

MovieNet: A Holistic Dataset for Movie Understanding

Qingqiu Huang*, Yu Xiong*, Anyi Rao, Jiaze Wang, and Dahua Lin

CUHK-SenseTime Joint Lab, The Chinese University of Hong Kong
 {hq016, xy017, ra018, dhlin}@ie.cuhk.edu.hk
 jzwang@link.cuhk.edu.hk

Abstract. Recent years have seen remarkable advances in visual understanding. However, how to understand a story-based long video with artistic styles, *e.g.* movie, remains challenging. In this paper, we introduce MovieNet – a holistic dataset for movie understanding. MovieNet contains 1,100 movies with a large amount of multi-modal data, *e.g.* trailers, photos, plot descriptions, *etc.*. Besides, different aspects of manual annotations are provided in MovieNet, including 1.1M characters with bounding boxes and identities, 42K scene boundaries, 2.5K aligned description sentences, 65K tags of place and action, and 92K tags of cinematic style. To the best of our knowledge, MovieNet is the largest dataset with richest annotations for comprehensive movie understanding. Based on MovieNet, we set up several benchmarks for movie understanding from different angles. Extensive experiments are executed on these benchmarks to show the immeasurable value of MovieNet and the gap of current approaches towards comprehensive movie understanding. We believe that such a holistic dataset would promote the researches on story-based long video understanding and beyond. MovieNet will be published in compliance with regulations at <https://movienet.github.io>.

1 Introduction

“You jump, I jump, right?” When Rose gives up the lifeboat and exclaims to Jack, we are all deeply touched by the beautiful moving love story told by the movie *Titanic*. As the saying goes, “Movies dazzle us, entertain us, educate us, and delight us”. Movie, where characters would face various situations and perform various behaviors in various scenarios, is a reflection of our real world. It teaches us a lot such as the stories took place in the past, the culture and custom of a country or a place, the reaction and interaction of humans in different situations, *etc.*. Therefore, to understand movies is to understand our world.

It goes not only for human, but also for an artificial intelligence system. We believe that movie understanding is a good arena for high-level machine intelligence, considering its high complexity and close relation to the real world. What’s more, compared to web images [16] and short videos [7], the hundreds

* Equal contribution

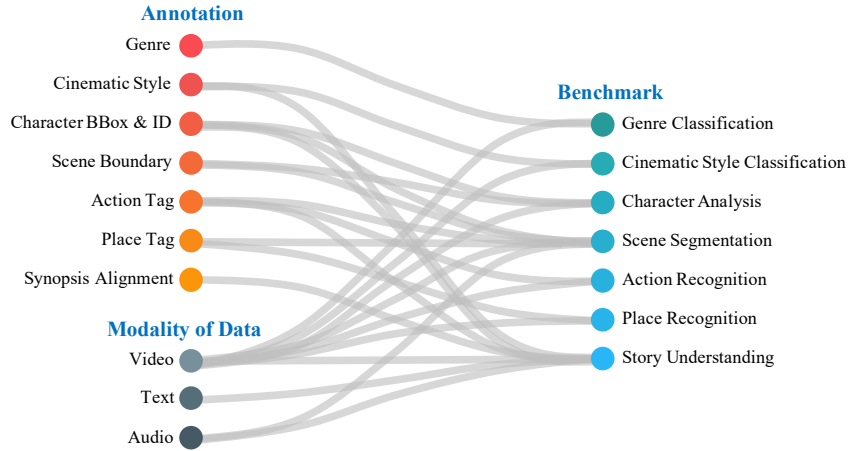


Fig. 1: The data, annotation, benchmark and their relations in MovieNet, which together build a holistic dataset for comprehensive movie understanding.

of thousands of movies in history containing rich content and multi-modal information become better nutrition for the data-hungry deep models.

Motivated by the insight above, we build a holistic dataset for movie understanding named *MovieNet* in this paper. As shown in Fig. 1, MovieNet comprises three important aspects, namely *data*, *annotation*, and *benchmark*.

First of all, MovieNet contains a large volume of data in multiple modalities, including movies, trailers, photos, subtitles, scripts and meta information like genres, cast, director, rating *etc.*. There are totally 3K hour-long videos, 3.9M photos, 10M sentences of text and 7M items of meta information in MovieNet.

From the annotation aspect, MovieNet contains massive labels to support different research topics of movie understanding. Based on the belief that middle-level entities, *e.g.* character, place, are important for high-level story understanding, various kinds of annotations on semantic elements are provided in MovieNet, including character bounding box and identity, scene boundary, action/place tag and aligned description in natural language. In addition, since movie is an art of filming, the cinematic styles, *e.g.*, view scale, camera motion, lighting, *etc.*, are also beneficial for comprehensive video analysis. Thus we also annotate the view scale and camera motion for more than 46K shots. Specifically, the annotations in MovieNet include: (1) 1.1M characters with bounding boxes and identities; (2) 40K scene boundaries; (3) 65K tags of action and place; (4) 12K description sentences aligned to movie segments; (5) 92K tags of cinematic styles.

Based on the data and annotations in MovieNet, we exploit some research topics that cover different aspects of movie understanding, *i.e.* genre analysis, cinematic style prediction, character analysis, scene understanding, and movie segment retrieval. For each topic, we set up one or several challenging benchmarks. Then extensive experiments are executed to present the performances



Fig. 2: MovieNet is a holistic dataset for movie understanding, which contains massive data from different modalities and high-quality annotations in different aspects. Here we show some data (in blue) and annotations (in green) of *Titanic* in MovieNet.

of different methods. By further analysis on the experimental results, we will also show the gap of current approaches towards comprehensive movie understanding, as well as the advantages of holistic annotations for throughout video analytics.

To the best of our knowledge, MovieNet is the first holistic dataset for movie understanding that contains a large amount of data from different modalities and high-quality annotations in different aspects. We hope that it would promote the researches on video editing, human-centric situation understanding, story-based video analytics and beyond.

2 Related Datasets

Existing Works. Most of the datasets of movie understanding focus on a specific element of movies, *e.g.* genre [89,63], character [1,3,31,48,65,22,35], action [39,21,46,5,6], scene [53,11,30,49,15,51] and description [61]. Also their scale is quite small and the annotation quantities are limited. For example, [22,65,3] take several episodes from TV series for character identification, [39] uses clips from twelve movies for action recognition, and [49] exploits scene segmentation with only three movies. Although these datasets focus on some important aspects of movie understanding, their scale is not enough for the data-hungry learning paradigm. Furthermore, the deep comprehension should go from middle-level elements to high-level story while each existing dataset can only support a single task, causing trouble for comprehensive movie understanding.

MovieQA. MovieQA [68] consists of 15K questions designed for 408 movies. As for sources of information, it contains video clips, plots, subtitles, scripts, and DVS (Descriptive Video Service). To evaluate story understanding by QA is a good idea, but there are two problems. (1) Middle-level annotations, *e.g.*, character identities, are missing. Therefore it is hard to develop an effective approach towards high-level understanding. (2) The questions in MovieQA come from the wiki plot. Thus it is more like a textual QA problem rather than story-based video understanding. A strong evidence is that the approaches based on textual plot can get a much higher accuracy than those based on “video+subtitle”.

LSMDC. LSMDC [57] consists of 200 movies with audio description (AD) providing linguistic descriptions of movies for visually impaired people. AD is quite different from the natural descriptions of most audiences, limiting the usage of the models trained on such datasets. And it is also hard to get a large number of ADs. Different from previous work [68,57], we provide multiple sources of textual information and different annotations of middle-level entities in MovieNet, leading to a better source for story-based video understanding.

AVA. Recently, AVA dataset [28], an action recognition dataset with 430 15-min movie clips annotated with 80 spatial-temporal atomic visual actions, is proposed. AVA dataset aims at facilitating the task of recognizing atomic visual actions. However, regarding the goal of story understanding, the AVA dataset is not applicable since (1) The dataset is dominated by labels like *stand* and *sit*, making it extremely unbalanced. (2) Actions like *stand*, *talk*, *watch* are less informative in the perspective of story analytics. Hence, we propose to annotate semantic level actions for both action recognition and story understanding tasks.

MovieGraphs. MovieGraphs [71] is the most related one that provides graph-based annotations of social situations depicted in clips of 51 movies. The annotations consist of characters, interactions, attributes, *etc.*. Although sharing the same idea of multi-level annotations, MovieNet is different from MovieGraphs in three aspects: (1) MovieNet contains not only movie clips and annotations, but also photos, subtitles, scripts, trailers, *etc.*, which can provide richer data for various research topics. (2) MovieNet can support and exploit different aspects of movie understanding while MovieGraphs focuses on situation recognition only. (3) The scale of MovieNet is much larger than MovieGraphs.

Table 1: Comparison between MovieNet and related datasets in terms of data.

	# movie	trailer	photo	meta	script	synop.	subtitle	plot	AD
MovieQA[68]	140						✓	✓	
LSMDC[57]	200				✓				✓
MovieGraphs[71]	51								
AVA[28]	430								
MovieNet	1,100	✓	✓	✓	✓	✓	✓	✓	

Table 2: Comparison between MovieNet and related datasets in terms of annotation.

	# character	# scene	# cine. tag	# aligned sent.	# action/place tag
MovieQA[68]	-	-	-	15K	-
LSMDC[57]	-	-	-	128K	-
MovieGraphs[71]	22K	-	-	21K	23K
AVA[28]	116K	-	-	-	360K
MovieNet	1.1M	42K	92K	25K	65K

3 Visit MovieNet: Data and Annotation

MovieNet contains various kinds of data from multiple modalities and high-quality annotations on different aspects for movie understanding. Fig. 2 shows the data and annotations of the movie *Titanic* in MovieNet. Comparisons between MovieNet and other datasets for movie understanding are shown in Tab. 1 and Tab. 2. All these demonstrate the tremendous advantage of MovieNet on both quality, scale and richness.

3.1 Data in MovieNet

Movie. We carefully selected and purchased the copies of 1,100 movies, the criteria of which are (1) colored; (2) longer than 1 hour; (3) cover a wide range of genres, years and countries.

Metadata. We get the meta information of the movies from IMDb and TMDb¹, including title, release date, country, genres, rating, runtime, director, cast, storyline, *etc.*. Here we briefly introduce some of the key elements, please refer to supplementary material for detail: (1) Genre is one of the most important attributes of a movie. There are total 805K genre tags from 28 unique genres in MovieNet. (2) For cast, we get both their names, IMDb IDs and the character names in the movie. (3) We also provide IMDb ID, TMDb ID and Douban ID of each movie, with which the researchers can get additional meta information from these websites conveniently. The total number of meta information in MovieNet is 375K. Please note that each kind of data itself, even without the movie, can support some research topics [37]. So we try to get each kind of data as much

¹ IMDb: <https://www.imdb.com>; TMDb: <https://www.themoviedb.org>

as we can. Therefore the number here is larger than 1,100. So as other kinds of data we would introduce below.

Subtitle. The subtitles are obtained in two ways. Some of them are extracted from the embedded subtitle stream in the movies. For movies without original English subtitle, we crawl the subtitles from YIFY². All the subtitles are manually checked to ensure that they are aligned to the movies.

Trailer. We download the trailers from YouTube according to their links from IMDb and TMDb. We found that this scheme is better than previous work [10], which use the titles to search trailers from YouTube, since the links of the trailers in IMDb and TMDb have been manually checked by the organizers and audiences. Totally, we collect 60K trailers belonging to 33K unique movies.

Script. Script, where the movement, actions, expression and dialogs of the characters are narrated, is a valuable textual source for research topics of movie-language association. We collect around 2K scripts from IMSDb and Daily Script³. The scripts are aligned to the movies by matching the dialog with subtitles.

Synopsis. A synopsis is a description of the story in a movie written by audiences. We collect 11K high-quality synopses from IMDb, all of which contain more than 50 sentences. Synopses are also manually aligned to the movie, which would be introduced in Sec. 3.2.

Photo. We collect 3.9M photos of the movies from IMDb and TMDb, including poster, still frame, publicity, production art, product, behind the scene and event.

3.2 Annotation in MovieNet

To provide a high-quality dataset supporting different research topics on movie understanding, we make great effort to clean the data and manually annotate various labels on different aspects, including character, scene, event and cinematic style. Here we just demonstrate the *content* and the *amount* of annotations due to the space limit. Please refer to supplementary material for details.

Cinematic Styles. Cinematic style, such as view scale, camera movement, lighting and color, is an important aspect of comprehensive movie understanding since it influences how the story is telling in a movie. In MovieNet, we choose two kinds of cinematic tags for study, namely view scale and camera movement. Specifically, the view scale include five categories, *i.e.* *long shot*, *full shot*, *medium shot*, *close-up shot* and *extreme close-up shot*, while the camera movement is divided into four classes, *i.e.* *static shot*, *pans and tilts shot*, *zoom in* and *zoom out*. The original definitions of these categories come from [26] and we simplify them for research convenience. We totally annotate 47K shots from movies and trailers, each with one tag of view scale and one tag of camera movement.

Character Bounding Box and Identity. Person plays an important role in human-centric videos like movies. Thus to detect and identify characters is a foundational work towards movie understanding. The annotation process of

² <https://www.yifysubtitles.com/>

³ IMSDb: <https://www.imsdb.com/>; DailyScript: <https://www.dailyscript.com/>

character bounding box and identity contains 4 steps: (1) Some key frames, the number of which is 758K, from different movies are selected for bounding box annotation. (2) A detector is trained with the annotations in step-1. (3) We use the trained detector to detect more characters in the movies and manually clean the detected bounding boxes. (4) We then manually annotate the identities of all the characters. To make the cost affordable, we only keep the top 10 cast in credits order according to IMDb, which can cover the main characters for most movies. Characters not belong to credited cast were labeled as “others”. In total, we got 1.1M instances of 3,087 unique credited cast and 364K “others”.

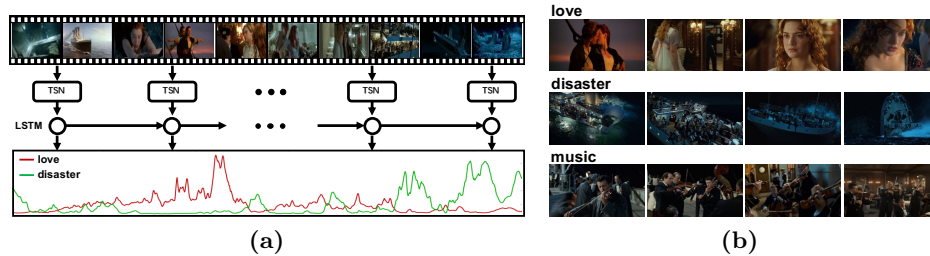
Scene Boundary. In terms of temporal structure, a movie contains two hierarchical levels – shot, and scene. Shot is the minimal visual unit of a movie while scene is a sequence of continued shots that are semantically related. To capture the hierarchical structure of a movie is important for movie understanding. Shot boundary detection has been well solved by [62], while scene boundary detection, also named scene segmentation, remains an open question. In MovieNet, we manually annotate the scene boundaries to support the researches on scene segmentation, resulting in 42K scenes.

Action/Place Tags. To understand the event(s) happened within a scene, action and place tags are required. Hence, we first split each movie into clips according to the scene boundaries and then manually annotated place and action tags for each segment. For place annotation, each clip is annotated with multiple place tags, *e.g.*, {deck, cabin}. While for action annotation, we first detect sub-clips that contain characters and actions, then we assign multiple action tags to each sub-clip. We have made the following efforts to keep tags diverse and informative: (1) We encourage the annotators to create new tags. (2) Tags that convey little information for story understanding, *e.g.*, *stand* and *talk*, are excluded. Finally, we merge the tags and filtered out 80 actions and 90 places with a minimum frequency of 25 as the final annotations. In total, there are 42K segments with 19.6K place tags and 45K action tags.

Description Alignment Since the event is more complex than character and scene, a proper way to represent an event is to describe it with natural language. Previous works have already aligned script [46], Descriptive Video Service (DVS) [57], book [91] or wiki plot [66,67,68] to movies. However, books cannot be well aligned since most of the movies would be quite different from their books. DVS transcripts are quite hard to obtain, limiting the scale of the datasets based on them [57]. Wiki plot is usually a short summary that cannot cover all the important events of the movie. Considering the issues above, we choose synopses as the story descriptions in MovieNet. The associations between the movie segments and the synopsis paragraphs are manually annotated by three different annotators with a coarse-to-fine procedure. Finally, we obtained 4,208 highly consistent paragraph-segment pairs.

Table 3: (a). Comparison between MovieNet and other benchmarks for genre analysis. (b). Results of some baselines for genre classification in MovieNet

(a)					(b)				
	genre	movie	trailer	photo	Data	Model	r@0.5	p@0.5	mAP
MGCD[89]	4	-	1.2K	-	Photo	VGG16 [64]	27.32	66.28	32.12
LMTD[63]	4	-	3.5K	-		ResNet50 [32]	34.58	72.28	46.88
MScope[10]	13	-	5.0K	5.0K	Trailer	TSN-r50[74]	17.95	78.31	43.70
MovieNet	21	1.1K	68K	1.6M		I3D-r50 [9]	16.54	69.58	35.79
						TRN-r50 [86]	21.74	77.63	45.23

**Fig. 3:** (a). Framework of genre analysis in movies. (b). Some samples of genre-guided trailer generation for movie *Titanic*.

4 Play with MovieNet: Benchmark and Analysis

With a large amount of data and holistic annotations, MovieNet can support various research topics. In this section, we try to analyze movies from five aspects, namely *genre*, *cinematic style*, *character*, *scene* and *story*. For each topic, we would set up one or several benchmarks based on MovieNet. Baselines with currently popular techniques and analysis on experimental results are also provided to show the potential impact of MovieNet in various tasks. The topics of the tasks have covered different perspectives of comprehensive movie understanding. But due to the space limit, here we can only touched the tip of the iceberg. More detailed analysis are provided in the supplementary material and more interesting topics to be exploited are introduced in Sec. 5.

4.1 Genre Analysis

Genre is a key attribute for any media with artistic elements. To classify the genres of movies has been widely studied by previous works [89,63,10]. But there are two drawbacks for these works. (1) The scale of existing datasets is quite small. (2) All these works focus on image or trailer classification while ignore a more important problem, *i.e.* how to analyze the genres of a long video.

MovieNet provides a large-scale benchmark for genre analysis, which contains 1.1K movies, 68K trailers and 1.6M photos. The comparison between different datasets are shown in Tab. 3a, from which we can see that MovieNet is much larger than previous datasets.

Based on MovieNet, we first provide baselines for both image-based and video-based genre classification, the results are shown Tab. 3b. Comparing the result of genre classification in small datasets [63,10] to ours in MovieNet, we find that the performance drops a lot when the scale of the dataset become larger. The newly proposed MovieNet brings two challenges to previous methods. (1) Genre classification in MovieNet becomes a long-tail recognition problem where the label distribution is extremely unbalanced. For example, the number of “Drama” is 40 times larger than that of “Sport” in MovieNet. (2) Genre is a high-level semantic tag depending on action, clothing and facial expression of the characters, and even BGM. Current methods are good at visual representation. When facing a problem that need to consider higher-level semantics, they would all fail. We hope MovieNet would promote researches on these challenging topics.

Another new issue to address is how to analyze the genres of a movie. Since movie is extremely long and not all segments are related to its genres, this problem is much more challenging. Following the idea of learning from trailers and applying to movies [36], we adopt the visual model trained with trailers as shot-level feature extractor. Then the features are fed to a temporal model to capture the temporal structure of the movie. The overall framework is shown in Fig. 3a. With this approach, we can get the genre response curve of a movie. Specifically, we can predict which part of the movie is more relevant to a specific genre. What’s more, the prediction can also be used for genre-guided trailer generation, as shown in Fig. 3b. From the analysis above, we can see that MovieNet would promote the development of this challenging and valuable research topic.

4.2 Cinematic Style Analysis

As we mentioned before, cinematic style is about how to present the story to audience in the perspective of filming art. For example, a *zoom in* shot is usually used to attract the attention of audience to a specific object. In fact, cinematic

Table 4: (a). Comparison between MovieNet and other benchmarks for cinematic style prediction. (b). Results of some baselines for cinematic style prediction in MovieNet

(a)				(b)		
	shot	video	scale move.	Method	scale acc.	move. acc.
Lie 2014 [4]	327	327	✓	I3D [9]	76.79	78.45
Sports 2007 [82]	1,364	8	✓	TSN [74]	84.08	70.46
Context 2011 [80]	3,206	4	✓	TSN+R ³ Net[17]	87.50	80.65
Taxon 2009 [72]	5,054	7	✓			
MovieNet	46,857	7,858	✓			

Table 5: Datasets for person analysis.

	ID	instance	source
COCO[43]	-	262K	web image
CalTech[19]	-	350K	surveillance
Market[85]	1,501	32K	surveillance
CUHK03[40]	1,467	28K	surveillance
AVA[28]	-	426K	movie
CSM[34]	1,218	127K	movie
MovieNet	3,087	1.1M	movie

**Fig. 4:** Persons in different data sources**Table 6:** Results of (a). Character Detection and (b).Character Identification

(a)			(b)			
Train Data	Method	mAP	Train Data	cues	Method	mAP
COCO[43]	FasterRCNN	81.50	Market[85]	body	r50-softmax	4.62
Caltech[19]	FasterRCNN	5.67	CUHK03[40]	body	r50-softmax	5.33
CSM[34]	FasterRCNN	89.91	CSM[34]	body	r50-softmax	26.21
MovieNet	FasterRCNN	92.13	MovieNet	body	r50-softmax	32.81
	RetinaNet	91.55		body+face	two-step[45]	63.95
	CascadeRCNN	95.17		body+face	PPCC[34]	75.95

style is crucial for both video understanding and editing. But there are few works focusing on this topic and no large-scale datasets for this research topic too.

Based on the tags of cinematic style we annotated in MovieNet, we set up a benchmark for cinematic style prediction. Specifically, we would like to recognize the view scale and camera motion of each shot. Comparing to existing datasets, MovieNet is the first dataset that covers both view scale and camera motion, and it is also much larger, as shown in Tab. 4a. Several models for video clip classification such as TSN [74] and I3D [9] are applied to tackle this problem, the results are shown in Tab. 4b. Since the view scale depends on the portion of the subject in the shot frame, to detect the subject is important for cinematic style prediction. Here we adopt the approach from saliency detection [17] to get the subject maps of each shot, with which better performances are achieved, as shown in Tab. 4b. Although utilizing subject points out a direction for this task, there is still a long way to go. We hope that MovieNet can promote the development of this important but ignored topic for video understanding.

4.3 Character Recognition

It has been shown by existing works [71,75,45] that movie is a human-centric video where characters play an important role. Therefore, to detect and identify characters is crucial for movie understanding. Although person/character recognition is not a new task, all previous works either focus on other data

Table 7: Dataset for scene analysis.

	scene	action	place
OVSD [59]	300	-	-
BBC [2]	670	-	-
Hollywood2 [46]	-	1.7K	1.2K
MovieGraph[71]	-	23.4K	7.6K
AVA [28]	-	360K	-
MovieNet	42K	45.0K	19.6K

Table 8: Datasets for story understanding in movies in terms of (1) number of sentences per movie; (2) duration (second) per segment.

Dataset	sent./mov.	dur./seg.
MovieQA [68]	35.2	202.7
MovieGraphs [71]	408.8	44.3
MovieNet	83.4	428.0

sources [85,40,43] or small-scale benchmarks [31,3,65], leading to the results lack of convincingness for character recognition in movies.

We proposed two benchmarks for character analysis in movies, namely, character detection and character identification. We provide more than 1.1M instances from 3,087 identities to support these benchmarks. As shown in Tab. 5, MovieNet contains much more instances and identities comparing to some popular datasets about person analysis. The following sections will show the analysis on character detection and identification respectively.

Character Detection. Images from different data sources would have large domain gap, as shown in Fig. 4. Therefore, a character detector trained on general object detection dataset, *e.g.* COCO [43], or pedestrian dataset, *e.g.* Cal-Tech [19], is not good enough for detecting characters in movies. This can be supported by the results shown in Tab. 6a. To get a better detector for character detection, we train different popular models [55,42,8] with MovieNet using toolboxes from [13,12]. We can see that with the diverse character instances in MovieNet, a Cascade R-CNN trained with MovieNet can achieve extremely high performance, *i.e.* 95.17% in mAP. That is to say, character detection can be well solved by a large-scale movie dataset with current SOTA detection models. This powerful detector would then benefit research on character analysis in movies.

Character Identification. To identify the characters in movies is a more challenging problem, which can be observed by the diverse samples shown in Fig. 4. We conduct different experiments based on MovieNet, the results are shown in Tab. 6b. From these results, we can see that: (1) models trained on ReID datasets are inefficient for character recognition due to domain gap; (2) to aggregate different visual cues of an instance is important for character recognition in movies; (3) the current state-of-the-art can achieve 75.95% mAP, which demonstrates that it is a challenging problem which need to be further exploited.

4.4 Scene Analysis

As mentioned before, scene is the basic semantic unit of a movie. Therefore, it is important to analyze the scenes in movies. The key problems in scene understanding is probably *where is the scene boundary* and *what is the content in a scene*. As shown in Tab. 7, MovieNet, which contains more than 43K scene boundaries and 65K action/place tags, is the only one that can support both

Table 9: Results of Scene Segmentation

Dataset	Method	AP(\uparrow)	M_{iou} (\uparrow)
OVSD [59]	MS-LSTM	0.313	0.387
BBC [2]	MS-LSTM	0.334	0.379
MovieNet	Grouping [59]	0.336	0.372
	Siamese [2]	0.358	0.396
	MS-LSTM	0.465	0.462

Table 10: Results of Scene Tagging

Tags	Method	mAP
action	TSN [74]	14.17
	I3D [9]	20.69
	SlowFast [23]	23.52
place	I3D [9]	7.66
	TSN [74]	8.33

scene segmentation and *scene tagging*. What’s more, the scale of MovieNet is also larger than all previous works.

Scene Segmentation We first test some baselines [59,2] for scene segmentation. In addition, we also propose a sequential model, named Multi-Semantic LSTM (MS-LSTM) based on Bi-LSTMs [27,52] to study the gain brought by using multi-modality and multiple semantic elements, including audio, character, action and scene. From the results shown in Tab. 9, we can see that (1) Benefited from large scale and high diversity, models trained on MovieNet can achieve better performance. (2) Multi-modality and multiple semantic elements are important for scene segmentation, which highly raise the performance.

Action/Place Tagging To further understand the stories within a movie, it is essential to perform analytics on the key elements of storytelling, *i.e.*, place and action. We would introduce two benchmarks in this section. Firstly, for action analysis, the task is multi-label action recognition that aims to recognize all the human actions or interactions in a given video clip. We implement three standard action recognition models, *i.e.*, TSN [74], I3D [9] and SlowFast Network [23] modified from [83] in experiments. Results are shown in Tab. 10. For place analysis, we propose another benchmark for multi-label place classification. We adopt I3D [9] and TSN [74] as our baseline models and the results are shown in Tab. 10. From the results, we can see that action and place tagging is an extremely challenging problem due to the high diversity of different instances.

4.5 Story Understanding

Web videos are broadly adopted in previous works [7,79] as the source of video understanding. Compared to web videos, the most distinguishing feature of movies is the story. Movies are created to tell stories and the most explicit way to demonstrate a story is to describe it using natural language, *e.g.* synopsis. Inspired by the above observations, we choose the task of movie segment retrieval with natural language to analyze the stories in movies. Based on the aligned synopses in MovieNet, we set up a benchmark for movie segment retrieval. Specifically, given a synopsis paragraph, we aim to find the most relevant movie segment that covers the story in the paragraph. It is a very challenging task due to the rich content in movie and high-level semantic descriptions in synopses. Tab. 8 shows the comparison of our benchmark dataset with other related datasets. We can see that our dataset is more complex in terms of descriptions

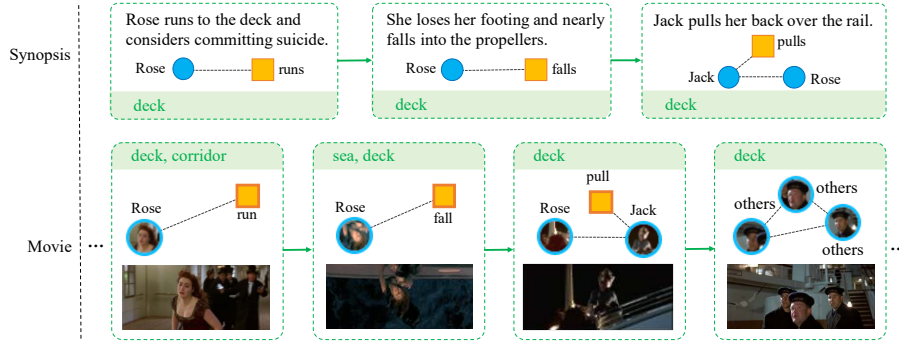


Fig. 5: Example of synopsis paragraph and movie segment in MovieNet-MSR. It demonstrate the spatial-temporal structures of stories in movies and synopses. We can also see that character, action and place are the key element for story understanding.

compared with MovieQA [68] while the segments are longer and contain more information than those of MovieGraphs [71].

Generally speaking, a story can be summarized as “*somebody do something in some time at some place*”. As shown in Fig. 5, both stories represented by language and video can be composed as sequences of {character, action, place} graphs. That being said, to understand a story is to (1) recognize the key elements of story-telling, namely, character, action, place *etc.*; (2) analyze the spatial-temporal structures of both movie and synopsis. Hence, our method first leverage middle-level entities (*e.g.* character, scene), as well as multi-modality (*e.g.* subtitle) to assist retrieval. Then we explore the spatial-temporal structure from both movies and synopses by formulating middle-level entities into graph structures. Please refer to supplementary material for details.

Using middle-level entities and multi-modality. We adopt VSE [25] as our baseline model where the vision and language features are embedded into a joint space. Specifically, the feature of the paragraph is obtained by taking the average of Word2Vec [47] feature of each sentence while the visual feature is obtained by taking the average of the appearance feature extracted from ResNet [32] on each shot. We add subtitle feature to enhance visual feature. Then different semantic elements including character, action and cinematic style are aggregated in our framework. We are able to obtain action features and character features thanks to the models trained on other benchmarks on MovieNet, *e.g.*, action recognition and character detection. Furthermore, we observe that the focused elements vary under different cinematic styles. For example, we should focus more on actions in a full shot while more on character and dialog in a close-up shot. Motivated by this observation, we propose a cinematic-style-guided attention module that predicts the weights over each element (*e.g.*, action, character) within a shot, which would be used to enhance the visual features. The experimental results are shown in Tab. 11. Experiments show that by considering different elements of the movies, the performance improves a lot. We can see that a holistic dataset

Table 11: Results of movie segment retrieval. Here, G stands for global appearance feature, S for subtitle feature, A for action, P for character and C for cinematic style.

Method	Recall@1	Recall@5	Recall@10	MedR
Random	0.11	0.54	1.09	460
G	3.16	11.43	18.72	66
G+S	3.37	13.17	22.74	56
G+S+A	5.22	13.28	20.35	52
G+S+A+P	18.50	43.96	55.50	7
G+S+A+P+C	18.72	44.94	56.37	7
MovieSynAssociation [77]	21.98	51.03	63.00	5

which contains holistic annotations to support middle-level entity analyses is important for movie understanding.

Explore spatial-temporal graph structure in movies and synopses. Simply adding different middle-level entities improves the result. Moreover, as shown in Fig. 5, we observe that stories in movies and synopses persist two important structure: (1) the temporal structure in movies and synopses is that the story can be composed as a sequence of events following a certain temporal order. (2) the spatial relation of different middle-level elements, *e.g.*, character co-existence and their interactions, can be formulated as graphs. We implement the method in [77] to formulate the above structures as two graph matching problems. The result are shown in Tab. 11. Leveraging the graph formulation for the internal structures of stories in movies and synopses, the retrieval performance can be further boosted, which in turn, show that the challenging MovieNet would provide a better source to story-based movie understanding.

5 Discussion and Future Work

In this paper, we introduce MovieNet, a holistic dataset containing different aspects of annotations to support comprehensive movie understanding.

We introduce several challenging benchmarks on different aspects of movie understanding, *i.e.* discovering filming art, recognizing middle-level entities and understanding high-level semantics like stories. Furthermore, the results of movie segment retrieval demonstrate that integrating filming art and middle-level entities according to the internal structure of movies would be helpful for story understanding. These in turn, show the effectiveness of holistic annotations.

In the future, our work would go on in two aspects. (1) **Extending the Annotation.** Currently our dataset covers 1,100 movies. In the future, we would further extend the dataset to include more movies and annotations. (2) **Exploring more Approaches and Topics.** To tackle the challenging tasks proposed in the paper, we would explore more effective approaches. Besides, there are more meaningful and practical topics that can be addressed with MovieNet from the perspective of video editing, such as movie deoldify, trailer generation, *etc.*

Supplementary Material

In the following sections, we provide overall details about MovieNet, including data, annotation, experiments and the toolbox. The content is organized as follows:

(1) We provide details about the content of particular data and how to collect and clean them in Sec. A:

- **Meta Data.** The list of meta data is given followed by the content of these meta data. See Sec. A.1.
- **Movie.** The statistics of the movies are provided. See Sec. A.2
- **Subtitle.** The collection and post-processing procedure of obtaining and aligning subtitles are given. See Sec. A.3.
- **Trailer.** We provide the process of selecting and processing the trailers. See Sec. A.4.
- **Script.** We automatically align the scripts to movies. The details of the method will be presented. See Sec. A.5.
- **Synopsis.** The statistics of synopsis will be introduced. See Sec. A.6.
- **Photo.** The statistics and some examples of photo will be shown. See Sec. A.7.

(2) We demonstrate annotation in MovieNet with the description about the design of annotation interface and workflow, see Sec. B.

- **Character Bounding Box and Identity.** We provide step by step procedure of collecting images and annotating the images with a semi-automatic algorithm. See Sec. B.1.
- **Cinematic Styles.** We present the analytics on cinematic styles and introduce the workflow and interface of annotating cinematic styles. See Sec. B.2.
- **Scene Boundaries.** We demonstrate how to effectively annotate scene boundaries with the help of an optimized annotating workflow. See Sec. B.3.
- **Action and Place Tags.** We describe the procedure of jointly labeling the action and place tags over movie segments. The workflow and interface are presented. See Sec. B.4.
- **Synopsis Alignment.** We provide the introduction of an efficient coarse-to-fine annotating workflow to align a synopsis paragraph to a movie segment. See Sec. B.5.
- **Trailer Movie Alignment.** We introduce a automatic approach that align shots in trailers to the original movies. This annotation facilitate tasks like trailer generation. See Sec. B.6.

(3) We set up several benchmarks on our MovieNet and conduct experiments on each benchmark. The implementation details of experiments on each benchmark will be introduced in Sec. C:

- **Genre Classification.** Genre Classification is a multi-label classification task build on MovieNet genre classification benchmark. See details at Sec. C.1.

- **Cinematic Styles Analysis.** On MovieNet cinematic style prediction benchmark, there are two classification tasks, namely *scale classification* and *movement classification*. See Sec. C.2 for implementation details.
- **Character Detection.** We introduce the detection task as well as model, implementation details on MovieNet character detection benchmarks. See Sec. C.3.
- **Character Identification.** We further introduce the challenging benchmark setting for MovieNet character identification. See details in Sec. C.4.
- **Scene Segmentation.** The scene segmentation task is a boundary detection task for cutting the movie by scene. The details about feature extraction, baseline models and evaluation protocols will be introduced in Sec. C.5.
- **Action Recognition.** We present the task of multi-label action classification task on MovieNet with the details of baseline models and experimental results. See Sec. C.6.
- **Place Recognition.** Similarly, we present the task of multi-label place classification task on MovieNet. See Sec. C.7.
- **Story Understanding.** For story understanding, we leverage the benchmark MovieNet segment retrieval to explore the potential of overall analytics using different aspects of MovieNet. The experimental settings and results will be found in Sec. C.8.

(4) To manage all the data and provide support for all the benchmarks, we build up a codebase for managing MovieNet with handy processing tools. Besides the codes for the benchmarks, we would also release this toolbox, the features of this tool box are introduced in Sec. D

A Data in MovieNet

MovieNet contains various kinds of data from multiple modalities and high-quality annotations on different aspects for movie understanding. They are introduced in detail below. And for comparison, the overall comparison of the data in MovieNet with other related dataset are shown in Tab. A1.

A.1 Meta Data

MovieNet contains meta data of 375K movies. Note that the number of metadata is significantly large than the movies provided with video sources (*i.e.* 1,100) because we belief that metadata itself can support various of tasks. It is also worth noting that the metadata of all the 1,100 selected movies are included in this metadata set. Fig. A1 shows a sample of the meta data, which is from *Titanic*. More details of each item in the meta data would be introduced below.

- **IMDb ID.** IMDb ID is the ID of a movie in the IMDb website⁴. IMDb ID is usually a string begins with “tt” and follows with 7 or 8 digital numbers,

⁴ <https://www.imdb.com/>

Table A1: Comparison between MovieNet and related datasets in terms of data.

	movie	trailer	photo	meta	genre	script	synop.	subtitle	plot	AD
MovieScope [10]	-	5,027	5,027	5,027	13K	-	-	-	5,027	-
MovieQA [68]	140	-	-	-	-	-	-	408	408	-
LSMDC [57]	200	-	-	-	-	50	-	-	-	186
MovieGraphs [71]	51	-	-	-	-	-	-	-	-	-
AVA [28]	430	-	-	-	-	-	-	-	-	-
MovieNet	1,100	60K	3.9M	375K	805K	986	31K	5,388	46K	-

<pre>"imdb_id": "tt0120338", "tmdb_id": "597", "douban_id": "1292722", "title": "Titanic (1997)", "genres": ["Drama", "Romance"], "country": "USA", "version": [{ "runtime": "194 min", "description": "" }], "imdb_rating": 7.7, "director": [{ "id": "nm0000116", "name": "James Cameron" }] }</pre>	<pre>"writer": [{ "id": "nm0000116", "name": "James Cameron", "description": "written by" }], "cast": [{ "id": "nm0000138", "name": "Leonardo DiCaprio", "character": "Jack Dawson" }, { "id": "nm0000701", "name": "Kate Winslet", "character": "Rose Dewitt Bukater" }, ...]</pre>	<pre>"overview": "84 years later, a 101-year-old woman named Rose Dewitt Bukater tells the story to her granddaughter Lizzy Calvert, ...", "storyline": "... And she explains the whole story from departure until the death of Titanic on its first and last voyage April 15th, 1912 at 2:20 in the morning ...", "plot": "... They recover a safe containing a drawing of a young woman wearing only the necklace dated April 14, 1912, the day the ship struck the iceberg ...", "synopsis": "... Also boarding the ship at Southampton are Jack Dawson (Leonardo DiCaprio), a down-on-his-luck sketch artist, and his Italian friend Fabrizio (Danny Nucci) ..."</pre>
--	--	--

Fig. A1: A sample of metadata from the movie *Titanic*.

e.g. “tt0120338” for the movie *Titanic*. One can easily get some information of a movie from IMDb with its ID. For example, the homepage of *Titanic* is “<https://www.imdb.com/title/tt0120338/>”. The IMDb ID is also taken as the ID of a movie in MovieNet.

- **TMDb ID.** TMDb ID is the ID of a movie in the TMDb website⁵. We find that some of the content in TMDb is of higher-quality than IMDb. For example, TMDb provides different versions of trailers and higher resolution posters. Therefore, we take it as a supplement of IMDb. TMDb provides APIs for users to search for information. With the TMDb ID provided in MovieNet, one can easily get more information if needed.
- **Douban ID.** Douban ID is the ID of a movie in Douban Movie⁶. We find that for some Asian movies, such as those from China and Japan, IMDb and TMDb contains few information. Therefore, we turn to a Chinese movie website, namely Douban Movie, for more information of Asian movies. We also provide Douban ID for some of the movies in MovieNet for convenience.

⁵ <https://www.themoviedb.org/>

⁶ <https://movie.douban.com/>

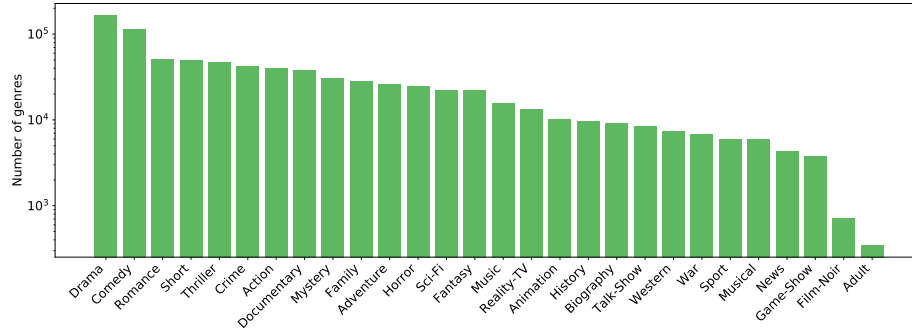


Fig. A2: Statistics of genres in metadata. It shows the number of genres for each genre category (y-axis in log-scale).

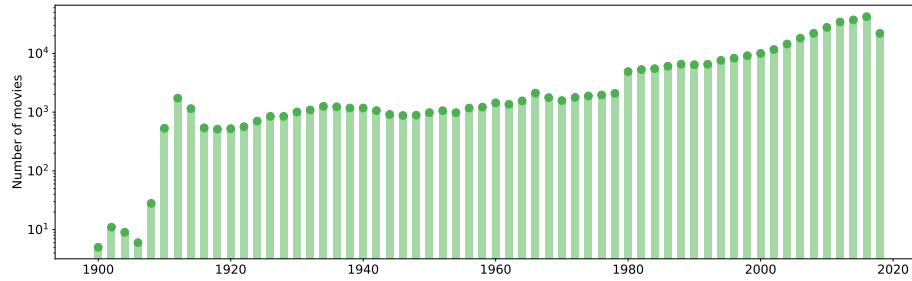


Fig. A3: Distribution of release date of the movies in metadata. It shows the number of movies in each year (y-axis in log-scale). Note that the number of movies generally increases as time goes by.

- **Version.** For movie with over one versions, *e.g.* normal version, director’s cut, we provide runtime and description of each version to help researchers align the annotations with their own resources.
- **Title.** The title of a movie following the format of IMDb, *e.g.*, *Titanic (1997)*.
- **Genres.** Genre is a category based on similarities either in the narrative elements or in the emotional response to the movie, *e.g.*, comedy, drama. There are totally 28 unique genres from the movies in MovieNet. Fig. A2 shows the distribution of the genres.
- **Release Date.** Release Date is the date when the movie published. Fig A3 shows the number of the movies released every year, from which we can see that the number of movies continuously grows every year.
- **Country.** Country refers to the country where the movie produced. The top-40 countries of the movies in MovieNet are shown in Fig. A4.
- **Version.** A movie may have multiple versions, *e.g.*, director’s cut, special edition. And different versions would have different runtimes. Here we provide the runtimes and descriptions of the movies in MovieNet.

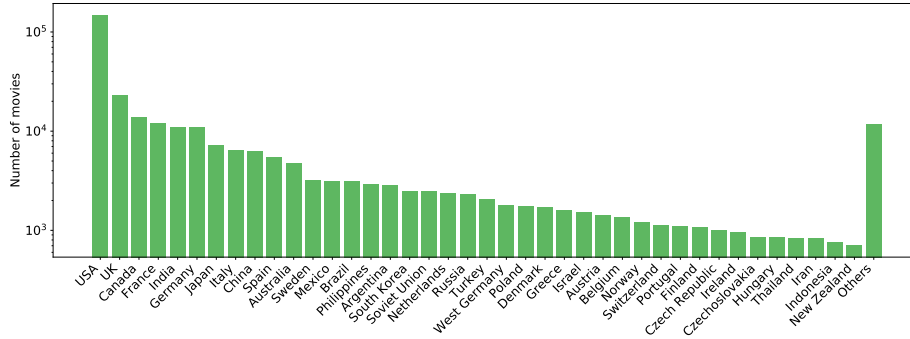


Fig. A4: The countries that the movies belong to in metadata. Here we show top 40 countries with the left as “Others”. The number of movies (y-axis) is in log-scale.

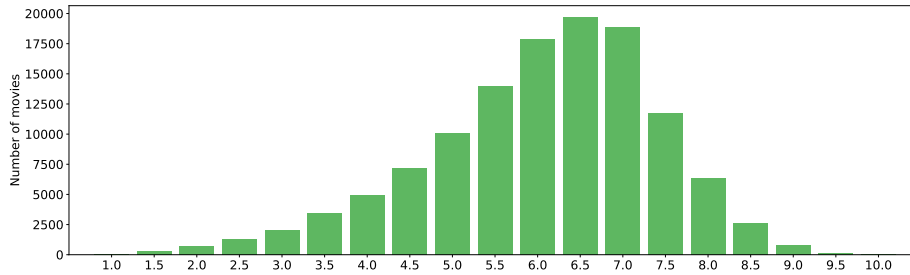


Fig. A5: Distribution of score ratings in MovieNet metadata.

- **IMDb Rating.** IMDb rating is the rating of the movie uploaded by the users. The distribution of different ratings are shown in Fig. A5.
- **Director.** Director contains the director’s name and ID.
- **Writer.** Writer contains the writer’s name and ID.
- **Cast.** A list of the cast in the movie, each of which contains the actor/actress’s name, ID and character’s name.
- **Overview.** Overview is a brief introduction of the movie, which usually covers the background and main characters of the movie.
- **Storyline.** Storyline is a plot summary of the movie. It is longer and contains more details than the overview.
- **Wiki Plot.** Wiki Plot is the summary of the movie from Wikipedia and is usually longer than overview and storyline.

A.2 Movie

As we introduced in our paper, there are 1,100 movies in MovieNet. Here we show some statistics of the 1,100 movies in Fig. A6, including the distributions of runtime and the shot number. As mentioned in Sec. A.1, in addition to these

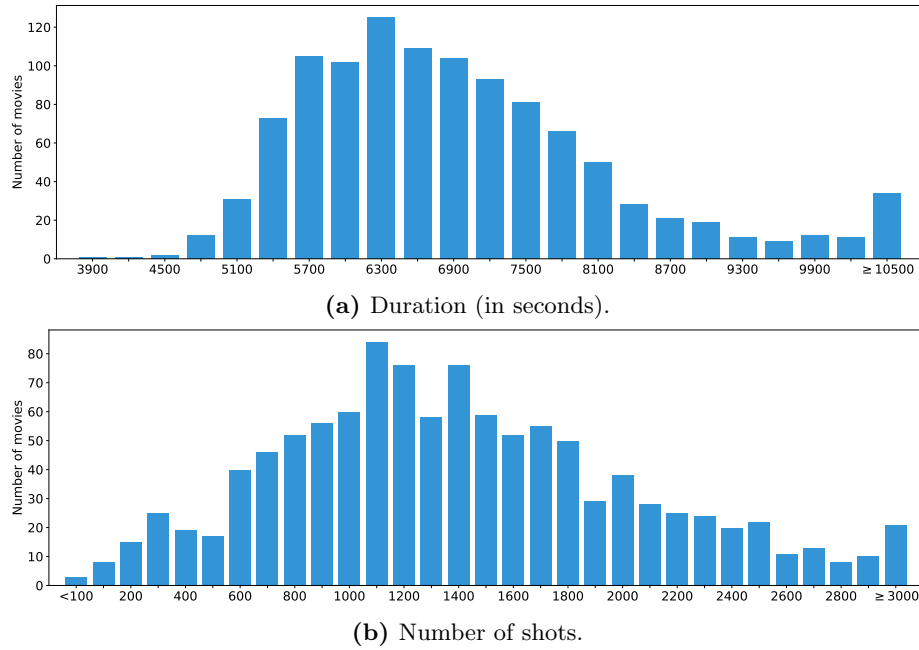


Fig. A6: Distribution of duration and number of shots for the 1,100 movies in MovieNet.

1,100 movies, we also provided metadata for other movies as much as we can. This also apply for other data like trailer and photo, and we would not clarify it in the next sections.

It is mentioned in the paper that we select the movie that covers a wide range of years, countries and genres. The distribution of these data are shown in Fig. A7. We can see that the movies are diversity in terms of year, country and genre.

Feature Representation. To play with a long video is nontrivial for the current deep learning framework and computational power. For the convenience of research, we propose multiply ways of feature representations for a movie.

- **Shot-based visual feature.** For most of the task, *e.g.* genre classification, shot-based representation is an efficient representation. A shot is a series of frames that runs for an uninterrupted period of time, which can be taken as the smallest visual unit of a movie. So we use shot-based representation for movies in our MovieNet. Specifically, we first separate each movie into shots with a shot detection tool [62]. Then, we sample three key frames and extract visual features using models pre-trained on ImageNet.
- **Audio feature.** For each shot, we also cut the audio wave within this shot and then extract audio feature [14] as the supplementary of visual feature.

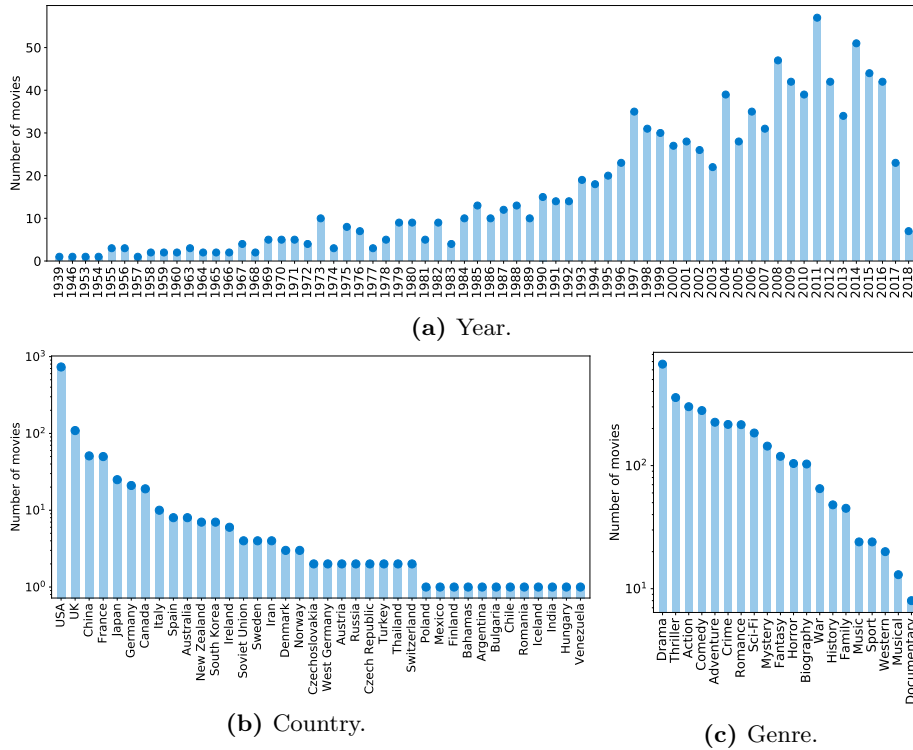


Fig. A7: Distribution of release year, countries and genres for the 1,100 movies in MovieNet (y-axis of country and genre in log scale).

- **Frame-based feature.** For those tasks like action recognition that need consider motion information, we also provide frame-based feature for these tasks.

A.3 Subtitle

For each movie from MovieNet, we provide an English subtitle that aligned with the movie. It is often the case that the downloaded subtitle is not aligned with the video because usually a movie has multiple versions, *e.g. director's cut* and *extended*. To make sure that each subtitle is aligned with the video source, before manually checking, we make the following efforts: (1) For the subtitle extracted from original video or downloaded from the Internet, we first make sure the subtitles are complete and are English version (by applying regular expression). (2) Then we clean the subtitle by removing noise such as HTML tags. (3) We leverage the off-the-shelf tool⁷ that transfers audio to text and matches text with the subtitle to produce a shift time. (4) We filtered out those subtitles with a

⁷ <https://github.com/smacke/subsync>

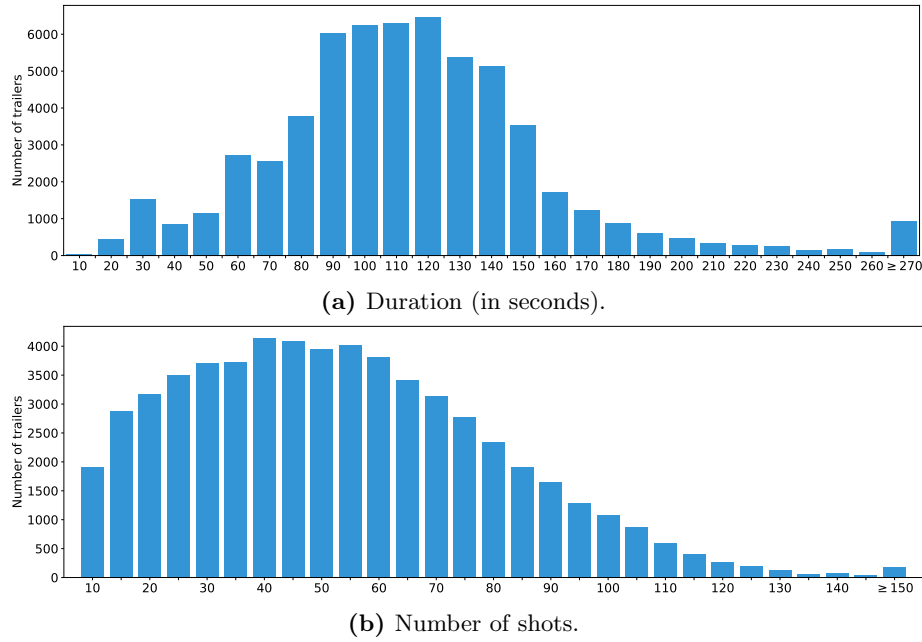


Fig. A8: Distribution of duration and number of shots for the trailers in MovieNet.

shift time surpasses a particular threshold and download another subtitle. After that, we manually download the subtitles that are still problematic, and then repeat the above steps.

Particularly, the threshold in step(4) is set to 60s by the following observations: (1) Most of the aligned subtitles have a shift within 1 seconds. (2) Some special cases, for example, a long scene without any dialog, would cause the tool to generate a shift of a few seconds. But the shift is usually less than 60s. (3) The subtitles that do not align with the original movies are usually either another version or crawled from another movie. In such cases, the shift will be larger than 60s.

After that, we ask annotators to manually check if the auto aligned subtitles are still misaligned. It turns out that the auto alignment are quite effective that few of the subtitles are still problematic.

A.4 Trailer

There are 60K trailers from 33K unique movies in MovieNet. The statistics of the trailers are shown in Fig. A8, including the distributions of runtime and shot number.

Besides some attractive clips from the movie, which we name as content-shots, trailers usually contains exact shots to show some important information,

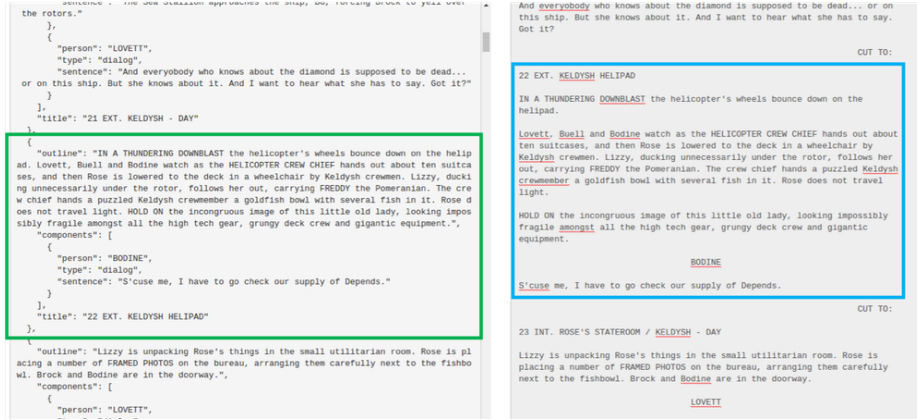


Fig. A9: Here we show an example of the parsed script. The left block shows the formatted script snippet and the right block shows the corresponding raw ones of “*Titanic*”.

e.g., the name of the director, release date, *etc.*. We name these shots as *info-shots*. Info-shots are quite different from other shots since they contain less visual content. For most of the tasks with trailers, we usually focus on content-shots only. Therefore, it is necessary for us to distinguish info-shots and content-shots. We develop a simple approach to tackling this problem.

Given a shot, we first use a scene text detector [69] to detect the text on each frame. Then we generate a binary map of each frame, where the areas covered by the text bounding boxes are set to 1 and others are set to 0. Then we average all the binary maps of a shot and get a heat map. By average the heat map we get an overall score s to indicate how much text detected in a shot. The shot whose score is higher than a threshold α and a the average contrast is lower than β is taken as a info-shot in MovieNet. Here we take the contrast into consideration by the observation that info-shots usually have simple backgrounds.

A.5 Script

We provide aligned scripts in MovieNet. Here we introduce the details of script alignment. As mentioned in the paper, we align the movie script to movies by automatically matching the dialog with subtitles. This process is introduced below.

Particularly, a movie script is a written work by filmmakers narrating the storyline and dialogs. It is useful for tasks like movie summarization. To obtain the data for these tasks, we need to align scripts to the movie timelines.

In the preprocessing stage, we develop a script parsing algorithm using regular expression matching to format a script as a list of scene cells, where scene cell denotes for the combination of a storyline snippet and a dialog snippet for a specific event. An example is shown in Fig. A9. To align each storyline snippet

Algorithm 1 Script Alignment

INPUT: $\mathbf{S} \in \mathbb{R}^{M \times N}$
 $R \leftarrow \text{Array}(N)$
 $val \leftarrow \text{Matrix}(M, N)$
 $inds \leftarrow \text{Matrix}(M, N)$
for $col \leftarrow 0, N - 1$ **do**
 for $row \leftarrow 0, M - 1$ **do**
 $a \leftarrow val[row, col - 1] + \mathbf{S}[row, col]$
 $b \leftarrow val[row - 1, col]$
 if $a > b$ **then**
 $inds[row, col] \leftarrow row$
 $val[row, col] \leftarrow a$
 else
 $inds[row, col] \leftarrow inds[row - 1, col]$
 $val[row, col] \leftarrow b$
 end if
 end for
end for
 $index \leftarrow M - 1$
for $col \leftarrow N - 1, 0$ **do**
 $index \leftarrow inds[index, col]$
 $R \leftarrow index$
end for
OUTPUT: R

to the movie timeline, we choose to connect dialog snippet to subtitle first. To be specific, we formulate script-timeline alignment problem as an optimization problem for dialog-subtitle alignment. The idea comes from the observation that dialog is designed as the outline of subtitle.

Let dig_i denote the dialog snippet in i^{th} scene cell, sub_j denote the j^{th} subtitle sentence. We use TF-IDF [50] to extract text feature for dialog snippet and subtitle sentence. Let $f_i = \text{TF-IDF}(dig_i)$ denote the TF-IDF feature vector of i^{th} dialog snippet and $g_j = \text{TF-IDF}(sub_j)$ denote that of j^{th} subtitle sentence. For all the M dialog snippets and N subtitle sentences, the similarity matrix \mathbf{S} is given by

$$s_{i,j} = \mathbf{S}(i, j) = \frac{f_i^T g_j}{|f_i| |g_j|}$$

For j^{th} subtitle sentence, we assume the index of matched dialog snippet i_j should be smaller than i_{j+1} , which is the index of matched dialog for $(j + 1)^{th}$ subtitle sentence. By taking this assumption into account, we formulate the dialog-subtitle alignment problem as the following optimization problem,

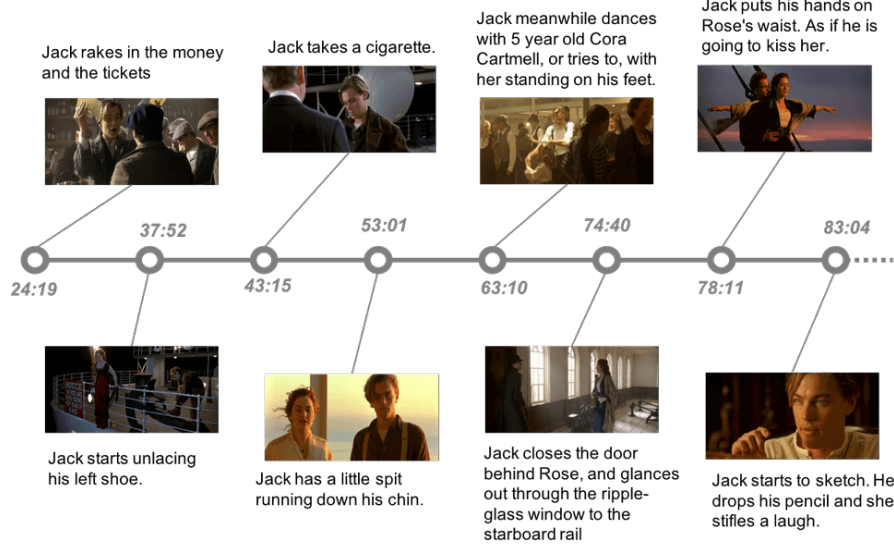


Fig. A10: Qualitative result of script alignment. The example comes from the movie *Titanic*. Each node marked by a timestamp is associated with a matched storyline snippet and a snapshot image.

$$\begin{aligned}
 & \max_{i_j} \sum_{j=0}^{N-1} s_{i_j, j} \\
 & \text{s.t.} \quad 0 \leq i_{j-1} \leq i_j \leq i_{j+1} \leq M-1.
 \end{aligned} \tag{1}$$

This can be effectively solved by dynamic programming algorithm. Let $L(p, q)$ denote the optimal value for the above optimization problem with \mathbf{S} replaced by its submatrix $\mathbf{S}[0, \dots, p, 0, \dots, q]$. The following equation holds,

$$L(i, j) = \max\{L(i, j-1), L(i-1, j) + s_{i, j}\}$$

It can be seen that the optimal value of the original problem is given by $L(M-1, N-1)$. To get the optimal solution, we apply the dynamic programming algorithm shown in Alg. 1. Once we obtain the connection between a dialog snippet and a subtitle sentence, we can directly assign the timestamp of the subtitle sentence to the script snippet who comes from the same scene cell as the dialog snippet. Fig. A10 shows the qualitative result of script alignment. It illustrates that our algorithm is able to draw the connection between storyline and timeline even without human assistance.

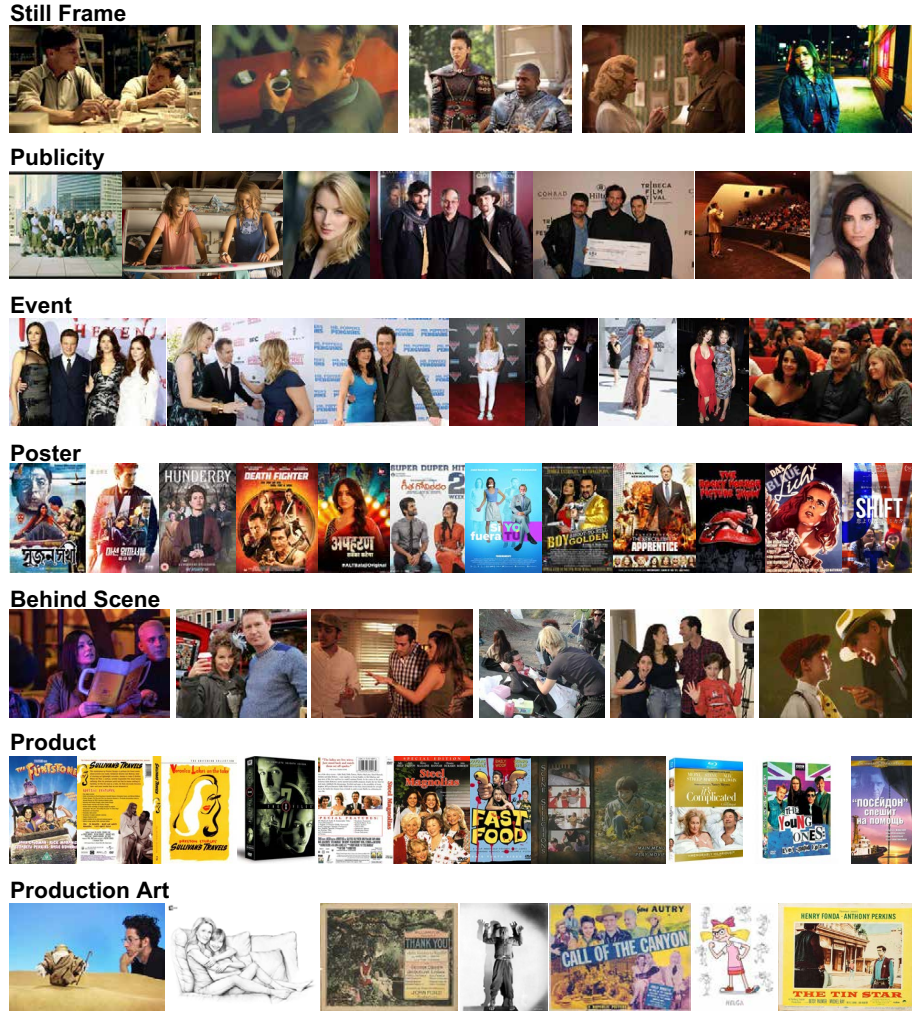


Fig. A13: Samples for different types of photos in MovieNet.

B Annotation in MovieNet

To achieve high-quality annotations, we have made great effort on designing the workflow and labeling interface, the details of which would be introduced below.

B.1 Character Bounding Box and Identity

Workflow and Interface. Annotation of the character bounding box and identity follows six steps. (1) We first randomly choose 758K key frames from the provided movies. Here the key frames are extracted by average sampling three

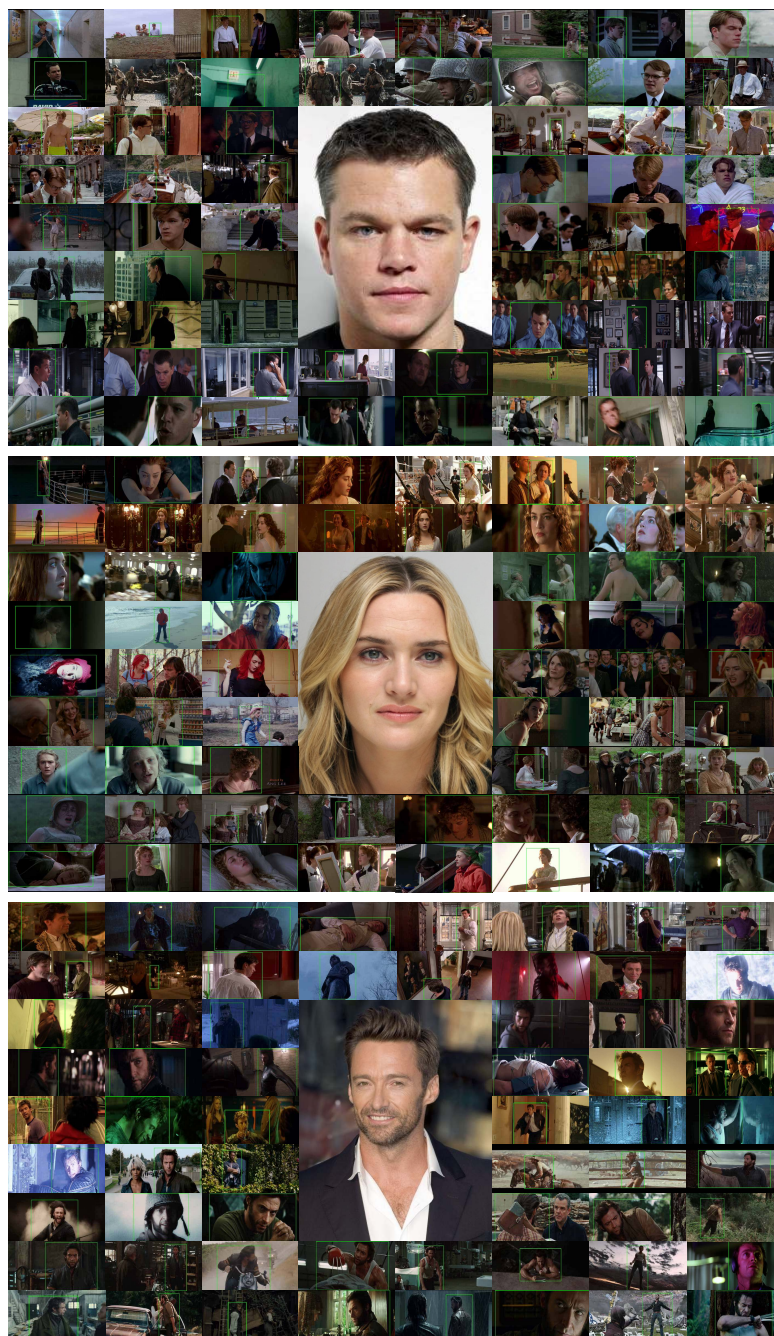


Fig. B1: Samples of the character annotations in MovieNet with portrait in the center.

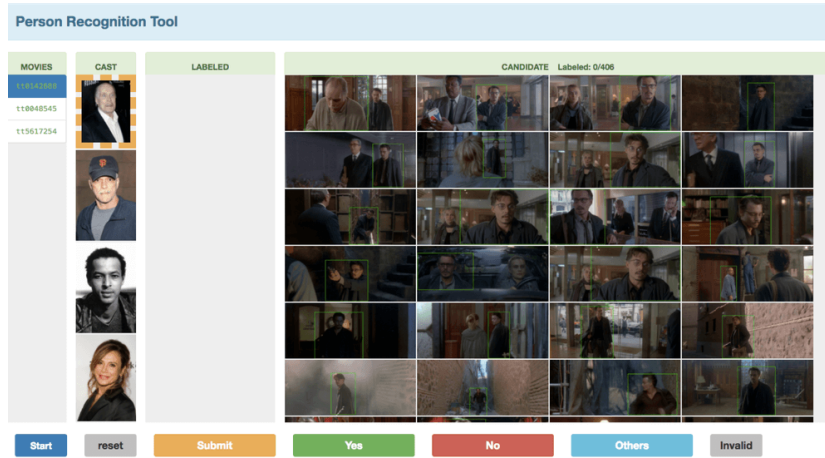
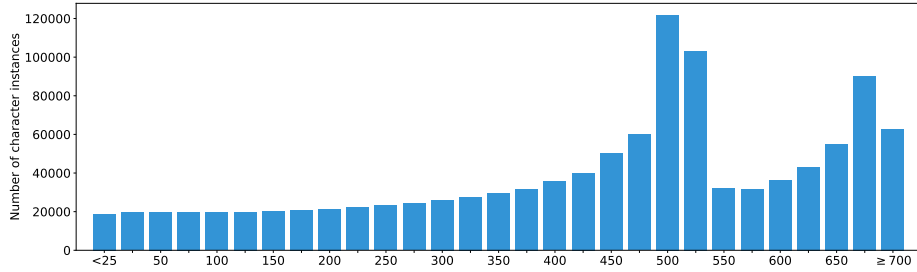
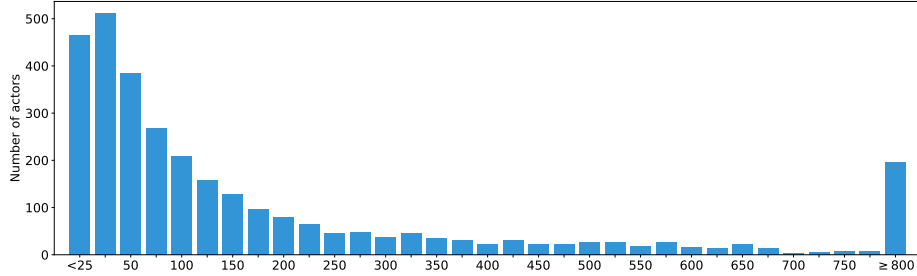


Fig. B2: Annotation interface of Character Identity (Stage 1). From left to right, they are (1) the movie list, (2) the cast list of the selected movie shown by their portraits, (3) the labeled samples annotated as the selected cast, which would be helpful for annotating more hard samples, and (4) the candidates of the selected cast, which is generated by our algorithm considering both face feature and body feature. The annotator can label positive samples (by clicking “Yes”) or negative samples (by clicking “No”). After several iterations, when they are familiar to all the cast in the movie, they can label the characters belong the credit cast list by clicking “Others”.

frames each shot. Then the annotators are asked to annotate the bounding box of the characters in the frames, after which we get 1.3M character bounding boxes. (2) With the 1.3M character bounding boxes, we train a character detector. Specifically, the detector is a Cascade R-CNN [8] with feature pyramid [41], using a ResNet-101 [32] as backbone. We find that the detector can achieve a 95% mAP. (3) Since the identities in different frames within one shot are usually duplicated, we choose only one frame from each shot to annotate the identities of the characters. We apply the detector to the key frames for identity annotation. Since the detector performs good enough, we only manually clean the false positive boxes in this step. Resulting in 1.1M instances. (4) To annotate the identities in a movie is a challenging task due to the large variance in visual appearances. We develop a semi-automatic system for the first step of identity annotation to reduce cost. We first get the portrait of each cast from IMDb or TMDb, some of which are shown in Fig. B1. (5) We then extract the face feature with a face model trained on MS1M [29] and extract the body feature with a model trained on PIPA [84]. By calculating the feature similarity of the portrait and the instances in the movie, we sort the candidate list for each cast. And the annotator is then asked to determine whether each candidate is the cast or not. Also, the candidate list would update after each annotation, which is similar to active learning. The interface is shown in Fig. B2. We find that this semi-automatic system can highly reduce the annotation cost. (6) Since the

**Fig. B3:** Height of character instance (pixel)**Fig. B4:** Number of character instance**Fig. B5:** Statistics of character bounding box and identity annotation, including the height of character instance and the number of character instance.

semi-automatic system may introduce some bias and noise, we further design a step for cleaning. At this step, the frames are demonstrated in time order and the annotating results at the first step are shown. The annotator can clean the results with temporal context.

Statistics and Samples. Here we show some statistics of the character annotation in Fig. B5, including the size distribution of bounding boxes and the distribution of instance number. From the statistics we can see that the number of character instance is a long-tail distribution. However, for those famous actors like *Leonardo DiCaprio*, they have much more character instances than others. Some samples are also shown in Fig. B1. We can see that MovieNet contains large-scale and diverse characters, which would be helpful for the researches on character analysis.

B.2 Cinematic Styles

We annotated the commonly used two kinds of cinematic tags of a shot [26]. Shot scale depict the portion of subject within the frames in a shot, while shot movement describe the camera movement or the lens change of a shot.

Shot Scale. Shot scale has 5 categories (as shown in Fig. B6): (1) *extreme close-up shot*: it only shows a very small part of a subject, *e.g.*, an eye or a mouth of a

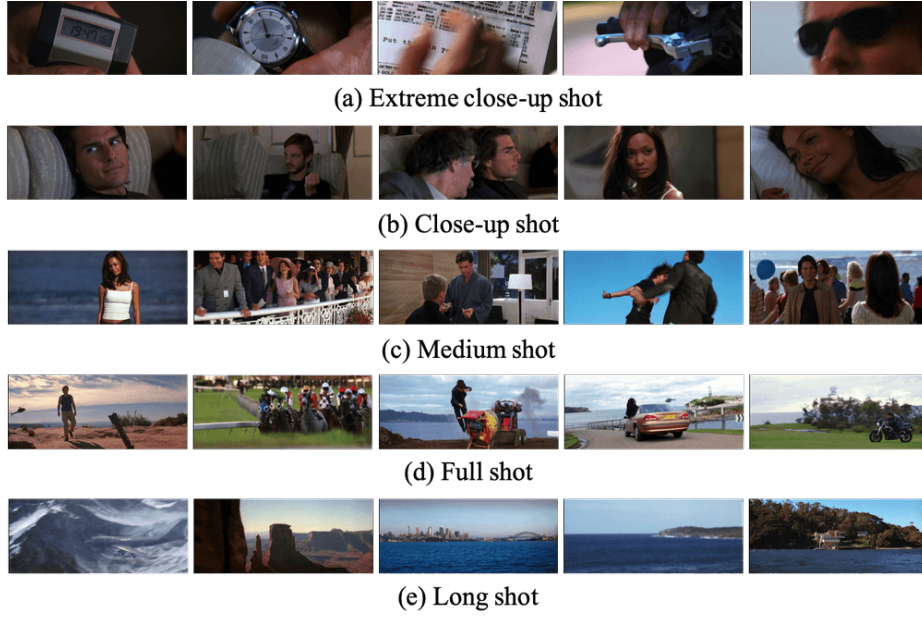


Fig. B6: Examples of the cinematic style (scale type) of shots in *Mission Impossible*. From (a) to (e), they are extreme close-up shot, close-up shot, medium shot, full shot and long shot.

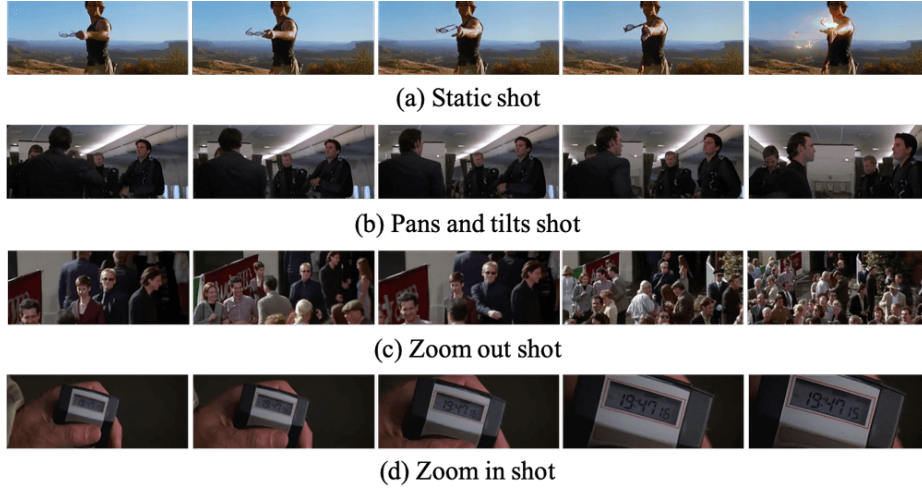


Fig. B7: Examples of the cinematic style (movement type) of shots in *Mission Impossible*. From (a) to (d), they are static shot, pans and tilts shot, zoom out shot, and zoom in shot.

person; (2) *close-up shot*: it concentrates on a relatively small part of a subject, *e.g.*, the face of the hand of a person; (3) *medium shot*: it contains part of a

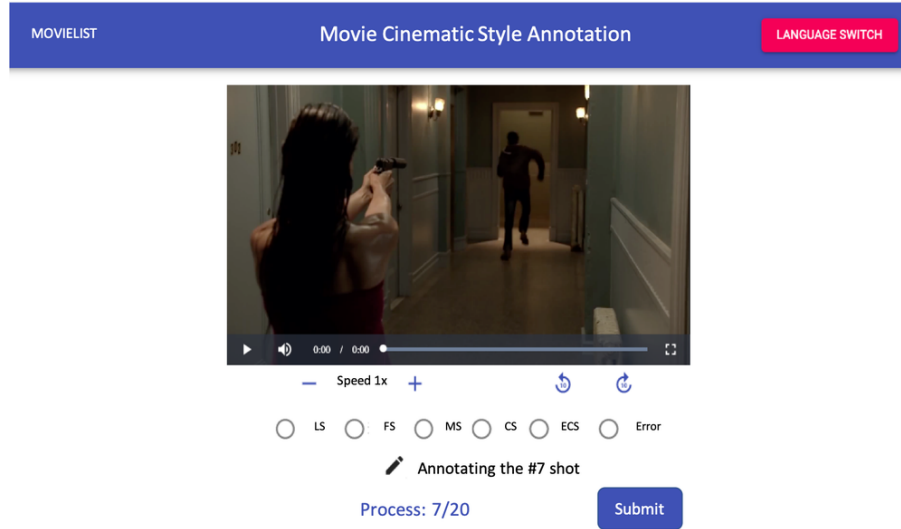


Fig. B8: Annotation interface of cinematic styles.

subject *e.g.*, a figure from the knees or waist up; (4) *full shot*: it includes the full subject; (5) *long shot*: it is taken from a long distance and the subject is very small within the frames.

Shot Movement. Shot movement has 4 categories (as shown in Fig. B7): (1) *static shot*: the camera is fixed but the subject is flexible to move; (2) *pans and tilts shot*: the camera moves or rotates; (3) *zoom out shot*: the camera zooms out for pull shot; (4) *zoom in shot*: the camera zooms in for push shot.

Annotation Categories. Compared with the definition in [26], we simplify the cinematic styles to make the annotation affordable. But we make sure the simplified categories are enough for most applications. For example, other cinematic styles like lighting are also important aspects. But the standard of lighting is hard to develop and we are now working on that with movie experts.

Annotation Workflow. To ensure the quality of the annotation, we make efforts as follows, (1) Instead of asking annotators to cut shot from movie or trailers and annotate their labels simultaneously, we cut shots from movies and trailers with the off-the-shelf method [62]. It is going to mitigate annotators' burdens considering the shot detection is well solved. (2) All the annotators went through a training phase first from cinematic professionals. And they are not allowed to annotate until they pass the professional test. They use our web-based annotation tool, as shown in Fig. B8. Each task is annotated three times to ensure a high annotation consistency.

Sample. We show five samples of each category of shot scale in one movie *Mission Impossible*. As we can see from Fig. B6. Extreme close-up shot reveals the detailed information about characters or objects. Close-up shot reveals information about characters' faces and their emotions. Medium shot shows character involved activities. Long shot shows the scenes setting of a character or an ob-

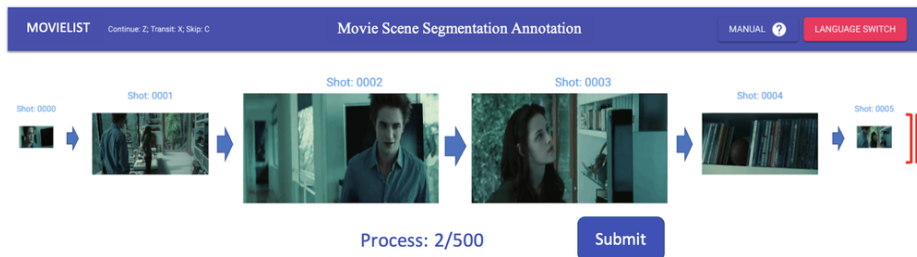


Fig. B9: Annotation interface of scene boundary.

ject. Full shot shows the landscapes that set up the movie. Additionally we show four examples of different movement type shots, as shown in Fig. B7.

B.3 Scene Boundary

Annotation Workflow. To increase the efficiency of annotating procedure, as well as to improve the label quality, we propose the following annotating strategy. It would be a prohibitive amount of work if the annotators are asked to go through the movies frame by frame. In this work, we adopt a shot-based approach, based on the observation that a shot is a part of exactly one scene. Hence, we consider a scene as a continuous subsequence of shots, and consequently the scene boundaries can be selected from shot boundaries. Also note that shot detection, as a task, is already well solved. Specifically, for each movie, we first divide it into shots using off-the-shelf methods [62]. This shot-based approach greatly simplifies and speeds up the annotation process.

Annotation Interface. We also developed a web-based annotation tool, as shown in Fig. B9 to facilitate human annotators to determine whether a scene transit or not between each pair of shots. On the web UI, annotators can watch two shots placed in the center, together with the frames preceding and succeeding these shots, respectively on the left and right sides. Annotators are required to make a decision as to whether the two shots belong to different scenes after watching both shots. The preceding and succeeding frames also provide a useful context.

Annotation Quality. A movie contains about $1K$ to $2K$ shots. It takes about one hour for an annotator to work through a whole movie, which can cause the difficulty to focus. To mitigate such issues, we cut each movie into 3 to 5 chunks (with overlaps), each containing about 500 to 700 shots. Then it only takes about 15 minutes to annotate one chunk. We found that it is much easier for annotators to focus when they work on chunks instead of the entire movie. All the annotators went through a training phase to learn how to use the annotation tool and how to handle various ambiguous cases before they work on the annotation tasks. We asked annotators to make careful decisions between each pair of shots. If an annotator is not sure about certain cases, they can choose unsure as an

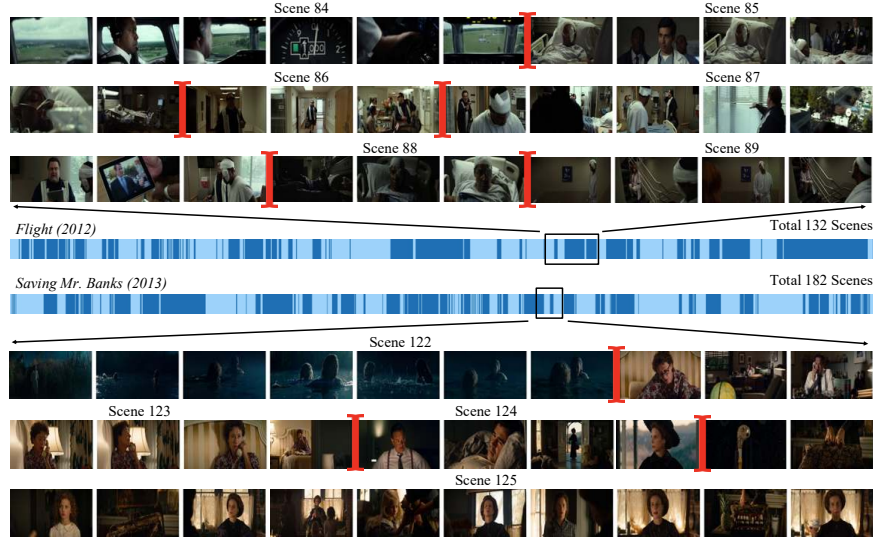


Fig. B10: Examples of the annotated scenes from two movies. The two lines in the middle correspond to the whole movie time line where the dark blue and light blue regions represent different annotated scenes, while the representative frames sampled from some scenes are also shown.

answer and skip. The collection constitutes two rounds. In the first round, we dispatch each chunk of movies to three independent annotators so that we can check consistency. In the second round, inconsistent annotations, *i.e.* the results collected from three annotators do not agree, will be re-assigned to two additional annotators. For such cases, we will get five results for each chunk.

Samples. We show two samples of scene boundaries in Fig. B10. Segmenting scenes is challenging since visual cues are not enough to recognize the scene boundaries. For example, in the first movie *Flight* scene 84 to scene 89 and the second movie *Saving Mr. Banks* scene 123 to scene 124, most of the frames look very similar to each other. Additional semantic information such as character, action, audio are needed to make the right prediction. The difficulty of segmenting vary among different scenes. Some easy cases are also listed, *e.g.* the scene boundary between scene 122 and scene 123 is easy to recognize since visual changes are obvious.

B.4 Action and Place Tags

In this section, we introduce the annotating procedure of action and place tags in MovieNet. We develop an interface to jointly label action and place tags. The detailed workflow and introduction of interface will be expanded as follows.

Annotation Workflow. We split each movie into segments according to scene boundaries. Each scene video lasts around 2 min. We manually annotated tags



(a) Eat.



(b) Kiss.



(c) Shoot Gun.

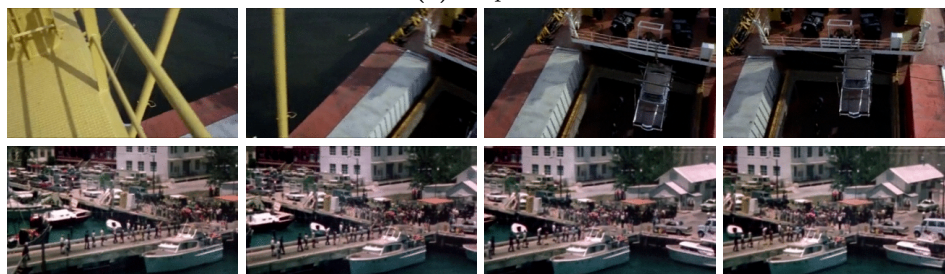


(d) Play Guitar.

Fig. B11: Example actions sample frames from category “eat”, “kiss”, “shoot gun” and “play guitar”.



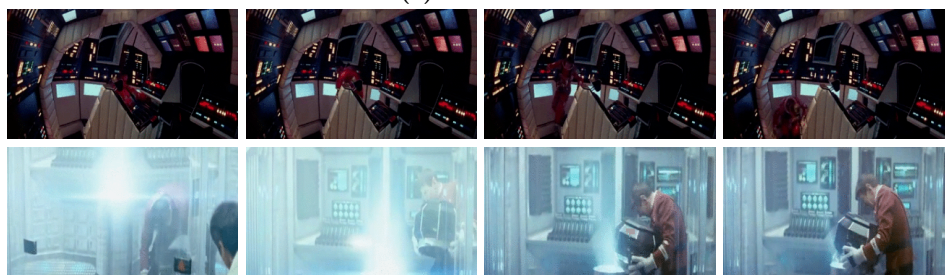
(a) Airport.



(b) Pier.



(c) Inside Car.



(d) Spaceship.

Fig. B12: Example place sample frames from category “airport”, “pier”, “inside car” and “spaceship”.

steps: (1) We collect the tags from previous works. The action tags are collected from datasets like AVA [28], Hollywood2 [46], *etc.* and the place tags are from datasets like Places [87], HVU [18], Placepedia [33], *etc.* (2) We leverage GoogleNLP tools to detect verbs that stand for actions and nouns that stand for location. Then we manually choose some useful action and place tags into the list. (3) We randomly annotate a few hours of videos to collect tags before we ask the annotators to do so.

Besides, in case that there are uncovered tags, we highly recommend the annotators to create tags by themselves. During annotation stage, we gradually collect and merge the created tags into our pre-defined list for improving the annotation efficiency.

Annotation Interface. We provide annotators with easy-to-use annotation interface shown in Fig. B15. At the beginning of annotating tags, annotators are able to get familiar with all the pre-defined tags by clicking the button “ACT. LABEL” and “PLACE LABEL” in the menu bar. Then they can carry out action and place annotation at “Action Annotation Zone” and “Place Annotation Zone” respectively. For the convenience of annotators and also for the consideration of annotation quality, we provide a function of replaying history action annotation that enable the annotators to replay what they just labeled and refine the original annotations. We also provide snapshot of shot keyframes to help the annotators quickly grasp the rough content of the current video. To help annotate temporal action boundaries, we provide workspace for annotators to set timestamps by moving forward or backward at a minimum stride of 0.1 seconds. By the observation that shot boundaries are often action boundaries, we provide shortcuts to set action boundary as shot boundary. The above strategies are beneficial for improving annotation quality and efficiency.

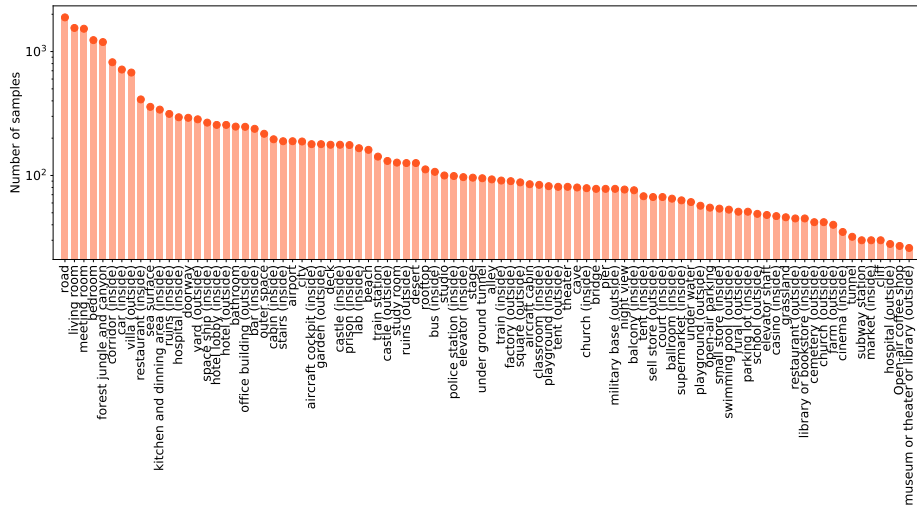


Fig. B14: Distribution of place annotations in MovieNet (y-axis in log scale).

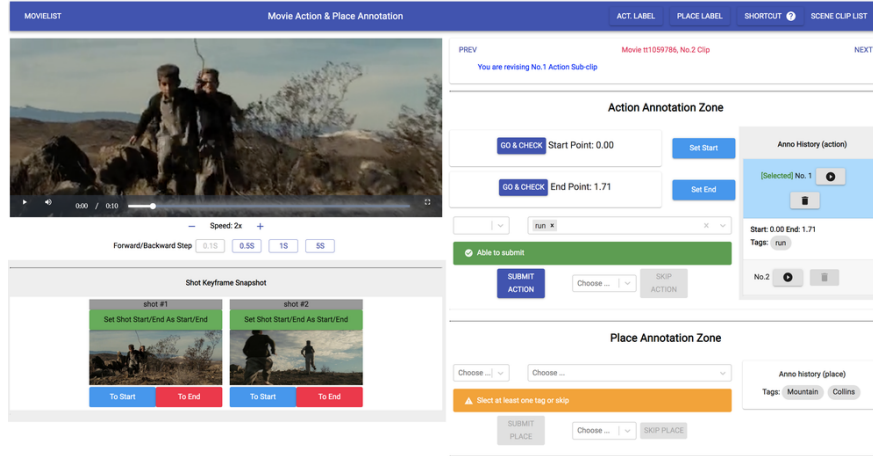


Fig. B15: Annotation interface of the action and place tagging in MovieNet.

Table B2: Comparison of action and place tags with related datasets. For AVA, the action tags are annotated every second, hence the number of tags are larger than expected (see *). So for AVA, we show the statistics after merging the person instance into tracklet for fair comparison, resulting in 116K character tracklet (each tracklet is taken as a clip) and 360K action tags.

Dataset	dura.(h)	action clip	action tag	place clip	place tag	source
Hollywood2 [46]	21.7	1.7K	1.7K	1.2K	1.2K	movie
MovieGraphs [71]	93.9	7.6K	23.4K	7.6K	7.6K	movie
AVA [28]	107.5	116K	360K(1.58M*)	-	-	movie
SOA [54]	-	308K	484K	173K	223K	web video
HVU [18]	-	481K	1.6M	367K	1.5M	web video
MovieNet	214.2	41.3K	45.0K	13.7K	19.6K	movie

Dataset Comparison. Here we compare our annotated tags with other datasets with action and place tagging. The comparison is shown in Tab. B2. Note that for fair comparison, we merge the tags of AVA because they are annotated every second. The de-duplication is done by merging the tags within the same character tracklet. After de-duplication, the number of tags is 360K in AVA. But most of them are common actions like stand. There are 116K tracklet with 426K bboxes in AVA. In a word, AVA is comparable to MovieNet in spatial temporal action recognition, but MovieNet can support much more research topics.

B.5 Synopsis Alignment

To support the movie segment retrieval task, we manually associate movie segments and synopsis paragraphs. In this section, we will present the following

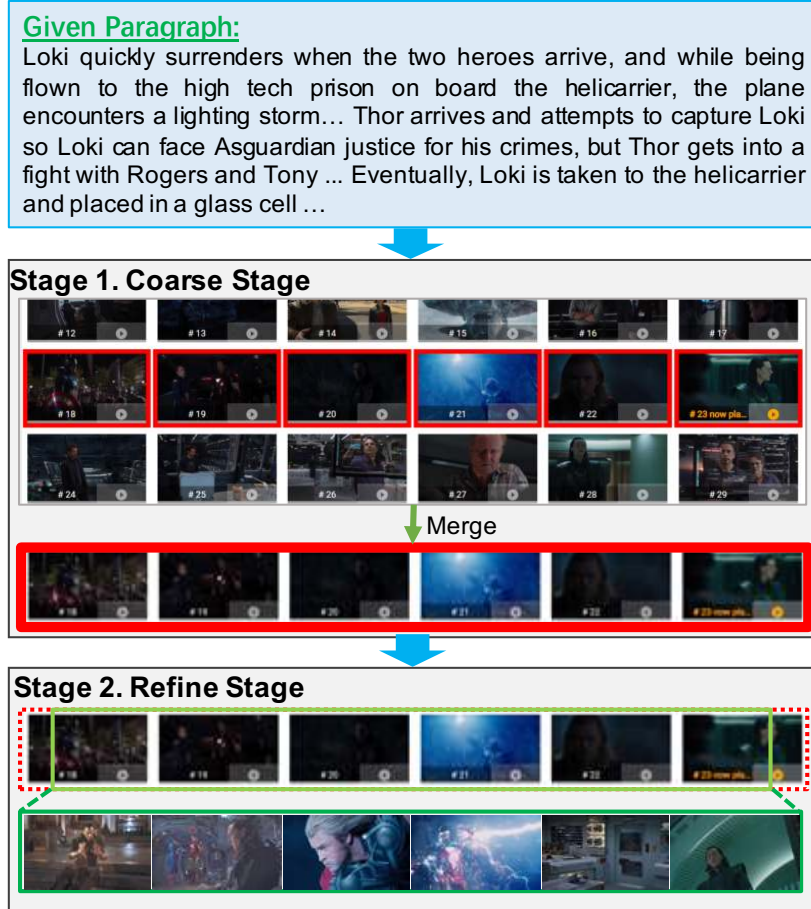


Fig. B16: Example of the annotating procedure for the movie *The Avengers*. At the coarse stage, annotator chooses consecutive clips. At the refine stage, boundaries are refined.

details about the movie-synopsis alignment, including annotation interface and workflow.

Annotation Workflow. After collecting synopses from IMDb, we further filtered out synopses with high quality, *i.e.* those contain more than 50 sentences, for annotating. Then we develop a coarse-to-fine procedure to effectively align each paragraph to its corresponding segment. (1) At the coarse stage, each movie is split into N segments, each lasting few minutes. Here we set $N = 64$. For each synopsis paragraph, we ask the annotators to select K consecutive clips that cover the content of the whole synopsis paragraph. (2) At the refine stage, the annotators are asked to refine the boundaries that make the resulting segment better aligned with the synopsis paragraph.

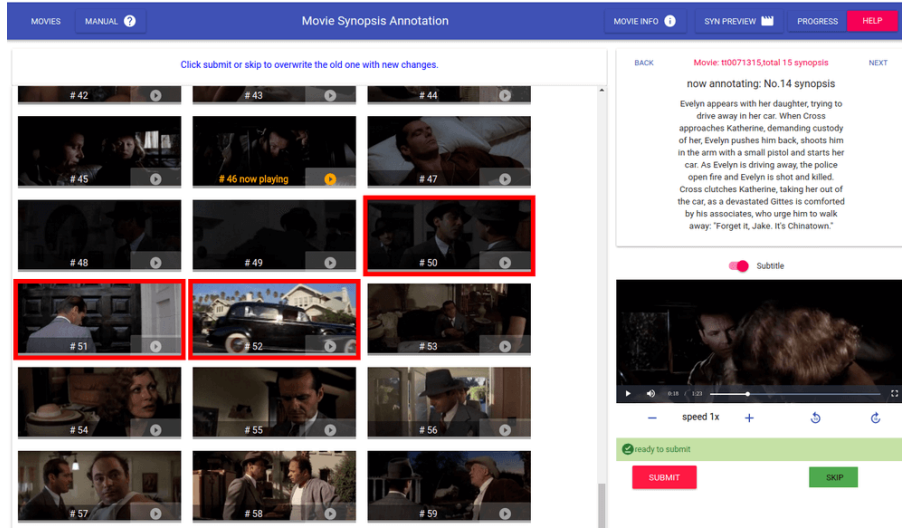


Fig. B17: User interface for the coarse stage. The left part display 64 movie clips; the upper right panel shows the synopsis paragraph; the bottom right player plays the selected movie clip.

To ensure the quality of the annotation, we make the following efforts. (1) The synopsis paragraphs from the same movie will be assigned to the same annotator. To ensure they are familiar with the movie, we provide the annotators with detailed overview of the movie, including character list, plot, reviews, *etc.* (2) Each synopsis paragraph is dispatched to 3 annotators. Then, we only keep those annotations with high consistency, *i.e.*, those with high temporal IoU among all the 3 annotations. Finally, 4208 paragraph-segment pairs are obtained.

Annotation Interface. The movie-synopsis alignment is collected in a coarse-to-fine manner shown in Fig. B16. We develop an online interface to carry out these two stages. The interface of the coarse stage is shown in Fig. B17. At the beginning of annotating each movie, annotators are required to browse the overview of this movie, which is available at “MOVIE INFO” and “SYN PREVIEW” in the menu bar. Then annotators select a subsequence of N consecutive clips that cover the corresponding synopsis description shown in the text panel. After receiving the annotations, the back-end server will merge the consecutive clips into a whole segment for refine stage. As shown in Fig. B18, at the refine stage, annotators adjust the temporal boundaries of the resultant segments. We allow annotators to set timestamps at the current playback location as start or end timestamps. To enable fine adjustment, users are able to control the video player by moving forward or backward at a minimum stride of 0.1 seconds.

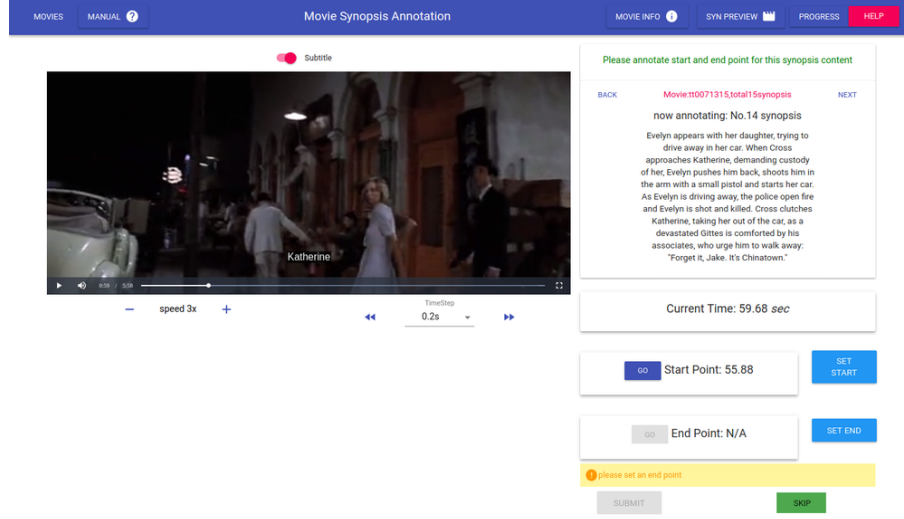


Fig. B18: User interface for the refine stage. The left part displays the merged movie segment; the upper right panel shows the synopsis paragraph; the bottom right buttons are for adjusting the boundaries.

B.6 Trailer Alignment.

To facilitate tasks like trailer generation, we provide an automatic process for matching a shot in a trailer to its movie counterpart. The process is introduced below.

Consider each shot in a trailer as a sequence of images $s^T = \{s_1^T, \dots, s_M^T\}$ where M is the number of frames in this shot. For a movie, we define it as a sequence of frame $s^M = \{s_1^M, \dots, s_N^M\}$ where N is the number of frame in the movie. To locate the position of the trailer shot in the movie is to find the most similar sub-sequence of the movie with the shot sequence. Let $Sim(s_i^T, s_j^M)$ denotes the similarity of i -th frame in trailer shot and j -th frame in the movie. The solution is to find the sub-sequence in the movie that maximize the similarities:

$$j^* = \underset{j}{\operatorname{argmax}} \sum_{i=1}^N Sim(s_i^T, s_{j+i-1}^M) \quad (2)$$

s.t. $j \leq N - M + 1.$

To obtain the similarity between two frames, we resort to features from low-level to high-level. For each frame, we extract GIST feature [20] and feature from *pool5* layer of ResNet-50 [32] that pre-trained on ImageNet [60]. We choose to use low-lwvel feature like GIST feature because we observe that most of the frames from trailers are alike with the original ones in movies, only with slightly changes in terms of color, size, lighting, boundary, *etc.*. For some harder cases

like cropped or carefully edited shots, we find high-level features like ResNet feature works. Hence, the similarity can be obtained by

$$Sim(s_a, s_b) = cosine(f_a^{gist}, f_b^{gist}) + cosine(f_a^{imagenet}, f_b^{imagenet}).$$

where f^{gist} stands for GIST feature while $f^{imagenet}$ stands for feature from ResNet.

Solving the optimization problem would result in the required alignment results. For those aligned shots with low optimized similarity score, we set it as misaligned shot. By observation, those misaligned shots are mostly transition shots or shots that only exist in the trailer. We then manually checked the alignment result to make sure the alignment results are accurate.

C Experiments

In this section, we introduce the detailed information for each benchmark. Note that unless specified, the 1100 movies are split according to a ratio of 3 : 1 : 1. The annotations are split according to the result of the split of movies, hence there are no overlapping movies among train, val and test sets for each task.

C.1 Genre Classification

Here we would introduce the benchmark setting, evaluation metrics, implementation details of the baseline models and show more experiment results.

Benchmark Setting. There are total 28 unique genres in MovieNet. Some rare genres, e.g. Adult, are ignored. And we further remove non-visual genres, e.g. News, to obtain a list with 21 genres. For image-based genre classification, we use the 3.9M photos of MovieNet as our data source. Since not all types of photos are related to the genres, e.g. publicity, here we only take 4 of them to build the benchmark for genre classification, namely poster, still frame, product and production art, resulting in 1.6M photos left. The 1.6M images are split into train, validation and test set that contains 1.1M, 160K and 321K images respectively. For video-based genres classification, we take trailers for experiments. We sample 32K trailers containing at least one of the 21 genres. They are split as train, validation and test set contains 22.5K, 3.2K and 6.4K videos respectively.

Evaluation Metric. Genre classification is a multi-label classification problem. We use mAP, recall@0.5 and precision@0.5 as evaluation metrics. Here 0.5 is the threshold to get the final prediction, which means that the final score (between 0 and 1) above 0.5 would be set as positive while the others would be set as negative.

Implementation Details. The models are trained with BCE Loss. The input size is set to 224×224 and we use SGD as optimizer. For the video-based model, we get 8 clips, each with 3 frames, on training. At the inference stage, we would predict the score of all the clips and average them to get the final prediction.

More Results. Here we show the per-genre results of the ResNet-50 model in Tab. C1. We can see that *animation* achieve the highest accuracy. This is

Table C1: Per-genre performance of genre classification.

	R@0.5	P@0.5	AP
Drama	79.42	71.16	79.95
Comedy	48.65	68.61	68.81
Thriller	14.50	64.98	49.80
Action	22.21	73.96	54.60
Romance	14.02	71.93	49.27
Horror	8.76	70.03	35.51
Crime	39.30	74.12	49.25
Documentary	4.79	85.49	21.03
Adventure	24.72	75.24	53.06
Sci-Fi	14.51	81.35	44.14
Family	27.11	82.55	52.19
Fantasy	13.51	69.83	39.12
Mystery	7.76	76.42	39.70
Biography	0.04	100.00	9.13
Animation	74.09	93.16	86.45
History	12.52	82.90	34.41
Music	27.24	89.04	47.13
War	12.80	86.27	34.41
Sport	21.99	94.97	39.59
Musical	4.45	73.58	22.88
Western	51.93	88.89	73.99

reasonable since the characteristic of animation is significant. And the AP of *Biography* and *Documentary* is much lower since these genres are determined by higher semantic elements.

C.2 Cinematic Style Analysis

Dataset Split. The MovieNet cinematic style prediction benchmark contains 46K shots coming from 8K trailers where 26K for training, 7K for validation and 13K for testing.

Details of baseline models for cinematic style prediction. We implement TSN [73] (3 segments) and I3D [9] with different backbones ResNet-18, ResNet-34, ResNet-50 [32]. The results are shown in the Tab. C2. We observe that 2D models achieve better results as the network becoming deeper both in terms of scale and movement classification. Deeper 3D models are a bit of over fitting on movement since the performance drops a little when the network becomes deeper.

Details of TSN+R³Net. As we point out in the paper, the subject is very important for cinematic style analysis. Thus a subject-based method is proposed to solve the problem. First, we adopt R³Net [17] to get the saliency map of each frame in the shot. Then, each video clip passes through a two-branch classification network, one branch is for video clips, the other branch is for video saliency

Table C2: Baselines for MovieNet cinematic style prediction.

Method	Backbone	Scale-Accuracy	Move-Accuracy
TSN [73]	ResNet18	79.31	68.02
	ResNet34	82.73	69.91
	ResNet50	84.08	70.46
I3D [9]	ResNet18	76.79	78.45
	ResNet34	82.10	82.17
	ResNet50	82.70	81.97
TSN+R ³ Net [17]	ResNet50	87.58	80.65

clips. For movement classification, the whole image and the background image are used as the inputs. For scale classification, the whole image and the subject image are used as the inputs of the two branches. The features from the two branches are concatenated and pass through a fully-connected layer to get the final prediction.

Implementation Details. We take cross-entropy loss for the classification. We train these models for 60 epochs with mini-batch SGD, where the batch size is set to 128 and the momentum is set to 0.9. The network weights are initialized with pretrained models from ImageNet [60]. The initial learning rate is 0.001 and the learning rate will be divided by 10 at the 20th and 40th epoch.

C.3 Character Detection

Dataset Split. We collect 1.3M bounding boxes from 758K keyframe images. The 758K images are split into train and test set with 692K and 66K respectively.

Evaluation Metric Following one of the most popular benchmark for object detection – COCO [43], we use the AP from 0.5 to 0.95 with a stride 0.05, namely mAP, as our evaluation metric.

Implementation Details. We take a Faster R-CNN with ResNet-50 as backbone, and train on three different datasets, COCO [43], CalTech [19] and our MovieNet-PDet to show the large domain gap between movie and other data source. And then we also try more models on MovieNet, including a single-stage model, namely RetinaNet, and a more powerful model, *i.e.* Cascade R-CNN [8] with ResNeXt-101 [76] backbone and feature pyramid [41].

C.4 Character Identification

Dataset Split. We annotate identities of more than 1.1M instances of 3K identities. In the MovieNet character identification benchmark. They are split into train, val, test set with 2088 identities with 639.9K instances, 821 identities with 336.6K instances and 876 identities with 364.2K instances respectively.

Benchmark Setting. The Character identification task is to search for all the instances of a character in a movie with just one portrait. To enable the task,

Table C3: Character identification: comparison of MovieNet and related datasets that used as the training datasets in this benchmark.

Dataset	ID	instance
Market [85]	1,501	32K
CUHK03 [40]	1,467	28K
CSM [34]	1,218	127K
MovieNet	3,087	1.1M

Table C4: Performance of different methods for character identification in MovieNet.

Train Data	cues	Method	mAP
Market [85]	body	r50-softmax	4.62
CUHK03 [40]	body	r50-softmax	5.33
CSM [34]	body	r50-softmax	26.21
MovieNet	body	r50-softmax	32.81
	body+face	LP [90]	8.29
	body+face	two-step [45]	63.95
	body+face	PPCC [34]	75.95

Table C5: Comparison of MovieNet scene segmentation with related datasets.

Dataset	# Scene	Duration(hour)	Source
OVSD [59]	300	10	MiniFilm
BBC [2]	670	9	Documentary
MovieNet	42K	633	Movie

we download a portrait from homepage of each credited cast, which will serve as the query portraits for the character identification tasks.

Evaluation Metric. We use mAP that average the AP on each query as the evaluation protocol.

Baseline Results. The results of character identification are shown in Tab. C4. The character identification task is similar to conventional person ReID task, however our dataset is much more challenging and larger than theirs. So in order to show the domain gap, we train ResNet-50 with softmax loss on three person/character identification dataset, namely, Market [85], CUHK03 [40] and MovieNet shown in Tab. C3. From the results, we see that due to the large domain gap, current ReID datasets cannot support the researches on character analysis in Movies. We also adopt methods – LP [90], PPCC [34] for comparison.

Implementation Details. The character identification task need to utilize both face feature and body feature. Here the face features are extracted by a ResNet-101 trained on MS1M [29] and the body features are extracted by a ResNet-50 trained by MovieNet identity annotation or other ReID dataset. The cues in Tab. C4 means that we retrieve the character by only the features mentioned by cues. The *Two-Step* method means that we would first retrieve by face features, and then add some instances with high confidence to the query set, after which we would do set-to-set retrieval by body features. This is widely used in the WIDER Challenge [45]. LP [90] means the naive label propagation method, which would be affected by noise and get a poor performance. And PPCC [34] improves LP by developing a competitive consensus scheme, which is the current state-of-the-art for character identification.

C.5 Scene Segmentation

Dataset Split. The overall 42K scenes are split into train, val, test sets with 25K, 8.9K and 8.1K scene segments respectively. There are no overlapped movies. The comparison of our dataset with other related scene segmentation datasets are shown in Tab. C5.

Evaluation Metrics. We take two commonly used metrics: (1) Average Precision (AP). Specifically in our experiment, it is the mean of AP of detected scene boundary for each movie. (2) M_{iou} : a weighted sum of intersection of union of a detected scene boundary with respect to its distance to the closest ground-truth scene boundary.

Baseline Models. We reproduce Grouping [59] and Siamese [2] according to their papers. For our baseline multi-semantic LSTM (MS-LSTM), we extract audio, character, action and scene feature from each shot as follows,

- **Audio feature.** We use tools from [81] that apply NaverNet [14] pre-trained on AVA-ActiveSpeaker dataset [58] to separate speech and background sound, and stft [70] to get their features respectively in a shot with 16K Hz sampling rate and 512 windowed signal length, and concatenate them to obtain audio features.
- **Character feature.** We firstly take the advantage of Faster-RCNN [56] pretrained on MovieNet character detection benchmark to detect character instances. And then we use a ResNet50 trained on MovieNet character identification benchmark to extract character features.
- **Action feature.** We utilize TSN [73] with AVA dataset [28] pretraining to get *action* features.
- **Place feature.** We take ResNet50 [32] with Places dataset [88] pretraining on key frame images of each shot to get place features.

More Results. From Tab. C6, we observe that (1) Benefited from large scale and high diversity, models trained on MovieNetSSeg achieve more than 40% improvement in performance. Specifically, Grouping [59] improves 98% from 0.170 to 0.336, Siamese [2] improves 34% from 0.268 to 0.358, MS-LSTM improves 39% from 0.334 to 0.465. (2) Multiple semantic elements are important for scene segmentation, which highly raise the performance. Jointly using audio, character, action and place information surpass any single element.

Implementation Details. We take cross entropy loss for the binary classification. We train these models for 30 epochs with SGD optimizer. The initial learning rate is 0.01 and the learning rate will be divided by 10 at the 15th epoch.

C.6 Action Recognition

We conduct experiments of action recognition on MovieNet action recognition benchmark. This task aims at predicting multiple action tags of a given video.

Dataset Split. The whole dataset is randomly split as train, val, test set with 23747, 7543, 9969 video clips respectively, without overlapping movies.

Table C6: Baseline results for scene segmentation.

Training Data	Method	AP (\uparrow)	M_{iou} (\uparrow)
OVSD [59]	Grouping [59]	0.170	0.301
	MS-LSTM	0.313	0.387
BBC [59]	Siamese [2]	0.268	0.358
	MS-LSTM	0.334	0.379
MovieNet-SSeg	Grouping [59]	0.336	0.372
	Siamese [2]	0.358	0.396
	MS-LSTM (Audio only)	0.210	0.341
	MS-LSTM (Character only)	0.213	0.348
	MS-LSTM (Action only)	0.227	0.368
	MS-LSTM (Place only)	0.442	0.421
	MS-LSTM	0.465	0.462

Table C7: Result of action classification. **Table C8:** Performance of different methods for place classification.

Method	mAP
TSN [73]	14.17
I3D [9]	20.69
SlowFast [23]	23.52

Method	mAP
I3D [9]	7.66
TSN [73]	8.33

Loss and Metrics. For training all models, we use binary cross-entropy loss as the loss function. For multi-label classification evaluation, we use mAP as the evaluation protocol.

Implementation of TSN. We adopt TSN [73] as one of our baseline models. To be specific, the TSN2D model adopt ResNet50 [32] as backbone. We sample 3 segments for each video and the consensus function is simply *Average*. We set batch size as 32 and dropout rate as 0.5. The model is trained using SGD for 100 epochs with an initial learning rate 0.01, momentum 0.9 and weight decay 0.0005. The learning rate is divided by 10 at the 60 and 90 epoch.

Implementation of I3D. For I3D [9], we adopt ResNet-I3D [9] with depth 50 as the baseline model. The inflate style is set to $3 \times 1 \times 1$ and the input length is 32 with stride 2. We set batch size as 8 and dropout rate as 0.5. The model is trained using SGD for 100 epochs with an initial learning rate 0.01, momentum 0.9 and weight decay 0.0001. The learning rate is divided by 10 at the 60 and 90 epoch.

Implementation of SlowFast. For SlowFast Network [23], the backbone is I3D with ResNet50. We set τ to 8, α to 8 and β to $1/8$. The input length is 32 with stride 2. We set batch size as 8 and dropout rate as 0.5. The model is also trained 100 epochs using a half-period cosine schedule [44] of learning rate decaying with n -th iteration learning rate as $0.5\eta[\cos(\frac{n}{n_{max}}\pi) + 1]$. n_{max} is the max iteration number, η is the basic learning rate set as 0.2.

Analysis. The experimental results are shown in Tab. C7. We see that the performance of TSN is the lowest while SlowFast is the best and outperform other baselines by a large margin. To further analysis the performance, we show per-class AP score of SlowFast Network in Fig. C1.

C.7 Place Recognition

We conduct experiments of place recognition on MovieNet place recognition benchmark. The task aims at predicting multiple place tags of a given video.

Dataset Split. The whole dataset is randomly split as train, val, test set with 8101, 2624, 2975 clips respectively, without overlapping movies.

Loss and Metrics. For training all models, we use binary cross-entropy loss as the loss function. For multi-label classification evaluation, we use mAP as the evaluation protocol.

Implementation of TSN. We adopt the same TSN structure as in MovieNet-Action, except that we use 12 segments here instead of 3. The training scheme is also the same as TSN in MovieNet-Action except that the batch size is changed to 8.

Implementation of I3D. We use the same I3D model and the same training scheme of I3D in MovieNet-Action.

Analysis. The experimental results are shown in Tab. C8. The backbone weights are adopted from ImageNet pretrained model. We see that the performance of TSN outperforms I3D probably because 3d convolution is harder to learn and the I3D model suffers over-fitting. We do not use SlowFast Network as one of the baseline models because SlowFast can not leverage the pretrain weight from ImageNet and our dataset is not large enough to support training SlowFast from scratch. To further analysis the performance, we show per-class ap score of TSN Network in Fig. C2.

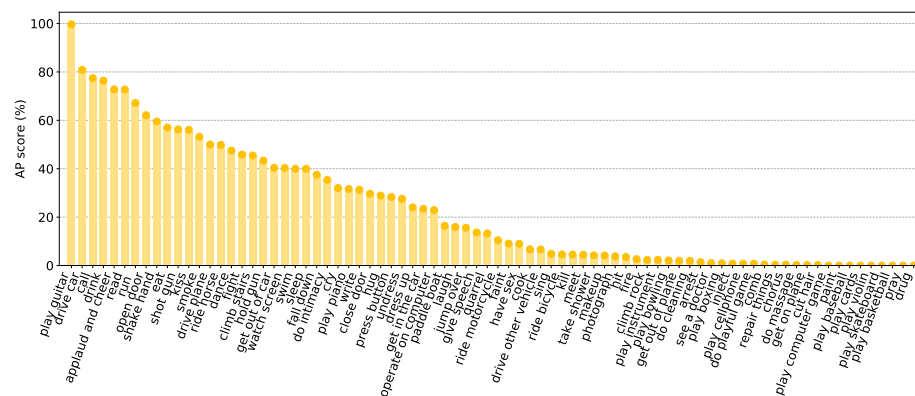


Fig. C1: Per-class AP score of SlowFast Network for action recognition task on MovieNet (sorted according to the performance of each class).

Baseline Models. We adopt VSE [25] as the base model in our experiments. In VSE model, we gradually add four kinds of nodes into baseline model, namely, *appearance*, *subtitle*, *action* and *character*. For each node, we extract a sequence of visual features from movie segment and a sequence of text features from synopsis paragraph. The detail of feature extraction will be introduced in the next section. In VSE, the input video features and paragraph features are first transformed with two-layer MLPs. For appearance, subtitle and action nodes, we first obtain the embedding of segment and paragraph by taking the average of the output sequence features. Then, the similarity score between segment and paragraph is computed by applying cosine similarity between two embeddings. For cast feature, we obtain the similarity score by applying *KuhnMunkres (KM)* algorithm [38] between the output sequence features from segment and paragraph. For training, we use the pairwise ranking loss with margin α shown below,

$$\begin{aligned} \mathcal{L}(S; \theta) = & \sum_i \sum_{j \neq i} \max(0, S(Q_j, P_i) - S(Q_i, P_i) + \alpha) \\ & + \sum_i \sum_{j \neq i} \max(0, S(Q_i, P_j) - S(Q_i, P_i) + \alpha) \end{aligned} \quad (3)$$

where $S(Q_j, P_i)$ denotes the similarity score between j^{th} segment Q_j and i^{th} paragraph P_i . θ denotes the model parameters.

Feature Extraction. The process of extracting input features for different modalities and different nodes are presented below.

- **Appearance feature from movie segment.** Appearance feature consists of a sequence of features extracted from each shot. For each shot, we extract the feature from *pool5* layer of ResNet-101 [32].
- **Appearance features from synopsis paragraph.** For paragraph, the appearance feature is represented as a sequence of Word2Vec [47] embeddings extracted from each sentence.
- **Subtitle feature from movie segment.** We also use Word2Vec embedding to extract features for subtitles in each shot. When adding to the model, we directly concatenate the subtitle feature of each shot to the appearance feature of each shot.
- **Action feature from movie segment.** The action features come from feature concatenated by TSN [73] pre-trained on AVA [28] and on MovieNet action recognition benchmark. We extract the action feature on each shot when there are actors appear in this shot.
- **Action feature from synopsis paragraph.** We detect verbs using part-of-speech tagging provided by GoogleNLP⁸. We select 1000 verbs with the highest frequency from the synopses corpus, and then retain those corresponding to visually observable actions, *e.g.* *run*. Action verbs are then represented by Word2Vec embeddings.

⁸ <https://cloud.google.com/natural-language/>

- **Character feature from movie segment.** We leverage the detector trained on MovieNet character detection benchmark to detect character instance in every shot. Then we use ResNet50 pretrained on PIPA [84] to extract the body feature and face feature as representation.
- **Character feature from synopsis paragraph.** We detect all the named entities (*e.g. Jack*) using StanfordNer [24]. With the help of IMDb, we can retrieve a portrait for each named character and thus obtain facial and body features using ResNet50 pre-trained on PIPA. This allows character nodes to be matched to the character instances detected in the movie.

Add Cinematic Style. As mentioned in the paper, cinematic style can help distinguish whether a node is important in a particular shot. Hence we design a module that take cinematic style and the node itself as input and produce an attention on this node. For each element in a shot, we concatenate its feature and the probability of the cinematic style as input, then this feature is passed through a MLP to produce a single attention score. This score is later used as the weight of output embedding.

Using Graph Formulation. We implement the algorithms in [78], both Event Flow Module and Character Interaction Module. The results are the combination of the two modules. That being said, we leverage the two modules to model spatial and temporal graph relations respectively. The difference between our implementation with theirs are the feature we used are different (see the Feature Extraction section).

Implementation Details. We train all the embedding networks using SGD with learning rate 0.001. The batch size is set to 16 and the margin α in pairwise ranking loss is set to 0.2.

D Toolbox

In this section, we introduce the toolbox designed for MovieNet, it will be released with the dataset and corresponding benchmark codes. The toolbox are mainly comprised of the following parts:

- **Crawlers.** The crawler for downloading metadata, subtitle and other useful data will be provided.
- **Preprocessing.** The preprocessing tools would provide functions that efficiently process multi-media resources. For example, extract audio waves, cut movies, resize movies.
- **Data generators.** The tools for generating the data, for example, shot detection will be included. Besides, we will also provide many tool for users to align their own movies with ours, if needed.
- **Data Parser.** The parsers are designed to easily access the required metadata or data. For example, to visualize the character bounding box or to read the genres of a movie.

References

1. Arandjelovic, O., Zisserman, A.: Automatic face recognition for film character retrieval in feature-length films. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2005) 4
2. Baraldi, L., Grana, C., Cucchiara, R.: A deep siamese network for scene detection in broadcast videos. In: 23rd ACM International Conference on Multimedia. pp. 1199–1202. ACM (2015) 11, 12, 46, 47, 48
3. Bauml, M., Tapaswi, M., Stiefelhagen, R.: Semi-supervised learning with constraints for person identification in multimedia data. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2013) 4, 11
4. Bhattacharya, S., Mehran, R., Sukthankar, R., Shah, M.: Classification of cinematographic shots using lie algebra and its application to complex event recognition. IEEE Transactions on Multimedia 16(3), 686–696 (April 2014) 9
5. Bojanowski, P., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Finding actors and actions in movies. In: Proceedings of the IEEE International Conference on Computer Vision (2013) 4
6. Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Weakly supervised action labeling in videos under ordering constraints. In: European Conference on Computer Vision. Springer (2014) 4
7. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 961–970 (2015) 1, 12
8. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6154–6162 (2018) 11, 29, 45
9. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017) 8, 9, 10, 12, 44, 45, 48
10. Cascante-Bonilla, P., Sitaraman, K., Luo, M., Ordonez, V.: Moviescope: Large-scale analysis of movies using multiple modalities. arXiv preprint arXiv:1908.03180 (2019) 6, 8, 9, 17
11. Chasanis, V.T., Likas, A.C., Galatsanos, N.P.: Scene detection in videos using shot clustering and sequence alignment. IEEE transactions on multimedia (2008) 4
12. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., Change Loy, C., Lin, D.: Hybrid task cascade for instance segmentation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 11
13. Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C.C., Lin, D.: MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07155 (2019) 11
14. Chung, J.S.: Naver at activitynet challenge 2019–task b active speaker detection (ava). arXiv preprint arXiv:1906.10555 (2019) 20, 47
15. Del Fabro, M., Böszörményi, L.: State-of-the-art and future challenges in video scene detection: a survey. Multimedia systems (2013) 4
16. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee (2009) 1

17. Deng, Z., Hu, X., Zhu, L., Xu, X., Qin, J., Han, G., Heng, P.A.: R3net: Recurrent residual refinement network for saliency detection. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 684–690. AAAI Press (2018) 9, 10, 44, 45
18. Diba, A., Fayyaz, M., Sharma, V., Paluri, M., Gall, J., Stiefelhausen, R., Van Gool, L.: Holistic large scale video understanding. arXiv preprint arXiv:1904.11451 (2019) 38, 39
19. Dollar, P., Wojek, C., Schiele, B., Perona, P.: Pedestrian detection: An evaluation of the state of the art. IEEE transactions on pattern analysis and machine intelligence 34(4), 743–761 (2011) 10, 11, 45
20. Douze, M., Jégou, H., Sandhawalia, H., Amsaleg, L., Schmid, C.: Evaluation of gist descriptors for web-scale image search. In: Proceedings of the ACM International Conference on Image and Video Retrieval. pp. 1–8 (2009) 42
21. Duchenne, O., Laptev, I., Sivic, J., Bach, F.R., Ponce, J.: Automatic annotation of human actions in video. In: Proceedings of the IEEE International Conference on Computer Vision (2009) 4
22. Everingham, M., Sivic, J., Zisserman, A.: Hello my name is... buffy – automatic naming of characters in tv video. In: BMVC (2006) 4
23. Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 6202–6211 (2019) 12, 48
24. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by gibbs sampling. In: Proceedings of the 43rd annual meeting on association for computational linguistics. pp. 363–370. Association for Computational Linguistics (2005) 52
25. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: Advances in neural information processing systems. pp. 2121–2129 (2013) 13, 51
26. Giannetti, L.D., Leach, J.: Understanding movies, vol. 1. Prentice Hall Upper Saddle River, New Jersey (1999) 6, 30, 32
27. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural networks 18(5-6), 602–610 (2005) 12
28. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., 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. pp. 6047–6056 (2018) 4, 5, 10, 11, 17, 37, 38, 39, 47, 51
29. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European Conference on Computer Vision. pp. 87–102. Springer (2016) 29, 46
30. Han, B., Wu, W.: Video scene segmentation using a novel boundary evaluation criterion and dynamic programming. In: IEEE International conference on multimedia and expo. IEEE (2011) 4
31. Haurilet, M.L., Tapaswi, M., Al-Halah, Z., Stiefelhausen, R.: Naming tv characters by watching and analyzing dialogs. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2016) 4, 11
32. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016) 8, 13, 29, 42, 44, 47, 48, 51

33. Huang, H., Zhang, Y., Huang, Q., Guo, Z., Liu, Z., Lin, D.: Placepedia: Comprehensive place understanding with multi-faceted annotations. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 38
34. Huang, Q., Liu, W., Lin, D.: Person search in videos with one portrait through visual and temporal links. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 10, 46
35. Huang, Q., Xiong, Y., Lin, D.: Unifying identification and context learning for person recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2018) 4
36. Huang, Q., Xiong, Y., Xiong, Y., Zhang, Y., Lin, D.: From trailers to storylines: An efficient way to learn from movies. arXiv preprint arXiv:1806.05341 (2018) 9
37. Huang, Q., Yang, L., Huang, H., Wu, T., Lin, D.: Caption-supervised face recognition: Training a state-of-the-art face model without manual annotation. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) 5
38. Kuhn, H.W.: The hungarian method for the assignment problem. *Naval research logistics quarterly* 2(1-2), 83–97 (1955) 51
39. Laptev, I., Marszałek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society (2008) 4
40. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 152–159 (2014) 10, 11, 46
41. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117–2125 (2017) 29, 45
42. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017) 11
43. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 10, 11, 45
44. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016) 48
45. Loy, C.C., Lin, D., Ouyang, W., Xiong, Y., Yang, S., Huang, Q., Zhou, D., Xia, W., Li, Q., Luo, P., et al.: Wider face and pedestrian challenge 2018: Methods and results. arXiv preprint arXiv:1902.06854 (2019) 10, 46
46. Marszałek, M., Laptev, I., Schmid, C.: Actions in context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society (2009) 4, 7, 11, 38, 39
47. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013) 13, 51
48. Nagrani, A., Zisserman, A.: From benedict cumberbatch to sherlock holmes: character identification in tv series without a script. *BMVC* (2017) 4
49. Park, S.B., Kim, H.N., Kim, H., Jo, G.S.: Exploiting script-subtitles alignment to scene boundary detection in movie. In: IEEE International Symposium on Multimedia. IEEE (2010) 4
50. Ramos, J., et al.: Using tf-idf to determine word relevance in document queries. In: Proceedings of the first instructional conference on machine learning. vol. 242, pp. 133–142. Piscataway, NJ (2003) 24

51. Rao, A., Wang, J., Xu, L., Jiang, Xuekun, H.Q., Zhou, B., Lin, D.: A unified framework for shot type classification based on subject centric lens. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020) [4](#)
52. Rao, A., Xu, L., Xiong, Y., Xu, G., Huang, Q., Zhou, B., Lin, D.: A local-to-global approach to multi-modal movie scene segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10146–10155 (2020) [12](#)
53. Rasheed, Z., Shah, M.: Detection and representation of scenes in videos. *IEEE transactions on Multimedia* (2005) [4](#)
54. Ray, J., Wang, H., Tran, D., Wang, Y., Feiszli, M., Torresani, L., Paluri, M.: Scenes-objects-actions: A multi-task, multi-label video dataset. In: The European Conference on Computer Vision (ECCV) (September 2018) [39](#)
55. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in neural information processing systems. pp. 91–99 (2015) [11](#)
56. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) *Advances in Neural Information Processing Systems* 28, pp. 91–99. Curran Associates, Inc. (2015) [47](#)
57. Rohrbach, A., Rohrbach, M., Tandon, N., Schiele, B.: A dataset for movie description. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3202–3212 (2015) [4](#), [5](#), [7](#), [17](#)
58. Roth, J., Chaudhuri, S., Klejch, O., Marvin, R., Gallagher, A., Kaver, L., Ramaswamy, S., Stopczynski, A., Schmid, C., Xi, Z., et al.: Ava-activespeaker: An audio-visual dataset for active speaker detection. *arXiv preprint arXiv:1901.01342* (2019) [47](#)
59. Rotman, D., Porat, D., Ashour, G.: Optimal sequential grouping for robust video scene detection using multiple modalities. *International Journal of Semantic Computing* **11**(02), 193–208 (2017) [11](#), [12](#), [46](#), [47](#), [48](#)
60. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. *International journal of computer vision* **115**(3), 211–252 (2015) [42](#), [45](#)
61. Shao, D., Xiong, Y., Zhao, Y., Huang, Q., Qiao, Y., Lin, D.: Find and focus: Retrieve and localize video events with natural language queries. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 200–216 (2018) [4](#)
62. Sidiropoulos, P., Mezaris, V., Kompatsiaris, I., Meinedo, H., Bugalho, M., Trancoso, I.: Temporal video segmentation to scenes using high-level audiovisual features. *IEEE Transactions on Circuits and Systems for Video Technology* **21**(8), 1163–1177 (2011) [7](#), [20](#), [32](#), [33](#)
63. Simões, G.S., Wehrmann, J., Barros, R.C., Ruiz, D.D.: Movie genre classification with convolutional neural networks. In: 2016 International Joint Conference on Neural Networks (IJCNN). IEEE (2016) [4](#), [8](#), [9](#)
64. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014) [8](#)
65. Tapaswi, M., Bäuml, M., Stiefelhagen, R.: knock! knock! who is it? probabilistic person identification in tv-series. In: IEEE Conference on Computer Vision and Pattern Recognition. IEEE (2012) [4](#), [11](#)
66. Tapaswi, M., Bäuml, M., Stiefelhagen, R.: Story-based video retrieval in tv series using plot synopses. In: Proceedings of International Conference on Multimedia Retrieval. p. 137. ACM (2014) [7](#)

67. Tapaswi, M., Bäumel, M., Stiefelwagen, R.: Aligning plot synopses to videos for story-based retrieval. *International Journal of Multimedia Information Retrieval* **4**(1), 3–16 (2015) [7](#)
68. Tapaswi, M., Zhu, Y., Stiefelwagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) [4](#), [5](#), [7](#), [11](#), [13](#), [17](#)
69. Tian, Z., Huang, W., He, T., He, P., Qiao, Y.: Detecting text in natural image with connectionist text proposal network. In: *European conference on computer vision*. pp. 56–72. Springer (2016) [23](#)
70. Umesh, S., Cohen, L., Nelson, D.: Fitting the mel scale. In: *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*. vol. 1, pp. 217–220. IEEE (1999) [47](#)
71. Vicol, P., Tapaswi, M., Castrejon, L., Fidler, S.: Moviegraphs: Towards understanding human-centric situations from videos. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018) [4](#), [5](#), [10](#), [11](#), [13](#), [17](#), [39](#)
72. Wang, H.L., Cheong, L.F.: Taxonomy of directing semantics for film shot classification. *IEEE Transactions on Circuits and Systems for Video Technology* **19**(10), 1529–1542 (2009) [9](#)
73. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Val Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: *ECCV* (2016) [44](#), [45](#), [47](#), [48](#), [51](#)
74. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: *European conference on computer vision*. pp. 20–36. Springer (2016) [8](#), [9](#), [10](#), [12](#)
75. Xia, J., Rao, A., Xu, L., Huang, Q., Wen, J., Lin, D.: Online multi-modal person search in videos. In: *Proceedings of the European Conference on Computer Vision (ECCV)* (2020) [10](#)
76. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1492–1500 (2017) [45](#)
77. Xiong, Y., Huang, Q., Guo, L., Zhou, H., Zhou, B., Lin, D.: A graph-based framework to bridge movies and synopses. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 4592–4601 (2019) [14](#)
78. Xiong, Y., Huang, Q., Guo, L., Zhou, H., Zhou, B., Lin, D.: A graph-based framework to bridge movies and synopses. In: *The IEEE International Conference on Computer Vision (ICCV)* (October 2019) [50](#), [52](#)
79. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 5288–5296 (2016) [12](#)
80. Xu, M., Wang, J., Hasan, M.A., He, X., Xu, C., Lu, H., Jin, J.S.: Using context saliency for movie shot classification. In: *2011 18th IEEE International Conference on Image Processing*. pp. 3653–3656. IEEE (2011) [9](#)
81. Xu, X., Dai, B., Lin, D.: Recursive visual sound separation using minus-plus net. In: *Proceedings of the IEEE International Conference on Computer Vision*. pp. 882–891 (2019) [47](#)
82. Yang, Y., Lin, S., Zhang, Y., Tang, S.: Statistical framework for shot segmentation and classification in sports video. In: *Computer Vision – ACCV 2007*. pp. 106–115. Springer Berlin Heidelberg (2007) [9](#)
83. Yue Zhao, Yuanjun Xiong, D.L.: Mmaction. <https://github.com/open-mmlab/mmaction> (2019) [12](#)

84. Zhang, N., Paluri, M., Taigman, Y., Fergus, R., Bourdev, L.: Beyond frontal faces: Improving person recognition using multiple cues. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4804–4813 (2015) [29](#), [52](#)
85. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., Tian, Q.: Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision. pp. 1116–1124 (2015) [10](#), [11](#), [46](#)
86. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 803–818 (2018) [8](#)
87. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017) [38](#)
88. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2018) [47](#)
89. Zhou, H., Hermans, T., Karandikar, A.V., Rehg, J.M.: Movie genre classification via scene categorization. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 747–750. ACM (2010) [4](#), [8](#)
90. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation (2002) [46](#)
91. Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., Fidler, S.: 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. pp. 19–27 (2015) [7](#)