Movie Genre Classification by Language Augmentation and Shot Sampling

Zhongping Zhang[1] Yiwen Gu[1] Bryan A. Plummer[1]
Xin Miao[2] Jiayi Liu[2] Huayan Wang[2]
[1]Boston University [2]Kuaishou Technology
[1]{zpzhang, yiweng, bplum}@bu.edu [2]wanghy514@gmail.com

Abstract

*Video-based movie genre classification has garnered considerable attention due to its various applications in recommendation systems. Prior work has typically addressed this task by adapting models from traditional video classification tasks, such as action recognition or event detection. However, these models often neglect language elements (*e.g.*, narrations or conversations) present in videos, which can implicitly convey high-level semantics of movie genres, like storylines or background context. Additionally, existing approaches are primarily designed to encode the entire content of the input video, leading to inefficiencies in predicting movie genres. Movie genre prediction may require only a few shots[1] to accurately determine the genres, rendering a comprehensive understanding of the entire video unnecessary. To address these challenges, we propose a Movie genre Classification method based on Language augmentatIon and shot samPling (Movie-CLIP). Movie-CLIP mainly consists of two parts: a language augmentation module to recognize language elements from the input audio, and a shot sampling module to select representative shots from the entire video. We evaluate our method on MovieNet and Condensed Movies datasets, achieving approximate* $6-9\%$ *improvement in mean Average Precision (mAP) over the baselines. We also generalize Movie-CLIP to the scene boundary detection task, achieving* 1.1% *improvement in Average Precision (AP) over the state-of-the-art. We release our implementation at this http URL.*

1. Introduction

Video-based movie genre classification facilitates a wide range of applications, including content recommendation [15, 27], genre-based video retrieval and filtering [20], and automatic tagging and annotation [21]. Early research on this task typically focused on domain-specific datasets with only a few categories [6, 40, 45, 63]. Restricted by the model's scalability and dataset size, these methods primarily predict movie genres using posters or still frames from videos. Recently, to leverage the prior knowledge of large-scale video-based datasets [1, 8, 20], researchers have explored adapting models from related video classification tasks (*e.g.*, Uniformer [29] or TRN [61]) to movie genre classification, as shown in Figure 1 (A).

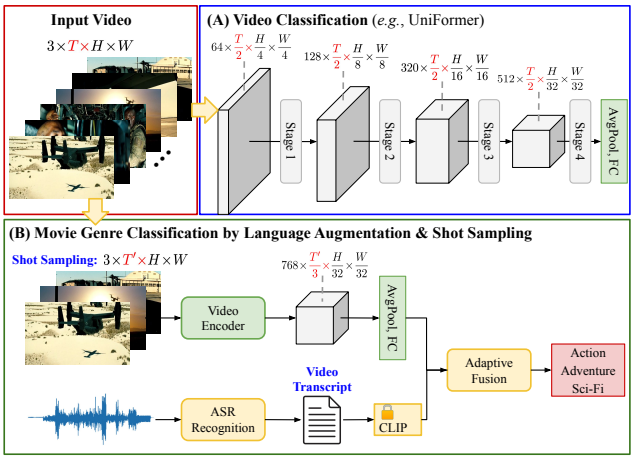

Figure 1. Our task aims at predicting movie genres based on input videos. Existing methods commonly approach this task by adapting models from related video-classification tasks (*e.g.*, action recognition [8, 29] or topic recognition [1]). As shown in (A), these methods typically incorporate the entire video as input and ignore the language elements in videos, making the prediction less efficient and accurate. To address these challenges, we introduce two components to our model, as shown in (B). We propose a language augmentation module to extract language information from the input video, circumventing the dependence on provided video captions and improving the prediction accuracy. Additionally, we apply shot sampling strategy to select representative shots from the entire video, leading to a notable reduction in computational cost.

Directly adapting these video classification models to movie genre prediction presents two challenges. First, prior work typically neglects the language information present in videos [8, 17, 29, 58, 61]. However, this information,

[1]A shot is defined as a series of frames captured from the same camera over an uninterrupted period of time [51].

such as narrations or conversations, can implicitly convey genre information. In some cases, movie genres can even be accurately predicted based solely on movie transcripts. For example, imagine a movie scene where characters engage in intense discussions about mysteries, accompanied by eerie music. The suspenseful dialogue and atmospheric setup convey the information of horror or thriller genre. In contrast, in a movie where characters engage in coincidental encounters and comedic misunderstandings, the lighthearted language will suggest that the film most likely belongs to the romantic or comedy genre. Second, video classification frameworks often encode the entire video to comprehend the events occurring in videos. A comprehensive understanding of the entire video is essential for tasks like action recognition or event detection. However, in movie genre classification, we observe that only a few shots can be sufficient to accurately predict movie genres. For example, humans can predict genres based solely on clips or trailers, without the need to watch the entire movie.

To introduce language information into models, a straightforward method is collecting text documents related to the input video and encoding them as part of the input. This strategy is often adapted by multimodal methods when provided with both videos and text documents [7, 9, 30]. However, collecting the text documents will introduce extra overhead, and these text documents may not always be available. For example, Condensed Movies [2] collected approximately 33,000 movie clips from YouTube, and captions were absent in half of these videos. To address this issue, we integrate an Automatic Speech Recognition (ASR) system, Whisper [37], into our model, as illustrated in Figure 1 (B). In this scenario, our model automatically extracts language information from the audio, eliminating the requirement of user-provided captions. Additionally, we introduce a shot sampling module to improve the efficiency of our model. Motivated by sparse sampling strategies [21, 28, 55], our method improves efficiency from two aspects. First, we segment the entire video into individual shots and select shots from different scenes as the video representation. Second, we subsample keyframes from each shot to use as the shot representation. Since the number of keyframes T' is significantly smaller than the total number of frames T in the video (Figure 1 (B)), our model notably reduces the computational cost.

In summary, the contributions of this paper are:

- We propose a language augmented approach for movie genre prediction. Compared to prior work [9, 20, 21], Movie-CLIP extracts language information from the input video and does not rely on external language sources, such as captions, metadata, or Wikipedia.
- We leverage a shot sampling strategy to select key frames from input video as the visual representations. This strategy notably reduces the computational cost, while achieving competitive performance compared to frameworks [29, 61] that encode the entire video.
- Experiments on MovieNet [20] and Condensed Movies [2] demonstrate that Movie-CLIP outperforms the baselines, improving approximately 6-9% mAP points on genre classification. We further show that Movie-CLIP generalizes to scene boundary detection task, achieving 1.1% improvement in AP over the state-of-the-art.
- We perform extensive experiments on movie genre classification, exploring the correlations between movie genres and different components across various modalities.

2. Related Work

Studies on Movies involve a great number of research topics, spanning genre classification [9, 20, 21], scene boundary detection [12, 38], shot boundary detection [46, 47], person re-identification [57], action recognition [5, 44, 59], alignment between movie and text descriptions [14, 49, 64], understanding relationships of film characters [3, 26, 32, 56], movie question answering [23, 50, 54], scene and event understanding [12, 43], among others. Existing approaches typically comprehend movies from a visual perspective [20, 21] or align the visual modality with corresponding labels across other modalities, such as actions [25] or text descriptions [49]. Therefore, the language elements in movies are either ignored or provided as part of the input. In this paper, we explore to automatically extract language elements from input videos to improve the performance of genre classification. Compared to previous multimodal methods (*e.g.*, Moviescope [9]), Movie-CLIP leverages the language information for free, eliminating the requirement of additional language annotations like Wikipedia or metadata.

Movie Genre Classification can be divided into two major categories: image-based (posters, still frames, etc.) [20, 45, 63] or video-based (trailers, movie clips, etc.) [20, 21]. Recently, researchers have adapted popular frameworks from traditional video classification tasks to movie genre classification, such as methods on action recognition [11, 31, 52, 55, 61] or video summarization [34, 53]. An obstacle for these frameworks is the computational cost. Methods that take all frames as input would be infeasible to handle videos with hours' duration [52, 55]. Though sparse sampling strategies have been proposed to process videos more efficiently [28, 61], the analysis of hour-long videos still costs significant resources. To address this issue, we use a shot sampling algorithm to first divide the entire video into individual shots, and then select representative shots from different scenes, predicting genres efficiently.

Scene Boundary Detection aims to identify the starting and ending points of different scenes in videos. Early methods [39, 42] primarily employed unsupervised learning to segment scenes, relying on the similarity in colors. With

the emergence of datasets with human-annotations [4, 41], supervised learning approaches [4,35,38,41] have been proposed. A notable advancement in this topic was marked by the introduction of MovieNet [20], which comprises 1,100 movies, with 318 of them annotated with scene boundaries.

3. Movie-CLIP: Movie genre Classification by Language augmentatIon & shot samPling

Given a user-provided video V, our task aims at predicting movie genres accurately and efficiently by leveraging language elements L and sparsely sampled video representation V'. Since the language elements (*e.g.*, narrations or conversations) in videos can convey high-level semantics of movie genres, we explore the incorporation of language information, such as movie transcripts, into genre classification models. To circumvent the dependence on user-provided text documents, we integrate an ASR system to automatically recognize language elements from the audio, as discussed in Section 3.1. To improve model efficiency, we apply a shot sampling module in Section 3.2. This module selects representative shots from the entire video as the visual representation. Additionally, we introduce a fusion strategy to concatenate the outputs of each modality, which is discussed in Section 3.3. Figure 2 provides an overview of our method.

3.1. Language Augmentation

As discussed in the Introduction, language information such as narrations or conversations can play an important role in genre prediction. For example, narrations of *documentary* genres are often richer than those in *action* genres. Existing multimodal approaches commonly incorporate language information from user-provided text documents [7, 9]. However, while some videos come with text data like captions, there is a considerable number of videos that do not have captions. To circumvent the dependence on provided text documents, we introduce a language augmentation module in Movie-CLIP. This module incorporates an ASR model, Whisper [37], to generate transcripts. In other words, the input to Movie-CLIP comprises the input video and its associated audio, from which the language modality is extracted and leveraged by our model without additional requirements. Thus, we named this mechanism as the language augmentation module. An overview of our language augmentation module is provided in Figure 2 (A), consisting of the following components:

Automatic Speech Recognition. Given the input audio A, we apply Whisper [37] to obtain the initial transcript L. The initial transcript consists of multiple language components, including narrations, conversations, voice activities, etc.

Keyword-aware Documents. To incorporate the language information, a straightforward method is to directly apply a language encoder to the transcript L. However, in our experiments, we observed that directly applying a language encoder like BERT [22] to L only resulted in a slight boost to the performance of our genre prediction model. We attribute this to the fact that the ASR system cannot perfectly recognize all language tokens, leading to the presence of noise in the initial transcript L, which might adversely affect the genre prediction results. In addition, as we will show, some words are predictive of a particular genre. Thus, we propose a keyword-aware algorithm based on these observations. Motivated by the intuition that Nouns, Pronouns, and Adjectives often contain important clues for describing events in videos, we first narrow the scope from the entire document to Nouns, Pronouns, and Adjectives. Each word's part-of-speech is identified by SpaCy [19]. We then define keyword K as tokens with high frequency that appear in captions. From these tokens, we select the top k[2] tokens with high frequency that appear in captions.

Language Representation. Our language encoder takes the concatenation of the initial transcript L and the keywords K as input. Language representations are obtained by $\mathrm{f}_l(L, K; \boldsymbol{\theta_l})$, where we apply the text encoder of CLIP [36] as our $f_l(\cdot)$. Based on $\mathrm{f}_l(L, K; \boldsymbol{\theta_l})$, we further apply a linear layer to get the prediction score for the language modality:

$$\rho_l = W\mathrm{f}_l(L, K; \boldsymbol{\theta_l}) + b, \tag{1}$$

where W and b are the linear layer's learnable parameters.

Language Modality Loss. As shown in Figure 2, we apply a multi-label classification head[3] to each modality, enabling our model to effectively handle both multi-modal features and each individual modality. We apply binary relevance to train our model. The language prediction head of Movie-CLIP comprises an ensemble of binary classifiers, with each classifier predicting the presence of a specific genre in the video. The loss of the j-th genre can be expressed as

$$L_{lj} = -\left[y_j \log(\rho_{lj}) + (1 - y_j) \log(1 - \rho_{lj})\right], \tag{2}$$

where y_j and ρ_{lj} denote the ground truth label and prediction scores for the j-th movie genre, respectively. We average Eq. 2 across each classifier to obtain the final text loss:

$$L_{text} = \frac{1}{K} \sum_{j=1}^{K} L_{lj}. \tag{3}$$

Audio Representation and Loss. Consistent with the language representation and loss, we obtain the prediction score of the audio modality, ρ_a, by applying PANNs [24], and use a binary relevance strategy to train the multi-label classifiers for the audio modality.

[2] In our experiments, we set k to 20.

[3] Since a movie often has multiple genre labels, movie genre prediction is formulated as a multi-label classification task.

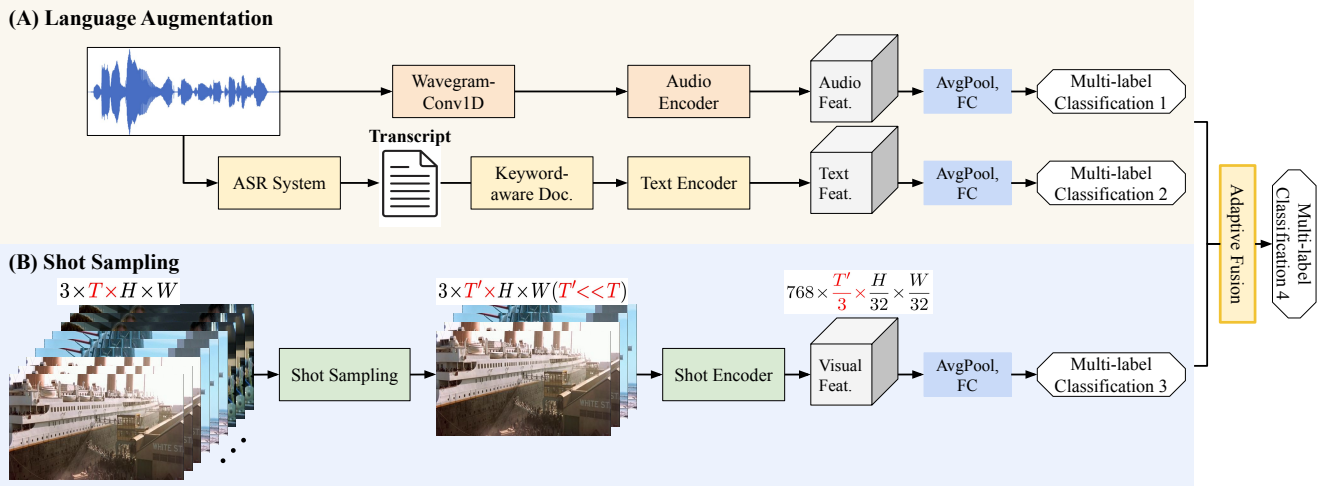

Figure 2. **Movie-CLIP Overview**. Our approach consists of two major components: (A) Language Augmentation: An ASR System, Whisper [37], is employed to automatically extract language elements for the input video, eliminating the requirement of provided captions. We further propose a keyword-aware mechanism to suppress the noise introduced by the ASR system. See Section 3.1 for detailed disccusion; (B) Shot Sampling: We introduce a sparse sampling strategy to select shots from different scenes as the video representation. In each shot, we sample key frames as the shot representation. See Section 3.2 for detailed discussion.

3.2. Shot Sampling

In movie genre classification, the input video can be quite lengthy in some cases (*e.g.*, a movie may last around two hours or longer). Consequently, encoding the frames of the entire video is computational expensive. Existing approaches [28, 55] applied sparse sampling strategies to reduce the computational cost. To sample frames with consistent semantic information, Huang *et al*. [21] proposed dividing the input video into a sequence of coherent shots and randomly select shots as the visual representation. Motivated by this work, we introduce our shot sampling strategy in Figure 2 (B), with the following steps:

Shot-based Video Representation. Given a video V, we initially divide it into separate shots $\{S_1, S_2, ..., S_N\}$ using a shot boundary detection framework [46]. Within each shot S_i, we uniformly sample m frames $\{I_{i1}, I_{i2}, ..., I_{im}\}$. The representation of shot S_i is computed by taking the average of the features extracted from the m frames:

$$\mathbf{f_v}(S_i; \boldsymbol{\theta_v}) = \frac{1}{m}\sum_{j=1}^{m} \mathbf{f_v}(I_{ij}; \boldsymbol{\theta_v}), \tag{4}$$

where $\mathbf{f_v}(\cdot)$ is the feature extractor, and $\boldsymbol{\theta}$ its the corresponding parameters. Consistent with our language encoder, we use CLIP as our image encoder $\mathbf{f_v}(\cdot)$. We further combine the shot representations to derive the video representation

$$\mathbf{f_v}(V; \boldsymbol{\theta_v}) = \frac{1}{N'}\sum_{i=1}^{N'} \mathbf{f_v}(S_i; \boldsymbol{\theta_v}), \tag{5}$$

where N' denotes the number of sampled shots. Correspondingly, the number of sampled frames is $T' = m \times N'$, which is much smaller than the number of frames T in the video, thus reducing our model's computational burden.

Shot Sampling Strategy. We observe that in movie videos, shots within the same scene are often semantically similar, resulting redundancy in movie genre classification. For example, as we will show in Figure 3, shots within scenes depicting weapons and soldiers always tend to be categorized as the genre of *War*, thus containing redundant information. To address this issue, we sample shots from different scenes instead of the random sampling done in prior work [21].

Visual Modality Loss. Consistent to our language modality loss, we apply binary relevance to train multi-label classifiers for the visual modality. Given the prediction score ρ_v, the loss is defined as:

$$L_{vj} = -\left[y_j \log(\rho_{vj}) + (1 - y_j)\log(1 - \rho_{vj})\right], \tag{6}$$

$$L_{visual} = \frac{1}{K}\sum_{j=1}^{K} L_{vj}, \tag{7}$$

where ρ_{vj} denotes the prediction score of visual modality for the j-th movie genre.

3.3. Adaptive Fusion

To combine multimodal features, we apply a weighted linear regression on the outputs of each modality. Let $\rho = \{\rho_v, \rho_l, \rho_a\}$, then the final prediction score is obtained by

$$\rho = \alpha\rho_v + \beta\rho_a + \gamma\rho_l, \tag{8}$$

where α, β, γ can be interpreted as the hyperparameters that control the contribution of each modality. Inspired by adaptive mechanisms that can automatically learn hyperparameters [60], we convert α, β, γ into extra learnable parameters. Thus, these parameters can be considered as attention weights assigned to each modality. Correspondingly, the loss function of multi-modal features is:

$$L_j = -\left[y_j \log(\rho_j) + (1 - y_j) \log(1 - \rho_j)\right], \quad (9)$$

$$L_{multi} = \frac{1}{K} \sum_{j=1}^{K} L_j, \quad (10)$$

where ρ_j denotes the prediction scores based on multi-modal features for the j-th movie genre. The loss for the whole model is given by:

$$L_{total} = L_{multi} + L_{visual} + L_{text} + L_{audio}. \quad (11)$$

4. Experiments

4.1. Datasets and Experimental Settings

Datasets. We evaluate Movie-CLIP on MovieNet [20] and Condensed Movies [2]. The released version of MovieNet contains 1.1K movies and 30K trailers, which are provided as URLs. Filtering out invalid links and unlabeled trailers resulted in 28,466 trailers remaining. As MovieNet [20] does not release their testing split, we randomly split the 28K trailers into training, validation, and test sets, following the ratio of 7:1:2 in [20]. Condensed Movies consists of 33K movie clips from 3,600 movies. After we processed Condensed Movies using the same procedure as MovieNet, we obtained 22,174 movie clips. We split the dataset into 15,521/2,217/4,436 train/val/test clips, respectively.

Metrics. Following [20], we adopt recall@0.5 and precision@0.5, which refers to using a 0.5 threshold to distinguish between positive and negative predictions as our evaluation metrics. We also use mean average precision (mAP). Given the considerable unbalanced distribution of movie genres, we report the assessment scores at "macro" and "micro" levels. The "macro" average weights the metrics equally for each genre, thereby ignoring the label imbalance. Within the "micro" average, the metrics are computed across all categories. It aggregates the contributions of all classes. In simpler terms, "macro" accentuates the influence of samples attributed to smaller categories, while "micro" assigns equal importance to every single sample.

Baselines. Following [20], we first compare Movie-CLIP with several established models, namely TSN [55], I3D [8], TRN [61], and Trailer-Storylines [21]. Our reported results for these models are taken from [20]. In addition, we extend our evaluation to include adaptations from video classification and retrieval: SlowFast [17], Uniformer [29], and Imagebind [18]. We reproduce these models, drawing from either their official code repositories or the MMAction2 [13] framework. To ensure a fair comparison, we finetune these models' video encoders on each dataset and replace their prediction heads with our custom-designed multi-label genre classifiers.

Implementation Details. Each input video is split into distinct shots by TransNet v2 [46]. We select 8 shots where each shot consists of 3 sampled frames as the visual representation of the input video. We sample audio waveforms at a rate of 16 kHz from each video as the input to both PANNs [24] and Whisper [37]. Our models are trained with a batch size of 256 and a maximum learning rate of 10^{-3} on NVIDIA RTX-3090 GPUs. We adopt the "ReduceLROnPlateau" [33] strategy to reduce the learning rate.

4.2. Genre Classification

Quantitative Results. In Table 1, we present the quantitative results of different models on MovieNet [20] and Condensed Movies [2]. We observe that Movie-CLIP significantly outperforms the baselines referenced in [20]. *E.g.*, Movie-CLIP outperforms Trailer-Storylines [21] by approximately 20∼30% in macro-mAP and 22∼24% in micro-mAP. The improvement in both macro and micro metrics demonstrates the capability of Movie-CLIP to enhance performance across the entire dataset, including samples within imbalanced genres. When comparing to baselines adapted from other video understanding tasks[4], Movie-CLIP achieves an improvement of 6∼9% in mAP. Though Movie-CLIP does not achieve the highest performance across all metrics, *e.g.*, Imagebind [18] achieving the highest r@0.5 score on Condensed Movies, we note that it has a better trade-off among various metrics, thus validating the effectiveness of our proposed modules.

Ablation Study. In Table 2, we present the ablation results of Movie-CLIP from three aspects. First, we compare our shot sampling strategy (*Shot*) to randomly sampling (*Random*). Our shot sampling strategy notably outperforms randomly sampling. Second, we evaluate the effectiveness of each modality. We note that while the performance does not match that of the *Shot* modality (*e.g.*, 62.2 vs. 42.5 vs. 37.8 in macro-mAP), the *Audio* and *Language* modalities still hold significant importance in movie genre prediction. Third, we investigate the incorporation of text features and observe that incorporating text features improves performance compared to models lacking such features (*e.g.*, 64.8 vs. 65.2 vs. 65.4 in macro-mAP, 75.0 vs. 74.8 vs. 75.2 in micro-mAP). The improvement validates that our language augmentation module can effectively extract pertinent linguistic information from the input audio, while also verifying our keyword-aware documents can suppress the noise in the initial transcript.

[4]We replace the prediction heads of these models with our multi-label classifiers to make them work better on movie genre classification.

Method	macro			micro		
	r@0.5	p@0.5	mAP	r@0.5	p@0.5	mAP
(A) MovieNet						
TSN [55]	17.95	78.31	43.70	-	-	-
I3D [8]	16.54	69.58	35.79	-	-	-
TRN [61]	21.74	77.63	45.23	-	-	-
SlowFast [17]	17.60	67.70	41.33	28.82	69.34	50.35
SlowFast-Multimodal [58]	21.05	68.12	43.62	35.16	65.55	52.96
Trailer-Storylines [21]	19.52	72.40	44.02	33.32	64.55	53.14
Uniformer [29]	38.72	70.31	58.21	48.61	74.19	69.00
Imagebind [18]	39.84	71.63	58.74	49.85	73.81	69.66
Movie-CLIP (ours)	**40.42**	**80.05**	**65.38**	**52.45**	**80.08**	**75.21**
(B) Condensed Movies						
SlowFast [17]	17.25	58.18	39.99	26.72	65.88	50.78
SlowFast-Multimodal [58]	20.57	59.29	42.51	34.45	67.36	53.89
Trailer-Storylines [21]	14.87	61.57	41.33	26.39	68.95	54.83
Uniformer [29]	30.25	74.92	55.58	45.68	72.94	67.19
Imagebind [18]	**42.79**	76.40	63.68	52.42	75.58	72.73
Movie-CLIP (ours)	41.30	**86.38**	**72.01**	**53.79**	**82.53**	**78.35**

Table 1. **Movie genre classification.** The scores of TSN [55], I3D [8], TRN [61] are cited from [20], which do not include the micro metrics. We implement the other baselines and report both macro and micro metrics to provide comprehensive comparison for unbalanced genre labels. Movie-CLIP not only notably outperforms the baselines reported in [20], but also achieves a 6∼9% improvement in mAP compared to methods adapted from different video classification tasks. See Section 4.2 for discussion.

Method	macro-mAP	micro-mAP
Random	52.81	63.90
Shot	62.22	72.31
Audio	42.48	58.92
Language	37.76	49.04
Shot+Audio	64.77	75.00
Shot+Audio+Raw Text	65.15	74.80
Shot+Audio+Language	**65.38**	**75.21**

Table 2. **Ablation study on MovieNet.** *Random* denotes that we randomly sample frames from the videos. *Shot* denotes our shot sampling strategy. *Raw Text* denotes that we directly use the initial transcript L as the text input. *Language* denotes that we use the keyword-aware document as the text input.

4.3. Movie Analysis based on Genre Classification

Genre-based Shot Retrieval. As discussed in the Introduction, comprehending movies is a challenging task, especially when dealing with videos of considerable duration. However, given that Movie-CLIP segments input videos into discrete shots, it can be easily applied to genre-based shot retrieval in lengthy videos through a sliding window approach. Correspondingly, Movie-CLIP extracts a sequence of genre labels from the video input.

In Figure 3, we present the application of genre classification in genre-based shot retrieval using "Transformers: Revenge of the Fallen" as an example. As shown in the figure, Movie-CLIP not only accurately identifies shots that correspond to the ground truth genres but also generalizes well to genres that are not part of the ground truth. Specifically, the movie "Transformers: Revenge of the Fallen" falls under the genres of *Sci-Fi* and *Action*, with their respective shots showcased in Figure 3 (A) and Figure 3 (B). Aligning with our expectations, shots classified as *Sci-Fi* encompass scenes depicting the universe, planets, or robot armies, while shots categorized as *action* show up together with typical elements found in *action* movies, such as explosions or dynamic movements. We further present two additional genres, *Romance* and *War*, in Figure 3 (C) and Figure 3 (D), respectively. Movie-CLIP effectively shows a sequence of shots related to these two genres. For example, in Figure 3 (D), shots featuring weapons or soldiers are more likely to associated with the *War* genre, whereas shots portraying daily life or romantic relationships are more likely to align with the *Romance* genre. Genre-based shot retrieval carriers practical implications, such as automated trailer generation or automatic clipping for highlighting movies. See the supplementary for additional examples.

Sound Event Analysis. We analyze audio waveforms on MovieNet to uncover the correlations between genres and sound events. The representative sound events of four different movie genres are presented in Figure 4. The figure substantiates that sound events in the audio modality are

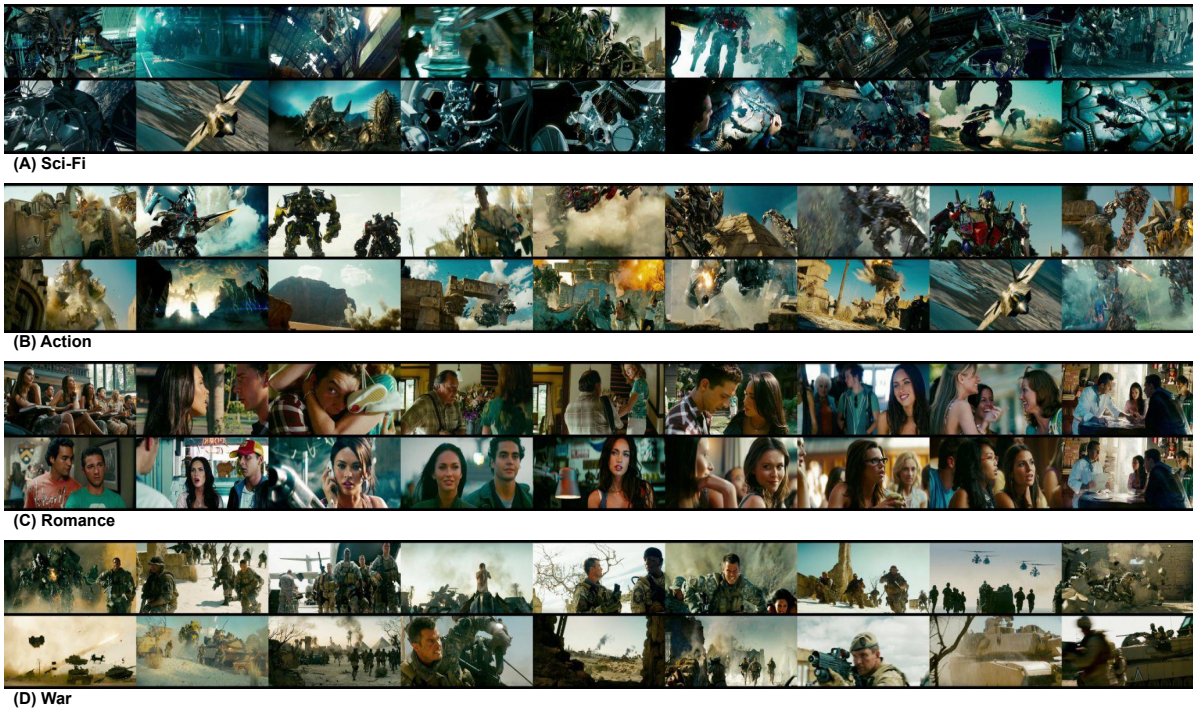

Figure 3. **Genre-based shot retrieval** on movie "Transformers: Revenge of the Fallen." Movie-CLIP effectively identified shots corresponding to various genres within a video spanning hours. See Section 4.3 for detailed discussion

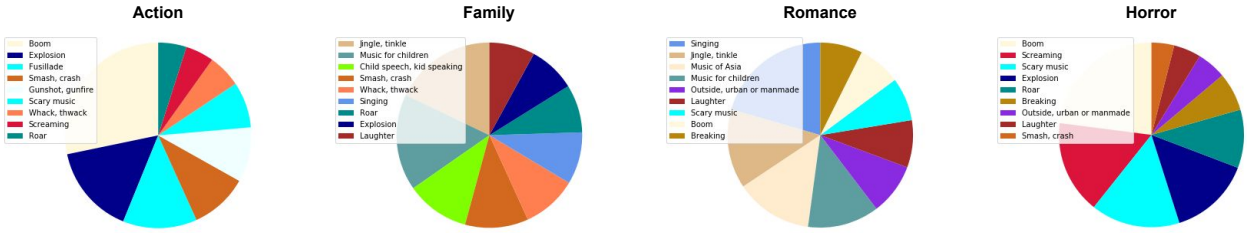

Figure 4. **Representative sound events** of *Action*, *Family*, *Romance* and *Horror*. Sound events exhibit discriminative characteristics acorss various genres, supporting our motivation for using sound events to improve our model. See Section 4.3 for further discussion.

discriminative attributes for genre recognition. For example, the prevalent sound events in the *Romance* genre include "Singing," "Music for children," and "jingle, tinkle," evoking feelings of relaxation and happiness. In contrast, elements characteristic of *Action* movies consistently involve "Gunshot," "Scary music," and "fusillade," making people feel thrilled and excited. Additional examples can be found in the supplementary material.

Keyword Analysis. We calculate the Term Frequency - Inverse Document Frequency (TF-IDF) to reveal the correlations between keywords and movie genres. Specifically, for each genre, we create table T with dimensions $n \times m$, where n is the number of movies in that genre and m represents the vocabulary size. T_{ij} denotes the TF-IDF value of word j in movie i, and the score of word j across entire genre is defined as $s_j = \sum_i^n T_{ij}$. Subsequently, we generate wordclouds for each genre. It is worth noting that some words such as "know," "man," "think" rank high among most of genres but do not carry real information. To address this issue, we introduce a mechanism where the top N words from all genres are aggregated into a list of size $N \times 21$, and their occurrences are counted. Any word surpassing the threshold of M occurrences within this list is excluded from the wordcloud plots. Here, we set N to 20 and M to 5. Figure 5 shows wordcloud plots of *Romance*, *Thriller*, *Music* and *War* genres on the trailer data of MovieNet. We observe that trailers containing keywords such as "singer," "applause," "blues" tend to align more with the *Music* genre. Conversely, *War* movies exhibit stronger associations with words such as "soldier," "country," "majesty." See supplementary for additional examples.

Figure 5. **Wordclouds** of *War*, *Music*, *Thriller* and *Romance* (MovieNet). Word clouds show varying distributions across genres, validating our keyword-aware approach to movie genre classification. See Section 4.3 for further discussion.

4.4. Generalization to Scene Boundary Detection

Datasets & Experiment Settings. The evaluation of the scene boundary detection task is conducted on MovieNet using 318 movies with annotated scene boundaries. Following ShotCoL [12], we split the 318 movies into 190, 64, 64 movies for training, validation and test sets, respectively. Average Precision (AP) and Recall@0.5 are applied as our evaluation metrics. For this task, we use a sequence of four consecutive shots as input, with the probability that a scene boundary exists between the second and third shots as output. Consistent with our approach in the genre classification task, we employ Binary Cross Entropy as the loss function. Due to the data imbalance, the weight for boundary versus non-boundary samples is 10:1.

Model Architecture. Following [12], our decoder uses a three-layer MLP classifier (number-of-shots $\times$ feature-dimension - 4096 - 1024 -2). Similar to the genre classification task, we representation each shot with 3 subsampled frames. As in [12], we encode each shot's frames using a model pretrained on Places[5] [62].

Evaluation Results. Table 3 reports the scene boundary detection results of Movie-CLIP and other baselines. From the table, we see that Movie-CLIP gets new state-of-the-art results with our shot representations and Places features.

4.5. Limitations and Future Work

In this paper, we adopt the binary relevance strategy to train Movie-CLIP. While this approach is straightforward to implement, it ignores the inter-dependencies among labels. Our observation reveals that movie genres indeed exhibit correlations with one another. For example, genres like *Thriller*, *Crime*, and *Horror* often co-occur. Similarily, *Family* genre frequently occurs with *Animation*. In contrast, negative Pearson correlation coefficients exist between genres like *Comedy* and *Thriller*, *Drama* and *Documentary*. Based on these findings, we conclude that effectively leveraging the correlations among different genres should be helpful for movie genre classification.

Moreover, our primary focus lies in the impact of pretrained features from different modalities in this paper. The encoders to extract these features remain frozen during the training of Movie-CLIP. As a result, a potential improvement can involve the development of refined training strategies, such as end-to-end learning method.

[5]Places is a large-scale dataset for the scene recognition task.

Models	AP	Recall@0.5
SCSA [10]	14.7	54.9
Story Graph [48]	25.1	58.4
Siamese [4]	28.1	60.1
ImageNet [16]	41.26	30.06
Places [62]	43.23	59.34
LGSS [38]	47.1	73.6
ShotCoL [12]	53.37	81.33
Movie-CLIP (ours)	**54.45**	**82.21**

Table 3. **Scene boundary detection** results on MovieNet. Baseline results are taken from [12]. We find Movie-CLIP generalizes well on the scene boundary detection task, outperforming the state-of-the-art. See Section 4.4 for discussion.

5. Conclusion

In this paper, we propose a movie genre classification model, Movie-CLIP, which consists of language augmentation and shot sampling modules. For language augmentation model, since our model's transcripts are extracted from input audios, Movie-CLIP enhances performance without requiring additional language annotations. Additionally, we introduce a shot sampling strategy designed to select representative shots from diverse scenes within a video. This approach reduces computational cost in comparison to encoding the entire video. Movie-CLIP outperforms existing benchmarks on movie genre classification, improving 6-9% mAP scores on MovieNet and Condensed Movies datasets. We also generalize Movie-CLIP to scene boundary detection task, achieving the new state-of-the-art by improving 1.1% AP scores. We perform extensive experiments to demonstrate the applications of Movie-CLIP on movie analysis and explore the correlations between genres and various movie elements across different modalities.

Acknowledgements This material is based upon work supported, in part, by DARPA under agreement number HR00112020054. Any opinions, findings, and conclusions or recommendations are those of the author(s) and do not necessarily reflect the views of the supporting agencies.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 1

[2] Max Bain, Arsha Nagrani, Andrew Brown, and Andrew Zisserman. Condensed movies: Story based retrieval with contextual embeddings. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 2, 5

[3] David Bamman, Brendan O'Connor, and Noah A Smith. Learning latent personas of film characters. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 352–361, 2013. 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. 3, 8

[5] Piotr Bojanowski, Francis Bach, Ivan Laptev, Jean Ponce, Cordelia Schmid, and Josef Sivic. Finding actors and actions in movies. In *Proceedings of the IEEE international conference on computer vision*, pages 2280–2287, 2013. 2

[6] Darin Brezeale and Diane J Cook. Using closed captions and visual features to classify movies by genre. In *Poster session of the seventh international workshop on Multimedia Data Mining (MDM/KDD2006)*. Citeseer, 2006. 1

[7] Zihui Cai, Hongwei Ding, Jinlu Wu, Ying Xi, Xuemeng Wu, and Xiaohui Cui. Multi-label movie genre classification based on multimodal fusion. *Multimedia Tools and Applications*, pages 1–18, 2023. 2, 3

[8] 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. 1, 5, 6

[9] Paola Cascante-Bonilla, Kalpathy Sitaraman, Mengjia Luo, and Vicente Ordonez. Moviescope: Large-scale analysis of movies using multiple modalities. *arXiv preprint arXiv:1908.03180*, 2019. 2, 3

[10] 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. 8

[11] Chun-Fu Richard Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6165–6175, 2021. 2

[12] 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. 2, 8

[13] MMAction2 Contributors. Openmmlab's next generation video understanding toolbox and benchmark. https://github.com/open-mmlab/mmaction2, 2020. 5

[14] Timothee Cour, Chris Jordan, Eleni Miltsakaki, and Ben Taskar. Movie/script: Alignment and parsing of video and text transcription. In *European Conference on Computer Vision*, pages 158–171. Springer, 2008. 2

[15] Yashar Deldjoo, Mehdi Elahi, Paolo Cremonesi, Franca Garzotto, Pietro Piazzolla, and Massimo Quadrana. Content-based video recommendation system based on stylistic visual features. *Journal on Data Semantics*, 5:99–113, 2016. 1

[16] 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. 8

[17] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019. 1, 5, 6

[18] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 5, 6

[19] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017. 3

[20] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 709–727. Springer, 2020. 1, 2, 3, 5, 6

[21] Qingqiu Huang, Yuanjun Xiong, Yu Xiong, Yuqi Zhang, and Dahua Lin. From trailers to storylines: An efficient way to learn from movies. *arXiv preprint arXiv:1806.05341*, 2018. 1, 2, 4, 5, 6

[22] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2, 2019. 3

[23] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8337–8346, 2019. 2

[24] 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, 5

[25] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 2

[26] Anna Kukleva, Makarand Tapaswi, and Ivan Laptev. Learning interactions and relationships between movie characters.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9849–9858, 2020. 2

[27] Joonseok Lee and Sami Abu-El-Haija. Large-scale content-only video recommendation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017. 1

[28] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 2, 4

[29] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 1, 2, 5, 6

[30] Rafael B Mangolin, Rodolfo M Pereira, Alceu S Britto Jr, Carlos N Silla Jr, Valéria D Feltrim, Diego Bertolini, and Yandre MG Costa. A multimodal approach for multi-label movie genre classification. *Multimedia Tools and Applications*, pages 19071–19096, 2022. 2

[31] Yue Meng, Chung-Ching Lin, Rameswar Panda, Prasanna Sattigeri, Leonid Karlinsky, Aude Oliva, Kate Saenko, and Rogerio Feris. Ar-net: Adaptive frame resolution for efficient action recognition. In *European Conference on Computer Vision*, pages 86–104. Springer, 2020. 2

[32] Seung-Bo Park, Yoo-Won Kim, Mohammed Nazim Uddin, and Geun-Sik Jo. Character-net: Character network analysis from video. In *2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 305–308. IEEE, 2009. 2

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 5

[34] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *European conference on computer vision*, pages 540–555. Springer, 2014. 2

[35] 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. 3

[36] 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

[37] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023. 2, 3, 4, 5

[38] 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. 2, 3, 8

[39] 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. 2

[40] Zeeshan Rasheed, Yaser Sheikh, and Mubarak Shah. On the use of computable features for film classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):52–64, 2005. 1

[41] 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. 3

[42] 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. 2

[43] Arka Sadhu, Tanmay Gupta, Mark Yatskar, Ram Nevatia, and Aniruddha Kembhavi. Visual semantic role labeling for video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5589–5600, 2021. 2

[44] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision*, pages 510–526. Springer, 2016. 2

[45] Gabriel S Simões, Jônatas Wehrmann, Rodrigo C Barros, and Duncan D Ruiz. Movie genre classification with convolutional neural networks. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 259–266. IEEE, 2016. 1, 2

[46] Tomáš Souček and Jakub Lokoč. Transnet v2: An effective deep network architecture for fast shot transition detection. *arXiv preprint arXiv:2008.04838*, 2020. 2, 4, 5

[47] Tomáš Souček, Jaroslav Moravec, and Jakub Lokoč. Transnet: A deep network for fast detection of common shot transitions. *arXiv preprint arXiv:1906.03363*, 2019. 2

[48] 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. 8

[49] 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. 2

[50] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 2

[51] Paul-Louis Thirard and Lorenzo Codelli. Robert sklar. film, an international history of the medium, 1993;; kristin thomp-

son, david bordwell. film history, an introduction, 1994. *1895, revue d'histoire du cinéma*, 17(1):170–170, 1994. 1

[52] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[53] Ba Tu Truong and Svetha Venkatesh. Video abstraction: A systematic review and classification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 3(1):3–es, 2007. 2

[54] Bo Wang, Youjiang Xu, Yahong Han, and Richang Hong. Movie question answering: Remembering the textual cues for layered visual contents. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[55] 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. 2, 4, 5, 6

[56] Chung-Yi Weng, Wei-Ta Chu, and Ja-Ling Wu. Rolenet: Movie analysis from the perspective of social networks. *IEEE Transactions on Multimedia*, 11(2):256–271, 2009. 2

[57] Jiangyue Xia, Anyi Rao, Qingqiu Huang, Linning Xu, Jiangtao Wen, and Dahua Lin. Online multi-modal person search in videos. In *European Conference on Computer Vision*, pages 174–190. Springer, 2020. 2

[58] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020. 1, 6

[59] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. *Advances in Neural Information Processing Systems*, 34:1086–1099, 2021. 2

[60] Zhongping Zhang, Youzuo Lin, Zheng Zhou, and Tianlang Chen. Adaptive filtering for event recognition from noisy signal: An application to earthquake detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3327–3331. IEEE, 2019. 5

[61] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 803–818, 2018. 1, 2, 5, 6

[62] 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. 8

[63] Howard Zhou, Tucker Hermans, Asmita V Karandikar, and James M Rehg. Movie genre classification via scene categorization. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 747–750, 2010. 1, 2

[64] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27, 2015. 2