USV: Towards Understanding the User-generated Short-form Videos

Haoyue Cheng[1†] Su Xu[2†] Liwei Jin[1†] Wayne Wu[2] Chen Qian[2] Limin Wang[1 ✉]
[1]State Key Laboratory for Novel Software Technology, Nanjing University, China
[2] SenseTime Research
{xusu,wuwenyan,qianchen}@sensetime.com
{chenghaoyue98,liwei.jin97}@gmail.com lmwang@nju.edu.cn

Abstract

Several large-scale video datasets have been published these years and have advanced the area of video understanding. However, the newly emerged user-generated short-form videos have rarely been studied. This paper presents USV, the User-generated Short-form Video dataset for high-level semantic video understanding. The dataset contains around 245K videos collected from UGC platforms by label queries without extra manual verification and trimming. Although video understanding has achieved plausible improvement these years, most works focus on instance-level recognition, which is not sufficient for learning the representation of the high-level semantic information of videos. Therefore, we further establish two tasks: topic recognition and video-text retrieval on USV. We propose two unified and effective baseline methods called Multi-Modality Fusion Network (MMF-Net) and Video-Text Contrastive Learning (VTCL) to tackle the topic recognition task and video-text retrieval respectively, and carry out comprehensive benchmarks to facilitate future researches.[1]

1. Introduction

Recently, user-generated short-form videos from platforms such as TikTok [8], Kwai [4], and Reels [6] have drawn much attention [9]. Understanding user-generated short-form videos is of great importance for practical usages such as video recommendation [17], venue analysis [11] and automated video summarization [30]. Take the video recommendation as an example. Zhu *et al*. [77] recommend videos by recalling videos that most fit the interested topic distribution of users. Deng *et al*. [18] take a real-time hot topic detection as the first step of recommendation. However, neither a dataset nor a benchmark exists in previous literature for understanding user-generated short-form videos from a multi-modality and high-level semantic perspective.

In these years, many video datasets have been proposed and pushed the boundary of video understanding. For example, some [33, 37, 43, 50, 54, 60, 63] are built for action recognition, others [25, 28, 31] are built for action localization. They mainly focus on recognizing instance-level actions and entities in long-term videos. None of the aforementioned works collect data purely from user-generated short-form videos, nor for leveraging multi-modality cues to facilitate understanding high-level semantic information of videos.

User-generated short-form videos have four main features that distinguished from other video forms. 1) *Topic Concentration*: User-generated short-form videos are more topic concentrated [52] compared to professional generated ones such as movies, TV series. Because short-form video platforms often have a duration constraint, short-form videos are forced to convey a single main topic in a few seconds. 2) *Text Richness*: User-generated short-form videos usually include a lot of text information, such as titles, subtitles, dialogue, and comments. These texts are all user-generated and have rich semantic information related to the videos. 3) *High Activity*: User-generated video platforms are highly active, with millions of videos uploaded every day, together with all kinds of new topics of videos. This difference makes manually filtering the noise for a clean dataset trivial compared with scaling up with the noise, which makes the traditional time-consuming data process with manual annotating impractical. 4) *Large Diversity*: User-generated short-form videos have more unique genres (*e.g*., slides, lectures, podcasts) and narratives (*e.g*., selfies, portraits) to demonstrate their topics. With the large diversity, the multi-modality cues are rather important for understanding.

In this study, we aim at moving towards understanding the user-generated short-form videos based on the aforementioned features. First, we introduce a new dataset named User-generated Short-form Video (USV-1.0). Specifically, USV-1.0 (the first version of USV) contains around 245K videos of 212 topic categories. Considering the first feature of *Topic Concentration*, we propose a new task named topic recognition. Regarding the second feature of *Text Richness*, we define the second task as video-text retrieval. Detailed definition and motivation of these two tasks can be found in Sec.4.1. For the third feature of *High*

[1]†: Equal contribution, ✉: Corresponding author.

Activity, we use no human labor for verification and trimming but directly assign the queried words as the topics and the user-generated titles as the text to retrieve. This webly supervised scheme is good at leveraging the large-scale user-generated video stream to facilitate video understanding in the real world. The fourth feature of *Large Diversity* identifies the main challenge of our tasks: how to integrate sufficient cues from various modalities to boost high-level semantic video understanding?

For topic recognition, we propose a simple yet effective baseline method Multi-Modality Fusion Network (MMF-Net) as a baseline. Specifically, MMF-Net is a three-branch network that fuses the predictions of models for three modalities to form a consensus on the topic. For video-text retrieval, we adopt a video-text contrastive learning (VTCL) framework, which has proven effective in self-supervised learning. For both tasks, we build a comprehensive benchmark to facilitate future researches. Benchmark experiments are conducted with our own implementation based on a common protocol and evaluated under a unified setting without bells and whistles.

To be brief, our contributions are three-fold:

1. New Data: we collect a new dataset: USV-1.0, which is the first large-scale dataset that aims at pushing the boundary of real-world short-form video understanding, to our knowledge.

2. New Tasks: we define a new task called topic recognition and we first try to perform video-text retrieval based on user-generated titles. Both tasks focus on understanding high-level semantic information.

3. New Methods: we propose MMF-Net and VTCL which are the first trials to utilize both audio and subtitles to tackle the topic recognition task and the video-text retrieval respectively, and build a comprehensive benchmark to facilitate future researches.

2. Related Work

Video Recognition Datasets. Several video datasets have been published these years. From initial trials [37, 63] to large-scale benchmarks [33, 50, 51]. Then datasets for specific fields emerged, such as fine-grained gym datasets [29, 60], human gestures [46], surveillance footage [54] and RGB-D camera [59]. Also, datasets of a tremendous scale [12, 20, 32] with the help of automatic annotation systems and web data have been derived. Most of these listed restrict their label space to be instance-level visual entities, and most videos are generated by professionals.

A similar idea of topical understanding is observed in YouTube8M [12]. However, they also restrict their topic entities to be visual and collect from YouTube with mostly long-term videos that are hard to infer a single topic or retrieve by a single title. This difficulty forces them to label a video with several topics and eventually degrades into visual-instance recognition.

A few works have been done towards understanding short-form videos [44, 52, 53, 69, 75]. However, most of their works are based on Vine, a video platform that has been shut down for years. Besides, the task designed on those datasets is scene/venue classification, which is still an instance-level recognition task.

Ours is not restricted to visual entities and is labeled video-wisely and purely from user-generated content.

Video-Text Retrieval. Learning videos with language has been a trending towards understanding videos and video-text retrieval is one of the fundamental task. Common datasets including [58, 71, 76] are relatively small, and [49] is too large to leverage, and restricted to tutorial videos. Many of them focus on specific domains, such as instructional videos, cooking videos, *etc*. Besides, most of them are well annotated and trimmed but not scalable.

As for methods, latent space-based models are common [34, 41, 49, 65]. Visual and textual representations are projected into a shared embedding space, where similarity can be measured directly. Typical visual encoding approach is to first extract frame-level features and then aggregate them into video-level representation [39, 65, 65]. A similar paradigm for textual encoding is to extract each word feature and aggregate them into sentence feature [19, 48, 55, 73, 74]. We aggregate frame-level visual feature and encode sentence feature directly to balance performance and computation cost.

Self-Supervised Video Representation Learning. Self-supervised representation learning constructs different kinds of supervision tasks from the data itself, to learn semantic representation to promote downstream tasks. For video, some of these tasks include temporal ordering of videos [23, 38], predicting motion and appearance [67], predicting the other parts of videos [26], *etc*.

Contrastive learning has been widely used in self-supervised representation learning, which aims at distinguishing the similar and dissimilar data pairs. For example, Radford *et al*. [56] learns visual and textual embedding based on image-text pair-wise contrastive learning, breaking through the limitation of classifying only on predefined categories in traditional image classification. Nowadays more and more works leverage multi-modality data based on contrastive learning [14, 35, 47, 49, 64]. Korbar *et al*. [35] uses visual and audio correspondence in semantic and temporal dimension to construct positive and negative samples. Some methods make use of abundant source of text, such as tags, labels, ASR, scripts, *etc*. Miech *et al*. [47] differs from previous approaches in employing multiple positive samples, denoted as MIL-NCE. Inspired by previous works, we conduct contrastive learning on video-text retrieval task.

3. USV: User-Generated Short-form Video Dataset

Our project aims to build a dataset for user-generated short-form video understanding, which is both intractable in

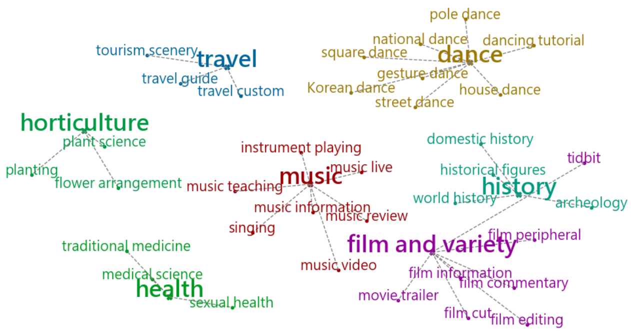

Figure 1. **The word embedding t-SNE of the taxonomy.** We select a part of the taxonomy for a better presentation. Different colors represent different macro-categories. Macro-categories are largely distant, while intra-category distance is short.

task and data itself. We will first demonstrate the procedure of building the dataset in Sec. 3.1 and illustrate the challenges within the dataset. Afterward, we will give statistics and comparison with other datasets in Sec. 3.2.

3.1. How the Dataset is Built

USV-1.0 is built by first pre-defining the taxonomy and using it as the query words to collect videos. We then extract the visual, audio, and textual modalities from the raw videos. We label the videos in an untrimmed and unverified manner that performs no extra manual annotating.

Stage I: Categories Taxonomy. To our best knowledge, no literature studies the semantic taxonomy for user-generated short-form videos. Most datasets [12, 28, 33] built their taxonomy regarding some former sociological researches and picked the visual-dependent ones. However, we do not adopt this as it will neglect the important feature of videos: *Large Diversity*. In addition, words from the knowledge graph or other references are out-dated. One important feature of user-generated short-form videos is that they are *Highly Active* with trending topics coming up daily such as *ASMR (Autonomous sensory meridian response)*, *blockchain*, *finger dance*, which can't be found in any knowledge graph of a sociological study on the taxonomy of videos. Therefore, we refer to the sector system of several online video platforms such as YouTube [10], BiliBili [1], and picked 32 macro topics including *anime*, *international affairs*, *sport*, *health*, *affection*, *etc*. as the root nodes of our taxonomy.

To step further, we investigate the top-watched categories of each macro topic in UGC short-form video platforms, and expand each of them into several micro topics as leaf nodes. As the result, we obtain the final 212 leaf nodes. Note that topics are not limited to visual-only ones, and they can be an abstract concept (*e.g.*, *affection*), audio- (*e.g.*, *ASMR*) or textual- (*e.g.*, *international news*) dependent, which requires understanding from multi-modality aspects. An overview of the topic taxonomy is demonstrated by t-SNE [45] in Fig 1.

Stage II: Collecting. We collect the videos by the 212 micro-categories. We use the words of the 212 micro topics and their synonyms as the queries to retrieve videos and their corresponding titles, and assign the queried topics as the labels of the videos. The rationale behind it is the feature *Topic Concentration* as the query word tends to be the one and only topic of the video. We are aware of the noise that emerges from this query-based collecting and labeling strategy. The candidate videos are recalled by the internal recommendation service of UGC platforms, therefore some videos may be hardly related to the queried topics. Therefore, our dataset is noisy and challenging. Then the unique video id is used to skip duplication within each micro topic. We collect 256700 videos in total.

Stage III: Modality Extraction. We extract three modalities: visual, audio, and text from all raw videos. We choose the raw soundtrack and the subtitles on the raw frames as the representation of audio and text modality. We extract the audio with FFMPEG [3]; as for text, we sample one frame each second and perform OCR detection with EasyOCR [2] to extract subtitles. Trivial words such as the water-print are filtered and then we approximate the subtitles of the sampled frames to their adjacent frames.

Stage IV: Human Verification. Human verification can be a challenge for video tasks. First, the total duration shown in Tab. 2 of our dataset is $>$100d for a human to go through, which will take more than years. Second, crowded source annotators have personal bias and recognition differences themselves, especially when asked to perform high-level inference rather than simply identify instances. Third, as stated in the introduction about the *High Activity* feature, the incremental ability of data possessing strategy is far more important than the correctness of supervision signals for highly active media like short-form UGC videos. We only verify the validation and test set for topic labels by two human annotators. The annotation user interface asks the annotators: By watching/listening/reading the video, whether the main topic of the video is the same as the query word. If both annotators choose *NO*, the video is then considered label noise and removed from the validation set. The original validation set of size 31260 is randomly split from the total dataset. After verification, 6788 videos are excluded. Since the validation set is with the same distribution as the training set, we can also assume that there are approximately 21.7% noisy label in the training set as well. Similar verification is also applied to the test set. After that, We remove the validation and test sets videos with empty title for video-text retrieval task.

3.2. Datasets Comparison

USV-1.0 is a large-scale dataset with various categories and contains rich modality data. More details are listed in Tab. 1 and Fig. 2. We compare ours with other intensively studied video recognition datasets [12, 20, 24, 28, 32, 33, 37, 50, 57, 63] and video-text retrieval datasets [13, 16, 36, 49, 58, 71, 76] in Tab. 2 and 3. We demonstrate the characteristics of our dataset: non-visual-only, topical, and user-generated.

Non-visual-only. In Tab. 2, **V** (Visual-only) represents whether the categorization can be done by visual modality

Table 1. **Statistics summary of USV-1.0 dataset.** We have at least 216 videos for each category and an average number of 1059.

Dataset Specifications	
Number of videos	245,672
Train set	200,000
Validation set	24,472
Test set	21,200
Number of micro-category labels	212
Average number of videos per micro-category	1,059
Number of macro-category labels	32
Average number of videos per macro-category	7,015
Range of training videos per micro-category	216-1,786
Number of videos with valid subtitles	151,598
Number of videos with valid titles	235,759
Average length of valid subtitles	106
Average length of valid titles	32
Number of videos with valid audio	245,650
Average duration of videos(in seconds)	55

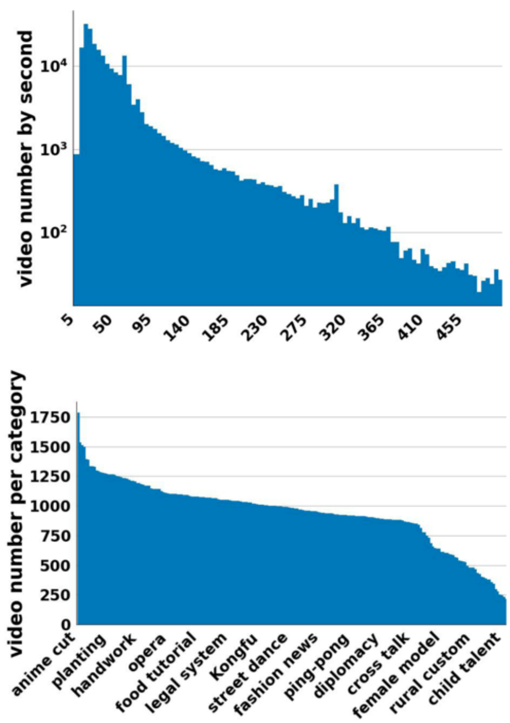

Figure 2. **Video number and duration distribution. Top:** distribution of the number of videos for each duration. **Bottom:** number of videos for each category.

only, and inversely, ¬**V** is non-visual-only. Note that despite some datasets such as Kinetics and YouTube8M preserve the audio soundtrack, they are also visual-only due to the videos or classes depending on other modalities for classification are removed by human annotators. UCF-101 is not visual-only, since it preserves audio-dependent classes and samples like *playing instruments*. Distinctly, categorization on the USV-1.0 dataset largely relies on multi-modality data since the data itself is *diverse*, thus our dataset is ¬**V**.

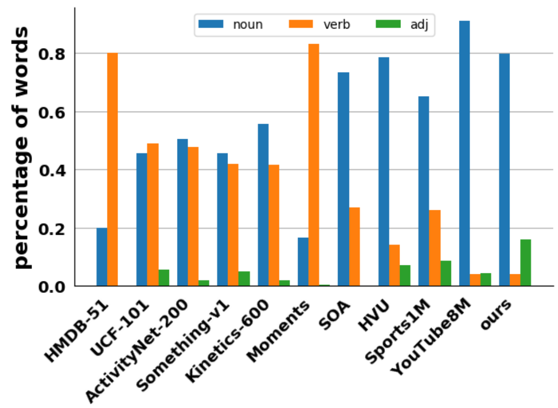

Figure 3. **Comparison of part-of-speech.** Our dataset taxonomy mainly consists of nouns and adjectives because of high-level semantics, so does YouTube8M. We use an NLP tool named textblob [7] to determine the part of speech for each word.

Topical. **T** (Topical) denotes whether the dataset taxonomy focusing on grabbing the topic or recognizing the existing instances in videos. As illustrated in Fig. 3, our label space has a different distribution compared with most other datasets. Former datasets focus on human actions and interaction between humans and objects, thus their label spaces are composed mostly of nouns and verbs. On the contrary, since our taxonomy is constructed by topics that summarize the main ideas of the videos and they usually consist of attributive adjuncts and the key objects of the video. Therefore, nouns and adjectives have much larger proportions. YouTube8M has a similar distribution to ours as it is also topical. Although nouns account for a large proportion of HVU and SOA[2], these labels are all instance-wise.

User-generated. **U** (User-generated) indicates whether all videos come purely from UGC video platforms. User-generated short-form videos have several unique characteristics that require to be studied respectively. USV-1.0 is constructed purely from the UGC platforms. In contrast, other datasets are collected from platforms like YouTube where the majority of videos are released by professional video producers. EPIC-KITCHENS is recorded by 32 individuals via headsets but not from a platform with a large number of users, so we consider it not user-generated.

4. User-Generated Short-Form Video Understanding

To benchmark user-generated short-form video understanding on USV-1.0 dataset, we propose two specific tasks called topic recognition and video-text retrieval. We will first demonstrate the detailed definition and motivation of these two tasks in Sec. 4.1. And then we describe our proposed *Multi-Modality Fusion Network (MMF-Net)* for topic recognition task in Sec. 4.2. Afterward, we provide a simple but effective *Video-Text Contrastive Learning (VTCL)*

[2]Since SOA can't be accessed publicly, we consider *scenes* and *objects* as nouns and *actions* as verbs.

Table 2. **Dataset comparison(topic recognition).** We compare the total number of videos and clips, the number of categories, total duration in time, whether the video categorization depends on other modality besides vision(¬**V**), whether the label taxonomy is Topical(**T**), and whether videos totally come from User-generated video platform(**U**).

Dataset	Videos	Clips	Categories	Duration	¬V	T	U
HMDB-51	3.3k	6.7k	51	5.7h	×	×	×
UCF-101	2.5k	13k	101	27h	✓	×	×
ActivityNet-200	20k	28k	200	27d	×	×	×
Something(v1)	108k	108k	174	121h	×	×	×
Kinetics-600	495k	495k	600	57d	×	×	×
Moments	1M	1M	339	31d	✓	×	×
SOA	562k	562k	65d	553	×	×	×
HVU	577k	577k	4378	66d	×	×	×
Sports1M	1M	1M	487	10y	×	×	×
YouTube8M	8M	8M	4800	57y	×	✓	×
USV-1.0 (**Ours**)	245k	245k	212	144d	✓	✓	✓

Table 3. **Dataset comparison(video-text retrieval).** We compare several common video-text retrieval datasets with ours in the number of videos/clips, average number of sentences per video/clip, total dataset duration, average duration per video, whether the videos are User-generated(**U**).

Dataset	Videos	Clips	Captions	Duration	U
MSR-VTT(v1)	7k	10k	200k	40h	×
LSMDC	200	128k	128k	150h	×
YouCook2	2k	14k	14k	176h	×
EPIC-KITCHENS	432	40k	40k	55h	×
DiDeMo	10k	27k	41k	87h	×
ANet-Captions	20k	100k	100k	849h	×
HowTo100M	1.2M	136M	136M	15.3y	×
USV-1.0 (**Ours**)	235k	235k	235k	138d	✓

framework for video-text retrieval task in Sec. 4.3. The methods we proposed are relatively simple but sufficient. The reason is that we intend to conduct preliminary explorations to provide insights on how multi-modality cues are beneficial for holistic user-generated short-form video understanding.

4.1. Task Definition and Motivation

Topic Recognition Task. Although both topic recognition and action recognition [28,33,37,50,63] can be categorized as a single-label multi-class classification problem, there are two crucial points to distinguish them. First, topic recognition uses topics as labels, which contain more high-level semantic information than most instance-level classification problems. Second, our proposed topic recognition encourages the use of multi-modality information inside videos for classification. To be specific, raw frames, audios, and subtitles are allowed to be used during training and evaluating stages. Modality-based tools like ASR or OCR are also not forbidden. Thus, topic recognition is not a purely instance-level visual task, but a multi-modal high-level semantic video classification task.

Video-Text Retrieval Task. Most user-generated short-form videos are paired with user-uploaded titles, which are usually strongly related to the corresponding videos. We view these collected titles as natural weak video captions. These "captions" are not annotated by professional annotators and easy to scale. Moreover, personal bias may be relieved with a large variety of "annotators". Formally, our defined title-based video-text retrieval consists of two subtasks: text-based video retrieval and video-based text retrieval. Suppose a test set of n pairs of videos and titles, text-based video retrieval aims to find the corresponding video for every given title in this set, and video-based text retrieval is vice versa. Similar to topic recognition, video-text retrieval encourages utilizing multi-modality information too. Our proposed video-text retrieval task forms a harder problem than topic recognition and its well annotated counterpart, since titles are usually of large diversity and sometimes related to videos at a high semantic level, e.g., a video of a beach traveling VLog with a title *"Happy holiday"*, in which case the title and visual frames are only related in high semantic level. To summarize, video-text (user-generated title) retrieval calls for high-level semantic video understanding rather than instance-level recognition.

Tasks Design Motivation. Our intention for building USV is to push the boundary of high-level semantic UGC short videos understanding. The reason why we choose topic recognition rather than existed tasks such as video object detection, action recognition or video classification is that, those tasks probe the ability of learning low-level video representation, while we figure it is more demanded by the industry to develop a sophisticated model to be able to reason along modalities and time to get a overall representation of videos.

Video-text retrieval steps further to some extent, as it abandons predefined labels the classical supervised learning scheme and uses natural language as supervisory signals. Video-text retrieval is not the only way to leverage natural language information to help video understanding, generating tasks like video captions or text-based video generation can also bind video with natural language. So why is retrieval? Here is an intuitive explanation: we notice that babies can learn concept by matching pictures and texts, but it is hard for them to write sentences or draw picture. So it is natural to assume video-text retrieval as a moderately difficult task for current video understanding.

4.2. Topic Recognition

We regard topic recognition as a fully-supervised learning problem and solve it with a general but effective network architecture named *Multi-Modality Fusion Network (MMF-Net)*. As is shown in Fig. 4, MMF-Net can be abstracted as a three-stream late fusion network to combine the results of three multi-modality streams. Late fusion has proven to be a simple but powerful technique used by many video recognition networks [22,62]. We detail the specific structures of these three branches in the following.

Branch I: Visual Encoder. Visual branch is implemented with classical 2D- and 3D-Conv networks such as TSN [68], I3D [15]. It consists of a feature extractor (backbone) built up with 2D- or 3D-Conv modules and a classifier (head) built with a linear layer. Those models have demonstrated

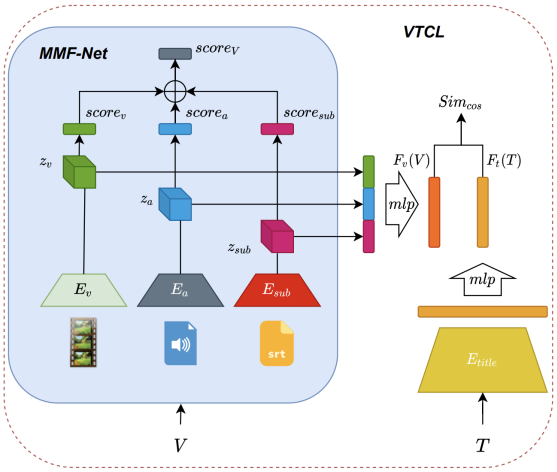

Figure 4. **The pipeline of our Multi-Modality Fusion Network(MMF-Net) and video-text contrastive learning (VTCL) framework for topic recognition and video-text retrieval.** First, the multi-modality signals are fed into modality-specific networks for feature extraction. For topic recognition, these features are used to predict 212-d classification scores separately and these scores are fused to form a video-level prediction. For video-text retrieval, multi-modality features are concatenated and projected by an MLP layer into a v-t joint-embedding, in which video features are matched with text features of user-generated titles via cosine similarity.

good representation learning ability on action recognition datasets [24, 33], which demands models to utilized both spatial and temporal information.

Branch II: Audio Encoder. Following the work of [14,70], we use a 2D image backbone to extract the audio features, namely ResNet-18 and log-Mel spectrograms of clips as input to generate the prediction logits. Log-Mel spectrograms have been used for audio recognition [72], multi-modality self-supervised learning [14, 61], *etc*., and have proven to be a discriminative and compact representation of audio. What's more, using spectrograms can largely benefit from the excellent designs of 2D image backbones like ResNet [27] from numerous image research areas.

Branch III: Text Encoder. Many new genres of videos have derived from short-form video platforms. For example, one popular type is those informative lecture videos produced by semi-professional self-media producers. We observe that most user-generated short-form videos are associated with subtitles, which can be viewed as an inner part of videos and are essential to holistic video understanding. Video subtitles are extracted by a powerful and robust open-source OCR toolkit named EasyOCR [2]. We use EasyOCR to extract frame-level subtitles and concatenate these subtitles in chronological order to form a video-level subtitle classification dataset. The textual branch consists of a pre-trained multilingual BERT model [19] appended with a linear classification head. We fine-tune this textual branch on the subtitle classification dataset.

4.3. Video-Text Retrieval

As topic labels are not available under the setting of video-text retrieval, we are only given n pairs of videos and titles, forming a scenario of cross-modality self-supervised learning.

Video-Text Contrastive Learning Framework. Following recent visual-language self-supervised learning methods [40, 47, 56], we propose an end-to-end contrastive learning framework called *Video-Text Contrastive Learning (VTCL)*. As is shown in Fig. 4, we aim at learning two mapping functions F_v and F_t that map videos and titles into a d-dimensional joint embedding space. In this d-dimensional joint embedding space, the embedding of a certain video is pulled closer to the embedding of its corresponding title, and the distance between embeddings of unrelated videos and titles should be extended. We measure the distance between video and title embeddings by cosine similarity:

$$s(V,T) = \frac{\langle \mathbf{F_v}(V), \mathbf{F_t}(T) \rangle}{\|\mathbf{F_v}(V))\|_2 \|\mathbf{F_t}(T)\|_2} \tag{1}$$

where V stands for sampled clips of a certain video and T stands for a certain title. Following [47, 56], we adopt an InfoNCE loss widely used in recent contrastive learning works:

$$\mathcal{L}_i = -\log \frac{e^{s(V_i,T_i)/\mathcal{T}}}{\sum_{j=1}^{n} (e^{s(V_i,T_j)/\mathcal{T}} + e^{s(V_j,T_i)/\mathcal{T}})} \tag{2}$$

where $\mathcal{T}$ is a temperature parameter and n is the size of mini-batch.

Video Encoder. The three branches of the video encoder share the same design as described in Sec. 4.2. There are two different characteristics worth mentioning: 1. The visual branch only uses 8-frame TSN [68] for its high performance in topic recognition and computing cost balance; 2. The parameters of the language model are frozen when training. Because the corpus for BERT pre-training is much larger and cleaner than ours, fine-tuning BERT directly on USV might cause language model crushing. The features output by three branches are concatenated and then projected by an MLP composed of 2 linear layers into d-dimensional joint embedding space.

Text Encoder. The text encoder also utilizes the above-mentioned multilingual BERT to extract 768-dimension features from user-generated titles. The BERT is also frozen during training for the same reason. These features are then projected by an MLP composed of 2 linear layers into d-dimensional joint embedding space, too.

5. Experiments

We design fundamental and important experiments for two tasks on USV-1.0, which are illustrated as follows. Besides, we have explored more different experiment settings and case studies in the supplementary materials due to the page limitation.

5.1. Experiment Setup

For supervised topic recognition, our experiment setup is unified for a fair comparison. Most of the training and testing follow the protocols in the original papers unless specified. We use 8 RGB frames of scale 224 for training, either sparsely or densely sampled. For evaluation, an identical amount of total frames as input is used, and mean class accuracy (mca) is adopted as the metric since our validation set is label unbalanced. We weight the class scores of vision, audio, and text branches according to empirical weight parameters of 1, 0.5, and 0.5 to obtain the ensemble classification scores, since vision plays a more important role in recognition than audio and text.

For self-supervised video-text retrieval, we use the same video input and training setting as topic recognition unless specified. The temperature parameter $\mathcal{T}$ is set as 0.05. To evaluate our learned VTCL embedding, we randomly select 20k valid video-title pairs from the validation set and divide them into 20 subsets evenly. Following the evaluating setting of previous retrieval datasets [58, 71, 76], we average the standard recall metrics R@1, R@5, R@10, and the median rank(Median R) on these 20 subsets.

5.2. Topic Recognition

Table 4. **Baseline Performance.** The column *Input frames* is formed by ($num\_crops \times clip\_length \times frame\_interval \times num\_clips$). $-f3$ denotes the frames used for training is 3.

Method	Type	pre-trained	Input frames	mca
TSN-f3	2D	Scratch	10x1x1x24	70.13
TSN-f5	2D	Scratch	10x1x1x24	71.15
TSN-f8	2D	Scratch	10x1x1x24	**73.51**
TSN-f8	2D	ImageNet	10x1x1x24	71.75
TSN-f8	2D	Kinetics-400	10x1x1x24	71.73
I3D	3D	Scratch	3x8x8x10	66.85
R(2+1)D	3D	Scratch	3x8x8x10	63.52
SlowFast	3D	Scratch	3x32x2x10	70.00

Baseline Models. For baseline experiments, we choose 5 mostly used models for video recognition, which are TSN [68], TSM [42], I3D [15], R(2+1)D [66], Slowfast [22]. All of them are using ResNet-50 as the base model. 3D-Conv models are trained with 8 densely sampled frames as input, while 2D models are trained with 8 sparsely sampled frames. The evaluation inputs are controlled to be the same (except for SlowFast which have double streams with different amounts of input) for a fair comparison. It is usually expected that 3D models may have better performance based on experience on several large-scale datasets [33]. However, according to Tab. 4, it is not the case on USV-1.0. The 3D models are surpassed by the 2D models with a great margin: The best-performed model is TSN trained with 8 frames, while the best 3D-model is SlowFast which requires much more input frames that still scores around 3.5% lower than TSN.

Training Frames. In addition, we also evaluate the result of TSN trained with different input frames, from 3 to 8. A steady increment can be found, which indicates that for a better understanding of the main topic of USV-1.0, more frames in training can be beneficial.

Pre-training. Furthermore, we pre-train the best performed TSN model on ImageNet and Kinetics-400 first and fine-tune it on USV-1.0 until convergence. It turns out the pre-trained weights even have a negative impact on USV-1.0. The performance of ImageNet pre-trained and Kinetics-400 pre-trained scores have a marginal difference and are lower than the from-scratch one by approximately 2%. This finding suggests that there may exist a considerable domain gap between ours and Kinetics-400 and ImageNet.

MMF-Net. In Tab. 5, we demonstrate the effectiveness of the designing of MMF-Net. For Branch I, the visual recognition branch, we have evaluated two typical video models of different mechanisms. We observe that multi-modality branches have a positive impact. To be more specific, audio and text branches score lower than 50%, however they bring no overall harm but benefit when fused with the visual branch. Note that the coefficients of all branches are set heuristically rather than by finding the optimized combination.

Table 5. **MMF-Net performance.** We evaluate the effect of each branch when fused to the visual branch as an ablation study.

Branch	Modality	mca(Δ)
I(TSN)	V	73.51
I(Slowfast)	V	70.00
II	A	40.88
I(TSN) + II	V + A	74.71(+1.20)
I(Slowfast) + II	V + A	71.81(+1.81)
III	T	46.61
I(TSN) + III	V + T	78.18(+4.67)
I(Slowfast) + III	V + T	76.07(+6.07)
I(TSN) + II + III	V + A + T	**78.84(+5.33)**
I(Slowfast) + II + III	V + A + T	**77.11(+7.11)**

5.3. Video-Text Retrieval

We adopt different multi-modality settings progressively by adding audio and subtitle stream to the video branch separately and together. All these variations use the same training setting for fair comparison.

Similar to topic recognition, we find multi-modality information also has positive effects on VTCL. As is shown in Tab. 6, additional audio and subtitle information fused with visual branch respectively both outperform visual-only baseline consistently. And combining the three modalities results in the highest performance among all the variations. It shows that integrating multi-modality information may help understand user-generated short-form videos from a holistic perspective.

Retrieval-Based Zero-Shot Topic Recognition. We evaluate our multi-modality embedding of VTCL with zero-shot classification on USV without any fine-tuning in Tab. 7. We transform class labels and videos into the same embedding space and recognize the video as the class with the

Table 6. **Video-text retrieval performance.** We evaluate the impact of using different combinations of three modalities in the video encoder of our proposed VTCL as an ablation study. In order to distinguish, we denote subtitles as T and titles as Title respectively.

Modality	Recall@1	Recall@5	Recall@10	Median R
V to Title	20.29	46.20	56.49	5.00
V + A to Title	22.19	48.58	59.08	4.00
V + T to Title	21.98	49.18	60.28	4.00
V + A + T to Title	**23.51**	**50.77**	**62.20**	**3.00**
Title to V	20.77	46.37	56.83	5.00
Title to V + A	22.33	48.90	59.85	4.00
Title to V + T	22.45	49.24	60.18	4.00
Title to V + A + T	**23.71**	**51.12**	**62.30**	**3.00**

Table 7. **Retrieval-based zero-shot topic recognition performance.** We evaluate the impact of using different combinations of three modalities in the video encoder as an ablation study.

Modality	mca(Δ)
V	27.21
V + A	27.23(+0.02)
V + T	27.23(+0.02)
V + A + T	**29.01(+1.80)**

highest cosine similarity. Although retrieval-based zero-shot topic recognition has lower performance than its fully-supervised counterpart, it is proposed as a generic self-supervised learning strategy using only user-generated titles to conduct topic recognition. This scheme is easy to scale up and generalize to the open world applications.

6. Conclusion

In this work, we build the first large-scale user-generated short-form video dataset, define two tasks called topic recognition and video-text retrieval, and propose MMF-Net and VTCL framework as simple but effective baselines for these two tasks. We conduct comprehensive experiments as preliminary explorations to facilitate future researches on user-generated short-form video understanding.

Appendix

A. Implementation Details

A.1. MMF-Net Formalization

To formalize MMF-Net, we define our dataset as $\mathcal{V} = \{v_i\}$, where v_i denotes the i-th video in the dataset. Similarly, we denote $x_i^{j(k)}$ as the j-th clip-wise input sampled with a fixed duration from a total of N clips of the i-th video's k-th modality. For the subtitle branch we use video-level text input namely $N = 1$. For example, $N = 5$, $i = 10$, $j = 3$, $k = 2$, $x_i^{j(k)}$ represents the 3-rd audio clip out of 5 uniformly divided clips of the 10-th video. With all annotations above, the inference procedure of *MMF-Net* can be formalized as:

$$\hat{s}_i^{(k)} = \frac{1}{N} * \sum_{j=1}^{N} \sigma(E^{(k)}(x_i^{j(k)})) \tag{3}$$

$$\hat{S}_i = \frac{1}{3} * \sum_{k=1}^{3} w_k \hat{s}_i^{(k)} \tag{4}$$

where $\hat{s}_i^{(k)}$ denotes the recognition score of the k-th branch on the i-th video, $E^{(k)}$ denotes the feature extractor of branch k, and w_k denotes the weight of the score of branch k. To be specific, we use 1, 0.5, 0.5 as the weights from Branch I to III. Eq. (3) can be viewed as a clip-wise consensus [68] with the average operation, and Eq. (4) is a branch-wise late-fusion for video topic recognition by averaging as well.

While training, we set $N = 1$ for each video and we train each branch separately, therefore the forward equation and loss function for each branch during training is:

$$\hat{s}_i^{(k)} = \sigma(E^{(k)}(x_i^{(k)})) \tag{5}$$

$$\mathcal{L}_{\text{rec}} = -\sum_{c=0}^{211} y_c * \ln \hat{s}_c \tag{6}$$

y_c denotes the ground truth for the c-th logits from 0 or 1, the loss is a plain Cross-Entropy loss with $Softmax$ activation.

A.2. Branch Instantiation Details

Branch I For 2D backbones such as TSN, we divide each video into 8 parts and randomly select one frame from each part for training. Frames are resized to short-side 256p first, and perform a multi-scale crop into a square, and resized to 224×224; While testing, we sample 24 frames uniformly, each perform a short-size resizing to 256 and a ten-crop to a square 256p image. TSN is instantiated by the backbone of ResNet-50 and a linear layer as the classification head mapping the global average pooled 2048-d vector to 212 class scores; TSM follows the original optimal setting is the original paper [42], with $1/8$ channels shifted on each $conv1$ layer of all child blocks in each ResNet stage.

For 3D backbones such as Slowfast, we randomly sample a clip of 8 frames of interval 8 and perform the identical augmentation as 2D backbones; For testing, 10 uniform clips are sampled and each performs a three-crop to 256p frames. Clip-wise predictions are then averaged to form a video-wise prediction. I3D is based on a ResNet-50 backbone and inflate the 2D-Convs into 3D ones. Note that we do not bootstrap 3D filters from 2D pretrains but train it from scratch; For R(2+1)D, it follows the instantiation of I3D, but only factorizes all 3D-Convs into (2+1)D ones; As for SlowFast, both branches are a ResNet3D backbone such as I3D. The input for the slow pathway is 8×8, while for the fast path it is 32×2, and the base channels of the fast are $1/8$ of the slow one. The 2048-d and 512-d feature vectors of the two pathways are concatenated and mapped by a linear layer to the 212-d score.

Branch II For each video, we randomly sample a corresponding clip of 2s, *i.e.* 2×16000 bins. Then the log-Mel transformation with 80 mel-filters, window size 32ms, hop size 16ms and fft size of 1280 will turn the 2-second clip into a spectrogram of 80 channels and 128 time stamp, *i.e.* a gray image of size $1 \times 80 \times 128$; When testing, we uniformly sample 10 clips per video and average all clips' logits to get the final prediction.

Branch III We use the pre-trained multilingual BERT model provided by Google-Research [19] with 104 languages, 12-layer, 768-hidden, 12-heads, and 110M parameters as the text-branch backbone. For each video, we use EasyOCR [2] to extract subtitles of one frame per second and filter out watermarks. Then we concatenate frame-level OCR results in chronological order to form video-level subtitles as the raw input. There are 12.3k videos with valid subtitles out of the total 20k training videos. We just leave them empty strings as BERT input. For topic recognition, we fine-tune the the pre-trained multilingual BERT model with initial learning rate is 5e-5 for 8 samples per batch; a linear scheduler with 500 warm-up steps is applied until saturation. For video-text retrieval, we freeze the pre-trained multilingual BERT model as an off-shelf feature extractor when training and evaluating.

A.3. Experiment Setup

Training Settings. For topic recognition, 2D-Conv models are trained for 100 epochs until saturation, the initial learning rate of 0.05 and a step schedule that decay the learning weight by 0.1 at step 40 and 80, the batch size is 128; 3d-Conv models are trained for 200 epochs, the initial learning rate is 0.1 for 64 samples per batch. The scheduler used is cosine annealing until saturation. This setting follows several public codebases including [5,21], in which 2D models use less epoch as the sparse sampling can lead to faster convergence, and smaller and plainer learning since 2D models have fewer parameters.

For video-text retrieval, TSN is applied for visual branch I in the video encoder. Frozen multilingual BERT is used as the feature extractors both in subtitle branch III and the title encoder. VTCL is trained for 100 epochs, with the initial learning rate of 0.03, a cosine annealing scheduler, and a batch size of 128. VTCL is under-optimized for computing cost limitation.

Evaluation Settings and Metrics. As examined by multiple pieces of research [15, 21, 68], the number of input frames has a great impact on accuracy. Therefore, we balance the number of crops and clips to control an identical number of frames as input for all experiments.

Since our validation set is of the same distribution as the training set, namely the numbers of videos of each topic category are not identical, it is unfair to use the top k accuracy. Therefore, we used the mean class accuracy for evaluation on USV, and use top-1 accuracy on downstream datasets.

To evaluate VTCL on video-text retrieval, we randomly select 20k valid video-title pairs from the validation set and divide them into 20 subsets evenly. Following the evaluating setting of previous retrieval datasets [58, 71, 76], we average the standard recall metrics R@1, R@5, R@10, and the median rank(Median R) on these 20 subsets.

B. Supplementary Analysis of Experiments

B.1. Confusion Analysis

In Tab. 8, 9 and 10, we demonstrate the confusions of the worst ten classes for three models: TSN , SlowFast, and MMF-Net (with TSN as Branch I), which are the representative methods for 2D-Conv, 3D-Conv, and ours. The confusion shows that fine-grained categories are easier to be confused, such as *planting* and *farm work*, *pet cat*, and *pet dog*. This observation raises the challenge of accurate spatial discrimination; Classes such as *restaurant reviews* and *food reviews* may have a very similar visual, audio, and textual appearance, and it further emphasizes the importance of the reasoning ability of the model.

Class 1	Class 2	Confusion
male model	layman handsome influencer	65%
restaurant review	food review	49%
planting	farm work	44%
movie information	movie review	31%
global military intelligence	domestic military intelligence	29%
science anecdote	cutting edge of science & technology	26%
rural performance	folklore	26%
luxury car	roadster	24%
roadster	luxury car	21%
military exercise	global military intelligence	21%

Table 8. **Top-10 class confusions in USV-1.0, using the TSN model.**

Class 1	Class 2	Confusion
restaurant review	food review	45%
planting	farm work	44%
movie information	movie review	34%
male model	layman handsome influencer	34%
roadster	luxury car	31%
rural performance	folklore	26%
domestic military intelligence	global military intelligence	24%
pet cat	pet dog	22%
layman beauty influencer	hairdressing	22%
financial management	stock market	21%

Table 9. **Top-10 class confusions in USV-1.0, using the Slowfast model.**

B.2. Quantitative Effect of Multi-modality

We demonstrate the quantitative effectiveness of the design of multi-modality by showing the easiest and hardest 10 classes for each multi-modality branch and their fusion.

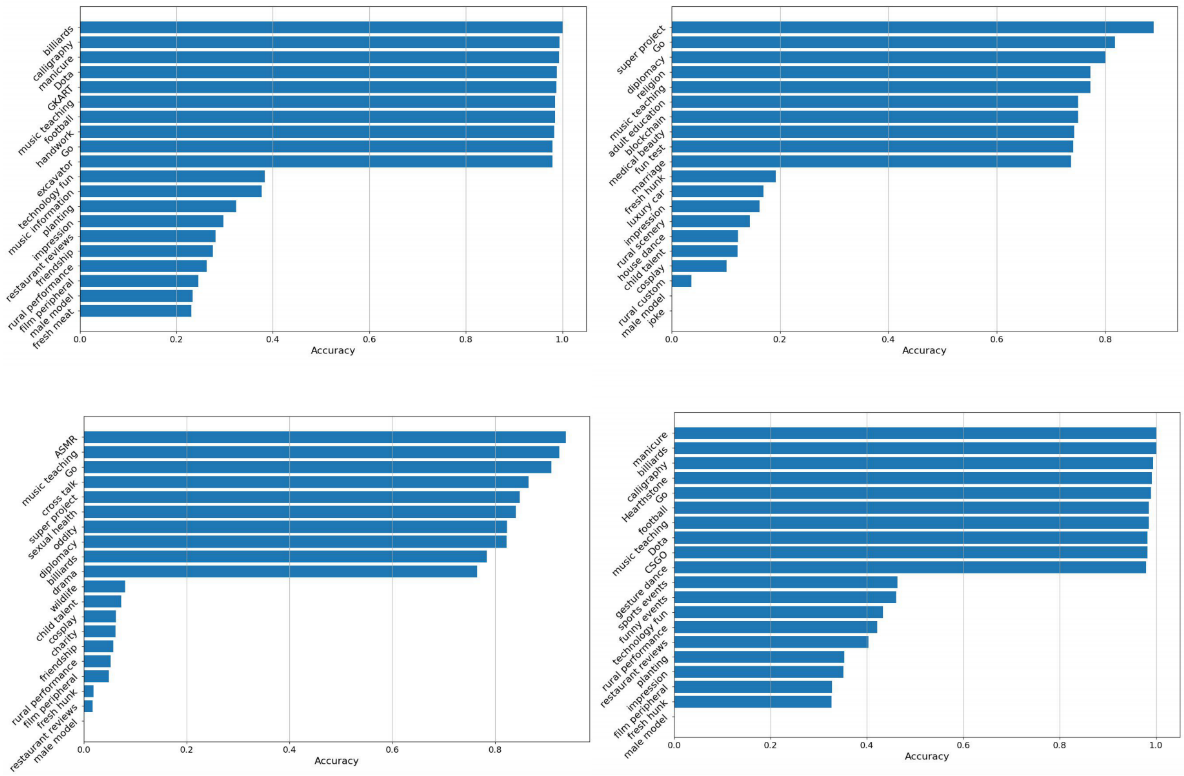

Figure 5. **Top-10 easy and hard classes. Upleft**: Visual branch. **Upright**: Textual Branch. **Downleft**: Audio Branch. **Downright**: Fused.

Class 1	Class 2	Confusion
male model	layman handsome influencer	93%
restaurant review	food review	49%
movie information	movie review	41%
planting	farm work	41%
luxury car	roadster	35%
science anecdote	cutting edge of science & technology	35%
parent-child interaction	children's sport	34%
domestic military intelligence	global military intelligence	24%
layman handsome influencer	layman beauty influencer	20%
global military intelligence	domestic military intelligence	20%

Table 10. **Top-10 class confusions in USV-1.0, using the MMF-Net model.**

Model	+Audio(Δmca)	+Text(Δmca)	+Both(Δmca)
TSN	movie mashup 0.15	blockchain 0.34	blockchain 0.39
	music video 0.14	religion 0.27	adult education 0.27
	impression show 0.13	adult education 0.25	hunting 0.24
	ACG dance 0.13	friendship 0.22	friendship 0.22
	tourist guide 0.10	hunting 0.22	sociality 0.20
	male model -0.21	male model -0.17	male model -0.23
	luxury car -0.18	grange -0.05	domestic military intelligence -0.11
	primary and secondary -0.12	luxury car -0.05	luxury car -0.09
	archaeology -0.09	domestic military intelligence -0.05	swim -0.04
	grange -0.07	child talent -0.04	tourism attraction -0.04
SlowFast	movie review 0.12	religion 0.40	adult education 0.39
	rural custom 0.11	adult education 0.35	blockchain 0.34
	ASMR 0.10	blockchain 0.28	religion 0.31
	singing 0.09	friendship 0.25	college 0.27
	K-pop dance 0.09	college 0.25	sociality 0.25
	male model -0.19	male model -0.14	male model -0.27
	luxury car -0.11	live music -0.03	luxury car -0.07
	restaurant review -0.08	spoof -0.03	sports star -0.03
	roadster -0.07	handsome influencer -0.01	spoof -0.03
	hunting -0.06	sports event -0.01	swim -0.02

Table 11. **Accuracy gain and loss.** Each column denotes the top-5 classes with the greatest gain and loss due to one or both modalities are fused.

It can be observed in Fig. 5 that classes with special visual clues such as *billiards* with similar green tables, *Dota* with similar game interface rank the best when using a single visual branch only; classes that have special audio features rank high for the audio branch. For example, *ASMR* which is a newly emerged videos in which authors make soft sounds with an extreme close sound field, and classes such as *music teaching* and *cross-talk* also rank higher than other branches; For the text branch, informative videos that usually possess detailed subtitles such as *diplomacy*, *adult education*, *blockchain* are also among the bests.

We further study the gain and loss caused by each modality in Tab. 11. We show 10 classes of the greatest accuracy gain and drop for Branch I when fused with Branch II, III, and both. Besides the gain for those acoustic and informative classes as mentioned before, we observe that the textual branch brings a dominant gain for the whole, which indicates the importance of visual-textual learning in UGC short-form videos.

C. Detailed Taxonomy

We list the full taxonomy of topic categories here. Words in bold are among the 32 macro topics, while those listed below are the micro topics expanded from the macro topic.

1. entertainment
- entertainment scene
- entertainment gossip

2. variety show
- variety show

3. film and television
- movie clip
- movie mashup
- titbits
- movie information
- movie review
- movie peripheral derivatives
- movie trailor

4. amusing
- spoof
- weirdo
- joke
- roast
- funny dubbing
- cross talk
- impression show
- autotune remix
- meme
- funny child
- funny animal
- funny event

5. beauty
- beauty influencer
- female model
- layman beauty influencer

6. handsome guy
- layman handsome influencer
- male model
- handsome influencer
- young hunk
- silver fox

7. science & technology
- science experiment
- science anecdote
- digital gadget
- automation
- scientific figure
- cutting edge of science & technology
- aerospace
- mechanical
- blockchain

8. sports
- fitness & diet
- basketball
- billiard
- yoga
- football
- water sports
- ping-pong
- run
- volleyball
- badminton
- tennis
- boxing
- kong-fu
- car racing
- swim
- cycling
- extreme sports
- chess & card game
- sports news
- sports event
- sports star
- children's sport

9. anime
- cosplay
- anime cut
- anime peripheral derivatives
- manga
- children's manga
- One Piece
- Naruto
- Detective Conan

10. game
- League of Legends
- PUBG
- Arena of Valor
- Speed Drifters
- Hearthstone
- CSGO
- Dota2
- Overwatch

11. vehicle
- motorcycle
- driving skill
- vehicle maintenance
- auto show
- excavator
- car tuning
- roadster
- luxury car
- traffic accident
- driving test

12. parenting
- pregnant mother
- child education
- child care
- child talent
- cute baby
- parent-child interaction

13. music
- singing
- live music
- music video
- instrument playing
- music teaching
- music information
- music review

14. fashion
hairdressing
nail art
skin care
outfit
make up
medical cosmetology
street shot
imitation makeup
men's fashion
fashion information
15. society
public welfare
natural disaster
anecdote
legal system
diplomacy
anti-corruption
regional affair
domestic affair
international news
16. pet
pet dog
pet cat
pet bird
pet reptile
17. tourism
customs
tourist attraction
tourist guide
18. nature
wild animal
wild plant
19. daily life
furniture
house decoration
good stuff
photo editing
ASMR
handwork
workplace
wedding ceremony
daily life tip
fun quiz
20. finance
stock market
real estate
financial management
financial information
entrepreneurship
financial figure
21. health
regimen
traditional medicine
medical science
sexual health
22. military
domestic military
intelligence
global military
intelligence
military figure
armed special police
weaponry equipment
military exercise
war history
23. history
historical figure
world history
domestic history
archaeology
24. education
primary and secondary school
college and university
adult education
vocational examination
language teaching
25. novelty seeking
over-fancy mind experiment
oddity
26. delicacy
mukbang
cooking tutorial
food review
restaurant review
snack
beverage
27. culture and art
calligraphy
painting
acrobatics
magic
Go chess
reading
antique collection
opera
folklore
live theatre
sculpture
building
region
28. horticulture
flower arrangement
planting
plant science
29. industry
construction
super project
manufacture
30. dance
K-pop dance
street dance
square dance
gesture dance

pole dance
ACG dance
folk dance
dance teaching

31. affection

love
marriage
family
sociality
friendship

32. countryside

cultivation
hunting
fishing
farm work
grange
rural performance
rural custom
rural scenery

References

[1] Bilibili. https://bilibili.com. 3

[2] easyocr. https://github.com/JaidedAI/EasyOCR. 3, 6, 9

[3] Ffmpeg. www.ffmpeg.com. 3

[4] Kwai. https://www.kwai.com/. 1

[5] mmaction2. https://github.com/open-mmlab/mmaction2/. 9

[6] Reels. https://about.instagram.com/blog/announcements/introducing-instagram-reels-announcement. 1

[7] textblob. https://github.com/sloria/TextBlob. 4

[8] Tiktok. https://www.tiktok.com/. 1

[9] Tiktok statistics. https://www.oberlo.ca/blog/tiktok-statistics. 1

[10] Youtube. https://youtube.com. 3

[11] Youtube revenue analysis. https://www.businessofapps.com/data/youtube-statistics/. 1

[12] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 2, 3

[13] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 3

[14] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 609–617, 2017. 2, 6

[15] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 5, 7, 9

[16] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736, 2018. 3

[17] James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, Taylor Van Vleet, Ullas Gargi, Sujoy Gupta, Yu He, Mike Lambert, Blake Livingston, et al. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296, 2010. 1

[18] Zhengyu Deng, Ming Yan, Jitao Sang, and Changsheng Xu. Twitter is faster: Personalized time-aware video recommendation from twitter to youtube. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 11(2):1–23, 2015. 1

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2, 6, 9

[20] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jurgen Gall, Rainer Stiefelhagen, and Luc Van Gool.

Holistic large scale video understanding. *arXiv preprint arXiv:1904.11451*, 2019. 2, 3

[21] Haoqi Fan, Yanghao Li, Bo Xiong, Wan-Yen Lo, and Christoph Feichtenhofer. Pyslowfast. https://github.com/facebookresearch/slowfast, 2020. 9

[22] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 6202–6211, 2019. 5, 7

[23] Basura Fernando, Hakan Bilen, Efstratios Gavves, and Stephen Gould. Self-supervised video representation learning with odd-one-out networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3636–3645, 2017. 2

[24] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, volume 1, page 5, 2017. 3, 6

[25] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 1

[26] Tengda Han, Weidi Xie, and Andrew Zisserman. Video representation learning by dense predictive coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 6

[28] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 961–970, 2015. 1, 3, 5

[29] Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1971–1980, 2016. 2

[30] Zhong Ji, Yaru Ma, Yanwei Pang, and Xuelong Li. Query-aware sparse coding for web multi-video summarization. *Information Sciences*, 478:152–166, 2019. 1

[31] Yu-Gang Jiang, Jingen Liu, A Roshan Zamir, George Toderici, Ivan Laptev, Mubarak Shah, and Rahul Sukthankar. Thumos challenge: Action recognition with a large number of classes, 2014. 1

[32] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014. 2, 3

[33] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 2, 3, 5, 6, 7

[34] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014. 2

[35] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *arXiv preprint arXiv:1807.00230*, 2018. 2

[36] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 3

[37] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011. 1, 2, 3, 5

[38] Hsin-Ying Lee, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Unsupervised representation learning by sorting sequences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 667–676, 2017. 2

[39] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. *arXiv preprint arXiv:2102.06183*, 2021. 2

[40] Tianhao Li and Limin Wang. Learning spatiotemporal features via video and text pair discrimination. *arXiv preprint arXiv:2001.05691*, 2020. 6

[41] Dahua Lin, Sanja Fidler, Chen Kong, and Raquel Urtasun. Visual semantic search: Retrieving videos via complex textual queries. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2657–2664, 2014. 2

[42] Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7083–7093, 2019. 7, 8

[43] Chunhui Liu, Yueyu Hu, Yanghao Li, Sijie Song, and Jiaying Liu. Pku-mmd: A large scale benchmark for continuous multi-modal human action understanding. *arXiv preprint arXiv:1703.07475*, 2017. 1

[44] Meng Liu, Liqiang Nie, Meng Wang, and Baoquan Chen. Towards micro-video understanding by joint sequential-sparse modeling. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 970–978, 2017. 2

[45] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008. 3

[46] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. The jester dataset: A large-scale video dataset of human gestures. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 2

[47] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 2, 6

[48] Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv preprint arXiv:1804.02516*, 2018. 2

[49] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE international conference on computer vision*, pages 2630–2640, 2019. 2, 3

[50] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 42(2):502–508, 2019. 1, 2, 3, 5

[51] Mathew Monfort, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Bowen Pan, Quanfu Fan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *arXiv preprint arXiv:1911.00232*, 2019. 2

[52] Liqiang Nie, Meng Liu, and Xuemeng Song. Multimodal learning toward micro-video understanding. *Synthesis Lectures on Image, Video, and Multimedia Processing*, 9(4):1–186, 2019. 1, 2

[53] Liqiang Nie, Xiang Wang, Jianglong Zhang, Xiangnan He, Hanwang Zhang, Richang Hong, and Qi Tian. Enhancing micro-video understanding by harnessing external sounds. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1192–1200, 2017. 2

[54] Sangmin Oh, Anthony Hoogs, Amitha Perera, Naresh Cuntoor, Chia-Chih Chen, Jong Taek Lee, Saurajit Mukherjee, JK Aggarwal, Hyungtae Lee, Larry Davis, et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*, pages 3153–3160. IEEE, 2011. 1, 2

[55] Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Learning joint representations of videos and sentences with web image search. In *European Conference on Computer Vision*, pages 651–667. Springer, 2016. 2

[56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 6

[57] Jamie Ray, Heng Wang, Du Tran, Yufei Wang, Matt Feiszli, Lorenzo Torresani, and Manohar Paluri. Scenes-objects-actions: A multi-task, multi-label video dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 635–651, 2018. 3

[58] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 123(1):94–120, 2017. 2, 3, 7, 9

[59] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1010–1019, 2016. 2

[60] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2020. 1, 2

[61] Abhinav Shukla, Stavros Petridis, and Maja Pantic. Learning speech representations from raw audio by joint audiovisual self-supervision. *arXiv preprint arXiv:2007.04134*, 2020. 6

[62] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014. 5

[63] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1, 2, 3, 5

[64] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Learning video representations using contrastive bidirectional transformer. *arXiv preprint arXiv:1906.05743*, 2019. 2

[65] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*, 2016. 2

[66] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 7

[67] Jiangliu Wang, Jianbo Jiao, Linchao Bao, Shengfeng He, Yunhui Liu, and Wei Liu. Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4006–4015, 2019. 2

[68] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016. 5, 6, 7, 8, 9

[69] Yinwei Wei, Xiang Wang, Weili Guan, Liqiang Nie, Zhouchen Lin, and Baoquan Chen. Neural multimodal cooperative learning toward micro-video understanding. *IEEE Transactions on Image Processing*, 29:1–14, 2019. 2

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

[71] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 2, 3, 7, 9

[72] Yong Xu, Qiuqiang Kong, Wenwu Wang, and Mark D Plumbley. Large-scale weakly supervised audio classification using gated convolutional neural network. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 121–125. IEEE, 2018. 6

[73] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018. 2

[74] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3165–3173, 2017. 2

[75] Jing Zhang, Yuting Wu, Jinghui Liu, Peiguang Jing, and Yuting Su. Low-rank regularized multimodal representation for micro-video event detection. *IEEE Access*, 8:87266–87274, 2020. 2

[76] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2, 3, 7, 9

[77] Qiusha Zhu, Mei-Ling Shyu, and Haohong Wang. Videotopic: Content-based video recommendation using a topic model. In *2013 IEEE International Symposium on Multimedia*, pages 219–222. IEEE, 2013. 1