mathematics

MDPI

Article

Boundary-Match U-Shaped Temporal Convolutional Network for Vulgar Action Segmentation

Zhengwei Shen [1,*], Ran Xu [1], Yongquan Zhang [2], Feiwei Qin [1], Ruiquan Ge [1], Changmiao Wang [3] and Masahiro Toyoura [4]

[1] School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China; 201050083@hdu.edu.cn (R.X.); qinfeiwei@hdu.edu.cn (F.Q.); gespring@hdu.edu.cn (R.G.)
[2] School of Data Sciences, Zhejiang University of Finance and Economics, Hangzhou 310018, China; zyq@zufe.edu.cn
[3] Medical Big Data Lab, Shenzhen Research Institute of Big Data, Shenzhen 518172, China; wangcm@sribd.cn
[4] Department of Computer Science and Engineering, University of Yamanashi, Kofu 400-8511, Japan; mtoyoura@yamanashi.ac.jp
* Correspondence: 211050079@hdu.edu.cn

Abstract: The advent of deep learning has provided solutions to many challenges posed by the Internet. However, efficient localization and recognition of vulgar segments within videos remain formidable tasks. This difficulty arises from the blurring of spatial features in vulgar actions, which can render them indistinguishable from general actions. Furthermore, issues of boundary ambiguity and over-segmentation complicate the segmentation of vulgar actions. To address these issues, we present the **B**oundary-**M**atch **U**-shaped **T**emporal **C**onvolutional **N**etwork (BMUTCN), a novel approach for the segmentation of vulgar actions. The BMUTCN employs a U-shaped architecture within an encoder–decoder temporal convolutional network to bolster feature recognition by leveraging the context of the video. Additionally, we introduce a boundary-match map that fuses action boundary inform ation with greater precision for frames that exhibit ambiguous boundaries. Moreover, we propose an adaptive internal block suppression technique, which substantially mitigates over-segmentation errors while preserving accuracy. Our methodology, tested across several public datasets as well as a bespoke vulgar dataset, has demonstrated state-of-the-art performance on the latter.

Keywords: vulgar action segmentation; boundary-match; U-shaped network; temporal convolutional network; adaptive internal block suppression

MSC: 90B20

Citation: Shen, Z.; Xu, R.; Zhang, Y.; Qin, F.; Ge, R.; Wang, C.; Toyoura, M. Boundary-Match U-Shaped Temporal Convolutional Network for Vulgar Action Segmentation. *Mathematics* **2024**, *12*, 899. https://doi.org/10.3390/math12060899

Academic Editors: Xiaojiang Peng, Linlin Shen and Yang You

Received: 8 February 2024
Revised: 14 March 2024
Accepted: 17 March 2024
Published: 18 March 2024

1. Introduction

Efficient action segmentation and recognition in videos are critical for fostering a healthy Internet environment. The proliferation of video content on the Internet has made vulgar and pornographic materials more accessible, posing potential risks to adolescent development [1–4]. Consequently, the vulgar action segmentation framework demonstrates potential applicability across diverse scenarios. Within the realms of social media platforms and online advertising, it possesses the capability to monitor and selectively filter out content deemed obscene or inappropriate that is contributed by users. This function is integral to upholding a positive and amicable environment on such platforms. Additionally, it serves a protective role, shielding children from exposure to potentially harmful or unsuitable content. In recent years, advancements in deep learning have yielded substantial improvements in video processing capabilities, and notable successes have been achieved in identifying general actions within extended video sequences [5–10]. Despite these

advances, the task of densely labeling every frame in a video remains a daunting challenge, particularly within complex datasets such as those containing vulgar content [11].

The primary obstacles to effective action segmentation and recognition in vulgar content can be attributed to the following factors: (1) Inadequate representational features of vulgar actions in spatial domains: Unlike pornographic actions, vulgar actions exhibit spatial characteristics that are akin to those of normal actions, making them less distinct. (2) Boundary ambiguity: During the video annotation process, even when labels are modified, there may be no significant change in the action information, leading to low frame accuracy at action boundaries. (3) Over-segmentation errors: Models tend to over-segment extended video clips to enhance accuracy, which inadvertently generates numerous fragmented short clips. These fragmented segments, referred to as internal blocks, disrupt the continuity and coherence of the video segmentation [12].

In this study, we introduce the Boundary-Match U-shaped Temporal Convolutional Network (BMUTCN), a novel framework designed for the effective detection and segmentation of vulgar actions in videos. As depicted in Figure 1, our approach incorporates a U-shaped structure into the temporal convolutional network (TCN) [13], harnessing the temporal dynamics of video content. Drawing inspiration from the Boundary-Matching Network, we have developed the Boundary-Match (BM) Module to precisely capture the boundary information of actions within the video. Moreover, we have enhanced the Local Burr Suppression technique from the Efficient Two-Step Networks [14], formulating the Adaptive Internal Block Suppression (AIBS). This innovative module integrates an adaptive mechanism into the Local Blur Suppression (LBS) architecture, enabling the automatic calibration of hyperparameters to suit diverse samples. Given the absence of publicly available vulgar datasets, we curated a vulgar action video dataset from public video platforms with expert guidance tailored specifically for vulgar action segmentation. Our contributions to the field are multifaceted and can be summarized as follows:

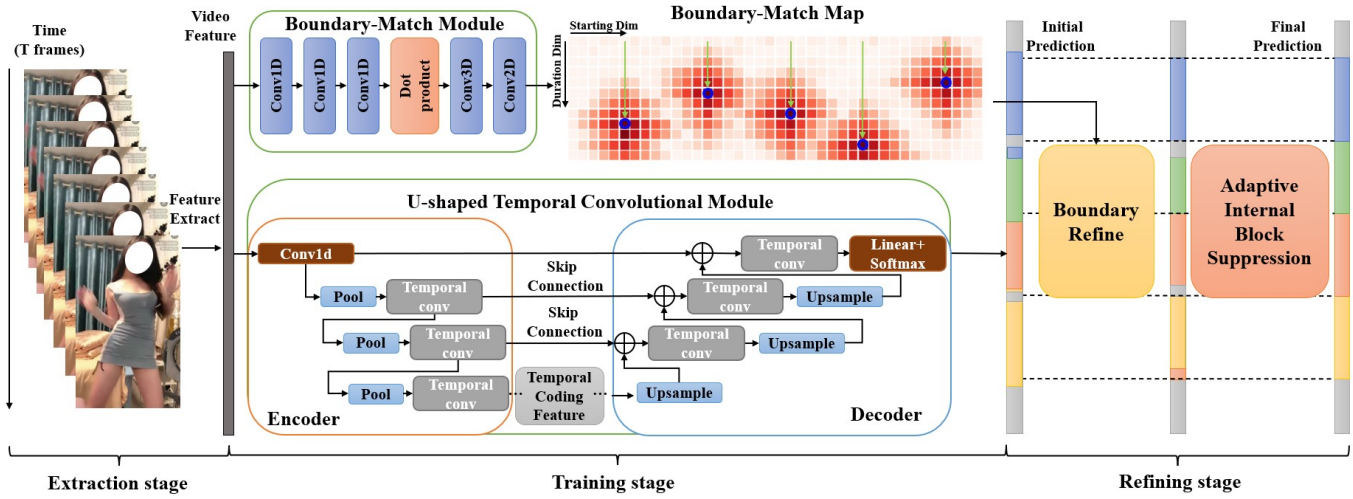

Figure 1. The architecture of BMUTCN. Given a video, the process begins with the extraction phase, where the video is processed into a sequential array of features. During the training phase, the Boundary-Match Module (BMM) constructs a boundary-match map that encapsulates the confidence levels of all potential action segments. Concurrently, the U-shaped Temporal Convolutional Module (UTCM) applies an Encoder–Decoder architecture to purify the input sequence of noise and subsequently delivers frame-level predictions. In the subsequent refinement phase, the preliminary predictions are meticulously refined twice—initially utilizing the BM map and subsequently via the AIBS method.

1. Development of a U-shaped TCN for enhanced temporal information utilization: We introduce a novel U-shaped TCN architecture that effectively leverages the tem-

poral information present in videos, specifically addressing the issue of ambiguous boundaries in vulgar content.
2. Introduction of the BM Module for precise action boundaries: Through the implementation of the BM Module, we refine predictions using boundary information, resulting in a marked improvement in the accuracy of action boundary delineation.
3. Creation of the AIBS for mitigating over-segmentation: Our AIBS technique, an unsupervised refinement method, is capable of adaptively adjusting hyperparameters, thereby effectively reducing over-segmentation errors while maintaining high accuracy.
4. Extensive model comparison across three challenging datasets: Our model has been rigorously evaluated against other methods on the 50Salads dataset [15], the Georgia Tech Egocentric Activities (GTEA) dataset [16], and our self-constructed vulgar action dataset. The results demonstrate that our approach sets a new benchmark for action segmentation on the vulgar action dataset.

2. Related Work

2.1. Action Recognition

In recent years, neural networks have played a pivotal role in video processing, particularly in the identification and detection of vulgar and pornographic content. Wehrmann et al. [17] introduced a deep learning framework that leverages convolutional networks and LSTM networks for the automated detection and obscuration of pornographic material. Perez et al. [10] suggested the fusion of optical flow and MPEG motion vectors to capture both dynamic and static information, enhancing the efficiency of pornographic content identification in videos. Mallmann et al. [18] designed the PPCensor architecture, which is implemented on specialized hardware and utilizes video-oriented streaming agents for real-time detection capabilities. Upon detecting pornographic frames, this system can obscure the offending content, enabling continuous monitoring. Song et al. [19] championed a multimodal strategy, incorporating image frames, optical flow, and audio data to more accurately identify pornographic content within videos. Furthermore, Vitorino et al. [2] investigated diverse transfer learning strategies for modeling CNNs, harnessing advanced data-driven methodologies and deep CNNs for the detection of child pornography.

Vulgar content, with its more ambiguous nature compared to explicit pornography, necessitates a more precise mechanism for feature extraction. Models developed for pornographic action recognition could serve as the foundational framework for vulgar action segmentation tasks, providing frame-level feature extraction in extended, untrimmed videos.

2.2. Action Detection

The task of temporal action detection involves identifying both the temporal boundaries and categories of action instances within untrimmed videos. Consequently, this task can be segmented into two distinct phases: action proposal generation and action classification. Gao et al. [20] introduced the Complementary Temporal Action Proposal (CTAP) generator, which capitalizes on the complementary strengths of sliding window ranking and action score grouping to generate action proposals. Lin et al. [21] presented the Boundary Sensitive Network (BSN), taking a "local-to-global" approach. The BSN aims to retrieve high-quality proposals with a limited number of candidates to achieve high recall and overlap for actual action instances. Yu et al. [22] proposed a novel approach to image segmentation using an adaptive bilateral filter-based local region model, which potentially offers insights for discerning action boundaries in video content by addressing the challenge of blurred boundaries. Liu et al. [23] introduced the Multi-Granularity Generator (MGG), which synergizes a segmental proposal generator with a frame action generator, thus allowing for the generation of temporal action proposals from various granular perspectives. Lin et al. [24] put forward the BM mechanism for the evaluation of confidence scores across densely distributed proposals. By representing proposals as matching pairs of start and end boundaries, their model is capable of producing proposals with precise

temporal demarcations and robust confidence assessments. Yu et al. [25] contributed a novel edge-based active contour model (ACM) for medical image segmentation aimed at addressing complex noise interference. This model may serve as a reference for enhancing the refinement process in action segmentation, especially in the presence of noise.

Feature extractors currently used for action proposal generation are often designed with trimmed action classification tasks in mind. Alwassel et al. [26] introduced a novel supervised pre-training paradigm tailored for feature extraction from untrimmed videos. This paradigm not only focuses on classifying activities but also incorporates background clips and global video information to enhance temporal sensitivity.

The task of action proposal generation may be viewed as a simplified variant of action segmentation, providing valuable insights for action detection within lengthy, untrimmed videos. Drawing inspiration from the Boundary-Matching Network (BMN), we propose the integration of a boundary-match mechanism into action segmentation as a refinement technique. This approach effectively mitigates over-segmentation errors and enhances the precision of action segmentation.

2.3. Action Segmentation

Within the current domain of action segmentation, researchers face three primary challenges: enhancing the accuracy of segmentation, discerning frames with indistinct features during transitions between actions, and mitigating over-segmentation errors within action sequences.

Method based on spatiotemporal information fusion. Previous research has established the challenges in achieving precise action segmentation in videos when relying solely on spatial information. Lea et al. [13] addressed this issue by proposing the TCN, which employs a succession of one-dimensional temporal convolution operations to capture the temporal dynamics within videos. With advancements in GPU capabilities, Farha et al. [27] observed that action segmentation accuracy is greatly enhanced when TCN modules are stacked and the prediction sequence is iteratively refined. Cao et al. [12] proposed an AU-TCN that leverages an adaptive receptive field convolution kernel. This approach integrates a U-shaped structure founded on the temporal convolutional network framework, facilitating an effective analysis of the model's high-level and low-level temporal features. Although leveraging spatiotemporal information can improve accuracy, it also has the potential to exacerbate the model's segmentation errors when dealing with frames that contain blurred action boundaries.

Method based on boundary information. Exploiting spatiotemporal information has been shown to enhance the accuracy of action segmentation; however, this approach can concurrently increase the incidence of segmentation errors in frames with blurred boundaries. To address this, Wang et al. [28] introduced a novel smoothing operation, termed local obstacle pooling, which leverages semantic boundary information to coalesce local predictions, thereby reducing the blurriness of frames near action transitions. Complementarily, Ishikawa et al. [29] suggested the incorporation of a dedicated branch for predicting action boundaries. The main network's output is refined using these predicted boundaries, resulting in a significant boost in segmentation performance.

Method based on refining over-segmentation errors. Striking an optimal balance between minimizing over-segmentation and maximizing accuracy is crucial in the realm of action segmentation tasks. Singhania et al. [30] introduced a novel temporal encoder–decoder framework to tackle the issue of sequence fragmentation. Notably, their decoder employs a coarse-to-fine architecture that implicitly ensembles predictions across multiple temporal resolutions. In a different approach, Li et al. [14] developed an unsupervised refinement method known as LBS, which has been demonstrated to substantially diminish over-segmentation errors. Park et al. [31] confronted the over-segmentation challenge using a divide-and-conquer strategy that first seeks to maximize the model's frame-level classification accuracy before proceeding to address and alleviate over-segmentation errors.

3. Method

As discussed in Section 1, the domain of vulgar action segmentation is currently beset by several critical challenges. Firstly, there is the issue of spatial features associated with vulgar actions being inherently ambiguous. Secondly, the segmentation process is often plagued by internal fragmentation due to over-segmentation. Thirdly, frames that are proximal to action boundaries tend to exhibit a high degree of ambiguity. To address these issues, as illustrated in Figure 1, we introduce a novel and effective BMUTCN, which encompasses three distinct stages: extraction, training, and refining.

In the extraction stage, we employed ResNet-101 as the backbone architecture to extract frame-level features from the original video content. Subsequently, during the training phase, these video features were fed into two separate modules: the BMM and the UTCM. The BMM was responsible for generating the boundary-match map, while the UTCM was devised to provide frame-level action predictions. Finally, the refining stage involved leveraging both the boundary-match map produced by the BMM and the frame-level action predictions from the UTCM for boundary refining, which yielded predictions that incorporated boundary information. Moreover, drawing inspiration from ETSN [14], we propose AIBS as an unsupervised technique for refining these predictions.

3.1. Boundary-Match Module

Over-segmentation and accurate action boundary determination remain persistent challenges within the field of action segmentation tasks. Traditional approaches have often relied on auxiliary loss functions designed to suppress short segments during segmentation tasks [24]; however, such strategies tend to be limited in scope and may inadvertently compromise accuracy. In an effort to surmount these limitations, we have adopted the Proposal Evaluation Module (PEM) from BMN [24] as our Boundary-Match Module. This module is tasked with generating a boundary-match map that is enriched with explicit boundary information, thereby facilitating more precise action segmentation.

3.1.1. Boundary-Match Intersection over Union

Drawing on methodologies from the action proposal generation domain, we aimed to harness the complete spectrum of boundary information inherent in video content. To this end, we employed Intersection over Union (IoU) as a measure of confidence. Furthermore, we have re-envisioned the traditional computation of IoU to better accommodate the nuances of multi-class action segmentation tasks. This recalibrated IoU calculation is designed to be more representative of the complex interactions between different action classes and their respective temporal boundaries.

The IoU is a metric derived by dividing the area of overlap between two regions by the area of the union of the same regions. The formula for computing the IoU is presented as follows:

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union}. \tag{1}$$

Utilizing this principle, we concurrently calculated the IoU between the prediction and various segments, ultimately adopting the greatest IoU value as the confidence measure for the prediction. The enhanced formula for this calculation is given by:

$$IoU(i,j) = \max\left(\frac{p_1 \cap y_j}{p_1 \cup y_j}, \frac{p_2 \cap y_j}{p_2 \cup y_j}, ..., \frac{p_i \cap y_j}{p_i \cup y_j}\right), \tag{2}$$

where p_i signifies the i-th prediction, while y_j represents the j-th action segment. The term $p_i \cap y_j$ indicates the intersection, or overlap, between the predicted segment p_i and the true action segment y_j. Conversely, $p_i \cup y_j$ refers to the union of p_i and y_j. The index i ranges from 1 to the total count of predictions pertaining to the action segment, and the index j spans from 1 to the aggregate number of action segments within the video. As depicted in Figure 2, the black anchor boxes symbolize the predictions, and only those segments

that perfectly align with the start and end points of the anchor boxes can attain the highest confidence levels.

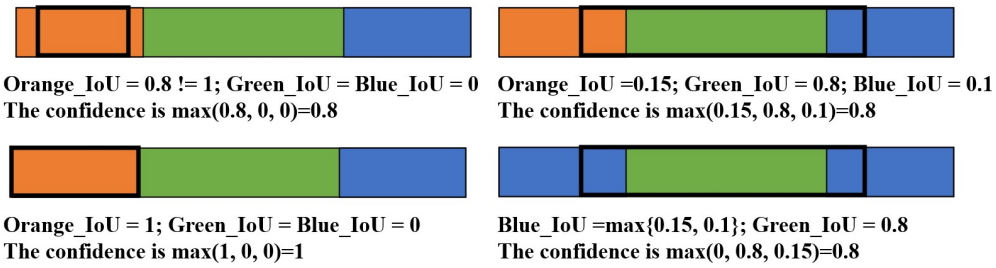

Figure 2. Illustration of boundary-match IoU. To quantify the prediction confidence, we compute the IoU between the predicted anchor box and all existing segments concurrently. Subsequently, the highest IoU value obtained from these comparisons is selected to represent the prediction confidence. This approach ensures a robust estimation of the anchor box's alignment with the ground truth segments, thereby offering a reliable metric for assessing the accuracy of the segmentation model.

3.1.2. Boundary-Match Map

In the BM map, we characterized a predicted action segment by a pair comprising the segment's start time and its duration. We integrated all conceivable predicted segments into a two-dimensional boundary-match map, which was organized according to the predicted segment's start boundary and length. Figure 3 illustrates that in this matrix representation, action segments within each column possess an identical start time, while those within each row share a uniform duration. Consequently, this two-dimensional boundary-match map encapsulates all potential action segments.

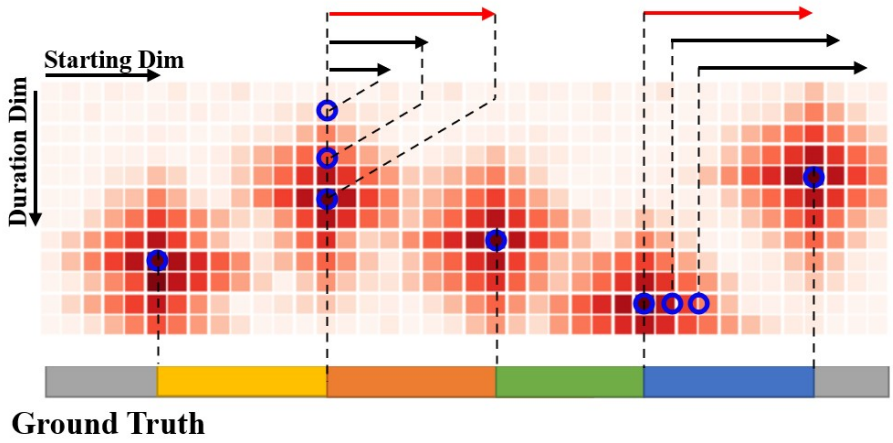

Figure 3. Illustration of BM map. In the provided schematic representation, segments aligned within the same row exhibit uniform duration, while those arranged in the same column share an identical starting time. The segment denoted by the red arrow within the figure is highlighted to represent the one with the highest confidence level. This visual delineation facilitates an intuitive understanding of the temporal structure of the segments and the identification of the segment with maximal prediction confidence.

Within the BM map, the value indicated by each matrix element corresponds to the confidence score of the associated action segment. By generating BM maps, we could produce confidence scores for all potential action segments simultaneously, representing their predicted likelihoods. As demonstrated in Figure 3, we predefined a set of weight matrices $W \in \mathbb{R}^{N \times T}$, each with dimensions $D \times T$, which are dot-multiplied with the transformed video feature sequence $S_F \in \mathbb{R}^{C \times T}$ to yield the segment feature $m_{i,j} \in \mathbb{R}^{C \times N}$. The computation for each weight matrix is governed by the following equation:

$$m_{i,j}(c,n) = \sum_{t=1}^{T} S_F(c,t) \cdot W_{i,j}(n,t), \tag{3}$$

where T denotes the number of clips in the video, D represents the maximum duration of an action segment, C is the channel count of the features, and N is the number of feature sampling points per segment. Through the matrix dot product operation, we can simultaneously generate corresponding features for all potential action segments.

3.2. U-Shaped Temporal Convolutional Module

To enhance the model's capability for action recognition and improve the precision of action segmentation, existing methods often opt to increase model complexity by stacking additional network layers [32]. However, the complexity of models is frequently counterproductive when applied to complex datasets, leading to an exacerbation of the over-segmentation issue as the model complexity increases. Thus, advancing the model's recognition ability while avoiding over-segmentation errors presents a significant challenge in the domain of action segmentation.

The TCN stands out as a formidable model within action segmentation research [13]. It demonstrates the ability to achieve high accuracy without necessitating the stacking of complex network layers. The Encoder–Decoder TCN (ED-TCN) exemplifies the implementation of TCN, and it has been adopted as the baseline for this study. Considering the similarities between image segmentation and action segmentation tasks, the U-Net architecture, which excels in medical image segmentation [33], has been leveraged. To further augment the feature recognition capabilities of the model, we introduce a U-Net-inspired enhancement to the ED-TCN, termed the UTCM. This integration of the encoder–decoder and U-shaped structures aims to attain high segmentation accuracy with a minimal stacking of network layers. Figure 4 delineates the composition of the UTCM, which accepts video frames processed by the backbone network as input, with the standard input dimension represented by $(T, 2048)$, where T signifies the total frame count. The module is comprised of an encoder and a decoder that compress and expand the temporal features of the video, respectively, incorporating skip connections at corresponding depths to enhance contextual comprehension and recognition of nuanced features within the action sequences.

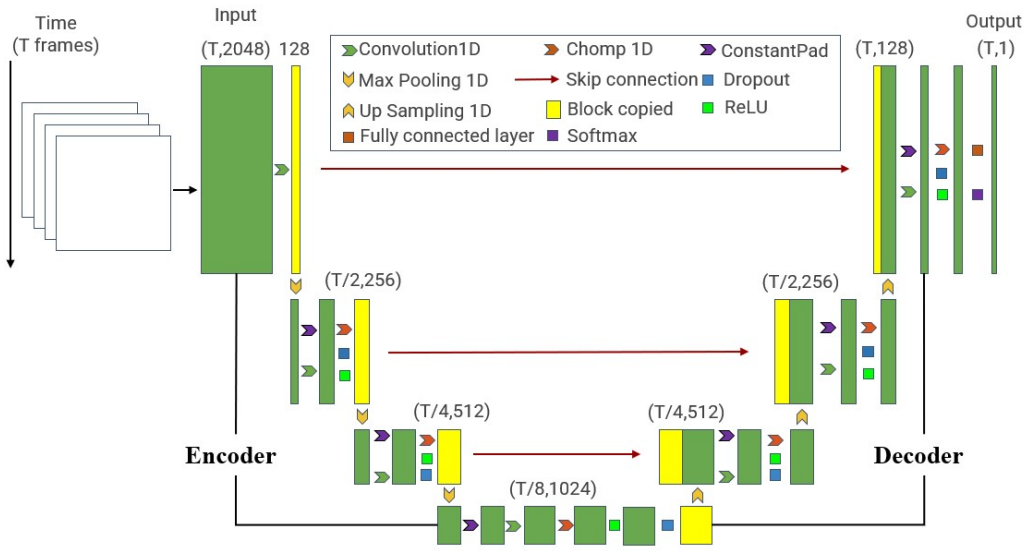

Figure 4. Structure of the U-shaped Temporal Convolutional Module. Each layer of the codec includes a series of operations such as convolutions, alignment procedures, activation functions, and pooling operations. A skip connection is established between the encoder and decoder at corresponding depths.

The encoder employs *Convolution*1*D* and *ConstantPad* for feature processing, followed by the *Chomp* operation for data alignment. The inclusion of *Dropout* and *ReLU* activation functions serves to mitigate model overfitting. Subsequent application of *MaxPooling* reduces the temporal feature dimensions by half. Mirroring the encoder, the decoder's structure incorporates similar layers, where each layer's input is concatenated with the matching encoder layer's output. However, *MaxPooling* operations are substituted with $UpSampling$ to restore the temporal dimensionality. Ultimately, upon expansion of temporal features to dimension T, a *FullyConnected* layer coupled with the $Softmax$ function generates the action segmentation probabilities.

3.3. Loss Function

3.3.1. Loss Function for Boundary-Match Module

To ensure optimal performance of the Boundary-Match Module, we refined the loss function employed in the BMN model [24]. This enhancement involved the integration of multi-class cross-entropy loss with temporal mean square error (TMSE) loss, as described by the following equation:

$$\mathcal{L}_{BMM} = \mathcal{L}_{cls} + \lambda_{BMM}\mathcal{L}_{TMSE}, \tag{4}$$

where λ_{BMM} represents a hyperparameter. It is noteworthy that the labels utilized for our classification loss diverge from the conventional multi-class labeling approach. As depicted in Figure 2, our label corresponds to the IoU value between the predicted and the ground truth segments. For the proposed model, which considers C classes and N latent segments within a video, the classification loss $\mathcal{L}_{cls}^{BMM}$ and the regression loss $\mathcal{L}_{TMSE}$ are defined as follows:

$$\begin{cases} \mathcal{L}_{cls}^{BMM} = -\dfrac{1}{N}\displaystyle\sum_{n=1}^{N}\sum_{c=1}^{C} y_{n,c}\log(p_{n,c}) \\ \mathcal{L}_{TMSE} = \dfrac{1}{N}\displaystyle\sum_{n=1}^{N}\sum_{c=1}^{C}\left(y_{n,c} - p_{n,c}\right)^2 \end{cases}, \tag{5}$$

where $y_{n,c}$ denotes the ground truth IoU value, and $p_{n,c}$ the predicted confidence. To balance the distribution of positive and negative samples within $\mathcal{L}_{TMSE}$, we designated samples with $y_{n,c} > 0.6$ as positive and those with $y_{n,c} < 0.2$ as negative. We aimed to maintain a training ratio of positive to negative samples in close proximity to 1:1.

3.3.2. Loss Function for U-Shaped Temporal Convolutional Module

The UTCM outputs a frame-level prediction sequence for the video. Consequently, we employed the conventional multi-class cross-entropy loss function for model training. Given C, the number of action categories, and T, the total number of frames in the video, the classification loss $\mathcal{L}_{cls}^{UTCM}$ is mathematically expressed as:

$$\mathcal{L}_{cls}^{UTCM} = -\frac{1}{N}\sum_{t=1}^{T}\sum_{c=1}^{C} y_{t,c}\log(p_{t,c}), \tag{6}$$

where $y_{t,c}$ represents the ground truth label for frame t for category c, while $p_{t,c}$ denotes the predicted probability that frame t belongs to category c. Note that the summation is conducted over all frames and all action categories to compute the loss, which encourages the model to make accurate frame-level predictions for each action category.

3.4. Refining Action Segmentation Prediction

Since the introduction of ASRF [29], which proposed post-processing model outputs to mitigate over-segmentation errors, researchers have intensified their efforts to refine video sequence predictions. Several methods leverage basic boundary information of

predicted actions to enhance model precision [28,29]. In contrast, our approach utilizes the informative BM map as an auxiliary instrument for the refinement of predictions.

As depicted in Figure 5, initial predictions were generated using the UTCM as detailed in Section 3.2. These predictions were subsequently refined with the boundary-match map produced by the BM Module, as discussed in Section 3.1. To further enhance performance, we propose the use of AIBS to suppress the fragmentation within the predictions, thereby improving the continuity and coherence of the segmented actions.

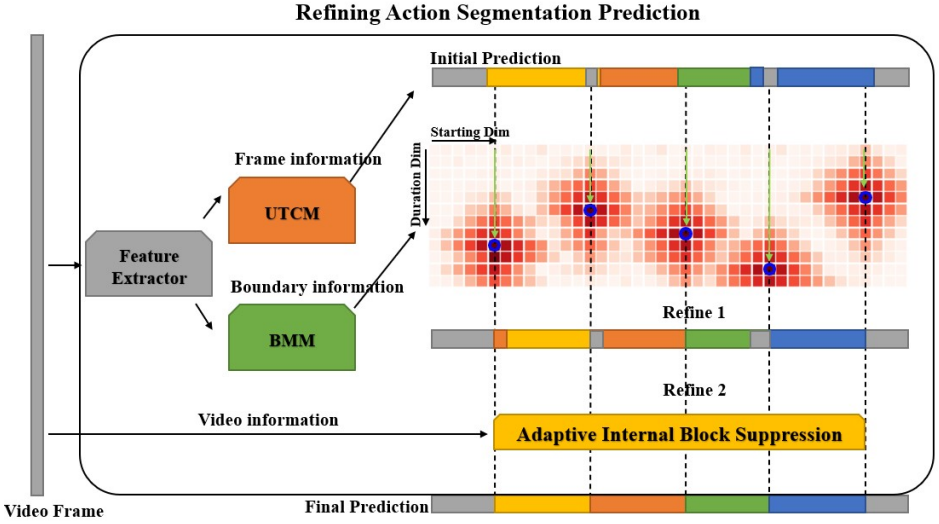

Figure 5. Illustration of the action segmentation prediction refinement process. During the refinement phase, we execute two procedures: boundary refinement, which employs the BM map from the BM Module to adjust frames near the initial predicted action boundaries, potentially introducing additional over-segmentation errors; and AIBS, which performs unsupervised refinement of the predictions from the previous step.

3.4.1. Boundary Refine

The BM map encapsulates the predicted probabilities and associated classes for all potential action segments within a video. In Section 3.3, we employed two distinct loss functions to derive the classification confidence S_{cc} and regression confidence S_{cr}. To obtain a more dependable score, these confidences are amalgamated to yield a fused score $S_f \in \mathbb{R}^{d \times t \times K}$, where d indicates the action's duration, t represents the action's starting point, and both are measured in terms of interpolated frame stacks.

To fully exploit the boundary information encapsulated in the BM map, we transform it into a confidence tensor $P_B \in \mathbb{R}^{T \times K}$, where T is the aggregate number of video frames and K is the count of action categories. The conversion process is formalized by the following equation:

$$\begin{cases} P_B[T_s\ T_e] = max(S_f[t_e - t_s, t_s]) \\ T_e = T_s + T \times \dfrac{t_e - t_s}{t}, \end{cases} \tag{7}$$

where t_e denotes the endpoint of an action that corresponds to the maximum confidence in S_f originating from t_s. The initial action prediction output by UTCM is denoted as $P_C \in \mathbb{R}^{T \times K}$. We integrate the corresponding confidences from both P_B and P_C sequences to compute P_{BR}:

$$P_{BR} = \sqrt{P_B P_C}. \tag{8}$$

Subsequently, we determine the predicted category for each frame by selecting the maximum category confidence value from P_{BR} for that frame, resulting in a final prediction $P_F \in \mathbb{R}^T$.

3.4.2. Adaptive Internal Block Suppression

Following the refinement step, the action boundaries within the predicted sequence are rendered more distinct. Nevertheless, the fusion of temporal features may still perturb the continuity of the actions, and over-segmentation errors can persist. Previous research has attempted to mitigate these errors using loss functions, yet the outcomes have been suboptimal [30].

Drawing inspiration from the Enhanced Temporal Segment Networks (ETSN), our study posited that over-segmentation errors predominantly arise from the excessive identification of local segments. By integrating the concept of non-maximum suppression (NMS), we employed unsupervised adaptive internal block suppression to eliminate superfluous short segments within the prediction sequence. This method also allowed for the adaptive determination of hyperparameters based on the input video information. As shown in Figure 6, short segments whose length fell below a predetermined threshold were merged with an adjacent longer segment—for example, combining segments "22" and "6". Subsequently, brief segments with low confidence were allocated to the nearest segment with high confidence, as exemplified by "1111". Should a short segment be flanked by high-confidence segments on both sides, it was apportioned proportionally, as demonstrated by "111". Through the execution of several unsupervised refinements on the preliminary prediction sequence, there was a marked diminution in over-segmentation errors. This process significantly enhanced the precision of the action segmentation. The procedural steps are delineated in Algorithm 1.

Algorithm 1 Adaptive Internal Block Suppression

Input: Action maximum length $A_{\max}$, minimum length $A_{\min}$; Initial prediction P_F, Pre-set window size $L_i(A_{\max}, A_{\min}, P_F) \in R$, $i = 1, \ldots, n$; Pre-set confidence threshold $\varphi_i(A_{\max}, A_{\min}, P_F) \in R$, $i = 1, \ldots, n$

Output: Refined prediction P_R

Remove the short action segments
Count the length (n) of each action
for $i = 1, \ldots, n$ **do**
    if $(n < L_i)$ and $(\text{Mean}(P_F) < \varphi_i)$ **then**
        Locate the short segment
        if Even-numbered continuous segments **then**
            Perform the even-numbered continuous segments process
        else if Odd-numbered continuous segments **then**
            Ignore the middle segment and perform the even-numbered continuous segments process
        end if
        if Isolated segment **then**
            Perform the isolated segments process
        end if
    end if
end for

where $A_{\max}$ represents the maximum allowable action length, $A_{\min}$ denotes the minimum allowable action length, P_F refers to the initial prediction output, L_i indicates the pre-defined window size, and φ_i is the adaptive confidence threshold employed in the refinement process.

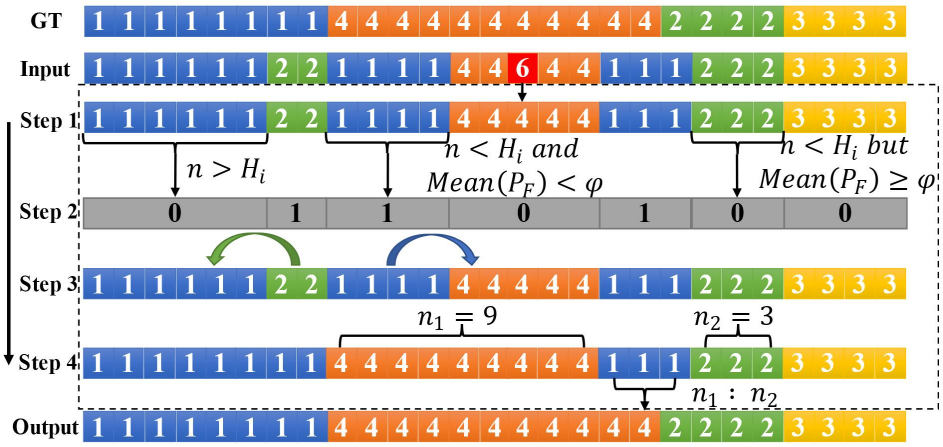

Figure 6. Illustration of the adaptive internal block suppression process.segments characterized by their brevity and low confidence levels are either allocated to the nearest segment with high confidence or divided proportionally between the adjacent high-confidence segments on either side.

4. Experiments

4.1. Datasets

Our method was evaluated on four challenging action segmentation datasets: 50Salads [15], GTEA [16], and the Vulgar datasets. A summary of the overall statistical metrics for these datasets is presented in Table 1 where "Videos" refers to the total count of videos in the dataset, "Classes" denotes the distinct number of action categories, "Instances" represents the average number of action instances per video, and "Cross-VAL" indicates the number of folds employed in cross-validation.

Table 1. The statistics of action segmentation datasets.

	50Salads	GTEA	Vulgar
Videos	50	28	264
Classes	17	11	4
Instances	20	20	13
Cross-val	5	4	5

Owing to the sensitive nature of vulgar content, there are no publicly available datasets for vulgar actions. To address this gap, we curated a dataset comprising 264 videos sourced from public video platforms through web crawling tools, with the collection process being supervised by domain experts. This dataset encompasses three types of vulgar actions along with non-vulgar background activities. Each video contains approximately 6000 to 7000 frames and, on average, 13 instances of action. The distribution of the various vulgar actions within the dataset is detailed in Table 2, which presents the size and proportion of each action class within the vulgar dataset. During the data collection process, we imposed criteria regarding the gender, clothing, body size, and appearance of individuals featured in the videos to ensure a diverse representation. Representative samples from the vulgar dataset are depicted in Figure 7.

Table 2. The statistics for the vulgar action dataset.

Category	Size	Counts	Proportion
Normal	555 MB	645	43.11%
Caress breast	208 MB	325	21.72%
Hip shake	400 MB	395	26.40%
Suck	82 MB	131	8.75%

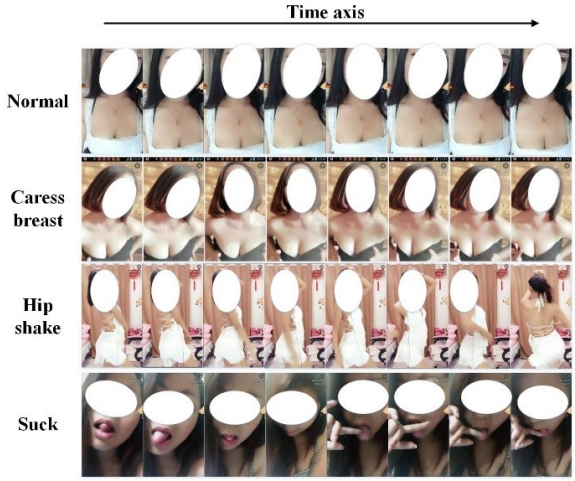

Figure 7. Sample frames from the vulgar action dataset.

4.2. Evaluation Metrics

We adopted three metrics to evaluate action segmentation performance, as described in [29]: frame-wise accuracy (Acc), segmental edit score (Edit), and segmental F1 score at IoU thresholds of 0.10, 0.25, and 0.50 (F1@10,25,50). Frame-wise accuracy assessed the classification capabilities of the model, while both the edit score and F1 score were utilized to quantify the error due to over-segmentation. Edit scores were computed using the Levenshtein Distance [34], with the calculation formula provided as follows:

$$S_{edit}(G, P) = (1 - D(G, P)/\max(M, N)) \times 100, \tag{9}$$

where $D(G, P)$ denotes the dissimilarity between the ground truth sequence $G = \{G_1, \ldots, G_M\}$ and the predicted sequence $P = \{P_1, \ldots, P_N\}$.

The computation of the F1 score hinges on the precision and recall metrics derived from frame-wise classification within the video. The mathematical expressions for these calculations are as follows:

$$\begin{cases} F1 = 2 * \frac{Precision * Recall}{Precision + Recall}, \\ Precision = \frac{TP}{TP+FP}, \\ Recall = \frac{TP}{TP+FN}, \end{cases} \tag{10}$$

The calculation process of TP, FP, and FN are illustrated in Figure 8. In this approach, the IoU for each ground truth segment $G = \{G_1, \ldots, G_5\}$ and predicted segment $P = \{P_1, \ldots, P_n\}$ was computed to determine whether they are classified as TP or FP.

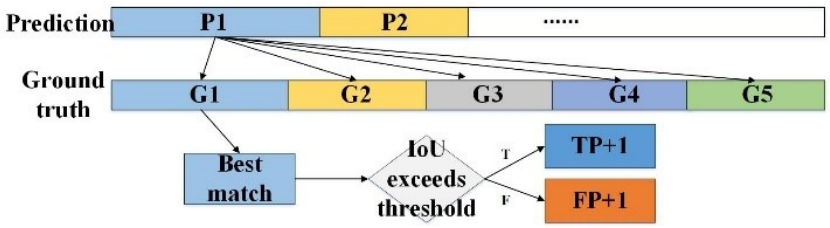

Figure 8. Calculation of TP and FP values based on IoU. 'P' represents the action prediction, while 'G' denotes the ground-truth action label.

4.3. Experimental Settings

Our experimental setup utilized the PyTorch framework and was executed on a machine equipped with a GeForce RTX 3090 graphics card, boasting 24 GB of memory and

CUDA acceleration capabilities. The datasets employed in our research were partitioned following an 80:20 ratio for the training and testing subsets, respectively. For feature extraction from video data, we relied on a ResNet architecture [35] as our feature extraction backbone.

In the implementation of the UTCM, the channel configurations of the encoder were specified as {256, 512, 1024}. Within the BM Module, the number of sampling points, denoted as N, was set to 32. Both the maximum duration of an action, represented by D, and the temporal length of the BM map, denoted by T, were configured to align with the unique attributes of each dataset. Specifically, for the 50Salads dataset, T was set to 400 and D to 70. In the case of the GTEA dataset and the bespoke vulgar dataset, these parameters were adjusted to $T = 300$ and $D = 50$, respectively.

We adopted a mini-batch size of 8 for our training procedure. The initial learning rate was established at 1×10^{-4}, and we applied the MultiStepLR scheduling strategy for dynamic learning rate adjustments. The AdamW optimizer was employed for parameter optimization across all models. The training process spanned a total of 200 epochs, with an average training duration of approximately 2.5 h for each model.

4.4. Comparison with State-of-the-Art

In this subsection, we assess the performance of our proposed BMUTCN by juxtaposing it with other cutting-edge action segmentation methodologies. Our selection of benchmark action segmentation methods encompasses Spatial-CNN [36], Bi-LSTM [37], Dilated TCN [13], ED-TCN [13], BCN [28], ETSN [14], and ASRF [29], all of which have demonstrated impressive experimental outcomes on public datasets.

The UTCM in our enhanced model signifies the U-shaped Temporal Convolutional Module bereft of the refinement step. Subsequent model iterations are augmented based on this network. BMM denotes the model utilizing the Boundary-Match Module for boundary refinement, and AIBS stands for the model applying adaptive internal block suppression to refine initial predictions. The BMUTCN integrates all the aforementioned improvements. We executed experiments across three challenging datasets, and the results are delineated in Table 3. Our findings indicate that BMUTCN effectively amalgamates the strengths of the BMM and AIBS approaches, not only elevating accuracy but also substantially enhancing the F1 score and Edit distance. On the 50salads dataset, BMUTCN boosts the frame-level accuracy by 1.7%; for the GTEA dataset, our method achieves a frame-level accuracy enhancement of 2.5%. On the Vulgar dataset, BMUTCN secures state-of-the-art results, with improvements of 6.6% and 3.4% in F1@10 and Edit distances, respectively, alongside a 5.9% increase in frame-level accuracy. In contrast to the 50Salads and GTEA datasets, the Vulgar dataset poses distinct challenges that include substantial variation in the duration of vulgar behaviors, limited discriminability of image features for such actions, complex features associated with vulgar behavior, and ambiguous boundaries between actions. The samples in the 50Salads and GTEA datasets are characterized by brief action durations and a high frequency of transitions, which starkly contrast with the traits observed in the Vulgar dataset. Consequently, these datasets are not fully representative of the specific challenges our algorithm is designed to address, leading to diminished performance of our algorithm when applied to the GTEA dataset. While the results on commonly used action segmentation benchmarks like 50Salads and GTEA may not achieve state-of-the-art status—given that these public datasets contain a wide variety of action types and detailed annotations—the experimental outcomes on the Vulgar dataset robustly demonstrate the reliability and effectiveness of the BMUTCN, particularly in the context of long-duration and complex action segmentation.

For a more intuitive elucidation of the role each module plays within BMUTCN, we have visualized the experimental outcomes. Figure 9 displays the segmentation results derived from our method on the three datasets. The various colored bars within the figure correspond to different actions in the video sequence. It is evident that the prediction sequence generated solely by the UTCM method contains numerous internal fragments, leading to an excess of over-segmentation errors, thereby resulting in lower F1 scores and

Edit distances in the evaluation metrics. Post-implementation of the BMM approach, the output predictions for ambiguous frames near action boundaries become more discernible, and frame-level predictive accuracy improves, albeit with some residual over-segmentation errors. When coupled with the AIBS module, there is an effective reduction in internal fragmentation of the prediction sequence; however, boundary prediction accuracy sees only marginal enhancement. Ultimately, BMUTCN, through the synergistic integration of BMM and AIBS, efficiently leverages the advantages of these modules to achieve high-quality video segmentation.

Table 3. Comparing the performance of common action detection methods.

Dataset	**50salads**					**GTEA**					**Vulgar**				
Method	**F1@{10,25,50}**			**Edit**	**Acc**	**F1@{10,25,50}**			**Edit**	**Acc**	**F1@{10,25,50}**			**Edit**	**Acc**
Spatial-CNN [36]	32.3	27.1	18.9	24.8	54.9	41.8	36.0	25.1	–	54.1	28.6	23.2	14.3	22.7	53.2
Bi-LSTM [37]	62.6	58.3	47.0	55.6	55.7	66.5	59.0	43.6	–	55.5	52.1	47.6	37.4	39.2	54.6
Dilated TCN [13]	52.2	47.6	37.4	43.1	59.3	58.8	52.2	42.2	–	58.3	46.3	41.3	32.7	39.8	58.7
ED-TCN [13]	68.0	63.9	52.6	59.8	64.7	72.2	69.3	56.0	–	64.0	53.6	50.2	43.8	41.6	68.5
BCN [28]	82.3	81.3	74.0	74.3	84.4	88.5	87.1	77.3	84.4	79.8	71.2	69.3	62.5	68.9	77.6
ETSN [14]	**85.2**	**83.9**	75.4	78.8	82.0	**91.1**	**90.0**	77.9	**86.2**	78.2	75.6	74.3	66.4	68.8	75.3
ASRF [29]	84.9	83.5	**77.3**	**79.3**	84.5	89.4	87.8	**79.8**	83.7	77.3	73.1	71.5	65.6	69.4	79.7
UTCM	59.3	57.9	47.7	49.6	76.3	78.2	74.8	58.7	70.6	76.4	45.4	43.2	33.8	31.1	80.1
UTCM + BMM	68.9	66.5	58.1	55.7	84.2	86.8	82.6	67.2	83.2	78.9	53.2	51.1	42.7	37.6	84.5
UTCM + AIBS	78.6	75.3	61.2	72.8	81.6	85.3	81.4	61.1	80.3	77.2	70.5	69.5	64.5	68.3	81.9
BMUTCN	79.8	76.7	63.1	73.3	**86.2**	84.6	82.4	76.4	80.8	**82.4**	**82.2**	**81.6**	**72.9**	**72.8**	**85.6**

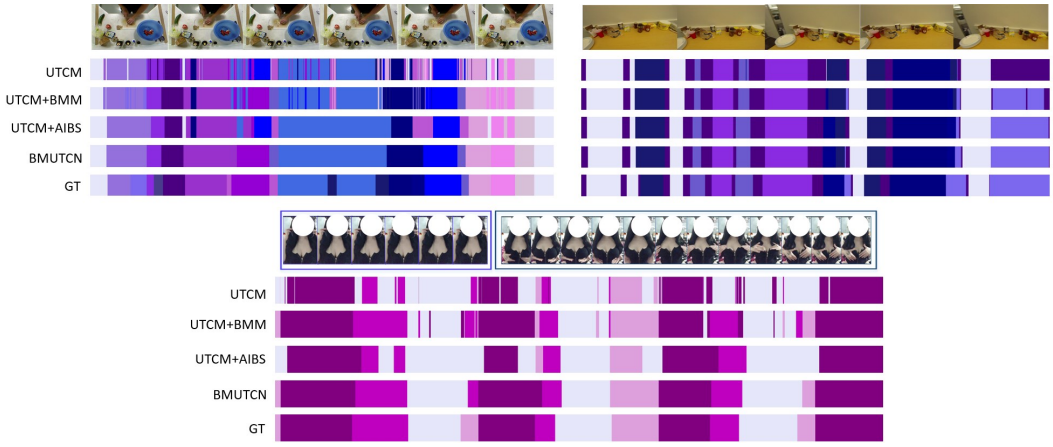

Figure 9. Visualization of segmentation results on the datasets. The bars in the diagram represent all video frames, with different colors indicating the model's predictions for each frame.

4.5. Ablation Studies

In this section, we explore the impact of hyperparameters on the performance of the UTCM and the BMM.

4.5.1. Impact of Codec Structure for U-Shaped Temporal Convolutional Module

The recognition capabilities of the UTCM principally arise from its codec structure. To this end, we designed several groups of encoder structures for our ablation studies, with corresponding decoders arranged in the reverse order of their encoders.

The experimental results, presented in Table 4, reveal that encoder structures of insufficient depth lack the feature extraction prowess necessary for high accuracy. Conversely, overly complex encoder structures can induce overfitting and over-segmentation errors, resulting in reduced frame-level accuracy and F1 scores. Optimal results were ultimately obtained using a codec structure consisting of [128, 256, 512, 1024].

Table 4. Impact of different encoder structures on the vulgar dataset.

Encoder Structures	F1@{10,25,50}			Edit	Acc
128, 256	70.2	68.3	62.5	66.5	75.6
128, 256, 512	76.7	73.9	69.2	70.2	82.4
128, 256, 512, 1024	**82.2**	**81.6**	**72.9**	**72.8**	**85.6**
128, 256, 512, 1024, 2048	74.8	71.3	66.8	67.9	80.3

4.5.2. Impact of T and D for Boundary-Match Module

The primary purpose of the Boundary-Match Module is to extract action boundary information from videos and construct a BM map. The dimensions of the BM map, denoted as $m \in \mathbb{R}^{D \times T}$, are governed by two critical parameters: D and T. These parameters significantly influence the precision with which action boundaries are identified. This section evaluates the effects of varying the hyperparameters T and D on the action segmentation task.

Empirical results, as presented in Table 5, indicate that the parameter T correlates with the temporal sampling resolution of the video. The enhancement effect attributable to increased T diminishes when T exceeds 300. In contrast, the parameter D is indicative of the maximal extent of potential action durations within a video sequence. A D value greater than 50 leads to a decrease in segmentation accuracy. This decline is presumably due to the model's diminished ability to accurately recognize shorter actions within the context of longer durations. Balancing model complexity and training efficiency, we opt for a configuration of $(T = 300, D = 50)$ as our optimal parameter setting.

Table 5. Impact of T and D on the vulgar dataset.

Impact of T	F1@{10,25,50}			Edit	Acc
BMUTCN ($T = 200$, $D = 50$)	79.7	77.2	70.5	73.3	83.2
BMUTCN ($T = 300$, $D = 50$)	**82.2**	**81.6**	**72.9**	**72.8**	**85.6**
BMUTCN ($T = 400$, $D = 50$)	81.4	79.3	72.5	72.5	84.9
BMUTCN ($T = 500$, $D = 50$)	82.5	81.2	72.8	73.0	85.7
Impact of D	**F1@{10,25,50}**			**Edit**	**Acc**
BMUTCN ($T = 300$, $D = 30$)	80.1	78.5	71.6	70.5	82.9
BMUTCN ($T = 300$, $D = 40$)	81.3	79.8	72.1	70.9	83.6
BMUTCN ($T = 300$, $D = 50$)	**82.2**	**81.6**	**72.9**	**72.8**	**85.6**
BMUTCN ($T = 300$,$D = 60$)	81.8	78.9	72.5	71.9	80.3

5. Conclusions

This article introduces a novel framework termed the BMUTCN, designed specifically for the segmentation of vulgar actions in video content. The BMUTCN framework incorporates a boundary-match map to enhance the precision of frame classification in the vicinity of action boundaries, which are often ambiguous. Utilizing a U-shaped encoder–decoder architecture within the temporal convolutional network, BMUTCN proficiently localizes and categorizes vulgar actions by leveraging contextual action information.

To mitigate the issue of over-segmentation errors in predictions, we introduce an Adaptive Internal Block Suppression technique inspired by Local Burr Suppression. This approach operates in an unsupervised manner and has been demonstrated to significantly elevate the F1 score, thereby enhancing model accuracy. Extensive evaluations of our model across various public datasets have established that our framework delivers state-of-the-art performance on datasets featuring vulgar content. The segmentation of vulgar actions holds considerable significance for digital environments such as online social networks, video-sharing platforms, content moderation systems, and emergent contexts like metaverse applications. Through the automatic detection and elimination of content deemed inappropriate or offensive, it is possible to preserve the integrity of both traditional platforms and the metaverse. This proactive approach not only alleviates the onerous

task of manual content moderation but also serves to augment the user experience while simultaneously elevating the quality and security of the content available. Looking ahead, the sensitive nature of vulgar video datasets presents substantial challenges in amassing large-scale collections. In addition, the task of manually labeling such datasets is remarkably resource-intensive. Consequently, our future research endeavors will be directed toward the development of weakly-supervised techniques for vulgar action localization.

To adhere to the principles of reproducibility and transparency in the scientific community, we affirm that all materials, data, computer code, and protocols pertinent to this publication shall be made accessible to readers. Any potential restrictions on the availability of these materials must be disclosed at the time of manuscript submission. It is imperative that new methods and protocols are described in sufficient detail to enable replication and extension of the published results, while established methods should be succinctly presented with appropriate citations.

Author Contributions: Conceptualization, Z.S.; methodology, R.X.; software, Z.S. and Y.Z.; validation, R.G. and C.W.; resources, F.Q.; data curation, M.T.; writing—original draft preparation, Z.S.; writing—review and editing, F.Q.; visualization, R.X.; supervision, F.Q.; project administration, C.W; funding acquisition, F.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data are contained within the present article.

Acknowledgments: This work was supported by the Natural Science Foundation of Zhejiang Province (No. LY21F020017), GuangDong Basic and Applied Basic Research Foundation (No. 2022A1515110570), Innovation Teams of Youth Innovation in Science and Technology of High Education Institutions of Shandong Province (No. 2021KJ088), Shenzhen Science and Technology Program (No.KCXFZ20201221173008022), Key Research & Development Program of Zhejiang Province, China (No. 2019C03127) and the Open Project Program of the State Key Laboratory of CAD&CG (No. A2304).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Owens, E.W.; Behun, R.J.; Manning, J.C.; Reid, R.C. The impact of Internet pornography on adolescents: A review of the research. *Sex. Addict. Compuls.* **2012**, *19*, 99–122. [CrossRef]
2. Vitorino, P.; Avila, S.; Perez, M.; Rocha, A. Leveraging deep neural networks to fight child pornography in the age of social media. *J. Vis. Commun. Image Represent.* **2018**, *50*, 303–313. [CrossRef]
3. Papadamou, K.; Papasavva, A.; Zannettou, S.; Blackburn, J.; Kourtellis, N.; Leontiadis, I.; Stringhini, G.; Sirivianos, M. Disturbed YouTube for kids: Characterizing and detecting inappropriate videos targeting young children. In Proceedings of the International AAAI Conference on Web and Social Media, Atlanta, GA, USA, 8–11 June 2020; Volume 14, pp. 522–533.
4. Du, H.; Shi, H.; Zeng, D.; Zhang, X.P.; Mei, T. The elements of end-to-end deep face recognition: A survey of recent advances. *Acm Comput. Surv. (CSUR)* **2022**, *54*, 1–42. [CrossRef]
5. Moustafa, M. Applying deep learning to classify pornographic images and videos. *arXiv* **2015**, arXiv:1511.08899.
6. Caetano, C.; Avila, S.; Schwartz, W.R.; Guimarães, S.J.F.; Araújo, A.d.A. A mid-level video representation based on binary descriptors: A case study for pornography detection. *Neurocomputing* **2016**, *213*, 102–114. [CrossRef]
7. Mei, M.; He, F. Multi-label learning based target detecting from multi-frame data. *IET Image Process.* **2021**, *15*, 3638–3644. [CrossRef]
8. Zeng, D.; Chen, S.; Chen, B.; Li, S. Improving remote sensing scene classification by integrating global-context and local-object features. *Remote. Sens.* **2018**, *10*, 734. [CrossRef]
9. Ge, S.; Li, C.; Zhao, S.; Zeng, D. Occluded face recognition in the wild by identity-diversity inpainting. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 3387–3397. [CrossRef]
10. Perez, M.; Avila, S.; Moreira, D.; Moraes, D.; Testoni, V.; Valle, E.; Goldenstein, S.; Rocha, A. Video pornography detection through deep learning techniques and motion information. *Neurocomputing* **2017**, *230*, 279–293. [CrossRef]
11. Arif, M. A systematic review of machine learning algorithms in cyberbullying detection: Future directions and challenges. *J. Inf. Secur. Cybercrimes Res.* **2021**, *4*, 01–26. [CrossRef]
12. Cao, J.; Xu, R.; Lin, X.; Qin, F.; Peng, Y.; Shao, Y. Adaptive receptive field U-shaped temporal convolutional network for vulgar action segmentation. *Neural Comput. Appl.* **2023**, *35*, 9593–9606. [CrossRef]
13. Lea, C.; Flynn, M.D.; Vidal, R.; Reiter, A.; Hager, G.D. Temporal convolutional networks for action segmentation and detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 156–165.

14. Li, Y.; Dong, Z.; Liu, K.; Feng, L.; Hu, L.; Zhu, J.; Xu, L.; Liu, S. Efficient two-step networks for temporal action segmentation. *Neurocomputing* **2021**, *454*, 373–381. [CrossRef]
15. Stein, S.; McKenna, S.J. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, Zurich, Switzerland, 8–12 September 2013; pp. 729–738.
16. Li, Y.; Ye, Z.; Rehg, J.M. Delving into egocentric actions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 287–295.
17. Wehrmann, J.; Simões, G.S.; Barros, R.C.; Cavalcante, V.F. Adult content detection in videos with convolutional and recurrent neural networks. *Neurocomputing* **2018**, *272*, 432–438. [CrossRef]
18. Mallmann, J.; Santin, A.O.; Viegas, E.K.; dos Santos, R.R.; Geremias, J. PPCensor: Architecture for real-time pornography detection in video streaming. *Future Gener. Comput. Syst.* **2020**, *112*, 945–955. [CrossRef]
19. Song, K.H.; Kim, Y.S. Pornographic video detection scheme using multimodal features. *J. Eng. Appl. Sci.* **2018**, *13*, 1174–1182.
20. Gao, J.; Chen, K.; Nevatia, R. Ctap: Complementary temporal action proposal generation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 68–83.
21. Lin, T.; Zhao, X.; Su, H.; Wang, C.; Yang, M. Bsn: Boundary sensitive network for temporal action proposal generation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
22. Yu, H.; He, F.; Pan, Y. A scalable region-based level set method using adaptive bilateral filter for noisy image segmentation. *Multimed. Tools Appl.* **2020**, *79*, 5743–5765. [CrossRef]
23. Liu, Y.; Ma, L.; Zhang, Y.; Liu, W.; Chang, S.F. Multi-granularity generator for temporal action proposal. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3604–3613.
24. Lin, T.; Liu, X.; Li, X.; Ding, E.; Wen, S. Bmn: Boundary-matching network for temporal action proposal generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3889–3898.
25. Yu, H.; He, F.; Pan, Y. A novel segmentation model for medical images with intensity inhomogeneity based on adaptive perturbation. *Multimed. Tools Appl.* **2019**, *78*, 11779–11798. [CrossRef]
26. Alwassel, H.; Giancola, S.; Ghanem, B. Tsp: Temporally-sensitive pretraining of video encoders for localization tasks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 3173–3183.
27. Farha, Y.A.; Gall, J. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3575–3584.
28. Wang, Z.; Gao, Z.; Wang, L.; Li, Z.; Wu, G. Boundary-aware cascade networks for temporal action segmentation. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 34–51.
29. Ishikawa, Y.; Kasai, S.; Aoki, Y.; Kataoka, H. Alleviating over-segmentation errors by detecting action boundaries. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 2322–2331.
30. Singhania, D.; Rahaman, R.; Yao, A. Coarse to fine multi-resolution temporal convolutional network. *arXiv* **2021**, arXiv:2105.10859.
31. Park, J.; Kim, D.; Huh, S.; Jo, S. Maximization and restoration: Action segmentation through dilation passing and temporal reconstruction. *Pattern Recognit.* **2022**, *129*, 108764. [CrossRef]
32. Ahn, H.; Lee, D. Refining action segmentation with hierarchical video representations. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 16302–16310.
33. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
34. Levenshtein, V.I. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys. Dokl.* **1966**, *10*, 707–710.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
36. Lea, C.; Reiter, A.; Vidal, R.; Hager, G.D. Efficient segmental inference for spatiotemporal modeling of fine-grained actions. *arXiv* **2016**, arXiv:1602.02995.
37. Singh, B.; Marks, T.K.; Jones, M.; Tuzel, O.; Shao, M. A multi-stream bi-directional recurrent neural network for fine-grained action detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1961–1970.