UNIVERSIDADE FEDERAL DE MINAS GERAIS
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

Edson Roteia Araujo Junior

An Audiovisual Approach for Video Summarization Using Psychoacoustic Features

Belo Horizonte
2023

Edson Roteia Araujo Junior

An Audiovisual Approach for Video Summarization Using Psychoacoustic Features

Final Version

Thesis presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Erickson Rangel do Nascimento

Belo Horizonte
2023

Araujo Junior, Edson Roteia.

G963c An audiovisual approach for video summarization using psychoacoustic features [recurso eletrônico] / Edson Roteia Araujo Junior – 2023.
1 recurso online (67 f. il, color.): pdf.il.il.

Orientador: Erickson Rangel do Nascimento.
Dissertação (mestrado) - Universidade Federal de Minas Gerais, Instituto de Ciências Exatas, Departamento de Ciência da Computação.
Referências: f. 61-67.

1. Computação – Teses. 2. Sumarização automática de vídeo – Teses. 3. Psicoacústica – Teses. 4. Aprendizagem multimodal – Teses. 5. Semântica I – Teses. I. Nascimento, Erickson Rangel do. II. Universidade Federal de Minas Gerais; Instituto de Ciências Exatas, Departamento de Ciência da Computação. III. Título.

CDU 519.6*84(043)

Ficha catalográfica elaborada pela bibliotecária Belkiz Inez Rezende Costa
CRB 6/1510 Universidade Federal de Minas Gerais - ICEx

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

AN AUDIOVISUAL APPROACH FOR VIDEO SUMMARIZATION USING PSYCHOACOUSTIC FEATURES

EDSON ROTEIA ARAUJO JUNIOR

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

Prof. Erickson Rangel do Nascimento - Orientador
Departamento de Ciência da Computação - UFMG

Prof. Michel Melo da Silva
Departamento de Informática - UFV

Prof. Flávio Luis Cardeal Pádua
Departamento de Computação - CEFET-MG

Belo Horizonte, 28 de fevereiro de 2023.

Documento assinado eletronicamente por **Flávio Luis Cardeal Pádua**, **Usuário Externo**, em 12/05/2023, às 14:35, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Michel Melo da Silva**, **Usuário Externo**, em 15/05/2023, às 09:50, conforme horário oficial de Brasília, com fundamento no art. 5º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufmg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **2299074** e o código CRC **9DA07EC3**.

Aos meus pais, que me apoiaram em cada decisão que tomei e me deram a liberdade de seguir meus sonhos, esta dissertação é dedicada com profunda gratidão e amor.

Acknowledgments

Quero expressar minha profunda gratidão a todas as pessoas que me ajudaram a chegar até aqui. Primeiramente, gostaria de agradecer aos meus pais, cujo amor e apoio inabaláveis me deram força e inspiração para perseguir meus objetivos. Sem a presença deles, nada disso seria possível. Muito obrigado por tudo que vocês fizeram por mim.

Não posso deixar de expressar minha profunda gratidão a meus irmãos, Renata e Robson. Ambos cientistas da computação, eles plantaram em mim a semente da curiosidade e do amor pela ciência desde a infância. Sua influência, tanto profissional quanto pessoal, tem sido um farol na minha vida, me moldando e direcionando. O apoio e a participação deles em todas as etapas da minha vida foram fundamentais na construção da pessoa que sou hoje.

Também quero agradecer à minha namorada, Ana Luiza, por sua presença constante e seu amor incondicional. Você é minha parceira e melhor amiga, e não poderia ter chegado aqui sem o seu apoio. Obrigado por tudo o que você faz por mim e por ser uma parte tão importante da minha vida.

Gostaria de agradecer aos meus amigos que encontrei ao longo da vida, aqueles que me deram risadas, conselhos, apoio e amor. Vocês são minha família escolhida e sempre estiveram ao meu lado, mesmo nos momentos mais difíceis. Agradeço também aos meus alunos de iniciação científica, Gustavo e Luis Gustavo, por seu trabalho árduo e dedicação em contribuir para este projeto. Vocês foram fundamentais para o sucesso deste trabalho e sou grato pela oportunidade de trabalhar com vocês.

Estendo um agradecimento especial ao meu orientador, Erickson Nascimento. Cada reunião com ele era marcada por um ambiente de leveza e inspiração. Ele acreditou fervorosamente no projeto desde a sua concepção, infundindo confiança e otimismo em cada etapa do caminho. Remontando aos meus dias de graduação, Erickson abriu-me inúmeras oportunidades, pavimentando meu caminho no universo acadêmico. Sua orientação e conselhos foram fundamentais para a realização deste trabalho. Agradeço por compartilhar sua sabedoria e experiência, e por me guiar nesta jornada.

Além disso, gostaria de expressar minha sincera gratidão às agências de fomento CNPq, FAPEMIG e CAPES. As bolsas que recebi e os equipamentos de laboratório que utilizei durante toda minha trajetória acadêmica foram adquiridos com recursos provenientes de projetos custeados por essas instituições. Seu apoio financeiro foi indispensável para a realização deste trabalho.

Este trabalho é dedicado a todos vocês - minha família, minha namorada, meus

amigos, meu orientador e ao meus colegas alunos de iniciação científica. Obrigado por me ajudarem a alcançar este marco importante. Vamos continuar lutando juntos para fazer a diferença e tornar este mundo um lugar melhor para todos.

Resumo

A sumarização de vídeo se refere à criação de uma versão resumida de um vídeo mais longo, destacando as partes mais informativas ou engajantes. Esta técnica é útil na área da recuperação de informação multimídia, permitindo que os usuários acessem facilmente informações importantes em grandes coleções de vídeos. Os métodos de sumarização de vídeo, que ajudam os usuários a consumir a crescente quantidade de dados visuais publicados, foram melhorados como resultado do avanço da pesquisa em visão computacional e aprendizado de máquina. Apesar do progresso realizado por *backbones* poderosos e designs de arquiteturas de redes neurais, a maioria dos métodos atuais negligencia as informações multimodais que estão ampla e naturalmente disponíveis na maioria dos cenários, como os sinais audiovisuais presentes em um vídeo. Neste trabalho, apresentamos um novo método baseado em informações audiovisuais para resumir vídeos. Ao contrário da maioria dos métodos atuais, nosso método aproveita as informações multimodais presentes nos vídeos, incluindo os sinais audiovisuais, para melhorar o desempenho da sumarização de vídeo. Nosso modelo incorpora essa informação em uma arquitetura baseada em *transformers* e demonstra uma melhora significativa como resultado. Além disso, propomos uma nova estratégia de treinamento usando pseudo-rótulos gerados a partir de características psicoacústicas do vídeo, o que nos permite alcançar resultados de ponta na configuração não-supervisionada. Por fim, introduzimos um novo *dataset* de sumarização de vídeo e avaliamos o desempenho de nosso método através de uma abordagem de avaliação de *zero-shot*. Nosso método supera as técnicas atuais estado da arte nesse domínio. Avaliamos as contribuições de cada componente do nosso método com estudos de ablação cuidadosos. Nossos experimentos mostram que nosso método é uma base de comparação forte tanto na configuração supervisionada quanto na não-supervisionada, alcançando o melhor desempenho na última com pontuação F1 de 52.6 no conjunto de dados SumMe.

Palavras-chave: Sumarização de Vídeo. Informação Semântica. Psicoacústica. Aprendizagem Multi-modal.

Abstract

Video summarization refers to the creation of a condensed version of a longer video, highlighting the most informative or engaging parts. This technique is useful in the field of multimedia information retrieval, allowing users to easily access important information from large video collections. Video summarization methods, which help users digest the increasing amount of published visual data, have been improved as a result of the advance in computer vision and machine learning research. Despite the remarkable progress that has been made by powerful backbones and clever architectural designs, most of the current methods neglect the multi-modal information that is widely and naturally available in most scenarios, such as the audiovisual signals present in a video. In this thesis, we present a novel method based on audiovisual information to summarize videos. In contrast to most current methods, our method leverages the multi-modal information present in videos, including both audiovisual signals, to improve the performance of video summarization. Our model incorporates this information in a transformer-based architecture and demonstrates significant improvement as a result. Additionally, we propose a new training schema using pseudo-labels generated from the psychoacoustic features of the video, allowing us to achieve state-of-the-art results in the unsupervised setting. Furthermore, we introduce a novel audiovisual video summarization dataset and assess our method's performance on it through a zero-shot evaluation approach. Our method surpasses the current state-of-the-art techniques in this domain. We evaluate the contributions of each of our method's components with thorough ablation studies. Our experiments show that our method is a strong baseline in both supervised and unsupervised settings, achieving the best performance in the latter with an F1 score of 52.6 on the SumMe dataset.

Keywords: Video Summarization. Semantic Information. Psychoacoustics. Multimodal Learning.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Video summarization is the process of generating a short summary of the content of a longer video by selecting and presenting the most informative or interesting materials for potential users. This task is crucial in the domains of computer vision and multimedia information retrieval, as it enables users to quickly access significant information from a vast collection of videos.

Navigating the plethora of videos that are available in today's digital age can be a daunting task. With the proliferation of online platforms for sharing and streaming video and the increasing use of video for a variety of purposes, the amount of published visual data is overwhelming and can be difficult for users to digest. Video summarization, illustrated in Figure 1.1, is an important task that helps to address this problem by condensing large amounts of video into a shorter, more manageable format while still preserving the key information of the original content. This task can be particularly useful for tasks such as video indexing, search, and retrieval, as well as for reducing the time and resources required to view and understand the content of a video. The importance of video summarization is further highlighted by the need for more efficient ways to navigate and make sense of the growing amount of video data.

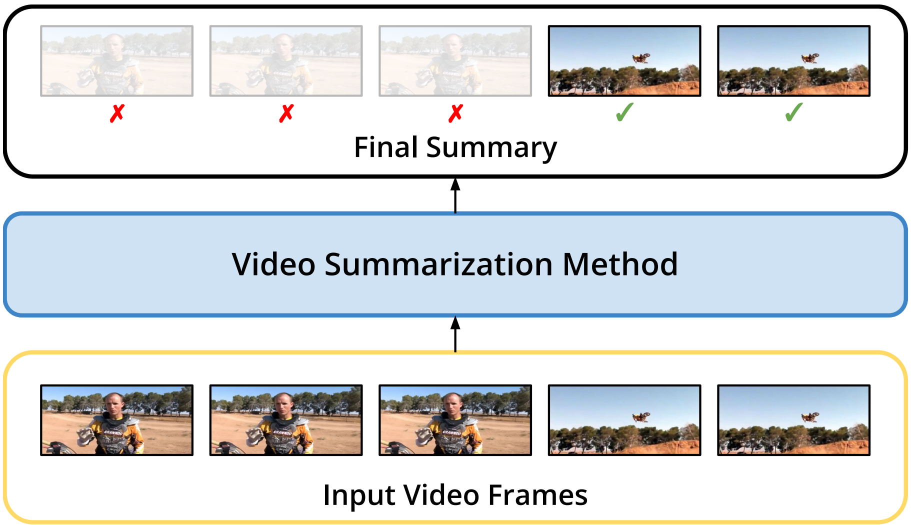
Over the years, there has been growing interest in using machine learning techniques for video summarization, with the aim of automating this process and making it more efficient. However, most existing approaches to video summarization rely on human-annotated labels, which are time-consuming and expensive to obtain. To obtain human-annotated labels for video summarization, typically, a dataset of videos is shown to human evaluators, who then select the most relevant or salient segments or frames of the video to create a summary. This process can be repeated multiple times with different evaluators to ensure that the summaries are representative of the consensus among multiple viewers. The method of obtaining human-annotated labels is both time-consuming and costly, as it requires the participation of multiple evaluators and the manual annotation of each video segment. This high cost has led to the development of unsupervised methods for video summarization, which do not require human-annotated labels and can be trained on large collections of unlabeled videos, thus reducing the cost and effort required to obtain labels.

One of the categorizations of video summarization methods can be defined as static or dynamic. Static video summarization involves generating a summary using still images, while dynamic video summarization involves selecting representative video segments that retain motion information to provide a more efficient and comfortable browsing experience [52]. By carefully selecting informative and representative video segments, a dynamic summary can provide a rich and engaging experience for the viewer. In this thesis, we focus on dynamic video summarization, which has the potential to offer a more comprehensive and immersive understanding of the content of a video.

Despite the abundance of methods for addressing the video summarization problem, many of these approaches fail to take into account the audio information present in the videos. Furthermore, none of the existing methods incorporate psychoacoustic features in their analyses. This lack of consideration for audio information and psychoacoustic features is particularly notable given the importance of both auditory and psychological factors in the human perception and understanding of videos. The inclusion of these elements could potentially lead to a more comprehensive and accurate representation of the video content and, subsequently, more effective and efficient video summarization. The objective of this thesis is to address this gap in the literature by proposing a novel approach for video summarization that incorporates both audio information and psychoacoustic features and evaluating its effectiveness through extensive experiments and analysis.

In this thesis, we propose an audiovisual approach to video summarization that uses psychoacoustic features to generate pseudo-labels for training the model. Psychoacoustics is the branch of science studying the psychological responses associated with sound, and

Figure 1.1: **Summarizing Videos.** Video summarization is the process of selecting frames from an input video to create a condensed and informative summary of the original content.



Source: Created by the author.

it has been shown to be effective in a variety of tasks related to audio processing. By using psychoacoustic features as pseudo-labels, our method avoids the need for human-annotated labels and can be trained in a self-supervised manner.

The investigation of psychological responses induced by sound has been an extensively studied topic [1, 49, 8, 48, 50]. When evaluating human audiovisual consumption, sounds should not only be accounted as a purely mechanical phenomenon but also as a response-inducing perceptual event. Psychoacoustic features are characteristics of sound that are related to how the human auditory system perceives and processes them. They can play an important role in machine learning methods that aim to model the perceived importance of different segments of a video, as they capture the physiological and psychological responses of the human auditory system to auditory stimuli. Auditory annoyance, for example, has both psychological and physiological effects on us. Some psychological effects, as pointed out by Abel *et al.* [1] and Saeki *et al.* [49], include inhibited memory, lengthened reaction times, increased errors in cognition, and selective attention. Physiologically, it leads to effects such as hypertension [8], increased blood pressure [48], and other symptoms of stress [18]. In a recent study by Sammler *et al.* [50], the effects of various musical signals on electroencephalogram (EEG) power and heart rate were examined using EEG recordings and physiological measures of heart rate. The results of the study provide insights into the ways in which different types of musical signals can affect brain and cardiovascular function.

Our method leverages both auditory and visual information from the input video to generate the frames' importance scores. These scores are used to select the frames that are included in the summarized video. The backbone of our approach is a transformer encoder architecture, which has proven to be highly effective for a variety of natural language processing tasks, including machine translation and sentiment analysis. In our case, the transformer encoder is trained to capture the relationships between the audio and visual information in the video and to use this information to predict the importance of each frame. Our results show that this approach is able to outperform or strongly compete against other unsupervised methods of video summarization, demonstrating the effectiveness of this architecture for this particular task. Furthermore, our approach is highly flexible and can be easily adapted to incorporate additional information, such as audiovisual attention mechanisms, that may further improve the performance of the summarization process.

1.1 Objectives and Contributions

This thesis presents a new method for video summarization that utilizes both audio and visual information. While previous approaches have made significant advancements in the field, they often overlook the availability of multi-modal information in most scenarios. Our method incorporates both auditory and visual modalities, leading to improved performance, as demonstrated in our experiments. Furthermore, we propose a training schema that utilizes pseudo-labels generated from the psychoacoustic features of the video, leading to state-of-the-art results in the unsupervised setting. Our model outperforms the state-of-the-art methods in the unsupervised setting on the SumMe dataset while demonstrating strong performance in both supervised and unsupervised settings for the SumMe and TVSum datasets.

The main contributions of this thesis can be summarized as follows:

1. We introduce a novel video summarization method that leverages audiovisual information to generate summaries;

2. We propose the use of psychoacoustic features as pseudo-labels to enable self-supervised training of video summarization methods;

3. We create a new audiovisual summarization dataset that is suitable for evaluating audiovisual video summarization models;

4. We conduct extensive experiments and ablation studies to demonstrate the effectiveness of our method and the challenges of our dataset.

To further advance research in this field and facilitate the reproducibility of our results, we are making our code and network weights publicly available.

1.2 Document Structure

The structure of this thesis is as follows: ***i)*** Chapter 2 provides an introduction and background information on video summarization and psychoacoustics, while Chapter 3 discusses related works in the field and their approaches, results, and contributions. ***ii)*** Chapter 4 presents our proposed approach to address the video summarization problem, including details on our proposed pseudo-label extraction and our audiovisual video summarization model. ***iii)*** Chapter 5 introduces our novel multi-modal dataset and explains the pipeline used to collect and prepare the data. ***iv)*** Chapter 6 presents our experimental protocol and results. ***v)*** Finally, Chapter 7 includes our conclusions and identifies potential avenues for future research to build upon and improve our work.

Chapter 2

Theoretical Background

This chapter aims to provide a comprehensive overview of the key concepts and techniques related to video summarization and psychoacoustics. This section is divided into several sub-sections that cover the following topics: video summarization, supervised video summarization, unsupervised video summarization, the process of converting frame scores to a video summary, and psychoacoustics, more specifically, the psychoacoustic annoyance metric that we use in this work. Together, these sub-sections provide a solid foundation for understanding the research presented in the subsequent chapters of this thesis.

2.1 Video Summarization

Video summarization is the process of creating a condensed representation of a video by identifying and extracting the most informative or representative segments. This task poses a significant challenge due to the vast volume of data present in videos, such as diverse visual content, complex temporal dynamics, and varying audio information. As a result, determining the most relevant information becomes difficult. There are several different approaches to video summarization, including supervised and unsupervised methods.

In addition, video summarization methods can be classified based on the type of summary they generate. Some methods, falling under the umbrella of dynamic video summarization, concentrate on generating an abridged version of the original video, which entails creating a shorter video that retains the most significant content while preserving its temporal continuity. Other methods create a set of key-frames that represent the most important segments of the video. Some methods also generate a textual summary that describes the main events or actions in the video.

Overall, video summarization is a complex task that benefits from the integration of multiple modalities and cues, such as visual, audio, and textual information. Additionally, it usually requires the use of sophisticated models and algorithms to extract the most relevant information from the video. The ultimate goal of video summarization is to create a condensed representation of the video that contains the most important information while being as short as possible within a certain budget. This condensed representation can be used for a variety of applications, such as video browsing, retrieval, and analysis.

In recent years, there has been a significant amount of research on video summarization, which has led to the development of several methods that can effectively summarize videos. However, there are still many open challenges and areas for improvement, such as addressing videos with complex or overlapping events, managing videos encompassing various modalities, and generating summaries customized to individual user preferences or specific applications.

2.1.1 Supervised Video Summarization

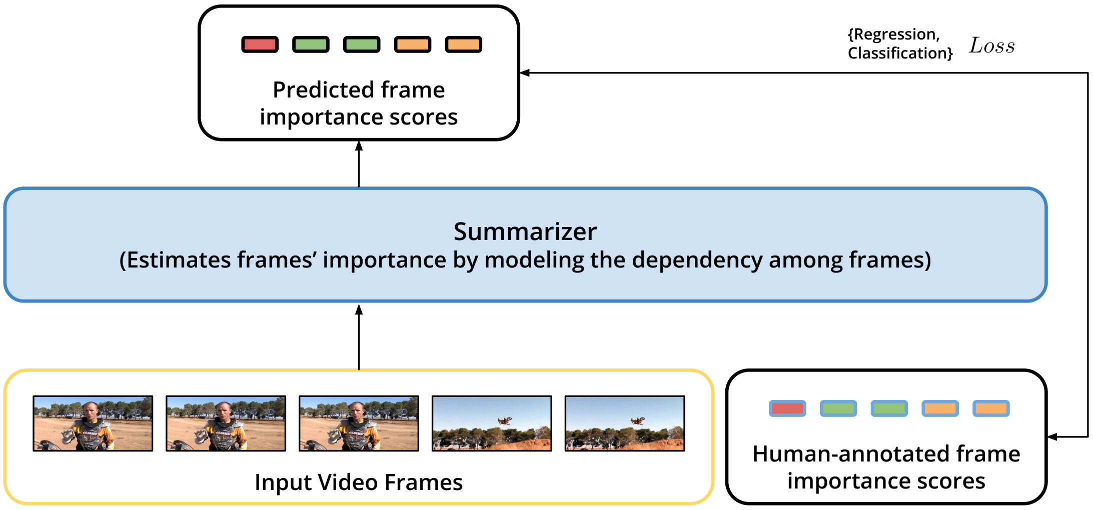
As illustrated in Figure 2.1, supervised video summarization methods rely on the use of annotated data, such as ground-truth scores, to train models that can predict the importance of video segments. These methods are based on the assumption that the annotated data provides a reliable indicator of the most important segments of the video. This approach is useful for datasets that have rich annotations, as it allows for the development of models that can accurately predict the importance of video segments.

The main advantage of supervised video summarization methods is that they can achieve high accuracy and precision as they are trained on annotated data. This characteristic makes them suitable for datasets that have rich annotations, such as the SumMe and TVSum datasets. Additionally, supervised methods can be used to identify patterns and features that are indicative of important segments, which can inform the development of unsupervised methods. However, these methods may exhibit limitations in their generalizability to other datasets or video content, as their performance heavily relies on the quality and relevance of the annotations used for training. Consequently, they may not adapt as effectively to new or diverse video content without extensive annotation work.

2.1.2 Unupervised Video Summarization

Unsupervised video summarization methods, portrayed in Figure 2.2, do not rely on annotated data and instead use other cues, such as the dependency among frames or

Figure 2.1: **High-level representation of supervised algorithms that perform summarization by learning the frames' importance after modeling their dependency.**


Predicted frame importance scores
{Regression, Classification}
$Loss$
Summarizer
(Estimates frames' importance by modeling the dependency among frames)
Input Video Frames
Human-annotated frame importance scores

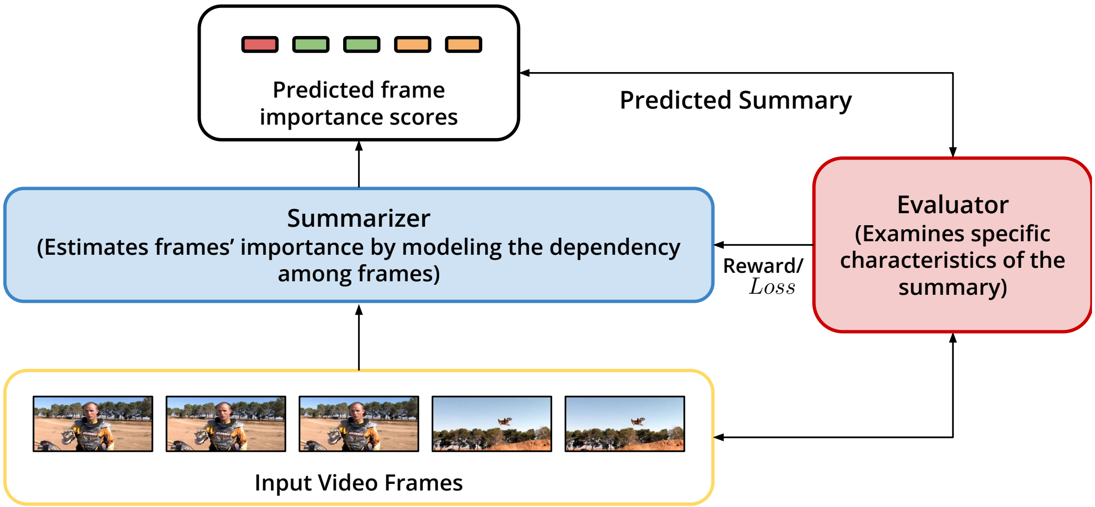
Source: Figure and caption adapted from Apostolidis *et al.* [4].

some consistency analyses between the final summary and input video, to determine the importance of video segments [61, 58, 3]. These methods are based on the assumption that there are underlying patterns and features in the video that can be used to identify the most informative segments. The main advantage of unsupervised video summarization methods is that they do not rely on annotated data, making them suitable for datasets that lack annotation. Additionally, unsupervised methods can be used to identify patterns and features that are indicative of important segments, which can inform the development of supervised methods. However, these methods exhibit the limitation of potentially not being as accurate as supervised methods and can be sensitive to the quality of the cues employed for summarization, such as the strength of the frame dependencies, the robustness of consistency analyses, or the effectiveness of feature extraction techniques in capturing meaningful video characteristics.

2.1.3 Frame scores to video summary

There are various approaches for converting frame scores to a final video summary; in this work, we focus on the method illustrated in Figure 2.3, following the footsteps of previous works [45]. The chosen method utilizes the Kernel Temporal Segmentation (KTS) technique due to its effectiveness in handling varying segment lengths and adaptability to different video content.

Figure 2.2: **High-level representation of unsupervised algorithms that perform summarization by learning the frames' importance after modeling their dependency.**


Predicted frame importance scores
Predicted Summary
Summarizer
(Estimates frames' importance by modeling the dependency among frames)
Reward/
Loss
Evaluator
(Examines specific characteristics of the summary)
Input Video Frames

Source: Figure and caption adapted from Apostolidis *et al.* [4].

The initial step entails generating frame scores using a summarizer method, which can incorporate inputs from diverse modalities, such as audio, visual, and text. A significant portion of the existing literature in the video summarization field concentrates on the development and optimization of summarizer methods. These frame scores indicate the importance of each frame within the video.

Subsequently, the video undergoes temporal segmentation using the KTS method. The method employs features extracted from the video frames to divide the video into segments of varying lengths. For each segment, the frame scores are averaged to produce a segment score, which represents the overall importance of the segment.

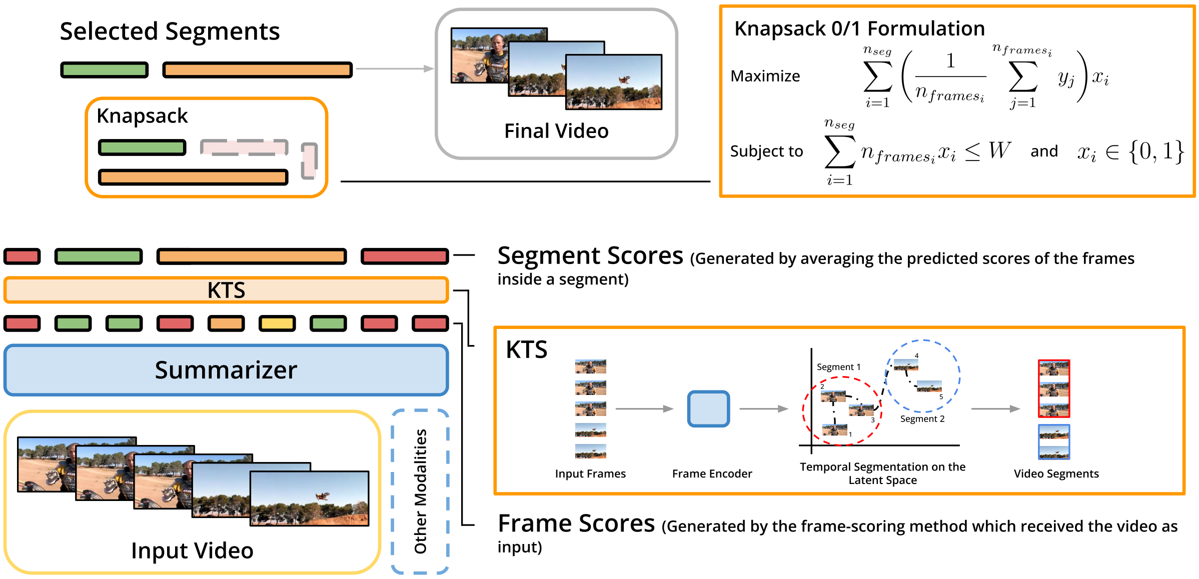
Finally, these segments, accompanied by their respective length and score, are input into the Knapsack 0/1 algorithm. This algorithm selects the segments that fit within a predefined budget, typically 15% of the total video length while striving to achieve the maximum possible score. We define the Knapsack 0/1 algorithm as:

$$\begin{gathered} \text{Maximize} \sum_{i=1}^{n_{seg}} \left(\frac{1}{n_{frames_i}} \sum_{j=1}^{n_{frames_i}} y_j \right) x_i \\ \text{subject to} \sum_{i=1}^{n_{seg}} n_{frames_i} x_i \leq W \quad \text{and} \quad x_i \in \{0,1\}, \end{gathered} \tag{2.1}$$

where n_{seg} is the number of segments of the video, n_{frames_i} is the number of frames in segment i, x_i indicates the presence of the segment in the final summary, and W is the maximum allowed size.

The selected segments are then concatenated to form the final video summary. This process balances the trade-off between the importance of the selected segments, as determined by the frame scores, and the length of the summary, as determined by the budget constraint.

Figure 2.3: **Process of converting frame scores to a final video summary.** Illustration of the various steps involved in the process, beginning with the generation of frame scores using a summarizer method, which can take visual inputs from the video along with other modalities. The frame scores represent the importance of each frame in the video. Next, the video is temporally segmented using the KTS method, which uses encoded frames to temporally divide the video into segments of different lengths. For each segment, the frame scores are then averaged to generate a segment score, which represents the importance of the segment as a whole. These segments, along with their corresponding length and score, are then passed to the Knapsack 0/1 algorithm. The algorithm selects the segments that will fit within a selected budget, usually 15% of the total video length, while trying to obtain the maximum possible score. The selected segments are then concatenated to form the final video summary, balancing the trade-off between the importance of the selected segments and the length of the summary.


Selected Segments
Knapsack
Final Video
Knapsack 0/1 Formulation
Maximize $\sum_{i=1}^{n_{seg}} \left(\frac{1}{n_{frames_i}} \sum_{j=1}^{n_{frames_i}} y_j \right) x_i$
Subject to $\sum_{i=1}^{n_{seg}} n_{frames_i} x_i \leq W$ and $x_i \in \{0, 1\}$
KTS
Segment Scores (Generated by averaging the predicted scores of the frames inside a segment)
Summarizer
KTS
Input Frames
Frame Encoder
Segment 1
Segment 2
1
2
3
4
5
Temporal Segmentation on the Latent Space
Video Segments
Input Video
Other Modalities
Frame Scores (Generated by the frame-scoring method which received the video as input)

Source: Created by the author.

In this section, we acknowledge the limitations associated with the Kernel Temporal Segmentation (KTS) and Knapsack 0/1 algorithm employed in our chosen method. KTS may group segments with visual similarities, which can lead to relevant information being lost if an important frame with highly important content is situated close to a cluster of unimportant frames. Consequently, the significance of such a frame may be diluted within the segment.

Regarding the Knapsack 0/1 algorithm, its global optimization nature can result in the selection of multiple smaller, less relevant segments over a single highly relevant segment to maximize the overall score within the predetermined budget. This can potentially reduce the summary's informativeness.

Despite these limitations, we have adopted this method in line with the existing literature [45, 51, 59], as it has demonstrated effectiveness in various video summarization tasks. This approach serves as a basis for future work in addressing these limitations and developing more sophisticated summarization techniques.

2.2 Psychoacoustics

Noise effects on humans have been extensively studied in a variety of specific domains, including transportation [19, 13, 56, 22, 9], sleep [23, 56, 37], and general health and well-being [38, 42, 27, 10]. Research has shown that noise can have negative impacts on these areas, with transportation noise being linked to decreased performance and increased accidents, sleep noise leading to sleep disturbance and negative impacts on mental health, and general health and well-being being affected by long-term exposure to noise. The effects of noise can be divided into primary effects, which are measurable immediately following exposure to noise, and after effects, which are the longer-term consequences of exposure [37]. In this work, we are focusing on modeling the primary effects of noise on humans. These effects are the immediate psychological responses to sound exposure that occurs when individuals watch and react to video segments. We aim to capture these responses in order to better understand how to create better video summaries.

According to Zwicker *et al.* [64], sound perception by a human listener can be estimated by psychoacoustic properties that are closely related to the relative degree of perceived auditory annoyance. The authors propose the psychoacoustic annoyance (PA) metric as a function of four other properties: *fluctuation, roughness, loudness, and sharpness*. These properties are designed to objectively approximate the effects of different sound stimuli on the human ear.

In this thesis, we leverage these estimations of psychological responses to sound for our video summarization pipeline. We extract the PA metric from each video to use them as pseudo-labels during training time. This training strategy allows us to achieve strong results in the unsupervised video summarization setting, outperforming all methods on the SumMe dataset and most of the baselines on the TVSum dataset.

2.2.1 Psychoacoustic Annoyance (PA)

This section describes the processes involved in the calculation of the Psychoacoustic Annoyance (PA) values for a sound. Psychoacoustic Annoyance is a measure of the subjective discomfort that a person experiences as a result of exposure to certain sounds [64]. It is an important concept in the field of psychoacoustics, which is the study of how the human auditory system perceives and processes sound. It is important to note that PA is different from the sound pressure level (SPL), which is a physical measure of the amplitude of sound waves. While SPL is an objective measurement of the intensity of a sound, PA focuses on the subjective perception of annoyance, taking into account various psychoacoustic factors that contribute to the discomfort experienced by a listener. In this work, we used the measurement model from Zwicker *et al.* [64]. In this model, the acoustic annoyance of a sound is related to the following psychoacoustic indices:

- *Fluctuation (F) and Roughness (R)*: these indices measure the modulation of a signal over time. A modulated signal with higher values for these indices tends to be more unpleasant. On high fluctuation signals, the listener can hear each individual rise and fall in the sound;
- *Loudness (N)*: this property is based on perceived loudness, and it is based on human subject studies. It measures how loud people with average hearing perceive a sound;
- *Sharpness (S)*: it is calculated by a weighted sum of specific loudness levels in different bands. A sound with higher sharpness is more unpleasant. Sharp sounds have a greater proportion of high frequencies than the rest of the energy in them.

The PA value can be calculated as follows:

$$PA = N_5\left(1 + \sqrt{\omega_S^2 + \omega_{FS}^2}\right), \tag{2.2}$$

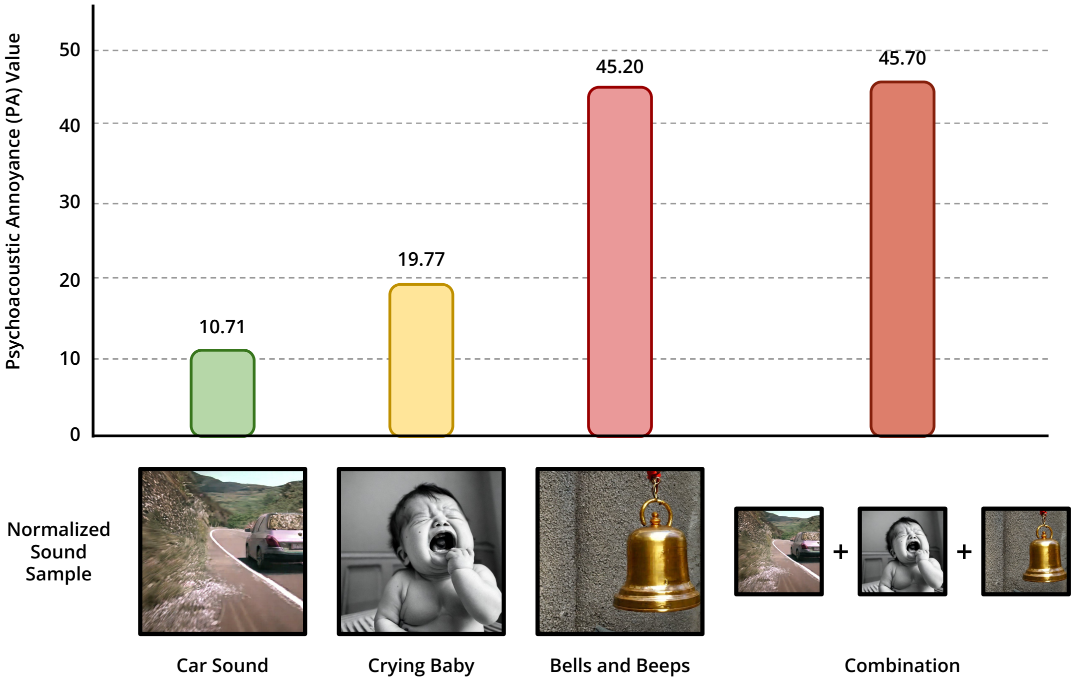
$$\omega_S = \mathbb{1}[S > 0] \times (S - 1.75)log(N_5 + 10), \tag{2.3}$$

$$\omega_{FS} = 2.18 \times N_5^{-0.4} \times (0.4F + 0.6R), \tag{2.4}$$

where N is the loudness, N_5 is the 95th percentile of loudness, S is the sharpness, F and R are fluctuation and roughness, respectively, and $\mathbb{1}[X]$ is the indicator function, that evaluates the predicate X, returning 1 if it is true and 0 otherwise.

In this work, we aim to model the importance of different video segments in order to compose a summary by examining the relationship between psychoacoustic and human attention. As shown in Figure 2.4, the semantics of sound can significantly impact human perception, as demonstrated by the various PA values of common sounds that have been

Figure 2.4: **Example Psychoacoustic Annoyance (PA) values for common sounds.** The sound samples were normalized to the same sound pressure level (SPL). The sounds have the following characteristics: **(i)** Car Sound: low, rumbling, continuous sound; **(ii)** Crying Baby: High-pitched, high-fluctuation sound; **(iii)** Bells and Beeps: unpredictable, high-pitched, multiple-source sound. The combination of the sounds results in a PA value close to the highest individual value. This highlights the complex nature of sound perception and the way in which sounds interact with one another to shape our experience of them.



Source: Created by the author.

normalized to the same sound pressure level. As seen in the data, the car sound is associated with a value of 10.71, the crying baby with 19.77, and bells and beeps with 45.20. Notably, the combination of all of these sounds yields a value of 45.70, which is not the additive result of the individual values. This highlights the complex nature of sound perception and the way in which sounds interact with one another to shape our experience of them. These examples also indicate that the Psychoacoustic annoyance is not simply the sum of the individual sounds but rather a result of the interactions between them. These interactions can manifest in a variety of ways, such as masking, the interaction of frequency components and temporal envelope, and cognitive factors. We aim to utilize this understanding of the role of sound semantics to accurately model the importance of video segments.

Chapter 3

Related Work

This chapter provides an overview of the related work in the field, including supervised and unsupervised learning methods and the use of psychoacoustics in video processing. The focus is on the recent progress made in video summarization using deep learning techniques and the contributions made by each method toward advancing the field. The chapter also highlights the key similarities and differences between the proposed methods and our proposed approach, which leverages psychoacoustic features as pseudo-labels for self-supervised training.

3.1 Video Summarization

The field of automatic video summarization has experienced significant growth in recent years, fueled in part by the widespread adoption of deep learning methods in computer vision. This has resulted in a proliferation of proposed methods for video summarization. The establishment of standard evaluation protocols for this task has been aided by the introduction of benchmark datasets such as SumMe [29], and TVSum [51]. These datasets have played a key role in advancing research in the field by providing a common platform for the evaluation of various video summarization methods. Early learning-based video summarization methods were focused on manipulating and extracting frame features such as clustering [36], and pairwise deep ranking models [57]. Sequence-based models followed by formulating the summarization problem as a structured prediction task on sequential data [59, 36]. These models used recurrent networks, such as LSTMs, to leverage the sequential structures in videos. In a parallel research direction, early methods also introduced the use of multi-modality by incorporating information from sensors, gaze [33], and text [44, 11]. Instead of using recurrent architectures, our method models the interaction between audio and video frames using a transformer network to leverage multi-modal signals.

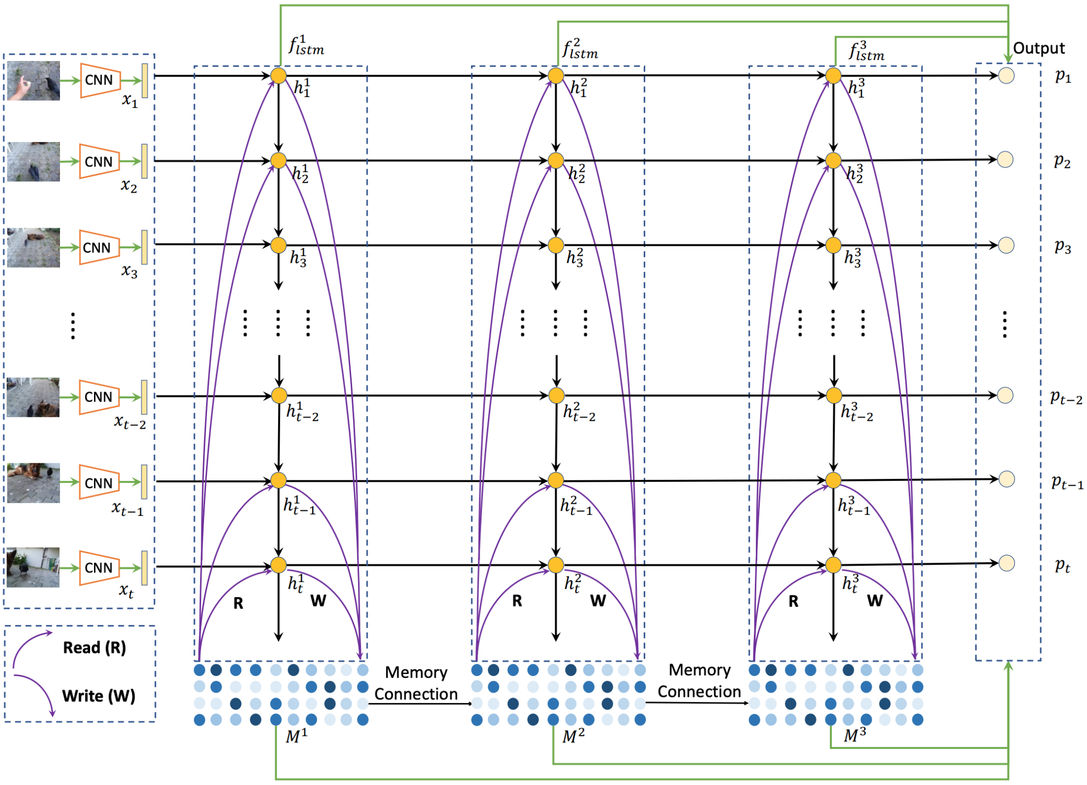
3.2 Supervised Video Summarization

Supervised learning methods have achieved state-of-the-art results in video summarization, thanks to the availability of well-annotated datasets. A high-level representation of supervised video summarization methods is portrayed in Figure 2.1. Fajtl *et al.* [20] proposed a method based on self-attention to replace the commonly used recurrent models. The authors demonstrated it was possible to use a simpler and more efficient model while achieving better performance. Apostolidis *et al.* [5] extended the self-attention method by combining local and global multi-headed attention mechanisms. Their model is able to model frames' dependencies at different granularity levels. Wang *et al.* [55] overcome the RNNs limitations by augmenting LSTM layers with a memory layer, enforcing the explicit modeling of the long dependency among video frames, and achieving state-of-the-art results on the SumMe dataset. Their method is depicted in Figure 3.1. In this thesis, we use psychoacoustic information from videos to generate pseudo-labels for the data rather than relying on human annotations. This strategy enables us to train the network in a self-supervised manner, using the data itself to provide supervision rather than explicit labels.

Attention mechanisms have been widely used in video summarization methods in order to capture important information from the video and generate a summary. Feng *et al.* [21] proposed the use of external memory to record visual information of the whole video and then predict the importance score of a video shot based on the global understanding of the video frames in order to generate a summarized version of the video. The method introduced by Liu *et al.* [34] uses a multi-concept self-attention mechanism to identify informative regions across temporal and concept video features while enforcing consistency between the video and summary. The MSVA method by Ghauri *et al.* [26] predicts importance scores in video summarization by combining three feature sets for visual content and motion. It utilizes a parallel attention mechanism before fusing these features, leveraging both visual content and motion features to generate the final summary. In our transformer-based approach, the self-attention layers use the attention mechanism to model the relationships between every frame and audio segment.

Zhu *et al.* [63] leverage temporal consistency in order to generate summarized videos. Their method uses a dense sampling of temporal interest proposals with multi-scale intervals to extract long-range temporal features for interest proposal location regression and importance prediction and assigns positive and negative segments for the correctness and completeness of the generated summaries. Alternatively, it can directly predict the importance scores of video frames and segment locations in an anchor-free approach. DSNet is formulated as a regression problem with temporal consistency and integrity constraints. In their more recent work, Zhu *et al.* [62] extract object-level and

Figure 3.1: **The overall framework of the Stacked Memory Network (SMN) video summarization method.** Given a video, they first employ the pretrained CNN network to extract video frame features. Then, they forward these features into their stacked memory networks to update the states of LSTM layers and memory layers. After combining the states from these LSTM layers and memory layers, they employ a fully-connected layer to predict each frame an importance score. In addition, they also explore different types of connections between two memory networks to fuse the learned representation from previous layers.



Source: Figure and caption adapted from Wang *et al.* [55].

relation-level information in order to capture spatial-temporal dependencies from videos. The method builds spatial graphs on object proposals and temporal graphs by aggregating spatial graphs. Then, it performs relational reasoning over the spatial and temporal graphs using graph convolutional networks and extracts spatial-temporal representations for importance score prediction and key shot selection.

3.3 Unsupervised video summarization

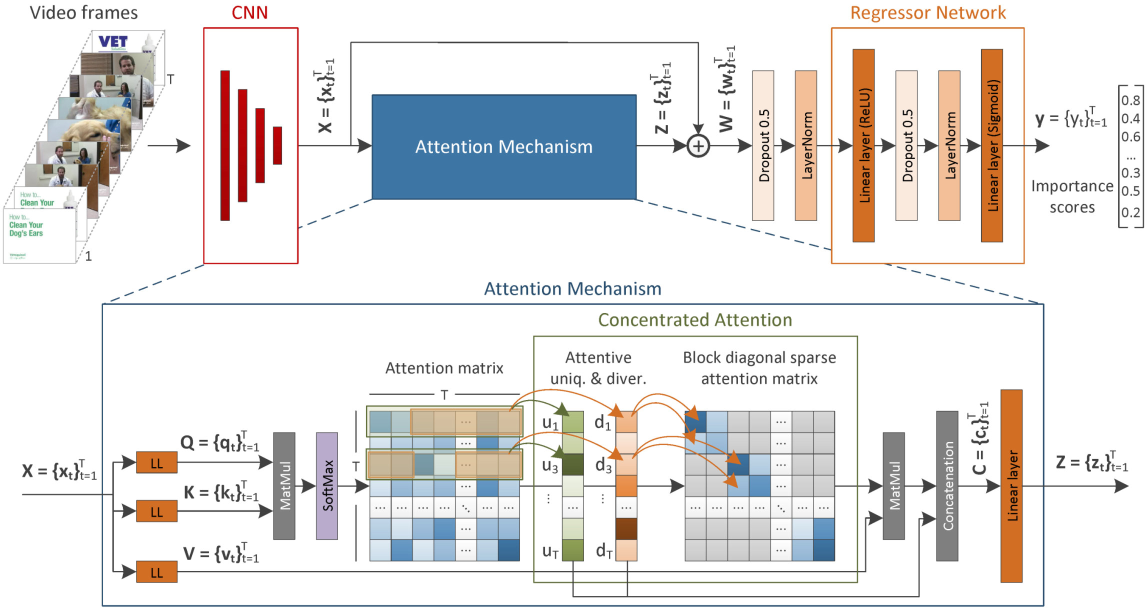
Unsupervised video summarization methods do not require ground-truth data and can be trained using only a large collection of original videos, making them useful for real-world applications where manual annotation of data may not be feasible. Zhou *et al.* [61] present a method for video summarization called DSN, which is formulated as a sequential decision-making process. DSN predicts a probability for each frame, indicating how likely it is to be selected, and uses these probabilities to select frames and generate a summary. The method is trained using an end-to-end, reinforcement learning-based framework with a novel reward function that evaluates the diversity and representativeness of the generated summaries. The reward function does not require labels or user interactions, and DSN aims to maximize the reward by learning to produce more diverse and representative summaries.

Generative Adversarial Network (GAN)-based video summarization has been shown to be a promising approach for generating condensed summaries of videos without the need for human-labeled data. Yuan *et al.* [58] proposed using a cycle-consistent adversarial LSTM architecture to maximize the information and compactness of the summary video. It consists of a frame selector, which is a bi-directional LSTM network that learns video representations that embed long-range relationships between frames, and an evaluator that defines a learnable information-preserving metric between the original video and summary video. The evaluator is composed of two GANs, and the consistency between their outputs is used as the information-preserving metric for video summarization. The evaluator supervises the selector to identify the most informative frames to include in the summary. Methods such as Apostolidis *et al.* [7] and Apostolidis *et al.* [3] incorporate an attention mechanism into the GAN framework, which improves the ability of the model to identify the most relevant parts of the video. Apostolidis *et al.* [2] use an actor-critic model to learn a sequence generation policy for selecting important video fragments, which further enhances the performance of the GAN-based model. Although these methods have shown promising results in unsupervised video summarization using GANs, this formulation is notoriously hard to train [39].

In the work by Jung *et al.* [31], the authors address two main challenges in the video summarization task: (1) flat distributions of output scores for each frame, which hinder feature learning, and (2) difficulty in training with long videos. To address these challenges, the authors propose a regularization loss term called variance loss and a two-stream network called the Chunk and Stride Network (CSNet), respectively. They also introduce an attention mechanism to handle dynamic information in videos. Our transformer-based approach is able to model the dependencies between visual and auditory features in long videos while also handling the dynamic information from these modalities.

Park *et al.* [43] proposed a video summarization method using a graph-based approach to identify and extract important segments from a video. It is based on the idea of modeling the video as a graph, where the nodes represent individual frames of the video and the edges represent the affinity between them. SumGraph has been shown to be effective at summarizing videos with complex structures and has been applied to a

Figure 3.2: **The analysis pipeline of the CA-SUM method.** The lower part illustrates the processing steps within their attention mechanism.



Source: Figure and caption adapted from Apostolidis *et al.* [6].

variety of video summarization tasks. It has also been shown to be complementary to other video summarization methods and can be used to improve their performance.

Hu *et al.* [30] proposed a two-stream LSTM network that leverages both spatial saliency and temporal semantic dependencies to improve the critical features of images in user-created videos. The method includes a mechanism to filter out irrelevant information and a system to extract temporal dependencies on semantic features. It also utilizes a multi-feature-based reward function to strengthen the model and employs the Deep Deterministic Policy Gradients (DDPG) algorithm for unsupervised training.

In their most recent work, Apostolidis *et al.* [6] tackled the limitations of existing approaches to unsupervised video summarization, including the unstable training of Generator-Discriminator architectures, the use of RNNs for modeling long-range dependencies, and the inability to easily parallelize the training process of RNN-based network architectures. They use a self-attention mechanism and a concentrated attention mechanism to estimate the importance of video frames and extract and exploit knowledge about the uniqueness and diversity of frames to make better estimates of the significance of different parts of the video. It also has fewer learnable parameters than other methods. Their method is illustrated in Figure 3.2.

It is important to notice that, similarly to this thesis, some works design a solution for a specific supervision modality, such as supervised or unsupervised learning, but ultimately evaluate their method in both modalities. This is why, although some methods may be described in a specific section of this chapter that is focused on a particular supervision modality, they are also compared in Chapter 6 in both supervised and unsupervised settings. This allows for a thorough evaluation of the performance of these methods in different scenarios and enables a more comprehensive understanding of their capabilities. By experimenting with both supervision modalities, these works are able to provide a complete picture of the effectiveness of their method.

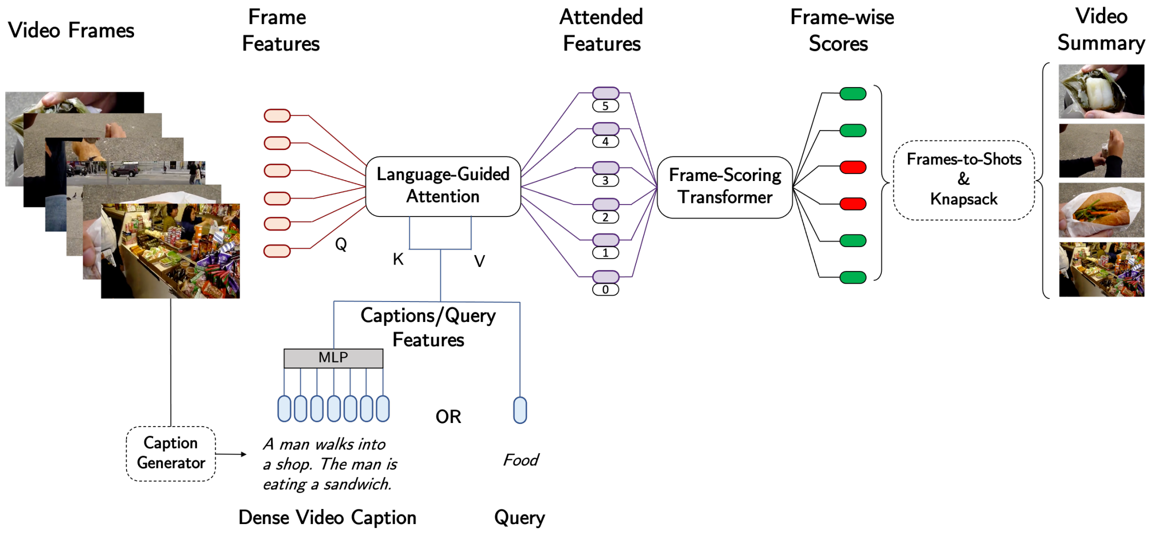
Our proposed method is primarily related to three recent video summarization approaches: the work of Narasimhan *et al.* [41], Zhong *et al.* [60], and another study by Narasimhan *et al.* [40]. The approach of Narasimhan *et al.* [41], shown in Figure 3.3, employs a multi-headed attention layer to attend to one modality using the other. In contrast, Zhong *et al.* [60] minimize the distance between video and text representations without any frame-level labels. Both of these approaches demonstrate promising results in video summarization tasks and share similarities with our proposed method. Narasimhan *et al.* [40] proposed a video summarization method for instructional videos using pseudo summaries and achieved state-of-the-art results on the WikiHow Summaries dataset. Our approach leverages psychoacoustic features as pseudo-labels to enable self-supervised training, which is a unique aspect of our method. Our video summarization method has been shown to achieve very good performance, comparable to that of Narasimhan *et al.* [41], outperforming them on both supervised and unsupervised settings on the SumMe dataset. However, unlike their method, our method does not require access to text data for training. Their training schema relies on either ground-truth text labels, which can be expensive and time-consuming to obtain, or automatically extracted video captions. In contrast, our method uses the audio psychoacoustic features extracted from the audio in the videos to generate pseudo-labels, allowing us to train our model in an unsupervised manner. Our experiments show that this approach outperforms other unsupervised methods of video summarization.

3.4 Psychoacoustics in Video Processing

In the field of video processing, there are two works that are most closely related to our approach when it comes to processing human responses to sounds. Their task is a sub-task of video summarization, called semantic hyperlapse (first defined by Ramos *et al.* [47]), which involves speeding up long first-person videos to make them more manageable to watch by adding the constraints of temporal continuity and visual stability to the final summarized video. These methods leverage human responses to sound in order to propose new methods for video processing beyond traditional semantic audio representations.

The first work, proposed by de Matos *et al.* [15], presents a new fast-forward

Figure 3.3: **Overview of CLIP-It!** Given an input video, CLIP-It generates a summary conditioned on either a user-defined natural language query or an automatically generated dense video caption. The Language-Guided Attention head fuses the image and language embeddings, and the Frame-Scoring Transformer jointly attends to all frames to predict their relevance scores. During inference, the video summary is constructed by converting frame scores to shot scores and using the Knapsack algorithm to select high-scoring shots.



Source: Figure and caption adapted from Narasimhan *et al.* [41].

method that considers both the information present in the video and the background music. The method uses neural networks to automatically recognize the emotions induced in the video and the background music and combines the contents in the accelerated video through a new method of frame selection that aims to maximize the similarity of the induced emotions. The method is evaluated on a large dataset with different videos and songs, and the results show that it achieves the best performance in matching emotion similarity while maintaining the video's visual quality.

Finally, it is also important to emphasize that Furlan *et al.* [24] proposed the only work, to the extent of our knowledge, that used psychoacoustic annoyance as a proxy for human interaction on a video processing task. Interestingly, in their work, they aim to minimize the PA in the final video so the users would have a more pleasant auditory experience. In our case, we associate the high PA values with high user attention, meaning more meaningful segments. Although their work and ours are optimizing opposite objective definitions, our findings discussed in Section 6.2.4 suggest that psychoacoustic annoyance is a good proxy for estimating frame importances.

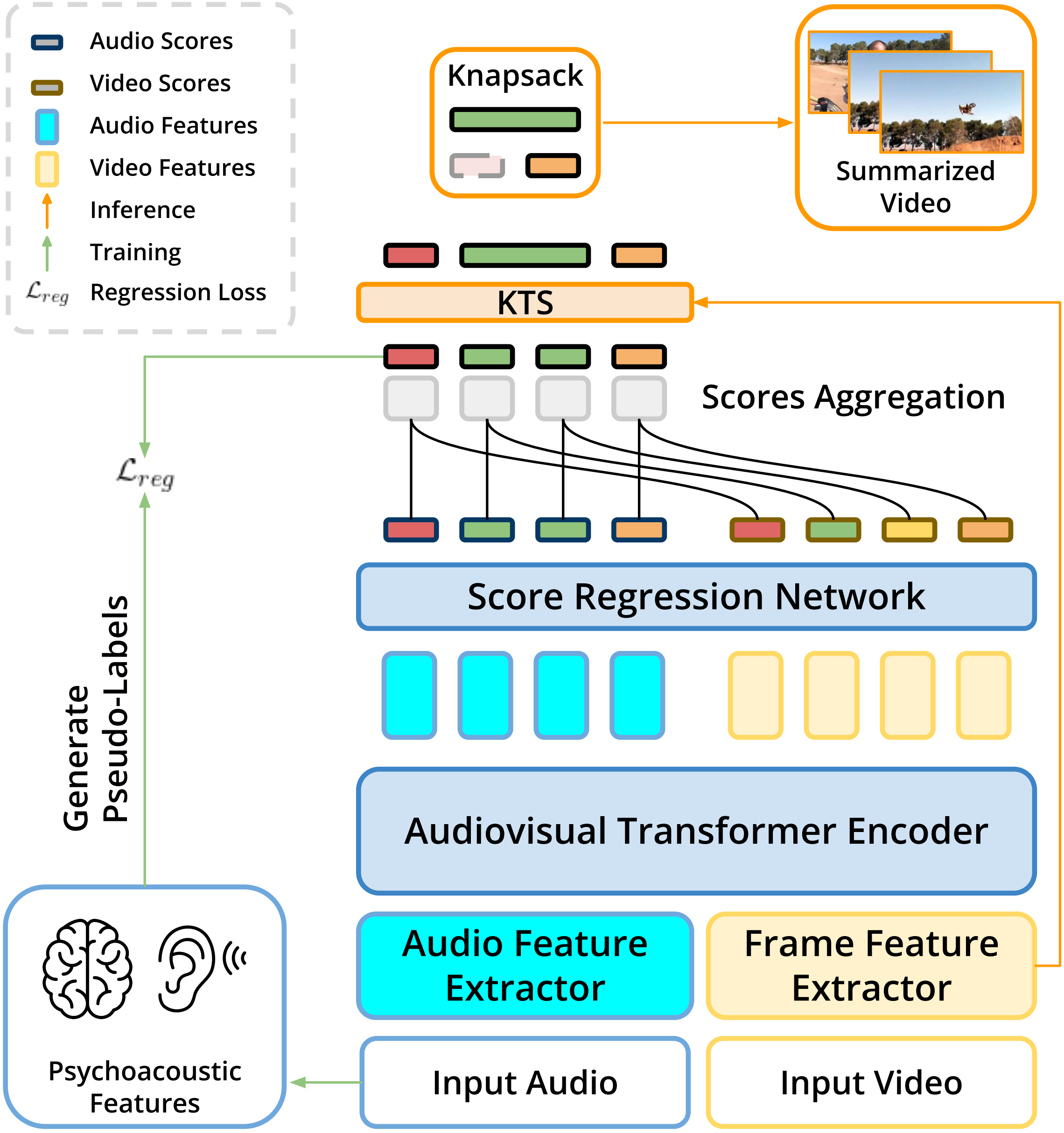
Chapter 4

Methodology

In this chapter, we present a comprehensive overview of the techniques and approaches that we have developed for tackling the video summarization problem. Specifically, we detail our method for generating psychoacoustic pseudo-labels, which we use to train our system in an unsupervised manner. We also describe our approach for integrating audio and visual signals to automatically generate a summary of a given video. To provide a clear and comprehensive understanding of our methods, we first provide an overview of our overall approach and then delve into the details of each individual section in our methodology.

Illustrated in Figure 4.1, our method is based on a self-supervised framework for video summarization. Given an input video, we extract its auditory and visual information using feature extractors to feed a transformer encoder that will incorporate the information from both modalities into its output tokens. Then, these tokens are fed into a score regression network which will generate the scores for each individual frame from both modalities, which are then aggregated. Since human annotations are rather expensive, leveraging information that is already in the data through self-supervised learning can be a useful alternative that allows the model to learn from large amounts of unlabeled data, saving time and resources. We introduce pseudo-labels generated from psychoacoustic features instead of using human-annotated labels. From the input audio, we extract psychoacoustic features that are used to supervise the frame-scoring pipeline. After the frame scores are predicted, the KTS algorithm is used to group the video into semantically-consistent segments. Finally, the 0-1 Knapsack algorithm is used to select the most important segments to include in the summary based on their number of frames and average frame importance.

Figure 4.1: **Overview of the proposed approach for unsupervised video summarization.** After being extracted, the psychoacoustic features from the video are used as pseudo-labels for training the frame-scoring pipeline. The model leverages auditory and visual information to generate the frames' importance scores, which are converted into the final summarized video.



Source: Created by the author.

4.1 Psychoacoustic Pseudo-Labels

Pseudo-labels are important as they allow us to train the model in an unsupervised manner, without the need for manual annotations, enabling us to make use of large amounts of data without the cost of manual annotation and allowing models to learn from a wider range of information. We define the pseudo-labels for a video as Y', with size T, the number of input tokens for the visual modality.

For each video in the dataset, we extract the audio and convert it to mono by averaging the original channels, if needed. Converting the audio to mono is a practical choice that simplifies the audio processing while preserving the necessary information for our task. This step reduces the dimensionality of the audio data and ensures compatibility with various audio formats present in the dataset. Next, the one-channel signal is split into P-second non-overlapping segments. The choice of P-second non-overlapping segments ensures that each segment captures distinct audio characteristics, while minimizing the computational complexity associated with processing overlapping segments.

After segmenting the audio, we compute the PA for each segment. Given an audio length of S seconds, we obtain a final list of $\frac{S}{P}$ values. The PA values correspond to the perceived annoyance the sound segment has. These values are usually bounded between 0 and 100 originally, but they are normalized to the 0-to-1 range to match the scale of human-annotated ground-truth frame score values of the datasets we used. Finally, the list of PA values is upsampled to match the length of the visual input sequence length T. The upsample method used for this process is explained in the next section, as it is used to align the pseudo-labels with the visual input. Proper alignment between the pseudo-labels and the visual input is essential to ensure that our model can effectively learn from both modalities and generate accurate video summaries.

4.2 Audiovisual Video Summarization

Frame Encoding. Encoding frames using a pre-trained neural network allows us to leverage the learned feature representations from large amounts of data, capturing the underlying patterns and relationships in the data and thus improving the performance of our video summarization system. To encode the video frames, we use a pretrained image network f_{img}. Given a set of frames from the input video, we extract the frame features $X_{img} = \{x_{img_t}\}_{t=1}^{T}$. The image network is kept frozen during our training process.

Audio Encoding. Encoding the audio signal with a pre-trained neural network allows for the extraction of high-level, semantically meaningful representations of the audio content, which can provide valuable information for video summarization tasks. The audio is encoded using a pretrained audio network f_{aud}. Then, we compute the log-power spectrograms of w-second windows using the Short-Time Fourier Transform and feed them to f_{aud}, generating the audio features $X_{aud} = \{x_{aud_s}\}_{s=1}^{S/w}$.

Given that the number of visual input tokens T is greater than the ratio of audio length S to the window size w, we need to upsample the audio features to match the length of the frame features. This is an important step to ensure that the audio and visual modalities have the same temporal resolution, which allows for better fusion and alignment during the training process.

To upsample the audio features, we apply a simple yet effective repetition-based approach. For each consecutive audio feature x_s, we calculate the repetition factor, r, as follows:

$$r = T/(\frac{S}{w}).$$

The repetition factor r represents the number of times each audio feature needs to be repeated in order to achieve the desired length. By repeating each audio feature x_s for r times, we create an upsampled audio feature sequence that has the same length as the frame features. This approach maintains the temporal structure of the audio features while ensuring that the audio modality is aligned with the visual modality in terms of sequence length.

It is worth noting that this repetition-based upsampling method is a simple yet effective technique that does not introduce any additional computation or complexity to the model. More advanced upsampling methods, such as interpolation or learned upsampling, could potentially be explored in future work to further improve the alignment between audio and visual modalities. However, the current approach provides a reasonable trade-off between simplicity and performance for the task of video summarization.

Audiovisual Transformer Encoder. To model and contextualize the information of video frames and audio, we use a transformer encoder [54]. The model takes a concatenated multimodal sequence of features $X_{multimodal} = X_{img} \frown X_{aud}$ and outputs the corresponding attended features. Since features from each modality can be extracted from a variety of pretrained networks with different feature dimensionality, we add modality-specific MLP heads to project both audio and frame features onto the same embedding space. We let these heads be trained in an end-to-end matter. Thus the common embedding space is shaped by the final task.

We also introduce modality-specific positional embeddings. Positional embeddings are crucial for providing the model with information about the relative or absolute positions of the input tokens in a sequence. By adding independent positional embeddings to each modality (audio and visual), we help the model to better understand the temporal relationships within each modality and enforce the correspondence between the audio and frame features at the same timestep.

In other words, modality-specific positional embeddings enable the model to differentiate between the positions of the audio and visual tokens in their respective sequences. This distinction allows the model to effectively capture the temporal structure and relationships within each modality, while also facilitating the alignment and fusion of the two modalities at corresponding timesteps.

Score Regression Network. The feature vectors that have passed through our transformer encoder have contextual information on the whole sequence for both modalities. We introduce a score regression network that takes the attended features and generates an importance score for each. As we understand the importance of the auditory events in capturing users' attention when watching videos, each modality's features generate separate scores. The final scores Y are obtained by aggregating the modality-specific scores

$$Y = \lambda_{img} Y_{img} + \lambda_{aud} Y_{aud},$$

where Y_{img} and Y_{aud} are the scores generated from the image and audio attended features, respectively, and $\lambda_{img} + \lambda_{aud} = 1$. The score regression network comprises fully connected layers with shared weights along the whole multimodal sequence.

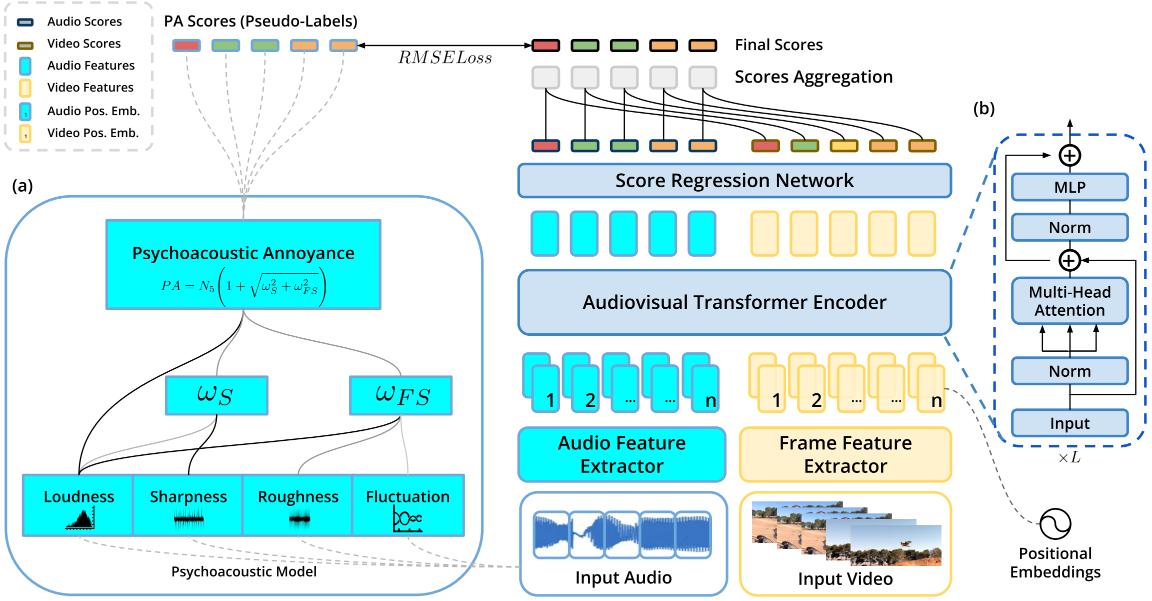
Learning. To train our summarization method, we minimize the Root Mean Squared Error between the predicted scores Y and the pseudo-labels Y':

$$RMSE = \sqrt{\sum_{i=1}^{T} \frac{(y'_i - y_i)^2}{T}},$$

given that $Y' = \{y'_0, y'_1, ..., y'_T\}$ and $Y = \{y_0, y_1, ..., y_T\}$.

In our unsupervised setting, we found that the use of auxiliary losses, such as in the work of Narasimhan *et al.* [41], hindered our model's convergence during training. We argue this could be due to the way in which the audio tokens interact with the frame tokens after they have been processed by the transformer encoder, and we leave this investigation to further extensions of this work. We found that in the unsupervised setting, the pseudo-labels generated from the psychoacoustic information in the videos provide sufficient signal for the model to obtain a good performance.

Figure 4.2: **Overview of our pseudo labeling generation and video summarization model training**. Our approach is composed of two main steps: (a) Processing the video audio to extract psychoacoustic information that is used to compose the pseudo-labels. (b) Then, using these generated labels, we train an audiovisual video summarization model. After the input modalities have their features extracted, they are added to the respective positional embedding and fed into the audiovisual transformer encoder. The attended features resulting from the attention mechanism are passed to a score regression network that generates scores for both modalities. Finally, The audio and visual scores are aggregated, composing the final score. At inference time, the final summary is composed by segmenting the video using Kernel Temporal Segmentation and then selecting the segments using the Knapsack 0/1 algorithm using the average score as "value" and the number of frames as "weight".



Source: Created by the author.

Chapter 5

The Audiovisual Summarization Dataset (AVSum)

This chapter presents the methodology for our proposed video summarization dataset. The main goal of this dataset is to explore the relationship between auditory stimuli and important events in videos in order to improve future video processing methods. This dataset is intended to supplement existing datasets, such as SumMe and TVSum, by providing diverse audiovisual settings and frame-level scores to be used in future video processing methods.

5.1 Data Collection

The Audiovisual Summarization Dataset (AVSum) was created with the aim of advancing the field of video summarization. The data collection process involved selecting videos based on specific criteria, which will be detailed in the next subsection, and then extracting labels for the important frames. The extracted labels consisted of importance scores and psychoacoustic pseudo-labels, which were used to train the summarization models. In this section, we describe the video selection criteria and the label extraction process and provide an overview of the dataset statistics.

5.1.1 Video Selection Criteria

The data collection for the AVSum dataset focused on first-person videos, which are known for their long-running recordings and potential for summarization (inspired by Ramos *et al.* [47]). Specifically, we chose videos of people touring around cities, as they provide a rich and variable environment for different visual and auditory stimuli. The selection process for the videos aimed to guarantee a diverse range of auditory and visual information, catering to various content types and styles. This diversity is crucial for training a robust summarization model that can generalize well across different scenarios. Additionally, the videos were chosen to be of sufficient length, ensuring that they contained enough content to be effectively summarized, thus providing meaningful challenges for the video summarization task. We provide some examples of the dataset scenes in Figure 6.1.

5.1.2 Label Extraction

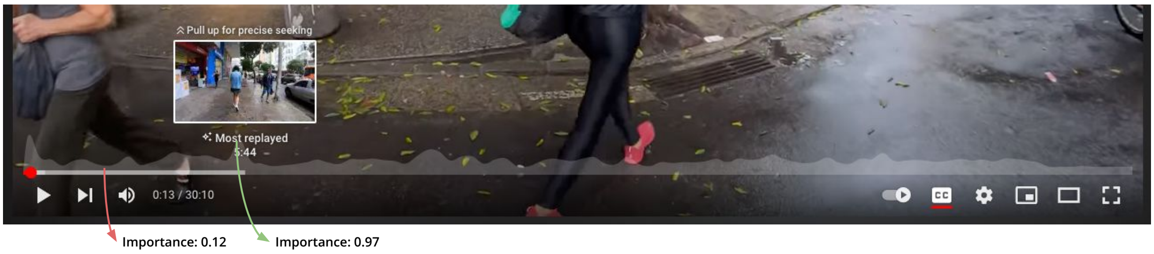
In this section, we explain how we extracted the labels for the AVSum dataset. The process includes the extraction of the frame-level importance scores and the psychoacoustic annoyance values.

Importance Scores. In May 2022, YouTube introduced a new feature named heatmap-a graph by the progress bar that shows the most replayed parts of a video [1]. For the channels that have this feature enabled, the graph provides values from 0 to 1 for each 1% increment of the video's duration, as illustrated in Figure 5.1. This feature is intended to help creators understand which parts of their videos are the most engaging and interesting to their audience and to use this information to improve their content. Additionally, the feature can also be used by viewers to quickly navigate to the most popular parts of a video and to easily identify which parts of the video are worth watching [2]. In the context of the video processing community, these values can be interpreted as ground-truth data

[1] https://twitter.com/TeamYouTube/status/1527024322359005189

[2] https://techcrunch.com/2022/05/18/youtubes-player-gains-new-features-including-most-replayed-video-chapters-single-loop-and-more

Figure 5.1: **The heatmap YouTube feature graph showing engagement levels throughout a video on YouTube, with values ranging from 0 to 1 for each 1% increment of the video's duration.**



Source: Created by the author.

collected from multiple users and be a good way forward to create bigger, less-costly datasets.

In this work, we use these importance values extracted from each YouTube video page to compose our ground-truth importance scores. As the annotations have a fixed length for every video of 100 samples, we resample the signal to fit the number of final frames of each video after our sampling, following the protocol of Gygli *et al.* [29].

As a result, we obtain frame scores in the same format as those present in other datasets in the literature. These frame scores serve as the ground-truth scores for our dataset and provide a consistent and reliable basis for training and evaluating video summarization models. By using the YouTube heatmap feature as a source of ground-truth data, we can efficiently create larger datasets with lower annotation costs, as these scores are derived from real-world user engagement and reflect the interests of a diverse audience. This approach not only facilitates the development of more accurate and robust summarization models but also ensures that the generated summaries align well with the preferences of actual users.

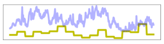
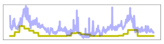
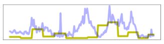
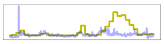
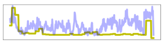
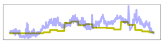
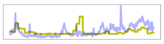
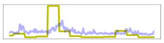
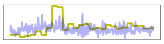
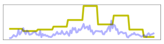
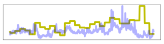
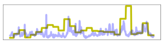
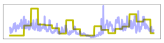
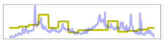
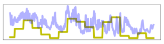
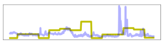
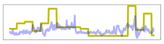
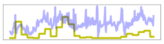
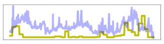
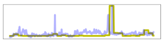
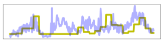
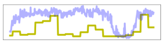
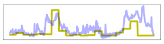
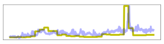
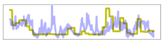
Psychoacoustic Pseudo-labels. After extracting the audio signals from each video, we segment them into 1s audio clips and resample these clips to $16Hz$, averaging left and right channels if the signal is stereo. Then, we extract the Psychoacoustic Annoyance (PA) values using the protocol described in Section 4.1, generating the pseudo-labels for each video.

5.2 Dataset Statistics

In this section, we present the statistics of the AVSum dataset, including the number of videos, the duration, and the content. The dataset is composed of 27 videos from YouTube [3] with their duration varying from 2 to 10 minutes long. The dataset has frame-level importance scores calculated from the heatmap extracted from youtube along with the videos. Table 5.1 describes the individual information for each video of the dataset, including the number of frames, frames per second, the average PA value, and the average ground-truth value.

[3] https://www.youtube.com

Table 5.1: **Summary of video features of our AVSum dataset.** The table shows the number of frames, frames per second (FPS), and the mean Psychoacoustic Annoyance (PA) Score and Ground-truth (GT) values for each video. In the last column, it is depicted a profile for both PA (in blue) and GT (in yellow) curves along the video. *Best seen in color and with zoom.*

Video Title	# Frames	FPS	PA Avg.	GT Avg.	PA vs GT
video_1	9600	30	0.575	0.295	
video_2	8400	30	0.343	0.395	
video_3	3600	30	0.158	0.270	
video_4	16200	30	0.253	0.424	
video_5	14400	30	0.223	0.333	
video_6	13500	30	0.309	0.356	
video_7	9600	30	0.308	0.373	
video_8	10800	30	0.425	0.291	
video_9	10800	30	0.210	0.255	
video_10	14400	30	0.200	0.271	
video_11	10800	30	0.447	0.202	
video_12	15000	30	0.157	0.158	
video_13	18000	30	0.405	0.175	
video_14	12300	30	0.170	0.146	
video_15	10200	30	0.240	0.193	
video_16	12600	30	0.618	0.265	
video_17	15000	30	0.280	0.179	
video_18	14400	30	0.239	0.200	
video_19	15000	30	0.300	0.261	
video_20	11400	30	0.317	0.227	
video_21	14400	30	0.333	0.150	
video_22	6000	30	0.336	0.203	
video_23	14400	30	0.128	0.232	
video_24	13500	30	0.386	0.178	
video_25	11400	30	0.268	0.196	
video_26	12900	30	0.167	0.151	
video_27	7200	30	0.257	0.217	

Chapter 6

Experiments

In this chapter, we present the results and discussion of our proposed method for video summarization. We begin by discussing the quantitative results for both unsupervised and supervised video summarization, followed by a discussion of the qualitative results. Additionally, we present the results of our zero-shot experiments, where we evaluate our method on a new audiovisual video summarization dataset. We then present the results of our analysis of the ground truth scores and the psychoacoustic annoyance. We also conduct an ablation study to investigate the impact of various components of our method on the performance of the system. This includes the use of psychoacoustic scores supervision versus fully supervised, shared score regression network between modalities, features aggregation before transformer, audiovisual vs. visual-only, and other factors. Finally, we provide a discussion of the results and their implications.

6.1 Experimental Setup

In this section, we describe the experimental setup used in this study to evaluate the proposed method for video summarization. We begin by describing the datasets used, which include SumMe, TVSum, and our proposed AVSum dataset. We then provide details on the evaluation protocol, including the use of the canonical setting when comparing to baselines, the use of the F-score as the evaluation metric, and the process for calculating the metrics for videos with multiple user-annotated summaries. Finally, we provide implementation details, including information on how the pseudo-labels were extracted, the resampling of audio, and the models used to extract features.

6.1.1 Datasets

In this work, we used three video summarization datasets, described in Table 6.1 and illustrated in Figure 6.1: SumMe [29], TVSum [51], and our proposed AVSum. SumMe and TVSum datasets are used to train our audiovisual model and for performance evaluation, as well as ablation studies. Our AVSum dataset is used in this work as a benchmark for the transfer learning setting analysis.

- **SumMe** comprises 25 videos, with an average video duration of 2m40s, and a diverse set of categories, e.g., sports, leisure, and travel. For each video, there are 15 to 18 human annotations on a segment level. The frame-level scores are obtained by averaging the segment-level annotations.
- **TVSum** is composed of 50 videos from YouTube[1]. The videos are distributed among a wide range of genres, e.g., documentaries, news, and historical lectures, and their duration varies from 1 to 4 minutes long. The dataset has frame-level importance scores which can be computed by averaging across 20 users' frame-level annotations.
- **AVSum** dataset is composed of 27 videos from YouTube. All videos have well-defined user-curated auditory events, and their duration varies from 2 to 10 minutes long. The frame-level importance scores were obtained by collecting data from the new (at the time of publication) YouTube heatmap, or *Most replayed*, feature which conveys the information about which segments of the video are being most watched by the platform's users.

[1] http://www.youtube.com

Table 6.1: **Datasets used in this thesis' experiments**.

Dataset	SumMe [53]	TVSum [51]	AVSum
# of videos	25	50	27
duration (min)	1 - 6	2 - 10	2-10
content	holidays, events, sports	news, how-to's, user-generated, documentaries	walking tours
type of annotations	multiple sets of key-fragments	multiple fragment-level scores	frame-level scores
# of annotators per video	15 - 18	20	-

6.1.2 Evaluation Protocol

In this study, we evaluate the performance of the proposed method using a standard approach. Specifically, we randomly divide the dataset into five parts, using four parts for training and one part for evaluation. The evaluation metric used is the F-score, which is a measure of the similarity between the generated summaries and the ground-truth summaries. This metric is computed by considering the precision and recall of the temporal overlap between the two.

We use the F-score to evaluate the similarity between the generated summaries (S_i) and the ground-truth summaries (S_i^*) for each video (i). The precision and recall are determined by calculating the temporal overlap between S_i and S_i^*:

$$Precision = \frac{|S_i \cap S_i^*|}{|S_i|}, \quad Recall = \frac{|S_i \cap S_i^*|}{|S_i^*|},$$

and the F-score is calculated by

$$F{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall}.$$

For videos that have been summarized by multiple users, we use the same method as described in [59] to compute the metrics.

6.1.3 Implementation Details

The pseudo-labels were extracted using a segment size of $P = 1$, and the audios were resampled to $16Hz$. For extracting the features, we use a pretrained CLIP ViT-B/32 [46] pretrained on ImageNet [16], and an ESResNeXt [28] model pretrained on Audioset [25], for the visual and audio respectively. In the given setup, the audio model is pretrained on Audioset to provide a strong foundation for learning audio representations. Audioset is a large-scale dataset consisting of diverse audio events, which enables the model to capture a wide range of audio features. The original task for which the model is pretrained involves audio event classification, where the goal is to identify the presence of various audio events within a given audio clip. The positional embeddings added to each modality have a maximum sequence length of 5000. The audiovisual transformer encoder has 4 layers with 8 heads each and an embedding size of 512. The score regression network consists of two fully-connected layers, which have input and internal dimensions 512 and output dimension 1 (for regression). The scores are aggregated using $\lambda_{img} = 0.5$

and $\lambda_{aud} = 0.5$. The video summarization training in all setups is made with a learning rate of $5e{-}4$, using the Adam optimizer [32] and a weight decay of $1e{-}6$. The training was done using a *full batch* setup for 1500 epochs. The complete training for one split takes around 6.5 hours on a single NVIDIA Tesla T4 GPU. The KTS algorithm utilizes the frame features extracted from the ViT to efficiently segment the video into meaningful parts. The Knapsack algorithm is applied with a constraint on the maximum weight, set to 15% of the video size.

6.2 Results & Discussion

This section presents and analyzes the results of our proposed approach for video summarization. This section is divided into several sub-sections that cover the following topics: Quantitative Results, Qualitative Results, Ground truth scores and Psychoacoustic Annoyance, and an Ablation Study. In the first sub-section, we will present the quantitative results of our approach, comparing it with state-of-the-art methods for both unsupervised and supervised video summarization. In the second sub-section, we will provide qualitative results, demonstrating the effectiveness of our approach on a variety of video samples. In the third sub-section, we perform the experiments using our dataset in the zero-shot scenario. In the fourth sub-section, we will analyze the relationship between the psychoacoustic annoyance and the annotated ground-truth scores. The last sub-section will present an ablation study, evaluating the impact of different design choices and features on the performance of our approach.

6.2.1 Quantitative Results

Unsupervised Video Summarization. We compare our method to unsupervised video summarization baselines on SumMe and TVSum in their canonical configuration [59]. Table 6.2 shows that our method outperforms all unsupervised baselines in the SumMe dataset while maintaining competitive results in TVSum. It is important to emphasize that CLIP-It!, which has a strong performance in both datasets, is also multi-modal, utilizing textual information instead of sound. This can be an indication that with the increasing availability of multi-modal data, methods that can efficiently leverage this incoming data will thrive. Our method achieved an F1 Score of 52.6 on the SumMe

Table 6.2: **Unsupervised**. Our complete model (*Ours*) achieves overall better results compared with other alternatives (best in bold).

Method	SumMe		TVSum		Avg. Rank
	F1	Rank	F1	Rank	
DR-DSN [61]	41.4	11	57.6	10	10.5
Cycle-SUM [58]	41.9	10	57.6	10	10
SUM-GAN-sl [7]	47.8	8	58.4	8	8
SUM-GAN-AAE [3]	48.9	7	58.3	9	8
DSAVS [60]	47.0	9	59.4	6	7.5
SumGraph [43]	49.8	6	59.3	5	5.5
CSNet [31]	51.3	3	58.8	7	5
AC-SUM-GAN [2]	50.8	5	60.6	4	4.5
CA-SUM [6]	51.1	4	61.4	2	3
CLIP-It! [41]	52.5	2	**63.0**	**1**	**1.5**
Ours	**52.6**	**1**	61.2	3	2

Table 6.3: **Supervised Quantitative Results**. Our complete model (*Ours*) achieves overall better results compared with other alternatives (best in bold).

Method	SumMe		TVSum		Avg. Rank
	F1	Rank	F1	Rank	
CSNet [31]	48.6	10	58.5	12	11
VASNet [20]	48.0	11	59.8	10	10.5
DSAVS [60]	48.9	9	59.8	10	9.5
DSNet [63]	50.2	8	62.1	7	7.5
MSVA [26]	53.4	5	61.5	8	6.5
MAVS [21]	44.4	12	**66.8**	**1**	6.5
PGL-SUM [5]	55.6	3	61.0	9	6
MC-VSA [34]	51.6	7	63.7	4	5.5
RR-STG [62]	53.4	5	63.0	5	5
CLIP-It! [41]	54.2	4	66.3	2	3
SMN [55]	**58.3**	**1**	64.5	3	**2**
Ours	56.7	2	62.5	6	4

dataset and of 61.2 on the TVSum dataset.

Supervised Video Summarization. To assess the capabilities of our proposed model, we also evaluate it in a fully-supervised setting against the SOTA methods in the literature. In this setting, human-provided ground truth labels for frame importances are used to train the model, replacing our self-generated pseudo-labels. We observe that our method shows competitive performance against more complex architectures and solutions. Table 6.3 shows that the method SMN performed the best overall, achieving the highest F1 score on both SumMe and TVSum datasets. CLIP-It! also performed well, achieving the second-highest average F1 rank. The method proposed in this thesis also performed well, achieving the second-highest F1 rank on SumMe and the sixth-highest on TVSum, resulting in an average of 4th rank. It is important to notice that the best-performing method, SMN, performed poorly on the SumMe dataset compared to the TVSum dataset. This suggests that the method is heavily tuned for the TVSum dataset and may not perform as well on other datasets. Our proposed method, on the other hand, achieves a balanced performance on both datasets, which suggests that it has a more generalizable and robust approach to video summarization.

6.2.2 Qualitative Results

In Figure 6.2(c), we can see that our self-supervised model had a significant boost in performance, as measured by the F1 score. This suggests that the pseudo-labels generated from the audio-psychoacoustic features were effective in training the model to identify important frames. In Figure 6.2(d), both supervised and self-supervised models had similar F1 scores, but our self-supervised model had a better qualitative result when visually inspecting the generated summary. This demonstrates the effectiveness of our approach in producing high-quality video summaries without the need for costly and time-consuming human annotation.

6.2.3 Zero-shot Results

In this section, we present our evaluation results for the zero-shot scenario in which we utilize pretrained models from available codebases to evaluate their performance on our Audiovisual Summarization Dataset (AVSum). Our aim is to compare our audiovisual model to state-of-the-art models and to demonstrate the effectiveness of our dataset in evaluating video summarization models.

In this zero-shot evaluation, all models were evaluated using the unsupervised pre-training approach. This was done to ensure a fair comparison between our proposed method and the pretrained competitors, as all models were evaluated under the same conditions. By utilizing the unsupervised pre-training, we were able to provide a comprehensive evaluation of each model's performance, highlighting its strengths and limitations. The results of this evaluation, presented in the form of F1 scores, provide valuable insights into the effectiveness of each model in a zero-shot scenario.

We selected a set of state-of-the-art video summarization models and evaluated their performance on the AVSum dataset. The results of our zero-shot evaluation showed that our AVSum dataset follows a similar distribution to well-established datasets such as SumMe and TVSum, but with a stronger correlation between audio and visual information. The results are shown in Table 6.4.

Our audiovisual model outperformed the pretrained competitors, demonstrating the effectiveness of our approach and the importance of incorporating both audio and visual information in video summarization models. These results highlight the significance of our contribution to the field of video summarization and the potential for our audiovisual model to be applied in real-world scenarios.

In conclusion, the zero-shot evaluation results reinforce the importance of incorporating both audio and visual information in video summarization models and demonstrate the effectiveness of our audiovisual model and the AVSum dataset in evaluating video summarization performance.

6.2.4 Ground truth scores and the Psychoacoustic Annoyance

In this section, we present a comprehensive analysis of the relationship between Psychoacoustic Annoyance (PA) and the annotated ground-truth scores. This analysis is based on the examination of data from the SumMe and TVSum datasets. Our findings indicate that there is a significant correlation between the PA and the ground-truth scores, suggesting that the PA can be used as an effective feature for predicting human attention. Furthermore, we also present a comparison between the PA and the sound pressure level (SPL) to evaluate the relative importance of these features in predicting human attention. The results of this comparison provide valuable insights into the relationship between psychoacoustic features and human attention and can inform the development of models

Table 6.4: **Zero-shot results on the AVSum dataset.** The evaluation of the pretrained video summarization methods was conducted using a zero-shot setting, in which the models were pre-trained on each split of the SumMe and TVSum datasets and then evaluated on the complete AVSum dataset. The reported F1 scores are the average values obtained across all models for each dataset.

Method	Pretrained on	
	SumMe	TVSum
CSNet [31]	50.43	48.35
CA-SUM [6]	50.99	50.07
Ours	**52.16**	**51.34**

for predicting attention in real-world scenarios.

Figure 6.3 illustrates the relationship between the ground-truth scores and the psychoacoustic annoyance for three videos of the SumMe dataset. In Figure 6.3-(a), it is possible to observe that the events that have a strong auditory annoyance signal ended up catching the attention of the human labelers. However, in the video portrayed in Figure 6.3-(b), kids playing in leaves, the audio does not provide much information regarding the relevant frames in the first high-importance segment. In this case, the pseudo-label is more likely to generate noise in the training process. In all videos from Figure 6.3, we also plotted the normalized sound pressure level (SPL) to investigate whether the psychoacoustic was indeed a better proxy for the user's attention than a simple sound measurement. Interestingly, we see that, although there are some overlaps in the three signals, the correlation between PA and GT is significantly stronger than between the other two pairs ({GT, SPL} and {SPL, PA}).

The plot presented in the Fig. 6.4, extracted and adapted from Chen *et al.* [12], illustrates the relationship between psychoacoustic features and EEG sub-bands associated with human behavior. Specifically, the figure shows the correlation between the psychoacoustic features of Psychoacoustic Annoyance (PA), Sound Pressure Level (SPL), Loudness, Sharpness, and Roughness, and the EEG sub-bands associated with attention (2), drowsiness (4), and attention marker (7). The data presented in the figure suggests that there is a correlation between these psychoacoustic features and EEG sub-bands, which implies that they can be used as markers for human attention and drowsiness. The methodology used for extracting these features and sub-bands is described in chapter 4. This figure illustrates the potential of using psychoacoustic features as a tool for understanding human behavior.

This relationship between psychoacoustic features and human behavior can be used to develop models that can predict attention levels or drowsiness in real-time applications such as driver assistance systems, human-computer interaction, and others. The ability to predict attention levels can be used to improve human-computer interaction, for example, by adapting the interface or the content to the user's attentional state. Additionally, the ability to predict drowsiness can be used in driver assistance systems to prevent accidents caused by drowsy driving.

To further investigate the relationship between PA, SPL, and ground-truth scores, we present Figure 6.5. The figure illustrates the correlation values between the psychoacoustic feature of PA and SPL and the ground-truth scores for each video in the SumMe and TVSum datasets. The data in the plots show the correlation values for each video, where the x-axis represents the video name and the y-axis represents the correlation values. The dashed lines in each plot represent the average correlation values for each dataset.

In the SumMe dataset (Figure 6.5a), we see that that are some videos in which the difference between PA and SPL is very pronounced such as *Air_Force_One*, *Saving_dolphines*, and *paluma_jump*. This is likely due to the nature of the activities depicted in these videos, which are characterized by sudden and sharp sounds that are associated with higher PA values. On the other hand, for some other videos, such as *Paintball* and *cooking*, it is the other way around; the SPL has a higher correlation. This is intrinsically connected to the type of activity, where these videos are characterized by a more continuous and moderate sound that is associated with higher SPL values.

In the TVSum dataset (Figure 6.5b), the difference between PA and SPL correlations seems smaller and more variable, although there are some examples in which the PA correlation is higher than any of SumMe videos, such as *uGu_10sucQo* or *xmEERLqJ2kU*. This smaller difference can explain why we perform worse in the unsupervised setting in the TVSum dataset. This can be attributed to the fact that the TVSum dataset contains a more diverse range of videos and activities as described in Table 6.1, which results in a more varied distribution of PA and SPL values.

In the AVSum dataset (Figure 6.5c), the difference between the correlation of SPL and PA is noticeable. This is likely due to the bias of the data collection process, as the videos chosen to compose the dataset were selected for having a strong audiovisual signals correlation. This is evident in strong examples such as *video_7*, *video_3*, and *video_11*. These videos exhibit a marked difference in the correlation between SPL and PA, which highlights the importance of incorporating both audio and visual information in video summarization models.

The two videos in SumMe that have zero correlation (*Scuba* and *St_Marteen_Landing*) have zeroed values because their audio signal was not available in the dataset.

6.2.5 Ablation Study

We evaluate the impact of the different components of our proposed method using the canonical training and evaluation procedure from the SumMe dataset. First, we discuss the impact of having pseudo-labels versus a fully human-annotated set of scores. Then, we proceed with analyzing the modules using the fully-supervised setting. Table 6.5 shows the results after removing each of the following components:

Psychoacoustic scores supervision versus fully supervised. As mentioned in the discussion of Figure 6.3, one of the key points of this work is understanding the benefits of using psychoacoustic features as pseudo-labels in a setting where manually annotating video frames can be rather costly. In this experiment, we compare the performance

Table 6.5: **Ablation study**. Our complete model (*Ours*) achieves overall better results compared with other alternatives.

Method	Supervision	Modality	F1(↑)
Randomly Generated Scores + Knapsack	None	-	41.43
Small Transformer + Random Inputs	Full	-	44.23
Small Transformer + Audio Features	Full	A	46.78
Small Transformer + GoogleNet	Full	V	51.16
Transformer + GoogleNet	Full	V	53.87
Transformer + CLIP	Full	V	55.27
Transformer + CLIP + Concat. features pre-transformer	Full	AV	52.16
Transformer + CLIP + Separate Score Regressor	Full	AV	53.49
Ours (Transformer + CLIP + Shared Score Regressor)	Full	AV	56.71
Ours (Transformer + CLIP + Shared Score Regressor)	PA	AV	52.58

between the two modes of supervision. We have seen that although the psychoacoustic features can help guide the model predictions, in some cases, the video's content is not intrinsically auditory. The inferior model's performance in the unsupervised setting can be justified by the nature of these pseudo-labels.

Shared score regression network between modalities. As described in Chapter 4, the score regression network shares its weights between the two modalities. To understand the influence this architectural decision has, we compare it to a model with specific score regression networks for each modality. This change yields a 3.22 drop in the F1 score. Our understanding is that since we are predicting modality-specific scores and aggregating them afterward, the shared MLP network enforces a stronger interaction in the backpropagation path between audio and frame features. This can result in a quicker convergence of the model training.

Features aggregation before the transformer. We understand that the size of a sequence that is fed to a transformer model can significantly influence its efficiency due to its quadratic complexity. For this reason, we investigate the need to have a larger sequence composed of two modalities versus aggregating the audio and frame feature vectors before being fed to the transformer. For aggregation, we used average pooling between the features. The model dropped from a score of 53.49 to 52.16 in our experiment. In a similar rationale as the last subsection, the way both modalities interact with each other can impact strongly on performance.

Audiovisual vs. visual-only. As we are one of the few utilizing more than visual features to perform video summarization, part of the objective of this work is to understand the impact auditory information can bring to a video summarization method. We see that there is a 1.44% improvement by adding the audio information in our best configuration (comparing lines 6 and 9 of Table 6.5).

CLIP features. As previously shown by Medhini *et al.* [41], the feature extractor used can have a significant impact on the performance of a video summarization method. In our experiments, we observed a similar effect when switching from using CLIP features to GoogLeNet features. This resulted in a drop of 1.40% in the F1 Score performance of our method.

Transformer encoder size. In our experiments, we trained two transformer models with different sizes and found that there was only a small difference in their performance. The larger model had a slightly higher F1 score of 53.87% than the smaller model. This suggests that the size of the transformer encoder may be an important factor in the performance of our method, but further research is needed to confirm this. Our experiments show that our proposed method is a strong baseline for video summarization and has the potential to improve the performance of existing approaches in the field.

Audio-only Features. In this experiment, we replaced the visual features with audio-only features to evaluate the contribution of each modality individually. Our objective was to investigate if the audio was enough information to perform competitively with other video summarization methods. However, our results show that while the audio features can be of great help as additional information, removing the visual signal altogether resulted in a significant drop in performance, from 51.16 to 46.78 in the F1 score. This highlights the importance of considering both modalities in video summarization and reinforces that visual information is currently the main modality for this task.

Random Inputs. To understand the role of the features being fed to the transformer model in our method, we fed the same transformer architecture with random inputs instead. Our results suggest that even with random inputs, the transformer network is able to learn some features, such as the common regions of all videos that are most likely to be highlighted and similar patterns. This is an indication that the transformer can extract meaningful information from the task supervision itself, even when the input is not informative. On the other hand, we see that this implicit information is not enough to perform well on the datasets we used in this work. Using random inputs yielded a drop in F1 score of 6.93 points when compared to the GoogleNet visual features and a drop of 2.55 compared to the audio features. As we will see in the next experiment, this robustness to random inputs when compared to the audio features input can be explained by the reduction of the optimization space that the *frame scores to video summary* method, described in Section 2.1.3, intrinsically performs when transitioning from frame-level to segment-level comparisons.

Random Scores. In this experiment, we generated random scores and applied the KTS and Knapsack methods to them to evaluate the impact of the KTS and Knapsack methods on the final summary. Our results show that even when using random scores, the KTS method is able to segment the video and generate probable meaningful summaries. This highlights the importance of the KTS and Knapsack methods in our proposed method and the potential of using visual features alone for video summarization.

Discussion. In conclusion, our results indicate that our proposed method, which utilizes both visual and audio information, outperforms its alternatives that use only visual information. Furthermore, the shared score regression network and feature aggregation before transformer architecture choices have a positive impact on the performance of our method. The use of CLIP features also had a significant impact on the performance of our method, while the transformer encoder size had only a small effect. Overall, our experiments demonstrate the effectiveness of our proposed method in video summarization and highlight the importance of considering both visual and audio information in this task. Further research is needed to further improve the performance of our proposed method and explore the use of psychoacoustic features in other areas of video processing.

Figure 6.1: **Frame samples of the datasets used in this thesis.** The examples from the SumMe dataset include sporting and culinary events, while the TVSum dataset encompasses a diverse range of subjects, including sports and dog competitions. The AVSum dataset primarily consists of walking videos from various cities, showcasing a diverse set of urban landscapes.

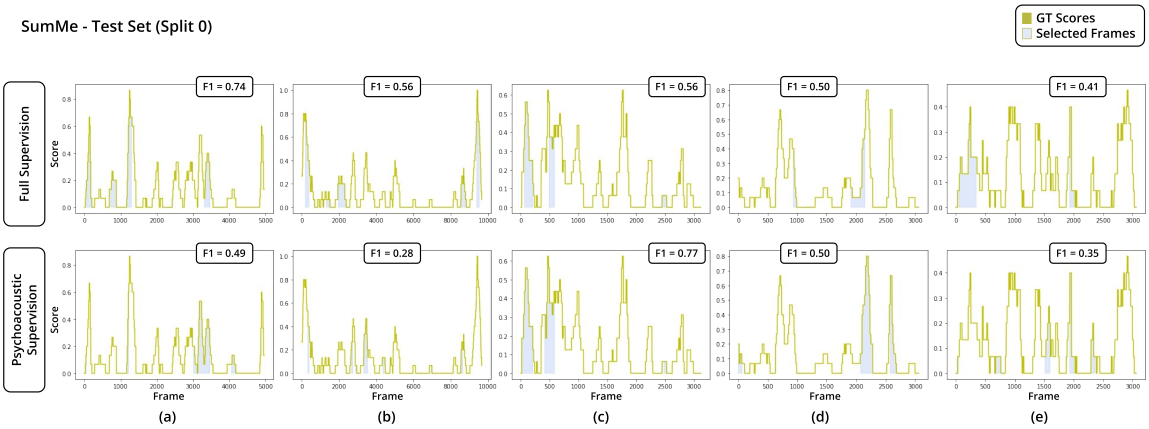
SumMe

TVSum

AVSum

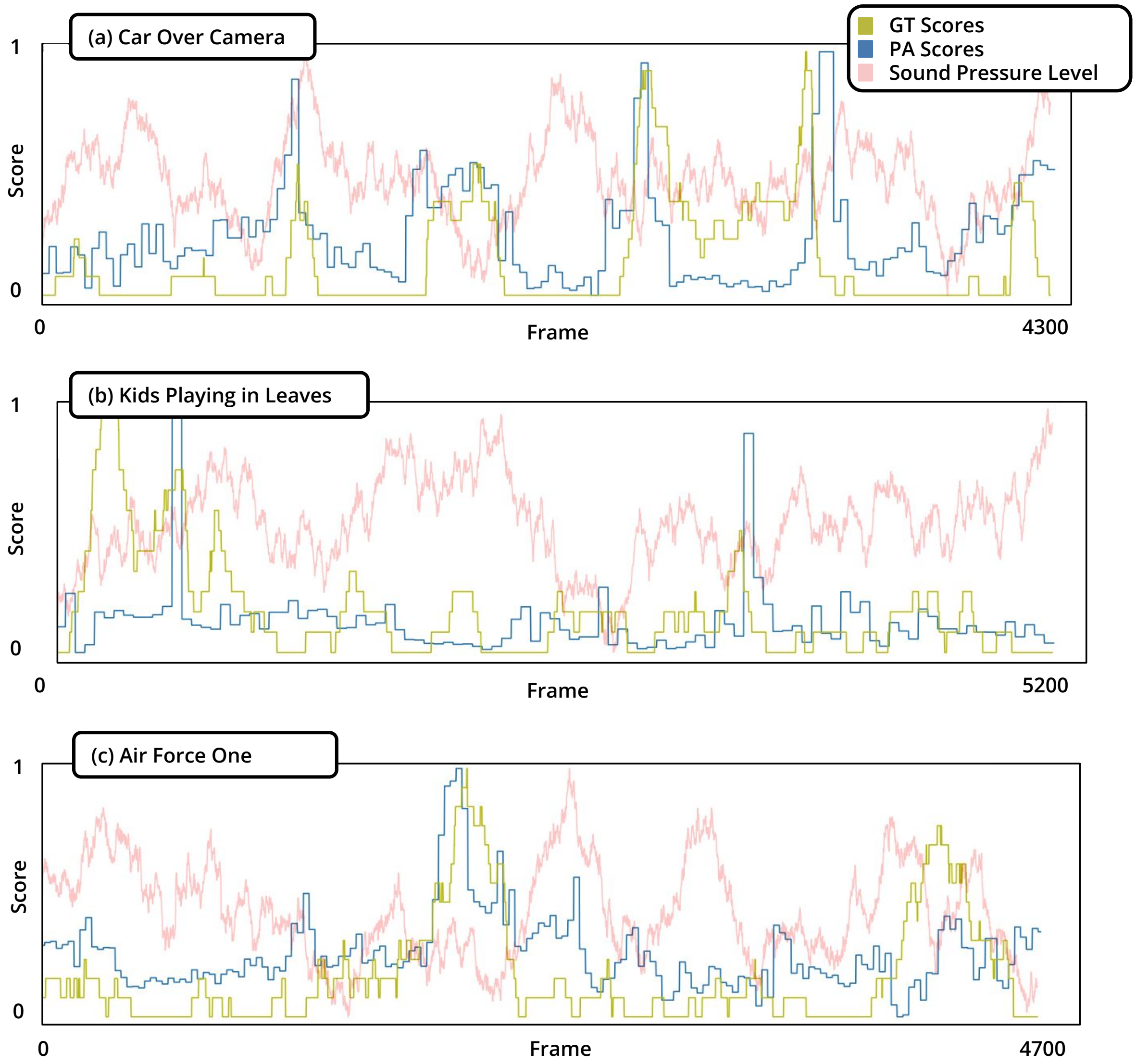
Source: Created by the author.

Figure 6.2: **Qualitative analysis on the test set of the SumMe dataset's *(split0).*** *Best viewed with zoom.* As expected, there is a drop in performance when switching from the supervised to the unsupervised setting. On videos (c) and (d), our unsupervised model outperforms the supervised model in terms of F1 score, with a significant boost on video (c). In addition, the unsupervised model also produces a better qualitative result on video (d).



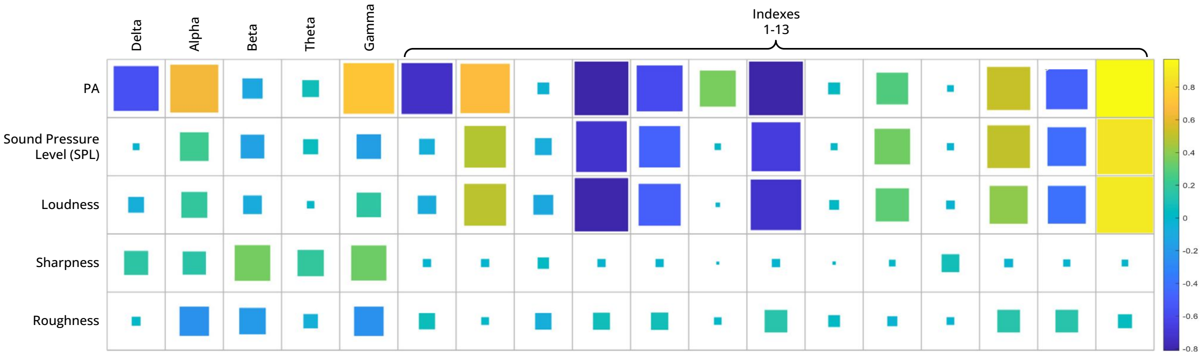
Source: Created by the author.

Figure 6.3: **Ground truth scores and the Psychoacoustic Annoyance.** Video samples from the SumMe dataset illustrate what was observed from the premise of linking human attention to psychoacoustic features, such as the PA. The PA scores (in blue) were extracted using the methodology described in Chapter 4 and are the same used for training in the self-supervised setting. Normalized sound pressure levels (in red) were extracted to help evaluate the influence of psychoacoustics when estimating frame importance scores.



Source: Created by the author.

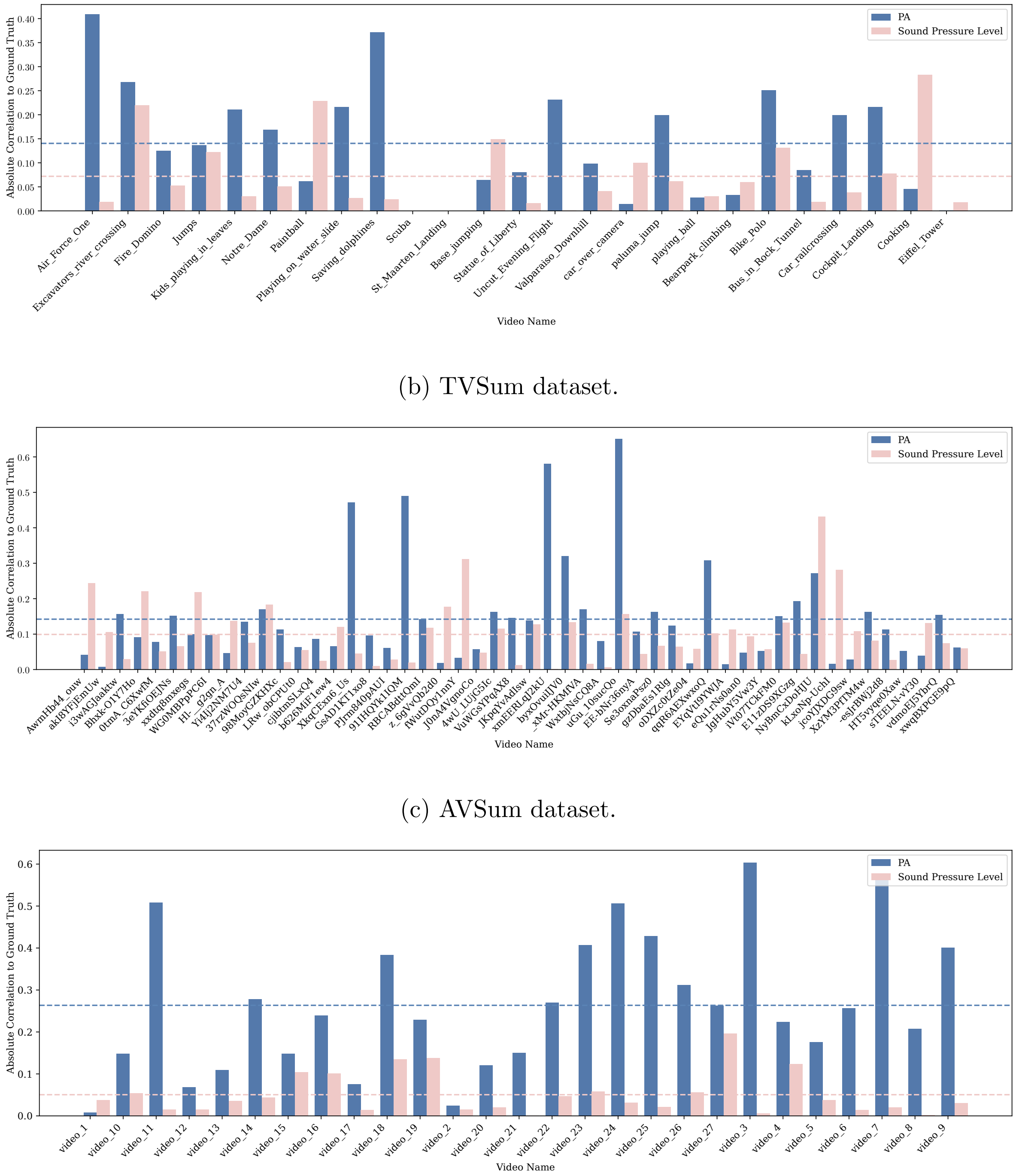
Figure 6.4: **Relation between psychoacoustic features and EEG sub-bands.** The figure illustrates the relationship between the psychoacoustic features of Psychoacoustic Annoyance (PA), Sound Pressure Level (SPL), Loudness, Sharpness, and Roughness, and the EEG sub-bands associated with human behavior such as attention (*index 2*), drowsiness (*index 4*), and attention marker (*index 7*). The data shows that there is a correlation between these psychoacoustic features and EEG sub-bands, suggesting that they can be used as markers for human attention and drowsiness. This figure illustrates the potential of using psychoacoustic features as a tool for understanding human behavior.



Source: Figure and caption adapted from Chen *et al.* [12].

Figure 6.5: **Correlation between psychoacoustic features and ground-truth scores for the SumMe, TVSum, and AVSum datasets.** This figure illustrates the correlation values between the psychoacoustic feature of PA and SPL and the ground-truth scores for each video in the (a) SumMe, (b) TVSum, and (c) AVSum datasets. The data in the plots show the correlation values for each video, where the x-axis represents the video name and the y-axis represents the correlation values. The dashed lines in each plot represent the average correlation values for each dataset. The comparison within each dataset allows us to evaluate the consistency of the relationship between PA and SPL with ground-truth scores and the generalizability of the results across different video contents. The methodology used for extracting these features and ground-truth scores is described in Chapter 4.

(a) SumMe dataset.



(b) TVSum dataset.

(c) AVSum dataset.

Source: Created by the author.

Chapter 7

Conclusion

In this thesis, we have presented a comprehensive study on the utilization of audiovisual information for video summarization. The exponential growth of multimedia data has led to increasing demand for effective and efficient summarization techniques. In this context, our proposed approach leverages both auditory and visual information to generate high-quality video summaries. Our method takes advantage of the correlation between audiovisual signals in videos, which is widely and naturally available in most scenarios, to enhance the performance of video summarization. We first presented a novel audiovisual model for video summarization that incorporates psychoacoustic features as pseudo-labels to help the model learn the most relevant parts of the videos without the need for human-labeled data. Our proposed model is based on a transformer architecture, which demonstrates significant improvement compared to state-of-the-art methods in the unsupervised setting on the SumMe dataset. Furthermore, we also conducted thorough ablation studies to evaluate the contributions of each component of our method. In addition to the proposed model, we also introduced a new audiovisual video summarization dataset called AVSum, which contains a significant amount of untrimmed videos showcasing a range of auditory stimuli. This new dataset provides a diverse and challenging benchmark for future video summarization research and highlights the need for more sophisticated models that can effectively leverage audiovisual information. Finally, our results show that the incorporation of psychoacoustic features can significantly improve the performance of video summarization models. Our experiments demonstrate that our method is a promising approach for generating accurate and effective summaries of videos, providing a strong foundation for future work in this area. The proposed model and dataset are of great significance for the field of video summarization and will have a lasting impact on future research in this area.

7.1 Limitations & Future Work

One possible direction for future work is to explore different methods of calculating the psychoacoustic measure, such as Zwicker's improved models proposed by Di *et al.* [17] and neural networks to calculate the PA values from the audio as proposed by Lopez *et al.* [35]. Another avenue we believe is worth exploring is the use of other emotion-based metrics, as they have been shown to be effective in similar applications such as Semantic Hyperlapse [14, 15]. Despite the promising results obtained by our model, there are some limitations that we identified during our experimentation. Firstly, the model did not converge using the auxiliary losses as used in Narasimhan *et al.* [41]. In order to resolve this issue, alternative auxiliary loss functions could be explored to stabilize the training process. Secondly, our performance on the TVSum dataset was not as good in the unsupervised setting due to the lower correlation between the psychoacoustic measure and the ground-truth labels. To address this, we could investigate methods for better aligning the psychoacoustic measure with the ground truth or using additional modalities in the self-supervision process. Lastly, it is worth mentioning that the self-supervision using psychoacoustic features may not be effective in datasets with no auditory stimuli. In these cases, alternative self-supervision techniques or a fully-supervised approach may be necessary. By expanding upon these methods, we hope that these future directions will improve the overall performance of our system and bring it closer to real-world applications.

Bibliography

[1] S.M. Abel. The extra-auditory effects of noise and annoyance: An overview of research. *The Journal of otolaryngology*, 19 Suppl 1:1–13, 1990.

[2] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. Ac-sum-gan: Connecting actor-critic and generative adversarial networks for unsupervised video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8):3278–3292, 2020.

[3] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. Unsupervised video summarization via attention-driven adversarial learning. In *International Conference on multimedia modeling*, pages 492–504. Springer, 2020.

[4] Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109(11):1838–1863, 2021.

[5] Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. Combining global and local attention with positional encoding for video summarization. In *2021 IEEE International Symposium on Multimedia (ISM)*, pages 226–234. IEEE, 2021.

[6] Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. Summarizing videos using concentrated attention and considering the uniqueness and diversity of the video frames. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, pages 407–415, 2022.

[7] Evlampios Apostolidis, Alexandros I Metsai, Eleni Adamantidou, Vasileios Mezaris, and Ioannis Patras. A stepwise, label-based approach for improving the adversarial training in unsupervised video summarization. In *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, pages 17–25, 2019.

[8] W. Babisch et al. Health status as a potential effect modifier of the relation between noise annoyance and incidence of ischaemic heart disease. *Occupational and Environmental Medicine*, 60(10):739–745, 2003.

[9] W Babisch, H Fromme, A Beyer, and H Ising. Increased catecholamine levels in urine in subjects exposed to road traffic noise: the role of stress hormones in noise research. *Environment international*, 26(7-8):475–481, 2001.

[10] Wolfgang Babisch, Göran Pershagen, Jenny Selander, Danny Houthuijs, Oscar Breugelmans, Ennio Cadum, Federica Vigna-Taglianti, Klea Katsouyanni, Alexandros S Haralabidis, Konstantina Dimakopoulou, et al. Noise annoyance—a modifier of the association between noise level and cardiovascular health? *Science of the total environment*, 452:50–57, 2013.

[11] Bor-Chun Chen, Yan-Ying Chen, and Francine Chen. Video to text summary: Joint video summarization and captioning with recurrent neural networks. In *Bmvc*, 2017.

[12] Xieqi Chen, Jianhui Lin, Hang Jin, Yan Huang, and Zechao Liu. The psychoacoustics annoyance research based on eeg rhythms for passengers in high-speed railway. *Applied Acoustics*, 171:107575, 2021.

[13] RH Chowns, I Abey-Wickrama, MF A'Brook, FEG Gattoni, and CF Herridge. Mental-hospital admissions and aircraft noise. *The Lancet*, 295(7644):467–468, 1970.

[14] Diognei de Matos, Washington Ramos, Luiz Romanhol, and Erickson R Nascimento. Musical hyperlapse: A multimodal approach to accelerate first-person videos. In *2021 34th SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 184–191. IEEE, 2021.

[15] Diognei de Matos, Washington Ramos, Michel Silva, Luiz Romanhol, and Erickson R Nascimento. A multimodal hyperlapse method based on video and songs' emotion alignment. *Pattern Recognition Letters*, 2022.

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[17] Guo-Qing Di, Xing-Wang Chen, Kai Song, Bing Zhou, and Chun-Ming Pei. Improvement of zwicker's psychoacoustic annoyance model aiming at tonal noises. *Applied Acoustics*, 105:164–170, 2016.

[18] I. Enmarker and E. Boman. Noise annoyance responses of middle school pupils and teachers. *Journal of Environmental Psychology*, 24(4):527 – 536, 2004.

[19] Charlotta Eriksson, Mats Rosenlund, Göran Pershagen, Agneta Hilding, Claes-Göran Östenson, and Gösta Bluhm. Aircraft noise and incidence of hypertension. *Epidemiology*, pages 716–721, 2007.

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

[21] Litong Feng, Ziyin Li, Zhanghui Kuang, and Wei Zhang. Extractive video summarizer with memory augmented neural networks. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 976–983, 2018.

[22] Maria Foraster, Ikenna C Eze, Danielle Vienneau, Mark Brink, Christian Cajochen, Seraina Caviezel, Harris Héritier, Emmanuel Schaffner, Christian Schindler, Miriam Wanner, et al. Long-term transportation noise annoyance is associated with subsequent lower levels of physical activity. *Environment international*, 91:341–349, 2016.

[23] Beate Fruhstorfer, Heinrich Fruhstorfer, and Peter Grass. Daytime noise and subsequent night sleep in man. *European journal of applied physiology and occupational physiology*, 53(2):159–163, 1984.

[24] Vinicius S. Furlan, Ruzena Bajcsy, and Erickson R. Nascimento. Fast forwarding egocentric videos by listening and watching. In *In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Sight and Sound*, page 2504–2507. IEEE Computer Society, 2018.

[25] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[26] Junaid Ahmed Ghauri, Sherzod Hakimov, and Ralph Ewerth. Supervised video summarization via multiple feature sets with parallel attention. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6s. IEEE, 2021.

[27] Rainer Guski, Dirk Schreckenberg, and Rudolf Schuemer. Who environmental noise guidelines for the european region: A systematic review on environmental noise and annoyance. *International journal of environmental research and public health*, 14(12):1539, 2017.

[28] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Esresne (x) t-fbsp: Learning robust time-frequency transformation of audio. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

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

[30] Min Hu, Ruimin Hu, Zhongyuan Wang, Zixiang Xiong, and Rui Zhong. Spatiotemporal two-stream lstm network for unsupervised video summarization. *Multimedia Tools and Applications*, pages 1–22, 2022.

[31] 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.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] Yong Jae Lee, Joydeep Ghosh, and Kristen Grauman. Discovering important people and objects for egocentric video summarization. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1346–1353. IEEE, 2012.

[34] Yen-Ting Liu, Yu-Jhe Li, and Yu-Chiang Frank Wang. Transforming multi-concept attention into video summarization. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[35] Jesus Lopez-Ballester, Adolfo Pastor-Aparicio, Jaume Segura-Garcia, Santiago Felici-Castell, and Maximo Cobos. Computation of psycho-acoustic annoyance using deep neural networks. *Applied Sciences*, 9(15):3136, 2019.

[36] 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.

[37] Sascha E Martin, Peter K Wraith, Ian J Deary, and Neil J Douglas. The effect of nonvisible sleep fragmentation on daytime function. *American journal of respiratory and critical care medicine*, 155(5):1596–1601, 1997.

[38] R Maynard, B Berry, IH Flindell, G Leventhall, B Shield, A Smith, and S Stansfield. Environmental noise and health in the uk: A report by the ad hoc expert group on noise and health. 2010.

[39] Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Which training methods for gans do actually converge? In *International Conference on Machine Learning (ICML)*, 2018.

[40] Medhini Narasimhan, Arsha Nagrani, Chen Sun, Michael Rubinstein, Trevor Darrell, Anna Rohrbach, and Cordelia Schmid. Tl; dw? summarizing instructional videos with task relevance and cross-modal saliency. In *European Conference on Computer Vision*, pages 540–557. Springer, 2022.

[41] Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in Neural Information Processing Systems*, 34:13988–14000, 2021.

[42] World Health Organization et al. *Burden of disease from environmental noise: Quantification of healthy life years lost in Europe.* World Health Organization. Regional Office for Europe, 2011.

[43] Jungin Park, Jiyoung Lee, Ig-Jae Kim, and Kwanghoon Sohn. Sumgraph: Video summarization via recursive graph modeling. In *European Conference on Computer Vision*, pages 647–663. Springer, 2020.

[44] Bryan A Plummer, Matthew Brown, and Svetlana Lazebnik. Enhancing video summarization via vision-language embedding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5781–5789, 2017.

[45] Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*, pages 540–555. Springer, 2014.

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[47] Washington LS Ramos, Michel M Silva, Mario FM Campos, and Erickson R Nascimento. Fast-forward video based on semantic extraction. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 3334–3338. IEEE, 2016.

[48] R. Rylander. Physiological aspects of noise-induced stress and annoyance. *Journal of Sound and Vibration*, 277(3):471 – 478, 2004. Fifth Japanese-Swedish Noise Symposium on Medical Effects.

[49] T. Saeki et al. Effects of acoustical noise on annoyance, performance and fatigue during mental memory task. *Applied Acoustics*, 65(9):913 – 921, 2004.

[50] Daniela Sammler, Maren Grigutsch, Thomas Fritz, and Stefan Koelsch. Music and emotion: electrophysiological correlates of the processing of pleasant and unpleasant music. *Psychophysiology*, 44(2):293–304, 2007.

[51] 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.

[52] Ba Tu Truong and Svetha Venkatesh. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(1):3–es, feb 2007.

[53] Erik Van der Burg, Christian Olivers, Adelbert Bronkhorst, and Jan Theeuwes. Pip and pop: Nonspatial auditory signals improve spatial visual search. *Journal of experimental psychology. Human perception and performance*, 34:1053–65, 11 2008.

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, undefinedukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[55] Junbo Wang, Wei Wang, Zhiyong Wang, Liang Wang, Dagan Feng, and Tieniu Tan. Stacked memory network for video summarization. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 836–844, 2019.

[56] Robert T Wilkinson and Ken B Campbell. Effects of traffic noise on quality of sleep: assessment by eeg, subjective report, or performance the next day. *The Journal of the Acoustical Society of America*, 75(2):468–475, 1984.

[57] Ting Yao, Tao Mei, and Yong Rui. Highlight detection with pairwise deep ranking for first-person video summarization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 982–990, 2016.

[58] 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.

[59] 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.

[60] Sheng-Hua Zhong, Jingxu Lin, Jianglin Lu, Ahmed Fares, and Tongwei Ren. Deep semantic and attentive network for unsupervised video summarization. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 18(2):1–21, 2022.

[61] 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.

[62] Wencheng Zhu, Yucheng Han, Jiwen Lu, and Jie Zhou. Relational reasoning over spatial-temporal graphs for video summarization. *IEEE Transactions on Image Processing*, 31:3017–3031, 2022.

[63] Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962, 2020.

[64] Eberhard Zwicker and Hugo Fastl. *Psychoacoustics: Facts and models*, volume 22. Springer Science & Business Media, 2013.