# Summary of Primary Paper: FineSports: A Multi-person Hierarchical Sports Video Dataset for Fine-grained Action Understanding

**https://openaccess.thecvf.com/content/CVPR2024/papers/Xu\_FineSports\_A\_Multi-person\_Hierarchical\_Sports\_Video\_Dataset\_for\_Fine-grained\_Action\_CVPR\_2024\_paper.pdf**

**Key Concepts and Taxonomy**

The research paper "FineSports: A Multi-person Hierarchical Sports Video Dataset for Fine-grained Action Understanding" introduces a new, large-scale video dataset designed to advance research in fine-grained human action analysis, particularly in complex, multi-person sports environments[cite: 5]. [cite\_start]The core of this work is the **FineSports** dataset, which consists of 10,000 video clips from NBA basketball games[cite: 39, 55]. [cite\_start]A key feature is its detailed, hierarchical annotation scheme, which defines 12 coarse-grained action categories (e.g., "Drive", "Shoot", "Pass") and 52 more specific, fine-grained sub-action types (e.g., "Drive Left", "Dribble Jumper", "High P&R")[cite: 57, 110, 238]. [cite\_start]These annotations provide not only semantic labels but also precise spatial-temporal information, including 123,000 bounding boxes for all players and the exact start and end frames for each sub-action performed by the ball handler[cite: 39, 58, 65].

To demonstrate the dataset's utility, the paper also proposes a novel model for **Prompt-driven Spatial-Temporal Action Localization (POSTAL)**[cite: 40]. [cite\_start]This end-to-end framework localizes a target player's actions in a video by using a natural language prompt that describes the player's appearance, such as "a player wearing a [white] jersey number [35]"[cite: 26, 278, 314]. [cite\_start]The **POSTAL** architecture is composed of two main components[cite: 61, 284]:

* **Prompt-driven Target Action Encoder (PTA):** This module uses a text encoder (like BERT) to interpret the descriptive prompt and integrates this information with video features via cross-attention, allowing the model to focus specifically on the target player[cite: 40, 285].
* **Action Tube-specific Detector (ATD):** This module takes the prompt-guided features from the PTA and predicts the "action tube"—a sequence of bounding boxes over time—as well as the corresponding fine-grained action class for the target player[cite: 40, 284].

**Main Contributions and Findings**

The paper's contributions are threefold[cite: 64]:

1. It introduces and makes publicly available the **FineSports dataset**, a significant resource for the computer vision community[cite: 64]. [cite\_start]Compared to existing sports datasets like MultiSports, FineSports offers a much larger scale for basketball (10,000 videos vs. 800) and a more refined hierarchy of action labels (52 sub-actions vs. 18), providing a challenging benchmark for understanding nuanced actions[cite: 263, 264].

2. It proposes **POSTAL**, a new prompt-driven methodology for the spatial-temporal action localization (STAL) task[cite: 66]. [cite\_start]This approach is distinct from prior methods that typically detect all actions and then classify them; instead, POSTAL is explicitly guided by a language description to find and classify a specific target's actions in one unified process[cite: 98].

3. Through extensive experiments, the paper demonstrates the effectiveness of the proposed **POSTAL** model, which achieves state-of-the-art results on the new FineSports benchmark, outperforming existing STAL methods like MOC and TubeR[cite: 41, 377]. [cite\_start]These results validate both the utility and difficulty of the FineSports dataset and the innovative nature of the prompt-driven localization approach[cite: 67].

**Limitations and Future Directions**

The authors identify a primary limitation in their work:

* **Domain Specificity:** The FineSports dataset is currently confined to a single sport—basketball (specifically NBA games)[cite: 418]. The complex dynamics, rules, and player interactions are specific to this context, which may limit the direct generalization of models trained on it to other sports.

To address this, the paper suggests a clear direction for future research:

* **Expansion to Other Sports:** The authors state that the dataset needs to be generalized to more multi-person sports like baseball, football, and volleyball to meet the broader need for fine-grained action understanding across various group activities[cite: 418].

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Key Referenced Papers

## 1. Diving48 (Li et al., ECCV 2018) [1]

**Key Concepts and Taxonomy**

This paper introduces **Diving48**, a fine-grained video dataset for action recognition focused on competitive diving[cite: 70]. [cite\_start]It features 48 fine-grained dive classes, which are defined by a combination of four semantic components: dive group (e.g., forward, back), number of somersaults, number of twists, and body position (e.g., tuck, pike)[cite: 70]. The work highlights the challenge of high intra-class variation and high inter-class similarity in fine-grained action recognition.

**Main Contributions and Findings**

The main contribution is the **Diving48** dataset itself, which provides a benchmark for fine-grained action understanding with a structured, compositional label space[cite: 70]. The authors also propose a model that learns to recognize these composite actions, demonstrating the dataset's utility for evaluating models on subtle visual distinctions.

**Limitations and Future Directions**

The dataset is limited to a single, highly structured activity (diving) with a mostly fixed camera view. Future work could involve expanding this compositional approach to less structured activities or more dynamic scenes.

**Relevance to Primary Paper**

The **FineSports** paper cites Diving48 as a key example of an existing fine-grained sports video dataset[cite: 262]. It is used to position FineSports within the landscape of action recognition and quality assessment datasets, highlighting how FineSports provides not just classification labels but also dense spatial-temporal annotations, which Diving48 lacks, to support the more complex task of localization.

## 2. FP-Basket (Bertasius et al., ICCV 2017) [2]

**Key Concepts and Taxonomy**

This work introduces **FP-Basket**, a dataset for assessing basketball player performance from first-person videos[cite: 76]. The task is to predict the outcome of a basketball shot (make or miss) and to evaluate the quality of a player's form. [cite\_start]The model, "Am I a Baller?", uses visual cues from the first-person perspective to make these assessments[cite: 423].

**Main Contributions and Findings**

The paper's primary contributions are the **FP-Basket** dataset and the novel task of first-person basketball performance assessment[cite: 76]. It demonstrated the feasibility of using computer vision to analyze sports performance from a player's own point of view, focusing on action quality rather than just recognition.

**Limitations and Future Directions**

The dataset is limited to the first-person perspective of a single action (shooting). This viewpoint is very different from the typical broadcast view used for sports analysis and does not capture team interactions or broader game context.

**Relevance to Primary Paper**

[cite\_start]**FineSports** cites FP-Basket to contrast its own contribution with prior work in basketball video analysis[cite: 76]. While FP-Basket focuses on first-person action quality assessment, FineSports uses a third-person, overhead view to tackle multi-person, fine-grained action localization and interaction understanding, representing a different and more complex challenge.

## 3. Kinetics (Carreira and Zisserman, CVPR 2017) [3]

**Key Concepts and Taxonomy**

[cite\_start]This paper introduces **Kinetics**, a large-scale, high-quality dataset of human action videos[cite: 425]. It contains hundreds of thousands of short, trimmed video clips from YouTube, covering 400-700 human action classes. [cite\_start]The paper also proposed the Inflated 3D ConvNet (I3D), a deep learning architecture that "inflates" successful 2D image classification models into 3D for video, achieving state-of-the-art results[cite: 92].

**Main Contributions and Findings**

[cite\_start]The creation of the **Kinetics** dataset was a major contribution that spurred significant progress in video-based action recognition[cite: 48]. It became the standard benchmark for pre-training video understanding models. The **I3D** model demonstrated that knowledge from large image datasets could be effectively transferred to the video domain.

**Limitations and Future Directions**

Kinetics consists of short, pre-trimmed clips focused on a single action, making it unsuitable for tasks requiring temporal localization in long, untrimmed videos. The annotations are also at the video level, lacking spatial information.

**Relevance to Primary Paper**

Kinetics is foundational to the **FineSports** paper. The authors of FineSports use a **CSN-152** network as their video backbone, which was pre-trained on the **Kinetics-400** dataset[cite: 364]. This demonstrates the common practice of leveraging powerful features learned from this large-scale dataset for more specific, downstream tasks.

## 4. SportsMOT (Cui et al., arXiv 2023) [4]

**Key Concepts and Taxonomy**

This paper presents **SportsMOT**, a large-scale, multi-object tracking (MOT) dataset focused on sports scenes, including basketball, volleyball, and football[cite: 428]. It provides dense annotations for tracking numerous players simultaneously in complex, interactive scenarios. The paper also introduces **MixSort-OC**, an effective tracking algorithm designed to handle the challenges present in sports videos, such as severe occlusions and rapid movements[cite: 222].

**Main Contributions and Findings**

The main contribution is the **SportsMOT** dataset, which serves as a challenging benchmark for multi-object tracking in the sports domain[cite: 428]. [cite\_start]The **MixSort-OC** tracker proposed alongside the dataset proved to be a powerful tool for accurately tracking players throughout a video sequence[cite: 222].

**Limitations and Future Directions**

While providing excellent tracking data, SportsMOT's annotations are focused on object trajectories (the "where") and do not include fine-grained action labels (the "what").

**Relevance to Primary Paper**

This work is highly relevant and was directly utilized by the **FineSports** authors. [cite\_start]They explicitly state that they used the **MixSort-OC** tracker from SportsMOT as a key tool in their annotation pipeline[cite: 222, 266]. It helped them automatically generate initial bounding box tracks for all players, which human annotators then refined and labeled with fine-grained actions, significantly speeding up the data creation process.

## 5. BERT (Devlin et al., NAACL 2019) [5]

**Key Concepts and Taxonomy**

This seminal paper introduces **BERT (Bidirectional Encoder Representations from Transformers)**, a language representation model that transformed Natural Language Processing (NLP)[cite: 429]. Unlike previous models that processed text in one direction, BERT uses a Transformer architecture to process the entire sequence of words at once, allowing it to learn deep bidirectional representations conditioned on both left and right context.

**Main Contributions and Findings**

BERT's key contribution was its pre-training approach on a massive text corpus, which allowed it to be fine-tuned with just one additional output layer to achieve state-of-the-art performance on a wide range of NLP tasks[cite: 429]. This established the pre-training/fine-tuning paradigm as dominant in NLP.

**Limitations and Future Directions**

BERT has a high computational cost for both pre-training and inference. Subsequent research has focused on creating more efficient versions (e.g., ALBERT, DistilBERT).

**Relevance to Primary Paper**

BERT is a crucial component of the proposed **POSTAL** model. The **Prompt-driven Target Action Encoder (PTA)** module in POSTAL uses a text encoder to understand the natural language prompt[cite: 285]. [cite\_start]The FineSports paper specifies using a BERT model (specifically, the frozen text encoder from BLIP, which is BERT-based) for this purpose, making it a foundational technology for their method's ability to link language to visual action[cite: 317, 370].

## 6. ActivityNet (Heilbron et al., CVPR 2015) [6]

**Key Concepts and Taxonomy**

This paper introduces **ActivityNet**, a large-scale video benchmark for human activity understanding from untrimmed videos[cite: 432]. It contains nearly 200 different activity categories with thousands of videos per category, all sourced from YouTube.

**Main Contributions and Findings**

ActivityNet's primary contribution was providing a large-scale, diverse benchmark of untrimmed videos, which pushed the research community to move beyond classifying short, pre-segmented clips and toward localizing events in long, realistic videos. It became a standard dataset for tasks like temporal action localization.

**Limitations and Future Directions**

The dataset's annotations are at the activity level and do not provide the fine-grained, procedural sub-action details or multi-person spatial bounding boxes that are the focus of FineSports.

**Relevance to Primary Paper**

**FineSports** cites ActivityNet in its introduction as a primary example of a publicly accessible action video dataset[cite: 48]. The authors argue that such datasets, while valuable, generally lack the high-quality, fine-grained annotations necessary to tackle the challenges of fine-grained action analysis, thereby motivating the creation of FineSports.

## 7. Abductive Ego-view Accident Video Understanding (Fang et al., arXiv 2024) [7]

**Key Concepts and Taxonomy**

This paper focuses on understanding traffic accidents from an ego-centric (first-person) perspective for autonomous driving applications. It proposes a framework for "abductive reasoning," where the goal is to infer the most likely explanation for an observed accident by analyzing the events leading up to it.

**Main Contributions and Findings**

The main contribution is a new framework for accident video understanding that moves beyond simple event recognition to causal reasoning. This is crucial for developing safer and more perceptive autonomous driving systems.

**Limitations and Future Directions**

The approach is highly specialized for the autonomous driving domain and ego-centric video, which has different characteristics from third-person sports videos.

**Relevance to Primary Paper**

This paper is cited in the introduction of **FineSports** as an example of a practical application for human action understanding, specifically in the domain of autonomous driving[cite: 46]. This helps to establish the broader importance and relevance of the research field that FineSports contributes to.

## 8. SlowFast Networks (Feichtenhofer et al., ICCV 2019) [8]

**Key Concepts and Taxonomy**

This work introduces **SlowFast Networks**, a dual-pathway architecture for video recognition. It consists of a "Slow" pathway, which operates at a low frame rate to capture spatial semantics, and a "Fast" pathway, which operates at a high frame rate to capture fine-grained temporal motion. The two pathways are fused to create a joint representation.

**Main Contributions and Findings**

The key contribution is the biologically-inspired SlowFast architecture, which explicitly models different temporal speeds within a single network. This design proved to be highly effective and efficient, achieving state-of-the-art results on major action recognition benchmarks like Kinetics.

**Limitations and Future Directions**

While a powerful backbone for recognition, it does not inherently perform localization. It would need to be integrated into a larger detection or localization framework.

**Relevance to Primary Paper**

**FineSports** cites SlowFast Networks as an example of a deep learning-based video understanding approach that has achieved remarkable performance on human action tasks[cite: 47]. This citation helps to set the stage and provide context for the current state of video understanding research, into which the FineSports paper introduces its own contributions.

## 9. Neural Bipartite Matching (Georgiev and Lió, arXiv 2020) [9]

**Key Concepts and Taxonomy**

This paper explores the use of neural networks to solve the bipartite matching problem, a classic task in graph theory. It proposes learning-based approaches to find optimal or near-optimal matchings in bipartite graphs, which can be more flexible than traditional combinatorial algorithms.

**Main Contributions and Findings**

The contribution is the exploration of neural methods for a classical optimization problem, suggesting that deep learning can be applied to solve combinatorial tasks where data-driven patterns exist.

**Limitations and Future Directions**

The effectiveness of neural approaches is often dependent on the distribution of the training data and may not provide the same formal guarantees as classical algorithms.

**Relevance to Primary Paper**

The **FineSports** paper cites this work in the implementation details for its proposed **POSTAL** model[cite: 372]. The POSTAL model uses bipartite matching to associate its predicted action tubes with the ground truth tubes during training in order to compute the matching loss. This is a common strategy in modern detector architectures inspired by DETR.

## 10. Finding Action Tubes (Gkioxari and Malik, CVPR 2015) [10]

**Key Concepts and Taxonomy**

This paper is an early and influential work in spatial-temporal action localization. It proposes a method to "find action tubes" by starting with static frame-level person detections and then linking them across time using motion information (optical flow) and appearance similarity.

**Main Contributions and Findings**

The paper's main contribution was a practical and effective framework for generating action tubes by connecting per-frame detections. It demonstrated a way to extend 2D object detection paradigms into the temporal domain for action localization.

**Limitations and Future Directions**

The method relies heavily on the quality of the initial per-frame person detector and can be brittle, as errors in one frame can propagate through the linking process. It primarily focuses on spatial information within frames rather than deep temporal modeling.

**Relevance to Primary Paper**

**FineSports** cites this paper in its related work section as an example of an early, frame-level localization paradigm[cite: 90]. This is used to contrast with more recent video-level paradigms (which POSTAL follows) that use 3D CNNs to jointly model space and time, showcasing the evolution of the field.

## 11. End-to-end Spatio-temporal Action Localisation with Video Transformers (Gritsenko et al., arXiv 2023) [11]

**Key Concepts and Taxonomy**

This paper proposes an end-to-end spatial-temporal action localization model based entirely on a Video Transformer architecture. It frames the task as a direct set prediction problem, where the model outputs a set of action tubes without needing region proposals or post-processing steps like non-maximum suppression.

**Main Contributions and Findings**

The main contribution is a fully Transformer-based, end-to-end framework for STAL. This demonstrates that the success of DETR-like models in image object detection can be effectively translated to the more complex video domain for action localization.

**Limitations and Future Directions**

Transformer-based models are known to be data-hungry and computationally expensive, often requiring long training schedules and large datasets.

**Relevance to Primary Paper**

This work is cited in the **FineSports** paper as an example of the current mainstream paradigm in STAL: video-level localization using powerful backbones like Video Transformers[cite: 92]. This provides context for the authors' own **POSTAL** model, which also follows a video-level paradigm and incorporates a Transformer-based detector (the ATD module).

## 12. AVA Dataset (Gu et al., CVPR 2018) [12]

**Key Concepts and Taxonomy**

This paper introduces the **AVA (Atomic Visual Actions)** dataset, a large-scale video dataset for spatio-temporally localized atomic visual actions[cite: 441]. It provides dense annotations on 15-minute movie clips, with bounding boxes for people at 1 frame-per-second, and labels each person with potentially multiple, concurrent actions (e.g., "talking" while "shaking hands").

**Main Contributions and Findings**

The key contributions of **AVA** are its scale, the density of its annotations, and its focus on atomic (simple) actions in realistic, multi-person scenarios. It enabled research into complex activity understanding where multiple actions occur simultaneously and interact.

**Limitations and Future Directions**

While densely annotated, the actions are "atomic" and may not capture the full procedural or hierarchical nature of more complex activities, which is a key focus of FineSports.

**Relevance to Primary Paper**

**FineSports** cites AVA in its introduction as a major publicly available action video dataset[cite: 48]. Like with ActivityNet, it is used as an example of existing datasets that, while valuable, lack the specific type of fine-grained, procedural annotations that FineSports was created to provide, thus motivating its development.

## 13. 3D ResNeXt (Hara et al., CVPR 2018) [13]

**Key Concepts and Taxonomy**

This paper investigates the performance of deep 3D Convolutional Neural Networks (3D CNNs) for video action recognition. The authors systematically explore architectures like 3D ResNet and introduce a 3D version of the successful ResNeXt model. They demonstrate that pre-training on large video datasets (like Kinetics) is crucial for these models to be effective.

**Main Contributions and Findings**

The main contribution is a thorough empirical study showing that well-designed, deep 3D CNNs can achieve state-of-the-art results in action recognition, retracing the success of 2D CNNs on ImageNet[cite: 443]. They also provided strong, pre-trained 3D models to the community.

**Limitations and Future Directions**

The primary limitation of these models is their very high computational cost in terms of memory and processing power, making them challenging to train and deploy.

**Relevance to Primary Paper**

This work is cited in the **FineSports** paper when discussing video-level localization backbones[cite: 94]. Specifically, the paper mentions that the YOWO method uses **3D-ResNext** as its backbone, providing context for the architectural choices made by previous state-of-the-art methods in the field of spatial-temporal action localization.

## 14. Human Action Recognition in Sports (Host and Ivašić-Kos, Heliyon 2022) [14]

**Key Concepts and Taxonomy**

This paper is a survey that provides an overview of human action recognition in sports using computer vision techniques. It reviews different methodologies, datasets, and challenges specific to the sports domain, such as fast motion, occlusions, and complex team interactions.

**Main Contributions and Findings**

The main contribution is a comprehensive review of the state of the art in sports-based action recognition. It categorizes existing work and identifies open research problems and future directions for the field.

**Limitations and Future directions**

As a survey paper, it describes the state of the field rather than introducing a new method.

**Relevance to Primary Paper**

**FineSports** cites this survey in its introduction to highlight that sports analysis is a significant and active application area for human action understanding[cite: 46]. This helps to frame the problem that FineSports is addressing as one with practical importance and significant research interest.

## 15. T-CNN (Hou et al., ICCV 2017) [15]

**Key Concepts and Taxonomy**

This paper proposes **T-CNN (Tube Convolutional Neural Network)**, a method for action detection in videos. The approach consists of two stages: first, generating "action tubelet" proposals (short sequences of linked bounding boxes), and second, classifying these tubelets and refining their boundaries. The core idea is to learn to link frame-level detections into tubelets.

**Main Contributions and Findings**

The main contribution is the concept of tubelet proposals and the T-CNN framework for action detection. This work provided a structured way to handle the temporal dimension of action localization by explicitly creating and classifying spatio-temporal proposals.

**Limitations and Future Directions**

The performance is highly dependent on the quality of the initial tubelet proposals. The two-stage design can be complex and less efficient than end-to-end models.

**Relevance to Primary Paper**

[cite\_start]This paper is cited in the **FineSports** paper in relation to the design of its learnable action tube queries[cite: 333]. The concept of using 3D cuboids or tubelets to represent actions, as popularized by T-CNN and other similar works, is a foundational idea that informs the design of the queries used in the **POSTAL** model's detector.

## 16. Hierarchical Deep Temporal Model for Group Activity Recognition (Ibrahim et al., CVPR'16) [16]

**Key Concepts and Taxonomy**

This paper proposes a hierarchical deep learning model for recognizing group activities in videos, specifically focusing on volleyball. The model uses LSTMs to capture the temporal dynamics of individual player actions and then aggregates these individual representations to predict the overall group activity (e.g., "right spike," "left set").

**Main Contributions and Findings**

The key contribution is a hierarchical model that explicitly reasons about individuals to understand the group's collective action. The paper also introduced the **Volleyball** dataset, with annotated player bounding boxes and group activity labels.

**Limitations and Future Directions**

The model requires bounding box annotations for all individuals, and its focus is on group-level classification rather than fine-grained individual action localization.

**Relevance to Primary Paper**

The **FineSports** paper includes this work in its comparison table of sports video datasets (Table 1)[cite: 79]. It is presented as an early example of a dataset with annotations for a team sport (volleyball), which helps to position FineSports and highlight its own advancements in terms of scale, annotation detail, and task focus (fine-grained individual STAL vs. group recognition).

## 17. FineFS / LUSD-NET (Ji et al., ACM MM 2023) [17]

**Key Concepts and Taxonomy**

This paper introduces **FineFS**, a large-scale fine-grained figure skating dataset containing both RGB videos and estimated skeleton sequences. To leverage this data, the paper proposes **LUSD-NET (Localization-assisted Uncertainty Score Disentanglement Network)** for action quality assessment. LUSD-NET first localizes the action in time and then assesses its quality, using an uncertainty score to improve robustness.

**Main Contributions and Findings**

The primary contributions are the **FineFS** dataset and the **LUSD-NET** model. The model's novelty lies in its fusion of spatial-temporal features with the original sequence and its application to multiple fine-grained tasks, including quality assessment.

**Limitations and Future Directions**

The work is primarily focused on action quality assessment in a single-person sport (figure skating), which has different dynamics than multi-person team sports.

**Relevance to Primary Paper**

[cite\_start]**FineSports** cites this work multiple times in its related work section[cite: 85, 88, 92, 95]. It's used as an example of recent work in spatial-temporal localization, a video-level localization paradigm, and a method that enhances feature perception for fine-grained understanding. This contextualizes **POSTAL** among recent, sophisticated approaches. [cite\_start]The **FineFS** dataset is also included in the comparison table[cite: 79].

## 18. Action Tubelet Detector (Kalogeiton et al., ICCV 2017) [18]

**Key Concepts and Taxonomy**

This paper proposes an **Action Tubelet Detector** that localizes actions in space and time by detecting "tubelets" (short sequences of bounding boxes). The method extends the Faster R-CNN object detector by adding a temporal dimension, allowing it to propose and classify spatio-temporal regions of interest directly from video frames.

**Main Contributions and Findings**

The main contribution is a framework that effectively integrates temporal information into a successful 2D detection pipeline. The paper also helped popularize the **frame-mAP** (mean Average Precision) metric for evaluating spatial-temporal action localization performance.

**Limitations and Future Directions**

The method relies on linking short tubelets together in a post-processing step, which can be complex and suboptimal compared to end-to-end approaches.

**Relevance to Primary Paper**

[cite\_start]This paper is cited in **FineSports** as the source for the **frame-mAP** evaluation metric[cite: 360]. This metric is one of the key benchmarks used to compare the performance of the proposed **POSTAL** model against other state-of-the-art methods, making this reference crucial for the experimental validation section.

## 19. The Kinetics Human Action Video Dataset (Kay et al., arXiv 2017) [19]

**Key Concepts and Taxonomy**

This paper provides a detailed description of the **Kinetics** dataset, a large-scale collection of human action videos designed for training and evaluating action recognition models. It contains hundreds of thousands of URL-linked, 10-second clips from YouTube across 400 (and later 600 and 700) action categories.

**Main Contributions and Findings**

The paper's contribution is the dataset itself, which became the de facto standard for pre-training deep learning models for video understanding. It enabled the development of powerful 3D CNNs by providing sufficient data to train them effectively.

**Limitations and Future Directions**

As with reference [3], the dataset consists of trimmed, single-action clips and lacks spatial or dense temporal annotations, making it unsuitable for localization tasks on its own.

**Relevance to Primary Paper**

[cite\_start]This paper is cited as the source of the **Kinetics-400** dataset, on which the **CSN-152** backbone used in the **POSTAL** model was pre-trained[cite: 364]. This citation is essential for reproducibility and for acknowledging the source of the foundational pre-trained model that their work builds upon.

## 20. YOWO (Köpüklü et al., arXiv 2019) [20]

**Key Concepts and Taxonomy**

This paper proposes **YOWO (You Only Watch Once)**, a single-stage, unified CNN architecture for real-time spatial-temporal action localization. It combines a 2D CNN branch for extracting spatial features from a keyframe with a 3D CNN branch for extracting temporal features from a clip, fusing them to predict bounding boxes and action classes in a single forward pass.

**Main Contributions and Findings**

The main contribution of YOWO is its efficiency and real-time performance. By adopting a single-stage approach inspired by YOLO (for images), it avoids the complex two-stage pipeline of proposal-then-classification, making it much faster.

**Limitations and Future Directions**

The single-stage approach can sometimes be less accurate than more complex two-stage models, especially for small or difficult-to-localize actions.

**Relevance to Primary Paper**

**FineSports** cites YOWO in its related work section as a prominent example of a video-level localization method[cite: 92, 94]. It is described as a two-branch framework that enhances feature aggregation, providing context for the architectural evolution in the STAL field and setting the stage for the introduction of the **POSTAL** model.

## 21. UniformerV2 (Li et al., ICCV 2023) [21]

**Key Concepts and Taxonomy**

This paper introduces **UniformerV2**, a video understanding model that aims to unlock the potential of Image Vision Transformers (ViTs) for video tasks. It proposes a novel approach that unifies convolution and self-attention in a dynamic, sparse manner, allowing the model to efficiently process both spatial and temporal information.

**Main Contributions and Findings**

The main contribution is a powerful and efficient video backbone that effectively adapts image-based ViT architectures for video. UniformerV2 achieved state-of-the-art performance on major action recognition benchmarks while being more computationally efficient than many competing models.

**Limitations and Future Directions**

The model is primarily a backbone for video classification/recognition and would need to be integrated into a larger framework for localization tasks.

**Relevance to Primary Paper**

**FineSports** cites UniformerV2 as a state-of-the-art, deep learning-based video understanding approach.This citation is used in the introduction to illustrate the significant recent advances in the field of human action understanding, providing a backdrop for the challenges and contributions presented by the FineSports dataset and the POSTAL model.

## 22. MOC (Li et al., ECCV 2020) [22]

**Key Concepts and Taxonomy**

This paper proposes **MOC (Actions as Moving Points)**, a framework for spatial-temporal action localization.It reframes the task by treating an action instance as a trajectory of moving points. The model simultaneously detects the action's center point in each frame, predicts the movement between frames, and regresses the bounding box size at each point.

**Main Contributions and Findings**

The key contribution is the novel formulation of detecting actions as tracking moving points, which simplifies the process compared to linking frame-level detections. This approach, implemented in a tri-branch structure, proved to be an effective and efficient method for generating action tubes.

**Limitations and Future Directions**

The model's performance can be sensitive to the accuracy of the initial center point detection. It does not incorporate external guidance, such as text prompts, to focus on specific actors.

**Relevance to Primary Paper**

**MOC** is cited as a state-of-the-art (SOTA) method for spatial-temporal action localization that the proposed **POSTAL** model is benchmarked against. The authors use the comparison to demonstrate that POSTAL's prompt-driven, query-based approach yields a 2.33% improvement on frame-mAP, highlighting its superior accuracy.

## 23. MultiSports (Li et al., ICCV 2021) [23]

**Key Concepts and Taxonomy**

This paper introduces **MultiSports**, a large-scale, multi-person video dataset for fine-grained spatial-temporally localized sports actions[cite: 463]. The dataset covers four different sports: aerobic gymnastics, volleyball, football, and basketball, and provides dense annotations of bounding boxes and action classes for multiple individuals.

**Main Contributions and Findings**

The main contribution is the creation of a diverse, multi-sport benchmark for fine-grained STAL, enabling the development and evaluation of models in complex, interactive scenarios.

**Limitations and Future Directions**

While comprehensive, the annotations for each individual sport may not be as extensive or as hierarchically detailed as a dataset focused on a single sport.

**Relevance to Primary Paper**

**MultiSports** is a key point of comparison for **FineSports**. [cite\_start]The authors cite it extensively in their dataset comparison sections (Table 1, Section 3.3)[cite: 79, 263]. They argue that while MultiSports is a valuable resource, the basketball portion of their **FineSports** dataset is significantly larger (10,000 vs. 800 videos) and more granular (52 vs. 18 sub-action types), thus providing a more challenging and comprehensive benchmark specifically for basketball analysis[cite: 264].

## 24. FSD-10 (Liu et al., arXiv 2020) [24]

**Key Concepts and Taxonomy**

This paper presents **FSD-10**, a figure skating dataset designed for fine-grained content analysis in competitive sports[cite: 465]. The dataset includes annotations for fine-grained actions (10 types) such as different kinds of jumps and spins, and is intended to support tasks like action recognition and technical evaluation.

**Main Contributions and Findings**

The primary contribution is the **FSD-10** dataset, which provides a resource for analyzing highly technical and visually similar actions in the domain of figure skating[cite: 72].

**Limitations and Future Directions**

The dataset is focused on a single-person sport and the primary task is recognition/analysis rather than multi-person localization.

**Relevance to Primary Paper**

**FineSports** includes FSD-10 in its comparative analysis of existing fine-grained sports datasets (Table 1)[cite: 79]. It is used as an example of a dataset for fine-grained action recognition, helping to contextualize FineSports' own contribution, which extends beyond recognition to the more complex task of spatial-temporal localization in a multi-person setting.

## 25. MCFS (Liu et al., AAAI 2021) [25]

**Key Concepts and Taxonomy**

This paper introduces **MCFS (Motion-Centered Figure Skating)**, a fine-grained dataset for temporal action segmentation. [cite\_start]The dataset features figure skating videos annotated with a three-level semantic hierarchy, allowing for detailed temporal parsing of performances[cite: 74]. The focus is on segmenting a long performance into its constituent fine-grained actions.

**Main Contributions and Findings**

The main contributions are the **MCFS** dataset and its focus on temporal action segmentation with hierarchical labels[cite: 74]. This enables research into understanding the procedural structure of complex activities.

**Limitations and Future Directions**

The dataset is for a single-person sport and focuses only on temporal segmentation, lacking the spatial bounding box annotations needed for STAL.

**Relevance to Primary Paper**

**MCFS** is included in the comparison table of sports video datasets (Table 1)[cite: 79]. It is categorized as a dataset for temporal action localization/segmentation, which helps differentiate the contribution of **FineSports**, which provides both spatial and temporal annotations for the STAL task.

## 26. RFSJ (Liu et al., MM 2023) [26]

**Key Concepts and Taxonomy**

This paper presents **RFSJ**, a figure skating jumping dataset that includes replay information and fine-grained annotations for action quality assessment. It leverages the slow-motion replays often shown in broadcasts to help analyze the fine-grained details of jump execution.

**Main Contributions and Findings**

The key contribution is a dataset that uniquely pairs main-view footage with replay footage for fine-grained analysis[cite: 86]. This supports tasks like action quality assessment and fine-grained recognition by providing multiple views of the same action.

**Limitations and Future Directions**

The dataset is specific to figure skating jumps and the task of quality assessment.

**Relevance to Primary Paper**

**RFSJ** is cited in the related work section and included in the dataset comparison table (Table 1)[cite: 79, 88]. It serves as an example of recent, highly specialized, fine-grained sports datasets. This helps to demonstrate the trend toward more detailed sports video analysis, a trend that **FineSports** contributes to significantly.

## 27. Video Swin Transformer (Liu et al., CVPR 2022) [27]

**Key Concepts and Taxonomy**

This paper introduces the **Video Swin Transformer**, an adaptation of the highly successful Swin Transformer from images to the video domain. The proposed architecture is a pure-transformer model that effectively captures spatio-temporal information by using 3D shifted windows for attention, making it both effective and computationally efficient compared to other video transformers.

**Main Contributions and Findings**

The main contribution is a powerful and efficient video backbone that achieves state-of-the-art results on major action recognition benchmarks[cite: 472]. The Video Swin Transformer became a popular choice as a backbone for various video understanding tasks.

**Limitations and Future Directions**

While a strong backbone, it is designed for recognition and requires integration into a larger framework for detection or localization tasks.

**Relevance to Primary Paper**

The **Video Swin Transformer** is cited in **FineSports** as a state-of-the-art, 3D CNN-based (in a broad sense, as it processes 3D data) backbone that represents the mainstream paradigm for video-level localization[cite: 47, 92]. This places the work of FineSports within the context of modern, powerful video architectures.

## 28. GolfDB (McNally et al., CVPRW 2019) [28]

**Key Concepts and Taxonomy**

This paper introduces **GolfDB**, a video database for the task of golf swing sequencing[cite: 475]. The dataset contains thousands of golf swing videos, each annotated with the temporal locations of eight distinct phases of the swing (e.g., "Address", "Top of Backswing", "Impact").

**Main Contributions and Findings**

The main contribution is the **GolfDB** dataset, which provides a benchmark for fine-grained temporal action localization and event sequencing in a specific, structured domain[cite: 79].

**Limitations and Future Directions**

The dataset focuses on a single-person activity with a relatively fixed structure and camera view. The task is temporal localization only, without spatial annotations.

**Relevance to Primary Paper**

**GolfDB** is included in the **FineSports** comparison table of fine-grained sports datasets (Table 1)[cite: 79]. It is presented as an example of a dataset for the temporal action localization task, which helps to differentiate the contribution of FineSports, which supports the more complex spatial-temporal action localization task.

## 29. MTL-AQA (Parmar and Morris, CVPR 2019) [29]

**Key Concepts and Taxonomy**

This paper addresses the task of **Action Quality Assessment (AQA)** by proposing a multi-task learning approach[cite: 477]. The model learns to predict an action's quality score by jointly training on the main AQA task and a secondary action recognition task. The idea is that learning to recognize the action helps the model learn better representations for assessing its quality.

**Main Contributions and Findings**

The key contribution is the demonstration that a multi-task learning framework can improve performance on AQA[cite: 77]. The paper also introduces a new diving dataset for this purpose.

**Limitations and Future Directions**

The approach is tailored for AQA and recognition in single-person sports like diving and gymnastics.

**Relevance to Primary Paper**

This work is included in the dataset comparison table (Table 1) in **FineSports**[cite: 79]. It serves as an example of a fine-grained sports dataset focused on the tasks of action quality assessment and recognition, thereby helping to position FineSports and its focus on multi-person spatial-temporal localization.

## 30. Multi-region two-stream R-CNN (Peng and Schmid, ECCV 2016) [30]

**Key Concepts and Taxonomy**

This paper proposes a method for action detection that extends the R-CNN framework. It uses a two-stream architecture (one for spatial appearance, one for motion) and introduces a multi-region approach that analyzes not just the person but also different parts of their body and the surrounding context to make a more informed detection.

**Main Contributions and Findings**

The main contribution is the multi-region design, which allows the model to capture more contextual information for action detection, leading to improved performance.

**Limitations and Future Directions**

This is a frame-level localization approach that requires a separate step to link detections across time to form action tubes. It is less efficient than modern end-to-end models.

**Relevance to Primary Paper**

**FineSports** cites this work in its related work section as an example of a frame-level localization method that utilizes 2D CNN-based detection networks[cite: 90]. This helps to trace the history of the field and contrast older frame-level paradigms with the video-level paradigm that the **POSTAL** model follows.

## 31. Fine-grained Activity Recognition in Baseball Videos (Piergiovanni and Ryoo, CVPRW 2018) [31]

**Key Concepts and Taxonomy**

This paper focuses on fine-grained activity recognition in baseball videos. It introduces a dataset and a model capable of distinguishing between very similar actions, such as different types of pitches (e.g., "fastball," "curveball") or swings.

**Main Contributions and Findings**

The primary contribution is a fine-grained dataset for baseball and a model that demonstrates the ability to perform fine-grained classification in this domain[cite: 79].

**Limitations and Future Directions**

The focus is on classification of trimmed clips, not spatial-temporal localization in untrimmed videos.

**Relevance to Primary Paper**

This paper is included in the dataset comparison table (Table 1) in **FineSports**[cite: 79]. It is listed as a fine-grained dataset for action recognition in a team sport (baseball), providing context for other datasets in the field and highlighting the research community's growing interest in fine-grained sports analysis.

## 32. FineGym (Shao et al., CVPR 2020) [32]

**Key Concepts and Taxonomy**

This paper introduces **FineGym**, a hierarchical video dataset for fine-grained action understanding in gymnastics[cite: 481]. It provides a three-level semantic hierarchy, with annotations ranging from coarse-grained gym elements to fine-grained sub-actions. The annotations are temporal, designed for recognition and temporal localization.

**Main Contributions and Findings**

The main contribution is the **FineGym** dataset, which provides a rich, hierarchically structured benchmark for fine-grained action analysis[cite: 71]. It uniquely provides temporal annotations at multiple levels of semantic granularity.

**Limitations and Future Directions**

The dataset focuses on a single-person sport and lacks spatial (bounding box) annotations, making it unsuitable for STAL.

**Relevance to Primary Paper**

**FineGym** is cited multiple times in **FineSports** as a key example of a fine-grained, hierarchical dataset for action recognition[cite: 71, 262]. The authors compare their own hierarchical structure to that of FineGym and use it to position **FineSports** as a dataset that extends the concept of fine-grained analysis to the more complex domain of multi-person spatial-temporal localization.

## 33. TAPOS (Shao et al., CVPR 2020) [33]

**Key Concepts and Taxonomy**

This paper introduces **TAPOS (Temporal Action Parsing of Olympic Sports)**, a dataset designed for studying the internal structure of actions through temporal action parsing[cite: 482]. It provides dense, fine-grained sub-action annotations for various Olympic sports, with a focus on understanding the sequence of procedural steps within a larger action. The task is temporal localization of these sub-actions.

**Main Contributions and Findings**

The main contribution is the **TAPOS** dataset, which enables research on "intra- and inter-action understanding" by providing detailed temporal annotations of action procedures[cite: 73].

**Limitations and Future Directions**

The dataset's focus is purely on temporal parsing and it lacks spatial bounding box annotations.

**Relevance to Primary Paper**

**FineSports** cites **TAPOS** in its related work section and dataset comparison table[cite: 73, 262]. It is highlighted as a dataset for temporal action localization with sub-action annotations. This helps differentiate **FineSports**, which adds the crucial spatial dimension to its fine-grained procedural annotations, supporting the full STAL task.

## 34. Real-time Elderly Monitoring (Sun and Chen, ISMICT 2022) [34]

**Key Concepts and Taxonomy**

This paper focuses on an application of human action recognition: real-time monitoring of elderly people for safety purposes. It proposes a lightweight model designed to recognize actions like falling, sitting, and standing, which can run efficiently on resource-constrained devices.

**Main Contributions and Findings**

The contribution is a practical, lightweight action recognition system tailored for a specific real-world application (elderly care), emphasizing efficiency and real-time performance.

**Limitations and Future Directions**

The model is designed for a limited set of simple actions in a controlled environment, not for complex, fine-grained actions in chaotic scenes.

**Relevance to Primary Paper**

This paper is cited in the introduction of **FineSports** as an example of a practical application of human action understanding, specifically in "abnormal monitoring"[cite: 46]. This citation helps to establish the broad societal importance of the research area to which FineSports contributes.

## 35. VideoMAE (Tong et al., NeurIPS 2022) [35]

**Key Concepts and Taxonomy**

This paper introduces **VideoMAE (Masked Autoencoders for Video)**, a self-supervised pre-training method for video understanding. Inspired by the success of MAE in images, VideoMAE applies a very high masking ratio to video tubelets (spatio-temporal cubes) and tasks the model with reconstructing the missing pixels.

**Main Contributions and Findings**

The key contribution is a highly effective and data-efficient self-supervised learning framework for video. **VideoMAE** demonstrated that one can achieve state-of-the-art performance on action recognition tasks by pre-training on unlabeled video data, reducing the reliance on massive, labeled datasets like Kinetics.

**Limitations and Future Directions**

The pre-training is focused on learning good representations for recognition; adapting these representations for localization tasks requires a dedicated downstream head and fine-tuning.

**Relevance to Primary Paper**

**FineSports** cites VideoMAE as an example of a state-of-the-art, deep learning-based video understanding approach[cite: 47]. This citation is part of the introduction's overview of recent advances in the field, setting the context for the authors' own work.

## 36. CSN (Tran et al., ICCV 2019) [36]

**Key Concepts and Taxonomy**

This paper introduces **CSN (Channel-Separated Convolutional Networks)**, an efficient architecture for video classification. CSNs are a type of 3D CNN that factorizes a standard 3D convolution into two separate and more efficient steps: a depthwise convolution (to process spatio-temporal information) and a pointwise convolution (to mix channel information).

**Main Contributions and Findings**

The main contribution is the **CSN** architecture, which significantly reduces the number of parameters and computational cost of 3D CNNs while maintaining high accuracy. This made it practical to build much deeper 3D networks.

**Limitations and Future Directions**

While efficient, the representational power might be slightly less than a full 3D convolution in some cases, though the paper shows strong empirical results.

**Relevance to Primary Paper**

**CSN** is a critical component of the methodology in the **FineSports** paper. [cite\_start]The authors explicitly state that they use a **CSN-152** network as the backbone for their video feature extractor in the **POSTAL** model[cite: 92, 97, 364]. This makes the CSN paper a foundational reference for their proposed architecture.

## 37. Learning to Track for Spatio-temporal Action Localization (Weinzaepfel et al., ICCV 2015) [37]

**Key Concepts and Taxonomy**

This paper tackles spatio-temporal action localization by framing it as a tracking problem. The method learns a tracker to follow actors in a video and then classifies the resulting tracks. It also introduces the concept of **video-mAP** as an evaluation metric.

**Main Contributions and Findings**

The key contribution is a "tracking-by-detection" approach tailored for actions. More importantly for future work, it introduced the **video-mAP** metric, which computes an IoU over the entire action tube (averaging per-frame spatial IoUs over the temporal intersection) to evaluate performance.

**Limitations and Future Directions**

The method relies on an external person detector and can be less integrated than modern end-to-end approaches.

**Relevance to Primary Paper**

This paper is cited in **FineSports** as the source of the **video-mAP** evaluation metric[cite: 360]. Along with frame-mAP, video-mAP is the primary metric used in the experiments to evaluate and compare the **POSTAL** model's performance, making this reference essential for understanding the paper's quantitative results.

## 38. Learning to Score Figure Skating (Xu et al., TCSVT 2019) [38]

**Key Concepts and Taxonomy**

This paper addresses the task of automatically scoring figure skating videos. It introduces the **Fis-V** dataset and proposes a model that learns to predict the scores given by professional judges by analyzing the visual content of the performance.

**Main Contributions and Findings**

The main contribution is a framework for automatic action quality assessment in the specific, highly technical domain of figure skating[cite: 79]. It demonstrated that visual features could be mapped to expert-level quality scores.

**Limitations and Future Directions**

The model is highly specialized for figure skating and the task of score regression (AQA).

**Relevance to Primary Paper**

This work and its associated **Fis-V** dataset are included in the **FineSports** comparison table (Table 1)[cite: 79]. It is listed as a dataset for action quality assessment, helping to survey the landscape of sports video analysis and differentiate the contribution of FineSports.

## 39. Action Recognition in Traffic Scenes (Xu et al., TITS 2022) [39]

**Key Concepts and Taxonomy**

This paper presents an action recognition framework specifically designed for traffic scenes to be used in autonomous driving systems. The goal is to recognize the actions of pedestrians, cyclists, and other vehicles to improve safety and navigation.

**Main Contributions and Findings**

The contribution is a specialized action recognition system for the traffic domain, demonstrating another practical and important application of computer vision for human/object action understanding.

**Limitations and Future Directions**

The system is tailored for the specific environment and action types found in traffic scenes.

**Relevance to Primary Paper**

This paper is cited in the introduction of **FineSports** as an example of a real-world application of human action understanding, specifically in autonomous driving[cite: 46]. This helps to motivate the research by showing its relevance to impactful technologies.

## 40. Unintentional Action Localization (Xu et al., TIP 2022) [40]

**Key Concepts and Taxonomy**

This work tackles the problem of **Unintentional Action Localization**, which involves identifying actions that are performed without a clear goal or intent. The authors propose a method that uses counterfactual examples to help the model distinguish between intentional and unintentional actions.

**Main Contributions and Findings**

The key contribution is the novel problem formulation of unintentional action localization and a counterfactual reasoning approach to solve it. This pushes action understanding beyond recognizing predefined, goal-oriented activities.

**Limitations and Future Directions**

The definition of "unintentional" can be ambiguous and subjective, making annotation and evaluation challenging.

**Relevance to Primary Paper**

This paper is cited in the introduction of **FineSports** as an example of an application of human action understanding in "abnormal monitoring"[cite: 46]. One of the authors of this paper, Jinglin Xu, is also the first author of the FineSports paper, showing a trajectory of research in nuanced action understanding.

## 41. FineDiving (Xu et al., CVPR 2022) [41]

**Key Concepts and Taxonomy**

This paper introduces **FineDiving**, a fine-grained dataset for procedure-aware action quality assessment. The dataset contains diving videos with detailed annotations of the procedural steps within each dive (e.g., take-off, flight, entry) and scores for each step. The goal is to assess quality based on an understanding of this procedure.

**Main Contributions and Findings**

The key contributions are the **FineDiving** dataset and the concept of "procedure-aware" AQA[cite: 77]. This moves beyond a single score for an action to a more detailed analysis of its component parts.

**Limitations and Future Directions**

The dataset is focused on a single-person sport (diving) and the task is quality assessment, not multi-person localization.

**Relevance to Primary Paper**

**FineDiving** is cited multiple times in **FineSports**[cite: 46, 77, 262]. It is used as a key example of recent work in fine-grained sports video analysis and is included in the dataset comparison table. The procedural nature of FineDiving's annotations is conceptually similar to the step-level sub-action annotations in FineSports, though FineSports applies this idea to a multi-person STAL task.

## 42. NBA Dataset / Social Adaptive Module (Yan et al., ECCV 2020) [42]

**Key Concepts and Taxonomy**

This paper tackles weakly-supervised group activity recognition, introducing a large-scale **NBA** basketball dataset for this task. It proposes a **Social Adaptive Module (SAM)** that models the interactions and relationships between different players to better recognize the collective group activity.

**Main Contributions and Findings**

The main contributions are the large-scale **NBA** dataset for group activity recognition and the SAM model, which explicitly models social/player interactions[cite: 501].

**Limitations and Future Directions**

The dataset provides only video-level labels for group activities (e.g., "3-point shot attempt"), lacking the fine-grained, individual-player, spatial-temporal annotations of FineSports.

**Relevance to Primary Paper**

This work is highly relevant and is cited as a key point of comparison[cite: 79, 262]. The **FineSports** authors contrast their dataset with this **NBA** dataset, noting that while both use NBA footage, FineSports provides much more detailed, fine-grained, and localized annotations for individual players, enabling a different set of research tasks (STAL vs. group recognition).

## 43. STEP (Yang et al., CVPR 2019) [43]

**Key Concepts and Taxonomy**

This paper proposes **STEP (Spatio-Temporal Progressive Learning)**, a method for video action detection. The key idea is to progressively refine action proposals through a series of stages. It starts with coarse proposals and iteratively improves their temporal boundaries and classification confidence.

**Main Contributions and Findings**

The main contribution is the progressive learning framework, which breaks down the difficult action detection problem into a series of simpler steps, leading to more accurate localization.

**Limitations and Future Directions**

The multi-stage design can be computationally expensive and complex compared to single-shot end-to-end models.

**Relevance to Primary Paper**

This work is cited in the **FineSports** paper in relation to the design of its learnable action tube queries[cite: 88, 333]. The concept of representing actions as spatio-temporal entities (tubes or cuboids) that are then reasoned about, as seen in STEP, is a foundational idea that informs the query-based design of the **POSTAL** model's detector.

## 44. LOGO (Zhang et al., CVPR 2023) [44]

**Key Concepts and Taxonomy**

This paper introduces **LOGO**, a long-form video dataset for group action quality assessment, based on artistic swimming competitions[cite: 84]. It provides detailed annotations for both individual actions and group formations, along with quality scores.

**Main Contributions and Findings**

The main contribution is the **LOGO** dataset, which addresses the challenge of evaluating synchronized group activities in long videos. It provides a unique benchmark for multi-person action and formation quality assessment.

**Limitations and Future Directions**

The dataset is specific to artistic swimming, a highly structured group sport. The primary task is AQA.

**Relevance to Primary Paper**

**LOGO** is included in the dataset comparison table in **FineSports** (Table 1)[cite: 79]. It is cited as a recent example of a multi-person sports dataset with detailed annotations, highlighting the research trend toward more complex group activity analysis, which FineSports also addresses, albeit for a different sport and task.

## 45. TubeR (Zhao et al., CVPR 2022) [45]

**Key Concepts and Taxonomy**

This paper introduces **TubeR (Tubelet Transformer for Video Action Detection)**, a Transformer-based approach for spatial-temporal action detection[cite: 508]. It formulates the task as a set prediction problem. [cite\_start]The model uses a set of learnable "tubelet queries" (short action tube segments) and a Transformer encoder-decoder architecture to directly predict the final action tubes and their classes from video features[cite: 97].

**Main Contributions and Findings**

TubeR's primary contribution is being one of the first end-to-end, query-based Transformer models for video action detection[cite: 508]. This approach eliminates the need for complex post-processing steps like non-maximum suppression, which were common in previous methods.

**Limitations and Future Directions**

The model's performance is tied to the design and number of learnable tubelet queries. It detects all actions rather than focusing on a specific, targeted action.

**Relevance to Primary Paper**

**TubeR** is cited as a leading state-of-the-art (SOTA) method against which the proposed **POSTAL** model is benchmarked[cite: 88, 360, 365]. [cite\_start]The comparison shows that POSTAL outperforms TubeR by 2.06% on frame-mAP[cite: 379]. The authors attribute this to POSTAL's ability to incorporate learnable language prompts, which guide the localization more accurately than the purely learnable (but unguided) queries in TubeR. [cite\_start]The query-based mechanism in POSTAL's ATD module is conceptually inspired by work like TubeR[cite: 333].

**Links to Referenced Papers & Sources**

1. **Diving48 (Li et al., ECCV 2018)**

**Paper Link:** [**https://openaccess.thecvf.com/content\_ECCV\_2018/html/Yingwei\_Li\_Diving48\_A\_Fine-grained\_ECCV\_2018\_paper.html**](https://www.google.com/search?q=https://openaccess.thecvf.com/content_ECCV_2018/html/Yingwei_Li_Diving48_A_Fine-grained_ECCV_2018_paper.html)

1. **FP-Basket (Bertasius et al., ICCV 2017)**

**Paper Link:** [**https://openaccess.thecvf.com/content\_ICCV\_2017/html/Bertasius\_Am\_I\_a\_ICCV\_2017\_paper.html**](https://www.google.com/search?q=https://openaccess.thecvf.com/content_ICCV_2017/html/Bertasius_Am_I_a_ICCV_2017_paper.html)

1. **Kinetics (Carreira and Zisserman, CVPR 2017)**

**Paper Link:** [**https://arxiv.org/abs/1705.07750**](https://arxiv.org/abs/1705.07750)

1. **SportsMOT (Cui et al., arXiv 2023)**

**Paper Link: https://arxiv.org/abs/2307.14929**

**GitHub: https://github.com/yml-lab/SportsMOT**

1. **BERT (Devlin et al., NAACL 2019)**

**Paper Link:** [**https://arxiv.org/abs/1810.04805**](https://arxiv.org/abs/1810.04805)

1. **ActivityNet (Heilbron et al., CVPR 2015)**

**Paper Link: https://openaccess.thecvf.com/content\_cvpr\_2015/html/Caba\_Heilbron\_ActivityNet\_A\_Large-Scale\_2015\_CVPR\_paper.html**

**Project Page: http://activity-net.org/**

1. **Abductive Ego-view Accident Video Understanding (Fang et al., arXiv 2024)**

**Paper Link:** [**https://arxiv.org/abs/2403.01896**](https://arxiv.org/abs/2403.01896)

1. **SlowFast Networks (Feichtenhofer et al., ICCV 2019)**

**Paper Link: https://arxiv.org/abs/1812.03982**

**GitHub: https://github.com/facebookresearch/SlowFast**

1. **Neural Bipartite Matching (Georgiev and Lió, arXiv 2020)**

**Paper Link:** [**https://arxiv.org/abs/2005.07923**](https://arxiv.org/abs/2005.07923)

1. **Finding Action Tubes (Gkioxari and Malik, CVPR 2015)**

**Paper Link:** [**https://openaccess.thecvf.com/content\_cvpr\_2015/html/Gkioxari\_Finding\_Action\_Tubes\_2015\_CVPR\_paper.html**](https://openaccess.thecvf.com/content_cvpr_2015/html/Gkioxari_Finding_Action_Tubes_2015_CVPR_paper.html)

1. **End-to-end Spatio-temporal Action Localisation with a Video Transformer (Gritsenko et al., arXiv 2021)**

**Paper Link:** [**https://arxiv.org/abs/2110.04650**](https://arxiv.org/abs/2110.04650)

1. **AVA Dataset (Gu et al., CVPR 2018)**

**Paper Link: https://arxiv.org/abs/1705.08421**

**Project Page: https://research.google.com/ava/**

1. **3D ResNeXt (Hara et al., WACV 2018)**

**Paper Link:** [**https://arxiv.org/abs/1711.09577**](https://arxiv.org/abs/1711.09577)

1. **Human Action Recognition in Sports (Host and Ivašić-Kos, Heliyon 2022)**

**Paper Link:** [**https://www.cell.com/heliyon/fulltext/S2405-8440(22)01548-X**](https://www.google.com/search?q=https://www.cell.com/heliyon/fulltext/S2405-8440(22)01548-X)

1. **T-CNN (Hou et al., ICCV 2017)**

**Paper Link:** [**https://openaccess.thecvf.com/content\_ICCV\_2017/html/Hou\_Tube\_Convolutional\_Neural\_ICCV\_2017\_paper.html**](https://www.google.com/search?q=https://openaccess.thecvf.com/content_ICCV_2017/html/Hou_Tube_Convolutional_Neural_ICCV_2017_paper.html)

1. **Hierarchical Deep Temporal Model for Group Activity Recognition (Ibrahim et al., CVPR 2016)**

**Paper Link:** [**https://openaccess.thecvf.com/content\_cvpr\_2016/html/Ibrahim\_A\_Hierarchical\_Deep\_CVPR\_2016\_paper.html**](https://openaccess.thecvf.com/content_cvpr_2016/html/Ibrahim_A_Hierarchical_Deep_CVPR_2016_paper.html)

1. **FineFS / LUSD-NET (Ji et al., ACM MM 2023)**

**Paper Link: https://dl.acm.org/doi/10.1145/3581783.3611956**

**GitHub: https://github.com/j-j-jim/LUSD-Net**

1. **Action Tubelet Detector (Kalogeiton et al., ICCV 2017)**

**Paper Link:** [**https://openaccess.thecvf.com/content\_ICCV\_2017/html/Kalogeiton\_Action\_Tubelet\_Detector\_ICCV\_2017\_paper.html**](https://www.google.com/search?q=https://openaccess.thecvf.com/content_ICCV_2017/html/Kalogeiton_Action_Tubelet_Detector_ICCV_2017_paper.html)

1. **The Kinetics Human Action Video Dataset (Kay et al., arXiv 2017)**

**Paper Link:** [**https://arxiv.org/abs/1705.06950**](https://arxiv.org/abs/1705.06950)

1. **YOWO (Köpüklü et al., arXiv 2019)**

**Paper Link:** [**https://arxiv.org/abs/1912.12356**](https://arxiv.org/abs/1912.12356)

1. **UniformerV2 (Li et al., ICCV 2023)**

**Paper Link:** [**https://openaccess.thecvf.com/content/ICCV2023/html/Li\_UniFormerV2\_Spatiotemporal\_Learning\_by\_Sharing\_Fresh\_Information\_and\_When\_ICCV\_2023\_paper.html**](https://www.google.com/search?q=https://openaccess.thecvf.com/content/ICCV2023/html/Li_UniFormerV2_Spatiotemporal_Learning_by_Sharing_Fresh_Information_and_When_ICCV_2023_paper.html)

1. **MOC (Li et al., ECCV 2020)**

**Paper Link:** [**https://arxiv.org/abs/2007.09484**](https://arxiv.org/abs/2007.09484)

1. **MultiSports (Li et al., ICCV 2021)**

**Paper Link: https://openaccess.thecvf.com/content/ICCV2021/html/Li\_MultiSports\_A\_Multi-Person\_Sports\_Video\_Dataset\_for\_Fine-Grained\_Spatio-Temporal\_Action\_ICCV\_2021\_paper.html**

**GitHub: https://github.com/MCG-NJU/MultiSports**

1. **FSD-10 (Liu et al., arXiv 2020)**

**Note: The summary cites Liu et al., but the original FSD paper is by Qi et al. (2022). The summary may be referring to an earlier or different work. Here is the primary FSD paper.**

**Paper Link: https://arxiv.org/abs/2207.13596**

1. **MCFS (Liu et al., AAAI 2021)**

**Paper Link:** [**https://ojs.aaai.org/index.php/AAAI/article/view/16307**](https://ojs.aaai.org/index.php/AAAI/article/view/16307)

1. **RFSJ (Liu et al., MM 2023)**

**Paper Link:** [**https://dl.acm.org/doi/abs/10.1145/3581783.3612269**](https://dl.acm.org/doi/abs/10.1145/3581783.3612269)

1. **Video Swin Transformer (Liu et al., CVPR 2022)**

**Paper Link:** [**https://arxiv.org/abs/2106.13230**](https://arxiv.org/abs/2106.13230)

1. **GolfDB (McNally et al., CVPRW 2019)**

**Paper Link:** [**https://openaccess.thecvf.com/content\_CVPRW\_2019/html/FGVC/McNally\_Fine-Grained\_Action\_Recognition\_in\_Golf\_CVPRW\_2019\_paper.html**](https://www.google.com/search?q=https://openaccess.thecvf.com/content_CVPRW_2019/html/FGVC/McNally_Fine-Grained_Action_Recognition_in_Golf_CVPRW_2019_paper.html)

1. **MTL-AQA (Parmar and Morris, CVPR 2019)**

**Paper Link:** [**https://openaccess.thecvf.com/content\_CVPR\_2019/html/Parmar\_What\_and\_How\_Well\_A\_New\_Look\_at\_Action\_Quality\_CVPR\_2019\_paper.html**](https://www.google.com/search?q=https://openaccess.thecvf.com/content_CVPR_2019/html/Parmar_What_and_How_Well_A_New_Look_at_Action_Quality_CVPR_2019_paper.html)

1. **Multi-region two-stream R-CNN (Peng and Schmid, ECCV 2016)**

**Paper Link:** [**https://link.springer.com/chapter/10.1007/978-3-319-46484-8\_43**](https://link.springer.com/chapter/10.1007/978-3-319-46484-8_43)

1. **Fine-grained Activity Recognition in Baseball Videos (Piergiovanni and Ryoo, CVPRW 2018)**

**Paper Link:** [**https://openaccess.thecvf.com/content\_cvpr\_2018\_workshops/w4/html/Piergiovanni\_Fine-Grained\_Activity\_Recognition\_CVPR\_2018\_paper.html**](https://www.google.com/search?q=https://openaccess.thecvf.com/content_cvpr_2018_workshops/w4/html/Piergiovanni_Fine-Grained_Activity_Recognition_CVPR_2018_paper.html)

1. **FineGym (Shao et al., CVPR 2020)**

**Paper Link: https://openaccess.thecvf.com/content\_CVPR\_2020/html/Shao\_FineGym\_A\_Hierarchical\_Video\_Dataset\_for\_Fine-Grained\_Action\_Understanding\_CVPR\_2020\_paper.html**

**Project Page: https://sdolivia.github.io/FineGym/**

1. **TAPOS (Shao et al., CVPR 2020)**

**Paper Link:** [**https://openaccess.thecvf.com/content\_CVPR\_2020/html/Shao\_Intra-\_and\_Inter-Action\_Understanding\_With\_Temporal\_Action\_Parsing\_CVPR\_2020\_paper.html**](https://www.google.com/search?q=https://openaccess.thecvf.com/content_CVPR_2020/html/Shao_Intra-_and_Inter-Action_Understanding_With_Temporal_Action_Parsing_CVPR_2020_paper.html)

1. **Real-time Elderly Monitoring (Sun and Chen, ISMICT 2022)**

**Paper Link:** [**https://ieeexplore.ieee.org/document/9904283**](https://ieeexplore.ieee.org/document/9904283)

1. **VideoMAE (Tong et al., NeurIPS 2022)**

**Paper Link: https://arxiv.org/abs/2203.12602**

**GitHub: https://github.com/MCG-NJU/VideoMAE**

1. **CSN (Tran et al., ICCV 2019)**

**Paper Link:** [**https://openaccess.thecvf.com/content\_ICCV\_2019/html/Tran\_Video\_Classification\_With\_Channel-Separated\_Convolutional\_Networks\_ICCV\_2019\_paper.html**](https://openaccess.thecvf.com/content_ICCV_2019/html/Tran_Video_Classification_With_Channel-Separated_Convolutional_Networks_ICCV_2019_paper.html)

1. **Learning to Track for Spatio-temporal Action Localization (Weinzaepfel et al., ICCV 2015)**

**Paper Link:** [**https://openaccess.thecvf.com/content\_iccv\_2015/html/Weinzaepfel\_Learning\_to\_Track\_ICCV\_2015\_paper.html**](https://openaccess.thecvf.com/content_iccv_2015/html/Weinzaepfel_Learning_to_Track_ICCV_2015_paper.html)

1. **Learning to Score Figure Skating (Xu et al., TCSVT 2019)**

**Paper Link:** [**https://ieeexplore.ieee.org/document/8712711**](https://ieeexplore.ieee.org/document/8712711)

1. **Action Recognition in Traffic Scenes (Xu et al., TITS 2022)**

**Paper Link:** [**https://ieeexplore.ieee.org/document/9796726**](https://ieeexplore.ieee.org/document/9796726)

1. **Unintentional Action Localization (Xu et al., TIP 2022)**

**Paper Link:** [**https://ieeexplore.ieee.org/document/9803362**](https://ieeexplore.ieee.org/document/9803362)

1. **FineDiving (Xu et al., CVPR 2022)**

**Paper Link: https://openaccess.thecvf.com/content/CVPR2022/html/Xu\_FineDiving\_A\_Fine-Grained\_Dataset\_for\_Procedure-Aware\_Action\_Quality\_Assessment\_CVPR\_2022\_paper.html**

**GitHub: https://github.com/xujing-ps/FineDiving**

1. **NBA Dataset / Social Adaptive Module (Yan et al., ECCV 2020)**

**Paper Link:** [**https://www.ecva.net/papers/eccv\_2020/papers\_ECCV/html/2069\_ECCV\_2020\_paper.php**](https://www.google.com/search?q=https://www.ecva.net/papers/eccv_2020/papers_ECCV/html/2069_ECCV_2020_paper.php)

1. **STEP (Yang et al., CVPR 2019)**

**Paper Link:** [**https://openaccess.thecvf.com/content\_CVPR\_2019/html/Yang\_STEP\_Spatio-Temporal\_Progressive\_Learning\_for\_Video\_Action\_Detection\_CVPR\_2019\_paper.html**](https://openaccess.thecvf.com/content_CVPR_2019/html/Yang_STEP_Spatio-Temporal_Progressive_Learning_for_Video_Action_Detection_CVPR_2019_paper.html)

1. **LOGO (Zhang et al., CVPR 2023)**

**Paper Link:** [**https://openaccess.thecvf.com/content/CVPR2023/html/Zhang\_LOGO\_A\_Long-Form\_Video\_Dataset\_for\_Group\_Action\_Quality\_Assessment\_CVPR\_2023\_paper.html**](https://openaccess.thecvf.com/content/CVPR2023/html/Zhang_LOGO_A_Long-Form_Video_Dataset_for_Group_Action_Quality_Assessment_CVPR_2023_paper.html)

1. **TubeR (Zhao et al., CVPR 2022)**

**Paper Link:** [**https://openaccess.thecvf.com/content/CVPR2022/html/Zhao\_TubeR\_Tubelet\_Transformer\_for\_Video\_Action\_Detection\_CVPR\_2022\_paper.html**](https://openaccess.thecvf.com/content/CVPR2022/html/Zhao_TubeR_Tubelet_Transformer_for_Video_Action_Detection_CVPR_2022_paper.html)