Master's Thesis

Multimodal recognition of dog social signals in human-animal interaction with pretrained model from a small-scale dataset

TRAN Anh Tuc

Supervisor OKADA Shogo

Division of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

August, 2025

Abstract

The study of dog social signals in dog-human interaction has significant academic and social importance; however, it remains relatively underexplored. Understanding dogs' emotional states enhances the human-dog connection and brings meaningful, practical benefits. In an effort to support continued exploration in this field, this thesis investigates effective multimodal machine learning approaches for classifying dog emotions based on vocal and behavioral patterns using a small-scale dataset. This study focuses exclusively on YouTube videos where dog social signals are detected and labeled by animal experts. This thesis is conducted using a dataset that includes 7 classes of emotions per voice type.

This research field receives limited research attention, and there are few dog social signal datasets. In many real-world scenarios, such as bioacoustic analysis or emotion classification, labeled datasets are limited and expensive to obtain. Data scarcity is both a motivation and a challenge, particularly for data-driven machine learning. Research on dog emotions mainly focuses on unimodality, such as dog faces, poses, or voices. However, multimodal machine learning, leveraging different modalities that provide additional context, presents an effective method for emotion classification. Even though showing its remarkable ability, multimodal classification tasks often require large volumes of labeled data to effectively train deep learning models. Pretrained models show their powerful ability to present robust representations while addressing limited dataset problems with transfer learning. However, in unexplored research areas such as this one, there are few available in-domain pre-trained models. Despite advances in multimodal classification, current models struggle to effectively classify dog social signal due to limited datasets, background noise, and limited resources.

This research aims to investigate effective methods for modeling dog social signals for classification on small-scale datasets using a multimodal machine learning approach. To achieve this goal, this study first develops a multimodal classification baseline for dog social signals using generic pretrained models as unimodal encoders. Additionally, a joint feature modulation (JFM) is proposed for multimodal fusion, which enables adaptively control the combination of pretrained representations, thereby preserving their rich information and even creating a more informative fused representation, which in turn enhances overall results. Although pre-trained models provide a strong starting point, they often fail to fully capture the domain-specific characteristics of inputs. This leads to suboptimal performance, particularly

in fine-grained classification tasks. To further improve the feature representations, this work proposes an approach to continue pretraining the unimodal encoder. The acoustic encoder is trained on available unlabeled data with initialization from cross-domain pretrained weights, enabling it to capture complex, discriminative acoustic features and make robust representations, thereby mitigating the problem caused by data scarcity and enhancing model robustness.

The multimodal model is built based on feature-level fusion with unimodal encoders. AudioSet pretrained AST (Audio Spectrogram Transformer) is used to encode spectrogram input to acoustic embedding, and S3D pretrained on Howto100M dataset is used to encode video into visual embedding. Further ablation studies are conducted using human speech and image-pretrained models to evaluate the effectiveness of cross-domain transfer learning in improving the audio unimodal feature representation. Data augmentation and PCA visualization will assist in generalization and interpretability. After training, models will be evaluated using accuracy and F1-score.

The expected outcome is this multimodal-based classification model with higher accuracy and robustness compared to the majority, non-parametric handcrafted feature model, and unimodal baselines. Results confirm the effectiveness of this proposed multimodal pretrained model based approach. Moreover, compared to simple early fusion methods, the proposed JFM fusion demonstrates improved performance across both metrics, proving to be more effective in our setting. However, this has not met our expectations. To enhance the acoustic encoder and ultimately improve multimodal performance, this work explores continued pretraining to inject in-domain discriminative features into unimodal embeddings based on cross-domain pretrained models. Although this experiment remains incomplete due to various limitations, it demonstrates strong potential for future development.

In conclusion, these results, achieved with a small and noisy dataset under simple settings and a transfer learning framework, nonetheless highlight the potential of multimodal classification of dog social signals and serve as the baseline for future references on this task. This work aims to contribute a reproducible research pipeline for dog social signal prediction to contribute to human-computer-animal research, hopefully guiding the development of more robust and efficient approaches.

Keywords: Dog social signal, dog emotions classification, multimodal classification, multimodal fusion, cross-domain transfer learning

Contents

1 Introduction **3**
- 1.1 Background . . . 3
- 1.2 Challenges . . . 4
- 1.3 Research Scope and Objective . . . 5
  - 1.3.1 Objectives . . . 5
  - 1.3.2 Scope . . . 5
- 1.4 Thesis organization . . . 6

2 Related work **8**
- 2.1 Multimodal machine learning . . . 8
  - 2.1.1 Concepts . . . 8
  - 2.1.2 Unimodal representations - Audio . . . 9
  - 2.1.3 Unimodal representations - Video . . . 11
  - 2.1.4 Fusion . . . 13
- 2.2 Tranfer learning . . . 13
- 2.3 Dog social signal in human-dog interaction . . . 15
- 2.4 Dog social signal processing . . . 17

3 Dataset **19**
- 3.1 Source dataset selection . . . 19
- 3.2 Dog social signal annotation . . . 19
- 3.3 Support dataset . . . 22
  - 3.3.1 Dog audio from AudioSet dataset . . . 22

4 Methology **23**
- 4.1 Model overview . . . 23
  - 4.1.1 Pretrained acoustic - AST . . . 23
  - 4.1.2 Pretrained video - S3D . . . 27
- 4.2 Feature extraction based on pre-trained model . . . 32
  - 4.2.1 Acoustic feature embedding . . . 32
  - 4.2.2 Visual feature embedding . . . 32

4.3 Multimodal social signal modeling 33

5 Experimentation **34**
5.1 Experiments: Pre-trained baseline for dog social signal classification 34
5.1.1 Baselines 34
5.1.2 Experimental Settings 35
5.1.3 Experimental results 35
5.2 Experiment: Multimodal fusion for effective representation . . 39
5.2.1 Results 41
5.3 Experiments: Improve performance with unimodal feature extractor 42
5.3.1 Model architecture overview 42
5.3.2 Experiment settings 43
5.3.3 Experiment results 44

6 Ablation Study **46**
6.1 Effect of Layer Unfreezing on Model Performance 47
6.1.1 Experiment results 47
6.2 Cross-domain adaptation for acoustic representation learning . 51
6.2.1 Experiment settings 52
6.2.2 Cross-domain transfer via pretrained feature representations results 52

7 Conclusion **56**
7.1 Summary 56
7.2 Contribution 57
7.3 Future work 57

List of Figures

2.1 Waveform, visualization of sound in time domain 9
2.2 (Frequency) Spectrum, visualization of sound in frequency domain 10
2.3 Spectrogram, visual representation of sound 10
2.4 Overview of transfer learning settings [38] 14

4.1 Transformer architecture 24
4.2 The scaled dot-product attention architecture (left) and the multi-head attention mechanism (right) are composed of multiple attention layers that function concurrently 25
4.3 Vision Transformer architecture 25
4.4 Audio Spectrogram Transformer architecture 26
4.5 The I3D architecture based on Inflated Inception-V1 (left) and a detailed view of its inception submodule (right). 28
4.6 Overview of the S3D model architecture (left), with details of the temporally separable convolution (Sep-Conv) module (middle) and the temporally separable inception blocks (Sep-Inc) (right). 29
4.7 MIL-NCE for learning video representations from uncurated datasets. Given a video x and a set of associated positive narration candidates $\mathcal{P}$, MIL-NCE leverages multiple positive pairs—such as $(x, y), (x, y_1), (x, y_2), (x, y_3), (x, y_4)$ (left), which better captures fine-grained object references like 'sander' in (x, y_3) and specific action descriptions in (x, y_4) that standard NCE may miss. This method promotes multiple correct positives while downweighting inaccurate ones using a discriminative ratio against negatives $\mathcal{N}$ (right). 30
4.8 An Overview of multimodal dog social signal classification . . 31

5.1 Comfusion maxtrix from eGeMAPS (acoustic) (a), pretrained acoustic (b), and multimodal(full-vid)(c) from test set. (Label 0: bark/aggression, 1: bark/attention, 2: bark/conflict, 3: whine/anxiety, 4: whine/attention, 5: growl/aggressive, 6: growl/conflict) 39

6.1 Performance across different unfreezing configurations. Subfigures show (a) accuracy and (b) macro F1-score, for each unfreezing setting. Settings range from unfreezing only the classifier layer (leftmost) to unfreezing all layers (rightmost). . 49
6.2 Testing loss across different unfreezing configurations. Settings range from unfreezing only the classifier layer (leftmost) to unfreezing all layers (rightmost). 50
6.3 Visualization from eGeMAPS features embedding and ViT finetuning embedding using PCA(2) 54

List of Tables

1 Glossary of terms and abbreviations (abbr.) 2

3.1 Statistics of the dataset 21
3.2 Statistics of dog voice types from AudioSet 22

5.1 Performance results of dog social signal classification. Numbers in bold are the best metric values recorded, while those underlined are the second-best ones. 36
5.2 Detailed performance results 37
5.3 Comparision results of dog social signal classification amongst differeren fusion methods. Numbers in bold are the best metric values recorded, while those underlined are the second-best ones. 41
5.4 Pretext evaluation of model continue pretraining for domain adaptation 44
5.5 Performance results of downstream dog social signal classification 45

6.1 Accuracy and F1-score results for progressive unfreezing experiments. 48
6.2 Cross domain transfer learning results of Image pretraining model (ViT) and human speech pretraining model (Wav2vec2). Both model's pre-trained weights are frozen for feature extraction and training classifier head for dog social signal recognition. 53
6.3 Dog vocalization results. A classification between bark, whine and growl on ViT model. 53

Glossary

Table 1: Glossary of terms and abbreviations (abbr.)

Term	**Abbr.**	**Definition**
Convolution Neuron Network	CNN	A type of deep neural network
Inflated 3D CNN	I3D	Model architecture, more detail in section 4.1
Separable 3D CNN	S3D	Model architecture, more detail in section 4.1
gated multimodal unit	GMU, gated.	More detail in section 5.2
Vision Transformer	ViT	Model architecture, more detail in section 4.1
Audio Spectrogram Transformer	AST	Model architecture, more detail in section 4.1
Spectrogram	Spec.	Visual representation of sound
Video	Vid.	
Concatenation	concat	
Linear probling	LP	Tranfer learning technique that we frozen pretrained weights and train linear classifier head for downstream dataset
Finetuning	FT	Tranfer learning technique that we train all model weights for downstream dataset

Chapter 1

Introduction

1.1 Background

For centuries, dogs have been more than loyal companions. They have stood by humans by playing vital roles, including military, security, therapy, and medical assistance. Recent studies have also shown that dogs play a positive role in supporting human mental health[1][2]. As our bond with dogs continues to grow, so does our responsibility to support not only their physical needs, but also their mental and emotional well-being. Understanding a dog's emotional state offers more than just insight into their inner world, it strengthens the human-dog relationship and leads to practical benefits. Moreover, it is commonly shared among dog trainers that when a dog exhibits signs of anxiety during training, implementing calming techniques, such as offering breaks, and using a soothing tone of voice, can help reduce stress and build the dog's confidence [56]. Additionally, distress is often expressed through several ways, such as fear and aggressive behaviors, yet many owners frequently overlook the more subtle signs of canine stress, even in their pets [55]. Improved emotional awareness can enhance dog training, support behavior correction, and aid in the early detection of stress or health issues. When machines play animal expert roles, they can help us to automatically understand dogs' emotional state, which not only assists with timely dog care but also reduces the reliance on human experts. These advances directly contribute to better dog welfare, more effective health monitoring, and more humane and adaptive care environments. Beyond its social impact, this

[1] "Research Results of a 40% Reduction in Risk of Developing Dementia in Elderly Dog Owners" (NHK Online, December 31, 2023, 6:25 a.m.)

[2] Taniguchi Y, et al. (2022) Evidence that dog ownership protects against the onset of disability in an older community-dwelling Japanese population. PLoS ONE 17(2): e0263791

work also holds strong academic value. By leveraging state-of-the-art technology such as machine learning to automatically recognize dog's emotions, this study opens up new avenues for research in cross-species communication and contribute to the growth of animal-computer interaction research. These contributions support not only the everyday lives of dog owners and professionals but also the growth of a field that bridges technology and animal welfare in meaningful ways. With this work, these findings serve as a baseline for future research, hopefully guiding the development of more robust and efficient approaches. This thesis aims to contribute to a more sustainable and harmonious future, the one that considers the well-being of all living beings, not just humans, but also our beloved animal companions.

These systems provide practical value and meaningful real-world applications. In today's era of rapidly advancing artificial intelligence, machine learning has shown exceptional capabilities in modeling complex data and delivering increasingly accurate predictions. Such progress is enabling the automation of a wide range of tasks and continues to reveal vast, seemingly limitless potential. Motivated by these developments, this work aims to explore and develop an automatic dog social signal classification system using state-of-the-art machine learning techniques.

1.2 Challenges

Unlike other research fields, this area has received limited attention, resulting in it being underexplored and lacking well-annotated datasets. While studies using large datasets can provide good estimations of real-world scenarios and yield reliable insights and results, the scarcity of data in this field imposes significant limitations. These range from constraints in comprehensive analysis to difficulties in developing effective and generalizable models. This is particularly challenging in the current data-driven machine learning era, where despite achieving remarkable results across diverse tasks and datasets and even with techniques like transfer learning, models trained on small datasets with large domain gaps are prone to severe overfitting and produce untrustworthy results.

Furthermore, in terms of affective expression, movements such as tail and ear movements, as well as posture, are important in the visual modality. However, social signal processing research lacks high-quality datasets for thorough analysis and study. My main data sources, which are online videos, are often of low resolution, unstable camera movements, motion blur, poor lighting conditions, and frequent occlusions. Moreover, these videos typically lack structured recording protocols. Such limitations restrict research efforts

to investigate distinct emotional traits necessary for effective emotion classification and hinder the development of robust feature representation learning methods. Specifically in this scenario, because the extraction rate of dog posture features was below 50%, it was impossible to conduct experiments with these features due to their high rate of detection failure.

Additionally, the species gap poses a significant challenge for researchers, especially those with limited expertise in handling bioacoustic data. It acts as a barrier not only to developing robust models but also to newcomers seeking to enter and make meaningful contribute to the field.

1.3 Research Scope and Objective

1.3.1 Objectives

This research studies to develop an effective dog social signal classification for a small-scaled dataset. To achieve this overall aim, the study focuses on the following objectives:

- Given the lack of prior multimodal studies on this dataset involving dog social signal, this research aims to investigate effective methods for modeling dog social signals using multimodal machine learning techniques that integrate both acoustic and visual modalities for classification.

- Due to the limited size and high noise levels in the available dataset, a common challenge across this field, this research develops efficient transfer learning approaches that leverage generic pretrained models and feature-level fusion, aiming to enhance both the robustness and generalization of the classification model.

- Generic pretrained models may lack domain-specific discriminative power for this task. Therefore, this study evaluates the use of self-supervised learning on additional in-domain datasets to enrich unimodal feature representations and improve overall classification performance.

1.3.2 Scope

- **Task**: This research will focus on multi-class classification where the objective is to assign a single class label based on information from multiple input modalities.

- **Modalities:** The research will primarily focus on the integration of audio and video an/or image data.
- **Fusion techniques:** This research will work on early fusion or feature-level fusion, where feature embedding in latent space are extracted and fused to formed multimodal representation for classification.
- **Dataset:** The study will specifically use small-scaled dog social signal dataset that has emotion per voice type annotation.
- **Transfer Learning Techniques:** The research will explore common transfer learning techniques including, but not limited to, full fine-tuning of the pre-trained model, freezing specific layers, and using the pre-trained model as a fixed feature extractor.
- **Evaluation Metrics:** Model performance will be evaluated using standard classification metrics such as accuracy, F1-score specifically tailored for multi-class classification on the chosen dataset.

1.4 Thesis organization

This thesis is demonstrated in 7 chapters in total.

Chapter 1 presents an overview of the study, describes the multi-class dog social signal classification problem, and explains the rationale for selecting the topic, as well as its research objectives and scope. It is followed by the proposed approach and the challenges encountered. Finally, it summarizes the practical significance and contributions of the study.

Chapter 2 provides an overview of key terms and methods that are used in later chapters. It also presents a brief summary of prior research in relevant areas on dog social signal classification using machine learning.

Chapter 3 introduces the main dataset, detailing the annotation procedures, label definitions, dataset statistics, and the supplementary datasets used in the experiments.

Chapter 4 describes the backbone architectures used for each modality, based on models from prior work, and explains how features are extracted, processed, and fused within the multimodal framework for classification.

Chapter 5 presents the experimental results, including a description of the conducted experiments, the outcomes, and a comparative analysis between the proposed multimodal approach and baseline methods on the dataset.

Chapter 6 presents ablation studies to validate the use of additional unimodal pretraining and frozen pretrained encoders in downstream tasks. This

chapter provides empirical evidence supporting the proposed approach to improve multimodal performance, particularly in scenarios involving small, noisy datasets.

Chapter 7 presents the conclusion, summarizing the key points of the thesis, and outlining potential directions for future work.

Chapter 2

Related work

This chapter provides a review of existing literature related to multimodal machine learning, dog social signal classification, and transfer learning. It begins by introducing the general concepts of each unimodal learning and multimodal fusion method, followed by an overview of transfer learning approaches. The chapter then discusses dog social signals and highlights prior studies related to dog social signal classification tasks. Finally, it identifies the limitations and research gaps in the current literature.

2.1 Multimodal machine learning

2.1.1 Concepts

A modality refers to a specific type of data or sensory input, for example, images (vision), speech signals (audio), or text (language). Human communication is inherently multimodal. We express intent, emotion, and social signals through language, voice, facial expressions, and body gestures [58, 45].

Over the past few years, multimodal learning has emerged as a powerful approach to enable machines to understand and reason across multiple sources of information, such as text, images, audio, and video. This paradigm aims to mimic human perception, which naturally integrates signals from various modalities to make sense of the world. For AI systems to advance in their understanding of the world, the ability to interpret and reason over multimodal information is essential [5]. The importance of multimodal integration was recognised early on through the work of McGurk and MacDonald [31], who showed that combining auditory and visual cues can substantially enhance speech perception. Later studies confirmed that incorporating visual information, such as lip movements, significantly improves speech recognition

performance [41]. Since then, the field has continued to evolve, culminating in the recent emergence of multimodal large language models that leverage diverse multimodal datasets and adapt to various tasks, marking a significant step towards achieving artificial general intelligence (AGI).

2.1.2 Unimodal representations - Audio

The waveform shows how audio sample values vary over time, representing changes in sound amplitude. This is referred to as the time-domain representation of sound, as illustrated in Figure 2.1. Alternatively, the frequency

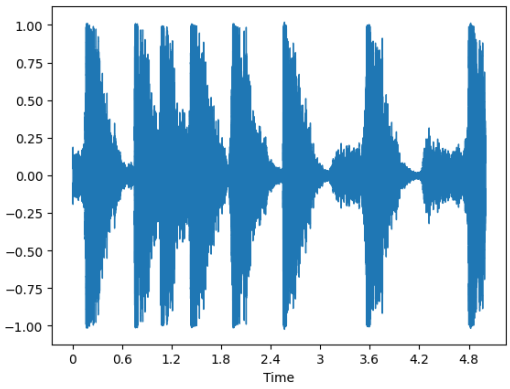


Figure 2.1: Waveform, visualization of sound in time domain

spectrum, or frequency-domain representation, is also commonly used to visualize audio, as shown in Figure 2.2. The spectrum captures a static snapshot of the frequencies present at a specific moment, and it can be derived from the time-domain waveform using the discrete Fourier transform (DFT).

By performing DFTs on successive short-time segments and stacking these frequency snapshots, we obtain a spectrogram. A spectrogram displays how the frequency content of a signal evolves over time, combining information on time, frequency, and amplitude in a single plot. The Short Time Fourier Transform (STFT) performs this computation, generating an informative visual representation of sound, as shown in Figure 2.3. A mel spectrogram is a type of spectrogram widely used in speech processing and machine learning. Unlike a regular spectrogram, which has a linear frequency axis measured in hertz (Hz), the mel spectrogram uses the mel scale, which better reflects how humans perceive sound. This is because our ears are more sensitive to lower frequencies, and this sensitivity decreases logarithmically as frequency increases. To generate a mel spectrogram, the Short Time Fourier

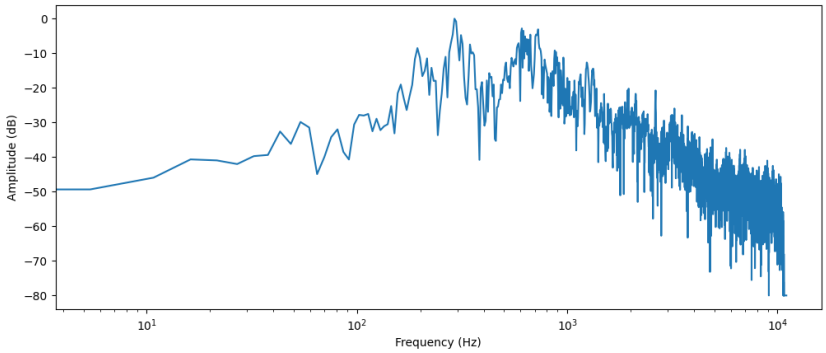


Figure 2.2: (Frequency) Spectrum, visualization of sound in frequency domain

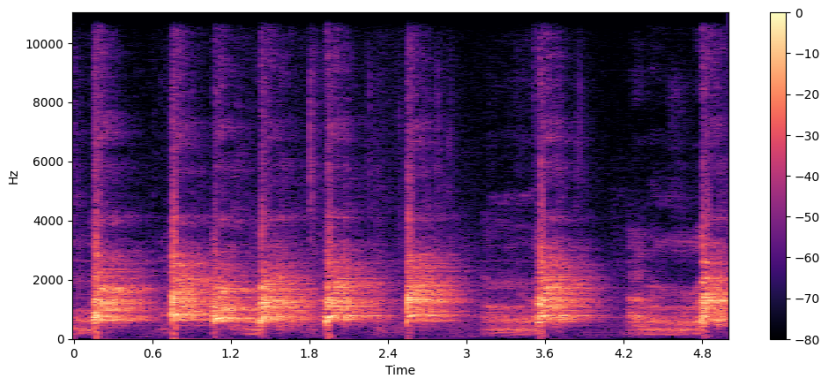


Figure 2.3: Spectrogram, visual representation of sound

Transform (STFT) is first applied to divide the audio into short frames and obtain their frequency spectra. These spectra are then passed through a mel filterbank, which converts the frequency values into the mel scale. Similar to standard spectrograms, it is common to represent the amplitude of mel frequency components in decibels, resulting in what is called a log-mel spectrogram due to the logarithmic transformation used in this conversion.

When working with audio, state-of-the-art models demonstrate impressive capabilities by utilizing various types of features. Applying deep learning to audio typically involves first converting the audio into an appropriate representation before feeding it into the model. These representations include raw audio features, as used in models like Wav2vec2 [4], and spectrogram-based features employed by models such as AVES [22] and AST [19]. Additionally, handcrafted features like MFCCs remain widely used and are particularly popular in many audio processing tasks. Recent studies have demonstrated that self-supervised learning approaches achieve outstanding performance by first training audio representation models on large unlabelled datasets, which are then applied to various downstream tasks. This strategy has yielded impressive results across multiple benchmarks, particularly when working with small-scale downstream datasets [22, 19, 4], and in some cases, strong performance is achieved even without additional fine-tuning. By leveraging the rich and transferable acoustic representations learned by these generic audio models, the need for large amounts of labelled data is significantly reduced. Moreover, this approach has shown strong transferability across domains, for example, when models pretrained on human speech are applied to bioacoustic tasks [52]. Given that data plays a critical role in the success of data-driven machine learning, this paradigm offers significant promise for the future. However, domain mismatch and data scarcity continue to pose major challenges.

2.1.3 Unimodal representations - Video

Video analysis and understanding are fundamental problems in computer vision. Unlike images, videos consist of sequences of frames, introducing an additional temporal dimension that has attracted significant research attention in recent years. Extracting effective features that capture both the spatial and temporal aspects of video remains a challenging task. Learning robust video representations is particularly difficult due to factors such as variations in viewpoint and background, changes in illumination, and occlusions, all of which can prevent models from fully capturing the intended scene [48]. Therefore, developing modern feature extraction methods is also essential to overcome these challenges and achieve robust and meaningful video

representations.

CNN-based model: Two-stream architectures are designed to model both the spatial and temporal components of video. For example, the SlowFast [15] network introduces a dual-pathway approach, where the slow pathway captures semantic information at a lower frame rate, while the fast pathway focuses on learning fine-grained temporal dynamics at a higher frame rate. With the rise of deep learning and the success of convolutional neural networks (CNNs) in image tasks, 3D CNNs have been extended to process videos by incorporating the temporal dimension into convolution operations. However, 3D convolutions are computationally expensive. To address this, the S3D[62] model factorizes 3D convolutions into separate spatial and temporal kernels, significantly reducing computational cost while also achieving notable improvements in accuracy.

Transformer-based model: Inspired by the Vision Transformer (ViT), which uses the attention mechanism to effectively model long-range contextual information in images, TimeSformer[6] extends this approach to videos. It divides video frames into patches and treats the entire video as a sequence of these patches, which are then fed into a transformer encoder to learn video representations. Although transformer-based models are highly effective in capturing long-range dependencies, they typically demand substantial computational resources and large-scale training data. These requirements can limit their practicality in real-world applications where resources are constrained.

Self-Supervised learning for video: Without access to large annotated datasets, it is challenging for models to learn robust and generalizable video representations for specific tasks. Self-supervised learning offers a promising solution by reducing the reliance on labelled data while enhancing performance in video-related domains. M. C. Schiappa et al. [53] categorize self-supervised learning approaches for videos into four main types: pretext tasks, generative learning, contrastive learning, and cross-modal agreement. Pretext tasks involve training models on carefully designed tasks that require a deep understanding of the input data, enabling them to learn generalizable features useful for downstream applications. Generative approaches, on the other hand, focus on reconstructing parts of the original input, such as predicting masked frames or generating future frames in a sequence. Contrastive learning trains models to bring positive pairs of inputs closer together in the feature space while pushing negative pairs apart, effectively learning discriminative representations. Cross-modal objectives typically use contrastive losses, such as Noise Contrastive Estimation (NCE) For instance, MIL-NCE [32] introduces an end-to-end framework for learning visual representations from uncurated instructional videos by aligning video and text embeddings.

This method produces robust representations that outperform many supervised approaches on downstream tasks, demonstrating the potential of learning generic video representations through joint video-text embedding.

2.1.4 Fusion

Multimodal fusion is one of the original and widely researched areas with a large number of approaches. It is the concept of integrating multiple sources of information and modalities to make a prediction. According to Baltrušaitis et al. [5], multimodal fusion methods can be broadly categorised into model-agnostic approaches and model-based approaches.

Model-agnostic approaches include early fusion, late fusion, and hybrid fusion techniques. Early fusion combines features from different modalities immediately after encoding, typically through simple operations such as concatenation. It can be further categorised into data-level fusion, which integrates data directly at the input stage, and feature-level (intermediate) fusion, which merges feature embeddings extracted from each modality. Late fusion integrates the outputs of separate unimodal models at the decision level, often using methods such as majority voting or weighted averaging to produce the final prediction. Hybrid fusion seeks to leverage the strengths of both early and late fusion by combining features at multiple levels within the model.

Model-based approaches include kernel-based methods, graph-based methods, and neural network-based approaches. Neural network-based methods, such as attention mechanisms and multimodal Transformers, have demonstrated strong capabilities in capturing complex inter-modal relationships. However, these approaches still face challenges, including data bias and high computational costs.

2.2 Tranfer learning

To build an effective model for a specific application, machine learning approaches typically require large-scale datasets to perform well. A common practical heuristic suggests that the number of training examples should exceed the number of trainable parameters by at least an order of magnitude to avoid severe overfitting. Traditionally, however, machine learning models are designed to operate in isolation and need to be rebuilt from scratch whenever the feature space distribution changes. This approach requires large datasets and incurs high computational costs for each new task, leading to inefficient use of knowledge and poor scalability. Transfer learning has emerged

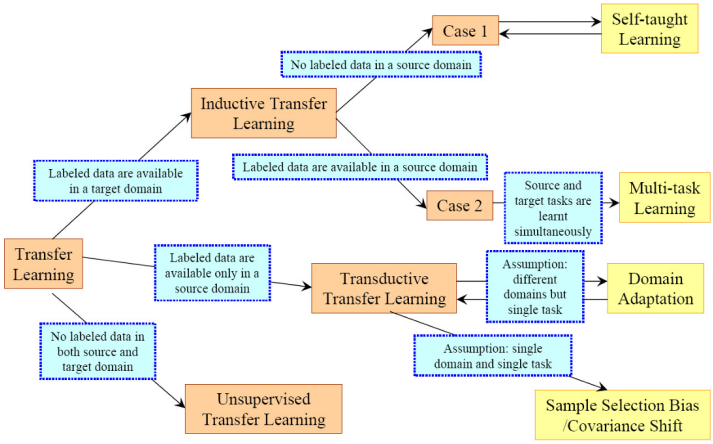


Figure 2.4: Overview of transfer learning settings [38]

to address these challenges by utilising knowledge learned from one task to improve performance on other related tasks.

Transfer learning has increasingly proven to be effective, enhancing model accuracy, reducing training time, and requiring less data, which makes it applicable to a wide range of tasks and datasets with both scholarly and industrial success. By leveraging knowledge from a large, high-quality source dataset and transferring it to a target task with a smaller dataset, transfer learning improves both accuracy and generalisation. This is often achieved by reusing features, weights, or representations from pretrained models to facilitate better learning on the target task. However, when applying transfer learning, it is important to consider which parts of the knowledge should be transferred, when transfer learning is appropriate, and which transfer techniques or algorithms will enhance performance without causing negative transfer. Depending on the domain and the availability of data, transfer learning strategies can be categorised as illustrated in Figure 2.4.

Two popular methods for deep transfer learning to downstream tasks are full fine-tuning (FT) and linear probing (LP). Fine-tuning updates all of the model parameters, whereas linear probing only trains the classifier head while keeping the lower layers frozen. However, when pretrained features are strong and there is a significant domain shift, fine-tuning can result in worse out-of-distribution (OOD) accuracy than linear probing, as it may distort the useful pretrained representations. In such cases, a two-step strategy combining LP followed by FT (LP-FT) can achieve higher performance by leveraging the strengths of both methods [26]. Many approaches aim to build generic and generalisable models across various modalities to create robust representa-

tions that achieve promising results in different conditions and tasks. In scenarios involving extremely small or noisy target datasets, linear probing with a strong, generic pretrained model is an effective strategy as it keeps the pretrained encoder frozen, preserving its high-quality generic feature embeddings and reducing the risk of overfitting to limited or low-quality data. This approach maintains the informative representations learned from the source data and provides a strong baseline for initial analysis and benchmarking.

2.3 Dog social signal in human-dog interaction

A social signal is a communicative or informative cue that conveys social information, such as social interactions, attitudes, relationships, or emotions [44]. This study aims to investigate whether dog social signals can be automatically extracted from audio and video data using multimodal machine learning approaches.

Vocal expressions of emotion follow basic rules that encode an animal's internal state into specific acoustic features, which human use these structural patterns to interpret and assign emotions to dog vocalizations that fit the context during cross-species interactions [14]. In particular, dogs growl and bark encode emotional states through specific acoustic features, allowing humans to perceive a dog's inner feelings across different contexts. According to the Source-Filter Framework, emotional arousal alters vocal parameters, such as call length, pitch (fundamental frequency), and formant dispersion, which signal changes in size perception and emotional valence. Farago et.al [14] investigated how human listeners recognize the dog growl types. From the experiment, human listeners could successfully recognize the emotional content (aggression, fear, despair, happiness, playfulness) and recognize the growl's context above chance levels. Play growls were rated higher on happiness and playfulness and lower on aggression and fear, while food guarding growls were judged most aggressive. Threatening growls were rated with both high aggression and fear. The study also clarified key acoustic features related to the vocal type. Jégh-Czinege et. al [25] also investigated how dog barks convey emotional states through their acoustic features, including primarily pitch, tonality, and inter-bark intervals, and they investigated how features influence how humans perceive both the dog's emotions and the level of annoyance caused by the barking. In a playback experiment with 153 Hungarian participants across different ages and living environments, listeners rated various dog bark sequences based on the dog's apparent emotional

state (happy-playful, scared-desperate, aggressive-angry) and the perceived annoyance. One of the important functions of barking is to evoke specific attention in humans. From these findings, "barking" as a vocal expression conveys various types of emotional inner-states of dogs. Acoustic features contained in barking sounds can be used as clues to identify the emotion categories of dogs. Automatically detecting a dog's social signal for specific attention-evoking to humans and others is a challenge in computational social signal processing. Lenkei et al. [28] investigated the role of "whine" as a vocal expression by focusing on the vocal activity of dogs with Separation-Related Problems (SRP). The SRP is used to refer to a set of problem behaviors shown by dogs when the owner or the attachment figure is absent [2]. The study found that separation-related problems (SRP) in dogs manifest through various behaviors, notably barking and whining, depending on the dog's inner emotional state and the owner's interaction style. Whining during separation is typically associated with fear and anxiety, while barking is linked more closely to frustration. Furthermore, the study by Lenkei et al. [29] provided empirical evidence that separation-related behaviors in dogs are not homogeneous but rather reflect distinct emotional profiles, underscoring the need for individualized intervention strategies based on the dog's temperament and behavioral tendencies.

It has long been recognized, since the work of Charles Darwin [37], that facial expressions provide crucial information for classifying emotional states. Boneh-Shitrit, Tali, et al. [7] investigated the automated recognition of dog emotional states from facial expressions using DogFACS [59], a facial action coding system developed specifically for dogs, together with deep learning methods. In their study on binary classification of dog emotions, specifically frustration (negative) versus anticipation (positive), the DogFACS-based approach showed promise. However, the deep learning model outperformed it, achieving an accuracy of over 89%, demonstrating the strength of deep learning for emotion classification tasks. DogFACS cannot exhaustively capture all possible behavioral variations, and many videos do not display identifiable DogFACS variables. Despite this, the deep learning model successfully classifies these cases based solely on images, highlighting its ability to detect fine-grained pixel-level details that may be imperceptible to the human eye. More works on dog visual features such as dog face or body [16] show that we can use visual features to predict dog emotions. Such proofs indicate the potential of using visual features to help us understand a dog's emotional state. In this thesis, both acoustic and visual features are incorporated in a multimodal pipeline for effective dog social signal classification.

2.4 Dog social signal processing

This research field receives limited research attention, and there are few dog social signal datasets. The current most related dataset to ours, Abzaliev [1] introduces a dataset with 14 dog vocalizations for 4 classifications tasks that are parallel with speech classification (dog recognition, breed recognition, gender identification, and context grounding). Mohandas, Prabu, et al [34] work on bark recorded two different dog species with various contexts for barking classification. However, apart from only dog bark, other dog vocalizations (growl, whine) also carry expressions of the dog's inner state. This work investigates a video dataset, including audio and video, with annotated emotion based on dog vocalization. This study not only includes a broader scope of vocalization but also enables multimodal study and analysis.

The current most related to this tasks is context classification using dog bark [35]. In this work, handcraft features extracted passed through a Naive Bayes classifier, even with substantial audio samples, the overall accuracy remains relatively low, which may be due to the simplicity of the baseline. Other work [23] also classifies dog bark for context classification with a small dataset using handcraft features but without an outstanding result. These methods apply traditional handcraft features of audio and simple classification methods without impressive results. Meanwhile, we can leverage a pre-trained model with minimum training efforts that produce praiseworthy results. The works of context classification using dog barks [18] yield notable performance using various deep learning models. However, creating a dataset of this size and richness is a non-trivial task, especially under real-world conditions where data availability and consistency are limited.

For a small and noisy dataset, transfer learning using a pre-trained model for downstream tasks is popular with efficient computing and outstanding performance. Several audio pre-train models have remarkable results on speech benchmarks. However, these are mostly audio tasks that are human speech-related, and few models are trained for bioacoustics data. Sarkar et al [52] investigates the benefits of different pretrained models, general purpose, human speech pre-trained ones for bioacoustic data. Results suggest the general-purpose audio pre-trained model is suitable for bioacoustic tasks without pre-training on animal vocalizations. This work investigates whether representations learned from general-purpose pretraining contain meaningful information for emotion classification. Moreover, experiment with cross domain transfer learning in section 6.2.2 is conducted in other to understand how transferable learned representations are across domains and species, and whether domain-specific fine-tuning can close the performance gap. This is a potential solution that opens up the possibility of leveraging knowledge

learned from a well-annotated source domain to improve performance in a target domain where data is limited or noisy. In addition, experiment with two general pre-trained models for multimodal classification for both visual and acoustic modality is conducted to gain more insights into the effectiveness of leveraging generic pre-trained models for dog social signal classification.

Chapter 3

Dataset

3.1 Source dataset selection

Dog social signal dataset is a portion of the dog video data contained in AudioSet [17]. AudioSet is a large-scale dataset with $1,789,621$ videos from YouTube, which was manually annotated with 632 audio event categories. All segments have 10 seconds long, except for those with shorter durations.

3.2 Dog social signal annotation

As detailed in Section 2.3, dog vocalizations serve as crucial social signals, with distinct barks conveying diverse functions. Consequently, dog vocal expressions are categorized into three primary types: bark, whine, and growl.

Definition of internal state labels

Based on animal behavior knowledge and extensive clinical experience, the internal state labels of dogs are defined by a professional animal scientist as follows:

"Attention": As a positive internal state, when a dog has a favorable attitude toward its owner or interaction partner and wants to seek their attention, this state is labeled as "Attention."

"Aggressive": As a negative internal state, when a dog feels aversion toward another entity and wants to avoid it, or when it wants the other entity to go away and is in an aggressive state toward it, this state is labeled as "Aggressive."

"Anxiety": When a dog appears to be feeling anxiety, this label is assigned.

"Conflict": When a dog experiences two opposing motivations—for example, wanting to socialize with a person but simultaneously feeling afraid of how the person might react—it may display signals through body language and vocalizations that seem contradictory. In such moments, the dog's barking reflects this internal motivational conflict. Additionally, when they are conflicted, they are also likely to feel frustration.

Annotation procedure

First, the expert who defined the labels was watching 7760 videos in AudioSet and selecting videos in which the dogs seemed to signal their internal states (defined in Section 3.2) with various vocal expressions. The expert also carefully annotates the segments (start and end times) of the scenes where the dogs is barking or whining or growling in the videos, expressing their internal state labels. Finally, total 550 videos were annotated with 8 labels. Table 3.1 shows the labels and number of samples. Labels indicate internal states corresponding to a dog vocal sound of the subject at the moment of audio recording. The videos were carefully labeled by annotation teams of professional animal scientists. Only a subset of videos containing sounds corresponding to specific vocal type labels and one dog per video is filtered and used for training.

Annotator agreement

The labels are subjectively annotated based on the experience of an expert. The validity of the labels defined in this study is verified by measuring the agreement rate of subjective labels. They independently commissioned an animal behavioral researcher (second coder) with less experience than the expert annotator to annotate a portion of the table samples and calculated the agreement rate. The second coder was provided with training to understand the label definitions and spent approximately 10 hours learning the relationship between the videos and the labels. As a validation test, the second coder annotated 42 independent samples which are not overlapped with samples the coder have learned in the training phase. As a result of the validation test, for seven class categories of the internal label, the agreement values (Cohen's Kappa) were 0.48, which is regarded as moderate agreement [27]. Although training more annotators is necessary, the validity of certain label definitions has been demonstrated.

Table 3.1: Statistics of the dataset

Type	**Label**	**Numerical label**	**#Segments**
Bark	Aggressive	0	175
	Attention	1	68
	Conflict	2	142
Whine	Anxiety	3	26
	Attention	4	50
Growl	Aggressive	5	37
	Attention	7	5
	Conflict	6	47
		Total	550

Statistics of annotated labels

At this point in this research, the dataset is characterized by significant class imbalance and a relatively small sample size as seen in Table 3.1. There are 31.82% labeled bark/aggressive, 25.82% labeled bark/conflict, and 12.37% labeled bark/attention, in a total of 70% samples belonging to the same group of bark. These labels are associated with the highest number of samples. On the other hand, labels such as growl/attention and whine/anxiety have less than 1% and 5% of the total number of samples respectively. The considerable imbalance in sample distribution between classes may bias the model towards the majority class labels, skewing predictions. To deal with this problem, I intentionally perform augmentation (such as adding noise, rir effect) to increase the number of the lesser label to reduce this negative impact. Future work on data collection and annotation could help with this problem. Moreover, due to the scarcity of the growl/attention label, I only do experiments with 7 labels (growl/attention is not used).

Additionally, various recording conditions, bad-quality videos, and noises, such as human voices or other object voices in between audio recordings, are also challenges for us.

Label	ID	Number of samples
bark	/m/05tny_	2292
yip	/m/07r_k2n	2022
howl	/m/07qf0zm	737
bow-wow	/m/07rc7d9	3325
growling	/m/0ghcn6	461
whimper	/t/dd00136	1157
bay	/m/07srf8z	0
Total		9994

Table 3.2: Statistics of dog voice types from AudioSet

3.3 Support dataset

3.3.1 Dog audio from AudioSet dataset

The dog audio data used in this study is a subset of the AudioSet, comprising approximately $13,705$ video annotations of familiar domesticated canid sounds, including Bark, Yip, Howl, Bow-Wow, Growling, Whimper, and Bay. These voice types are statistically listed as below table 3.2

Chapter 4

Methology

4.1 Model overview

4.1.1 Pretrained acoustic - AST

Transformer

The Transformer architecture is a sequence-to-sequence model consisting of an encoder-decoder structure, as illustrated in Figure 4.1. The vanilla Transformer is constructed from blocks that combine multi-head attention mechanisms with position-wise feed-forward networks.

Attention mechanism The introduction of attention mechanisms and the Transformer architecture has revolutionized natural language processing (NLP), leading to major advancements not only in NLP but also in computer vision and other fields [Vaswani et al., 2017]. The attention mechanism enables a model to focus selectively on different parts of the input sequence, enhancing its ability to capture contextual relationships. The Scaled Dot-Product Attention is formulated as:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.1}$$

where Q, K, and V denote the query, key, and value matrices, respectively, and d_k is the dimensionality of the key vectors, used to scale the dot product [57].

In the Transformer encoder, self-attention is applied with $Q = K = V = X$, where X is the output from the previous layer, forming the core of the autoencoder-like architecture utilized by models such as BERT [10]. Transformer decoder employs both masked self-attention, ensuring that each position in the output sequence attends only to earlier positions (thus preserving the autoregressive property crucial for generation tasks, as in GPT

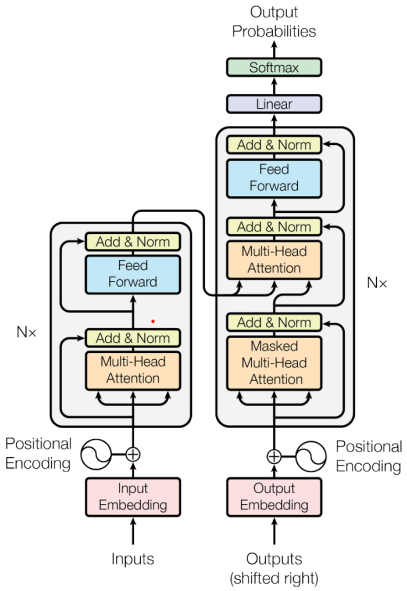


Figure 4.1: Transformer architecture

[46]), and cross-attention, where the queries come from the decoder's previous layer outputs, while the keys and values are derived from the encoder outputs. Together, the encoder-decoder architecture forms a sequence-to-sequence model, exemplified by models such as T5 [47]. Since its introduction, the generative Transformer paradigm has driven continuous breakthroughs in NLP.

Multi-Head Attention utilizes multiple sets of (Q, K, V) projections for a single input, enabling the model to capture different types of relationships within the sequence and thereby achieve a more comprehensive understanding of the data [57].

Since their introduction in 2017, Transformers with attention mechanisms have revolutionized natural language processing (NLP), fundamentally transforming approaches to a wide range of NLP tasks and driving remarkable progress across the field. More recently, this architecture has also been successfully extended to other domains.. For example, the introduction of the Vision Transformer (ViT) has demonstrated the great potential of applying Transformer architectures to CV tasks [11].

Audio Spectrogram Transformer

The Vision Transformer (ViT) utilizes a transformer architecture that processes images by segmenting them into sequences of patches, embodying the

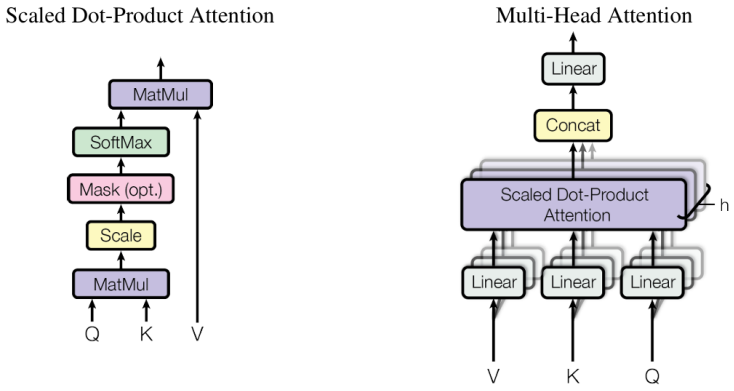


Figure 4.2: The scaled dot-product attention architecture (left) and the multi-head attention mechanism (right) are composed of multiple attention layers that function concurrently

idea that "an image is worth 16×16 words". Figure 4.3 shows the overall ViT architecture and how the image is divided into patches and then flattened before being fed into the transformer. This method has demonstrated outstanding performance across a wide range of computer vision benchmarks [39]. Audio classification methods that operate on spectrogram inputs typi-

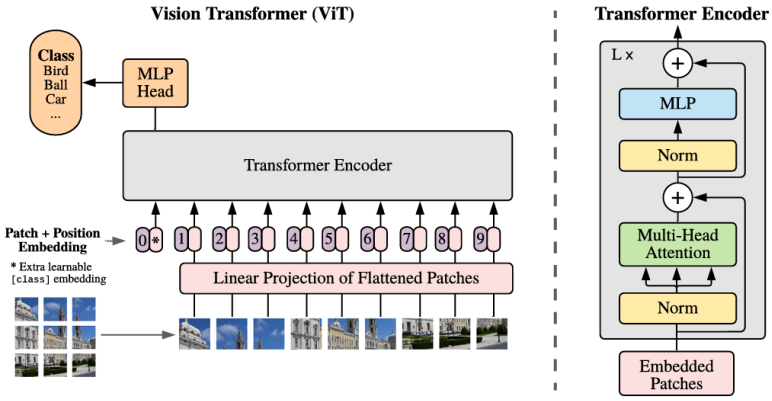


Figure 4.3: Vision Transformer architecture

cally rely on convolutional neural networks (CNNs) to exploit their inductive bias, capturing local spatial patterns and maintaining translation equivariance. These CNN-based models often incorporate attention mechanisms on

top to capture long-range global dependencies, achieving state-of-the-art results in various audio classification tasks. [30, 20, 49]. Inspired by ViT's success in computer vision, the Audio Spectrogram Transformer (AST) was proposed as a fully attention-based, convolution-free model for audio classification, as illustrated in Figure 4.4.

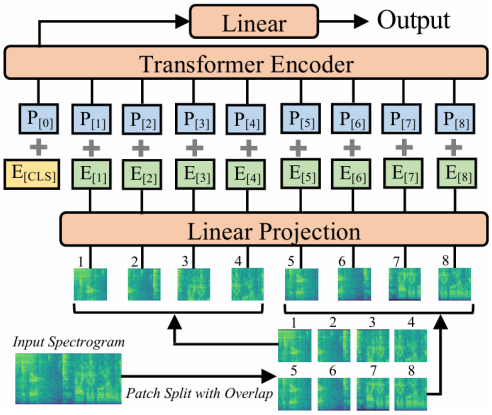


Figure 4.4: Audio Spectrogram Transformer architecture

An audio waveform of t seconds is first converted into log Mel filterbank features with 128 frequency bins, computed using a 25 ms Hamming window and a 10 ms hop size, resulting in a spectrogram of size $128 * 100t$. This spectrogram is then divided into 16×16 patches with an overlap of 6 frames in both frequency and time dimensions. Consequently, the number of patches becomes $12[(100t–16)/10]$, which forms the input sequence to the AST model. Experimental results showed that using a patch overlap of 6 increases the sequence length, leading to a quadratic rise in computational cost. However, this configuration achieved the highest performance among all tested settings on the AudioSet benchmark, including non-overlap, 2, 4, and 6 overlap.

The AST model adopts a standard transformer architecture similar to ViT. However, while ViT is designed for 3-channel RGB image inputs, AST operates on single-channel spectrogram inputs. To align the patch embedding layer with ViT's pretrained weights, the weights corresponding to the three input channels in ViT are averaged, effectively adapting them for AST's single-channel input by replicating the averaged weights across the single channel. In ViT with an input resolution of 384×384 and a patch size of 16×16, the image is divided into 24×24 patches. In contrast, AST processes spectrograms that yield $12[(100t–16)/10]$ overlapping patches, depending on the audio duration t. Because of this difference in the number and arrangement of patches, AST adapts ViT's positional embeddings by cropping along

the first dimension and applying bilinear interpolation along the second dimension. The [CLS] token positional embedding is directly reused without modification. This positional embedding adaptation, involving cropping and bilinear interpolation, plays a critical role in enabling effective knowledge transfer from ViT to AST, as demonstrated in the paper's experiments.

Once the AST architecture is established, the model is initialized with ViT weights pretrained on ImageNet, specifically using weights from DeiT. Experiments conducted on three benchmark datasets: AudioSet[17], ESC-50 [43], and Speech Commands v2 [60], demonstrate its superior performance compared to CNN-attention hybrid models and SOTA benchmarks on these datasets. The ESC-50 dataset contains 2,000 five-second environmental audio recordings spanning 50 classes, while Speech Commands includes 105,829 one-second recordings across 35 command labels. Remarkably, AST achieves strong results even on these relatively small datasets, highlighting the substantial benefits of transfer learning from ImageNet-pretrained weights. This study demonstrates that ImageNet pretraining significantly reduces the need for large-scale in-domain audio data for AST, supporting its potential as a generic and effective audio classifier capable of handling varying audio lengths.

Transformer architecture with attention mechanism using spectrogram feature produces a concise and meaningful representation for audio [63] but requires a large amount of data to train. Since image and spectrogram have similar formats, many researchers have explored cross-domain transfer learning from vision to audio domain. Audio Spectrogram Transformer [19] shows the effectiveness of ImageNet pretraining, which can help reduce the necessity for in-domain audio data and produce remarkable results even without AudioSet [17] pretraining for speech classification tasks. Given the limited data availability and scarcity of pretrained models in this field, this approach shows strong potential as a generic audio representation. Therefore, this study adopts it as the audio encoder.

4.1.2 Pretrained video - S3D

Inflated 3D ConvNet

Two-dimensional convolutional neural networks (2D CNNs) have achieved remarkable success in learning image representations and performing a wide range of image tasks. Given that videos can be viewed as sequences of consecutive images capturing motion over time, many studies have explored extending 2D CNN architectures to video analysis. Carreira and Zisserman[8] introduced the I3D (Inflated 3D ConvNet) architecture, which extends the

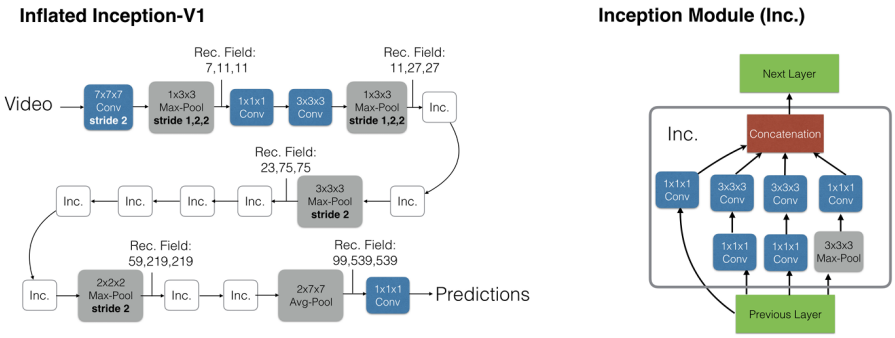


Figure 4.5: The I3D architecture based on Inflated Inception-V1 (left) and a detailed view of its inception submodule (right).

ImageNet-pretrained Inception V1 model by inflating its 2D convolutional filters into 3D convolutions along the temporal dimension. This architecture is illustrated in Figure 4.5. This approach enables I3D to perform convolutions across both spatial and temporal dimensions, effectively capturing motion dynamics in videos. Their experiments demonstrated that I3D achieves strong performance on multiple action recognition benchmarks, highlighting its ability to learn rich spatiotemporal representations. Furthermore, they showed that initializing I3D with inflated weights from ImageNet-pretrained models consistently outperformed training from scratch, underscoring the significant benefits of leveraging transfer learning for video-based tasks.

Separable 3D CNN

However, 3D CNNs are computationally intensive and susceptible to overfitting due to their large number of parameters. To address these limitations, Xie et al. [62] proposed the S3D (Separable 3D CNN) architecture, which factorizes standard I3D convolutions (with kernel size $k_t \times k \times k$) into separate spatial and temporal convolutions. Specifically, S3D replaces each 3D convolution with a temporal convolution of size $k_t \times 1 \times 1$ followed by a spatial convolution of size $1 \times k \times k$, where k_t denotes the kernel size along the temporal, and k denotes the kernel denotes width/height of kernel in spatial dimensions. The overview of S3D model architecture is illustrated in Figure 4.6 This factorization significantly reduces the number of parameters and computational cost while improving accuracy compared to conventional 3D CNNs. Furthermore, their results suggest that replacing lower-layer 3D convolutions with more efficient 2D convolutions yields top-heavy models that are not only more computationally efficient but also achieve higher accuracy.

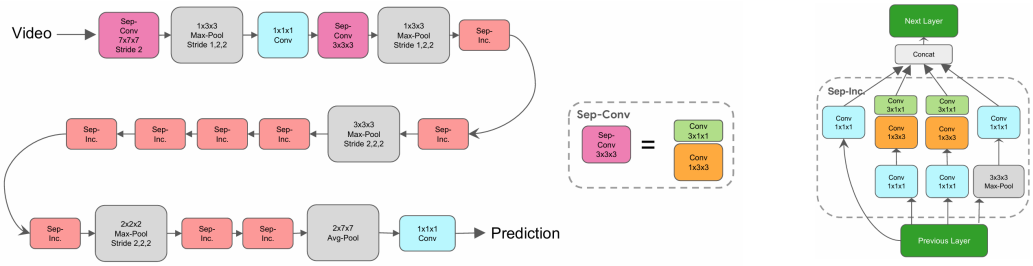


Figure 4.6: Overview of the S3D model architecture (left), with details of the temporally separable convolution (Sep-Conv) module (middle) and the temporally separable inception blocks (Sep-Inc) (right).

Importantly, retaining 3D convolutions in higher layers remains beneficial, as they effectively capture temporal dependencies among high-level semantic features.

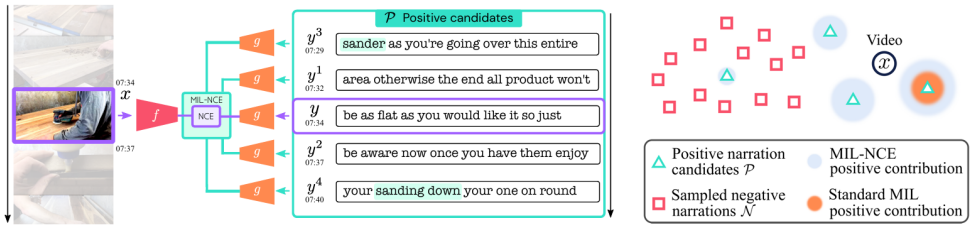
The Fast-S3D architecture improves computational efficiency by replacing the 3D convolutions in the lower layers with 2D convolutions while retaining separable 3D convolutions in the top two layers, achieving an optimal trade-off between speed and accuracy. Additionally, the S3D-G variant enhances the original S3D by incorporating context gating immediately after each temporal convolution of size $[k, 1, 1]$. In context gating, the output features Y are computed as:

$$Y = \sigma(W \cdot \text{avg\_pool}(X) + b) \odot X \tag{4.2}$$

where σ denotes the sigmoid activation function, W and b are learnable parameters, and $\odot$ represents element-wise multiplication along the channel dimension. The average pooling is performed across both spatial and temporal dimensions. This mechanism enables the model to adaptively upweight informative dimensions of the input X while downweighting less relevant ones, functioning similarly to a lightweight self-attention mechanism. The addition of context gating leads to improved accuracy with only a minimal increase in computational cost.

Video representation from Uncurated Instruction Videos[32]

Miech et al. [32] introduced the MIL-NCE (Multiple Instance Learning and Noise Contrastive Estimation) training loss to learn strong video representations from uncurated instructional videos in the HowTo100M dataset. This approach enables training high-quality video representations from scratch by



(a) **Examples of positive candidates** (b) **Illustration of *MIL-NCE***

Figure 4.7: MIL-NCE for learning video representations from uncurated datasets. Given a video x and a set of associated positive narration candidates $\mathcal{P}$, MIL-NCE leverages multiple positive pairs—such as $(x, y), (x, y_1), (x, y_2), (x, y_3), (x, y_4)$ (left), which better captures fine-grained object references like 'sander' in (x, y_3) and specific action descriptions in (x, y_4) that standard NCE may miss. This method promotes multiple correct positives while downweighting inaccurate ones using a discriminative ratio against negatives $\mathcal{N}$ (right).

effectively leveraging the weak and noisy supervision inherent in available large-scale datasets. One significant challenge in HowTo100M is that approximately 50% of clip-narration pairs are not temporally aligned, making it difficult for standard methods to associate narration content with corresponding visual events. MIL-NCE addresses this by treating the task as multiple instance learning: For each video clip, multiple candidate captions are considered as potential positive matches, thereby increasing the likelihood that caption accurately describes the visual content. During training, a 3.2-second clip is randomly sampled from a video, and a bag of positive candidate captions (P) is constructed using the captions temporally nearest to the clip, as illustrated in Figure 4.7. Negative samples are formed by pairing sampled clips with narrations that do not belong to its positive bag. The MIL-NCE objective is formulated as:

$$\max_{f,g} \sum_{i=1}^{n} \log \left(\frac{\sum_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)}}{\sum_{(x,y)\in\mathcal{P}_i} e^{f(x)^\top g(y)} + \sum_{(x',y')\sim\mathcal{N}_i} e^{f(x')^\top g(y')}} \right) \tag{4.3}$$

For specific sample indexed by i, the sampled video x and narration y, $\mathcal{P}_i$ denotes the set of positive pairs, $\mathcal{N}_i$ is the set of negative pairs. Two paramaterized mapping $f : \mathcal{X} \to R^d$, and $g : \mathcal{Y} \to R^d$ map sampled video and text narration into d-dimension vector space. Joint probability of a pair of video and narration is estimated by $e^{f(x)^\top g(y)}$. . This objective encourages learning a shared embedding space in which videos and texts semantically

related are positioned closely together, while unrelated pairs remain distant, enabling robust representation learning from noisy supervision.

By employing 3 to 5 positive candidates and 512 negative candidates along with a simple bag-of-words language model, this approach achieves superior performance on text-to-video retrieval tasks, effectively learning meaningful video representations solely from paired video and narration data. Remarkably, it accomplishes this without relying on any manually annotated datasets, demonstrating the power of MIL-NCE pretraining for leveraging weak supervision. The resulting S3D-G representations outperform prior methods, including both self-supervised and fully supervised approaches, across various downstream video tasks such as action recognition and action segmentation. This highlights the effectiveness of MIL-NCE in capturing rich semantic information directly from large-scale uncurated instructional videos. Given the critical need for robust and generalizable visual encoders in video understanding, S3D-G[32] pretrained with MIL-NCE on HowTo100M[33] is utilized as the visual modality encoder in this study.

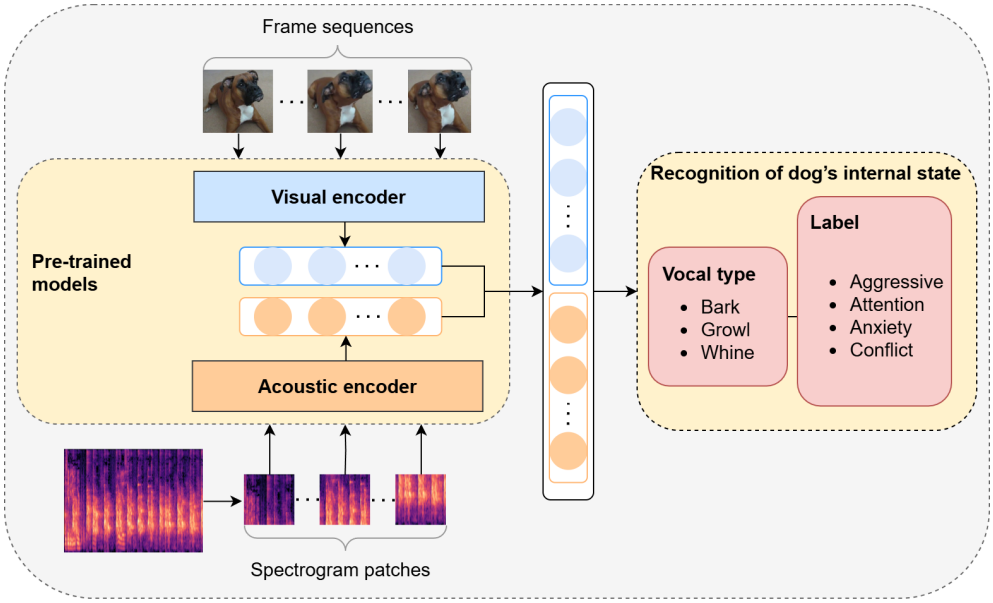


Figure 4.8: An Overview of multimodal dog social signal classification

4.2 Feature extraction based on pre-trained model

4.2.1 Acoustic feature embedding

Humans can recognize some of the most motivational dogs' inner states using audio features [36]. Dog vocalizations alone have acoustic properties related to emotions, physiological reactions, attitudes, or some particular internal states [18]. Animal sound pre-trained models are capable of capturing species-specific distinctive characteristics, which leads to a better understanding of animal communicational expression and contributes to improved performance in bioacoustic tasks. Nevertheless, the bioacoustic pre-trained model and general purpose pre-trained model show comparable results overall [52] on several different bioacoustic tasks. Furthermore, upon investigating the ImageNet pre-trained ViT model in section 6.2, I observe a remarkable dog vocalization ability despite the huge domain gap. Additionally, AST [19] is the modified architecture of ViT and pre-trained on AudioSet [17], a large-scale dataset with a wide range of categories. Given its strong potential as an audio representation model, I ultimately selected the AST model as audio embedding model. For acoustic feature embedding, a 10.24-second audio segment is processed through AST feature extraction to generate a 1024×124 spectrogram, which is then fed into the pre-trained AST model.

4.2.2 Visual feature embedding

Even though vocal features contain context and emotional information, I have answers from animal experts that visual features also provide useful indicators and are critical in several scenarios that help to annotate those of dog emotions in this dataset. Considering that S3D model [62] with 3D CNN architecture is a well-known model for video understanding tasks, I integrate S3D-G [32], a pre-trained models on HowTo100M [33], into the multimodal pipeline to encode video for visual feature extraction, and how its visual representations influence task performance. In visual modality, video is put through an extraction pipeline, in which 32 224x224 frames sre extracted to input into the S3D pre-trained model.

I experimented with various types of visual features, such as dog poses and object-centric representations, such as dog bounding boxes from video frames. To make the input more subject-centered, I applied object detection models to identify dogs and extracted features based on their bounding boxes. While these visual cues are proven effective indicators of emotion, this approach yielded suboptimal results on the dog social signal dataset. This

may be attributed to several factors, including the noisy nature of YouTube videos, frequent occlusions, and diverse camera angles that often fail to center on relevant objects. As a result, the detection models could successfully extract these features from only around 50% of the videos. Given these limitations, the current video understanding approach proves more reliable. For future work, I suggest curating more targeted or controlled video datasets, which could improve object-based feature extraction and potentially yield better performance.

4.3 Multimodal social signal modeling

This study addresses the task of classification dog social signals, aiming to identify 7 of dog voiced-based emotion classes (bark/conflict, bark/aggressive, bark/attention, whine/anxiety, whine/attention, growl/aggressive, growl/conflict). In this work, I conduct experiments with unimodal classification (audio, video) and multimodal classification (audio + video) to evaluate the performance of the proposed methods and analyze the contribution of these modalities.

As stated in [42], "Emotions are physical, and there are visible or audible signs of emotions. For example, if a computer tries to understand emotion, not just by the name, but by listening to the human voices, noticing their gesture, and appraising the situation that they are in". Since visual features provide rich contextual information for identifying emotions, while acoustic features capture subtle vocal emotion cues, I establish a multimodal baseline for analyzing dog emotions. An overview of this approach is presented in Figure 4.8. In this multimodal model, I leverage pretrained weights as unimodal encoders to construct a feature-based fusion classification pipeline. Specifically, each modality's input data is passed through its respective encoder to obtain a latent embedding, and these embeddings are then fused via concatenation. The resulting multimodal representation is fed into a classifier composed of three non-linear fully connected layers. To accelerate training and improve stability, I apply a normalization layer after the embedding concatenation. Additionally, to evaluate the contribution of each modality, I conduct experiments using the unimodal models independently, allowing us to assess their individual and combined effectiveness within the multimodal pipeline.

Chapter 5

Experimentation

5.1 Experiments: Pre-trained baseline for dog social signal classification

Understanding the challenges posed by the dataset, I conduct experiments using unimodal (audio and video) and multimodal approaches, leveraging transfer learning with powerful pretrained models for each modality. These experiments serve as baselines for this task, supporting future research and facilitating the development of more effective methods.

5.1.1 Baselines

To provide a more comprehensive evaluation, I conduct experiments using baseline models alongside the proposed method:

- Majority: As a basic reference, I use the majority class baseline to establish a minimum performance benchmark for evaluating the proposed methods.
- eGeMAPS: For the second baseline, I train a non-parametric Decision Tree model using 88 features extracted from the OpenSmile [13] eGeMAPSv02 [12] feature set.
- Unimodal classification: AudioSet pre-trained AST [19] is used for audio unimodal classification (Pretrain acoustic), and Howto100M [33] pre-trained S3D-G [32] for video unimodal classification (Pretrain visual).

- Multimodal classification model: For multimodal approaches, I select pre-trained AST for the audio encoder and pre-trained S3D for the visual encoder.

5.1.2 Experimental Settings

With a relatively small dataset, I employ a transfer learning approach on general pre-trained models to establish minimal training baselines for this benchmark. I keep the feature extraction module as in the pre-trained pipeline, more details about feature extraction and feature embedding are described in section 4.2, and transfer learning those pre-trained models with frozen pre-trained weights. Those pre-trained models used in this work are publicly available in Huggingface [61] hub.

Parameters: I optimize all models using Adam optimizer for cross-entropy loss. For all settings, I select the learning rate in $[1e-2, 1e-3]$ and batch size in $[32, 128]$, and models are trained for 30 epochs. For multimodal classification pipelines, 32 frames are extracted from video segments with frame step = 9.

Experimental results are recorded as five-fold cross-validation using the StrategiedFold object of the scikit-learn library [40]. This object splits the dataset into consecutive folds and preserves the percentage of samples for each class in folds. In the 5-fold cross-validation setup, each fold uses a hold-out set as the test set, while the remaining data serves as the training set. Approximately 30% of the training set is further allocated for validation. As one video can produce multiple data samples with annotation (start time, end time), and label ID. In order to prevent data leaks, the splits are made based on video IDs and labels, ensuring that clips with the same video ID and label are assigned to the same train, validation, or test set. For training splits within each fold, data augmentation methods (rir affects, or adding noise) are added to the lesser labels to reduce the effect of dataset imbalance.

Evaluation Metrics: In this paper, to evaluate model results, I adopted two common metrics accuracy and macro F1 score. Results are reported as the mean and standard deviation across all 5 folds. Accuracy is computed based on the total number of correctly predicted samples in the test set. The macro F1 score is calculated as the average of the F1 scores for each individual label in the test set.

5.1.3 Experimental results

Overall results are reported in Table 5.1, and Table 5.2 reports more details of the f1 score comparison of each label for each model. I also observe that the

Table 5.1: Performance results of dog social signal classification. Numbers in bold are the best metric values recorded, while those underlined are the second-best ones.

Feature	**Model**	**Accuracy$_7$**	**F1$_7$**
Majority	-	0.321 ± 0.000	0.069 ± 0.000
eGeMAPS (acoustic)	Decision Tree	0.406 ± 0.032	0.323 ± 0.072
Spectrogram	AST	$\underline{0.503 \pm 0.051}$	$\underline{0.472 \pm 0.070}$
Video	S3D	0.294 ± 0.060	0.220 ± 0.046
Video + Spec.	S3D-AST	$\mathbf{0.533 \pm 0.027}$	$\mathbf{0.477 \pm 0.033}$

multimodal model outperforms all baselines across all labels, while the machine learning approach using a pretrained acoustic model shows promising results as the second-best performer.

Detail results show the highest F1 score of each voice types group:

- Bark: bark/aggressive achived highest average f1 of 63.4% on multimodal (full-vid)
- Whine: whine/attention achived highest average f1 of 68.6% in pretrained acoustic.
- Growl: growl/conflict achived highest average f1 of 53.3% on multimodal (full-vid)

These labels correspond to those with the highest number of samples in each group, which points to the severe imbalance of this dog social signal dataset.

Compared to other pre-trained machine learning models, acoustic handcraft feature eGeMAPS shows limited performance. Though handcraft feature shows its ability in various emotion classification tasks, pretrained acoustic has 23.9% accuracy higher (46.1% f1 higher) than the eGeMAPS baseline, and multimodal (full-vid) has 31.3% accuracy higher (47.7% f1 higher) than the eGeMAPS baseline. This evidence shows effective of the proposed methods leverages the ability of pre-trained models to capture discriminative features from data, and produce good results. The domain gap between general-purpose pre-trained and the task-specific, limited dataset, initially seems unsuitable for training. However, the results demonstrate the effectiveness of leveraging robust pre-trained representations to achieve strong performance, surpassing traditional handcrafted acoustic features. Moreover, results from the confusion matrices: Figure 5.1 shows that, in pretrained acoustic (5.1.b),

Table 5.2: Detailed performance results

Model	**F1 score**						
	Bark			**Whine**		**Growl**	
	Aggressive	**Attention**	**Conflict**	**Anxiety**	**Attention**	**Aggression**	**Conflict**
Majority	0.486 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
eGeMAPS (acoustic)	0.503 ± 0.064	0.131 ± 0.088	0.418 ± 0.106	0.189 ± 0.144	0.507 ± 0.142	0.184 ± 0.189	0.328 ± 0.088
Pretrained acoustic	0.580 ± 0.057	0.342 ± 0.110	0.400 ± 0.180	0.428 ± 0.245	0.686 ± 0.086	0.352 ± 0.277	0.516 ± 0.128
Pretrained visual	0.397 ± 0.069	0.195 ± 0.097	0.333 ± 0.084	0.090 ± 0.131	0.164 ± 0.138	0.130 ± 0.083	0.228 ± 0.143
Multimodal (full-vid)	0.634 ± 0.036	0.301 ± 0.110	0.467 ± 0.100	0.284 ± 0.166	0.618 ± 0.128	0.502 ± 0.117	0.533 ± 0.122

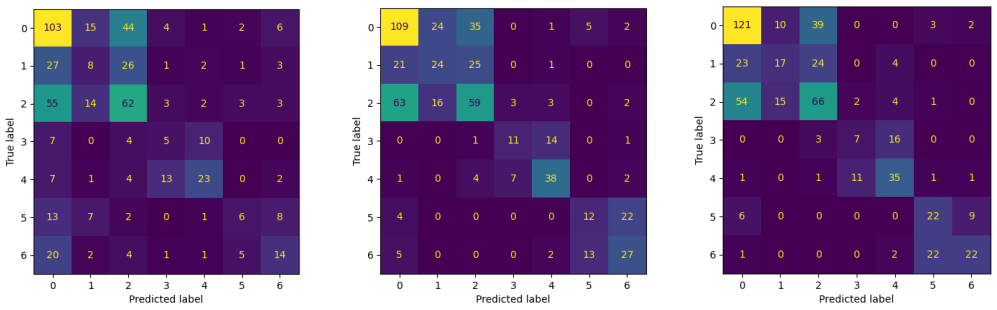
I observe clear vocal distinctions. Similarly, in the multimodal pipeline, Figure 5.1.c, I see that these models are more likely to confuse bark/aggression with bark/conflict than with growl/aggression. This trend supports the claim that vocal features play a dominant role. It also aligns with the hierarchical structure of the labels, as child labels are more likely to be confused with their sibling categories rather than unrelated ones. This is also challenging for non-professional scientists to distinguish. However, this pattern is not observed in eGeMAPS acoustic model, despite its learning from the acoustic features of audio. These impressive results stem from training a single classifier head on models which pre-trained on general data that may not be ideally suited for such a fine-grained task. They demonstrate the proposed methods' potential. I anticipate a significant performance boost in future work with the addition of more annotated data.

The pretrained acoustic model (AST) achieved the second-best performance across both metrics, surpassing 56.7% accuracy higher than the majority baseline (over 500% F1 score) and 23.9% accuracy (46.1% F1 score) higher than the eGeMAPS acoustic feature baseline. This result marks a strong initial performance. Additionally, AST shows a performance boost compared to ViT, with 23.9% higher accuracy and a 29.3% higher F1 score. While ViT demonstrates significant capability in vocalization recognition, emotional cues may lie in finer details that are more species-specific and less visually apparent than those in vocalizations. AST model, pre-trained on AudioSet with diverse audio data, has learned to distinguish different dog emotional states. Despite the promising results compared to the baseline, there is still considerable room for further improvement.

The multimodal (full-vid) achieve best results of all baselines and models, it achieves a 1.1% higher F1 score and higher 6% accuracy than AST, which suggests it is better suited for this dataset. Context adds more information for identifying correct emotions, and AST may be biased toward larger labels due to the imbalanced dataset. Furthermore, using full video inputs allows the model to exploit dynamic visual signals, such as movement and contextual interactions, that can contribute significantly to overall performance. However, the results remain relatively modest. One possible explanation is the large number of occluded videos in this dataset, where dogs are obscured by factors such as furniture or rear-facing camera angles. These occlusions hinder the visual encoder's ability to fully and effectively extract visual features on dogs' bodies. Additionally, this performance gap may stem from the pre-trained visual model itself, which is optimized for general context understanding and instructional videos that may struggle with fine-grained features, such as subtle emotional cues.

Although I anticipated that the multimodal pipeline would benefit from

the context provided by the visual features, this did not result in a significant performance boost. I also face challenges of noise such as video shaking, blurring, etc. Since the data comes from YouTube, there are no strict recording policies in place, which limits the ability to fully exploit those visual features. Another reason could be due to the late fusion strategy, the audio modality plays a more dominant role than the visual modality in this context of dog emotion prediction (pre-trained acoustic is over 70% accuracy and 100% f1 score higher than pre-trained visual). In this work, I assume that each modality equally contributes to classification results by using concatenation, and future research on fusion strategy and the unequal contribution of these modalities presents great opportunities. Expanding the dataset size and quality with visual annotation could improve these results for further finetuning.



(a) (b) (c)

Figure 5.1: Comfusion maxtrix from eGeMAPS (acoustic) (a), pre-trained acoustic (b), and multimodal(full-vid)(c) from test set. (Label 0: bark/aggression, 1: bark/attention, 2: bark/conflict, 3: whine/anxiety, 4: whine/attention, 5: growl/aggressive, 6: growl/conflict)

5.2 Experiment: Multimodal fusion for effective representation

To effectively integrate information from multiple modalities, I investigate several fusion strategies for multimodal classification. Specifically, I compare three commonly used techniques: concatenation (concat), element-wise addition (add), and gated fusion (gated). Concatenation merges modality features by simply appending them, treating all modalities equally. While it is easy to implement, it fails to capture the intrinsic correlations between

modalities. Additive fusion combines modality embeddings via element-wise addition, aiming to integrate complementary information into a unified representation. However, both of these approaches may struggle to model complex cross-modal dependencies effectively. More advanced methods, such as cross-attention, allow one modality to attend to relevant features in another, enabling richer interaction. Yet, such methods often require large and clean datasets to achieve high performance, which may not be feasible in this case. Given the limited size and noisy nature of this dog social signal dataset, overly complex fusion strategies may not yield optimal results. Arevalo et al. [3], introduced Gated Multimodal Unit (GMU), which introduces learnable gates that dynamically regulate the contribution of each modality to the output of hidden units and benefit the model's final decision. Motivated by these insights, we designed a joint feature modulation mechanism to our multimodal fusion that enables the model dynamically learns to control the combination features based on the input data and effectively fuse multimodal features from pretrained models. Generic pretrained models often generate feature representations that include redundant or irrelevant information for emotion classification, which can negatively impact performance. However, both modalities still contain valuable complementary information. Our approach aims to amplify the shared features among modalities that express the target emotion, while suppressing irrelevant or conflicting signals. To achieve these goals, we proceed as follows: the unimodal embeddings are first projected into a shared high-dimensional space, a joint feature modulation (JFM) combines these high-dimensional features and the fused representation is fed through one linear layer classifier to make prediction. Our JFM is fomulated as:

$$
\begin{aligned}
h_v &= \sigma_v(W_v x_v) \\
h_a &= \sigma_a(W_a x_a) \\
z &= \sigma_z(W_z[h_v, h_a]) \\
h &= z \cdot (h_v + h_a)
\end{aligned}
$$

For visual embedding x_v and audio embedding x_a, with corresponding transformation weights W_v and W_a, the gated fusion mechanism computes hidden representations h_v and h_a through non-linear activations σ_v and σ_a, respectively. A gating vector z is computed using a joint transformation over the concatenated representations $[h_v, h_a]$, and the final representation h is obtained as a weighted combination of h_v and h_a, controlled by the learned gate z. The resulting joint multimodal representation is fed through a one-layer linear classifier to make predictions. This approach introduces learnable

Table 5.3: Comparision results of dog social signal classification amongst differeren fusion methods. Numbers in bold are the best metric values recorded, while those underlined are the second-best ones.

Feature	**Model**	**Pretrained**	**Fusion**	**Accuracy$_7$**	**F1$_7$**
Vid. + Spec.	S3D-AST	Howto100M-AudioSet	concat.$_{1024}$	0.532 ± 0.031	0.481 ± 0.051
Vid. + Spec.	S3D-AST	Howto100M-AudioSet	add.$_{1024}$	$\underline{0.546 \pm 0.028}$	$\underline{0.492 \pm 0.033}$
Vid. + Spec.	S3D-AST	Howto100M-AudioSet	GMU.$_{1024}$	0.526 ± 0.024	0.452 ± 0.022
Vid. + Spec.	S3D-AST	Howto100M-AudioSet	JFM.$_{1024}$	$\mathbf{0.553 \pm 0.020}$	$\mathbf{0.496 \pm 0.029}$

gates that dynamically scale the combination from both modalites enhance fused embedding and benefit the model's final decision. This makes it particularly suitable for my setting with constrained data and varying modality quality.

To evaluate the impact of the proposed fusion strategies on this dataset, I use models with simple concatenation as baselines. Each unimodal encoder is pretrained and frozen for transfer learning, following the setup described in Section 5.1. Unimodal embeddings are first projected into a shared latent space by a nonlinear mapping layer. All experiment uses the same architecture: a unimodal mapper, a specific fusion technique, and a final classifier head. The fused multimodal embedding, which has dimension d, is then passed to the classifier. Our notation reflects this dimension; for instance, concatenation with a latent dimension of 1024 is denoted as concat_{1024}. This design enables a controlled comparison of fusion strategies, allowing us to assess how effectively each method integrates information from multiple modalities.

5.2.1 Results

The results are presented in Table 5.3. JFM fusion consistently achieves the best performance across all the experimental settings. This approach outperforms the concat baseline by 3.9% accuracy and 3.1% in terms of F1 score, and it exceeds additive fusion by 1.3% in accuracy and about 0.8% in F1 score. We also experimented with different latent multimodal embedding sizes $[128, 256, 512, 1024]$, and in all settings, the fusion with joint feature modulation consistently outperformed the additive approach with respect to both metrics. While gated fusion benefits from being trained on a high-quality dataset that teaches the model how to control the contribution of each modality and shows remarkable results on various multimodal tasks, our setting involves a small dataset, which, combined with transfer learning, limits its ability to further learn and refine this control. Our JFM has a 5.1%

increase in accuracy and a 9.7% increase in F1 score over GMU fusion. These results suggest that JFM fusion is the most suitable and effective strategy for this dataset.

5.3 Experiments: Improve performance with unimodal feature extractor

From previous results and insights from ablation studies in section 6.1 and section 6.2, I made a few following observations. First, generic models propose promising results; however, they may lack domain specific discriminative power for this task, and in-domain data is key for good performance. Additionally, learning representation typically demands vast amounts of data and significant computational resources. Moreover, strong image pretrained models such as ViT is promising for dog voice representation learning. This motivates my approach that leverages powerful cross-domain pretrained models for initialization, apply in domain pretraining, and then freezes them for downstream classification. This strategy not only reduces the need for extensive pretraining data and computational cost but also has the potential to yield more robust representations, and ultimately leads to improved overall results.

5.3.1 Model architecture overview

The following models are selected as backbones for encoder models::

- Audio modality: For audio data, I continue to use the AST model to process spectrogram representations. Following the same configuration as in earlier experiments, each 10.24-second audio clip is converted into a 1024×128 log-Mel spectrogram and passed through the AST model, producing an audio embedding of shape (B,768), where B is the batch size.

- Video modality: To capture spatiotemporal information from videos, I use the S3D model. A video clip consisting of 32 frames is input into the model to generate a rich video embedding, which captures both spatial and temporal dynamics, resulting in a (B, 1024) video embedding.

To provide a more comprehensive evaluation, I select models with their original pretrained weights as baselines in this experiment. Specifically, unimodal AST pretrained on AudioSet and S3D pretrained on Howto100M are used as baseline models, against which I compare the in-domain pretraining checkpoints.

5.3.2 Experiment settings

In this experiment, I use pretext tasks for in-domain pretraining and evaluate the resulting checkpoints on downstream dog social signal classification tasks. This setup provides a clearer view of the impact of pretraining and facilitates more effective analysis and discussion. The processes of feature extraction and embedding are described in greater detail in sections 4.2.

Pretext Tasks

To further pretrain the unimodal models, I apply modality-specific pretext tasks tailored to each model. We first conduct study with audio modality .

For the audio modality, AST was initialised with ImageNet pretrained weights [54]. I continue pretraining the AST model using a joint discriminative–generative masked spectrogram patch modeling (MSPM) objective, following the design proposed in [21]. This is performed using unlabeled audio data from the AudioSet-Dog subset, as described in Section 3.3.1. For the spectrogram X input, it is first divided into 512 patches x, which are then projected into patch embeddings E. A random set I of N patch indices is selected for masking. For each patch to be masked, its embedding is replaced with a learnable mask embedding E_{mask}. After adding positional embeddings to the resulting patch embeddings, the sequence is fed into the transformer encoder. For each masked patch x_i, the corresponding encoder output $o_i \in R^{768}$, is passed through two two-layer MLP heads that both map it to the same dimension as the original patch $x_i \in R^{256}$. A classification head then maps o_i to a prediction c_i, with the objective of selecting the correct patch for each masked position among all masked candidates. Negative samples are drawn from the same spectrogram, and training is guided by a discriminative InfoNCE loss objective $\mathcal{L}_d$:

$$\mathcal{L}_d = \frac{1}{N}\sum_{i=1}^{N} log(\frac{exp(c_i^T x_i)}{\sum_{j=1}^{N} exp(c_i^T x_j)}) \tag{5.1}$$

A reconstruction head maps o_i to r_i, with expectation that r_i to be close to x_i and using MSE $\mathcal{L}_g$ loss:

$$\mathcal{L}_g = \frac{1}{N}\sum_{i=1}^{N}(r_i - x_i)^2 \tag{5.2}$$

Model total loss is calculated with weights λ =10:

$$\mathcal{L} = \mathcal{L}_d + \lambda\mathcal{L}_g \tag{5.3}$$

Downstream tasks

After continued pretraining with the pretext task, the best checkpoint model was chosen and further evaluated with the downstream task, the dog social signal classification task, which includes 7 classes of dog emotion per voice types. The downstream evaluation follows a similar experimental setup to that described in section 5.1. Pretrained models are frozen and used as feature extractors, while a simple concatenation strategy is applied for multimodal fusion. A classification head is then trained on top of the model for this downstream task.

5.3.3 Experiment results

Experiment results on pretext tasks

Model	Task	**Dataset**	**Pretrained**	**Accuracy**	**F1**	**Loss**
AST	classification	AudioSet dog	ImageNet	0.142	-	0.046

Table 5.4: Pretext evaluation of model continue pretraining for domain adaptation

The early results for the pretext task are summarized in Table 5.4. Currently, the performance is limited, with accuracy reaching only 14% on validation set. This is significantly lower than the 80% reported in the original paper [21]. The primary reasons for this underperformance are time and resource constraints. Specifically, pretraining was conducted with a batch size 16 times smaller and using a dataset over 100 times smaller than the one used in prior work. Additionally, due to time limitations, I were unable to run sufficient training trials to select optimal checkpoints. Future work can address these limitations by allocating more time and computational resources to thoroughly explore the pretraining phase, which is expected to lead to improved performance on both the pretext and downstream tasks.

Experiment results on downstream tasks

The results for the downstream task are shown in Table 5.5. A noticeable drop in performance is observed, which I attribute to the use of suboptimal checkpoints. I believe that with longer pretraining, greater computational resources, and more targeted pretraining tailored to the task, the AST model could learn more robust representations. This, in turn, is expected to enhance performance on the downstream task.

Feature	**Model**	**Pretrained**	**Accuracy$_7$**	**F1$_7$ macro**
Spec.	AST	AudioSet	0.503 ± 0.051	0.472 ± 0.070
Spec.	AST	AudioSet Dog	0.374 ± 0.038	0.301 ± 0.068
Video + Spec.	S3D- AST	Howto100M - AudioSet	0.532 ± 0.027	0.477 ± 0.033
Video + Spec.	S3D- AST	Howto100M - AudioSet dog	0.420 ± 0.063	0.351 ± 0.081

Table 5.5: Performance results of downstream dog social signal classification

Chapter 6

Ablation Study

This section aims to investigate strategies for improving multimodal performance. In particular, I focus on enhancing the unimodal encoders through pretraining to obtain better representations, which can in turn benefit multimodal learning.

The motivation for this approach is that pretrained models have demonstrated strong performance on unseen data (see Sections 5.1, 6.2). Moreover, high-quality annotated bio-data is inherently rare, difficult, and expensive to obtain, which limits the effectiveness of fine-tuning for full adaptation and feature learning. Furthermore, domain adaptation through fine-tuning is computationally expensive, and in cases of significant domain shift, fine-tuning on small datasets can distort useful pretrained features, potentially leading to degraded performance. In contrast, unlabeled data is easier to acquire, and several available datasets can be leveraged to build strong representations. Recent studies have also shown that unsupervised representation learning on large-scale online unlabeled datasets achieves remarkable performance across various benchmarks. However, building a large-scale pretraining dataset is challenging in this field, even when labels are not required. Additionally, recent findings [52] suggest that general-purpose audio pretrained models can perform well on bioacoustic tasks, even without being specifically trained on animal vocalizations. This highlights the potential of cross-domain adaptation to transfer knowledge from robust, high-quality source domains to target domains with limited data. Therefore, I focus on continued pretraining unimodal encoders to address these limitations, enhance unimodal representation and improve overall multimodal learning. To assess this approach, I conduct the following experiments to verify its effectiveness.

6.1 Effect of Layer Unfreezing on Model Performance

To understand the impact of fine-tuning strategies, I record classification results when unfreezing different portions of the pretrained acoustic, AST model. This allows us to evaluate the transferability of learned representations across layers, optimize adaptation strategies, and balance performance gains with computational cost.

It is important to note that I do not apply unfreezing experiments to the multimodal setting at this stage for the following reasons:

- The dataset contains audio-based labels without corresponding visual annotations. From previous experiments, when extracting visual features, I observed that the videos are highly noisy, making the visual features potentially unreliable for current use. As a result, I plan to postpone visual-based experimentation until a later phase when annotated visual data becomes available.
- Fine-tuning a large multimodal model requires significantly more computational resources, which are currently not available for this stage of the project.

Settings are defined by which layers are unfrozen during fine-tuning, where "Frozen" mean the only the classfier head is unfrozen, "L11" indicates only the final transformer block with classifier head is unfrozen, "L10-11" indicates the last two layers with classifier head are unfrozen, and so on, up to "All layers", where all model layers are fine-tuned. In this experiments, the optimal configuration was found to be LP-FT, consisting of 10 epochs of linear probing (LP) with a learning rate of 1e-3, followed by fine-tuning (FT) for up to 30 epochs with early stopping and a learning rate of 3e-5. To accommodate hardware limitations, I employed a batch size of 4 across all settings, with gradient accumulation over 4 steps to simulate a larger effective batch size. Results were evaluated using 5-fold cross-validation.

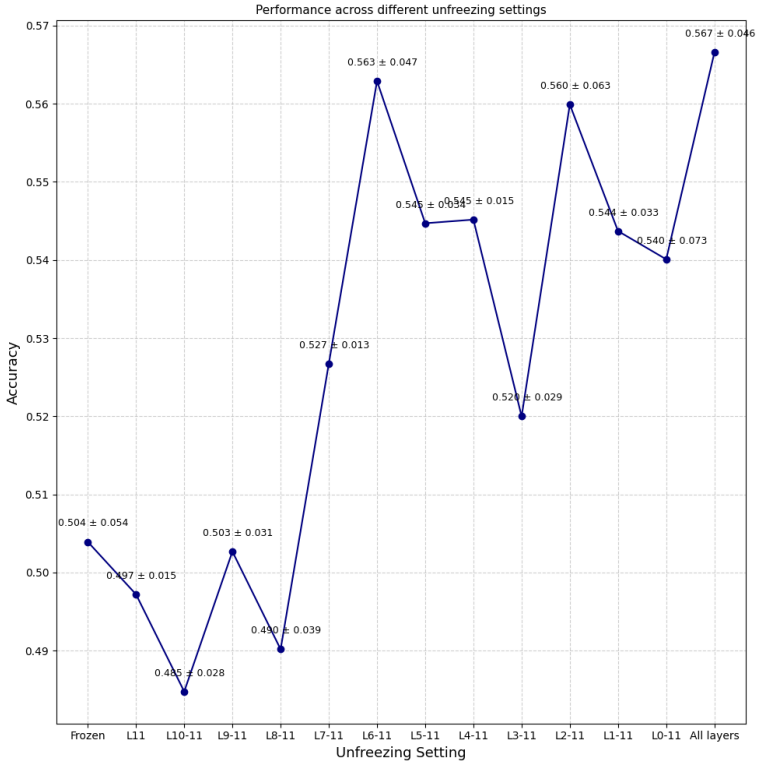
6.1.1 Experiment results

For detailed results, please refer to Table 6.1, and for a visual demonstration, see Figure 6.1, 6.2.

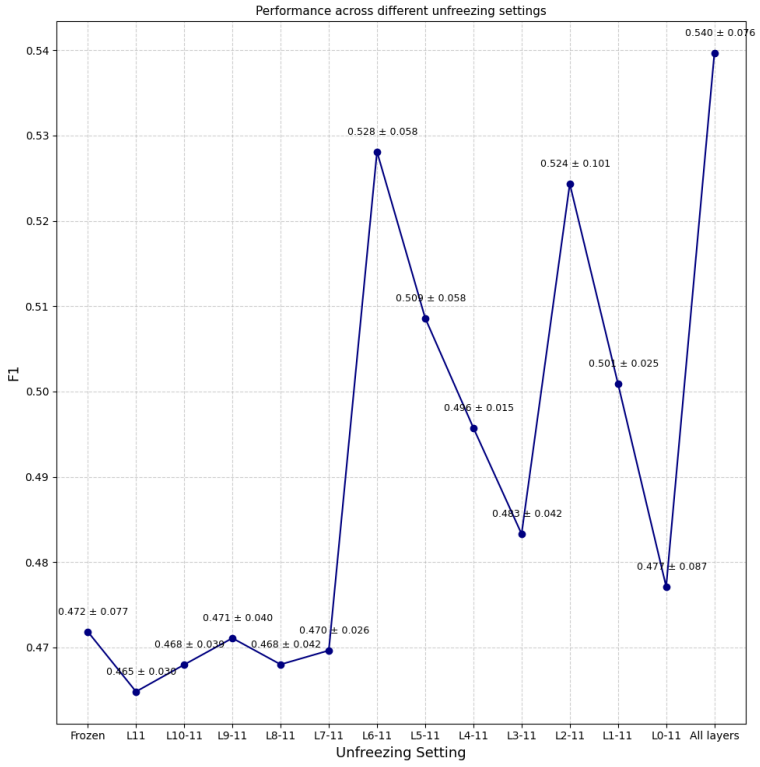
As expected, fully unfreezing all layers of AST yields the best performance, achieving a 12.0% improvement in accuracy and a 14.4% increase in F1 score over the frozen baseline. However, I note that the L6–11 setting (unfreezing layers 6 to 11) performs comparably, with only slightly lower

Table 6.1: Accuracy and F1-score results for progressive unfreezing experiments.

Settings	**Accuracy**$_7$ ↑	**F1**$_7$ ↑	**Loss** ↓
Frozen	0.504 ± 0.054	0.472 ± 0.077	1.163 ± 0.119
L11	0.497 ± 0.015	0.465 ± 0.030	**1.153 ± 0.084**
L10-11	0.485 ± 0.028	0.468 ± 0.039	1.282 ± 0.213
L9-11	0.503 ± 0.031	0.471 ± 0.040	1.444 ± 0.174
L8-11	0.490 ± 0.039	0.468 ± 0.042	1.286 ± 0.237
L7-11	0.527 ± 0.013	0.470 ± 0.026	1.714 ± 0.292
L6-11	0.563 ± 0.047	0.528 ± 0.058	1.669 ± 0.250
L5-11	0.545 ± 0.034	0.509 ± 0.058	1.544 ± 0.325
L4-11	0.545 ± 0.015	0.496 ± 0.015	1.561 ± 0.338
L3-11	0.520 ± 0.029	0.483 ± 0.042	1.623 ± 0.340
L2-11	0.560 ± 0.063	0.524 ± 0.101	1.696 ± 0.359
L1-11	0.544 ± 0.033	0.501 ± 0.025	1.628 ± 0.288
L0-11	0.540 ± 0.073	0.477 ± 0.087	1.592 ± 0.407
All layers	**0.567 ± 0.046**	**0.540 ± 0.076**	1.717 ± 0.346



(a)



(b)

Figure 6.1: Performance across different unfreezing configurations. Subfigures show (a) accuracy and (b) macro F1-score, for each unfreezing setting. Settings range from unfreezing only the classifier layer (leftmost) to unfreezing all layers (rightmost).

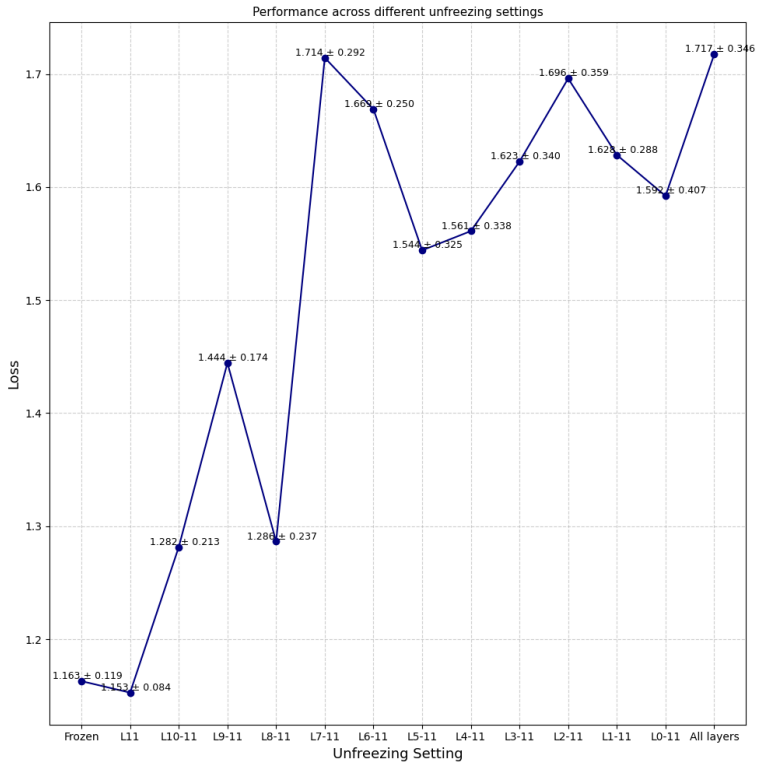


Figure 6.2: Testing loss across different unfreezing configurations. Settings range from unfreezing only the classifier layer (leftmost) to unfreezing all layers (rightmost).

gains (11.7% in accuracy and 11.9% in F1) while requiring approximately six fewer transformer blocks to be fine-tuned. Despite these performance improvements, both settings exhibit significantly higher losses: the fully unfrozen model has a 47.6% increase in loss, and L6–11 has a 43.5% increase, compared to the frozen encoder. This suggests that while the models predict more correct samples, they also become more prone to confidently making incorrect predictions, an undesirable trait for a well-generalized model.

Moreover, the frozen setting requires significantly less training time and computational resources compared to all other configurations. Although the reported results use the same training setup, such as batch size and number of epochs, to allow fair comparison of performance, I also conducted experiments adapted to available hardware with 16GB of VRAM to assess practical efficiency. Under these constraints, the frozen setting completed training approximately three times faster than L6–11 and supported a batch size twice as large. Moreover, L6–11 required only about 10% less training time than the fully unfrozen model, and it also accommodated a batch size twice as large.

More fine-tuning does not necessarily lead to better performance. Excessive fine-tuning can distort the useful representations learned during pretrain-

ing, while also significantly increasing computational cost. Prior studies have explored how many and which layers should be fine-tuned for downstream tasks. In this case, fine-tuning layers L6–11 appears to strike a good balance, preserving general knowledge from pretraining while allowing effective task-specific adaptation. However, fine-tuning on a small and imbalanced dataset can lead to overfitting, resulting in minimal performance gains. Furthermore, unfreezing additional layers, particularly those closer to the input (i.e., lower layers), further increases the computational burden without necessarily improving generalization.

While the LP-FT approach yields a modest performance gain, it also increases the model's confidence in incorrect predictions. This is a sign of reduced generalization. These suggest that the only viable path toward substantial improvement for this approach is the introduction of a significantly larger dataset. This highlights the challenges of working with limited data and the lack of abundant pretrained models in this domain. In light of these constraints, linear probing or using a frozen pretrained model proves to be a strong and promising baseline across modalities and domains. By leveraging the rich and generalizable acoustic representations learned by large-scale audio models, this approach reduces data requirements, computation resources, and potentially achives good results, offering a practical foundation for further development. However, challenges remain, such as the mismatch between the pretraining domain and the target task, and the possible loss of important task-specific information. Improving the feature extractor may help address these problems. I also explored using self-supervised learning with unlabeled data to learn dog-specific voice representations through cross-domain adaptation. This approach can reduce the amount of labeled data needed for downstream tasks and lower the need for a very large pretrained model and dataset. It shows promise for improving performance in settings with limited data, which is especially helpful for underexplored areas. Although this part of my work is still ongoing, early results are shown in Experiments 6.2 and 5.3.

6.2 Cross-domain adaptation for acoustic representation learning

Since image and spectrogram have similar formats, many researchers have explored cross-domain transfer learning from vision to audio domain. Audio Spectrogram Transformer [19] shows the effectiveness of ImageNet pretraining, which can help reduce the necessity for in-domain audio data and pro-

duce remarkable results even without AudioSet [17] pretraining for speech classification tasks. Moreover, recent studies on self-supervised learning (SSL) models pre-trained on human speech have shown remarkable success towards bioacoustic data [51]. These models capture latent features to produce meaningful representations of human speech. Given that both humans and animals have a similar voice production system, even without finetuning, SSL pre-trained models can produce latent embeddings for effective bioacoustic classification [50, 9]. While the bioacoustic domain faces data scarcity, domains such as image, and human speech are well-supported with extensive datasets and readily available pre-trained models. This part is dedicated to analyze how different domain pre-trained models model dog vocalization, to assess whether their vocalization capabilities can be effectively leveraged for emotion classification. Additionally, since I intend to develop an audio-pretrained model for this task using unlabeled data, I are motivated to study the effects of leveraging both ImageNet-pretrained models applied to spectrograms like ViT[11] and human speech-pretrained models like Wav2Vec 2.0 [4], to assess whether they can provide strong starting points and enhance the effectiveness of audio pretraining.

In this part, I adopt Vision Transformer (ViT) pre-trained on ImageNet, and Wav2vec2 , a SOTA self-supervised model for Automatic Speech Recognition, for a dog emotion classification study to explore the ability of these pre-trained models on capture abstract representation of cross-domain acoustic data. I aim to investigate the potential application of these cross domain pretrained models as initializations for dog social signal classification.

6.2.1 Experiment settings

Both models are frozen and train only the classifier head to see how I can transfer pre-trained knowledge on bioacoustic data. The Wav2Vec2 model takes raw audio as input, whereas for the ViT model, I preprocess the spectrogram into an approximate 3-channel RGB format using Matplotlib [24]. To evaluate model performance, I establish a majority-class baseline, commonly applied when dealing with imbalanced data. Both models were trained for 30 epochs with 5-fold cross-validation, following the same settings as in the experiment 5.1.

6.2.2 Cross-domain transfer via pretrained feature representations results

Results from Table 6.2 show impressive results for both ViT and Wav2vec2 with respectively 26.5%, and 36.8% improvement over majority baseline.

Table 6.2: Cross domain transfer learning results of Image pretraining model (ViT) and human speech pretraining model (Wav2vec2). Both model's pretrained weights are frozen for feature extraction and training classifier head for dog social signal recognition.

Domain	**Accuracy**	**F1**	**%Improve**
Majority	0.321 ± 0.0	0.069 ± 0.0	-
Image (ViT)	0.406 ± 0.048	0.365 ± 0.082	26.480%
Speech (Wav2vec2)	0.439 ± 0.046	0.393 ± 0.056	36.760%

Table 6.3: Dog vocalization results. A classification between bark, whine and growl on ViT model.

Layer Finetuning	**Accuracy**	**F1**
classifier head	0.898	0.825
full finetuning	0.933	0.904

Even for dog audio, Wave2vec is able to capture discriminative acoustic features of bioacoustic data, and satisfactorily classify a dog's inner state using those latent presentations with a minimum of only training classification. Despite the huge gap between the audio spectrogram and ImageNet image, ViT performs relatively well with more than 25% higher than the baseline, suggesting that this image pretrained model can capture some distinct traits that help differentiate between categories. In further investigation, I experiment with 3 classes of dog vocalization (bark, whine and growl) on ViT. The results from Table 6.3 shows that ViT yields a remarkable performance of 89.8% accuracy and 82.5% f1 score (macro average) when training only classification head. I conclude that vocal differences are visually discriminated in spectrogram, and distinguishing between emotion classes remains a challenge, which I aim to address through further refinement of my classification approach.

Even though both of the models have never experimented with such data, these results of only training classifier heads of cross-domain pre-trained weights have relatively impressive results. This could serve as a strong baseline for further investigation. This also raises the question of how performance could be further improved. One possibility is the inclusion of more labeled data for further finetuning; however, collecting such data is particu-

larly challenging and beyond the scope of my current work. I consider this an important direction to revisit in future research. While cross-domain transfer learning yields interesting results, the limited data currently available poses challenges for in-depth analysis, the field remains an exciting direction for continued research.

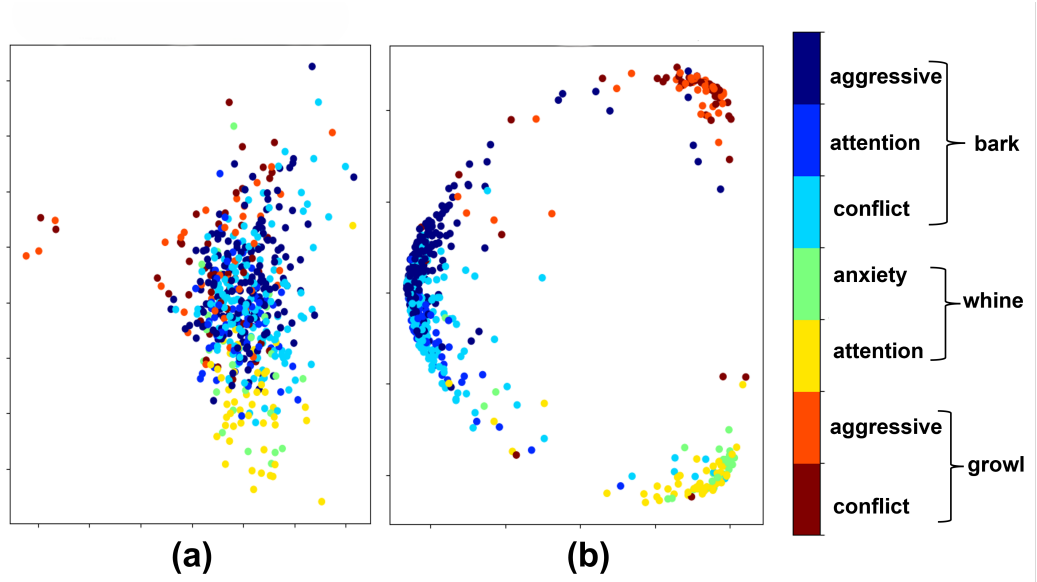


Figure 6.3: Visualization from eGeMAPS features embedding and ViT fine-tuning embedding using PCA(2)

Through experiments, I made the following observations: Even though the pre-trained acoustic (AST model) cannot classify dog social signals perfectly, this model demonstrates strong performance in dog vocalization. This phenomenon is observed not only in AST but also in its pre-trained backbone, ImageNet pre-trained ViT, as evidenced by the results presented in Table 6.3 from the study of cross-domain transfer learning in section 6.2.2. These results provide evidence that vocal differences are visually displayed in the spectrogram and effectively captured by these models. Audio processing using machine learning is highly applicable and currently achieves high results on several benchmarks. However, they are mostly centered on humans, and from these results, I understand that these techniques could also applied well to non-human species such as canine vocalization. For deeper understanding, I perform dimension reduction on ViT embedding which is finetuned with the dog social signal dataset, and open smile features embedding, then visualize using the PCA technique. From Figure 6.3, I notice that embedding from the ViT model creates 3 parent clusters for bark, whine, and

growl. These clusters are well-separated, meanwhile, the barrier between intraclasses is ambiguous. This naturally explains barking/aggression is more similar to barking/conflict than growl/aggression. Unfortunately, I can not detect such behavior in the Opensmile features set. These results provide additional evidence that makes these models a compelling choice for learning discriminative dog vocalization embedding.

These results show that strong pretrained models, such as those trained on images, can serve as effective initializations for domain adaptation, consistent with findings from prior work [19].

Chapter 7

Conclusion

7.1 Summary

Working with a small-scale and noisy dataset, I proposed a promising multimodal approach that leverages pretrained models to improve performance. Experimental results demonstrate the effectiveness of this method, as the multimodal setup using pretrained unimodal feature extractors outperforms all baselines. Additionally, the proposed joint feature modulation allows the model to dynamically control the combined feature representation. This scaled additive fusion enhances complementary modality information, suppresses irrelevant or conflicting signals, and produces a richer multimodal representation from the pretrained model, which ultimately leads to more accurate predictions. On the other hand, the experimental results reveal that current state-of-the-art multimodal approaches perform suboptimally on this dog social signal dataset, underscoring the complexity and challenges of this task. The limited performance reflects issues related to both dataset quality and model generalization. Key contributing factors include data scarcity, class imbalance, and high levels of noise, all of which hinder effective learning. The ablation study with gradual unfreezing reveals that, for small and noisy datasets, frozen pretrained models deliver superior performance. Additionally, experiments using audio and image-pretrained models independently show strong results in capturing the discriminative characteristics of dog vocalizations, despite a significant domain gap. These findings suggest considerable potential for further improvement through the use of unlabeled data to enable more effective domain adaptation via pretraining from cross-domain pretrained models and transfer learning for downstream classification, leading to more robust and generalizable representations and reducing the need of a extremely large pretraining dataset and resources. However, due to time

and resource constraints, I were unable to complete these experiments in the current study. Future research with sufficient resources may build on this approach to fully realize its benefits.

7.2 Contribution

In conclusion, contributions of this thesis is as following:

- This work proposes multimodal approach that models dog social signals through acoustic and visual modalities for classification. To address the limitations of a small-scale and noisy dataset, this work leverages transfer learning with frozen generic pretrained models that are publicly available.
- This study conduct experiments with other baselines, which not only establish an initial benchmark for the task but also demonstrate the effectiveness of the proposed method.
- This work proposes a joint feature modulation (JFM) for information fusion, a scaled additive fusion to enhance the fused information, and improve multimodal model performance.
- Through ablation studies, this research proposes an approach to improve the multimodal pipeline by enhancing unimodal feature representations via self-supervised in-domain pretraining from cross-domain pretrained model. This method aims to produce more robust and generalizable representations, while reducing the reliance on extremely large pretraining datasets and computational resources.

7.3 Future work

While the initial results highlight the challenges posed by the dataset, they also provide a foundation for future animal-human-computer research. Our benchmark models establish a starting point, and there is significant room for improvement. Future work may explore advanced deep learning methods, domain-specific adaptations, and improved fusion strategies to enhance model performance in challenging scenarios, such as missing or low-quality modalities. Moreover, in-domain pretraining from cross-domain pretrained models is encouraged, as it leverages the strength of existing models to build more robust representations while reducing the need for extensive pretraining resources and large datasets. Additionally, refining and expanding the

dataset could enhance generalization and robustness. We hope this study serves as a valuable reference for continued exploration toward more robust and efficient methods in this domain.

Acknowledgements

This thesis was prepared according to the curriculum for the Collaborative Education Program organized by Japan Advanced Institute of Science and Technology and Ho Chi Minh University Of Science, Vietnam National University.

To complete this graduation thesis, I would like to express my sincere gratitude for the support of individuals, collectives, and organizations both inside and outside the university who have facilitated and assisted us throughout my study and training period at the university.

With my deepest appreciation, I would like to sincerely thank the lecturers and staff at the Faculty of Information Technology, University of Science – Vietnam National University Ho Chi Minh City, and the School of Information Science, JAIST, for their support and for creating favorable conditions during my studies, enabling us to gather necessary information to prepare, implement, and complete this graduation thesis to the best of my abilities.

Finally, I would like to express my heartfelt thanks to my family, friends, and everyone around who have always been by my side, supporting us both materially and spiritually whenever I faced difficulties during my studies and while conducting this thesis.

This graduation thesis was carried out during the second year of the master's program while the student was studying at JAIST. As this is my first work on this problem and given the time constraints, there are inevitably many limitations and shortcomings. I sincerely hope to receive valuable feedback and suggestions from my lecturers, friends, and families to further improve my knowledge in this topic and field.

I sincerely thank you all!

Bibliography

[1] Abzaliev, A., Espinosa, H.P., Mihalcea, R.: Towards dog bark decoding: Leveraging human speech processing for automated bark classification. arXiv preprint arXiv:2404.18739 (2024)

[2] Amat, M., Le Brech, S., Camps, T., Manteca, X.: Separation-related problems in dogs: A critical review. Advances in Small Animal Care **1**, 1–8 (2020)

[3] Arevalo, J., Solorio, T., Montes-y Gómez, M., González, F.A.: Gated multimodal networks. Neural Comput. Appl. **32**(14), 10209–10228 (Jul 2020). https://doi.org/10.1007/s00521-019-04559-1, `https://doi.org/10.1007/s00521-019-04559-1`

[4] Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)

[5] Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence **41**(2), 423–443 (2018)

[6] Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Icml. vol. 2, p. 4 (2021)

[7] Boneh-Shitrit, T., Feighelstein, M., Bremhorst, A., Amir, S., Distelfeld, T., Dassa, Y., Yaroshetsky, S., Riemer, S., Shimshoni, I., Mills, D.S., et al.: Explainable automated recognition of emotional states from canine facial expressions: the case of positive anticipation and frustration. Scientific reports **12**(1), 22611 (2022)

[8] Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)

[9] Cauzinille, J., Favre, B., Marxer, R., Clink, D., Ahmad, A.H., Rey, A.: Investigating self-supervised speech models' ability to classify animal vocalizations: The case of gibbon's vocal signatures. In: Interspeech 2024. pp. 132–136. ISCA; ISCA (2024)

[10] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)

[11] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. ICLR (2021)

[12] Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., et al.: The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE transactions on affective computing **7**(2), 190–202 (2015)

[13] Eyben, F., Wöllmer, M., Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor. In: Proceedings of the 18th ACM international conference on Multimedia. pp. 1459–1462 (2010)

[14] Faragó, T., Takács, N., Miklósi, Á., Pongrácz, P.: Dog growls express various contextual and affective content for human listeners. Royal Society open science **4**(5), 170134 (2017)

[15] Feichtenhofer, C., Fan, H., Malik, J., He, K.: Slowfast networks for video recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6202–6211 (2019)

[16] Franzoni, V., Biondi, G., Milani, A.: Advanced techniques for automated emotion recognition in dogs from video data through deep learning. Neural Computing and Applications **36**(28), 17669–17688 (2024)

[17] Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 776–780. IEEE (2017)

[18] Gómez-Armenta, J.R., Pérez-Espinosa, H., Fernández-Zepeda, J.A., Reyes-Meza, V.: Automatic classification of dog barking using deep learning. Behavioural Processes **218**, 105028 (2024)

[19] Gong, Y., Chung, Y.A., Glass, J.: AST: Audio Spectrogram Transformer. In: Proc. Interspeech 2021. pp. 571–575 (2021). https://doi.org/10.21437/Interspeech.2021-698

[20] Gong, Y., Chung, Y.A., Glass, J.: Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3292–3306 (2021)

[21] Gong, Y., Lai, C.I., Chung, Y.A., Glass, J.: Ssast: Self-supervised audio spectrogram transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 10699–10709 (2022)

[22] Hagiwara, M.: Aves: Animal vocalization encoder based on self-supervision. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)

[23] Hantke, S., Cummins, N., Schuller, B.: What is my dog trying to tell me? the automatic recognition of the context and perceived emotion of dog barks. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5134–5138. IEEE (2018)

[24] Hunter, J.D.: Matplotlib: A 2d graphics environment. Computing in Science & Engineering **9**(3), 90–95 (2007). https://doi.org/10.1109/MCSE.2007.55

[25] Jégh-Czinege, N., Faragó, T., Pongracz, P.: A bark of its own kind – the acoustics of 'annoying' dog barks suggests a specific attention-evoking effect for humans. Bioacoustics **29**, 1–16 (02 2019). https://doi.org/10.1080/09524622.2019.1576147

[26] Kumar, A., Raghunathan, A., Jones, R.M., Ma, T., Liang, P.: Fine-tuning can distort pretrained features and underperform out-of-distribution. In: International Conference on Learning Representations (2022), `https://openreview.net/forum?id=UYneFzXSJWh`

[27] Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. biometrics pp. 159–174 (1977)

[28] Lenkei, R., Alvarez Gomez, S., Pongracz, P.: Fear vs. frustration ? possible factors behind canine separation related behaviour. Behavioural Processes **157**, 115–124 (2018). https://doi.org/https://doi.org/10.1016/j.beproc.2018.08.002, https://www.sciencedirect.com/science/article/pii/S0376635718300603

[29] Lenkei, R., Faragó, T., Bakos, V., Pongrácz, P.: Separation-related behavior of dogs shows association with their reactions to everyday situations that may elicit frustration or fear. Scientific Reports **11**(1), 19207 (2021)

[30] Li, P., Song, Y., McLoughlin, I., Guo, W., Dai, L.: An attention pooling based representation learning method for speech emotion recognition. In: Interspeech 2018. pp. 3087–3091 (2018). https://doi.org/10.21437/Interspeech.2018-1242

[31] McGurk, H., MacDonald, J.: Hearing lips and seeing voices. Nature **264**(5588), 746–748 (1976)

[32] Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In: CVPR (2020)

[33] Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In: ICCV (2019)

[34] Mohandas, P., Anni, J.S., Hasikin, K., Velauthapillai, D., Raj, V., Murugathas, T., Azizan, M.M., Thanasekaran, R.: Machine learning approach regarding the classification and prediction of dog sounds: A case study of south indian breeds. Applied Sciences **12**(20), 10653 (2022)

[35] Molnár, C., Kaplan, F., Roy, P., Pachet, F., Pongrácz, P., Dóka, A., Miklósi, Á.: Classification of dog barks: a machine learning approach. Animal Cognition **11**, 389–400 (2008)

[36] Molnár, C., Pongrácz, P., Miklósi, A.: Seeing with ears: Sightless humans' perception of dog bark provides a test for structural rules in vocal communication. Quarterly Journal of Experimental Psychology **63**(5), 1004–1013 (2010)

[37] Newmark, C.: Charles darwin: the expression of the emotions in man and animals. In: Schlüsselwerke der Emotionssoziologie, pp. 111–115. Springer (2022)

[38] Panigrahi, S., Nanda, A., Swarnkar, T.: A survey on transfer learning. In: Intelligent and Cloud Computing: Proceedings of ICICC 2019, Volume 1, pp. 781–789. Springer (2020)

[39] Papa, L., Russo, P., Amerini, I., Zhou, L.: A survey on efficient vision transformers: algorithms, techniques, and performance benchmarking. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)

[40] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[41] Petajan, E.D.: Automatic lipreading to enhance speech recognition. In: Proc. IEEE-CS Conference on Computer Vision and Pattern Recognition. pp. 40–47 (1985)

[42] Picard, R.W.: Affective computing. MIT press (2000)

[43] Piczak, K.J.: Esc: Dataset for environmental sound classification. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 1015–1018 (2015)

[44] Poggi, I., D'Errico, F.: Social signals: A psychological perspective. In: Computer analysis of human behavior, pp. 185–225. Springer (2011)

[45] Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Information fusion **37**, 98–125 (2017)

[46] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)

[47] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of machine learning research **21**(140), 1–67 (2020)

[48] Ravanbakhsh, E., Liang, Y., Ramanujam, J., Li, X.: Deep video representation learning: a survey. Multimedia Tools and Applications **83**(20), 59195–59225 (2024)

[49] Rybakov, O., Kononenko, N., Subrahmanya, N., Visontai, M., Laurenzo, S.: Streaming keyword spotting on mobile devices. In: Interspeech 2020. pp. 2277–2281 (2020). https://doi.org/10.21437/Interspeech.2020-1003

[50] Sarkar, E., Magimai.-Doss, M.: Can Self-Supervised Neural Representations Pre-Trained on Human Speech distinguish Animal Callers? In: Proc. INTERSPEECH 2023. pp. 1189–1193 (2023). https://doi.org/10.21437/Interspeech.2023-1968

[51] Sarkar, E., Magimai.-Doss, M.: On the utility of speech and audio foundation models for marmoset call analysis. In: 4th International Workshop on Vocal Interactivity In-and-between Humans, Animals and Robots (VIHAR2024) (2024). https://doi.org/10.5281/zenodo.13935495

[52] Sarkar, E., Magimai.-Doss, M.: Comparing self-supervised learning models pre-trained on human speech and animal vocalizations for bioacoustics processing. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025). https://doi.org/10.1109/ICASSP49660.2025.10889684

[53] Schiappa, M.C., Rawat, Y.S., Shah, M.: Self-supervised learning for videos: A survey. ACM Computing Surveys **55**(13s), 1–37 (2023)

[54] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: International conference on machine learning. pp. 10347–10357. PMLR (2021)

[55] Townsend, L., Gee, N.R.: Recognizing and mitigating canine stress during animal assisted interventions. Veterinary Sciences **8**(11), 254 (2021)

[56] Training, N.L.D.: What is emotional intelligence focus in dog training? (2024), `https://nextleveldogtraining.co.uk/emotional-intelligence-focus-in-dog-train` last accessed 02 May 2025

[57] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[58] Vinciarelli, A., Pantic, M., Bourlard, H., Pentland, A.: Social signal processing: state-of-the-art and future perspectives of an emerging domain. In: Proceedings of the 16th ACM international conference on Multimedia. pp. 1061–1070 (2008)

[59] Waller, B., Correia Caeiro, C., Peirce, K., Burrows, A., Kaminski, J.: Dogfacs: the dog facial action coding system (2013)

[60] Warden, P.: Speech commands: A dataset for limited-vocabulary speech recognition. arxiv 2018. arXiv preprint arXiv:1804.03209 (1804)

[61] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. Association for Computational Linguistics, Online (Oct 2020), https://www.aclweb.org/anthology/2020.emnlp-demos.6

[62] Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 305–321 (2018)

[63] Zaman, K., Sah, M., Direkoglu, C., Unoki, M.: A survey of audio classification using deep learning. IEEE Access **11**, 106620–106649 (2023)