Relative Boundary Modeling: Technical Report

Jun Yu[1], Leilei Wang[1], Renjie Lu[1],Shuoping Yang[1], Renda Li[1], Lei Wang[1]
Minchuan Chen[2], Qingying Zhu[2], Shaojun Wang[2], Jing Xiao[2]
[1]University of Science and Technology of China, [2]Ping An Technology

Abstract

In this report, we present our solution for the Phase-II of the Cricket Bowl Release Detection challenge, in MMSports 2023. We already published a paper "Relative Boundary Modeling: A High-Resolution Cricket Bowl Release Detection Framework with I3D Features" [4] for Phase-I (Paper submission) of that challenge. In this technical report, we improved our feature extraction method and achieved better results than in Phase-I. Our proposed method achieves a PQ score of 0.703, an SQ score of 0.901, and an RQ score of 0.780 on the challenge set of the DeepSportradar Cricket Bowl Release Dataset. Through this approach, our team, USTC_IAT_United, won the first place in Phase-II of the DeepSportradar Cricket Bowl Release Challenge. Here is our GitHub repository: https://github.com/haohantianchen/2023-winners-Cricket-Bowl-Release-Detection-challenge

1. Dataset

The provided dataset for this challenge consists of two main parts: the Challenge Set and the Training/Validation set. Both parts consist of videos with a frame rate of 30 frames per second. Annotations are provided for the Training/Validation set. To ensure successful implementation of our method on the dataset, we processed the data and converted the format of the Training/Validation set to the standard Thumos14-30fps format.

Although the original dataset annotations for cricketing actions were divided into two categories, namely, "is bowling"" and "ball release", we found that "ball release" only had one frame, making it challenging to effectively recognize the action. Therefore, in our attempts, we unified all events and labeled them as a single category "CricketBowling" thereby enhancing data usability and result accuracy.

2. Method

In this section, we adopt a high-performance temporal action detection model called TriDet [2]. The adopted model consists of three main components: Backbone, Scalable-Granularity Perception (SGP) Feature Pyramid, and Detection Head. For the backbone, we adopt the VideoMAEv2 [3] to extract features from the input videos, which are then concatenated to form the final feature representation.

2.1. Feature Extraction

We believe that efficient feature extraction from the target videos is essential for achieving more accurate detection results. Therefore, in the first stage, we used the feature extraction method of the Inflated 3D ConvNet (I3D) [1]. However, with our subsequent research, we found that the feature extraction method of VideoMAEv2 [3], as shown in Fig. 1, is more suitable for this task. It was used in our Phase-II approach.

VideoMAEv2 employs a dual-mask strategy based on prior knowledge of video data redundancy, simultaneously masking input tokens for both the encoder and decoder. Following an intermediate fine-tuning approach in the training strategy, it adopts a progressive training paradigm. It begins with video mask pretraining on a multimillion-level unlabeled, multi-source mixed video dataset, and then proceeds with post-pretraining fine-tuning on a labeled mixed video dataset. This method can be effectively applied to a variety of downstream tasks, and it has achieved leading performance in various downstream tasks, including action recognition, spatiotemporal action detection, and temporal action detection.

2.2. TriDet

TriDet is a one-stage framework designed to address the boundary prediction problem in video action detection, as shown in Fig. 2. We have provided a more detailed discussion of it in our previous paper [4]. This framework

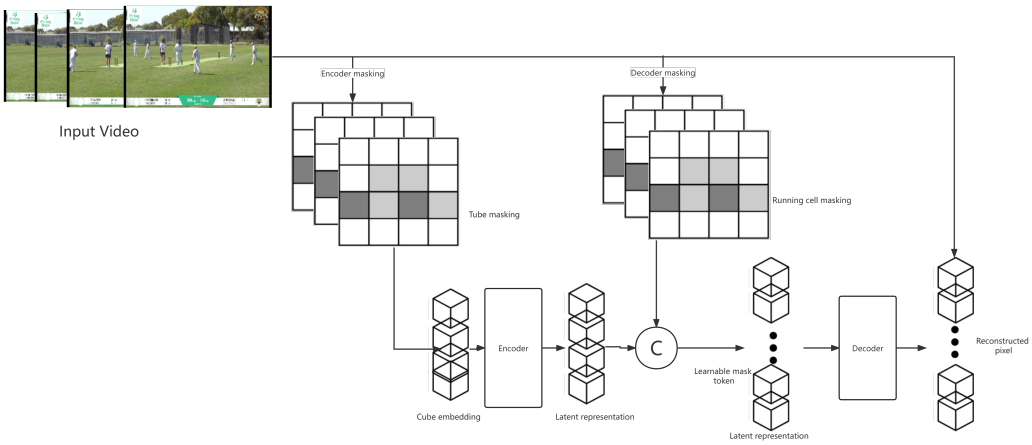

Figure 1. VideoMAE with dual masking. To improve the overall efficiency of computation and memory in video masked autoencoding, it proposes to mask the decoder as well and devise the dual masking strategy. Similar to the encoder, it also applies a masking map to the deocoder and simply reconstruct a subset of pixel cubes selected by the running cell masking. The final reconstruction loss only applies for the invisible tokens dropped by the encoder.

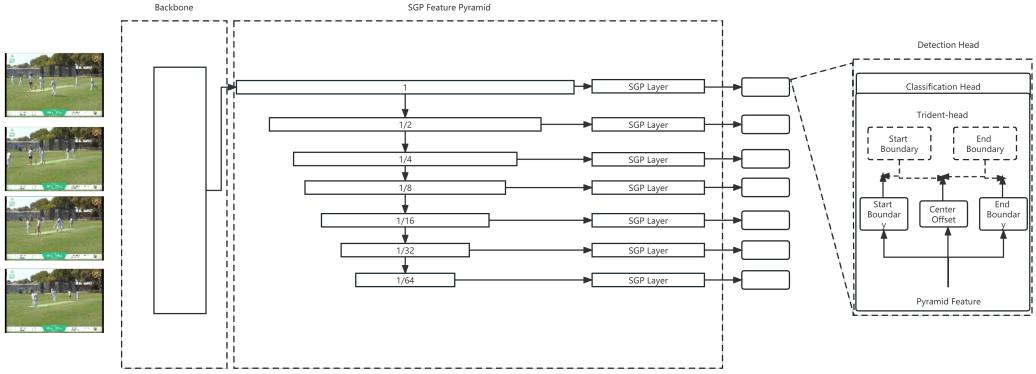

Figure 2. Illustration of TriDet. We construct the feature pyramid using the Scalable-Granularity Perception (SGP) layer. The features corresponding to each level are then input into a shared-weight detection head to obtain the detection result, which comprises a classification head and a Trident-head. The Trident-head estimates the boundary offset based on a relative distribution predicted by three branches: Start Boundary, End Boundary, and Center Offset.

consists of the Trident-head and Scalable-Granularity Perception (SGP) layer [2]. The Trident-head models action boundaries by estimating the relative probability distribution around the boundaries, thereby alleviating the boundary prediction issue. The SGP layer is used to mitigate the ranking loss problem generated by self-attention in video features and aggregates information at different temporal granularities.

3. Results

We present the results of the two phases of challenge here, and you can see the huge improvement in switching from feature extraction to VideoMAEv2.

Phase-I. In Phase-I, we employed the I3D+Tridet approach, which is elaborated in detail in our paper [4], and the results are shown in the Table. 1.

Phase-II. In Phase II, by replacing I3D's feature extraction approach with VideoMAEv2, we achieved better results, as shown in the Table. 2.

4. Conclusions

In this technical report, we propose a new approach using VideoMAE+TriDet for the Cricket Bowl Release Detection challenge. This approach outperforms our previous solution, I3D+TriDet, on the challenge set of the Cricket Bowl Release Dataset. It ultimately resulted in us achieving SOTA during the Prize submission phase of the Cricket Bowl Release Detection challenge, in MMSports '23 at ACM MM 2023.

Table 1. Results on the Challenge-Set (Paper submission).

Participation team	PQ	SQ	RQ
b703	0.643	0.927	0.693
snowdistance	0.591	0.932	0.643
USTC_IAT_United	0.519	0.822	0.632

Table 2. Results on the Challenge-Set (Prize submission).

Participation team	PQ	SQ	RQ
USTC_IAT_United	0.703	0.901	0.780
b703	0.643	0.927	0.693
snowdistance	0.591	0.932	0.643

References

[1] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[2] Dingfeng Shi, Yujie Zhong, Qiong Cao, Lin Ma, Jia Li, and Dacheng Tao. Tridet: Temporal action detection with relative boundary modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18857–18866, 2023. 1, 2

[3] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023. 1

[4] Jun Yu, Leilei Wang, Renjie Lu, Shuoping Yang, Renda Li, Lei Wang, Minchuan Chen, Qingying Zhu, Shaojun Wang, and Jing Xiao. Relative boundary modeling: A high-resolution cricket bowl release detection framework with i3d features. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, MMSports '23, page 151–159, New York, NY, USA, 2023. Association for Computing Machinery. 1, 2