scale Activity Recognition Challenge 2016 in video classification. He has served as a Senior PC or an Area Chair for CVPR, AAAI, IJCAI, and WACV.

**Yali Wang** received the Ph.D. degree in computer science from Laval University, QC, Canada, in 2014. He is currently an Associate Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests are deep learning, computer vision, machine learning, and pattern recognition.

**Xiao Ma** received the B.Eng. degree from the Southeast University of Computer Science and Engineering, Nanjing, China, in 2018, and the master’s degree from the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, in 2021. Her research interests include action classification, temporal action detection, and video understanding.

**Yu Qiao** (Senior Member, IEEE) is currently a Professor with the Shenzhen Institutes of Advanced Technology (SIAT), Chinese Academy of Science, and the Director of the Institute of Advanced Computing and Digital Engineering. He has published more than 180 articles in international journals and conferences, including IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IJCV, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON SIGNAL PROCESSING, CVPR, and ICCV. His research interests include computer vision, deep learning, and bioinformation. He received the First Prize of the Guangdong Technological Invention Award and the Jiaxi Lv Young Researcher Award from the Chinese Academy of Sciences. His group achieved the first runner-up at the ImageNet Large Scale Visual Recognition Challenge 2015 in scene recognition and the Winner at the ActivityNet Large Scale Activity Recognition Challenge 2016 in video classification. He has served as the Program Chair for the IEEE ICIST 2014.