Realistic Video Summarization through VISIOCITY: A New Benchmark and Evaluation Framework

Vishal Kaushal
IIT Bombay
vkaushal@cse.iitb.ac.in

Suraj Kothawade
UT Dallas
suraj.kothawade@utdallas.edu

Rishabh Iyer
UT Dallas
rishabh.iyer@utdallas.edu

Ganesh Ramakrishnan
IIT Bombay
ganesh@cse.iitb.ac.in

ABSTRACT

Automatic video summarization is still an unsolved problem due to several challenges. We take steps towards making it more realistic by addressing the following challenges. Firstly, the currently available datasets either have very short videos or have few long videos of only a particular type. We introduce a new benchmarking dataset called VISIOCITY which comprises of longer videos across six different categories with dense concept annotations capable of supporting different flavors of video summarization and other vision problems. Secondly, for long videos, human reference summaries, necessary for supervised video summarization techniques, are difficult to obtain. We present a novel recipe based on pareto optimality to automatically generate multiple reference summaries from indirect ground truth present in VISIOCITY. We show that these summaries are at par with human summaries. Thirdly, we demonstrate that in the presence of multiple ground truth summaries (due to the highly subjective nature of the task), learning from a single combined ground truth summary using a single loss function is not a good idea. We propose a simple recipe VISIOCITY-SUM to enhance an existing model using a combination of losses and demonstrate that it beats the current state of the art techniques. We also present a study of different desired characteristics of a good summary and demonstrate that a single measure (say F1) to evaluate a summary, as is the current typical practice, falls short in some ways. We propose an evaluation framework for better quantitative assessment of summary quality which is closer to human judgment than a single measure. We report the performance of a few representative techniques of video summarization on VISIOCITY assessed using various measures and bring out the limitation of the techniques and/or the assessment mechanism in modeling human judgment and demonstrate the effectiveness of our evaluation framework in doing so.

KEYWORDS

video summarization, benchmark, dataset, automatic evaluation, challenges

1 INTRODUCTION

Videos have become an indispensable medium for capturing and conveying information in many domains like entertainment (TV shows, movies, etc.), sports, personal events (birthday, wedding etc.), education (HOWTOs, tech talks etc.), to name a few. The increasing availability of cheaper and better video capturing and video storage devices have led to the unprecedented growth in the amount of video data available today. Most of this data, however, has a lot of redundancy, partly because of the inherent nature of videos (as a set of *many* images) and partly due to the 'capture-now-process-later' mentality. Consequently this has given rise to the need of automatic video summarization to produce shorter videos without significantly compromising the quality and quantity of information contained in them. What makes this especially challenging though is that an *ideal* summary is a) highly context dependent (depends on the purpose behind getting a video summary), b) subjective (even for the same purpose, preferences of two persons don't match) and c) depends on high-level semantics of the video (two visually different scenes could capture the same semantics or visually similar looking scenes could capture different semantics). Thus, there is no single right answer. For example, as far as context is concerned, one may want to summarize a surveillance video either to see a 'gist' of what all happened or to quickly spot any 'abnormal' activity. Similarly, as far as personal preferences or subjectivity is concerned, while summarizing a *Friends* video (a popular TV series), two users may have different opinion on what is 'important' or 'interesting'. As another example, when it comes to semantics at a higher level than what is captured by visual appearance, closeup of a player in soccer can be considered important if it is immediately followed by a goal, while not so important when it occurs elsewhere. These and other issues discussed below make it an interesting research problem with several papers pushing the state-of-the-art for newer algorithms and model architectures [6, 10, 19, 32, 34, 36–38, 40] and datasets [8, 26, 29]. However, as noted by a few recent works several fundamental issues remain to be addressed. We summarize these below.

Dataset: For a true comparison between different techniques, a benchmark dataset is critical. Almost all recent techniques have reported their results on TVSum [30] and SumMe [9] which have emerged as benchmarking datasets of sorts. However, since the average video length in these datasets is of the order of *only* 1-5 minutes, they are far from being effective in real-world settings. The effect and need of automatic video summarization is desired and more pronounced especially in long videos. While there have been many attempts at creating datasets for video summarization, they either a) have very short videos, or b) they have very few long videos and often only a particular type. A large dataset with a lot of different types of full-length videos with rich annotations to be able to support different techniques was one of the recommendations in [31], is still not a reality and is clearly a need of the hour [12]. One of the contributions of this work is to address this gap. We present a large and diverse video summarization dataset VISIOCITY (Video Summarization Based on Continuity, Intent and Diversity) with

a total of 67 videos spanning six categories: TV shows (*Friends*) , sports (soccer), surveillance, education (tech-talks), birthday videos and wedding videos. The videos have an average duration of about 50 minutes. In Table 1 and Table 2 we provide a comparison of VISIOCITY with other existing datasets. Furthermore, different flavors of video summarization, for example, query focused video summarization [28, 32, 33], are often treated differently and on different datasets. With its rich annotations (described in section 3.2), VISIOCITY can be used for other flavors of video summarization and also other video analysis tasks like captioning or action detection. Also, since the videos span across different well-defined categories, VISIOCITY is also suitable for more in-depth domain specific studies on video summarization [26, 31].

Nature of supervision: Supervised techniques tend to work better than unsupervised techniques [12, 38] because of learning directly from user summaries. In a race to achieve better performance, most state of the art techniques are based on deep architectures and are thus data hungry. The larger the dataset and more the number of reference summaries to learn from, the better. However, it is very costly to create ground truth summaries when it comes to long videos. It becomes increasingly expensive and, beyond a point, infeasible to get these reference summaries from humans. Also, this is not scalable to experiments where reference summaries of different lengths are desired [10]. We address this need by providing a recipe for ground truth creation from indirect ground truth provided by humans. Indirect ground truth is annotations in the form of concepts present in each shot. We argue, getting indirect ground truth for long videos is more feasible and much more objective. As elaborated in section 4.2, this allows for automatically generating as many different ground truth summaries as desired by a machine learning model and more importantly, with different characteristics. Since there is no single right answer, indirect ground truth annotations, being at a meta level, can be seen as a *generator* of ground truth summaries [31].

A related problem is the fact that current supervised techniques are trained using a 'combined' ground truth summary, either in form of combined scores from multiple ground truth summaries or scores [4, 12, 37] or in form a set of ground truth selections, as in dP-PLSTM [37]. However, since there can be multiple correct answers, a reason for low consistency between user summaries [13, 23] combining them into one misses out on the separate flavors captured by each of them. Combining many into one set of scores also runs the risk of giving more emphasis to 'importance' over and above other desirable characteristics of a summary like continuity, diversity etc. This is also noted by [1, 40] where they argue that supervised learning approaches, which rely on the use of a combined ground-truth summary, cannot fully explore the learning potential of such architectures. The necessity to deal with different kind of summaries in different ways was also observed by [31]. [1, 40] use this argument to advocate the use of unsupervised approaches. Another alternative however is to make a model learn directly from multiple ground truth summaries instead of a 'combined' ground truth summary. The learning can be further enhanced by employing a combination of loss functions each measuring the deviation from different desired characteristics of good summaries. Using this principle, we demonstrate a simple extension to a mixture model (see section 5) which performs better than other state of the art techniques on VISIOCITY.

Evaluation: With a desire to be comparable across techniques, almost all recent work evaluates their results using F1 score [4, 12, 39]. This approach of assessing a candidate summary vis-a-vis a ground truth summary sounds good, but it has following limitations: 1) The user summaries are themselves inconsistent with each other, as already noted above [13, 23]. As a workaround, the assessment is done with respect to the nearest neighbor [10, 30]. This is still not free from the problem of having limited ground truth

summaries. The number of right answers can be large, especially in case of long videos and a good candidate may get a low score just because it was not fortunate to have a matching user summary. Furthermore, F1 seems to be good to measure the 'closeness' with a user summary, but the numbers can be deceptive as it is affected by the segmentation used as a post processing step in typical video summarization pipeline [23]. F1 is also not well suited to measure other characteristics of a summary like continuity or diversity. Two summaries may have same F1 score, and yet one may be more continuous (and hence visually more pleasurable to watch) than another (more discussion on this in section 4.1). While F1 has its utility in measuring the goodness of a summary, instead of over dependence on one measure, we propose using a suite of measures to capture various aspects of a summary like continuity, diversity, redundancy, importance etc. As discussed in section 4.1, these use the annotation provided in VISIOCITY (indirect ground truth), as against comparing the candidate to ground truth summaries.

In what follows, we first talk about the related work in the areas outlined above. Then we give details of the VISIOCITY benchmark dataset (section 3). In section 4 we introduce different supervised scoring functions that can be used to characterize a good summary followed by a recipe to generate ground truth summaries and the proposed evaluation framework. Thereafter, in section 5 we introduce a simple recipe to enhance a model which beats the state of the art by making use of multiple ground truth summaries and by learning using the multiple loss functions. Finally, in section 6 we present the extensive experiments and analysis demonstrating the different characteristics of good summaries, the effectiveness of our recipe, necessity of proposed evaluation framework, a better approach to learning and an analysis of the performance of a few representative state of the art techniques on our dataset. This is followed by conclusion and future work.

2 RELATED WORK

2.1 Datasets

As noted earlier, the currently available annotated datasets for video summarization have either too few videos or too short videos or videos of only one type. Table 1 summarizes the important statistics of the different datasets. MED Summaries Dataset [26] consists of 160 annotated videos of length 1-5 minutes, with 15 event categories like birthday, wedding, feeding etc. The annotation comprises of segments and their importance scores. TVSum [29] consists of 50 videos (average length 4 minutes) from 10 categories. with importance scores provided by 20 annotators for each 2 second snippet. The videos correspond to very short events like 'changing vehicle tires', 'making sandwich' etc. The UT Egocentric Dataset [16] consists of long and annotated videos captured from head mounted cameras. However, though each video is very long, there are only 4 videos and they are of one type, i.e. egocentric. SumMe [8] consists of 25 videos with an average length of about 2 min. The annotation is in form of user summaries of length between 5% to 15%. Each video has 15 summaries. The VSUMM dataset [2] consists of two datasets. Youtube consists of 50 videos (1-10 min) and OVP consists of 50 videos of about 1-4 min from the Open Video project. Each video has 5 user summaries in the form of set of key frames. Tour20 [24] consists of 140 videos with a total duration of 7 hours and is designed primarily for multi video summarization. It is a collection of videos of a tourist place. The average duration of each video is about 3 mins. TV Episodes dataset [35] consists of 4 TV show videos, each of 45 mins. The total duration is 3 hours. A recent dataset, UGSum52 [17] offers 52 videos with 25 user generated summaries each. LOL [5] consists of online eSports videos from the League of Legends. It consists of 218 videos with each video being between 30-50 mins. The associated summary videos are 5 - 7 mins long. While this dataset is significantly larger compared to the other datasets, it is limited only to a single domain, i.e. eSports.

Name	# Videos	Duration of Videos	Total Duration	# summ	# cat
SumMe [8]	25	Avg: 2 min, 9	1 Hour, 10 min	15-18	-
TVSum [29]	50	Avg: 4min, 11 sec	3.5 Hours	20	10
MED Summaries [26]	260	Dur: 1-5 mins, Avg: 2.5min	9 Hours	2-4	15
UT Egocentric [16]	4	Avg: 254 mins	16 Hours	-	1
Youtube 1 [2]	50	Dur: 1-10 min, Avg: 1 min, 39 sec	1.38 Hours	5	-
Youtube 2 [2]	50	Dur: 1-4 min, Avg: 2min, 54sec	2.41 Hours	5	-
Tour20 [24]	140	Avg: 3 min	7 Hours	-	-
TV Episodes [35]	4	Avg: 45 min	3 Hours	-	1
LOL [5]	218	Dur: 30 to 50 min	-	-	1
VISIOCITY (OURS)	**67**	**Dur: 14-121 mins, Avg: 55 mins**	**71 hours**	**160**	**5**

Table 1: Key statistics of available datasets. "-" means the corresponding information was not available. # summ here means # reference summaries per video. "Dur" stands for Duration and #cat is # categories or domains.

[27] have extended the UTE dataset to 12 videos and have provided concept annotations, but they are limited to only egocentric videos and do not support any concept hierarchy. The scores annotations, as in TVSum etc. are superior indirect ground truth annotations, but are limited only to importance scores. VISIOCITY on the other hand comes with dense concept annotations for each shot. To the best of our knowledge, VISIOCITY is one of its kind large dataset with many long videos spanning across multiple categories and annotated with rich concept annotations for each shot.

2.2 Techniques for Automatic Video Summarization

A number of techniques have been proposed to further the state of the art in automatic video summarization. Most video summarization algorithms try to optimize several criteria such as diversity, coverage, importance and representation. [10, 14] proposed a submodular optimization framework with combining weighted terms for each of these criteria in a mixture and trained using loss augmented inference. Several recent papers have used deep learning to model importance. [38] propose a LSTM based model which they call vsLSTM. They also propose to add diversity using a Determinantal Point Process along with an LSTM (which they call dppLSTM). [40] uses deep reinforcement learning and model with diversity and representation based rewards. [12] uses an encoder-decoder based attention model and has obtained the best results on TV-Sum and Summe datasets. [32] studied query based summarization where they incorporate a query relevance model along with their submodular framework with diversity, representation and importance. [6] attempts to address video summarization via attention-aware and adversarial training.

2.3 Evaluation

As presented above, evaluation is challenging task owing to the multiple definitions of success. Early approaches [20, 21] involved user studies with the obvious problem of cost and reproducibility. With a move to automatic evaluation, every new technique of video summarization came with its own evaluation criteria making it difficult to compare results different techniques. Some of the early approaches included VIPER [3], which addresses the problem by defining a specific ground truth format which makes it easy to evaluate a candidate summary, and SUPERSEIV [11] which is an unsupervised technique to evaluate video summarization algorithms that perform frame ranking. VERT [18] on the other hand was inspired by BLEU in machine translation and ROUGE in text summarization. Other techniques include pixel-level distance between keyframes [15], objects of interest as an indicator of similarity [16] and precision-recall scores over key-frames selected by human annotators [7]. It is not surprising thus that [31] observed that researchers should at least reach a consensus on what are the best procedures and metrics for evaluating video abstracts. They

concluded, that a detailed research that focuses exclusively on the evaluation of existing techniques would also be a valuable addition to the field. This is one of the aims of this work. More recently, computing overlap between reference and generated summaries has become the standard framework for video summary evaluation. However, all these methods which require comparison with ground truth summaries suffer from the challenges highlghted above. Yeung et al. observed that visual (dis)similarity need not mean semantic (dis)similarity and hence proposed a text based approach of evaluation called VideoSet. The candidate summary is converted to text and its similarity is computed with a ground truth textual summary. That text is better equipped at capturing higher level semantics has been acknowledged in the literature [25] and form the motivation behind our proposed evaluation measures. However, our measures are different in the sense that a summary is not converted to text domain before evaluating. Rather, how important its selections are, or how diverse its selections are, is computed from the rich textual annotations in VISIOCITY. This is similar in spirit to [27], but there it has been done only for egocentric videos. As noted by [23] "the limited number of evaluation videos and annotations further magnify this ambiguity problem". Our VISIOCITY framework precisely hits the nail by not only offering a larger dataset but also in proposing a richer evaluation framework better equipped at dealing with this ambiguity.

3 VISIOCITY: A NEW BENCHMARK

We introduce VISIOCITY, a new challenging benchmark dataset for video summarization. VISIOCITY stands for Video Summarization based on Continuity, Intent and Diversity. As noted earlier, VISIOCITY is a large challenging dataset, which is an order of magnitude larger (both in terms of total number of hours as well as the duration of each video) compared to existing datasets for video summarization.

3.1 Videos

VISIOCITY is a diverse collection of 67 videos spanning across six different categories: TV shows (*Friends*) , sports (Soccer), surveillance, education (Tech-Talks), birthday videos and wedding videos. The videos have an average duration of about 50 mins. VISIOCITY is compared to existing datasets in Table 1 and Summary statistics for VISIOCITY are presented in Table 2. Publicly available Soccer, Friends, Techtalk, Birthday and Wedding videos were downloaded from internet. TV shows contains videos from a popular TV series *Friends*. They are typically more aesthetic in nature and professionally shot and edited. In sports category, VISIOCITY contains Soccer videos. These videos typically have well defined events of interest like goals or penalty kicks and are very similar to each other in terms of the visual features. Under surveillance category, VISIOCITY covers diverse settings like indoor, outdoor, classroom, office and lobby. The videos were recorded using our own surveillance

Domain	# Videos	Duration	Total Duration
Sports(Soccer)	12	(37,122,**64**)	12.8
TVShows (Friends)	12	(22,26,**24**)	4.8
Surveillance	12	(22,63,**53**)	10.6
Educational	11	(15,122,**67**)	12.28
Birthday	10	(20,46,**30**)	5
Wedding	10	(40,68,**55**)	9.2
All	67	(26,75,**49**)	54.68

Table 2: Key Statistics of VISIOCITY. Third Column is in minutes (min/max/avg) and fourth column is in hours.

cameras. These videos are in general very long and are mostly from static continuously recording cameras. Under educational category, VISIOCITY has tech talk videos with static views or inset views or dynamic views. In personal videos category, VISIOCITY has birthdays and wedding videos. These videos are typically long and unedited.
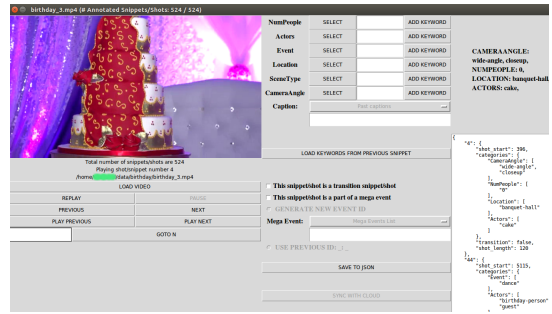
3.2 Annotations

The annotations are in the form of concepts marked for each unit of annotation as against asking annotators to prepare actual summaries. Indirect ground truth offers several advantages. Firstly, being at a higher level it can be seen as a 'generator' of ground truth summaries and thus allows for multiple solutions (reference summaries) of different lengths with different desired characteristics and is easy to scale. Secondly, it is more informative. And thirdly, it makes the annotation process more objective and easier than asking the users to directly produce reference ground truth summaries.

As far as unit of annotation is concerned, it had to be small enough to not contain redundancy within it and large enough to make the annotation process viable. Thus VISIOCITY has annotations at the level of shots. Wherever the output of the shot detector was not satisfactory (for example in surveillance videos coming from static cameras), or in Soccer videos where visual content change is too frequent in spite of no change in semantic information, we used fixed length (2-5 seconds) segments for annotation.

The concepts are organized in categories instead of a long flat list. Example categories include 'actor', 'entity', 'action', 'scene', 'number-of-people', etc. The concept keywords within each category are carefully selected based on the category of the video through domain expertise. Categories provide a natural structuring to make the annotation process easier and also support for at least one level hierarchy of concepts for concept-driven summarization.

In addition to the concept annotations for every shot, there are additional annotations to mark consecutive shots which *together* constitute a cohesive event (we call such occurences as *mega-events*). To better understand the idea of *mega-events*, consider the case of a 'goal' event in Soccer. A model trained to learn importance scores (only) would do well to pick up the 'goal' snippet. However, such a summary will not be pleasing at all because what is required in a summary in this case is not just the ball entering the goal post but the build up to this event and probably a few shots as a followup. In this way, *mega-event* annotations capture the notion of continuity.

While past work has made use of other forms of indirect ground truth like asking annotators to give a score or a rating to each shot [30], using textual concept annotations offers several advantages. First, it is easier and more accurate for annotators to mark all keywords applicable to a shot/snippet than for them to tax their brain and give a rating (especially when it is quite subjective and requires going back and forth over the video for considering what is *more important* or *less important*). Second, when annotators are asked to provide ratings, they often suffer from chronological bias. One work addresses this by showing the snippets to the annotators in random

Figure 1: Annotation and visualization tool developed by us used in VISIOCITY framework

order [29] but it doesn't work for long videos because an annotator cannot remember all of these to be able to decide the relative importance of each. Third, the semantic content of a shot is better captured through text [25, 35]. This is relevant from an 'importance' perspective as well as 'diversity' perspective. As noted earlier, two shots may look visually different but could be semantically same and vice versa. Text captures the right level of semantics desired by video summarization. Also, when two shots have same rating, it is not clear if they are semantically same or they are semantically different but equally important. Textual annotations brings out such similarities and dissimilarities more effectively. Fourth, as already noted, textual annotations make it easy to adapt VISIOCITY to a wide variety of problems.

Annotation Protocol: A group of 13 professional annotators were tasked to annotate videos (without listening to the audio) by marking all applicable keywords on a snippet/shot using a Python GUI application developed by us for this task. The tool allows an annotator to go over the video unit by unit (shot/snippet) and select the applicable keywords using a simple and intuitive GUI and offers some convenience features like copying the annotation from previous snippet, which comes in handy where there are are a lot of consecutive identical shots, for example in surveillance videos (Figure 1).

Special caution was exercised to ensure high quality annotations. Specifically, the guidelines and protocols were made as objective as possible, the annotators were trained through sample annotation tasks, and the annotation round was followed by two verification rounds where both precision (how accurate the annotations are) and recall (whether all events of interest and continuity information has been captured in the annotations) were verified by another set of annotators. Whatever inconsistencies or inaccuracies were found and could be automatically detected, were included in our automatic sanity checks which were run on all annotations.

4 GROUND TRUTH SUMMARIES AND EVALUATION FRAMEWORK

4.1 What characterizes a good summary?

Diversity - A summary which does good on diversity is non redundant. It contains segments quite *different* from one another. *different* could mean different things in terms of content alone (i.e. one doesn't want two similar looking snippets in a summary) or in terms of content and time (i.e. one doesn't want visually similar *consecutive* snippets, but *does* want visually similar snippets that are separated in time) or in terms of the concepts covered (one does not want too many snippets covering the same concept and would rather want a few of all concepts). In surveillance videos for example, one would like to have a summary which doesn't have too many visually similar consecutive and hence redundant

snippets, but does have visually similar snippets that are separated in time. For instance, consider a video showing a person entering her office at three different times of the day. Though all three look similar (and will have identical concept annotations as well), all are desired in the summary. Let V be the set of shots/snippets in a video and $X \subset V$ be a candidate summary. With regards to the quantitative formulation, we define the first flavor of diversity as $Div_{sim}(X) = \max \min_{i,j \in X} d_{ij}$ where d_{ij} is IOU measure between snippets i and j based on their concept vectors. For the other two flavors of diversity, we define diversity clustered: $Div(X) = \sum_{i=1}^{|C|} \max_{j \in X \cap C_i} r_j$ where C are the clusters, which can be defined over time (all consecutive similar snippets form a cluster) giving Div_{time} or concepts (all snippets covering a concept belong to a cluster) giving $Div_{concept}$ and r_j is the importance rating of a snippet j. When optimized, this function leads to the selection of the best snippet from each cluster. This can be easily extended to select a finite number of snippets from each cluster instead of the best one.

MegaEventContinuity - element of continuity makes a summary pleasurable to watch. There is a thin line between modelling redundancy and continuity when it comes to visual cues of frames. Some snippets might be redundant but are important to include in the summary from a continuity perspective. To model the continuity, VISIOCITY has the notion of mega-events as defined above. To ensure no redundancy within a mega event, we define mega-events to be as tight as possible, meaning they should contain bare minimum snippets just enough to indicate the event. A non-mega event snippet is continuous enough to exist in the summary on its own and a mega event snippet needs other adjacent snippets to be included in the summary for visual continuity. We define this aspect quantitatively as follows $MegaCont(X) = \sum_{i=1}^{E} r^{mega}(M_i)|X \cap M_i|^2$ where, E is the number of mega events in the video annotation, $r^{mega}(M_i)$ is the rating of the mega event M_i and is equal to $\max_{\forall s \in M_i} r(s)$, A is the annotation of video V, that is, a set of snippets such that each snippet s has a set of keywords K^s and information about mega event, M is a set of all mega events such that each mega event M_i ($i \in 1, 2, \cdots E$) is a set of snippets that constitute the mega event M_i

Importance / Interestingness - This is the most obvious characteristic of a good summary. For some domains like sports, there is a distinct importance of some snippets (for eg. score changing events) over other snippets. This however is not applicable for some other domains like tech talks where there are few or no distinctly important events. With respect to the annotations available in VISIOCITY, importance of a shot or snippet is defined by the ratings of the keywords of a snippet. These ratings come from a mapping function which maps keywords to ratings for a domain. The ratings are defined from 0 to 10 with 10 rated keyword being the most important and 0 indicated an undesirable snippet. We assign ratings to keywords based on their importance to the domain and average frequency of occurence. Given the ratings of each keyword, rating of a shot is defined as $r_s = 0$ if $\exists i : r_{K_i^s} = 0$, and $r_s = \max_i r_{K_i^s}$ otherwise. Here K^s is the set of keywords of a snippet s and $r_{K_i^s}$ is the rating of a particular keyword K_i^s. Thus importance function can be defined as: $\text{Imp}(X) = \sum_{s \in X \cap A \setminus M} r(s)$. Note that when both importance and mega-event-continuity is measured, we define the importance only on the snippets which are not mega-events since the mega-event-continuity term above already handles the importance of mega-events.

4.2 Recipe for automatic generation of ground truth summaries

The indirect ground truth annotations in VISIOCITY allows us to generate ground truth summaries as follows.

The above functions modeling different characteristics of a good summary are a natural choice to be used as the building blocks of an ideal summary generator. A good summary should have these different characteristics to different degrees (and that is what makes one good summary different from another good summary, generating multiple correct answers). Thus we define a scoring function as a weighted mixture of above components. This composite scoring function takes an annotated video *keywords* and *mega-events* defined over snippets/shots) and generates a set of candidate (*ground-truth*) summaries which supervised or semi-supervised summarization algorithms can use.

Given X, a set of snippets of a video V, let $score(X)$ be defined as: $score(X, \Lambda) = \lambda_1 \text{MegaCont}(X) + \lambda_2 \text{Imp}(X) + \lambda_3 \text{Div}(X)$ This scoring function is parameterized on Λ's which assign the relative importance of these terms. This scoring function is approximately optimized via a greedy algorithm [22] to arrive at the candidate ground truth summaries.

Different configuration of λ generates different summaries. We observe (as demonstrated in section 6) that maximizing one characteristic doesn't necessarily and automatically lead to the maximization of another characteristic. In this sense these are orthogonal characteristics modelling different good aspects of a summary. Hence, some combinations would fare well on all while some would fare well on some, not so well on others. To understand what *are* the best combinations corresponding to the multiple right answers, we borrow from the notion of Pareto optimality. Pareto optimality is a situation that cannot be modified so as to make any one individual or preference criterion better off without making at least one individual or preference criterion worse off. Beginning with a random element in the pareto-optimal set, we iterate over remaining elements to decide whether a new element should be added or old should be removed or new element should be discarded. This is decided on the basis of the performance on various measures. A configuration is better than another when it is better on all measures, otherwise it is not.

We show in section 6 that the automatic ground truth summaries so generated are at par with the human summaries both qualitatively and quantitatively.

For assessing a candidate summary we advocate against the use of a single measure. Since different measures model different characteristics, true and wholesome assessment of a candidate summary can only be done when all measures (including the existing measures like F score) are used. Results and observations from our extensive experiments corroborate this fact.

5 A RECIPE TO ENHANCE AN EXISTING MODEL

Following [10] we formulate the problem of automatic video summarization as a subset selection problem where a weighted mixture of set functions is maximized to produce an optimal summary for a desired budget. In our experiments we generate summaries in the range of 1% to 5% of the video length as for long videos it yields summaries of reasonable length. The weights of the model are learnt using the following large margin framework as described in [10]. We train a simple model having only a submodular facility location term and a modular importance term with many automatic ground truth summaries and a margin loss which combines the feedback from different evaluation measures. The facility location function is defined as $f_{fl}(X) = \sum_{v \in V} \max_{x \in X} sim(v, x)$ where v is an element from the ground set V and $sim(v, x)$ measures the similarity between element v and element x. Facility Location governs the representativeness aspect of the candidate summaries and is monotone submodular.. The importance scores are taken from the VASNet model [4] and the vsLSTM model [37] trained on VISIOICTY. We call this proposed method VISIOCITY-SUM. We demonstrate that a simple model like this out performs the current techniques (state

of the art on TVSum and SumMe) on VISIOCITY dataset because of learning from multiple ground truth summaries and learning from mutliple loss functions, better equipped at capturing different characteristics of a summary.

6 EXPERIMENTS AND RESULTS

6.1 Experiment Setup and Implementation Details

For analysis of and comparison with human summaries, F1 score of any candidate summary is computed with respect to all human ground truth summaries following [37]. We report both avg and max across such scores. Thus max-F1 score can be seen as the F1 score with respect to the most matching (nearest) human summary. To calculate F1 scores of human summaries with respect to human summary, we compute max and avg in a leave-one-out fashion.

For analysis of different techniques on the VISIOCITY dataset, we report their F1 scores computed against the automatically generated summaries as a proxy for human summaries. We generate 100 automatic summaries for each video. All target summaries are generated such that their lengths are 1% to 5% of the video length. We test the performance of three different representative state of the art techniques on the VISIOCITY benchmark. vsLSTM [37] is a supervised technique that uses BiLSTM to learn the variable length context in predicting importance scores. VASNet [4] is a supervised technique based on a simple attention based network without computationally intensive LSTMs and BiLSTMs. DR-DSN [40] is an unsupervised deep-reinforcement learning based model which learns from a combined diversity and representativeness reward on scores predicted by a BiLSTM decoder. All three learn from a single combined ground truth summary per video. To generate a candidate machine generated summary from the importance scores predicted by vsLSTM, VASNET and DR-DSN, we follow [37] to convert them into machine generated summary of desired length (max 5% of original video). Our proposed model, VISIOCITY-SUM learns from multiple ground truth summaries and outputs a machine generated summary as a subset of shots.

In all tables, AF1 refers to Avg F1 score, MF1 refers to Max F1 score (nearest neighbor score), IMP, MC, DT, DC and DSi refer to the importance score, mega-event continuity score, diversity-time score, diversity-concept score and diversity-similarity score respectively, as measured by the functions introduced above. All figures are in percentages.

6.2 Degree of consistency and inconsistency in human summaries

We asked a set of 11 users (different from the annotators) to generate human summaries for two randomly sampled videos of each domain. The users were supposed to look at the video and mark segments they feel should be included in the summary such that the length of the summary remains between 1% to 5% of the original video. The procedure followed was similar to that of SumMe [9].

On closely examining the human selections, we observed that they are characterized by consistency to the extent there are important scenes in the video, for example a goal in Soccer videos. On the other hand, in the absence of such clear interesting events, the interestingness is more spread out and is more subjective, leading to higher inconsistency. We present more elaborate results and analysis in the supplementary material.

6.3 Orthogonal characteristics of good summaries and our measures are good in capturing in those characteristics

We assessed the human summaries using the set of performance measures presented above and found that while our measures get good scores on the human summaries, thus ascertaining their utility, there is also a lot of variation among the scores of different human summaries and this is captured well by our measures. We present the performance of human summaries on different measures in Table 3. We also plot the scores of different measures for different types of summaries and report them in Figure 2 for Soccer videos. We see that the measures defined by us get high values for human summaries as compared to random and uniform baselines. We include similar plots for other domains in the supplementary material. We also observe that a summary could be low on one measure, and yet is a good summary and this is captured by some other measure. Further, we also observe that the expected characteristics differ across different domains and the type of a particular video. It is also important to note that optimizing for one measure doesn't necessarily help the other measures thereby ascertaining that these measures are individually important. In the supplementary material we present a more detailed depiction of this interplay among different measures. As a simple example, a summary maximizing importance will do well to capture the goals in a soccer video, but unless there is an element of mega-event continuity in it, some shots preceeding the goal and following the goal will not be in the summary and the summary will not be visually pleasing.

6.4 Automatic summaries are at par with human summaries

Domain	Technique	AF1	MF1	IMP	MC	DT	DC	DSi
Soccer	Human	30	45	56	55	75	84	85
	Uniform	6	9	30	19	30	52	82
	Random	5	9	30	22	30	51	81
	Auto	27	37	83	88	82	90	80
Friends	Human	24	38	55	46	72	70	88
	Uniform	5	9	31	7	89	66	90
	Random	6	13	31	16	28	38	85
	Auto	25	41	87	69	76	85	81
Surveillance	Human	35	56	58	65	45	79	80
	Uniform	6	8	12	9	12	49	55
	Random	6	8	13	12	13	46	55
	Auto	31	40	80	85	82	99	88
TechTalk	Human	20	43	55	-	52	91	67
	Uniform	7	9	49	-	29	45	60
	Random	6	10	51	-	32	49	56
	Auto	25	43	86	-	78	93	96
Birthday	Human	21	31	56	38	48	70	83
	Uniform	6	9	48	12	79	73	82
	Random	6	10	48	16	47	57	78
	Auto	17	30	86	81	63	91	84
Wedding	Human	21	39	57	39	46	69	76
	Uniform	5	8	42	11	87	73	80
	Random	5	9	42	18	40	54	78
	Auto	14	21	81	79	71	95	88

Table 3: Performance of Human and Auto summaries for different domains. TechTalk videos do not have MegaEvents

To compare how the automatically generated ground truth summaries fare with respect to the actual human summaries, we generated 100 automatic summaries per video of about the same length as the human summaries. We then compute scores of different measures and report the numbers in Table 3 and also report their performance pictorially in figure 2 as against uniform and random summaries.

We see that automatic and human summaries are both much better than random on all the evaluation criteria. Next, we see that both the human and the automatic summaries are close to each other in terms of the F1 metric. We also see that the automatic summaries have the highest Importance, Continuity and Diversity
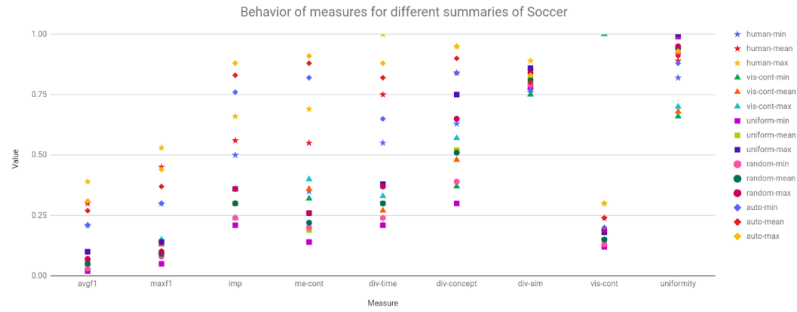
Figure 2: Behavior of different measures across for different types of summaries for Soccer videos. 'vis-cont' summaries are visually continuous summaries made by just picking a set of continuous shots

scores. This is not surprising as they are obtained at the first place by optimizing a combination of these criteria.

We also compare the human and automatic summaries qualitatively and see that a) it is very hard to distinguish the automatic summaries from humans though they are automatically generated, and b) they form very good visual summaries. We observe, though, that a perfect match is neither possible nor expected in the spirit of multiple right answers. We compare their selections in the supplementary material.

6.5 Analysis of performance of different techniques on VISIOCITY

As another contribution of this work, we present the performance of some representative summarization techniques tested on VISIOC-ITY benchmark and report their numbers in Table 4. We make the following observations: a) DR-DSN tries to generate a summary which is diverse. As we can see in the results, it almost always gets high score on the diversity term. Please note that the way we have defined these diversity measures, diversity-concept (DC) and diversity-time (DT) have an element of importance in them also. On the other hand, diversity-sim (DSi) is a pure diversity term where DR-DSN almost always excels. b) Due to this nature of DR-DSN, when it comes to videos where the interestingness stands out and importance clearly plays a more important role, DR-DSN doesnt perform well. In such scenarios, vsLSTM is seen to perform better, closely followed by VASNET. c) It is also interesting to note that while two techniques may yield similar scores on one measure, for example vsLSTM and VASNET for Soccer videos (Table 9), one of them, in this case vsLSTM, does better on mega-event continuity and produces a desirable characteristic in the summary. This further strengthens our claim of having a set of measures evaluating a technique or a summary rather than over dependence on one, which may not fully capture all desirable characteristics of good summaries. d) We also note that even though DR-DSN is an unsupervised technique, it is a state of the art technique when tested on tiny datasets like TVSum or SumMe, but when it comes to a large dataset like VISIOCITY, with more challenging videos, it doesn't do well, especially on those domains where there are clearly identifiable important events for example in Soccer (goal, save, penalty etc.) and Birthday videos (cake cutting, etc.). In such cases, models like vsLSTM and VASNET perform better as they are geared towards learning importance. In contrast, since the interstingness level in videos like Surveillance and Friends is more spread out, DR-DSN does relatively well even without any supervision.

We also report the performance of our simple model extension recipe VISIOCITY-SUM and compare them with the different models. We experimented with four flavors of our model. Three using the margin loss coming from importance, mega event continuity

Domain	Technique	AF1	MF1	IMP	MC	DT	DC	DSi
Soccer	Auto	59.3	93.3	83.2	84.3	82.6	85.9	76.2
	DR-DSN	2.8	8.9	23.7	20.3	23.2	30.4	**83.4**
	VASNET	28.4	43.4	**63**	49.3	62.1	67.4	75.2
	vsLSTM	**31.9**	**48.2**	62.2	60.1	62	69.5	76.5
	Ours	**32.6**	**50.3**	**64.2**	**62.6**	**63.4**	**72.2**	78.7
	Random	3.4	9.3	25.7	18.5	25.5	39.2	80.5
Friends	AUTO	66.3	96.9	87.8	84.6	80.3	89.8	83.1
	DR-DSN	4.3	9.4	19.1	6.9	65.7	51.5	**98.5**
	VASNET	**17**	**29.6**	**41**	39.3	49	60.6	86.7
	vsLSTM	15.5	27.2	40.4	39.2	64.7	59	91.1
	Ours	**17.4**	**31.2**	**42.5**	**40.5**	50.2	**64**	90.3
	Random	7.7	17.9	31.5	19.8	34.8	45.2	85.9
Surveillance	Auto	62.4	96.8	81.8	83.2	78.6	98	85.2
	DR-DSN	10	17.7	33.6	20.2	21.8	54.5	**57.2**
	VASNET	**19.4**	**31.4**	**39.5**	42.6	28.4	65.4	37.6
	vsLSTM	10.3	23.6	34.4	18.4	22.8	55.2	**58.4**
	Ours	**20.5**	**32.6**	**41.7**	**44.3**	**29.6**	**68.2**	38.5
	Random	3.9	8	16.6	12	15.3	49.4	69.4
TechTalk	Auto	64.7	91.5	79.8	-	80.5	88.4	94
	DR-DSN	13.5	22.5	49.3	-	24.8	29.9	35.2
	VASNET	**18.2**	**35.7**	52.1	-	**47.3**	43.3	43.2
	vsLSTM	15.1	32.2	**60.3**	-	38.8	35.3	41.7
	Ours	**18.7**	**37.5**	53.2	-	**50**	45.8	45.5
	Random	4.5	9.7	38.5	-	28	44	40.6
Birthday	Auto	67.3	97.2	89.7	88.6	68.1	90.8	81.3
	DR-DSN	8.1	14.2	54.7	14.1	**79.4**	63.6	**74.9**
	VASNET	21.6	37.6	50.1	30	36.2	47	48.7
	vsLSTM	**27.3**	**42.1**	**72.1**	57.2	59.6	**67.1**	73.6
	Ours	**28**	**44.3**	**74.8**	**60.3**	62	69.5	77.6
	Random	6.9	14.2	51.8	16.9	49.2	54.8	70.3
Wedding	Auto	55.4	94.4	83.9	74.7	67	88	85.7
	DR-DSN	4.2	8.9	40.7	14.4	**76.6**	62	**88.4**
	VASNET	4.5	14.4	46.5	22	44	52.7	84.9
	vsLSTM	**9**	**17.3**	50.2	29.5	50.1	56.9	80.7
	Ours	**9.4**	**17.9**	**52.8**	**30.3**	**51.8**	**58.6**	82.8
	Random	3.5	10	41.1	16.3	40.6	51.6	80

Table 4: Comparison of different techniques on VISIOCITY. TechTalk videos do not have MegaEvents

and diversity one at a time, and the fourth one learnt using a combined loss function. We found the fourth model to perform the best among these four and we have reported the numbers from this model (VISIOCITY-SUM) in the tables. A quick comparison with other models reveal that almost on all different characteristics, this new model manages to do well and better than all other techniques.

7 CONCLUSION

Identifying the need of the community, we presented VISIOCITY, a large benchmarking dataset and demonstrated its effectiveness in real world setting. To the best of our knowledge, it is the first of its kind in the scale, diversity and rich concept annotations. We introduce a recipe to automatically create ground truth summaries typically needed by the supervised techniques. We also extensively discuss and demonstrate the issue behind multiple right answers making the evaluation of video summaries a challenging task. Motivated by the fact that different good summaries have different characteristics and are not necessarily better or worse than the other, we propose an evaluation framework better geared at modeling human judgment through a suite of measures than having to overly depend on one measure. Finally through extensive and rigorous experiments we report the strengths and weaknesses of some representative state of the art techniques when tested on this new benchmark. Motivated by a fundamental problem in current supervised approaches, of learning from a single combined ground truth summary and/or learning from a single loss function tailored and optimizing *one* characteristic, our attempt to make simple extension to an existing mixture model technique gives encouraging results. We hope our attempt to address the multiple issues currently surrounding video summarization as highlighted in this work, will help the community advance the state of the art in video summarization.

ACKNOWLEDGEMENTS

REFERENCES

[1] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. 2020. Unsupervised Video Summarization via Attention-Driven Adversarial Learning. In *International Conference on Multimedia Modeling*. Springer, 492–504.

[2] Sandra Eliza Fontes De Avila, Ana Paula Brandão Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. 2011. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters* 32, 1 (2011), 56–68.

[3] David Doermann and David Mihalcik. 2000. Tools and techniques for video performance evaluation. In *icpr*. IEEE, 4167.

[4] Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. 2018. Summarizing Videos with Attention. In *Asian Conference on Computer Vision*. Springer, 39–54.

[5] Cheng-Yang Fu, Joon Lee, Mohit Bansal, and Alexander C Berg. 2017. Video highlight prediction using audience chat reactions. *arXiv preprint arXiv:1707.08559* (2017).

[6] Tsu-Jui Fu, Shao-Heng Tai, and Hwann-Tzong Chen. 2019. Attentive and Adversarial Learning for Video Summarization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1579–1587.

[7] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. 2014. Diverse sequential subset selection for supervised video summarization. In *Advances in Neural Information Processing Systems*. 2069–2077.

[8] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating Summaries from User Videos. In *ECCV*.

[9] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. 2014. Creating summaries from user videos. In *European conference on computer vision*. Springer, 505–520.

[10] Michael Gygli, Helmut Grabner, and Luc Van Gool. 2015. Video summarization by learning submodular mixtures of objectives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3090–3098.

[11] Mei Huang, Ayesh B Mahajan, and Daniel F DeMenthon. 2004. *Automatic performance evaluation for video summarization*. Technical Report. MARYLAND UNIV COLLEGE PARK INST FOR ADVANCED COMPUTER STUDIES.

[12] Zhong Ji, Kailin Xiong, Yanwei Pang, and Xuelong Li. 2019. Video summarization with attention-based encoder-decoder networks. *IEEE Transactions on Circuits and Systems for Video Technology* (2019).

[13] Sivapriyaa Kannappan, Yonghuai Liu, and Bernie Tiddeman. 2019. Human consistency evaluation of static video summaries. *Multimedia Tools and Applications* 78, 9 (2019), 12281–12306.

[14] Vishal Kaushal, Sandeep Subramanian, Suraj Kothawade, Rishabh Iyer, and Ganesh Ramakrishnan. 2019. A Framework towards Domain Specific Video Summarization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 666–675.

[15] Aditya Khosla, Raffay Hamid, Chih-Jen Lin, and Neel Sundaresan. 2013. Large-scale video summarization using web-image priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2698–2705.

[16] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. 2012. Discovering important people and objects for egocentric video summarization. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 1346–1353.

[17] Zhuo Lei, Chao Zhang, Qian Zhang, and Guoping Qiu. 2019. FrameRank: A Text Processing Approach to Video Summarization. *arXiv preprint arXiv:1904.05544* (2019).

[18] Yingbo Li and Bernard Merialdo. 2010. VERT: automatic evaluation of video summaries. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 851–854.

[19] Yandong Li, Liqiang Wang, Tianbao Yang, and Boqing Gong. 2018. How local is the local diversity? Reinforcing sequential determinantal point processes with dynamic ground sets for supervised video summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 151–167.

[20] Zheng Lu and Kristen Grauman. 2013. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2714–2721.

[21] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. 2002. A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*. ACM, 533–542.

[22] Michel Minoux. 1978. Accelerated greedy algorithms for maximizing submodular set functions. In *Optimization Techniques*. Springer, 234–243.

[23] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila. 2019. Rethinking the Evaluation of Video Summaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7596–7604.

[24] Rameswar Panda, Niluthpol Chowdhury Mithun, and Amit K Roy-Chowdhury. 2017. Diversity-aware multi-video summarization. *IEEE Transactions on Image Processing* 26, 10 (2017), 4712–4724.

[25] Bryan A Plummer, Matthew Brown, and Svetlana Lazebnik. 2017. Enhancing video summarization via vision-language embedding. In *Computer Vision and Pattern Recognition*, Vol. 2.

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

[27] Aidean Sharghi, Ali Borji, Chengtao Li, Tianbao Yang, and Boqing Gong. 2018. Improving sequential determinantal point processes for supervised video summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 517–533.

[28] Aidean Sharghi, Jacob S Laurel, and Boqing Gong. 2017. Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2127–2136.

[29] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. TVSum: Summarizing web videos using titles.. In *CVPR*. IEEE Computer Society, 5179–5187. http://dblp.uni-trier.de/db/conf/cvpr/cvpr2015.html#SongVSJ15

[30] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. 2015. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5179–5187.

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

[32] Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool. 2017. Query-adaptive video summarization via quality-aware relevance estimation. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 582–590.

[33] Shuwen Xiao, Zhou Zhao, Zijian Zhang, Xiaohui Yan, and Min Yang. 2020. Convolutional Hierarchical Attention Network for Query-Focused Video Summarization. *arXiv preprint arXiv:2002.03740* (2020).

[34] Bo Xiong, Yannis Kalantidis, Deepti Ghadiyaram, and Kristen Grauman. 2019. Less is More: Learning Highlight Detection from Video Duration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1258–1267.

[35] Serena Yeung, Alireza Fathi, and Li Fei-Fei. 2014. Videoset: Video summary evaluation through text. *arXiv preprint arXiv:1406.5824* (2014).

[36] Li Yuan, Francis EH Tay, Ping Li, Li Zhou, and Jiashi Feng. 2019. Cycle-SUM: Cycle-consistent Adversarial LSTM Networks for Unsupervised Video Summarization. *arXiv preprint arXiv:1904.08265* (2019).

[37] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Summary transfer: Exemplar-based subset selection for video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1059–1067.

[38] Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. 2016. Video summarization with long short-term memory. In *European Conference on Computer Vision*. Springer, 766–782.

[39] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.

[40] Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2018. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Thirty-Second AAAI Conference on Artificial Intelligence*.