Multimodal High-order Relation Transformer for Scene Boundary Detection

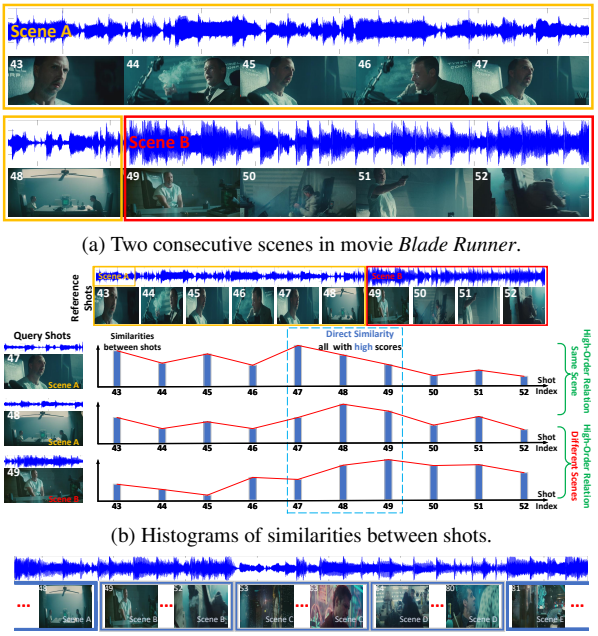
Xi Wei[1], Zhangxiang Shi[1], Tianzhu Zhang[1,*], Xiaoyuan Yu[2], Lei Xiao[2]
[1] University of Science and Technology of China
[2] Huawei Cloud
{wx33921, szxszx}@mail.ustc.edu.cn, tzzhang@ustc.edu.cn,
yuxiaoyuan@huawei.com, xiaolei199311@126.com

Abstract

Scene boundary detection breaks down long videos into meaningful story-telling units and plays a crucial role in high-level video understanding. Despite significant advancements in this area, this task remains a challenging problem as it requires a comprehensive understanding of multimodal cues and high-level semantics. To tackle this issue, we propose a multimodal high-order relation transformer, which integrates a high-order encoder and an adaptive decoder in a unified framework. By modeling the multimodal cues and exploring similarities between the shots, the encoder is capable of capturing high-order relations between shots and extracting shot features with context semantics. By clustering the shots adaptively, the decoder can discover more universal switch pattern between successive scenes, thus helping scene boundary detection. Extensive experimental results on three standard benchmarks demonstrate that the proposed model performs favorably against state-of-the-art video scene detection methods.

1. Introduction

Cognitive science has discovered that humans usually comprehend a long video by breaking down it into meaningful units and reasoning about their relationships [20]. Therefore, dividing a long video into a series of meaningful story-telling video scenes, i.e. video scene detection, turns out to be a crucial task towards high-level video understanding. In filmmaking and video production, the term 'scene' is a basic unit for story-telling, consisting of a series of semantic cohesive shots, while the term 'shot' is a series of frames captured by the same camera over an uninterrupted period of time [4]. The task of video scene detection has drawn remarkable attention and can be widely adopted to various tough applications, including video caption, content-driven video search, scene classification, human-centric storyline construction and so on [8, 15, 31]. Although significant progress has been achieved in recent years, video scene detection remains challenging due to requiring semantic understanding in a long-term video.



(a) Two consecutive scenes in movie *Blade Runner*.

(b) Histograms of similarities between shots.

(c) Diverse scene boundaries in movie *Blade Runner*.

Figure 1: (a) There is minimal visual alteration at the scene boundary, while the audio content undergoes significant changes. (b) Direct similarities between shots near the boundary can all have high scores, while the trend of similarity variation can provide more discriminative signals. (c) Scene boundaries can vary greatly even in the same movie. Therefore, learning shot-level representations alone is insufficient for this class-agnostic task.

Recently, researchers have proposed methods like [4, 15, 20, 38] that leverage visual cues in unlabeled videos to model scene boundary. These approaches mainly employ an unsupervised contrastive learning strategy to differentiate shots from distinct scenes. Although these methods have taken a big step forward, they simply use visual appearance cues to generate pseudo labels for contrastive

learning, resulting in the model learning the differences of shots at the visual apparent level rather than semantic level. In order to address above limitations, approaches such as [18, 19, 23, 27] introduce various expert networks to model multimodal signals like visual, audio, cast and place in various video understanding tasks. These methods still struggle to accurately identify scene boundaries in long-term videos, particularly those with complex story-telling structures. This is due to that modeling context of the shots plays a crucial role in identifying boundaries. Furthermore, all previous methods [1, 4, 23, 20, 27, 38] primarily emphasize learning discriminative shot features and overlook that scene boundary detection is a class-agnostic task. Thus they fail to identify more universal switch pattern between different scenes which is less dependent on appearance.

Based on the above analysis and the characteristics of scene boundary detection, we believe following aspects are important in this task: (1) The multimodal shot representations can provide comprehensive cues to detect scene boundary. (2) The high-order relation provides corrections for the context of shots. (3) The scene-level representations, derived from adaptively aggregation of shots features, can effectively indicate the switch pattern between different scenes. Firstly, as shown in Figure 1a, there is minimal alteration in the visual appearance between *Scene A* and *scene B*. However, the audio signal undergoes a significant change, making it easier to discern the transition between scenes. Secondly, as shown in Figure 1b, the shots before and after the scene boundary, such as shot 48 and shot 49, exhibit high similarities, suggesting that the scene switch can occur seamlessly. Neither the environment nor the characters change, only the two men's behavior changes. Under this circumstances, considering only the similarities between shots (first-order relation) can lead to undetectable blurred boundaries. Fortunately, by representing shots 47-49 as 10-dimensional similarity vectors based on their similarities with reference shots 43-52, we can observe a significant change near the scene boundary. This motivates us that the similarities between these similarity vectors, denoted as high-order relation, can correct for the first-order relation. Thirdly, unlike the common video action detection task, video scene detection is a category-agnostic task [5]. As shown in Figure 1c, even within the same movie, the appearances of the 'scene boundary shots' (i.e. shots 48, 52, 63, 80) still vary greatly. It is more important to discover the switch pattern between successive scenes rather than learn shot-level representations. This encourages us to learn more contextual and category-agnostic scene-level representations, which can be adaptively aggregated from shot-level features.

Take these observations together, we propose the Multimodal High-order Relation Transformer for video scene detection, which can model multimodal cues, high-order relation and scene adaptive clustering in a unified structure. Specifically, we first take multimodal expert networks and clip encoders to extract multimodal clip-level shot embeddings for the input shot sequence. Then our designed multimodal high-order relation transformer, consisting of a high-order encoder and an adaptive decoder, is employed to model the high-order relation for the shots and generate contextual scene-level representation in a unified framework. In detail, the high-order encoder primarily employs the multi-head and self-attention mechanism on the multimodal shot embeddings to produce high-order relationships ('relation in relation'), thereby enhancing the appropriate correlations between shots and suppressing erroneous ones. And the adaptive decoder merges contextual shot embeddings into scene representations with learnable scene prototypes and cross-attention. Finally, we take those contextualized shot embeddings for scene boundary classification and scene embeddings for boundary position regression.

The major contributions of this work can be summarized as follows: (1) We propose the Multimodal High-order Relation Transformer, which can model multimodal cues, high-order relation and scene adaptive clustering in a unified structure. (2) In our transformer, we design the high-order encoder to enhance the context of shots and rectify noises in the correlation map of shots caused by their imperfect features. We further employ the adaptive decoder to aggregate shot features, which is better at capturing switch pattern between scenes rather than visual appearance. (3) Extensive experimental results on three standard benchmarks including MovieNetSSeg [12], OVSD [27] and BBC planet earth [1] demonstrate that the proposed model can outperform previous works by a large margin, and can be even competitive with those pre-training methods which consume more than 5 times training data.

2. Related Work

The basic assumption of video scene detection is that one shot can only belong to one scene, thus scene boundaries are a subset of shot boundaries [30]. As a result, the pre-task of scene detection is usually shot detection [14], which divides a video into sequences of time-continuous, non-overlapping shots. Given shot sequences, the video scene segmentation problem can be modeled as shot grouping, which clusters consecutive shots into scenes, or binary classification, which predicts whether a shot is a scene boundary. Many methods have been proposed [1, 4, 27, 20, 38], which can be categorized as unsupervised methods, supervised methods and self-supervised methods.

Unsupervised Methods. Early works utilize a variety of unsupervised methods. Among them, [24, 29] cluster shots according to spatiotemporal video features or shot color similarities. [27] obtains a consecutive segmentation by dynamic programming. [28] presents a novel normalized cost function for optimally grouping shots into scenes. Other

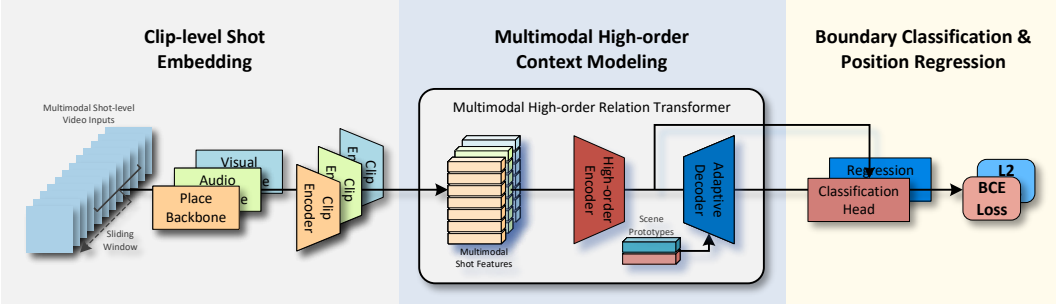


Figure 2: Our method follows a three-stage pipeline. Firstly, we extract multimodal features with short-term context using multimodal backbones and clip encoders. In the second stage, our high-order encoder models complex multimodal cues and context for each shot, while the decoder adaptively aggregates shots within the same scene. Finally, we employ separate heads for central shot classification and scene boundary position regression. Please refer to the text for more details.

works make use of multimodal information [27]. For instance, [30] uses both low-level and high-level audio features, visual features to build scene transition graph and extend the unimodal STG-based temporal segmentation technique to multimodal scene segmentation. However, these methods are trained on small-scale datasets [1, 27], leading to poor generalization. Our method turns to deep learning and is trained with newly proposed large dataset [12], obtaining much better performance and generalization.

Supervised and Self-Supervised Methods. As more and larger datasets are presented [4, 5, 20, 23, 38, 32, 37, 39], researchers are turning to supervised methods. With large-scale datasets, the deep neural network has a great advantage on extracting semantic features. [22] uses Alexresnet365 to extract scene features, while [13] achieves video scene segmentation using an image captioning model to extract semantic information of shots. [23] proposes a supervised baseline utilizing multimodal features to perform a shot-level binary classification, followed by an iterative optimal grouping. Recently, self-supervised methods achieve the best performance. Among them, [4] proposes a shot contrastive pre-training strategy to learn distinguishable shot representation. [38] achieves scene consistency representation learning using a novel positive sample selection strategy in contrast learning. And [20] mainly focuses on designing boundary-aware self-supervised tasks. However, these methods ignore multimodal information and simply use visual appearance cues to contrast different shots, while scene segmentation relies on semantic context. On the contrary, we can jointly model multimodal cues, high-order relation enhanced context with our proposed high-order encoder. And our proposed decoder can adaptive merge shot features into scene-level features to discover the switch pattern between video scenes.

3. Multimodal High-order Relation Transformer

As shown in Figure 2, our method deals with the multimodal shot-level video inputs in a pipeline of clip-level shot embedding, multimodal high-order context modeling and boundary classification. Given a full-length input video, we first split it into a constituent set of shots with the standard shot detection techniques [30]. We then use the sliding window to get a continuous subsequence of shots $\{s_0, s_1, ..., s_l\}$, and extract its multimodal representations (i.e. visual, place and audio) with the various expert networks. And the clip encoder is employed to model each shot's short-term context. Based on these multimodal shot features, we then perform multimodal high-order context modeling to exploit story-telling semantics for each shot. Specifically, the proposed multimodal high-order relation transformer first adopts the high-order encoder to take both multimodal cues and high-order relationships into consideration. Then the followed adaptive decoder is employed to aggregate shots into scenes. Finally, we adopt both shot features with contextual information and dynamically aggregated scene features for scene boundary prediction.

3.1. Clip-level Shot Embedding

Multimodal Expert Networks. A long-term video is a typical multimodal data containing different high-level semantics [23]. Thus visual features extracted by classical backbones like I3D [2] or TSN [36] are insufficient for the video scene detection task. Instead, the story-telling high semantics can be better described and complemented by multimodal representations. Observing that, in addition to visual contents, a 'scene' is closely related to shots sharing common place, audio elements. Following [23] we adopt visual, place and audio expert networks to extract multimodal shot features.

Following [1, 20, 23], we first extract the key frame for each shot in the sequence $\{s_0, s_1, ..., s_l\}$. Then we (1) adopt the ViT [7] to get visual features $\mathbf{R^v} = \{r_0^v, r_1^v, ..., r_l^v\}$, (2) employ the ResNet50 [10] pretrained on Places dataset [40] to get place features $\mathbf{R^p} = \{r_0^p, r_1^p, ..., r_l^p\}$, and (3) take stft [34] followed by the pretrained VGGish [11] to get audio features $\mathbf{R^a} = \{r_0^a, r_1^a, ..., r_l^a\}$. The superscript v, a, p denote for the different modality. Noting that all the expert

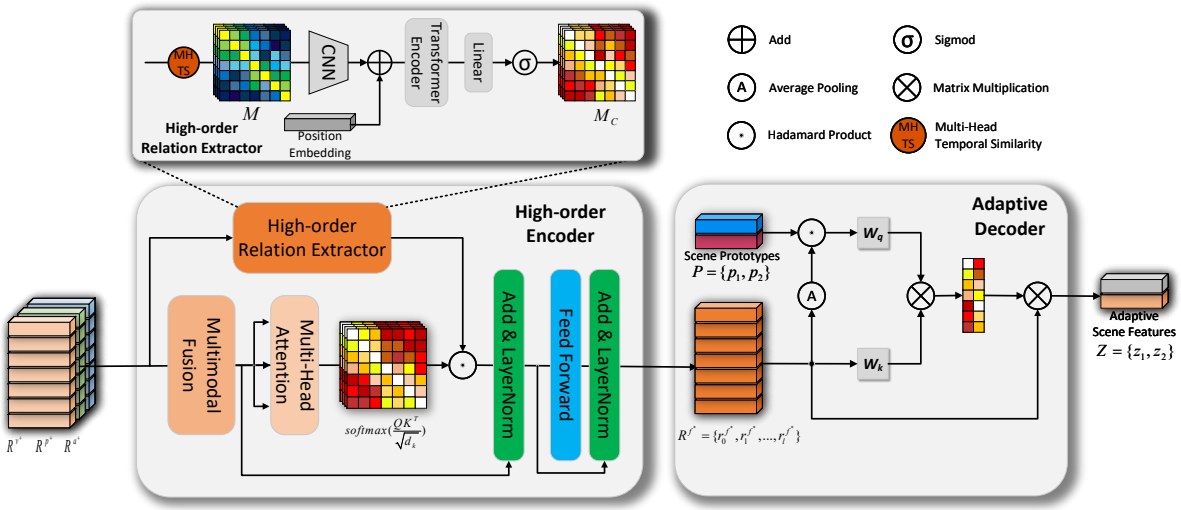


Figure 3: The proposed Multimodal High-order Relation Transformer is designed to enhance shot features with high-order context and discover the switch pattern between scenes. Specifically, the high-order encoder can fuse multimodal cues and model context for each shot under the guidance of high-order relation extractor. Additionally, the subsequent decoder adaptively aggregate shots in the same scene and captures the switch pattern between scenes. Best viewed with colors.

networks are frozen during the training process.

Aiming to enhancing shot features with high-order context and discovering the switch pattern between scenes, we design a transformer-based architecture. As shown in Figure 3, it consists of a high-order encoder and an adaptive decoder. The encoder is applied to fuse complex multimodal cues and model context for each shot with the help of high-order relation extractor, while the decoder is responsible for adaptively aggregating shots in the same scene.

Clip Encoder. Following [23], we add a clip encoder on a local window for each shot's representations, aiming to obtain their clip-level representations. Specifically, taking the visual feature r_i^v of shot i for example, we consider the differences $\mathbf{S_D}$ and relations $\mathbf{S_R}$ of its nearby shots in a window with length $2w_s$, i.e. $\mathbf{V}_i = \left\{r^v_{i-(w_s-1)}, ..., r^v_i, ..., r^v_{i+w_s}\right\}$. This process can be performed by temporal convolution layers Ψ and fully connected layers $\mathbf{FC}$, as shown in Equation (1):

$$\begin{aligned}
\mathbf{S_D}(\mathbf{V}_i) &= \Psi(\{r^v_{i-(w_s-1)}, ..., r^v_i\}) \cdot \Psi(\{r^v_{i+1}, ..., r^v_{i+w_s}\}), \\
\mathbf{S_R}(\mathbf{V}_i) &= \Psi(\{r^v_{i-(w_s-1)}, ..., r^v_{i+w_s}\}), \\
r_i^{v^+} &= \mathbf{FC}(concat(\mathbf{S_D}(\mathbf{V}_i), \mathbf{S_R}(\mathbf{V}_i))),
\end{aligned} \tag{1}$$

where $concat$ denotes the concatenation operation, $\cdot$ denotes the inner product. Similar operations are performed on the place and audio modality, and the updated shot features can be described by $\mathbf{R^{v^+}} = \{r_0^{v^+}, r_1^{v^+}, ..., r_l^{v^+}\}$, $\mathbf{R^{p^+}} = \{r_0^{p^+}, r_1^{p^+}, ..., r_l^{p^+}\}$, $\mathbf{R^{a^+}} = \{r_0^{a^+}, r_1^{a^+}, ..., r_l^{a^+}\}$. Consequently, these multimodal shot features are enhanced with short-term context.

3.2. Multimodal High-order Context Modeling

Aiming to enhancing shot features with high-order context and discovering the switch pattern between scenes, we design a transformer-based architecture. As shown in Figure 3, it consists of a high-order encoder and an adaptive decoder. The encoder is applied to fuse complex multimodal cues and model context for each shot with the help of high-order relation extractor, while the decoder is responsible for adaptively aggregating shots in the same scene.

High-order Relation Extractor. As discussed above, directly adopting the standard transformer encoder [35] over the shot sequence features may introduce noises in the correlation map due to imperfect shot features, and ignore comprehensive 'high-order relation' signal. Thus, we innovatively design the high-order relation extractor to correct first-order relations and fuse the multimodal cues jointly. As shown in the upper part of Figure 3, taking the multimodality shot sequence features $\mathbf{R^{v^+}}, \mathbf{R^{p^+}}, \mathbf{R^{a^+}}$ as inputs, we first calculate the multi-head based temporal similarity matrix for each modality, generating a first-order multimodal relation map $\mathbf{M} \in \mathcal{R}^{(l+1)\times(l+1)\times 3h}$, which naturally contains temporal positional signal. Note that $l+1$ is the length of shot sequence and h is the number of heads. Then a shallow CNN is applied on $\mathbf{M}$ to fuse multimodal cues and model neighboring first-order relations. And one standard transformer layer followed by linear layers and sigmoid function is adopted to model the global associations of each shot's relation, producing the guidance map $\mathbf{M_C}$.

High-order Encoder for Context Modeling. As shown in the bottom left part of Figure 3, we fuse multimodal fea-

tures with concatenation operation and a simple linear layer:

$$r_i^f = \mathbf{FC}(concat(r_i^{v^+}, r_i^{p^+}, r_i^{a^+})) \in \mathcal{R}^{d_f}, \quad (2)$$

producing fused shot sequence features $\mathbf{R}^f \in \mathcal{R}^{(l+1)\times d_f}$, where the superscript f denotes for fusion. Then we apply the guidance map $\mathbf{M_C}$ to enhance context modeling in a conventional transformer encoder. In detail, we first generate transformed queries and key-values pairs by: $\mathbf{Q} = \mathbf{R}^f\mathbf{W}^Q, \mathbf{K} = \mathbf{R}^f\mathbf{W}^K, \mathbf{V} = \mathbf{R}^f\mathbf{W}^V$, where $\mathbf{W}^Q \in \mathcal{R}^{d_f\times d_k}, \mathbf{W}^K \in \mathcal{R}^{d_f\times d_k}, \mathbf{W}^V \in \mathcal{R}^{d_f\times d_v}$. Noting that we omit the multi-head scripts to simplify the expression here. Then the modified attention process $\mathbf{HoAttn}(\mathbf{R}_f)$, under the guidance from the high-order multimodal relation cues $\mathbf{M_C}$, can be described by:

$$\mathbf{HoAttn}(\mathbf{R}^f) = (\mathbf{M}_C \odot softmax(\frac{\mathbf{QK}^T}{\sqrt{d_k}}))\mathbf{V}, \quad (3)$$

where $\odot$ denotes Hadamard product, indicating $\mathbf{M}_C$ works as a rectifier to enhance the appropriate correlations and suppress the erroneous ones. The residual connections, normalization operation and feed-forward layer are concatenated after this attention block, as shown in Figure 3. As a result, the high-order encoder can produce shot representations $\mathbf{R}^{f^*} = \{r_0^{f^*}, r_1^{f^*}, ..., r_l^{f^*}\}$. And $\mathbf{R}^{f^*}$ contains comprehensive multi-modality contextual information. The above two parts fit together as our High-order Encoder and can serve as a basic block, offering a flexible high-order relation perception function.

Adaptive Decoder for Shot Clustering. Although the high-order encoder successfully models both multimodal cues and high-order relations into the shot representations, it is still challenging to detect scene boundaries. Because even appearances of 'boundary shots' within the same video can vary dramatically. It is essential to learn the switch pattern between scenes, rather than simple shot features. Consequently, we design an adaptive decoder to dynamically merge shots within the same scene, resulting in aggregated contextual scene features. Our decoder also follows the query-key-value design in the vanilla transformer [35], making the full model unified.

As shown in the right part of Figure 3, our adaptive decoder consists of a pair of learnable scene prototypes $\mathbf{P} = \{p_1, p_2\}$ and an attention block. In order to reduce the domain gap between scene prototypes and shot features, we first employ the average pooled feature of all shots to modify the scene prototypes. The attention interaction between the shots and scene prototypes can be formulated by:

$$\begin{aligned} &\mathbf{P}^* = \mathbf{P} \odot avg(\mathbf{R}^{f^*}), \\ &\mathbf{A}_S = (\mathbf{P}^*\mathbf{W}_q)(\mathbf{R}^{f^*}\mathbf{W}_k)^T, \\ &\mathbf{Z} = \{z_1, z_2\} = \mathbf{A}_S\mathbf{R}^{f^*}, \end{aligned} \quad (4)$$

where $\mathbf{W}_q \in \mathcal{R}^{d_s\times d_q}, \mathbf{W}_k \in \mathcal{R}^{d_f\times d_q}$ are learnable parameters. The affinity matrix $\mathbf{A}_S \in \mathcal{R}^{2\times(l+1)}$ measures the relevance between each prototype and the shots, which directs the aggregation of shots into scenes. And $\{z_1, z_2\}$ indicate the dynamically aggregated scene representations, which is used to regress the scene boundary positions.

3.3. Objective Function and Data Augmentation

Based on above discuss, we learn a set of shot representations $\mathbf{R}^{f^*} = \{r_0^{f^*}, r_1^{f^*}, ..., r_l^{f^*}\}$ and two adaptively merged scene features $\mathbf{Z} = \{z_1, z_2\}$ for the input shot sequence $\{s_0, s_1, ..., s_l\}$. For each shot sequence, we make prediction on the central shot $r_{\lfloor l/2\rfloor}^{f^*}$, i.e. determining whether it's a scene boundary or not. The subscript $\lfloor l/2\rfloor$ stands for the floor function. The classification loss can be formulated as:

$$\mathbf{L}_{pre} = -[y_c log(\phi(r_{\lfloor l/2\rfloor}^{f^*})) + (1-y_c) log(\phi(r_{\lfloor l/2\rfloor}^{f^*}))], \quad (5)$$

where y_c is the label for the central shot, deciding whether it's a scene boundary, and ϕ is a binary classification head. Moreover, ordering to learn adaptive scene representations, we adopt the adaptively merged scene features $\mathbf{Z} = \{z_1, z_2\}$ and a regression head γ to regress the position of the boundary in the shot sequence:

$$\mathbf{L}_{reg} = \mathbf{MSE}[p, \gamma(concat(z_1, z_2))], \quad (6)$$

where p is the ground-truth position of the scene boundary in the shot sequence, and $p = -1$ or $p = l + 1$ stands for there are no boundaries or multiple boundaries in the sequence. $\mathbf{MSE}$ denotes the mean squared error. The overall training loss can be $\mathbf{L} = \mathbf{L}_{pre} + \alpha\mathbf{L}_{reg}$, where α is a hyperparameter.

Moreover, there is a serious class-imbalance problem in the scene detection task, because boundaries are rare in a long video. Thus, we construct shot sequence samples with positive boundary on the central position by concatenating two distant subsequences in the same video or two subsequences from different videos:

$$\{s_0, ..., s_{l/2}, ..., s_l\}_{new} = \{s_0, ..., s_{l/2}\}_{v_0} + \{s_{l/2}, ..., s_l\}_{v_1}. \quad (7)$$

The ratio of constructed samples to origin samples is adjustable according to the data distribution.

4. Experimental Results

4.1. Experimental Setup

Dataset. To demonstrate the effectiveness of our proposed method, we carry out extensive experiments on three public available datasets, including MovieNet-SSeg [12], BBC planet earth [1] and OVSD [27]. MovieNet [12], containing 1,100 videos with 1.6M shots, is the largest dataset for video understanding analysis by far. And the MovieNet-SSeg is a sub-dataset of that, with 190, 64 and 64 videos labeled with scene boundaries for training, validating and

Table 1: Results on the MovieNetSSeg dataset. The best results are in bold, and the second best are in bold and underlined. Note that self-supervised methods are pretrained with 1100 videos in MovieNet (***MS-1100***), and fine-tuned with 190 training videos in MovieSSeg (***MS-190***). Using less than 1/5 of the training data, our model can achieve competitive results.

Method	Training Data	AP (↑)	mIoU (↑)	AUC-ROC (↑)	F1 (↑)
Supervised Learning					
Siamese [1]	***MS-190***	35.8	39.6	-	-
MS-LSTM [12]	***MS-190***	46.5	46.2	-	-
LGSS [23]	***MS-190***	47.1	48.8	-	-
Ours	***MS-190***	**<u>54.8</u>**	**51.2**	**90.3**	**<u>46.3</u>**
Unsupervised Learning					
GraphCut [25]	***None***	14.1	29.7	-	-
SCSA [3]	***None***	14.7	30.5	-	-
DP [9]	***None***	15.5	32.0	-	-
Story Graph [33]	***None***	25.1	35.7	-	-
Grouping [27]	***None***	33.6	37.2	-	-
Self-supervised Learning					
ShotCoL [4]	***MS-1100***	53.4	-	-	-
SCRL [38]	***MS-1100***	**<u>54.8</u>**	-	-	**51.4**
BaSSL [20]	***MS-1100***	**56.3**	**<u>49.5</u>**	**90.3**	45.7

testing. Besides, we also carry out experiments on BBC planet earth [1] and OVSD [27] dataset. BBC planet earth contains 11 documentaries and OVSD contains 21 various videos. And we just take them as the testing set to verify the generalization of our model, i.e. our model trained on MovieScenes-SSeg is directly tested them, because of their limited scale and unavailable splitting strategy.

Evaluation Metrics. We make comprehensive validations on our model with several metrics: (1) Average Precision($\mathbf{AP}$), it's the mean of $\mathbf{AP}$ of $y_c = 1$ for each video. (2) $\mathbf{mIoU}$, the averaged intersection over union (IoU) between detected scene boundary with their closest ground-truth scene boundary. (3) $\mathbf{F1}$ score, the harmonic mean of the precision and recall, taking both recall and precision into consideration. (4) $\mathbf{AUC-ROC}$, short for area under the receiver operating characteristics.

Implementation Details. The proposed High-Order Relation Aware Multi-modality Network is implemented in PyTorch framework [21] with NVIDIA V100 GPUs.

To begin with, we set the shot sequence length as 7, which is able to contain enough contextual information while requiring relative low memory resource. In the clip-level shot embedding stage, we only take visual, place and audio modalities mentioned in [23], as they play the most crucial roles in scene boundary detection. In specific, we use ViT-B/16 pretrained on ImageNet [6], ResNet50 pretrained on Place dataset [40], and VGGish pretrained on AVA-ActivaSpeaker [26] to extract multimodal shot features with 768, 2048, 128 dimensions respectively. All of them were frozen during training. The clip encoder employs the 1D-CNN [16] and the FC layer to transform them into 384, 256, 128 dimensional vectors. This indicates the importance of each modality, which will be verified in the ablation studies. In the followed multimodal high-order context modeling stage, the high-order relation extractor first maps the multimodal features as the correlation map $\mathbf{M}$ with 8 heads. Feeding $\mathbf{M}$ into the shallow CNN with 4 layers, one transformer encoder layer and the followed linear layer, we get the high-order relation-aware guidance map $\mathbf{M_C}$. We set the corresponding high-order context modeling encoder with 8 heads, 1 layer, and hidden units as 768. Finally, in the Adaptive Decoder, the dimensions of the merged scene features and the shot features remain the same. As a result, the scale of learnable parameters in our model is approximately equivalent to the sum of two layers of transformer encoder, four layers of CNN and several linear layers, which is comparable to that of BaSSL [20] without its' backbone.

Table 2: Results of AP on BBC planet earth and OVSD datasets. Performances are achieved by the model trained on the MovieNetSSeg dataset without any fine-tuning.

Method	BBC	OVSD
SimCLR(instance) [4]	32.3	25.5
SimCLR(temperal) [4]	34.2	24.9
SimCLR(NN) [4]	32.9	25.0
BaSSL(More Training Data) [20]	**40.0**	**28.7**
Ours	**<u>38.7</u>**	**<u>27.5</u>**

The model is trained for 10 epochs with the Adam optimizer [17]. We start training with a learning rate of 0.0002, and decay the learning rate using a cosine schedule. The batch-size is set to 128 for all experiments. And the hyper-parameter α in loss is set to 0.2. At last, for evaluation on

the testing set, we tackle the over-fitting by choosing the snapshot of the model based on the validation set. Noting that during inference, for each shot sequence we only make prediction on the central shot. Only if the classification head predicts to be positive, or the regression head outputs the central position, we take it as a positive boundary.

4.2. Performance Comparison

We compare with several recent state-of-the-art methods on the MovieNet-SSeg [12] in Table 1. And verifying the generation ability of our method by carrying out experiments on BBC planet earth [1] and OVSD [27] datasets in Table 2. We can find that our proposed High-Order Relation Aware Multi-modality Network achieves much better performance compared with the supervised methods, and can be competitive with the self-supervised ones while consuming much less training resources.

Results on MovieNet-SSeg. Table 1 presents the quantitative results on MovieNet-SSeg. We compare our method with various types of methods, including supervised, unsupervised and self-supervised approaches. The results demonstrate that under the same settings, our methods can outperform previous works by a large margin and can be even competitive with those which employ more than 5 times training data. Particularly, our method achieves 54.8 for $\mathbf{AP}$, 51.2 for $\mathbf{mIoU}$, 90.3 for $\mathbf{AUC-ROC}$ and 46.3 for $\mathbf{F1}$. Noting that our method can offer a much more convenient way for training and testing compared with BaSSL, ShotCoL and SCRL [20, 4, 38]. For fair comparison, we also conduct experiments using identical multi-modality settings, specifically by the same features generated by expert networks. As shown in Tabel 3, using less external knowledge, our approach can still outperform LGSS by a large margin. Different from previous works which neglect the high-order relations among the shot sequence and only focus on learning shot features, our method jointly models both of them in a unified deep model.

Results on BBC planet earth and OVSD. As shown in Table 2, we further validate performance of our method on BBC planet earth [1] and OVSD [27]. The training and testing splits are not available and the scale of dataset size are quite small (11 for BBC, and 21 for OVSD respectively), in addition, 2 out of 21 videos in OVSD is not available. Consequently, we just directly employ the model trained on MovieNetSSeg without any fine-tune, and take inference on these two datasets. SimCLR(x) corresponds to ShotCoL [4] using different contrastive learning strategies reproduced by [20]. The results demonstrate advanced generalization ability of our methods. It's competitive with shot-level pretraining methods.

4.3. Ablation Studies and Analysis

To carry out analysis and evaluate the contribution of each component and setting in our method, we conduct a series of ablation studies. All the following comparative experiments are conducted on the MovieNetSSeg dataset.

Table 4: Impact of multimodal semantics, where three elements are studied including visual, place, and audio. † denotes that the results are copied from [23].

Method	place	audio	visual	AP (↑)
Grouping [27]†	✓	✓	✓	23.8
StoryGraph [33]†	✓	✓	✓	33.2
Siamese [1]†	✓	✓	✓	34.1
LGSS [23]	✓	✓	✓	47.1
Ours	✓			50.9
Ours		✓		16.3
Ours			✓	52.4
Ours	✓	✓		48.4
Ours	✓		✓	53.4
Ours		✓	✓	53.1
Ours	✓	✓	✓	54.8

Impact of Multimodal Semantics. First of all, we revisit the impact of each modality in Table 4. We utilize the whole pipeline with clip-level shot embedding and multimodal high-order context modeling to make predictions on the testing set. Experiment starts from the model using only one modality features, and gradually adds different modality cues. Observing the second block of Table 4, we get the conclusion that visual modality plays the most important role in video scene detection. It's hard for single audio features to obtain satisfied results. From the third part of Table 4, we learn that both audio and place modalities can give complementary cues to the visual modality, and improves the final results. In particularly, audio improves 2.7 in terms of $\mathbf{AP}$, place improves 3.5 for $\mathbf{AP}$. Combining three modalities can produce the best result, i.e. 54.8 for $\mathbf{AP}$. Compared with previous works in Table 4, our method gets the best results. Because our approach takes a more comprehensive way to fuse the multimodal features and drive them to complement with each other in a unified model. This can prove the hypothesis we presented above, i.e. multimodal content can complement for each other in the task of scene boundary detection.

Impact of Different Components. We take experiments to validate the effect of different components in our method, including the clip encoder, the high-order encoder and adaptive decoder in our multimodal high-order relation transformer. Specifically, taking the full model as baseline, we remove the clip encoder which is denoted as 'w/o CE'. Removing the high-order relation extractor can verify the effectiveness of modeling high-order context, and we denote it as 'w/o HR', i.e. without high-order relation guidance. And the model without the full high-order encoder in the multimodal high-order relation transformer is denoted as 'w/o CR' for 'without context relation', as it lacks the context relationship modeling. And the model without the

Table 3: Results on the MovieNetSSeg dataset with different multimodal semantics settings, which demonstrate the effectiveness of our approach under fair conditions.

Method	*LGSS (audio)*	*LGSS (place)*	*LGSS (audio+place)*	*LGSS (audio+place +act+cast)*	*Ours (audio)*	*Ours (place)*	*Ours (audio+place)*	*Ours (audio+place +visual)*
AP(↑)	17.5	39.0	43.4	47.1	16.3	50.9	48.4	**54.8**

Table 5: Impact of different components. We take the full model as baseline.

Modality	AP (↑)	mIoU (↑)	AUC-ROC (↑)	F1 (↑)
w/o CE	52.3	48.3	90.0	45.5
w/o HR	51.1	50.3	88.4	44.4
w/o CR	48.6	45.9	88.6	45.2
w/o Cluster	54.3	50.9	90.3	46.2
Full Model	54.8	51.2	90.3	46.3

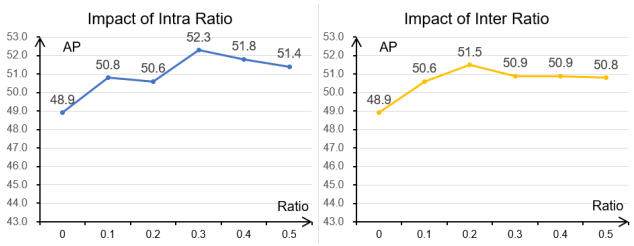


Figure 4: Impact of different data augmentation strategies.

proposed adaptive decoder is formulated as 'w/o Cluster', which does not need to cluster shots within the same scene. As shown in the first line of Table 5, removing the clip encoder leads to 2.5 drop on **AP**, and 0.8 on **F1**. If the model taken out the high-order relation extractor **AP** drops 3.7. Furthermore, if we remove the high-order encoder, the performance can degrade a lot, i.e. 6.2 on **AP**, 5.3 on **mIoU**. This demonstrates that modeling context and high-order relationships among the video shots is of great significance for the video scene detection task. Besides, comparing the last two lines in Table 5 can prove the necessity of clustering shots within the same scene. Lacking the guidance from scene-level representations, the results on **AP** drops 0.5, and **mIoU** drops 0.3.

Impact of Data Augmentation. Figure 4 demonstrates the impact of different data augmentations. We add shot sequence samples from one video to another video with positive boundary on the central position, to make the model to learn robust scene boundary pattern. If the two videos are from the same video, we call this augmentation method as 'Intra'. Otherwise, as 'Inter'. Hyperparameter ratio controls the rate of newly constructed samples. When the ratio equals zero, we only adopt shot sequence samples by sliding window on every training video without any data augmentation. Having appropriate numerical values for ratio is crucial since setting it too low can prevent the model from acquiring positive samples, while setting it too high can lead to overfitting. Optimal outcome is obtained with a ratio of 0.3 for 'Intra' augmentation and 0.2 for 'Inter' augmentation. Therefore, we incorporate these two hyperparameters in our complete model. Figure 4 shows that both of 'Intra' and 'Inter' can improve the performance, while 'Intra' is better. The reason may be that 'Intra' helps constructing harder samples.

Qualitative Analysis. In this part, we visualize similarity matrixes of 50 consecutive shot representations from multimodal features, which can intuitively prove the importance of multimodal cues in the task of scene boundary detection. Specifically, the models are trained with different multimodal semantic settings, and output multimodal shot features respectively. Then we visualize the matrix of cosine similarity between them. The ground-truth labels of scenes (i.e. shot clusters) are marked by red rectangles. It is obvious in Figure 5, shots are clearly clustered into scenes by multimodal features, i.e. the similarities between the shots before and after the boundaries are quite small. Besides, we can see that audio features can hardly separate the shots into different scenes, and the similarity matrix of place or visual features can roughly cluster the shots into scenes near the ground-truth. For single audio or place features, the similarities around boundaries are smooth, which may lead to confusing boundary detection results. Thus, we believe that audio, place and visual features can complement with each other to achieve better performance.

5. Conclusion

In this paper, we propose a unified multimodal framework for video scene boundary detection. And we design a novel Multimodal High-order Relation Transformer, which can model multimodal cues, high-order relation and scene adaptive clustering in a unified structure. We extract multimodal shot representations and model their clip-level context using expert networks and a clip encoder. Then we use a high-order encoder to uncover complex associations among the features and exploit contextual semantics. Then an adaptive decoder in our transformer is proposed to dynamically merge shots in the same scene, which is effective at discovering switch pattern between scenes rather than visual appearance. We have systematically studied the influence of our idea and carried out experiments. The results demonstrate the effectiveness of our model by achieving significant performance.

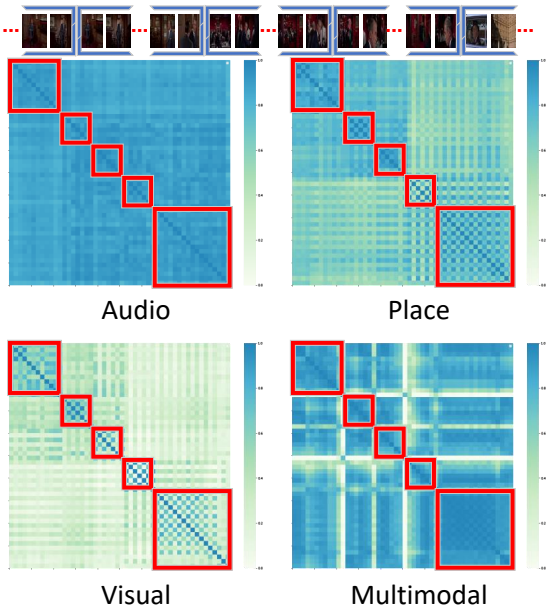


Figure 5: Visualization of similarity matrixes between shot representations in randomly sampled 50 consecutive shots. The ground-truth labels of scenes (shot clusters) are marked by red rectangles.

6. Acknowledgment

This work was partially supported by National Nature Science Foundation of China (Grant 62022078, 62121002) and National Defense Basic Scientific Research Program (2021601B013).

References

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

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

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

[4] Shixing Chen, Xiaohan Nie, David Fan, Dongqing Zhang, Vimal Bhat, and Raffay Hamid. Shot contrastive self-supervised learning for scene boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9796–9805, 2021.

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Zexing Du, Xue Wang, Guoqing Zhou, and Qing Wang. Fast and unsupervised action boundary detection for action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3323–3332, 2022.

[9] Bo Han and Weiguo Wu. Video scene segmentation using a novel boundary evaluation criterion and dynamic programming. In *2011 IEEE International conference on multimedia and expo*, pages 1–6. IEEE, 2011.

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

[11] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE, 2017.

[12] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *European Conference on Computer Vision*, pages 709–727. Springer, 2020.

[13] Hyesung Ji, Danial Hooshyar, Kuekyeng Kim, and Heuiseok Lim. A semantic-based video scene segmentation using a deep neural network. *Journal of Information Science*, 45(6):833–844, 2019.

[14] Xuekun Jiang, Libiao Jin, Anyi Rao, Linning Xu, and Dahua Lin. Jointly learning the attributes and composition of shots for boundary detection in videos. *IEEE Transactions on Multimedia*, 2021.

[15] Hyolim Kang, Jinwoo Kim, Taehyun Kim, and Seon Joo Kim. Uboco: Unsupervised boundary contrastive learning for generic event boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20073–20082, 2022.

[16] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[18] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019.

[19] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018.

[20] Jonghwan Mun, Minchul Shin, Gunsoo Han, Sangho Lee, Seongsu Ha, Joonseok Lee, and Eun-Sol Kim. Boundary-aware self-supervised learning for video scene segmentation. *arXiv preprint arXiv:2201.05277*, 2022.

[21] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[22] Stanislav Protasov, Adil Mehmood Khan, Konstantin Sozykin, and Muhammad Ahmad. Using deep features for video scene detection and annotation. *Signal, Image and Video Processing*, 12(5):991–999, 2018.

[23] Anyi Rao, Linning Xu, Yu Xiong, Guodong Xu, Qingqiu Huang, Bolei Zhou, and Dahua Lin. A local-to-global approach to multi-modal movie scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10146–10155, 2020.

[24] Zeeshan Rasheed and Mubarak Shah. Scene detection in hollywood movies and tv shows. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–343. IEEE, 2003.

[25] Zeeshan Rasheed and Mubarak Shah. Detection and representation of scenes in videos. *IEEE transactions on Multimedia*, 7(6):1097–1105, 2005.

[26] Joseph Roth, Sourish Chaudhuri, Ondrej Klejch, Radhika Marvin, Andrew Gallagher, Liat Kaver, Sharadh Ramaswamy, Arkadiusz Stopczynski, Cordelia Schmid, Zhonghua Xi, et al. Supplementary material: Ava-activespeaker: An audio-visual dataset for active speaker detection. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3718–3722. IEEE, 2019.

[27] Daniel Rotman, Dror Porat, and Gal Ashour. Optimal sequential grouping for robust video scene detection using multiple modalities. *International Journal of Semantic Computing*, 11(02):193–208, 2017.

[28] Daniel Rotman, Dror Porat, Gal Ashour, and Udi Barzelay. Optimally grouped deep features using normalized cost for video scene detection. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 187–195, 2018.

[29] Yong Rui, Thomas S Huang, and Sharad Mehrotra. Exploring video structure beyond the shots. In *Proceedings. IEEE International Conference on Multimedia Computing and Systems (Cat. No. 98TB100241)*, pages 237–240. IEEE, 1998.

[30] Panagiotis Sidiropoulos, Vasileios Mezaris, Ioannis Kompatsiaris, Hugo Meinedo, Miguel Bugalho, and Isabel Trancoso. Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(8):1163–1177, 2011.

[31] Jing Tan, Yuhong Wang, Gangshan Wu, and Limin Wang. Temporal perceiver: A general architecture for arbitrary boundary detection. *arXiv preprint arXiv:2203.00307*, 2022.

[32] Yunlong Tang, Siting Xu, Teng Wang, Qin Lin, Qinglin Lu, and Feng Zheng. Multi-modal segment assemblage network for ad video editing with importance-coherence reward. In *Proceedings of the Asian Conference on Computer Vision*, pages 3519–3535, 2022.

[33] Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. Storygraphs: visualizing character interactions as a timeline. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 827–834, 2014.

[34] Srinivasan Umesh, Leon Cohen, and D Nelson. Fitting the mel scale. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 1, pages 217–220. IEEE, 1999.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[36] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[37] Zhenzhi Wang, Zhimin Li, Liyu Wu, Jiangfeng Xiong, and Qinglin Lu. Overview of tencent multi-modal ads video understanding. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4725–4729, 2021.

[38] Haoqian Wu, Keyu Chen, Yanan Luo, Ruizhi Qiao, Bo Ren, Haozhe Liu, Weicheng Xie, and Linlin Shen. Scene consistency representation learning for video scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14021–14030, 2022.

[39] Zhongping Zhang, Yiwen Gu, Bryan A Plummer, Xin Miao, Jiayi Liu, and Huayan Wang. Effectively leveraging multi-modal features for movie genre classification. *arXiv preprint arXiv:2203.13281*, 2022.

[40] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.