OS-MSL: One Stage Multimodal Sequential Link Framework for Scene Segmentation and Classification

Ye Liu∗
Tencent YouTu Lab
Shanghai, China
rafelliu@tencent.com

Lingfeng Qiao∗
Tencent YouTu Lab
Shanghai, China
leafqiao@tencent.com

Di Yin
Tencent YouTu Lab
Shanghai, China
endymecyyin@tencent.com

Zhuoxuan Jiang
Tencent YouTu Lab
Shanghai, China
jzlx1211@gmail.com

Xinghua Jiang
Tencent YouTu Lab
Hefei, China
clarkjiang@tencent.com

Deqiang Jiang
Tencent YouTu Lab
Hefei, China
dqiangjiang@tencent.com

Bo Ren
Tencent YouTu Lab
Hefei, China
timren@tencent.com

ABSTRACT
Scene segmentation and classification (SSC) serve as a critical step towards the field of video structuring analysis. Intuitively, jointly learning of these two tasks can promote each other by sharing common information. However, scene segmentation concerns more on the local difference between adjacent shots while classification needs the global representation of scene segments, which probably leads to the model dominated by one of the two tasks in the training phase. In this paper, from an alternate perspective to overcome the above challenges, we unite these two tasks into one task by a new form of predicting shots link: a link connects two adjacent shots, indicating that they belong to the same scene or category. To the end, we propose a general One Stage Multimodal Sequential Link Framework (OS-MSL) to both distinguish and leverage the two-fold semantics by reforming the two learning tasks into a unified one. Furthermore, we tailor a specific module called DiffCorrNet to explicitly extract the information of differences and correlations among shots. Extensive experiments on a brand-new large scale dataset collected from real-world applications, and MovieScenes are conducted. Both the results demonstrate the effectiveness of our proposed method against strong baselines. The code is made available.1

CCS CONCEPTS
• Computing methodologies → Computer vision tasks.

1Both authors contributed equally to this research.

1https://github.com/TencentYouTuResearch/VideoStructuring-OSMSL

KEYWORDS
Video structuring, Scene segmentation and classification, Sequence model, Multimodal embeddings

ACM Reference Format:
Ye Liu, Lingfeng Qiao, Di Yin, Zhuoxuan Jiang, Xinghua Jiang, Deqiang Jiang, and Bo Ren. 2022. OS-MSL: One Stage Multimodal Sequential Link Framework for Scene Segmentation and Classification. In Proceedings of the 30th ACM International Conference on Multimedia (MM ’22), October 10–14, 2022, Lisboa, Portugal. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3503161.3547839

1 INTRODUCTION

With massive video data generated everyday, video structuring and understanding are significant work for media resource management. For example in a video clips search engine, long videos need to be pre-processed and segmented into different topics for later fast retrieval. Therefore, long video segmentation and topic classification are fundamental techniques for various senior media applications, which, however, is still less-studied in the community.

As known, usually we have two types of video segments: shot and scene for different scales. Shot is a set of successive frames taken without interruption by a single camera [3], and Scene is the combination of adjacent shots which are related with similar background or environment. As illustrated in Figure 1, the first row shows that we can segment six shots from a long video where each shot is composed of frames shot by the same camera. The second row shows the partition of similar shots which constitute scenes, e.g. studio and interview. The red lines represent the segment boundaries. To this end, this paper conducts a pioneer study to investigate the problem that how to segment long videos and classify each segment in a joint manner, and we name it as Scene Segmentation and Classification (SSC).

As mentioned above, video segmentation at the shot level is a relatively simple task, and adjacent frames can be easily identified to different shots or the same due to the distinct low-level visual features [8, 27]. However, the scene segmentation (SS) is pretty tough
We observe that simply building a multi-task learning model could be intuitively, SS focuses more on the local difference between two segments is also a crucial step for video understanding. In our framework, segmentation and classification tasks are unified into one-stage task by predicting shots link.

To overcome the above challenges, we tackle the SS and SSC problems in a new perspective. We propose to decompose long videos into two elements, namely shot and link. A link connects a pair of adjacent shots, indicating that they belong to the same scene. Under the above definitions, as illustrated in Figure 3, the task of SS is reformed to predict whether there is a link between current shot and next shot. In detail, by enumerate all the possible links, we have a 5-class link label set: \{B->I, I->I, I->E, B->E, N\}, where B is Beginning, I is Intermediate, E is End, and N means there is no link between these two shots.

SSC is an extension of SS, whose label set is denoted as \(\mathcal{C} \times \{B->I, I->I, I->E, B->E, N\}\), where \(\mathcal{C}\) is the number of classes. Next, multimodal information, visual and audio, are leveraged to extract powerful representation of shots [29]. It should be pointed out that text modality is not used because the linguistic contents in video are little related to SSC tasks. More discussions about these modalities could refer to the Experiments section. Then, to better model the difference and correlation between shots which is much important for the SSC, a tailored network (called DiffCorrNet) is proposed for particularly extracting shot features. At last, a transformer structure and a conditional random fields (CRF) module are utilized to make predictions on link tags. We name the proposed model with One Stage Multimodal Sequential Link Framework (OS-MSL).

Our method provides a general framework for other similar video sequential modeling problems, e.g., action, activity or event recognition, which can be categorized to a unified definition: find a boundary between two consecutive segments/clips with auxiliary information such as label, key frame, and caption.

In addition, we contribute a new large-scale dataset called TI-News for scene understanding community. Existing related datasets, such as MovieScenes [18], only focus on movie domain and they solely offer the SS labels. To expand the diversity of video domains and broaden the application scenarios, we collected real world news videos with sufficient SSC annotations. TI-News can enrich the dataset resources in scene understanding community and can promote the researches and applications in this area.

We summarize the main contributions as follows:

- We provide an alternate perspective to define the SS and SSC problems to a linking form, and propose one stage multimodal sequential link framework, which can unify multiple tasks in SSC into one task by predicting shot links.
- We propose a feature extraction network DiffCorrNet to simultaneously extract the difference and correlation between adjacent shots.
• We construct a new dataset TI-News which include hundreds of news videos with both segment and category labels.
• Extensive experiments on two datasets TI-News and MovieScenes demonstrate the effectiveness of our method and achieve state-of-the-art results.

2 RELATED WORK

2.1 Video Segmentation

Traditionally, graph based methods are widely proposed to group shots where each shot is regarded as node in a graph and segmentation is implemented by graph cut algorithms [19, 25, 36]. Later graph convolution network is applied to this line [17]. On the other hand, similarity measurement, such as siamese network [2], and shot clustering methods are proposed for video segmentation [20, 22, 23]. Recently, with a large-scale dataset released [12], end-to-end models are designed to detect boundaries directly. LGSS [18] raises B-Net to identify boundaries and uses multimodal features, and ShotCoL [7] leverages contrastive learning. However, their work performed a 2-classes classification which refers to whether the shot should be segmented or not, ignoring richer relation information between shots within a scene. Different from their work, we perform SS task to a link form, and explicitly reveal the sequential relation between shots under a 5-classes richer link labels, which can promote the SS task.

2.2 Video Classification

Existing video classification works are mostly for single task and usually applied to short videos or video snippets segmented from long videos, such as hand-crafted feature engineering method [33], deep learning based Two-stream CNNs [26, 34], 3D-CNN [31], and vision transformer based method [1]. As to in a joint task for long videos, existing classification works may suffer from a decreased performance of the preceding segmentation in a pipeline way. This paper tries to alleviate this issue in an ingenious way.

2.3 Video Segmentation and Classification

It is easily confusing that our SSC task is similar to the Temporal Action Detection (TAD) task, because both conduct segmentation and classification on long videos. Proposal generation and action classification are the two main tasks in TAD, and two-stage or multi-task methods are proposed with particular techniques such as sliding window [24], anchor-based method [10], Boundary Sensitive Network [15] and transformer structure with an end-to-end set-prediction-based detector [16]. However, the segmentation in our SSC task should process the whole video rather than only detecting actions, which is different from TAD. Specially, the two-stage and multi-task based algorithms in TAD field are able to handle SSC problem, while the two schemes have limitations which are mentioned in introduction section. Different with them, our Sequential Link framework can overcome their shortcomings and is proved effective in SSC experiments.

3 PROBLEM DEFINITION

SSC includes two tasks, i.e. scene segmentation and scene classification. It needs to partition the scene segments on the basis of the predefined scene categories. SS aims to extract all the segments satisfying the scene definition from the entire video. Different scene segments need to cover whole video and must be exclusive, without overlapping. Besides, the goal of scene classification is to assign correct scene label to each segment.

Because scenes are composed of consecutive shots, to simplify the scene problem, algorithms for SSC can be designed on the basis of the shot segmentation results. For an input video, firstly, it is segmented into several shots by shot algorithm [28]. The shots are denoted as $S_{shot} = \{a_1, a_2, \ldots, a_n\}$, where $n$ is the number of shots and the subscript numbers are indexed by time order. Scene algorithm takes efforts to group these shots into $n$ scenes $S_{scene} = \{b_1, b_2, \ldots, b_n\}$. Each scene $b_i$ consist several consecutive shots. The index of the start shot among them is denoted as $a_s$ and similarly, $a_e$ indicates the end shot. As a result, $b_i = [a_s, a_{s+1}, \ldots, a_e]$. Because each shot needs to be assigned to exactly one scene, it should be satisfied that $s_{i+1} = s_i + 1, i = 1, 2, \ldots, n - 1$. Furthermore, each scene $b_i$ needs to be classified. The label prediction of $b_i$ is denoted as $c_i = 1, 2, \ldots, n$. To sum up, $S_{scene} = \{b_1, b_2, \ldots, b_n\}$ is the result of scene segmentation. $L_{scene} = \{l_1, l_2, \ldots, l_n\}$ is the result of scene classification, where $l_i$ is the class label of scene $b_i$.

4 METHODOLOGY

The schematic diagram of our OS-MSL is shown in Figure 4. The main idea is extracting the multimodal feature of shots and then using sequential link framework with link tagging to solve segmentation and classification simultaneously. The architecture of OS-MSL consists of 3 parts. Firstly, unimodal features of shots are extracted. Then, features of different modalities are fused and the multimodal representations of shots are obtained. Finally, a transformer followed by CRF is used to tag the shots so as to determine the results of segmentation and classification.

4.1 Sequential Link Framework with Link Tagging

We use a link tagging to describe the linking relations between adjacent shots. In SS task, the link label set is: \{B->I, I->E, B->E, N\}. Therefore, e.g., a scene with 5 adjacent shots is tagged as \{B->I, I->I, I->I, I->E, N\}, which means the beginning shot is linked with 3 intermediate shots and ended with an ending shot that is not linked with the next scene.

In SSC task, the link label set has $5 \times C$ classes denoted as $C \times \{B->I, I->I, I->E, B->E, N\}$. For example, one Meeting scene with 4 adjacent shots is tagged as $C \times \{B->I, I->I, I->I, I->E, N\}$, which means the beginning shot is linked with 3 intermediate shots and ended with an ending shot that is not linked with the next scene.

In SSC task, the link label set has $5 \times C$ classes denoted as $C \times \{B->I, I->I, I->I, I->E, B->E, N\}$. Therefore, e.g., a scene with 5 adjacent shots is tagged as $C \times \{B->I, I->I, I->I, I->E, N\}$, which means the beginning shot is linked with 3 intermediate shots and ended with an ending shot that is not linked with the next scene.

4.2 Unimodal Feature Extraction

To assign shots with scene labels, the features of shots should be extracted. A shot, as a segment of the video, has its visual and audio
stream signals. The multimodal signals can provide abundant information so that we need to extract the multimodal representation of the shots. We use the visual and audio modalities. Nevertheless, because there exists large gap between visual and audio signals, unimodal features should be obtained first.

For visual modality, ResNet-18 [11] is used as the backbone of feature extractor. The visual feature is denoted as:

$$f^{(1)}_{vis} = F_{vis}(s_{vis}, \theta_{vis})$$

where $F_{vis}$ is the network architecture of ResNet-18, $\theta_{vis}$ is the parameters of $F_{vis}$ and $s_{vis}$ is the key frame of the shot. Currently, the frame at the middle time of the shot is chosen as the key frame.

Correspondingly, ResNet-VLAD [35] is used as the backbone of audio stream. The audio feature is denoted as:

$$f^{(1)}_{aud} = F_{aud}(s_{aud}, \theta_{aud})$$

where $F_{aud}$ is the network architecture of ResNet-VLAD, $\theta_{aud}$ is the parameters of $F_{aud}$ and $s_{aud}$ is the audio spectrogram of the shot. The spectrogram is obtained by processing the audio wave with short-time Fourier transform and Mel-scale filter banks.

### 4.3 Shot Difference and Correlation Network

The shot feature is used to fulfill two tasks. One is to decide whether a segmentation point is placed at the end of the shot. The other is to express the category information. In order to reinforce the feature representation, the local and global relations among adjacent shots are crucial. Therefore, DiffCorrNet is proposed to model the relations. For a certain shot $a_j$, consider its adjacent shots $a_{j-(k-1)}, a_{j-(k-2)}, \ldots, a_{j+k}$. The total $2k$ shots are used to generate the enhanced feature of $a_j$.

At first, the boundary modeling is constructed by measuring the difference between the former $k$ shots and the latter $k$ shots. The difference can describe the confidence of whether the scene boundary is placed at the end of shot $a_j$. It is intuitional because if the former shots and the latter one are dissimilar, they likely belong to different scenes. Inspired by BNet module proposed in [18], we also use $\cos$ metric to model the boundary information $g_{modal}$, where $modal \in \{vis, aud\}$. The model is shown in Figure 5.

$$h(a_j) = \sum_{i \in A(a_j)} w(a_j, a_i) f^{(1)}_{modal}(a_i)$$

where $A(a_j) = \{j-(k-1), j-(k-2), \ldots, j+k\} \setminus \{j\}$. 

To describe the global correlation between shots, the $f^{(1)}_{modal}$ of $2k$ shots are aggregated into a feature $h_{modal}$. Features of $2k - 1$ neighboring shots are weighted by shot attention mechanism and then summed to get the fused feature. The concept is illustrated in Figure 6.

$$w(a_j, a_k)$$

is denoted as the weight between shot $a_j$ and $a_k$. Concretely, the shot features are transformed by a fully connected network and then do subtraction to get a representation measuring the shot difference. Finally, a multilayer perceptron (MLP) is applied to get the weight.

After calculating all the weights between $a_j$ and its neighboring shots, the aggregated feature is obtained:

$$h(a_j) = \sum_{i \in A(a_j)} w(a_j, a_i) f^{(1)}_{modal}(a_i)$$

where $A(a_j) = \{j-(k-1), j-(k-2), \ldots, j+k\} \setminus \{j\}$. 

The purpose of this aggregated feature is to acquire the semantic information in adjacent shots with highly relation. The abundant information can be beneficial for better shot classification. To sum up, the proposed DiffCorrNet including the sub networks mentioned above takes the 2k adjacent shots around shot \( a_j \) into account and generates boundary feature \( g(a_j) \) and aggregated feature \( h(a_j) \). The final unimodal features of \( a_j \) are:

\[
f^{(2)}_{\text{modal}} = \text{concat}(f^{(1)}_{\text{modal}} \cdot g_{\text{modal}}, h_{\text{modal}})
\]

where \( \text{concat}() \) means the operation of feature concatenation. The parameters in DiffCorrNet are denoted as \( \theta_{dc} \). Each modality has its own DiffCorrNet and the parameters in different DiffCorrNets are not shared.

By this means, DiffCorrNet can guide the extraction of shot feature for better scene segmentation and classification.

### 4.4 Multimodal Feature Fusion

Before feature fusing, batch normalization (BN) [14] is used to uniform the scales of different modalities, as the distributions and amplitudes of visual and audio signals are quite different.

It is significant to fuse single-modality features to get the multimodal representation of the shots. In theory, there are several ways to fuse \( f^{(2)} \), e.g. early fusion or late fusion. In OS-MSL, early fusion is adopted so that the transformer may take the multimodal feature into account. As a result, the multimodal features of shots \( f^{(3)} \) can be obtained by:

\[
f^{(3)} = \text{concat}(BN(f^{(2)}_{\text{vis}}), BN(f^{(2)}_{\text{aud}}))
\]

where the fused feature is obtained by concatenating all the provided unimodal features. This framework can be compatible with more modalities if necessary.

### 4.5 Task head

Task head aims to complete the SS and SSC tasks. In this paper, a transformer [32] followed by conditional random fields (CRF) [13] is chosen as the task head. Denote the head as \( F_{\text{head}} \), we give the shot label \( y \) as:

\[
y = F_{\text{head}}(f^{(3)}, \theta_{\text{head}})
\]

where \( \theta_{\text{head}} \) is the parameters of \( F_{\text{head}} \).

The equations (1)-(6) represent the entire network of OS-MSL with trainable parameters \( \theta_{\text{vis}}, \theta_{\text{aud}}, \theta_{\text{dc}}, \theta_{\text{head}} \) which are trained by CRF loss [13]. After training, each shot of a testing video can be classified by OS-MSL with \( 5 \times C \) link labels. The final SSC results \( \{b_1, b_2, \cdots, b_n\} \) of segmentation boundaries and \( \{l_1, l_2, \cdots, l_n\} \) of category labels will be inferred by the predicted shots links \( \{y_1, y_2, \cdots, y_m\} \). In detail, consecutive shots or one shot with link labels \( \{B\rightarrow I, I\rightarrow L, I\rightarrow E, N\} \) or \( \{B\rightarrow E, N\} \) or \( \{N\} \) in category \( c \) are treated as one scene, and its classification label is \( c \).

### 5 EXPERIMENTS

#### 5.1 Dataset

**TI-News** To promote the video structuring research for news application, a new dataset, TI-News, is presented. It is one of the significant contributions in this paper. The dataset has over 500 news videos, including about 253,000 shots and 26,500 scenes. In training and validation phase, all the videos come from 8 news programs. In testing phase, 44 videos are selected for evaluation. 34 videos of them are chosen from aforementioned 8 programs. There is no overlap between the 34 testing videos and the training videos. Furthermore, 10 more videos collected from other 10 distinct programs are used for generalization testing. The detailed dataset splits is shown in supplementary material.

With carefully analyzing the characteristics of news videos, we established a label system for news SSC task. Seven representative categories are defined. They are *Studio*, *Outdoor*, *Person Interview*, *Remote*, *Meeting*, *Speech* and *News Board*. In addition, class *Others* is used as negative class denoting the scenes which not belong to above-mentioned seven categories. Some graphical illustrations of TI-News dataset is shown in Figure 7. In addition, detailed proportions of scene quantities are illustrated in Table 1.

**MovieScenes** MovieScenes is a SS dataset especially for movie. It’s frequently used for evaluating the algorithms in scene boundary detection field. The dataset includes 318 movies from a large-scale movie understanding dataset MovieNet [12]. In this paper, to fairly evaluate the scene algorithms, the experimental setup of MovieScenes keeps identical with the work in [18].

#### 5.2 Evaluation Criteria

According to the definitions of segmentation and classification, the algorithm is evaluated in multiple aspects. Denote the predicted scene segmentation results as \( S_{\text{pre}} = \{b_1, b_2, \cdots, b_{n_{\text{pre}}}\} \) and classification results as \( L_{\text{pre}} = \{l_1, l_2, \cdots, l_{n_{\text{pre}}}\} \). The ground truth results are \( S_{\text{gt}} = \{b_1^{(gt)}, b_2^{(gt)}, \cdots, b_{n_{\text{gt}}}^{(gt)}\} \) and \( L_{\text{gt}} = \{l_1^{(gt)}, l_2^{(gt)}, \cdots, l_{n_{\text{gt}}}^{(gt)}\} \).
For SS task, the evaluation criteria should measure the difference between \( \text{Sp}_{\text{pre}} \) and \( \text{Sp}_{\text{gt}} \). In our experimental settings, the precision and recall of segmentation points (Seg-points) are adopted. For \( b_i \) in \( S_{\text{Scene}} \), the end shot index \( e_i \) is denoted as the Seg-point. As a result, the list of Seg-points in \( S_{\text{pre}} \) is \( \{e_1, e_2, \ldots, e_{\text{pre}}\} \) and the Seg-points list in \( S_{\text{gt}} \) is \( \{e_1^{(gt)}, e_2^{(gt)}, \ldots, e_{\text{gt}}^{(gt)}\} \).

Based on the definition, true positive (TP) is counted by enumerating every predicted and ground truth Seg-points. Concretely, for \( \forall b_i \in S_{\text{pre}}, \text{if } 3b_i^{(gt)} \in S_{\text{gt}} \) satisfy \( e_i = e_i^{(gt)} \), the case will be counted into \( TP_{\text{seg}} \). The mathematical formulation of \( TP_{\text{seg}} \) is:

\[
TP_{\text{seg}} = \sum_{i=1}^{n_{\text{pre}}} \sum_{j=1}^{n_{\text{gt}}} \langle \{e_i = e_j^{(gt)}\}\rangle
\]  

(7)

where \( \langle \{\cdot\}\rangle \) is the operator to judge \( x \) is true or not, i.e. \( \langle \{x\} \rangle = 1 \) if \( x \) is true or 0.

Then, the precision (P), recall (R) and F1 score (F1) for SS can be obtained.

Furthermore, in SSC task, classification is taken into account. Classification aims to assign correct scene label to each segment. A correct predicted scene should have precise scene boundary and accurate label. As a result, the counting of true positive \( TP_{\text{seg} \& \text{cls}} \) will consider category results in addition, which is defined as:

\[
TP_{\text{seg} \& \text{cls}} = \sum_{i=1}^{n_{\text{pre}}} \sum_{j=1}^{n_{\text{gt}}} [\{e_i = e_j^{(gt)}\}] \cdot [\{l_i = l_j^{(gt)}\}] 
\]  

(8)

Similarly, the precision, recall and F1 score for segmentation and classification are calculated. Because of uneven class sizes, micro and macro metrics are both evaluated.

In previous works [7, 18], mean average precision (mAP) and recall are used for evaluation SS. However, the metrics are not consistent and they need enumerating thresholds to calculate while the thresholds are not common in various algorithms. The proposed evaluation criteria don’t have these limitations and it can directly measure the segmentation and classification outputs. As a result, the proposed criteria is more reasonable and it is used in this paper for fairly methods comparison.

### 5.3 Compared Methods

In this paper, the experiments are conducted on both SS task and SSC task. For SS task, LGSS [18] and ShotCoL [7] are currently the SOTA algorithms, so we use them as our strong baselines. Other methods such as Siamese [2], StoryGraph [30], Grouping [21] and etc., are not fresh enough to compare. In recent literatures, it is shown that these algorithms perform worse than LGSS to a great extent. Therefore, we conduct comparison experiments with the latest algorithms to reveal the ability of proposed OS-MSL. Note that origin version of LGSS uses late fusion, to utilize the multimodal signals. In this paper, a variant of LGSS is developed with early fusion. Early fusion means the features of multi modalities are fused before the LSTM in LGSS. The method is denoted as LGSS-early.

For SS task, our OS-MSL uses a 5-classes label set \( \{\text{B->I}, \text{I->I}, \text{I->E}, \text{B->E}, \text{N}\} \) to address it, which is denoted as OS-MSL(SS).

For SSC task, Two-stage and Multi-task are used as baselines. Two-stage consists of two models to address the segmentation and classification problems respectively. Multiple kinds of Two-stage methods can be established by choosing different pairs of segmentation and classification models. In this paper, LGSS-early and our OS-MSL for scene segmentation only are selected as the segmentation model. The corresponding methods are denoted as Two-stage (LGSS-early) and Two-stage (OS-MSL). On the other hand, Multi-task is built on LGSS-based framework and combined with classification branch to handle the multiple tasks. Detailed models of these methods will be introduced in supplementary material. For SSC task, our OS-MSL uses label set \( C \times \{\text{B->I}, \text{I->I}, \text{I->E}, \text{B->E}, \text{N}\} \) to address it, which is denoted as OS-MSL(SSC).

### 5.4 Results on TI-News Dataset

In this subsection, OS-MSL is evaluated on TI-News dataset with standard testing set and generalized testing set. The standard testing set refers to 34 videos from 8 programs which are included in training set. On the contrary, in generalized testing set, 10 more testing videos are selected from distinct news programs and their formats of program editing are very different with training samples. Therefore, the generalized testing set can evaluate the generalization of scene methods.

The results of SS experiments are shown in Table 2. Our OS-MSL(SS) outperforms the strong baselines, i.e. LGSS, LGSS-early and ShotCoL. It reflects that the scheme of Sequential Link is superior for solving the segmentation problem. To explore the reasons, we summarize that OS-MSL(SS) reforms SS task to a link form, and explicitly reveal richer sequential relation between shots using a 5-classes link label set. In generalized testing set, OS-MSL(SS) also shows its splendid performance. It indicates OS-MSL(SS) has great generalization ability on SS task. Furthermore, the SSC methods, Multi-task and OS-MSL(SSC), are examined their ability of SS. As the Table 2 shows, Multi-task outperforms LGSS-early and OS-MSL(SSC) achieves the best results, which illustrate that utilizing scene category information can promote SS task.

SSC results are illustrated in Table 3. The micro precision and micro recall of OS-MSL(SSC) reach 86.20% and 85.40% for the evaluation of segmentation and classification. Both precision and recall exceed 85%, which lead to an excellent performance of SSC task. In supplementary material, some visualized examples are presented. It can be found that the visualization results are nearly in accord with the human’s feeling. Compared with Two-stage and Multi-task, proposed OS-MSL(SSC) shows great superiority. It is verified that the Sequential Link framework can better exploit the relation between segmentation and classification tasks.

In generalized testing set, for OS-MSL(SSC), micro F1 score of OS-MSL(SSC) exceed 86%. It illustrates the good generalization of our algorithm. Relatively, the performance of LGSS, ShotCoL, Two-stage and Multi-task decreases to a large extent. These phenomena further support the superiority of our method.

However, the results are quite bad under the macro evaluation. The macro F1 scores of these methods are in the range of 48.93% ∼ 63.88%. The reason is that the algorithms perform poorly to find Outdoor, Remote, Speech categories, which greatly affect the macro results. The generalization ability of recognizing hard categories is a point for improvement.
Table 2: Scene segmentation results on TI-News standard testing set and generalized testing set.

| Methods               | Standard Testing Set | Generalized Testing Set |
|-----------------------|----------------------|-------------------------|
|                       | P        | R        | F1       | P        | R        | F1       |
| LGSS [Rao et al., 2020] | 77.68   | 54.94   | 64.36   | 72.96   | 40.62   | 52.19   |
| ShotCol. [Chen et al., 2021] | 78.10   | 67.48   | 72.40   | 75.10   | 53.55   | 62.52   |
| LGSS-early [Rao et al., 2020] | 84.98   | 81.08   | 82.98   | 88.40   | 74.72   | 80.99   |
| OS-MSL(SS)             | 90.14   | 88.70   | 89.41   | 90.48   | 87.78   | 89.11   |
| Multi-task             | 88.79   | 83.50   | 86.06   | 87.82   | 78.84   | 83.09   |
| OS-MSL(SSC)            | 90.40   | 89.37   | 89.98   | 90.75   | 89.35   | 89.54   |

Table 3: Scene segmentation and classification results on TI-News standard testing set and generalized testing set.

| Methods             | Standard Testing Set | Generalized Testing Set |
|---------------------|----------------------|-------------------------|
|                     | Micro | Macro   | Micro | Macro   |                 |
|                     | P    | R      | F1    | P       | R       |
| Two-stage(LGSS-early) | 77.0  | 73.47  | 75.19 | 69.06   | 65.01   | 66.97   | 77.31   | 65.34   | 76.82   | 53.27   | 43.98   | 48.18   |
| Two-stage(OS-MSL)    | 83.21 | 81.87  | 82.53 | 76.44   | 75.39   | 75.91   | 83.60   | 81.11   | 82.34   | 59.59   | 59.56   | 59.57   |
| Multi-task           | 81.40 | 76.55  | 78.90 | 80.20   | 64.04   | 71.21   | 79.11   | 71.02   | 74.85   | 57.70   | 43.47   | 48.93   |
| OS-MSL(SS)           | 86.20 | 85.40  | 85.80 | 83.74   | 78.60   | 81.09   | 87.16   | 86.79   | 86.97   | 71.56   | 57.69   | 63.88   |

Table 4: Scene segmentation results on MovieScenes.

| Methods               | Segmentation |
|-----------------------|--------------|
|                       | P | R | F1 |
| LGSS-early [Rao et al., 2020] | 45.56 | 44.97 | 45.26 |
| LGSS [Rao et al., 2020] | 45.64 | 46.36 | 46.00 |
| ShotCol. [Chen et al., 2021] | 51.02 | 47.97 | 49.45 |
| OS-MSL(SSC)           | 49.89 | 50.56 | 50.22 |

Table 5: Ablation study on TI-News standard testing set.

| Methods               | Segmentation and Classification |
|-----------------------|---------------------------------|
|                       | Micro | Macro |
|                       | P    | R    | F1   | P    | R    |
| OS-MSL(SSCC)           | 86.20 | 85.40 | 85.80 | 83.74 | 78.60 | 81.09 |
| w/o Audio             | 83.26 | 78.30 | 80.18 | 76.30 | 70.41 | 73.33 |
| w/o Visual            | 81.57 | 71.36 | 76.12 | 69.44 | 60.39 | 61.24 |
| w/o BB Training       | 85.86 | 81.90 | 83.83 | 82.09 | 73.07 | 77.32 |
| w/o BN                | 85.27 | 79.95 | 82.52 | 80.24 | 69.09 | 74.25 |
| w/o DiffCorrNet       | 86.24 | 79.63 | 82.80 | 83.36 | 71.61 | 77.04 |
| w/o CRF               | 84.60 | 82.11 | 83.34 | 82.65 | 74.17 | 78.18 |

Table 6: Ablation study on TI-News generalized testing set.

| Methods               | Segmentation and Classification |
|-----------------------|---------------------------------|
|                       | Micro | Macro |
|                       | P    | R    | F1   | P    | R    |
| OS-MSL(SSCC)           | 87.16 | 86.79 | 86.97 | 81.56 | 57.69 | 63.88 |
| w/o Audio             | 78.39 | 69.03 | 73.41 | 54.55 | 47.37 | 50.70 |
| w/o Visual            | 82.76 | 71.59 | 76.77 | 57.55 | 32.95 | 41.91 |
| w/o BB Training       | 85.56 | 78.27 | 81.75 | 60.54 | 32.82 | 56.42 |
| w/o BN                | 86.40 | 76.70 | 81.26 | 63.98 | 53.55 | 58.30 |
| w/o DiffCorrNet       | 86.20 | 78.98 | 82.43 | 67.39 | 60.41 | 63.71 |
| w/o CRF               | 83.46 | 80.26 | 81.83 | 66.59 | 36.76 | 61.29 |

5.5 Results on MovieScenes Dataset

The experiments on public available data set of MovieScenes are conducted to evaluate our performance. MovieScenes is established just for SS task so that only segmentation performance is evaluated. In Table 4, it illustrates our OS-MSL(SS) outperforms the latest algorithms, i.e. LGSS, LGSS-early and ShotCol. As a result, the scheme of our Sequential Link is also effective on movie domain. It can further support that the proposed Sequential Link framework is superior for solving the segmentation problem alone.

5.6 Ablation Study

To further explore the effectiveness of several modules designed in OS-MSL, the ablation study experiments of 1) w/o Backbone (BB) Training, 2) w/o BN, 3) w/o DiffCorrNet and 4) w/o CRF are conducted. Table 5 shows the ablation results of SSC task on TI-News standard testing set. OS-MSL(SSC) w/o BB Training, w/o BN and w/o DiffCorrNet all perform worse F1 score compared with OS-MSL(SSC). The phenomenon illustrates all these modules have their positive effects for SSC to different extent. The ablation study of OS-MSL(SS) on TI-News has the similar conclusion. Its detailed experiments are described in supplementary material.

It is feasible to use other sequence model to tag shots. However, CRF is an effective module because it can utilize the transition relations between sequence labels. For example, it’s impossible that a Meeting_{B→I} follows a Meeting_{I→J}. As Table 5 shows, OS-MSL(SSC) w/o CRF performs 1.6%/3.3% on precision/recall lower than proposed OS-MSL(SSC). In addition, we also substituted transformer with LSTM to explore the sequence model selection. The model achieves 83.87%/78.63% micro precision and recall, which is much worse than the results of transformer. The experiments illustrate the model of transformer with CRF is the better choice.

In Table 6 and Table 7, the ablation experiments on TI-News generalized testing set and MovieScenes testing set are conducted.
Table 7: Ablation study on MovieScenes.

| Methods                  | Segmentation |
|--------------------------|--------------|
|                          | P  | R  | F1  |
| OS-MSL(SS)               | 49.89 | 50.56 | 50.22 |
| w/o Audio               | 51.06 | 46.59 | 48.72 |
| w/o Visual              | 28.57 | 13.51 | 18.34 |
| w/o BB Training         | 49.73 | 49.60 | 49.67 |
| w/o BN                  | 49.25 | 50.38 | 49.81 |
| w/o DiffCorrNet         | 47.94 | 47.83 | 47.89 |
| w/o CRF                 | 48.15 | 50.43 | 49.27 |

The results also reflect that all the modules are effective to improve the performance of SS and SSC task.

6 DISCUSSION

6.1 Multimodal Fusion

In this subsection, we try to explore the effects of different modalities on the performance. Firstly, in news video scenario, as Table 5 and 6 show, our proposed multimodal method outperforms either unimodality method. In fact, audio and visual modalities have their specific roles, which should be complementary. Audio contains strong clues for segmentation because the switch of voices tends to be the Seg-points. However, classification only with audio modality is almost infeasible. Some scenes categories (such as Studio, Meeting and News Board) can not be distinguished because most of them have the same voice. As a result, the experimental results of OS-MSL w/o Visual are quite low. On the other hand, visual unimodality is also insufficient. The visual expression has large inner-class variance. For example, the frames of Person Interview include not only person talking but also the things describing what the person is talking about. The Seg-points of this case are hard to determine.

Secondly, in movie scenario, as Table 7 shows, OS-MSL w/o Visual performs poorly in MovieScenes dataset. On the contrary, OS-MSL w/o Audio achieves decent results, which are slightly worse than OS-MSL. The phenomenons illustrate that the audio signals are insufficient and visual modality are dominated to segment movie scenes. Nevertheless, audio modality also contains its exclusive clues to assist the task, which makes the multimodal method outperform the unimodal ones.

It may be argued that text is another useful modality. With the techniques of optical character recognition (OCR) and automatic speech recognition (ASR), text content in frames and audio can be recognized and offers the detail semantic information about the video. However, for the SSC task, the text modality is almost useless. The text content is little related to the definition of SSC task and sometimes it is noise which may cause negative effect.

Some cases may support this opinion. In news video scenario, presenter may report several news in the studio, but these news are unrelated to Studio category and we also can’t judge the scene boundary by the content of these news. In movie scenario, the words of background music or voice are meaningless. As a result, we abandon the text information and just use visual and audio modalities to address SSC task.

6.2 Learning Curves

In this subsection, we focus on the model collapse problem. During the training phase, the segmentation loss and classification loss of multi-task method are recorded individually. For comparison, the learning curve of OS-MSL(SSC) is also plotted, which is denoted as OS-MSL(SSC). Each curve is drawn according to the losses after normalization due to the different magnitudes of various tasks.

In Figure 8, it is illustrated that the learning curve of classification oscillates greatly. The reason is that the model is dominated by segmentation task. The gradients during training tend to optimize the segmentation task, which is inconsistent with classification task. As a result, the conflict between the two tasks may affect the model training, which finally harms the performance. Comparatively, the learning curve of OS-MSL(SSC) converges more rapidly and is more stable. The phenomenon reflects our one-stage method can alleviate the asynchronization issue existing in multi-task learning and obtain better performance for SSC task.

6.3 Sequential Link Framework

The tasks of SS and SSC can be characterized in a unified manner: a link form with link tagging. Furthermore, our framework can be easily generalized to other similar video sequential modeling tasks, e.g., action, activity or event recognition, which can be categorized to a unified definition: find a boundary between consecutive segments/clips with auxiliary information such as label, key frame, and caption. For example, video action recognition aims to recognize human actions in a video, thus the task is to find consecutive video frames with the same action labels, which can be obviously completed by our framework.

7 CONCLUSION

The paper raises a novel video structuring problem, i.e. SSC. To address the model collapse problem existing in two-stage and multi-task methods, Sequential Link framework is introduced to reform the two tasks into a unified one using a novel link form. Based on this, OS-MSL is proposed to leverage common information and avoid disharmony of the two tasks. To strength the shot representation, a new module DiffCorrNet is developed to extract differences and correlations among successive shots. For news program application, a specific dataset, TI-News, is established. The experiments on TI-News and MovieScenes illustrate its effectiveness.
REFERENCES

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. 2021. Vivit: A video vision transformer. arXiv preprint arXiv:2103.15691 (2021).

[2] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. 2015. A deep siamese network for scene detection in broadcast videos. In Proceedings of the 23rd ACM international conference on Multimedia. 1199–1202.

[3] Giuseppe Becquartono, Angelo Chianese, Vincenzo Moscati, and Antonio Piscariello. 2005. Feature-based video shot detection for video segmentation. IEEE Transactions on Circuits and Systems for Video Technology 15, 3 (2005), 365–377.

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

[5] Rich Caruana. 1997. Multitask learning. Machine learning 28, 1 (1997), 41–75.

[6] Vasileios T Chasanis, Aristidis C Likas, and Nikolaos P Galatsanos. 2008. Scene detection in videos using shot clustering and sequence alignment. IEEE transactions on multimedia 11, 1 (2008), 89–100.

[7] Shixing Chen, Xiaohan Nie, David Fan, Dongjing Zhang, Vimal Bhat, and Raffay Hamid. 2021. Shot Contrastive Self-Supervised Learning for Scene Boundary Detection. arXiv preprint arXiv:2104.13537 (2021).

[8] Costas Cotsaces, Nikos Nikolaidis, and Ioannis Pitas. 2006. Video shot detection and condensed representation: a review. IEEE signal processing magazine 23, 2 (2006), 26–37.

[9] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In Proceedings of the IEEE/CVF international conference on computer vision. 6202–6211.

[10] Jiyang Gao, Zhengheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. 2017. Turn tap: Temporal unit regression network for temporal action proposals. In Proceedings of the IEEE international conference on computer vision. 3628–3636.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition. 770–778.

[12] Qinqiu Huang, Yu Xiong, Anyi Rao, Jiawei Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 10146–10155.

[13] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. 2020. Movienet: A holistic dataset for movie understanding. In Proceedings of the IEEE conference on computer vision and pattern recognition. 1049–1058.

[14] Panagiotis Sidirooulos. 2006. Video shot detection and condensed representation. IEEE transactions on Multimedia 11, 1 (2009), 89–100.

[15] Hari Sundaram and Shih-Fu Chang. 2000. Video scene segmentation using video clustering and graph analysis. Computer vision and image understanding 77, 6 (2000), 562–587.

[16] Xiaolong Liu, Qimeng Wang, and Yao Hu. 2018. Temporal segment networks for action recognition in videos. In Proceedings of the IEEE conference on computer vision and pattern recognition. 4489–4497.

[17] Hari Sundaram and Shih-Fu Chang. 2000. Video scene segmentation using video clustering and graph analysis. Computer vision and image understanding 77, 6 (2000), 562–587.

[18] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Shih-Fu Chang. 2016. Temporal action localization in untrimmed videos via multi-stage cnns. In Proceedings of the IEEE conference on computer vision and pattern recognition. 6202–6211.

[19] Minerva Yeung, Boon-Lock Yeo, and Bede Liu. 1998. Segmentation of video by clustering and graph analysis. Computer vision and image understanding 71, 1 (1998), 94–109.