AudViSum: Self-Supervised Deep Reinforcement Learning for Diverse Audio-Visual Summary Generation

Sanjoy Chowdhury*[1]
schowdhury671@gmail.com
Aditya P. Patra*[2]
aditya.prakash.patra1997@gmail.com
Subhrajyoti Dasgupta[3]
subhrajyotidg@gmail.com
Ujjwal Bhattacharya[3]
ujjwal@isical.ac.in

[1] ShareChat, Bangalore, India
[2] Indian Institute of Technology, Patna, India
[3] Indian Statistical Institute, Kolkata, India
*indicates equal contributions

Abstract

A brief yet comprehensive summary of a lengthy video helps us understand the key insights about it. Video summarization aims to generate a 'video-thumbnail' from a given input video. Although the field has been widely studied in the literature, to the best of our knowledge, all the existing works in this area have majorly emphasized on visual modality only, although its audio component may carry crucial information for efficient video summarization. To this end, we introduce a novel self-supervised audio-visual summarization network *AudViSum*, that leverages both audio and visual information and employs Deep Reinforcement Learning to reward the model to generate diverse yet semantically meaningful summaries. Our experiments establish the fact that combining audio-visual information helps to generate realistic summaries from relatively lengthy input videos. To ensure diverse summary generation we report the top-3 summaries for each video. Since there is no publicly available annotation to evaluate audio-visual summaries, we annotate the TVSum & OVP datasets comprising 50 videos each. Experimental results indicate that *AudViSum* achieves promising performance in the audio-visual summary generation task when compared against human annotations.

1 Introduction

According to a recent study, more than 500 million hours of videos are being watched on YouTube alone every single day. We could refer to innumerable such studies to establish the fact that video contents are the new means of information sharing. A video is typically multi-dimensional. It is not just a sequential grouping of some frames but also a combination of audio, motion and time series dimensions. To make this enormously large volume of video data easily browsable, the requirement of an automatic summarization tool is highly imperative. Although, considerable amount of work on video summarization has been done

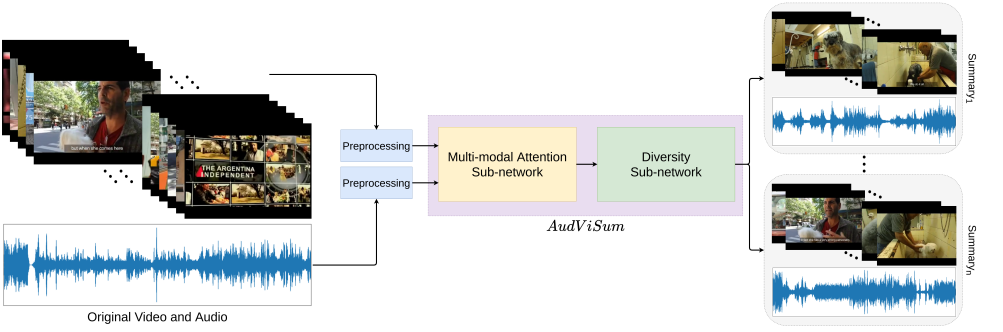


Figure 1: An overview of the audio-visual summarization task and the proposed architecture. *AudViSum* generates *n* diverse and concise summaries from the input video.

over the years [14, 27, 33, 39], surprisingly enough, to the best of our knowledge, none of them have claimed to emphasize equally on the audio content (as visual) in order to obtain a holistic abridged version of a lengthy video. In this study, we have focused our attention on the audio-visual summarization problem i.e. we attempt to produce a concise summary of a video by examining both its visual as well as auditory information. Here audio is not merely used as an auxiliary source of information but also plays a pivotal role in determining which shots should be selected in the summary videos.

In the literature of video summarization, supervised, unsupervised, and weakly supervised methods could be found. The supervised strategy [14, 16, 17, 50] requires a significant amount of annotated data thereby limiting its practical applicability. In this scheme of things every video sample is associated with its ground truth summary. Networks learn from these cues and generally tend to outperform its unsupervised or weakly supervised counterparts. Unsupervised methods [33, 39, 45, 58] on the other hand, tries to minimize the distance between training videos and a distribution of their summaries. They propose to follow some heuristic based steps to rank and select key frames from input video streams. Zhang *et al.* [51] attempt to use transfer learning for unsupervised domain adaptation. Some works [3, 37] provide weak supervision through supplementary web image or video cues.

In this paper, we propose *AudViSum*, a novel self-supervised deep reinforcement learning method for audio-visual summarization. After some initial pre-processing for quantization of the audio and image frames, the respective sequences are passed through Bi-directional LSTM [41] to capture the sequential flow of both the streams. Thereafter, we apply self-attention to generate potential summaries for the input video. To make sure the generated summaries are diverse enough, we then pass them through the Diversity Sub-network to finally produce 3 short summaries chosen from top as many models based on training reward, for each video. The duration of the output synopses are roughly 15% of the input video.

To summarise, our novel contributions are *five folds*: (1) We are the first to propose a holistic deep learning based audio-visual summarization method by utilizing the semantic interplay between audio as well as visual information from a video. We intend to propose a generic framework compared to prior task specific methods. (2) To the best of our knowledge, this is the first reported work to introduce self-supervision in the context of audio-visual summarization. (3) The task of generating highlights by nature is subjective, staying true to this we propose a novel reward scheme to produce 3 short summaries corresponding to each input video, we show that this method is more efficient to gauge the diversity of the model

and produces more realistic summaries. (4) Since there is no publicly available annotation to evaluate audio-visual summaries, we annotate the TVSum and OVP[1] datasets to facilitate this task. We will release our annotation consisting of 100 video summaries to help benchmark future research in the community[2]. (5) Lastly, although not directly comparable, our method outperforms SOTA visual-only summarization works by considerable margin on quantitative analysis.

2 Related Works

Video Summarization [17, 22, 23, 31, 32, 33, 50, 52, 54] has been extensively studied in the recent past. Works following a supervised learning strategy leverage the availability of ground truth data to guide the network training. Proposals to summarise videos by understanding the temporal dependency in the video frames were designed by [39, 51, 55]. Sequential determinantal point processes (SeqDPPs) model is used by [28, 43] to increase the diversity in the generated summary. On the other hand, [53] proposes a method that takes user queries into consideration to learn more user-oriented summaries. However, these works are heavily dependent on human-annotations. Methods involving unsupervised learning strategies also have been explored in the literature. Adversarial approaches to reconstruct videos using Auto-encoders and GANs were studied by [19, 24, 33, 49]. Rochan *et al.* proposed a method to learn video summarization from unpaired data, thus, alleviating the limitations of ground truth requirement [38]. A few investigations involving weakly-supervised learning strategies [8, 21, 37] have also been presented. Although these works have brought in substantial improvement to the video summarization task, a noteworthy observation is that harnessing the audio modality efficiently has not yet been studied in the literature.

Reinforcement learning (RL) is the area of machine learning that deals with sequential decision-making [8]. Reinforcement learning has been very widely used in a variety of vision tasks namely image restoration [48], recognition [5], and others [25, 26, 30, 44, 47]. RL has also been previously used in the video summarization task. Zhou *et al.* proposed a reward function that accounts for diversity and representativeness of the generated summaries [58]. Further, [15] uses RL to preserve the spatio-temporal features of the original video in the summary. Recently, Zhao *et al.* proposed a dual task setup (Video-to-summary and Summary-to-video), where the summary generator is rewarded under the assistance of the video reconstructor [56].

Ngiam *et al.* [34] studied multi-modal feature learning and demonstrated the benefits of information sharing and joint learning. PixelPlayer [57] was introduced by Zhao *et al.* that leverages multi-modal audio-visual learning to locate regions in an image that are producing sounds and perform sound-source separation from each pixel. This task gained other significant contributions from [2, 7, 40, 42]. Gao *et al.* proposed a novel approach for action recognition using multi-modal learning where audio is used as a preview mechanism to eliminate both short-term and long-term visual redundancies [13]. Multi-modal reinforcement learning has been studied well for navigation by exploiting audio-visual information [4, 12]. In the literature, we also find explorations using self-supervision to learn from audio-visual modalities [1, 9, 10, 11]. To the best of our knowledge, we are the first to propose a non task-specific, deep learning based approach which performs a holistic audio-visual summarization by exploiting the semantic interplay between audio and visual streams.

[1] Open video project: https://open-video.org/
[2] https://github.com/schowdhury671/AudViSum

3 Proposed Methodology

Our novel multi-modal audio-visual summarization network consists of two sub networks: Multi-modal Attention Sub-network and Diversity Sub-network. Inspired by the fact that an ideal summary video should be both diverse as well as semantically similar to the relatively lengthy input video, we formulate the task of summary generation as a sequential decision making process. As opposed to some prior works [29, 33, 38] who tend to assign frame level importance scores, for efficient encapsulation of contextual information we first quantize the visual and the audio streams into small chunks and use these shots for further processing. The algorithm is described in section 3.1. We use Bidirectional LSTM [41] based feature extractor on both the visual and audio streams parallelly followed by self-attention blocks [46]. We find that late fusion fared better compared to early fusion mechanisms while combining multi-modal features. In the next step, these high dimensional fused features are passed into the Diversity Sub-network, elaborated in section 3.3. Finally, we introduce a self-supervised Deep RL paradigm to ensure that all the generated short highlights are intrinsically diverse but semantically coherent with the input video. Fig.1 demonstrates the end-to-end process where *AudViSum* takes in pre-processed videos and produces multiple summaries.

3.1 Audio-Visual Preprocessing

In this quantization or pre-processing step the visual and the audio streams are separated and construed as small chunks. The extracted visual frames are passed through Resnet-50 [18] to obtain flattened feature embeddings of 1000 dimensions. The average of the feature vectors corresponding to every 3 second interval is taken and used as a representative for that shot. Audio stream processing is particularly interesting here. We split the entire audio signal into subsets of non silent intervals and remove the ones where audio amplitude is at least 4 decibel lesser than the mean amplitude. We empirically found this threshold to be suitable for removing the silent intervals. Issues pertaining to breathing gaps might still persist in which case we follow a heuristic to make the groupings for effective audio processing. Details about the heuristic is presented in the **supplementary material**.

Subsequently, towards our goal of dividing the audio stream into equal-sized chunks we break the sequence into meaningful intervals (pad wherever necessary). Again, on empirical analysis we observed that an interval of 3 seconds is optimal to capture audio-visual coherence. A pre-trained VGGish [20] network is used for audio feature extraction. This model takes mini-batches of 128 inputs, so, for a 3 second audio the output is a 384 dimensional feature vector. The intuition behind the heuristic based quantization algorithm is if the gap between two consecutive segments is less than 1.5 seconds we consider it to be a breathing gap else we process them independently.

3.2 Multi-modal Attention Sub-network

In this section we elaborate on our novel Multi-modal Attention Sub-network (MASN). It is the first part of the two-step pipeline. An overview of the network is provided in Fig. 2. Both the audio and visual frame sequences are passed through bidirectional LSTM layers separately to generate $L \times F_a$ and $L \times F_v$ dimensional feature maps respectively. As we are dealing with sequential data, the use of a bidirectional framework is very intuitive and it has also shown promising performance in related problems. At this stage self-attention is applied to both the modalities individually. Self-attention [46] is used to compute a representation

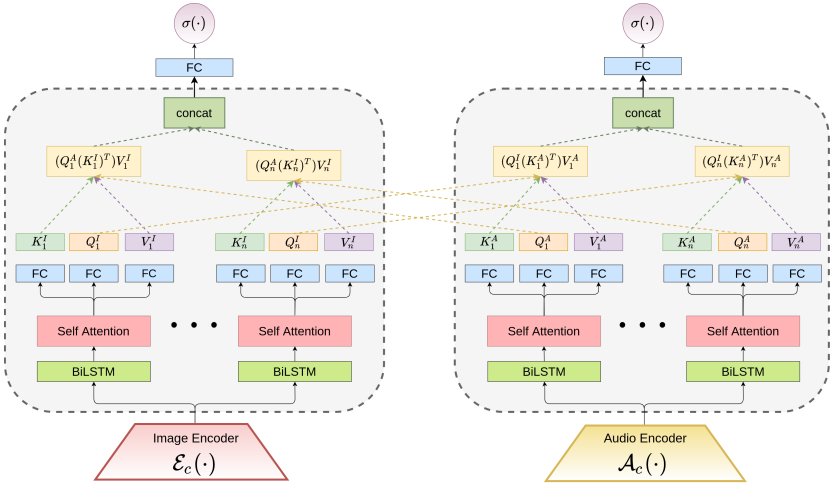


Figure 2: The Multi-modal Attention Sub-network helps to understand the intricate inter-dependencies between the visual and audio features, and finally computes the probability scores to include the best shots in the summary.

of the sequence by relating to its different parts. In order to attend to distinct positions of the multi-modal sequences, we first perform self-attention thereby finding intrinsic nuances of the individual modalities followed by their combination which results in much sound understanding of their inter dependencies compared to when we apply early fusion.

Let $h_i^A \in \mathbb{R}^{\mathrm{L}\times\mathrm{F_a}}$ and $h_i^I \in \mathbb{R}^{\mathrm{L}\times\mathrm{F_v}}$ be the audio and visual features extracted by the Bi-LSTM layers respectively, where $i = \{1,...,n\}$ and n denotes number of summary models. The self-attention layer is computed in accordance with [46], to obtain a representation of the relationship amongst the features. Subsequently, 3 fully-connected networks are used to compute query, key and value triplets (Q_i^I, K_i^I, V_i^I) and (Q_i^A, K_i^A, V_i^A) for the i^{th} model. Next, as shown in Fig. 2, to capture the multi-modal cues between the visual and audio channels, we calculate the attention as:

$$I_{att}(Q_i^A, K_i^I, V_i^I) = (Q_i^A(K_i^I)^T)V_i^I \tag{1}$$

$$A_{att}(Q_i^I, K_i^A, V_i^A) = (Q_i^I(K_i^A)^T)V_i^A \tag{2}$$

These features from the corresponding sub-networks are then concatenated and passed through fully-connected network to obtain an aggregated feature representation:

$$f(I_{att}) = W_f[(Q_1^A(K_1^I)^T)V_1^I, ..., (Q_n^A(K_n^I)^T)V_n^I] + b_f \tag{3}$$

$$f(A_{att}) = W_f[(Q_1^I(K_1^A)^T)V_1^A, ..., (Q_n^I(K_n^A)^T)V_n^A] + b_f \tag{4}$$

Finally, probability score for each of the z-frames are computed using a sigmoid function over the feature representation as:

$$p_t = \sigma(W f_z) \tag{5}$$

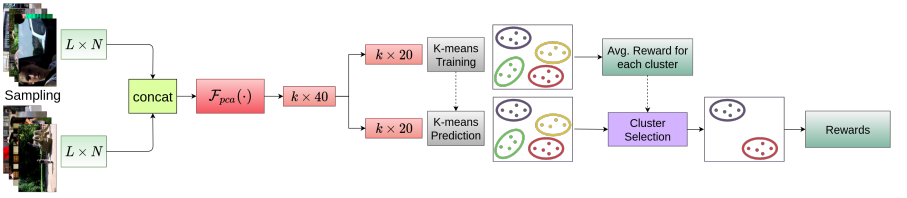


Figure 3: The Diversity Sub-network allows us to bring sufficient uncorrelatedness between different summaries.

3.3 Diversity Sub-network

Next, the Diversity Sub-network (DSN) feeds on the concatenated audio-visual features after they are passed through a FC (fully connected) layer. The role of this sub-network is to incorporate diversity in the generated summaries. We apply sigmoid function after the FC layer to indicate the probability of a video-shot to be selected in the final summary. Bernoulli sampling is applied to select video shots $a_t \sim$ Bernoulli (p_t), where $\{p_t\}_{t=1}^{T}$ represents shot wise importance score, and $a_t \in \{0,1\}$ represents the selection of the t^{th} video shot in the final summary. As K-means typically does not work well on binary data, we apply Principal Component Analysis (PCA) to calculate the eigen vectors. We multiply this truncated eigen vector matrix with the obtained samples to produce a non binary representation which is more suitable in our case for efficient processing.

K-means is applied on the first model $(\mathcal{M}_1)$ to form $\mathcal{K}$ clusters among which the generated summaries are distributed. At this stage, an average reward is calculated for each cluster. For the subsequent models, we use these $\mathcal{K}$ representatives for cluster assignment. For the $(j+1)$-th model we remove the top $j*\ \mathcal{K}'(<\mathcal{K})$ clusters, where $j = \{0,...,4\}$ and use the remaining clusters to calculate the reward for that particular model. Thus, once a model has generated a short summary video we want to restrict other models from generating similar summaries for that input video anymore. Fig. 3 represents how diverse yet meaningful summaries are obtained through proper reward assignment. We discuss different reward functions and how we propose to use it under unsupervised and self-supervised settings in more details under section 3.4.

3.4 Reward Functions

Following Reinforcement Learning (RL) strategy, the *AudViSum* network will receive a reward $\mathcal{R}(\mathcal{S})$ to evaluate the quality of the generated summaries. The better the summaries are w.r.t to human annotation, higher the reward is. So, we are to maximize the expected reward over time. Ideally a good synopsis of a lengthy video sequence should be both diverse as well as a good representative i.e. it should capture the important moments and should be concise at the same time. In the following sections we will discuss about the proposed weighted representative $\mathcal{R}_{rep}$ and diversity $\mathcal{R}_{div}$ rewards under unsupervised and self-supervised settings and compare their performance.

3.4.1 Unsupervised Reward

Since there is no prior work that we can directly compare with, we define our own baseline. To this end, we design and experiment with two unsupervised schemes:

Baseline Zhou *et al.* [] proposed an unsupervised diversity-representativeness reward to guide the RL agent for video summarization. Taking cue from them we introduce a weighted reward function. In this, $\mathcal{R}_{rep}$ measures how good a representative the summary video is and $\mathcal{R}_{div}$ stands for the degree of diversity in the generated summaries. These two rewards could be expressed as:

$$\mathcal{R}_{div} = \frac{1}{\mathcal{X}(\mathcal{X}-1)} \sum_{t \in \mathcal{X}} \sum_{\substack{t' \in \mathcal{X} \\ t' \neq t}} d(s_t, s_{t'}) \tag{6}$$

$$d(s_t, s_{t'}) = 1 - \frac{s_t^T s_{t'}}{\|s_t\|_2 \|s_{t'}\|_2} \tag{7}$$

$$\mathcal{R}_{rep} = \exp\left(-\frac{1}{T} \sum_{t=1}^{T} \min_{t' \in \mathcal{X}} \|s_t - s_{t'}\|_2\right) \tag{8}$$

where s_t stands for t^{th} shot level feature representation, $d(\cdot,\cdot)$ is the measure of temporal distance, and $\mathcal{X} = \{x_i \,|a_{x_i} = 1, i = 1, \ldots, |\,\mathcal{X}\,|\}$ are the indices of the selected frames.

Reward tends to get higher with increasing summary length. In order to eliminate this, we propose an experimentally found weighted reward scheme. While the diversity reward $\mathcal{R}_{div}$ is weighted by a factor of 1.5, the representative reward $\mathcal{R}_{rep}$ is weighted by 0.2. By assigning less weightage to representative reward we force the agent to generate semantically diverse summaries over time. Hence the overall reward is:

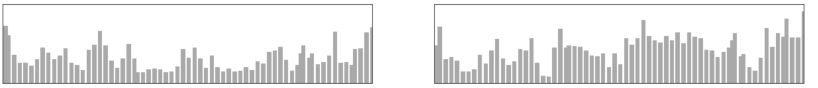
$$\mathcal{R}(\mathcal{S}) = 1.5 * \mathcal{R}_{div} + 0.2 * \mathcal{R}_{rep} \tag{9}$$

Teacher-Student Model Knowledge distillation is a very prominent way of teaching a part of the network to align according to some specific task. In this Teacher-Student (TS) setup we follow a sequential knowledge transfer strategy where the first model ($\mathcal{M}_1$) acts as a teacher to the second model ($\mathcal{M}_2$) which in turn acts as a teacher to $\mathcal{M}_3$, and so on. The feedback received from the 'teacher' model is used to eliminate the top $\mathcal{K}'$ cluster samples thereby ensuring that the similar summaries are penalised by reinforcing summaries with diverse semantics. This TS benchmark model gains significant improvements over the baseline version as shown in Table 1.

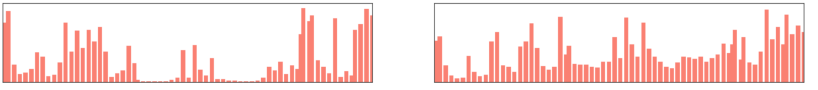
3.4.2 Self-supervised Reward

Self-supervision plays an important role in the current landscape of the computer vision community. The scarcity of labelled data at a large scale is one of the primary reasons for the prominence of this particular discipline of research. Although we could see a considerable amount of supervised, weakly-supervised video summarization techniques being proposed lately, our work is the first to learn video summarization in a self-supervised manner. Thereby achieving two objectives (i) the model could be trained without the aid of any explicitly annotated data (ii) it is possible to produce diverse yet highly meaningful summaries by exploiting the interplay of multi-modal information without any human intervention.

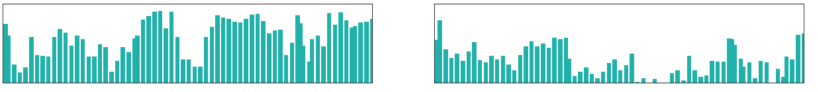
Contrastive Loss Following Eq.1 and 2 we perform multi-modal attention by applying late fusion. Through self-supervision, we want the network's visual attention to closely align with its audio counterpart. Let x^I and x^A represent the image and audio features respectively, we implement contrastive loss (Eq.10) so that the distribution gap between corresponding



a. Importance scores of shots at different timestamps predicted by AudViSum using weighted diversity and representative loss



b. Importance scores of shots at different timestamps predicted by AudViSum using teacher-student learning strategy



c. Importance scores of shots at different timestamps predicted by AudViSum using contrastive loss

Figure 4: Graphical representation of the shot-level importance scores as predicted by different variants of *AudViSum* for top-2 summaries on a sample video(3eYKfiOEJNs) from the TVSum dataset.

audio and visual modalities is reduced while the same between two different summaries are maximised as much as possible:

$$\mathcal{L}_N = -E_X \left[\log \frac{\exp\left(f(x^I)^T g\left(x^A\right)\right)}{\exp\left(f(x^I)^T g\left(x^A\right)\right) + \sum_{j=1}^{n-1} \exp\left(f(x^I)^T g\left(x_j^A\right)\right)} \right] \tag{10}$$

$$\mathcal{R}(\mathcal{S}) = 1.5 * \mathcal{R}_{div} + 0.2 * \mathcal{R}_{rep} + 0.5 * \mathcal{L}_N \tag{11}$$

We use InfoNCE [35] loss for this purpose, this is intuitive also as we want the summaries to attend to multi-modal information and also generate diverse summaries at the same time. Eq. 11 represents the overall reward under the self-supervised setting. Fig. 4 compares the relative shot-level importance scores assigned by three different variants of *AudViSum*.

4 Experiments

4.1 Dataset

Annotation Since none of the existing annotations considered audio information while producing summaries, we annotated two very popular datasets TVSum [45] and OVP to satisfy the demands of the problem and subsequently carry out the evaluation process efficiently. We chose to annotate these two datasets due to their meaningful audio contents along with the visuals. TVSum contains 50 video samples spanning over 10 different categories from the TRECVid MED dataset involving news, how-to's, documentaries, etc. Here the video duration ranges between 2 to 11 minutes. OVP comprises 50 relatively shorter videos (1 to 4 minutes) of documentaries, educational, historical, lectures among others.

In this work, as we have focused on audio-visual summary generation, the existing annotations were deemed infeasible. To this end, we define our own annotation rules and annotate the two above mentioned datasets as follows: we empirically found that 3 second shots are appropriate for capturing local contextual information with proper audio-visual coherence. To facilitate a holistic summary generation we record 3 different scores for each of these uniform length shots over the entire video. We collect a total of 500 annotations (10 per video).

Figure 5: Visual results of original video(4wU_LUjG5Ic) from TVSum dataset followed by summaries generated by *Contrastive* model and *Teacher* − *Student* model respectively.

To ensure a fair and unbiased scoring scheme, in half of those cases we make the annotators provide the visual-only scores first followed by audio-only scores and in the other half of the cases this order is reversed i.e. audio-only scores followed by the visual-only scores. In either case the last step is to provide aggregated shot-level audio-visual scores. We purposefully follow an intuitive three-pass annotation strategy to capture the actual essence of the video sequence, thereby making sure that if there is any unique yet important shot present, then that is deservedly included in the video synopsis. Finally for each shot we take a weighted average of these three scores (weights assigned to visual-only, audio-only, and audio-visual annotation are 25%, 25%, and 50% respectively). The scoring is done in a range of 1 (not important) to 5 (very important). The final summary annotation is selected by choosing top 15% shots after averaging out the scores provided by the 10 annotators. Details about the annotation statistics and scores are provided in the **supplementary material**.

4.2 Evaluation and Ablation

Following Otani *et al.* [36] we use Kendall's τ and Spearman's ρ correlation coefficients to measure the association between the predicted summaries and GT annotations. As we generate three distinct summaries for each video, we report the best of three scores for the evaluation purpose. The objective is to maximize the expected reward:

$$J(\theta) = E_{p_\theta(a_{1:T})}[\mathcal{R}(\mathcal{S})] \tag{12}$$

for the summary generator agent where the policy function is represented by π_θ with parameters θ. Here, $p_\theta(a_{1:T})$ denotes the probability distributions over possible action sequences and $\mathcal{R}(\mathcal{S})$ is the weighted reward function. Table 1 illustrates the performances of *AudViSum* under different settings. Baseline and *TS* models are compared against 4 variants of the Self-supervised model (Contrastive). To justify the importance of a holistic audio-visual study we report the performance when only visual information is used to produce the summaries. Observe that models which use multi-modal data significantly outperform the setup where purely visual stream is used to generate the summaries. It should also be noted that on removal of the DSN module (which incorporates diversity) or self-attention from the MASN module (which ensures effective multi-modal semantic interplay) the performance drops considerably. The complete model with self-supervised contrastive reward achieves the best results. Although there is no prior method that we can directly compare our work with due to the scope of the problem under consideration, in order to contextualize our contributions we perform a quantitative comparison to illustrate the importance of this study. For fair assessment we report the scores obtained by the SOTA models when evaluated against visual-only annotations whereas our model is evaluated against the audio-visual

Table 1: Ablation Study showing comparison of different variants of *AudViSum*.

Method	TVSum		OVP	
	τ	ρ	τ	ρ
Baseline	0.092	0.131	0.083	0.124
Teacher-Student	0.095	0.138	0.087	0.132
Visual only	0.044	0.059	0.035	0.048
w/o DSN	0.090	0.130	0.083	0.121
MASN w/o self-attn	0.087	0.128	0.082	0.120
Contrastive[proposed]	**0.101**	**0.146**	**0.094**	**0.141**

Table 2: Quantitative Comparison with increasing count of *Contrastive* summaries.

Top-Y summaries	TVSum		OVP	
	τ	ρ	τ	ρ
1	0.092	0.138	0.083	0.128
2	0.095	0.143	0.087	0.135
3	<u>0.101</u>	<u>0.151</u>	<u>0.094</u>	<u>0.141</u>
4	0.101	0.152	0.095	0.143
5	0.102	0.152	0.095	0.144

Table 3: Quantitative Comparison between *AudViSum* and a few existing SOTA architectures on our annotation of TVSum dataset.

Method	Annotation Used		TVSum	
	Visual only	Audio-visual	τ	ρ
DR-DSN [58]	✓	–	0.023	0.030
dppLSTM [51]	✓	–	0.040	0.054
Hierarchical RL [6]	✓	–	0.079	0.114
Li *et al.* [29]	✓	–	0.091	0.118
***AudViSum*$_{\text{contrastive}}$[ours]**	–	✓	**0.101**	**0.146**

annotation. As seen, the inclusion of audio information while producing summaries is beneficial as our best model tends to outperform other SOTA visual-only summarization methods by considerable margin. Fig. 5 draws comparison between the best summaries generated by the Contrastive and TS variants respectively for the same video. Table 2 compares the performance when summaries generated from top-Y models in a self-supervised setting are considered for evaluation. The models are chosen by comparing the corresponding rewards obtained by them individually during training. For benchmarking, we compare summaries from these Y models against human annotation and report the best scores. Observe that there is a significant gain by opting to choose till 3 top summaries. On selecting more than 3 summaries there is no considerable improvement implying the best summary is most likely to be within top-3 ones. The self-supervised model tends to highly correlate with the human annotations. This could be attributed to the efficient multi-modal information mining and semantically meaningful reward function.

5 Conclusions

In this study, we present our recent exploration on multi-modal audio-visual summarization. Prior SOTA works were massively focused on visual-only summary generation techniques. However, our novel *AudViSum* network, to the best of our knowledge, is the first reported work that leverages the self-supervision scheme and produces holistic audio-visual summaries. We intend to propose a generic framework in contrast to prior task-specific methods or approaches that consider audio as an auxiliary source of information. To this end, we report three distinct summaries for each input video to ensure that generated summaries are diverse among themselves. We will release the code and our summary annotations on TVSum and OVP datasets to encourage further research in this area. We believe our research will open avenues to a wide range of applications in relevant areas like automated summary generation, video browsing, monitoring systems among others.

References

[1] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. Self-supervised learning of audio-visual objects from video. In *Proceedings of the 16th European Conference on Computer Vision (ECCV), Part XVIII*, pages 208–224. Springer, 2020.

[2] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 435–451, 2018.

[3] Sijia Cai, Wangmeng Zuo, Larry S Davis, and Lei Zhang. Weakly-supervised video summarization using variational encoder-decoder and web prior. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 184–200, 2018.

[4] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Soundspaces: Audio-visual navigation in 3D environments. In *Proceedings of the 16th European Conference on Computer Vision (ECCV), Part VI*, pages 17–36, 2020.

[5] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin. Recurrent attentional reinforcement learning for multi-label image recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[6] Yiyan Chen, Li Tao, Xueting Wang, and Toshihiko Yamasaki. Weakly supervised video summarization by hierarchical reinforcement learning. In *Proceedings of the ACM Multimedia Asia*, pages 1–6. 2019.

[7] Sanjoy Chowdhury, Subhrajyoti Dasgupta, Sudip Das, and Ujjwal Bhattacharya. Listen to the pixels. In *IEEE International Conference on Image Processing (ICIP)*, pages 2568–2572, 2021.

[8] Vincent François-Lavet, Peter Henderson, Riashat Islam, Marc G Bellemare, and Joelle Pineau. An introduction to deep reinforcement learning. *arXiv preprint arXiv:1811.12560*, 2018.

[9] Chuang Gan, Hang Zhao, Peihao Chen, David Cox, and Antonio Torralba. Self-supervised moving vehicle tracking with stereo sound. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7053–7062, 2019.

[10] Chuang Gan, Deng Huang, Peihao Chen, Joshua B Tenenbaum, and Antonio Torralba. Foley music: Learning to generate music from videos. In *Proceedings of the 16th European Conference on Computer Vision (ECCV), Part XI*, pages 758–775. Springer, 2020.

[11] Chuang Gan, Deng Huang, Hang Zhao, Joshua B Tenenbaum, and Antonio Torralba. Music gesture for visual sound separation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10478–10487, 2020.

[12] Chuang Gan, Yiwei Zhang, Jiajun Wu, Boqing Gong, and Joshua B Tenenbaum. Look, listen, and act: Towards audio-visual embodied navigation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 9701–9707, 2020.

[13] Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10457–10467, 2020.

[14] Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse sequential subset selection for supervised video summarization. *Advances in Neural Information Processing Systems*, 27:2069–2077, 2014.

[15] N Gonuguntla, B Mandal, NB Puhan, et al. Enhanced deep video summarization network. *British Machine Vision Conference (BMVC)*, 2019.

[16] Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European Conference on Computer Vision (ECCV)*, pages 505–520. Springer, 2014.

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

[18] 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*, pages 770–778, 2016.

[19] Xufeng He, Yang Hua, Tao Song, Zongpu Zhang, Zhengui Xue, Ruhui Ma, Neil Robertson, and Haibing Guan. Unsupervised video summarization with attentive conditional generative adversarial networks. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 2296–2304, 2019.

[20] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, et al. CNN architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017.

[21] Hsuan-I Ho, Wei-Chen Chiu, and Yu-Chiang Frank Wang. Summarizing first-person videos from third persons' points of view. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 70–85, 2018.

[22] Cheng Huang and Hongmei Wang. A novel key-frames selection framework for comprehensive video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(2):577–589, 2019.

[23] Zhong Ji, Fang Jiao, Yanwei Pang, and Ling Shao. Deep attentive and semantic preserving video summarization. *Neurocomputing*, 405:200–207, 2020.

[24] Yunjae Jung, Donghyeon Cho, Dahun Kim, Sanghyun Woo, and In So Kweon. Discriminative feature learning for unsupervised video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8537–8544, 2019.

[25] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, pages 651–673. PMLR, 2018.

[26] Katie Kang, Suneel Belkhale, Gregory Kahn, Pieter Abbeel, and Sergey Levine. Generalization through simulation: Integrating simulated and real data into deep reinforcement learning for vision-based autonomous flight. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 6008–6014. IEEE, 2019.

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

[28] Yandong Li, Liqiang Wang, Tianbao Yang, and Boqing Gong. 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)*, pages 151–167, 2018.

[29] Zutong Li and Lei Yang. Weakly supervised deep reinforcement learning for video summarization with semantically meaningful reward. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pages 3239–3247, 2021.

[30] Xiaodan Liang, Tairui Wang, Luona Yang, and Eric Xing. Cirl: Controllable imitative reinforcement learning for vision-based self-driving. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 584–599, 2018.

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

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

[33] Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 202–211, 2017.

[34] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *ICML*, 2011.

[35] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[36] Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila. Rethinking the evaluation of video summaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7596–7604, 2019.

[37] Rameswar Panda, Abir Das, Ziyan Wu, Jan Ernst, and Amit K Roy-Chowdhury. Weakly supervised summarization of web videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3657–3666, 2017.

[38] Mrigank Rochan and Yang Wang. Video summarization by learning from unpaired data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7902–7911, 2019.

[39] Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 347–363, 2018.

[40] Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. Self-supervised audio-visual co-segmentation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2357–2361. IEEE, 2019.

[41] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[42] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018.

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

[44] Xinhui Song, Ke Chen, Jie Lei, Li Sun, Zhiyuan Wang, Lei Xie, and Mingli Song. Category driven deep recurrent neural network for video summarization. In *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2016.

[45] Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. TVSum: Summarizing web videos using titles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5179–5187, 2015.

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[47] Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 37–53, 2018.

[48] Ke Yu, Chao Dong, Liang Lin, and Chen Change Loy. Crafting a toolchain for image restoration by deep reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2443–2452, 2018.

[49] Li Yuan, Francis EH Tay, Ping Li, Li Zhou, and Jiashi Feng. Cycle-sum: cycle-consistent adversarial lstm networks for unsupervised video summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9143–9150, 2019.

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

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

[52] Ke Zhang, Kristen Grauman, and Fei Sha. Retrospective encoders for video summarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 383–399, 2018.

[53] Yujia Zhang, Michael Kampffmeyer, Xiaodan Liang, Min Tan, and Eric P Xing. Query-conditioned three-player adversarial network for video summarization. *British Machine Vision Conference (BMVC)*, 2018.

[54] Bin Zhao and Eric P Xing. Quasi real-time summarization for consumer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2513–2520, 2014.

[55] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Hsa-rnn: Hierarchical structure-adaptive RNN for video summarization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7405–7414, 2018.

[56] Bin Zhao, Xuelong Li, and Xiaoqiang Lu. Property-constrained dual learning for video summarization. *IEEE Transactions on Neural Networks and Learning Systems*, 31(10): 3989–4000, 2019.

[57] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 570–586, 2018.

[58] Kaiyang Zhou, Yu Qiao, and Tao Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.