MEGA: Multimodal Alignment Aggregation and Distillation For Cinematic Video Segmentation

Najmeh Sadoughi[1], Xinyu Li[1], Avijit Vajpayee[1], David Fan[1], Bing Shuai[2],
Hector Santos-Villalobos[1], Vimal Bhat[1], Rohith MV[1]
[1]Amazon Prime Video, [2]AWS AI Labs
{nnnourab,xxnl,avivaj,fandavi,bshuai,hsantosv,vimalb,kurohith}@amazon.com

Abstract

*Previous research has studied the task of segmenting cinematic videos into scenes and into narrative acts. However, these studies have overlooked the essential task of multimodal alignment and fusion for effectively and efficiently processing long-form videos (> 60min). In this paper, we introduce **M**ultimodal alignm**E**nt a**G**gregation and distill**A**tion (MEGA) for cinematic long-video segmentation. MEGA tackles the challenge by leveraging multiple media modalities. The method coarsely aligns inputs of variable lengths and different modalities with alignment positional encoding. To maintain temporal synchronization while reducing computation, we further introduce an enhanced bottleneck fusion layer which uses temporal alignment. Additionally, MEGA employs a novel contrastive loss to synchronize and transfer labels across modalities, enabling act segmentation from labeled synopsis sentences on video shots. Our experimental results show that MEGA outperforms state-of-the-art methods on MovieNet dataset for scene segmentation (with an Average Precision improvement of +1.19%) and on TRIPOD dataset for act segmentation (with a Total Agreement improvement of +5.51%).*

1. Introduction

In the world of video production, movies are composed of smaller units called shots, scenes, and acts. Shots are a continuous set of frames, a scene is a sequence of shots that tell a story, and an act is a thematic section of a narrative [14]. While computer vision has made significant strides in shot detection [39], scene and act segmentation remain a challenge, despite their potential for smart video navigation, advertisement insertion, and movie summarization. Cinematic content comprises of different data sources, including audio, visual, and text data, as well as derivative data sources from the narrative, including location, appearance, tone, or acoustic events. In this work, we will refer to all of these input components as “modalities” of cinematic content. Previous work has not fully explored how to align and aggregate these modalities which have different granularities.

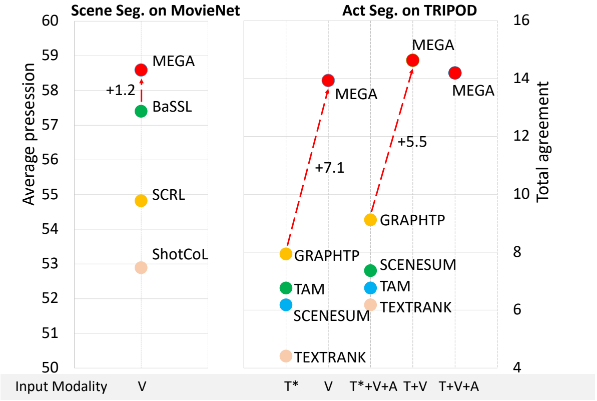

Figure 1: MEGA works well on both scene segmentation and act segmentation tasks, outperforming previous work with significant margin. V,T*,T,A denotes video, screenplay, subtitle and audio respectively.

We propose to address scene and act segmentation tasks with an unified multimodal Transformer. However, this approach presents two main challenges. Firstly, there is the issue of cross modality information synchronization and fusion at the shot level. Previous studies which use multimodal fusion for scene and act segmentation perform early [8, 32] or late fusion of features [35], and have not explored fusion strategies which utilize multimodal temporal alignment. Additionally, the fusion strategies that utilize temporal alignment such as merged attention or cross modality attention [9, 21] are computationally expensive and not generalizable to a large number of modalities. Secondly, due to the challenges associated with labeling a long video on act segmentation, the labels for act segmentation are provided on synopsis sentences [30] which do not provide timestamps. To avoid the more challenging task of cross-modal synchronization, previous studies on act segmentation [30, 32] rely on textual screenplay to transfer the labels from synopsis to movie, ignoring the rich multimodal information from the video, and introducing an additional dependency on screenplay data which is not always available.

To address these challenges, we introduce **M**ultimodal alignm**E**nt a**G**gregation and distill**A**tion (MEGA). MEGA includes a novel module called *alignment positional encoding* which aligns inputs of variable lengths and different modalities at a coarse temporal level. To fuse the aligned embeddings of different modalities in an efficient and effective manner, we adopt the bottleneck fusion tokens [28] and append a set of fusion tokens to each modality. These tokens share the same sequence length as the normalized positional encoding for different modalities, allowing us to inject them with the coarse temporal information, enabling information fusion in a better aligned embedding space. To address the issue of cross-domain knowledge transfer, we introduce a cross-modal synchronization approach. This method allows us to transfer the manually labeled act boundaries from synopsis level to movie level using rich multimodal information, enabling us to train MEGA directly on videos without relying on screenplay – which was a hard requirement for previous works [30, 32].

We test our proposed alignment and aggregation modules on the Movienet-318 [18] and the TRIPOD datasets [30], and we test our cross modality synchronization module on TRIPOD alone, as the labels are provided on a different modality during training. Our proposed MEGA outperforms previous SoTA on scene segmentation on the Movienet-318 dataset (by +1.19% in AP) and on act segmentation on the TRIPOD dataset (by +5.51% in TA). Our contributions are:

1. Alignment positional encoding module and a fusion bottleneck layer that performs multimodal fusion with aligned multi-modality inputs.
2. A cross-domain knowledge transfer module that synchronizes features across-domain, and enables knowledge distillation without requiring extra information.
3. SoTA performance on scene and act segmentation tasks, with detailed ablations, which can be used as reference for future work.

2. Related Work

Scene Segmentation in Cinematic Content: Recent works on scene segmentation have explored self-supervised learning (SSL) [8, 27, 45]. Self-supervised pretext tasks have included maximizing the similarity between nearby shots compared to randomly selected ones [8], maximizing the similarity between pairs of images selected according to scene consistency [45], and maximizing the similarity between pairs of images selected according to pseudo scene boundaries [27]. While several previous works have used multimodal inputs for this task [8, 35, 45], they have either utilized late fusion of features with predefined weights for each modality [35] or have utilized early integration of features derived via SSL [8, 45]. In this paper, we explore how to better align and integrate features from different modalities for scene segmentation.

Act Segmentation in Cinematic Content: Movie screenplays follow a general narrative structure on how a plot unfolds across the story. Several theories have been proposed in this domain dating as far back as Aristotle, who defined a 3 act structure with a beginning (protasis), middle (epitasis), and end (catastrophe) [33]. Modern screenplays are usually written with a 6 act structure [14], named as "the setup", "the new situation", "progress", "complications and higher stakes", "the final push", and "the aftermath", separated by five turning points (TPs). Prior approaches in narrative segmentation on movies have adopted the aforementioned 6 act structure and posed the problem as identifying the 5 TPs that separate these 6 acts. [32] is to our knowledge the only prior work that utilizes visual domain in act segmentation by using a pre-trained teacher model trained on textual input to train a student Graph Convolutional network with audio-visual-text representations as input. In contrast, our work uses a new multimodal fusion and distillation applied on the modalities which are available with the movie.

Multimodal Aggregation: Previous SoTAs on multimodal fusion with transformers perform early fusion of the features as inputs to the transformer [19], merge attention between them requiring more memory [9, 21], use cross attention between two modalities [9, 44], or add cross-modal interactions more efficiently via bottleneck tokens or exchanging some of their latents [15, 28]. [28] provides information flow between modalities efficiently by utilizing bottleneck tokens to tame the quadratic complexity of pairwise attention. This global exchange between modalities may not be enough for long videos, which require an adaptive fusion in different temporal locations. Our model considers [28] as baseline and extends it to incorporate local information during information exchange between modalities.

Positional Encoding: Previous studies on improving the positional encoding in long sequence modeling have mostly focused on adding relative distance positional encoding [24, 38, 40]. However, they do not offer solutions on better maintaining the relative position of latent tokens with respect to their starting point in time in long sequences with variable lengths, which is important for long movie narrative understanding [5, 14]. We propose Alignment Positional Encoding to bridge this gap.

Cross-Modality Synchronization & Distillation: To transfer the labels from synopsis to movie shots, we use cross-modality distillation. Previous cross-modality distillation studies for label transfer across modalities are focused on parallel data with the same granularity [2, 3, 12], or where the alignment is known [31]. Alignment of features at different granularities in the same modality, such as screenplay scenes to synopsis sentences [26, 49] and across modalities such as aligning synopsis sentences to visual shots [42, 47], book chapters to visual scenes [43] have been previously explored. While majority of these works rely on unsupervised optimization techniques [26, 42, 43, 49], there are studies that use supervised labels to improve the representations used for optimization [47]. We present an alignment approach with self-supervised loss for synchronizing data in different modalities of cinematic content to enable distillation.

3. Methodology

MEGA processes long videos and performs video segmentation in three major steps (Fig. 2). First, a video V is chunked into shots, and multiple features such as scene related features and sound event features are extracted at the shot-level (Sec. 3.1). The system is built on shot-level representation for two reasons: (1) scene and act boundaries always align with shot boundaries, and (2) the content within a shot remains similar, which allows efficient yet meaningful sampling without losing important information. Second, the embeddings from different input samples and each modality are coarsely aligned with a proposed alignment embedding (Sec. 3.2), and the alignment positional tokens are used to refine the commonly used bottleneck fusion tokens for cross-modal feature fusion (Sec. 3.2). Third, a linear layer is applied on top of the fused representations to generate scene and act boundaries (Sec. 3.3). Finally, to address the challenge of cross-domain knowledge transfer where labels from one domain may not directly align with another domain (e.g. act labels on synopsis sentences do not have movie-level timing information), we propose a cross-modal synchronization module that is simple yet effective (Sec. 3.4).

3.1. Preprocessing

We chose to utilize Transnet-v2 [39] for shot segmentation due to its superior performance and efficiency. We list our selection of pre-trained models and associated parameters in Tab. 1. As each pretrained feature extraction model has different requirements for input resolution and sampling rate, we first sample the input at various sample rates (as shown in Table 1). It is worth noting that the $\text{CLIP}_{\text{movie}}$ model is the CLIP [34] model with ViT-B/32 backbone fine-tuned on paired IMDB-image/text dataset. IMDB-image dataset comprises 1.6M images from 31.3K unique movies/TV series with 761 unique textual labels. The features attributed to each shot are the ones with overlap with the shot time stamp (More details are in Appendix). After feature extraction, we aggregate the features for each shot and normalize the feature dimension with linear projection as:

$$E_i^m = g^m \left(\frac{1}{T_i^m} \sum_{j=1}^{T_i^m} S_{ij}^m \right) \tag{1}$$

Feature extractor	Input Freq.	Input Res.	Feature dim.
Visual input			
BASSL [27] (bassl)	Varying	224^2	2048
$\text{ResNet}_{\text{Place}}$ [50] (place)	$1Hz$	224^2	2048
ResNeXt101 [46] (appr)	$1Hz$	224^2	2048
I3D [6] (action)	$16Hz$	16×224^2	2048
$\text{CLIP}_{\text{movie}}$ (clip)	$1Hz$	224^2	768
Acoustic input			
PANNs [23] (audio)	$1Hz$	$10 \times 32K$	2048
Linguistic input			
all-MiniLM-L6-v2 [1] (text)	Varying	-	384

Table 1: Sampling strategy and feature extraction backbones used for different modalities.

where $E_i^m \in \mathbb{R}^C$ denotes the embedding from the i-th shot of m-th modality, $S_{ij}^m \in \mathbb{R}^{D^m}$ is the j-th sampled feature of m-th modality from i-th shot, and g^m is a linear projection layer that projects feature dimension for m-th modality to a common dimension C across all modalities.

While it is possible to create an end-to-end system starting from raw shot inputs and training the model from scratch, pre-extracting the features from pretrained models is generally more scalable and efficient in actual industrial scenarios, hence we rely on the latter.

3.2. Cross-modality Alignment and Fusion

For long video segmentation into scenes and acts it is important to model short and long term context and perform effective multimodal fusion. However, the commonly used learnable positional embedding only provides fine-grained granularity and is not suitable for high-level semantic alignment across modalities. Furthermore, for tasks such as act segmentation we expect consistent patterns at normalized position of temporal inputs as the theory suggests approximate locations for each turning point (i.e., 10%, 25%, 50%, 75%, 95% [5, 14, 30]). Hence, in addition to using the traditional positional encoding, we introduce Alignment Positional Encoding layer $\in \mathbb{R}^{L_n \times C}$, which is a learnable embedding layer, for which the index at i-th temporal unit (e.g., shot) is derived by:

$$i_{align} = \text{floor}\left(\frac{L_n}{L} i \right) \tag{2}$$

where L is the temporal dimension (e.g., number of shots) and L_n is the length of alignment positional encoding which is a hyper-parameter ($L_n < L$). We add the alignment positional encoding to the features in conjunction with the conventional positional encoding (see Fig 2). This module is shared across different modalities. This module provides extra information to the network that can be helpful in learning from long training samples with varying lengths, and in coarsely aligning inputs from different modalities before information fusion.

Inspired by previous works [9, 28, 44], we choose cross-modal feature fusion, which has been shown to be more effective than early or late fusion. To make our approach scale to

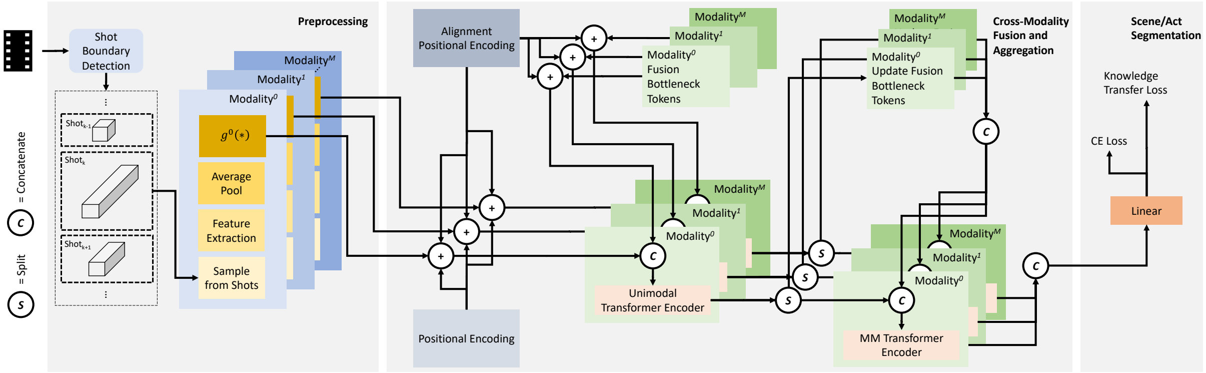

Figure 2: The pipeline of the proposed method includes 1) Preprocessing: splitting the video into shots, extraction of features from each shot, pooling and normalization 2) Cross modality fusion and alignment: with the help of alignment positional encoding and bottleneck fusion tokens, 3) Scene/Act Segmentation comprising the segmentation heads. For Scene Segmentation, CE loss is used and for Act Segmentation knowledge transfer loss is used (refer to Sec. 3.4 for details)

.

multiple modalities, we propose an efficient temporal bottleneck fusion based on [28], and follow their mid-fusion strategy, which comprises of an unimodal transformer encoder followed by a multimodal transformer encoder (See Fig. 2). While [28] proposes to use bottleneck tokens for fusing information across different modalities, these tokens learn to integrate the information across modalities in a global manner. We propose to use L_n fusion tokens, and then integrate them with the same Normalized Positional Encoding to align them with features on the coarse temporal scale (Fig. 3).

The transformer layer per modality then takes in an extra set of aligned fusion tokens concatenated with its input (Fig 2), making it much more efficient compared to other methods such as merged attention or pairwise cross-attention with respect to the number of modalities [9, 21]. Finally, the latent tokens per modality from the last fusion layer (i.e., Z^m for m-th modality) are concatenated as the fused representation:

$$Z^{\text{fused}} = \text{concat}_C\left(Z\right) = \text{concat}_C\left(\left[Z^1, ..., Z^M\right]\right) \tag{3}$$

where $\text{concat}_C(*)$ stands for channel-wise concatenation and M is the number of modalities.

3.3. Scene/Act Segmentation

MEGA adopts a similar approach to previous works [8, 27, 45], where scene segmentation is framed as a binary classification task using a key-shot representation from a window of $2 \times k + 1$ shots (k-shots before and after):

$$y_i = \varphi\left(Z_i^{\text{fused}}; \theta_s\right) \tag{4}$$

where y_i is the logit prediction for the i-th key-shot. φ denotes the linear layer with learnable parameters θ_s and a cross entropy loss is utilized [8, 27, 45].

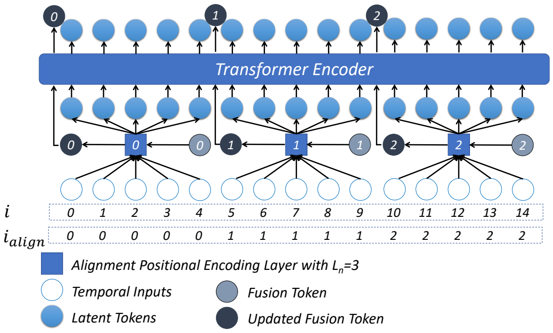

Figure 3: Illustration of Normalized Positional Encoding integration to the temporal tokens in one modality. This figure shows the integration of normalized Positional Encoding with 1) temporal shots for one modality and 2) bottleneck fusion tokens, where $L = 15$ and $L_n = 3$. For (1) i_{align} is obtained per shot index, i, and then each shot is integrated with normalized PE. For (2) each randomly initialized bottleneck token is integrated with normalized PE for its corresponding index.

The act segmentation task is formulated as N_{tp} linear prediction heads (N_{tp} is the number of turning points, and θ_{a_n} denotes the head parameters for n-th turning point [32]), for each individual shot from a temporal model that takes all the shots of the movie as input. To make a prediction at the $i-$th shot for the n-th act boundary ($n \in \{1, \ldots, N_{tp}\}$), we use:

$$y_{in} = \varphi\left(Z_i^{\text{fused}}; \theta_{a_n}\right) \tag{5}$$

where y_{in} is the logit prediction for i-th shot and n-th turning point.

3.4. Cross-domain Knowledge Transfer

It is quite common during machine learning that certain modalities may lack annotations that are directly available in other modalities with different information granularity. To address this, we propose a cross-modality synchronization scheme that enables cross-modal distillation. We utilize this module for act segmentation where we aim to transfer act labels from synopsis sentences to movie-level timestamps. Importantly, our approach does not require additional information, such as screenplay [32], to bridge the gap.

Our knowledge distillation utilizes (1) an individual network to learn the synopsis-based act segmentation in a supervised manner with cross-entropy loss similar to [31]; and (2) a novel synchronization approach between synopsis and movie. For (1) we use the same architecture mentioned in Secs. 3.1, 3.2, setting C_{fused} equal to the multimodal shot model setting. Additionally, similar to shot level linear prediction head (See Eq. 5), we use a sentence level linear prediction head, resulting in q_{in} logits for the i-th sentence of n-th TP. A supervised Cross Entropy loss is used to learn the synopsis labels from predictions for each turning point ($\mathscr{L}_{ce}$). For (2) we seek a synchronization matrix $W \in \mathbb{R}^{L_{sh} \times L_{syn}}$ between L_{sh} shots and L_{syn} synopsis sentences for a sample, where $w_{ij} = 1$ if the i-th shot matches with the j-th synopsis sentence and $w_{ij} = 0$, otherwise. Assuming $F(.;\theta)$ represents a parametric reward function (with parameters θ), to find W, we define an objective as:

$$\max_{W,\theta} \sum_{i,j} w_{ij} F(.;\theta) - \lambda \sum_{i,j} |w_{i,j}| \qquad s.t.\ 0 \leq w_{ij} \leq 1 \tag{6}$$

Expectation-Maximization algorithm is used to solve the objective in Eq. 6. We estimate the target variable W via fixed parameters (i.e., θ) in the E-step, and update the parameters while the target variable is known in the M-step.

E-step: Assuming $F(.;\theta)$ returns the similarity of input shot and synopsis sentence pair, the E-step has a closed form solution. In the E-step, following [26], we reduce the search space during optimization to only the pairs which are inside a diagonal boundary (see the proof for E-step and visualization of expected synchronization matrix for different examples in Appendix).

M-step: Using all samples in a batch with L_{SH}, L_{SYN} total number of shots and synopsis sentences, we form their cosine similarity matrix of dimension $L_{\text{SH}} \times L_{\text{SYN}}$. For each query (synopsis sentence/movie shot, respectively), the positive keys (movie shot/synopsis sentence, respectively) are derived from the expectation step. Negative keys are the keys lying outside the diagonal boundary of the similarity matrix of shot-synopsis pairs for one movie [26], and all the keys from other movies within the batch. Here each query (synopsis sentence/movie shot, respectively) can have more than one positive key (movie shot/synopsis sentence, respectively) attached to it. Following [20], we adopt a modified version of the InfoNCE loss and combine it with the symmetric contrastive loss [34] as:

$$\mathscr{L}_c = -\sum_{i=1}^{L_{\text{SYN}}} \frac{1}{|\hat{y}_i|} \sum_{k=1}^{L_{\text{SH}}} \hat{y}_{ik} \log \frac{exp(v_i u_k/\tau)}{\sum_{j=1}^{L_{\text{SH}}} exp(v_i u_j/\tau)} - \sum_{i=1}^{L_{\text{SH}}} \frac{1}{|\hat{y}_i^T|} \sum_{k=1}^{L_{\text{SYN}}} \hat{y}_{ik}^T \log \frac{exp(u_i v_k/\tau)}{\sum_{j=1}^{L_{\text{SYN}}} exp(u_i v_j/\tau)} \tag{7}$$

where τ is a learnable temperature parameter, and $\hat{y}_{ij}$ is a binary indicator of positive vs. negative pairs, u_i is the normalized feature for the i-th shot and v_i is the normalized feature for the i-th synopsis sentence.

Knowledge distillation: Knowledge distillation is used to transfer the knowledge available for the training samples on synopsis. The predictions from the synopsis model are mapped to shots using a matrix of their similarities as calculated in the maximization step, for each sample. The similarity scores for each shot are normalized along the synopsis sentences with softmax. The logit predictions from synopsis model are transferred to shots by multiplication with the normalized similarity matrix. A softmax along the shots is applied to the transferred logits to derive the probability scores for each shot. Following [31], we use a Kullback–Leibler divergence loss between predicted outputs for each shot and the transferred probabilities ($\mathscr{L}_{kd}$) (More details are provided in Appendix.).

The cross-domain knowledge transfer module can be trained by simply adding the losses together as:

$$\mathscr{L} = \alpha_c \mathscr{L}_c + \alpha_{ce} \mathscr{L}_{ce} + \alpha_{kd} \mathscr{L}_{kd} \tag{8}$$

where α_c, α_{ce}, and α_{kd} are hyperparameters that control the weights of the three losses.

4. Experiments

4.1. Dataset

We test our model on two commonly used dataset: **Movienet-318 [18]:** consists of 1100 movies, out of which 318 movies are annotated for the task of scene segmentation. The annotated dataset is split into 190, 64, and 64 movies for train, validation and test splits, respectively. We report the Average Precision (AP) and F1-score (F1) on the test split following previous work [35, 45].

TRIPOD [32] includes 122 movies, split into 84, and 38 movies for train and test, respectively. This dataset includes the annotations of 6 act boundaries (“the setup”, “the new situation”, “progress”, “complications and higher stakes”, “the final push”, “the aftermath”) on the movie synopsis sentences for the training set, and on the movie screenplay scenes for the test set. The authors also have released soft probability scores (silver annotations) for the training set, using [30][1]. To find the timestamps for the screenplay scenes in the movie, following [32] we used Dynamic Time Warping (DTW) to align the timed subtitles from the movie to the monologue lines in the screenplay. Following [32], we use total agreement (TA), partial agreement (PA) and distance (D) as evaluation metrics.

4.2. Implementation Details

For scene segmentation: We train our model with 8 V100 GPUs with total batch size of 1024. The Adam [22] optimizer is used with learning rate of 1e-4. We train the model for 20 epochs. GeLU [16] is used as activation function by default, we use weighted cross entropy to balance the positive and negative samples at batch level. We choose shot sequence length of 17 ($k = 8$) following [27]. We set $L_n = 2$ for this model.

For act segmentation: We train our model with 4 V100 GPUs with total batch size of 4. The SGD optimizer is used with learning rate of 1e-3. We train the model for 10 epochs. λ in Eq. 6 is empirically set differently for each synopsis sentence of each sample, by finding 99% percentile of the similarity scores between the synopsis sentence and all the shots corresponding to that sample. α_c, α_{ce}, α_{kd} are set to 1, 1, and 10. We set $L_n = 100$ for this the shot model and $L_n = 20$ for the synopsis model. We use max pooling to aggregate the shot-level predictions to scene level. We use all shots from a video for act segmentation.

4.3. Main Results

SoTA on Scene Segmentation. We first show MEGA outperforms previous SoTA on MovieNet318 [27] for scene segmentation (+1.19% on AP and +8.28% on F1). With the same input visual features, MEGA outperforms previous SoTA [27] (Tab. 2 +0.52% on AP and +3.69% on F1), which indicates that the proposed approach is effective. Thanks to the proposed cross-modality fusion module, the MEGA generalizes and benefits from additional information extracted from visual signals. MEGA with 3 visual modalities (clip, place, and bassl) outperforms single modality model by +0.67% on AP and +4.59% on F1, which shows the proposed fusion works as expected. It is worth mentioning that the proposed method is scalable and generalizes to various number of modalities at different scales, which makes it flexible for real-world applications.

SoTA on Act Segmentation. We then show that MEGA establishes the new SoTA performance on act segmentation on TRIPOD dataset (Tab. 3). We first show that MEGA outperforms previous SoTA on TRIPOD [32] dataset With only visual signals as input. Comparing to other works that take textural inputs [29, 32], MEGA is able to achieve better performance. Furthermore, in real-world applications, the visual input (the video) is often easier to obtain than textual inputs such as screenplay [32]. We further show that MEGA* (Tab. 3), which swaps our synchronization module to use similar synchronization as SoTA [32], outperforms GRAPHTP, which demonstrates the effectiveness of proposed approach including the alignment and fusion modules. It is worth mentioning that, with multiple features extracted from **visual** media modality alone, MEGA outperforms previous SOTA which makes the proposed model applicable for real-world scenarios, as it is usually harder to get additional media modalities (e.g., screenplay) which are used in other works.

Approach	*Modality*	*Pretrained on*	*AP* ↑ *[%]*	*F1*↑ *[%]*
Random [35]	-	-	8.2	-
Visual only input				
GraphCut [36]	V	Places [50]	14.1	-
SCSA [7]	V	Places [50]	14.7	-
DP [13]	V	Places [50]	15.5	-
Grouping [37]	V	Places [50]	17.6	-
StoryGraph [41]	V	Places [50]	25.1	-
Siamese [4]	V	Places [50]	28.1	-
LGSS [35]	V	Places [35]	39.0	-
LGSS [35]	V	Cast [17, 48]	15.9	-
LGSS [35]	V	Action [11]	32.1	-
ShotCoL [8]†	V	Movienet [18]	52.89	49.17
SCRL [45]	V	Movienet [18]	54.82	51.43
BaSSL [27]	V	Movienet [18]	57.4	47.02
MEGA	V	Movienet [18]	57.92	50.71
MEGA	V	M+P+I	58.59	55.30

Table 2: Scene boundary detection: comparison with SoTA. †means the numbers are copied from [45]. M+P+I denotes pre-trained on Movienet [18], Places [50] and IMDB.

By further aggregating the results from text input, MEGA establishes the new state-of-the-art on TRIPOD dataset (Tab. 3). MEGA almost doubles the performance of previous work [29, 32] with +5.51% TA, +9.15% PA, -%0.81 D. This shows the proposed multimodal fusion scales and generalizes well to multiple modalities. The cross-modality distillation also works robustly for various settings. We noticed that adding the acoustic features is not always helpful to the performance (Tab. 3), probably because acoustic information provides redundant or not useful information for the task of act boundary segmentation.

4.4. Ablations

We perform ablations to examine the effectiveness of major building blocks of MEGA. We use features extracted from visual media modality (for scene segmentation: clip, place, bassl and for act segmentation: clip, appr, action, place), with the proposed normalized positional encoding, multi-modality bottleneck fusion, and cross-modality synchronization by default unless specified. The training and evaluation follows same protocols as mentioned in Sec. 4.3.

Alignment Positional Encoding. We report the ablations on Alignment PE in Tab. 4a. We notice that the proposed

[1] https://github.com/ppapalampidi/GraphTP

Approach	*Modality*	*Modality for synch.*	*TA↑ [%]*	*PA↑ [%]*	*D↓ [%]*
Random (Evenly dist.) [32]	-	T*	4.82	6.95	12.35
Theory [14, 30]	-	T*	4.41	6.32	11.03
Distribution position [32]	-	T*	5.59	7.37	10.74
Single modality input					
TEXTRANK [25]	T*	T*	6.18	10.00	17.77
SCENESUM [10]	T*	T*	4.41	7.89	16.86
TAM [29]	T*	T*	7.94	9.47	9.42
GRAPHTP [32]	T*	T*	6.76	10.00	9.62
MEGA*	V	T*	10.51	14.54	8.98
MEGA	V	V	13.93	20.72	9.19
Multi-modality input					
TEXTRANK [25]	T*+A+V	T*	6.18	10.00	18.90
SCENESUM [10]	T*+A+V	T*	6.76	11.05	18.93
TAM [29]	T*+A+V	T*	7.36	10.00	10.01
GRAPHTP [32]	T*+A+V	T*	9.12	12.63	9.77
MEGA*	T+V	T*	11.14	15.20	**8.96**
MEGA	T+V	T+V	**14.63**	21.78	**8.96**
MEGA*	T+A+V	T*	10.00	14.08	8.96
MEGA	T+A+V	T+A+V	14.19	**22.10**	9.68

Table 3: TP identification: comparison with SoTA. MEGA* denotes the MEGA using the same synchronization as [30] for fair comparison. T*,V,T,A denote Textual-screenplay, Visual, Textual-subtitle and Acoustic features, respectively.

	Scene Seg.	*Act Segmentation*		
case	*AP↑*	*TA↑*	*PA↑*	*D↓*
w/o align. PE	57.77	5.29	7.37	31.04
w. align. PE	**58.59**	**13.93**	**20.72**	**9.19**

(a) Effectiveness of Alignment Positional Encoding.

		Scene Seg.	*Act Segmentation*		
case	*modality*	*AP↑*	*TA↑*	*PA↑*	*D↓*
w/o align.PE	V	58.31	13.60	20.53	9.47
w. align.PE	V	**58.59**	13.93	20.72	9.19
w/o align.PE	V + T	-	13.01	20.13	9.56
w. align.PE	V + T	-	**14.63**	**21.78**	**8.96**

(b) Effectiveness of Normalized Positional Encoding in bottleneck tokens.

	Scene Seg.	*Act Segmentation*		
MM. integ. type	*AP↑*	*TA↑*	*PA↑*	*D↓*
LateFusion	58.24	12.57	19.21	10.00
Bottleneck	**58.59**	**13.93**	**20.72**	**9.19**

(c) Multi-modal fusion strategies.

change	*AP↑*
-clip	58.09
-place	57.51
-bassl	51.88
-clip-place	57.92
-	**58.59**

(d) Impact from input modalities on scene seg.

change	*TA↑*	*PA↑*	*D↓*
-clip	6.09	10.66	21.81
-place	13.57	19.87	9.22
-action	13.31	20.20	10.38
-appr	13.42	20.59	**8.85**
-	13.93	20.72	9.19
+subtitle	**14.63**	**21.78**	8.96

(e) Impact from input modalities on act segmentation.

synopsis synch. by	*M for synch.*	*TA↑*	*PA↑*	*D↓*
[30]	T*	10.51	14.54	8.98
MEGA	V	13.93	20.72	9.19
MEGA	V + T	**14.63**	**21.78**	**8.96**

(f) Impact of Synchronization with multimodal video features on act segmentation.

Approach	*Feature Set Pretrained on*	*AP↑*	*Params↓*	*SPS↑*
BaSSL [27]	Movienet	57.4	**15.77M**	**6244.99**
LGSS [35]	M+P+I	52.93	66.16M	206.36
MEGA	M+P+I	**58.59**	67.57M	1736.13

(g) Impact from feature set and model size on scene seg. SPS denotes # of samples per second.

Approach	*Feature Set*	*TA↑*	*PA↑*	*D↓*	*Params↓*	*SPS↑*
GRAPHTP [32]	Set1 [32]	9.12	12.63	9.77	**0.745M**	**25.40**
GRAPHTP [32]	Set2	4.72	7.37	9.69	6.78M	14.36
MEGA	Set2	**14.19**	**22.10**	**9.68**	6.78M	18.24

(h) Impact from feature set and model size on act seg. Set1 includes Visual (appr), Audio (YAMNet), Textual (script-USE). Set2 has Visual (appr,clip,action,place), Audio (audio), Textual (text from subtitle).

Table 4: Ablation studies on MEGA components.

alignment PE consistently improves the performance on both scene and act segmentation tasks, while act segmentation benefits more significantly from it. This is because of two reasons: 1) inputs to the act segmentation model have variable lengths (all the shots from a video) as opposed to scene segmentation which has fixed length inputs, hence the alignment PE adds more information to the act segmentation model; and 2) Alignment PE is shared across all modalities and fusion layer and its absence causes more harm to sequences with longer lengths, as act segmentation takes longer shot sequences (entire video) compared to scene segmentation (17 shots). Overall, the drop in performance indicates that the proposed Alignment P.E. is an essential component for video segmentation tasks.

We further remove the proposed Alignment PE from bottleneck fusion tokens and show results in Tab. 4b. We observe a performance drop when removing the Alignment PE. It is worth mentioning that the performance drop is more noticeable when multiple modalities are involved, (e.g. V + T model), which suggests that the information from subtitles requires precise temporal alignment in order to be effective during multimodal fusion.

Multimodal Fusion Strategies. Tab. 4c compares our proposed fusion tokens with the commonly used late fusion strategy. The results show that the temporal bottleneck fusion clearly outperforms late fusion, demonstrating the effectiveness of aligned bottleneck tokens in improving the performance across both tasks.

Different Input Modalities. We study the impact of different modalities by removing them from the input and present results on scene segmentation and act segmentation. In **scene segmentation (Tab. 4d)**, we find that the bassl feature followed by place are the most important. This is because the BaSSL [27] model is pretrained for scene segmentation and place consistency is critical for scene segmentation. In **act segmentation (Tab. 4e)**, we find that the CLIP feature pre-trained on IMDB, followed by subtitle are the most important features. This is because the clip model is pre-trained on abstract concepts (e.g. genre, micro genre, character type and coarse key places), thus the CLIP feature contains richer semantics, and subtitle provides complementary rich semantic information. These high level semantics are considered useful for act segmentation. Overall, when all the features are included, the model is able to leverage the unique information provided by each feature and yields the best performance.

Cross-modality Synchronization. Tab. 4f studies the effectiveness of the proposed cross modality synchronization on act segmentation. To establish a baseline, we use the probability scores provided by [32] and derived by aligning synopsis sentences to scenes using screenplay [30] (T* in Tab. 4f denotes screenplay). For a fair comparison, we repeat the scores provided for each scene on all of its shots and then re-normalize[2]. MEGA with visual only input outperforms [30] across two metrics (PA and TA) and when we add the subtitle features (T in Tab. 4f denotes subtitle), MEGA outperforms [30] across all the metrics. The results demonstrate that the proposed cross-domain synchronization works effectively and generalize well to various modalities. It is worth mentioning that the proposed method generates act segmentation without requiring the screenplay, which makes it practical for various industrial applications.

Impact of Feature Set. In Tab. 4g, we examine whether feature set plays an outsized role in MEGA's improved performance over other methods. On the same feature set of M + P + I, we outperform LGSS [35] while only introducing a small number of additional parameters. BaSSL [27] achieves slightly lower performance using only Movienet features, indicating that the use of additional features is not the primary reason for our improved performance.

Impact of Model Size. We finally look into the impact of model size on the performance. To establish the fair comparison, we first expand GRAPHTP [32] which shares the same input as MEGA to roughly the same number of parameters as MEGA. Our results on act segmentation (Tab. 4h) show that MEGA outperforms the previous SoTA, GRAPHTP [32], which indicates the MEGA has efficient and effective design.

4.5. Fusion with Audio Modality

We also experimented with the effect of adding audio. Movienet [18] has released *Short Time Fourier Transform* (STFT) features extracted from audio files. We use the same audio backbone as [35]. However, by adding audio features in MEGA, we see a drop in the performance (see Tab. 5). Although [8, 35] have shown improvements across multiple models by adding the audio modality across Movienet-150 dataset [35] (where the split is not publicly available) or a private dataset: AdCuepoints [8], Wu *et al.* [45] have observed a similar trend as our experiments, where adding the released audio features in Movienet [18] to SCRL [45] and ShotCoL [8] drops the performance (see Tab. 5). Possible reasons can be 1) the audio features published by Movienet via STFT[3] are an incomplete view of the shot from audio modality either in representation or in terms of the audio chunk from each shot they used, and the raw audio files are not available. 2) our multimodal fusion strategy cannot exploit the possible complementary information or filter the harmful or confusing signals from the audio modality.

Approach	*Modality*	*AP*↑ [%]
LGSS [35]	V(place)	39.00
ShotCoL [8]‡	V	46.77
SCRL [45]	V	54.55
MEGA	V(place,clip,bassl)	58.59
LGSS [35]	V(place)+A	43.4
ShotCoL [8]‡	V+A	44.32
SCRL [45]	V+A	50.80
MEGA	V(place,clip,bassl)+A	55.36

Table 5: Scene Seg. with audio. ‡denotes copying from [45].

5. Discussion and Conclusion

Limitations. The explorations in this work are limited to appearance, location, activity, acoustic and textual features. For long movie segmentation, however, providing the name of actors (tabular data) and having a specific component for actor identification in the movie can help both the synchronization and the act/scene segmentation models. We will explore the use of this data.

The results demonstrated that richer semantic representations from the clip features enhanced the performance for long video segmentation. To obtain better performing representations for long video understanding one can use large amount of unlabeled data with carefully selected pretext tasks for understanding long context. We will investigate the use of SSL to train a rich multimodal representation from videos and will examine the learned representations across multiple long video understanding tasks.

Conclusion. This paper introduces MEGA, a unified solution for long video segmentation. Our design of normalized positional encoding, and their integration into fusion tokens allows MEGA to learn consistent patterns from inputs with variable lengths and efficiently and effectively align and fuse them across different modalities. Our synchronization schema further allows the use of rich multimodal tokens to

[2] This strategy maintains the rank of probability scores for different turning points across different segments of the movie, and is consistent with the max-pooling of prediction scores on shot level to derive the scene level predictions during evaluation (see Sec. 4.2).

[3] https://github.com/movienet/movienet-tools

be used in transferring the labels from synopsis sentences to movie shots, facilitating the knowledge distillation from synopses to movies. MEGA achieves state-of-the-art performance compared with previous works.

References

[1] Sentence-transformers model: all-minilm-l6-v2 from Hugging Face. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. 3

[2] Yuki Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. Labelling unlabelled videos from scratch with multi-modal self-supervision. *Advances in Neural Information Processing Systems*, 33:4660–4671, 2020. 2

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *Advances in neural information processing systems*, 29, 2016. 2

[4] 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. 6

[5] Jessica Brody. *Save the Cat! Writes a Novel: The Last Book On Novel Writing You'll Ever Need*. Ten Speed Press, 2018. 2, 3

[6] 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

[7] 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. 6

[8] 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. 1, 2, 4, 6, 8

[9] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18166–18176, 2022. 1, 2, 3, 4

[10] Philip Gorinski and Mirella Lapata. Movie script summarization as graph-based scene extraction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1066–1076, 2015. 7

[11] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 6

[12] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2827–2836, 2016. 2

[13] 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. 6

[14] Michael Hauge. *Storytelling Made Easy: Persuade and Transform Your Audiences, Buyers, And Clients-Simply, Quickly, and Profitably*. BookBaby, 2017. 1, 2, 3, 7

[15] Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021. 2

[16] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*, 2016. 6

[17] Qingqiu Huang, Yu Xiong, and Dahua Lin. Unifying identification and context learning for person recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2217–2225, 2018. 6

[18] 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. 2, 5, 6, 8

[19] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 2

[20] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020. 5

[21] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 1, 2, 4

[22] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020. 3

[24] Shengjie Luo, Shanda Li, Tianle Cai, Di He, Dinglan Peng, Shuxin Zheng, Guolin Ke, Liwei Wang, and Tie-Yan Liu. Stable, fast and accurate: Kernelized attention with relative positional encoding. *Advances in Neural Information Processing Systems*, 34:22795–22807, 2021. 2

[25] Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411, 2004. 7

[26] Paramita Mirza, Mostafa Abouhamra, and Gerhard Weikum. Alignarr: Aligning narratives on movies. In *The 59th Annual*

Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 427–433. ACL, 2021. 3, 5

[27] 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. 2, 3, 4, 6, 7, 8

[28] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021. 2, 3, 4

[29] Pinelopi Papalampidi, Frank Keller, Lea Frermann, and Mirella Lapata. Screenplay summarization using latent narrative structure. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1920–1933, 2020. 6, 7

[30] Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Movie plot analysis via turning point identification. *arXiv preprint arXiv:1908.10328*, 2019. 1, 2, 3, 6, 7, 8

[31] Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Film trailer generation via task decomposition. *arXiv preprint arXiv:2111.08774*, 2021. 2, 5

[32] Pinelopi Papalampidi, Frank Keller, and Mirella Lapata. Movie summarization via sparse graph construction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13631–13639, 2021. 1, 2, 4, 5, 6, 7, 8

[33] Patrice Pavis. *Dictionary of the theatre: Terms, concepts, and analysis*. University of Toronto Press, 1998. 2

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 3, 5

[35] 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. 1, 2, 5, 6, 7, 8

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

[37] 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. 6

[38] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 464–468, 2018. 2

[39] Tomáš Souček and Jakub Lokoč. Transnet v2: an effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. 1, 3

[40] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. 2

[41] 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. 6

[42] Makarand Tapaswi, Martin Bäuml, and Rainer Stiefelhagen. Aligning plot synopses to videos for story-based retrieval. *International Journal of Multimedia Information Retrieval*, 4(1):3–16, 2015. 3

[43] Makarand Tapaswi, Martin Bauml, and Rainer Stiefelhagen. Book2movie: Aligning video scenes with book chapters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1827–1835, 2015. 3

[44] Yikai Wang, Wenbing Huang, Fuchun Sun, Tingyang Xu, Yu Rong, and Junzhou Huang. Deep multimodal fusion by channel exchanging. *Advances in Neural Information Processing Systems*, 33:4835–4845, 2020. 2, 3

[45] 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. 2, 4, 5, 6, 8

[46] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 3

[47] Yu Xiong, Qingqiu Huang, Lingfeng Guo, Hang Zhou, Bolei Zhou, and Dahua Lin. A graph-based framework to bridge movies and synopses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4592–4601, 2019. 3

[48] Ning Zhang, Manohar Paluri, Yaniv Taigman, Rob Fergus, and Lubomir Bourdev. Beyond frontal faces: Improving person recognition using multiple cues. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4804–4813, 2015. 6

[49] Chao Zhao, Wenlin Yao, Dian Yu, Kaiqiang Song, Dong Yu, and Jianshu Chen. Learning-by-narrating: Narrative pre-training for zero-shot dialogue comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 212–218, Dublin, Ireland, May 2022. Association for Computational Linguistics. 3

[50] 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. 3, 6