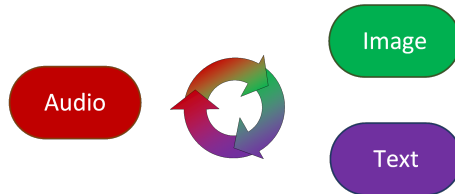


Dissertation

Audio Analysis for Multimodal Learning



Thesis approved by the Department of Computer Science
University of Kaiserslautern-Landau for the award of the Doctoral Degree

Doctor of Natural Sciences
(Dr. rer. nat.)

to

Andrey Guzhov

Reviewers Prof. Dr. Prof. h.c. Andreas Dengel
 Prof. Dr. Sebastian Vollmer

Date of Defense 20.09.2024
Dean of the Dept. Prof. Dr. Christoph Garth

Andrey Guzhov

Audio Analysis for Multimodal Learning

Dissertation, 20.09.2024

Reviewers: Prof. Dr. Prof. h.c. Andreas Dengel, Prof. Dr. Sebastian Vollmer

Chair: Prof. Dr. Leo van Waveren

Supervisor: Prof. Dr. Prof. h.c. Andreas Dengel

University of Kaiserslautern-Landau

AG Artificial Intelligence

Department of Computer Science

Gottlieb-Daimler-Straße

67663 Kaiserslautern

Abstract

Multimodal learning has received much attention in past years, and many advances in multimedia analysis and search have been accomplished. However, most focus was on the visual content. The audio signal, being arguably the second most important modality, also carries a lot of useful information that allows to improve the general performance of machine learning models. This has been indeed the case for a variety of applications, e.g., action recognition or scene classification. This work is dedicated to the challenging aspects found in environmental audio-visual analysis, and how one can design cross-modal fusion methods that solve tasks like classification or retrieval in a truly multimodal fashion.

The thesis presents a comprehensive overview of the deep learning landscape in audio-visual learning, encompassing essential tasks such as representation learning, correspondence learning, audio-visual object localization, sound source separation, and audio-visual generation. Additionally, specific audio-centric learning objectives, including audio-visual multimodal classification, regular and zero-shot environmental sound classification, and multimodal retrieval, are extensively explored to provide a holistic understanding of the current state-of-the-art in deep learning for audio-visual analysis.

Building upon this foundation, the thesis delves into specific techniques and methods tailored for solving these tasks, leveraging insights gained from the current research landscape. The research objective is being addressed by investigating effects of using visual-domain models for audio-specific tasks, the effectiveness of audio-specific architecture enhancements in deep learning models, and the shift from transfer learning to audio-visual contrastive learning. The experimental results demonstrate the significant performance and robustness improvements achieved through these approaches.

In conclusion, this doctoral thesis contributes to the field of audio analysis in multimodal learning by discovering the untapped potential of audio signals and highlighting future research directions that can further advance the field. By bridging the gap between audio and visual modalities, this thesis opens up new avenues for more effective and comprehensive multimodal learning systems.

Acknowledgement

Pursuing this thesis was a marathon of discovery, challenges, and growth. I want to express my heartfelt gratitude to everyone who made this journey not only feasible but enriching. First and foremost, I would like to thank my advisor Prof. Andreas Dengel for his guidance, support, and for the opportunity to evolve as a professional. Special thanks to Christian whose guidance throughout first steps in my PhD set the foundation for shaping the research. I would like to thank my co-authors Jörn and Federico whose experience and insightful advices have left an enduring impact on my growth as a researcher. I extend my gratitude to Sebastian, Fatemeh, Christiano, Joachim, Dayananda, Stanislav, and other colleagues from the German Research Center for Artificial Intelligence (DFKI) for fruitful discussions and valuable feedback. Turning to my family, I thank all of them for their unwavering support, belief, and love. To my fiancée, Sofya, and my parents, Svetlana and Alexey, your belief in me, your support, and encouragements have been my driving force. I would like to say a special thank you for the apt advice and life-affirming witticisms to my cousin Elena, whose cats grace the pages of this thesis. I love you all dearly.

List of Publications

1. A. Guzhov, F. Raue, J. Hees and A. Dengel, “ESResNet: Environmental Sound Classification Based on Visual Domain Models,” 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 4933-4940, doi: 10.1109/ICPR48806.2021.9413035.
2. A. Guzhov, F. Raue, J. Hees and A. Dengel, “ESResNe(X)t-fbsp: Learning Robust Time-Frequency Transformation of Audio,” 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 2021, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9533654.
3. A. Guzhov, F. Raue, J. Hees and A. Dengel, “AudioCLIP: Extending CLIP to Image, Text and Audio,” ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 976-980, doi: 10.1109/ICASSP43922.2022.9747631.

Contents

1	Introduction	1
1.1	Motivation and Problem Statement	1
1.1.1	Transfer Learning from the Visual Domain	3
1.1.2	Audio-specific Architecture Enhancement	4
1.1.3	Audio-visual Correspondence Learning	4
1.1.4	Research Focus and Objectives	5
1.1.5	Contributions	6
1.2	Thesis Structure	7
2	Background	9
2.1	Neural Networks for Audio-visual Processing	9
2.1.1	Building Blocks	10
2.1.2	Convolutional Neural Networks	22
2.1.3	Vision Transformer	23
2.1.4	Complex-valued Neural Networks	25
2.1.5	Knowledge Distillation	26
2.1.6	Siamese Networks	27
2.1.7	Contrastive Learning	28
2.1.8	Transfer Learning	30
2.1.9	Zero-shot Learning	32
2.2	Time-frequency Analysis of Audio	33
2.2.1	Fourier Analysis	34
2.2.2	Wavelet Analysis	37
3	Knowledge Framework and Conceptualization	41
3.1	Multimodal Learning with Audio Signals	41
3.1.1	Taxonomy of Audio-visual Learning Systems	44
3.1.2	Representation Learning	46
3.1.3	Correspondence Learning	48
3.1.4	Audio-visual Object Localization	52
3.1.5	Sound Source Separation	53

3.1.6	Audio-visual Generation	55
3.2	Optimization Objectives for Audio-visual Learning	57
3.2.1	Audio-visual Multilabel Classification	58
3.2.2	Environmental Sound Classification	59
3.2.3	Multimodal Retrieval	60
3.2.4	Zero-shot Sound Classification	61
3.2.5	Foundations for Experimental Exploration	62
4	Methods and Applications	65
4.1	Task Analysis	65
4.1.1	Audio-visual Multilabel Classification	66
4.1.2	Environmental Sound Classification	68
4.1.3	Multimodal Retrieval	70
4.1.4	Zero-shot Sound Classification	71
4.2	Experimental Design	73
4.2.1	Datasets	73
4.2.2	Transfer Learning from the Visual Domain	75
4.2.3	Audio-specific Architecture Enhancement	78
4.2.4	Audio-visual Correspondence Learning	82
5	Results and Discussion	85
5.1	Audio-visual Multilabel Classification	85
5.2	Environmental Sound Classification	87
5.2.1	Performance Evaluation	87
5.2.2	Robustness Evaluation	91
5.3	Multimodal Retrieval	94
5.4	Zero-shot Sound Classification	97
5.5	Experimental Findings	98
6	Conclusion	99
6.1	Transfer Learning from the Visual Domain	99
6.2	Audio-specific Architecture Enhancement	100
6.3	Audio-visual Correspondence Learning	101
6.4	Research Outcome	102
6.5	Outlook	102
	Bibliography	105

Introduction

“ *A complex system that works is invariably found to have evolved from a simple system that worked. A complex system designed from scratch never works and cannot be patched up to make it work. You have to start over with a working simple system.*

— **John Gall**

(Author, scholar, and pediatrician)

1.1 Motivation and Problem Statement

Research on machine learning models for audio and multimodal signals is rapidly evolving, spurred by the need to extend the successes achieved in image processing to challenges presented in the audio domain. This evolution is a natural response to the increasing demand for processing large and complex audio data in a variety of applications, posing intriguing questions about how established methodologies in other domains can be adapted and innovated to unlock the full potential of audio and multimodal signal processing [205]. These models open the possibility for solving exciting yet challenging tasks such as audio classification, sound event detection, speech recognition, and multimodal retrieval. With the advent of deep learning, researchers have introduced a plethora of architectures and enhancements, contributing to a new data-driven learning paradigm that has accelerated progress in machine learning, surpassing previous state-of-the-art models. Moreover, the availability of large annotated datasets—such as AudioSet [65]—has facilitated the development of these models. Therefore, the growing interest in multimodal data analysis is driving advances in various domains such as entertainment [205], health care, and security [13]. Thanks to deep learning, models made for audio and multimodal signals are starting to revolutionize the way humans communicate with machines, allowing for using voice and gestures to interact with computers and other smart devices [240].

These developments have opened up new avenues for exploring the rich and complex nature of audio signals, along with their interplay with other modalities such as images and text [124, 32, 141, 220]. The research community has made significant strides in leveraging deep neural networks to extract high-level representations from audio data [102, 58, 155], allowing for improved performance in various tasks. However, despite the remarkable progress made in audio and multimodal signal processing, multimodal learning with audio signals still presents many challenges.

One of such lies in effectively modeling and understanding relationships between audible, visual, and textual information. While considerable efforts have been devoted to developing models that exploit the complementary nature of these modalities, there is still a need for more sophisticated techniques that can seamlessly integrate and leverage the multimodal cues present in complex real-world scenarios [251]. Furthermore, the availability of labeled data remains a severe bottleneck. Although in safety-critical applications data acquisition has received deserved attention—and investments [174], data availability has been a sore spot for less sensitive use cases related to general-purpose multimodal signal processing, as they have suffered from a lack of large-scale annotated datasets [65].

Additionally, aspects related to the robustness of audio-centric models have rarely been addressed [205]. While impressive results have been achieved on benchmark datasets, these models often struggle to perform well when faced with real-world challenges such as varying acoustic conditions, background noise [214], or temporal or spatial misalignment [161]. The lack of robustness in these models is indicative of the need for effective representations of raw signals, regardless of modality. In turn, development of techniques that enhance the robustness of audio analysis models is, therefore, an essential area of research on bridging the gap between idealized benchmark scenarios and the complexities of real-world audio data.

Accounting for the existing challenges in multimodal learning with audio, this thesis addresses these limitations to develop improved deep learning methods. Here, the goal is to effectively model and leverage relationships between audio, visual, and textual information. Also, it seeks to overcome the bottleneck of limited labeled data and tackle the issue of robustness in audio-centric models, with a focus on improving their performance in difficult conditions.

Thus, the main research question of this thesis concerns the improvement of deep learning models by leveraging audio-specific and audio-agnostic methods as well as any combination thereof, and can be condensed as:

$\mathcal{Q}$: “Can the accuracy and robustness of current multimodal models be improved by combining methods that are unique to the audio domain with existing audio-agnostic techniques?”

This thesis decomposes the problem statement into three related subquestions formulated later in this chapter. To support answering them, the current state of the field of multimodal deep learning has to be evaluated. An overview of the deep learning landscape allows for identifying the most representative challenges. From low- to high-level semantics, these challenges are reflected by representation learning, correspondence learning, audio-visual object localization, sound source separation, and audio-visual generation. A thorough review of the learning objectives listed above allows for selecting specific tasks that bridge the gap between the theoretical basics and experimental results to answer the aforementioned subquestions.

Next, by considering audio-centric and general-purpose architecture modifications, their influence on the performance of deep learning models is evaluated. This thesis studies the incorporation of classical time-frequency analysis methods [10] to improve the model’s capabilities in processing audio. In the meantime, it investigates ways to enrich generalization abilities of deep learning models by jointly using and sharing abstract and low-level semantics across modalities based on transfer learning and knowledge distillation techniques.

Based on the theoretical basis and specific needs corresponding to audio-centric multimodal tasks, the evaluation of the experimental results allows for answering the following subquestions, which, in turn, build a basis for articulating an answer to the main question.

1.1.1 Transfer Learning from the Visual Domain

$\mathcal{Q}^{(1)}$: “How does leveraging models from the visual domain contribute to performance in audio-specific tasks?”

The first subquestion focuses on the benefits of using visual-domain models for solving audio-specific tasks such as audio-visual multilabel classification and environmental sound classification. By combining visual and audible information, the models learn better audio representations, leading to more accurate predictions. Moreover, transfer learning setup allows for leveraging pre-trained models and reducing the need for large amounts of labeled data, making it a suitable approach for audio-centric tasks [15, 125, 189, 68].

This thesis is addressing the subquestion $\mathcal{Q}^{(1)}$ by evaluating effects of incorporating visual-domain models in audio analysis task in a series of computational experiments. For the general-purpose and environmental sound classification tasks, the model’s ability to leverage low-level visual concepts transferred from the visual domain is being assessed, and the obtained results are summarized to provide a basis for answering $\mathcal{Q}^{(1)}$.

1.1.2 Audio-specific Architecture Enhancement

$\mathcal{Q}^{(2)}$: “How can one leverage audio-specific modules within existing deep learning backbones to obtain more accurate and robust models?”

The second subquestion considers the effects of using audio-specific architecture enhancements. Similarly to the subquestion $\mathcal{Q}^{(1)}$, it is also connected to the sound classification task, specifically, audio-visual multilabel classification and environmental sound classification. The subquestion $\mathcal{Q}^{(2)}$ concerns neural network layers that are specifically tailored for audio analysis and designed to capture and leverage special characteristics of audio signals represented by diverse temporal and spectral properties. From a more abstract—the whole model architecture—standpoint, a typical audio processing workflow differs from those for the visual input, as the audible modality assumes consideration of unique signal properties, different from those for images such as dimensionality or number of parallel channels. By incorporating audio-specific architectural modifications that include learnable time-frequency transformation and multichannel processing, models can extract more informative audio features, which, in turn, leads to increased classification accuracy.

In order to address the subquestion $\mathcal{Q}^{(2)}$, the experimental design is being extended to the assessment of model robustness against additive white Gaussian noise and reduction of sample rate, which provides additional insight into model performance orthogonal to conventional accuracy. By examining the impact of these enhancements, this thesis deepens the understanding of neural-network-based audio analysis in the context of multimodal learning.

1.1.3 Audio-visual Correspondence Learning

$\mathcal{Q}^{(3)}$: “Are there better training strategies beyond transfer learning for multimodal models?”

The third subquestion focuses on the benefits of going from the transfer-learning-based setup to the audio-visual correspondence learning that involves the use of a combined “student – teacher” model [91], which is also referred to as knowledge distillation. As the multimodal correspondence learning setup helps transfer learned semantics from one model to another, such combination of audible, visual, and textual modalities in a knowledge distillation setup contributes to the model performance [32, 61, 157]. Specifically, the models can learn from each other and benefit from the strengths of each modality, leading to more accurate results in sound classification and multimodal retrieval tasks.

This thesis is generalizing the experimental workflow to the explicit use of the visual modality, allowing for revealing findings associated with the knowledge-distillation-based training of audio-centric neural networks. By employing audio, visual, and textual modalities together, the thesis presents an extensive evaluation of model performance in a variety of multimodal tasks. Sound classification, including audio-visual multilabel classification, environmental sound classification, and zero-shot environmental sound classification, as well as multimodal retrieval help connect the subquestion $\mathcal{Q}^{(3)}$ to both the primary question $\mathcal{Q}$ and the solutions identified to address the subquestions $\mathcal{Q}^{(1)}$ and $\mathcal{Q}^{(2)}$.

1.1.4 Research Focus and Objectives

Guided by a thorough review of existing tasks and methods in Chapter 3, this thesis focuses on enhancing deep learning models for audio-dedicated tasks through the strategic integration of both audio-specific and audio-agnostic methods, along with their synergistic combinations. The motivation behind tackling these challenges lies in the imperative need for more accurate and efficient audio analysis methods, particularly for practical applications in diverse fields such as multimodal classification and retrieval.

The investigation comprises three key areas. First, this thesis explores the benefits of employing visual-domain methods to solve audio-domain tasks, highlighting their potential for improving performance. Second, the research delves into enhancing audio-specific architectures to further optimize the efficiency and robustness of sound classification models. Lastly, the thesis explores multimodal learning strategies based on knowledge distillation, aiming to leverage the power of combining audio, visual, and text information for improved results.

The sound classification tasks under consideration encompass a spectrum of scenarios that include audio-visual multilabel classification as well as regular and zero-shot environmental sound classification. By systematically addressing these challenges, the

research contributes to theoretical advancements in audio analysis as well as provides practical solutions for real-world applications that base on these methods. Thus, the ultimate goal of this thesis is to propel the development of more accurate and efficient audio analysis methods, paving the way for advancements in multimodal deep learning and in the related domains.

1.1.5 Contributions

The thesis presents the following contributions:

- Providing an overview of multimodal audio-related tasks in Chapter 3, this thesis selects a spectrum of high- and low-level learning objectives to build a knowledge framework helping to answer the subquestions $\mathcal{Q}^{(1)} - \mathcal{Q}^{(3)}$ by summarizing and discussing existing methods.
- Addressing the subquestion $\mathcal{Q}^{(1)}$, this thesis focuses on an isolated task of the environmental sound classification and proposes in Chapter 4 a transfer learning workflow that leverages powerful models from the visual domain, resulting in a successful ESResNet [79] model dedicated to solve audio-centered tasks. In addition, a technique adapting visual models to the time-frequency input is presented, namely spectrochromatic mapping.
- Considering audio-domain challenges such as Gaussian white noise and limited bandwidth, this thesis introduces in Chapter 4 an advanced time-frequency transformation layer (fbsp-layer) that improves both performance and non-adversarial robustness of the proposed ESResNeXt [78] model by employing trainable complex-valued signal processing, thereby addressing the subquestion $\mathcal{Q}^{(2)}$.
- Taking one step towards more general architectures in Chapter 4, the thesis proposes a multimodal framework allowing for a joint use of audible, textual, and visual input within a single model named AudioCLIP [77] to address the subquestion $\mathcal{Q}^{(3)}$. While improving the model performance in audio tasks, the introduced model design enriches AudioCLIP with extended capabilities of cross-modal retrieval and zero-shot classification.

Deep learning models and techniques proposed in this thesis were thoroughly evaluated and validated through extensive experiments on a broad range of datasets of various scales in Chapter 5.

1.2 Thesis Structure

Chapter 2

This chapter provides a theoretical foundation for audio-centric tasks employing neural network models and signal processing techniques. It covers the underlying architectures of neural networks and their training methods, as well as time-frequency analysis approaches, providing the essential background knowledge of the theoretical and practical basics for audio analysis in multimodal learning.

Chapter 3

This chapter presents an overview of the deep learning landscape in audio-visual learning, covering topics such as representation and correspondence learning, audio-visual object localization, sound source separation, and sound and video generation. It examines literature related to these topics and connects the subjects discussed to concrete tasks that include audio-visual classification, environmental sound classification, and multimodal retrieval.

Chapter 4

Based on the relationships between the branches of audio-visual visual multimodal learning and the specific tasks identified in Chapter 3, this chapter formulates a series of hypotheses to answer the research questions of this thesis. It also discusses the core techniques and practical applications employed in the study of audio analysis within the realm of multimodal learning. In particular, it proposes solutions to effectively leverage visual domain models and audio-specific architectural enhancements of deep neural networks for audio analysis, with a focus on the effects of using a transfer-learning- and knowledge-distillation-based multimodal setup.

Chapter 5

This chapter presents the results of the experimental assessment of cross-modal and audio-specific architecture enhancements for solving audio tasks. It gathers evidences to verify the hypotheses proposed in Chapter 4 by evaluating transfer-learning-based model training and audio-visual correspondence learning setting and discusses the gradual out-performance of the past contenders in the tasks such as audio-visual classification, environmental sound classification, cross-modal retrieval, and zero-shot environmental sound classification.

Chapter 6

The final chapter summarizes the main findings and key insights obtained throughout this thesis. It consolidates the most significant contributions and highlights novel aspects and advancements made in the field of audio analysis in multimodal learning. At the end, this chapter delves into outlining future research directions by identifying the current challenges and suggesting solutions to address them.

This chapter provides a theoretical foundation for understanding the neural network models and signal processing techniques employed in audio-centric tasks. Neural networks are a powerful tool for solving high-dimensional problems that are difficult to model mathematically [82]. In the context of audio processing, they have been successfully applied to tasks such as speech recognition [46], sound classification [88], and music generation [48]. The chapter introduces the basic architectural solutions for neural networks and their training techniques.

Another important aspect of audio processing using neural networks is time-frequency analysis. It is a set of methods that enable the transformation of time-domain signals into the frequency domain, which is often a more informative representation for analysis [198]. The chapter provides an overview of commonly used time-frequency analysis techniques such as short-time Fourier transform (STFT) and continuous wavelet transform (CWT), and their applications in various audio-centric tasks.

2.1 Neural Networks for Audio-visual Processing

The success of deep learning in audio signal processing is attributed to the ability of deep neural networks to learn complex, semantically rich representations of audio signals [102, 94, 145, 133]. This has led to significant progress in various audio-centric tasks, such as audio classification, audio segmentation, and speech recognition. This section provides a theoretical overview of essentials relevant for both general-purpose and audio-domain neural networks. It starts by introducing the basic building blocks of neural network architectures followed by optimization techniques and key operations, and then moves on to discuss convolutional and attention-based networks. By providing insights into particular architectural patterns and learning approaches, this section provides a comprehensive understanding of the neural network models used in audio processing tasks and their design choices and trade-offs.

2.1.1 Building Blocks

The field of deep learning is constantly evolving, with new advancements and techniques being developed to improve the performance of neural networks for various tasks. In order to understand these advances, it is important to have foundational knowledge of the basic building blocks of neural networks. Starting with backpropagation and optimization, this section discusses loss and activation functions as well as linear and convolutional layers, finalizing with the description of normalization layers.

Backpropagation

Backpropagation [222] is an algorithm that is commonly used to train neural networks in supervised learning tasks. It is a method for computing the multidimensional derivative—the gradient—of the loss function with respect to the parameters of the neural network, which is used to update the parameters during the optimization process. Figure 2.1 shows an example of a one-dimensional function $f(x)$ and its derivative $f'(x)$.

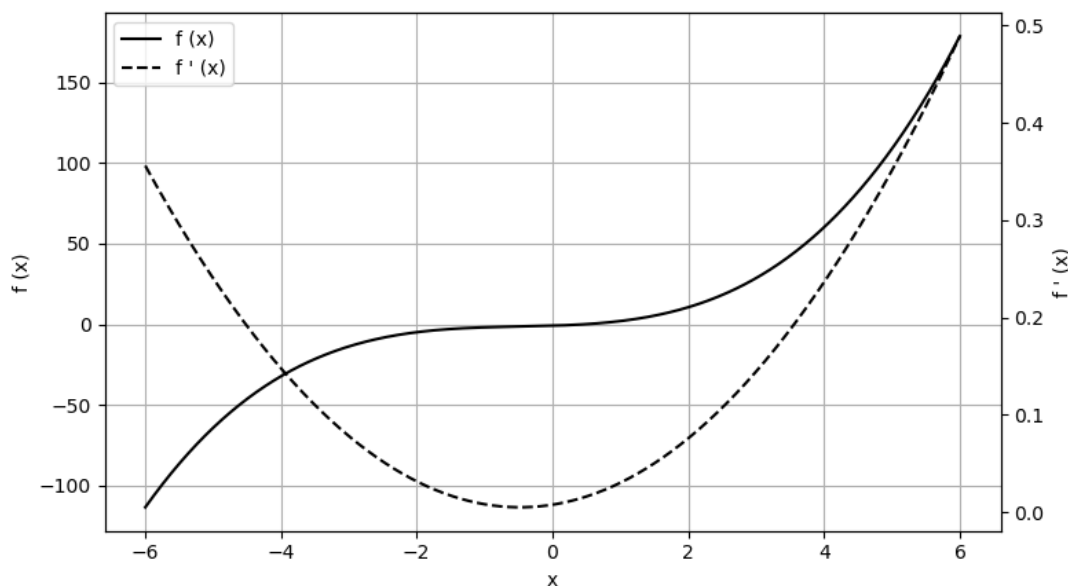


Fig. 2.1: Sample smooth function and its derivative.

The backpropagation algorithm works by propagating the error back through the network, from the output layer to the input layer. To begin, the network is first initialized with random weights and biases. Then, a forward pass is made through the network, where the input data is fed into the network, and the output is calculated by passing the input through a series of transformations, defined by the weights and biases of

the network. The difference between the predicted output and the true output is then calculated, which is referred to as the loss or cost function.

The backpropagation algorithm computes the gradient of the loss function with respect to the weights and biases of the network, by applying the chain rule of differentiation (see Equation 2.1). The gradient is then used to update the weights and biases of the network in the opposite direction of the gradient, with a step size determined by the learning rate hyperparameter. The pseudocode for the backpropagation is provided by Algorithm 1.

$$h'(x) = f'(g(x)) \cdot g'(x) \quad (2.1)$$

Algorithm 1 Backpropagation

Require: $\mathcal{D}$: training dataset

Require: α : learning rate

Require: θ : model parameters

while not converged **do**

 Randomly select a training example $(\mathbf{x}, y) \in \mathcal{D}$

 Feedforward pass: compute activations $\mathbf{a}$ for each layer

 Compute output error: $\delta^{(L)} \leftarrow \nabla_{\mathbf{a}^{(L)}} \mathcal{L}(\mathbf{a}^{(L)}, y)$

for $l = L - 1$ **to** 1 **do**

 Compute layer error $\delta^{(l)}$

 Compute weight gradients $\nabla_{\mathbf{W}^{(l)}} \mathcal{L}$

 Compute bias gradients $\nabla_{\mathbf{b}^{(l)}} \mathcal{L}$

 Update weights and biases according to the optimizer's rule

end for

end while

The backpropagation algorithm is typically performed using an optimization algorithm such as stochastic gradient descent or one of its variants, as further described in Section 2.1.1. The process of forward propagation followed by backpropagation is repeated multiple times until the network has learned to make accurate predictions on the training data.

The backpropagation algorithm, while a cornerstone in training neural networks, exhibits several limitations. One critical challenge is the issue of vanishing and exploding gradients, especially in deep networks, where gradients can become excessively small or large during backpropagation [85]. Additionally, the algorithm is susceptible to getting trapped in local minima during optimization, hindering its ability to find the global minimum [241]. Furthermore, the computational complexity of backpropagation increases with the size of the neural network and the amount of data, necessitating substantial computational resources [131]. The latter issue appears to become more

important in the recent studies, as the scale of deep learning models tends to increase, so the research community aims to find more efficient learning methods [89].

Optimization

The model optimization procedure is the process of finding the optimal set of parameters that minimize the loss function, which represents the difference between the predicted outputs of the model and the ground-truth labels. This process is typically done using an optimization algorithm, such as Stochastic Gradient Descent (SGD) [114] (Algorithm 2) or Adam [119] (Algorithm 3), which iteratively updates the model parameters given gradients of the loss function with respect to these parameters.

Algorithm 2 Stochastic Gradient Descent [114]

Require: α : learning rate
Require: $f(\theta)$: stochastic loss function
Require: θ : model parameters
 while not converged **do**
 $g_t \leftarrow \nabla_{\theta} f_t(\theta_{t-1})$ (compute gradients)
 $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot g_t$ (update parameters)
 end while
return θ_t

At each iteration, the algorithm computes the gradients of the loss function as a part of the backpropagation algorithm, which propagates the error from the output of the model back through the layers to compute the gradients of the parameters. These gradients are then used to update the model parameters in the direction that minimizes the loss function.

The optimization procedure typically involves a training set and a validation set. The model is trained on the training set, and the performance of the model is evaluated on the validation set. If the performance on the validation set starts to degrade, the training should be stopped [221], as the model is considered to be overfitting. The final model is selected based on its performance on the validation set. Once the model has been optimized, it can be used to make predictions on new data.

Although being commonly used in practice, the aforementioned model optimization algorithms exhibit certain limitations. SGD can suffer from slow convergence, particularly around saddle points, necessitating the use of learning rate schedules [138] or momentum [178] to enhance efficiency. On the other hand, Adam is known to be sensitive to hyperparameters and might exhibit suboptimal generalization [295]. To address these issues, practitioners often engage in careful hyperparameter tuning,

Algorithm 3 Adam [119]

Require: α : learning rate

Require: $\beta_1, \beta_2 \in [0, 1)$: exponential decay rates for the moment estimates

Require: $f(\theta)$: stochastic objective function

Require: θ_0 : model parameters

$m_0 \leftarrow 0$ (initialize first moment vector)

$v_0 \leftarrow 0$ (initialize second moment vector)

$t \leftarrow 0$ (initialize time step)

while not converged **do**

$t \leftarrow t + 1$

$g_t \leftarrow \nabla_{\theta} f_t(\theta_{t-1})$ (compute gradients)

$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (compute biased first moment estimate)

$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (compute biased second raw moment estimate)

$\hat{m}_t \leftarrow \frac{m_t}{1 - \beta_1^t}$ (compute bias-corrected first moment estimate)

$\hat{v}_t \leftarrow \frac{v_t}{1 - \beta_2^t}$ (compute bias-corrected second raw moment estimate)

$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \frac{\hat{m}_t}{\sqrt{\hat{v}_t + \epsilon}}$ (update parameters)

end while

return θ_t

exploring alternative optimizers like RMSprop [90] or NAdam [51], and incorporating regularization techniques such as weight decay [152]. However, it is important to note that the model is only as good as the data it has been trained on [265], and therefore, the quality and representativeness of the training data are critical for the success of the optimization procedure.

Loss Function

In deep learning, the loss function is a crucial component used to measure the difference between the predicted output of the model and the actual ground-truth labels. The goal of training a deep learning model is to minimize the loss function over the entire training dataset.

One commonly used loss function in binary or multilabel classification problems is binary cross-entropy, also known as log loss. This loss function measures the difference between the predicted probability distribution and the true probability distribution of the labels, as shown in Equation 2.2.

$$\mathcal{L}_{BCE} = -(y \cdot \log p + (1 - y) \cdot \log (1 - p)) \quad (2.2)$$

where y is the ground-truth label (either 0 or 1) and p is the predicted probability that the input belongs to class 1.

In multi-class classification problems, the cross-entropy loss is typically used. This loss function is an extension of binary cross-entropy to $N > 2$ classes and is defined by Equation 2.3.

$$\mathcal{L}_{CE} = - \sum_{i=1}^N y_i \cdot \log(p_i) \quad (2.3)$$

where y_i is a one-hot vector representing the ground-truth label, and p is the predicted probability for the i^{th} class.

The cross entropy loss can also be applied indirectly to enable the contrastive training setup. Specifically, the symmetric cross entropy loss aims to diagonalize an $N \times N$ similarity matrix for N pairs of representation embeddings by treating the pair index as the target class. Such learning objective found its use in building a contrastive text-image model CLIP [206] and its multimodal descendant AudioCLIP [77], as described later in Section 2.1.7.

Cross entropy tends to heavily penalize confident yet incorrect predictions, which can be undesirable in scenarios where a more nuanced approach to misclassifications is needed or the data labels are known to be noisy [294]. Additionally, it may struggle when dealing with imbalanced datasets, as Equation 2.3 prioritizes the majority class. Strategies like class weighting or the use of alternative loss functions, such as focal loss [139], have been employed to mitigate these limitations and improve the performance of models, especially in contexts where the data distribution is uneven.

Activation Function

Activation functions play a crucial role in deep learning models as they introduce non-linearity, enabling the model to learn complex relationships between inputs and outputs.

The sigmoid function (black line in Figure 2.2a) is a commonly used activation function in neural networks. It maps any real-valued number to a value between 0 and 1 (Equation 2.4), thus squeezing the input values into a fixed range, which can be of use for particular applications [164]. However, due to properties of its derivative (dotted black line in Figure 2.2b), sigmoid suffers from the vanishing gradient problem (see Equation 2.5), which can make it difficult to train deep neural networks.

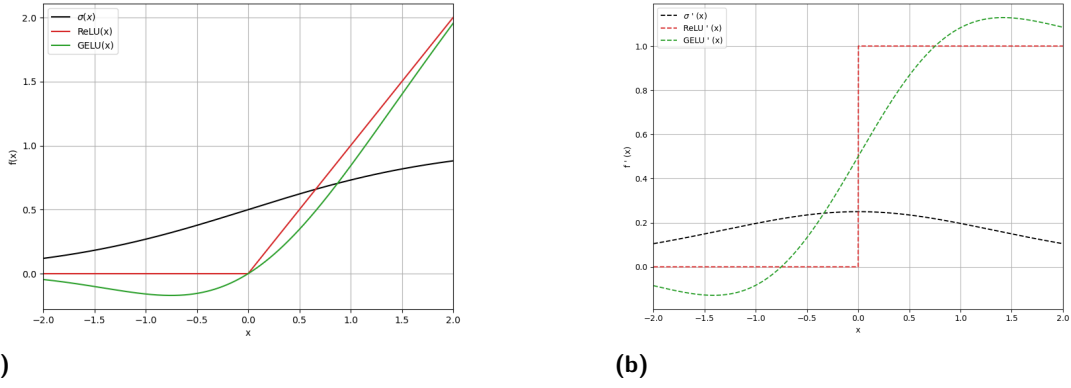


Fig. 2.2: Sample activation functions and the corresponding derivatives.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (2.4)$$

$$\sigma'(x) = \sigma(x) \cdot (1 - \sigma(x)) \quad (2.5)$$

The rectified linear unit (ReLU) [176] is another popular activation function. It is a simple and computationally efficient function (red line in Figure 2.2a) that sets all negative values to 0 and leaves positive values unchanged (Equation 2.6). ReLU has been shown to perform well in deep neural networks and helps alleviate the vanishing gradient problem [99]. However, it suffers from the dying ReLU problem, where some neurons can become inactive during training and never recover due to zero gradient for negative values (dotted red line in Figure 2.2b), as follows from Equation 2.7.

$$\text{ReLU}(x) = \max(0, x) \quad (2.6)$$

$$\text{ReLU}'(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \end{cases} \quad (2.7)$$

The Gaussian error linear unit (GELU) [87] is a relatively new activation function that has been shown to perform well in deep neural networks. It is a smooth approximation of the ReLU function and introduces non-monotonicity (green line in Figure 2.2a), which is beneficial for certain applications [267]. Although the exact definition is done by Equation 2.8, an approximation by Equation 2.10 is used in practice [87].

$$GELU(x) = \frac{x}{2} \cdot \left[1 + \operatorname{erf} \frac{x}{\sqrt{2}} \right] \quad (2.8)$$

where

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp^{-t^2} dt \quad (2.9)$$

$$GELU_{\approx}(x) = \frac{x}{2} \cdot \left[1 + \tanh \left(\sqrt{\frac{2}{\pi}} (x + 0.044715 x^3) \right) \right] \quad (2.10)$$

Equation 2.11 and Figure 2.2b (dotted green line) show the corresponding derivative of $GELU_{\approx}$ [181]. Gaussian error linear unit has been shown to be a good replacement for ReLU [87] and is widely used as a part of the Transformer architecture (see Section 2.1.3).

$$\begin{aligned} GELU_{\approx}'(x) = & \frac{1}{2} \cdot \tanh(0.0356774 x^3 + 0.797885 x) \\ & + 0.5 + (0.0535161 x^3 + 0.398942 x) \\ & \cdot \cosh^{-2}(0.0356774 x^3 + 0.797885 x) \end{aligned} \quad (2.11)$$

Linear Layer

A linear layer, also known as a fully-connected layer, is a fundamental component that performs a linear transformation on the input data. It is characterized by a set of learnable parameters (weights and biases), which are applied to the input features. At a high level, the linear layer can be viewed as a mapping from the input space of dimensionality N to the output space of dimensionality M .

The linear layer operates by performing a matrix multiplication between the input data $X \in \mathbb{R}^{B \times N}$ and the weight matrix $W \in \mathbb{R}^{M \times N}$, followed by an element-wise addition of the biases $b \in \mathbb{R}^M$, thus producing the output $Y \in \mathbb{R}^{B \times M}$, as shown in Equation 2.12. This computation allows the layer to learn complex relationships between the input features and the desired output. Each output of the layer represents a neuron that computes a weighted sum of the inputs and applies an activation function to produce the resulting linear transformation.

$$Y = W \cdot X + b \quad (2.12)$$

During training, parameters of the linear layer are being optimized to minimize the loss function L using an optimizer of choice, for which weight (Equation 2.13) and bias (Equation 2.14) gradients need to be computed.

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial Y} \cdot X^T \quad (2.13)$$

$$\frac{\partial L}{\partial b} = \sum \frac{\partial L}{\partial Y} \quad (2.14)$$

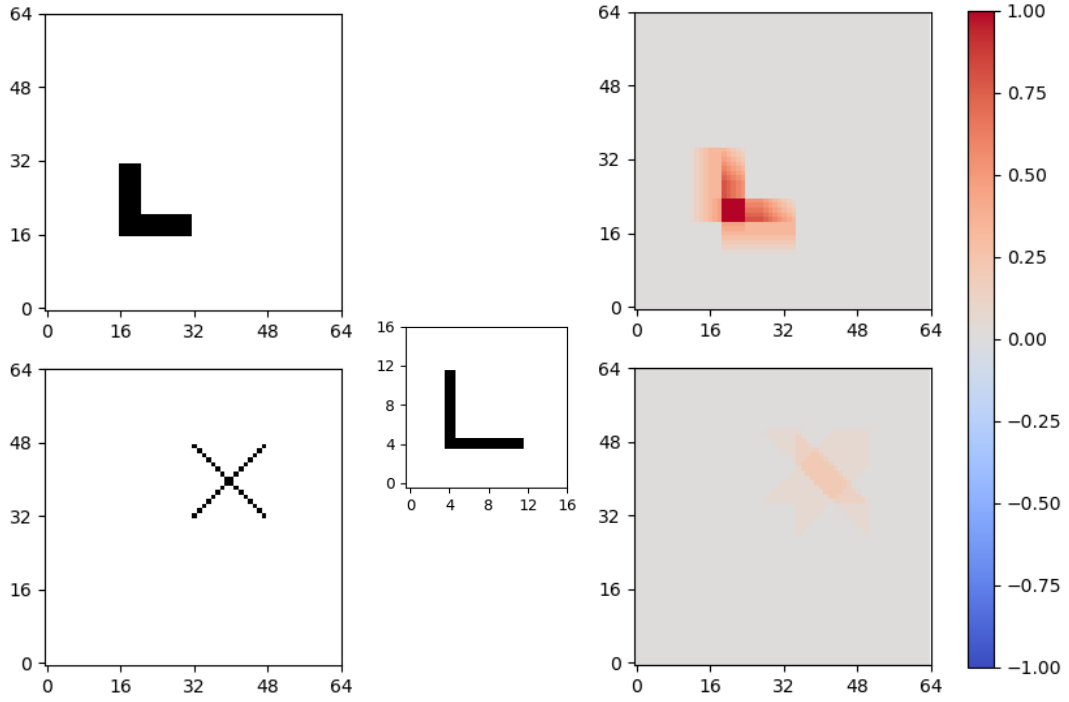
The linear layer plays an essential role in neural networks as it provides the capacity to model non-linear relationships by learning appropriate weight values. It acts as a transformation function that maps the input data to a higher-dimensional space, enabling the network to learn more complex representations [70]. By stacking multiple linear layers together with non-linear activation functions in between, deep neural networks can learn increasingly complex and abstract representations of the input data [53].

While theory says that a multilayer neural network can effectively approximate any smooth Euclidean function [93], its computational complexity rapidly grows with the size of the model [187]. To overcome this limitation, one can reuse layer parameters in a shared manner by applying convolutional layers described in Section 2.1.1.

Convolutional Layer

A convolutional layer [130] is a fundamental building block used primarily in convolutional neural networks for processing grid-like input data such as images or spectrograms in audio analysis. It relies on generalized dot product of tensors of rank 2 and higher and combines it with the weight sharing approach [130]. The convolutional layer operates by performing local spatial operations, applying filters or kernels to extract features from the input data.

At a high level, the convolutional layer convolves the input data with a set of learnable filters (Equation 2.15), which are small-sized matrices or tensors. These filters slide across the input in a systematic way, computing a dot product between the filter weights and a local region of the input at each position. The dot product operation captures the similarity between the filter and the corresponding input region, producing an activation map or feature map, as demonstrates Figure 2.3.



(a) Sample binary images. (b) Sample kernel. (c) Sample feature maps.

Fig. 2.3: Example of a two-dimensional convolution.

$$Y = W * X + b \quad (2.15)$$

A parameter optimization procedure of a two-dimensional convolutional layer uses Equation 2.16 and Equation 2.17 for computing gradients:

$$\frac{\partial L}{\partial W} = \frac{\partial L}{\partial Y} * X \quad (2.16)$$

$$\frac{\partial L}{\partial b} = \sum \frac{\partial L}{\partial Y} \quad (2.17)$$

Convolutional layers, while highly effective in capturing local features through shared filters, come with inherent limitations. One notable drawback is the lack of rotational and scale invariance, meaning that these layers might struggle with variations in orientation and size of objects in an image. Since the convolutional kernels are fixed in size and orientation, they may not robustly handle transformations like rotation [33] or changes in scale [278].

Another limitation is the difficulty in capturing global patterns. Convolutional layers are designed to focus on local information, making it challenging for them to discern overarching structures or relationships between distant elements in an input [231]. While strides and pooling layers can partially address these issues, achieving full rotational and scale invariance, as well as capturing global context, often requires more advanced architectures or additional techniques such as attention mechanisms discussed in Section 2.1.1.

Normalization Layer

Batch normalization [100] is a technique commonly used in deep learning models to improve training stability and accelerate convergence. At a high level, the batch normalization layer normalizes the activations of a neural network layer by transforming them to have zero mean and unit variance, as demonstrated by Figure 2.4.

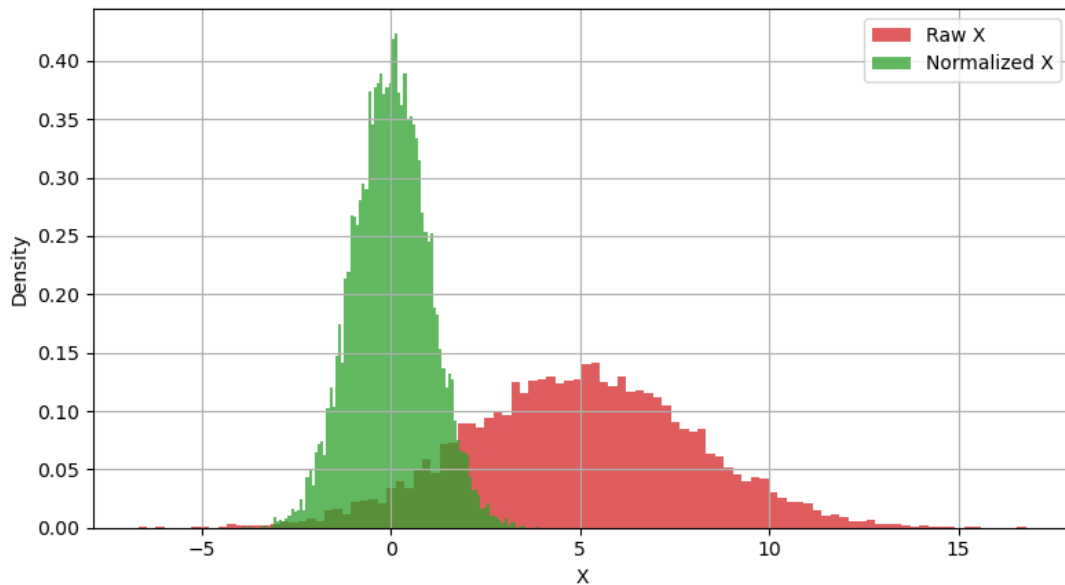


Fig. 2.4: Example of the distribution normalization.

Layer normalization [18] is a technique used in deep learning models to normalize the activations within a layer independently. Being an alternative to batch normalization, layer normalization operates on a single example or sequence independently, rather than on a mini-batch, evading dependency on the batch size, and is particularly relevant in architectures like the Transformer [18]. This makes it suitable for scenarios where batch sizes are small [285] or when dealing with variable-length sequences [39].

The normalization is described by Equation 2.18

$$Y = \gamma \odot \hat{X} + \beta \quad (2.18)$$

where

$$\hat{X} = \frac{X - \mu_X}{\sqrt{\sigma_X^2 + \epsilon}} \quad (2.19)$$

The corresponding gradient computations for the backward propagation are defined as follows:

$$\frac{\partial L}{\partial \gamma} = \frac{\partial L}{\partial Y} \odot \hat{X} \quad (2.20)$$

$$\frac{\partial L}{\partial \beta} = \sum \frac{\partial L}{\partial Y} \quad (2.21)$$

Results of the layer normalization may exhibit sensitivity to the choice of hyperparameters such as bias and gain [277] as well as to the variability of the input signal, thus impacting the effectiveness. To address the latter issue, adaptive layer normalization techniques are being developed [118].

Attention

The attention mechanism is a fundamental component in deep learning models that enables the model to focus on different parts of the input sequence when making predictions or performing computations. Figure 2.5 demonstrates sample self- (a) and “encoder – decoder” (b) attention matrices. Attention has been widely used in natural language processing tasks and is a key element in the Transformer architecture [253].

At a high level, the attention mechanism projects the input $X \in \mathbb{R}^{l_x \times d_x}$ sequence into Query $Q \in \mathbb{R}^{l_x \times d_{attn}}$, Key $K \in \mathbb{R}^{l_x \times d_{attn}}$, and Value $V \in \mathbb{R}^{l_x \times d_h}$ matrices, and assigns a corresponding relevance score $S_i \in \mathbb{R}^{l_x \times l_x}$ to each Value element, as presented in the pseudocode for a single attention head (Algorithm 4). The multi-head attention combines individual attention outputs $h_i \in \mathbb{R}^{l_x \times d_h}$ into a final attention vector $A \in \mathbb{R}^{l_x \times d_{out}}$ by using an output projection matrix $W^O \in \mathbb{R}^{(h \cdot d_h) \times d_{out}}$, as described by Equation 2.22.

In practice, there are two common cases how attention is used: self- and “encoder – decoder” attention. In the self-attention mechanism, the input sequence is divided into queries, keys, and values, which are transformed using linear projections. Then,

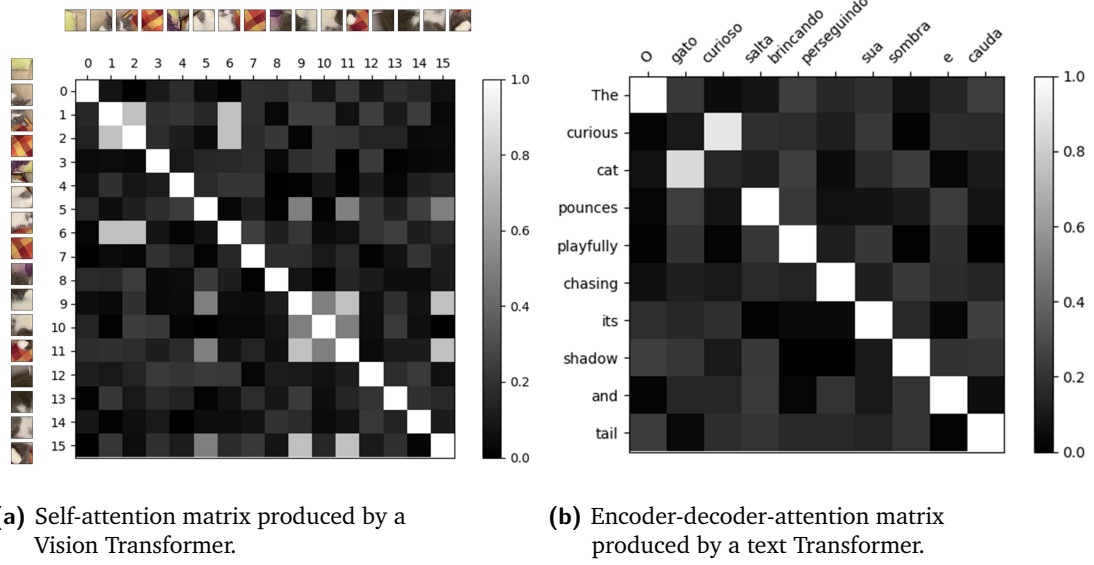


Fig. 2.5: Example of attention matrices.

attention scores are computed and the weighted sum of the values is obtained, thus providing a self-attended representation for each element in the sequence. For the “encoder – decoder” attention, the queries are derived from the previous decoder layer, while the keys and values are from the encoder output. This mechanism allows the decoder to focus on different parts of the input sequence during the generation of each output.

Algorithm 4 Single attention head [253]

Require: $X \in \mathbb{R}^{l_x \times d_x}$: input sequence

Require: $W_i^Q \in \mathbb{R}^{d_x \times d_{attn}}$, $W_i^K \in \mathbb{R}^{d_x \times d_{attn}}$, $W_i^V \in \mathbb{R}^{d_x \times d_h}$: input projections

$Q_i \leftarrow XW_i^Q$ (compute Query)

$K_i \leftarrow XW_i^K$ (compute Key)

$V_i \leftarrow XW_i^V$ (compute Value)

$S_i \leftarrow \frac{Q_i K_i^T}{\sqrt{d}}$ (compute relevance score)

$S_i \leftarrow \text{softmax}(S_i)$ (normalize the relevance score to sum up to 1)

$h_i \leftarrow S_i V_i$ (compute the weighted attention)

return h_i

While attention is highly effective in capturing contextual information, it has limitations, the important one being its quadratic computational complexity concerning sequence length [253], making it resource-ineffective for long sequences. To address this issue sub-quadratic reformulations of attention have been proposed to improve efficiency and enhance the model’s ability to handle long-range dependencies in sequences [259, 159].

$$A(X) = \begin{bmatrix} h_1(X) \\ \vdots \\ h_h(X) \end{bmatrix} W^O \quad (2.22)$$

2.1.2 Convolutional Neural Networks

Convolutional Neural Networks (CNN) [130] are a type of deep learning model that naturally exploit an explicit inductive bias for processing grid-like data [168]. They have been widely used in various computer vision and audio analysis tasks due to their ability to effectively identify spatial patterns and capture dependencies within the input data.

At a high level, CNNs consist of multiple layers that perform convolutions, non-linear activations, and pooling operations [66], as illustrated by Figure 2.6. The key idea behind CNNs is to learn local feature representations that are progressively combined to form more abstract and high-level representations.

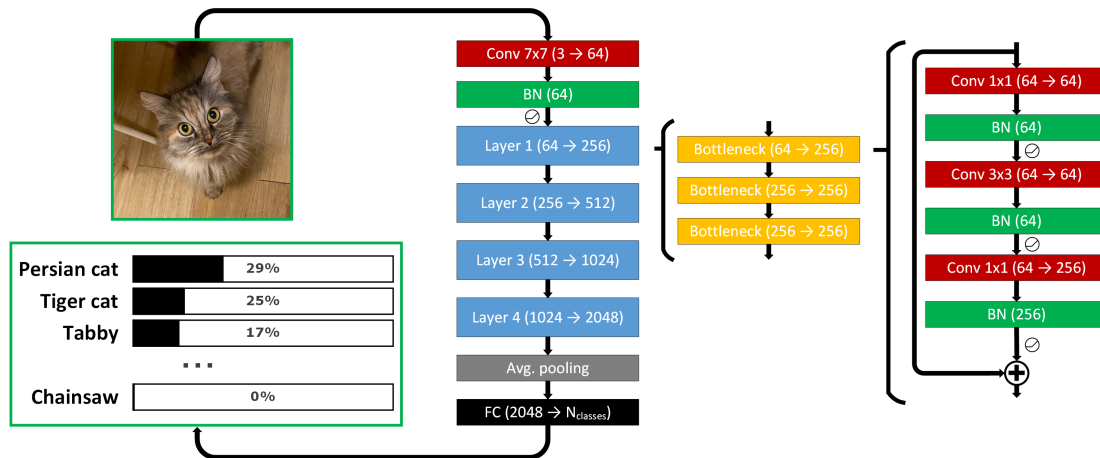


Fig. 2.6: Architecture of a residual CNN (ResNet-50).

ResNet (Residual Network) [85] is a popular CNN architecture that addresses the vanishing gradient problem encountered in deep networks. It introduces residual—or skip—connections that allow the network for learning residual mappings, capturing the difference between the input and the desired output. This enables the network to effectively propagate gradients through deep layers and improves the training and optimization process [85]. Depending on the specific implementation, ResNet has a number of residual blocks, each containing multiple convolutional layers, batch normalization, and non-linear activation functions, making them capable of learning complex representations.

ResNeXt (Residual Next) [274] is another CNN architecture that extends the idea of residual connections by introducing a new dimension called the “cardinality.” This dimension allows for parallel paths of information flow within each residual block, enabling the model to capture diverse and complementary features. By increasing the cardinality, ResNeXt models can effectively balance model complexity and computational cost, leading to improved performance [274].

Both ResNet and ResNeXt architectures have demonstrated state-of-the-art performance in various image and audio analysis tasks [79, 78]. In the context of audio analysis in multimodal learning, these architectures can be adapted to process audio spectrograms or other audio representations. By effectively distinguishing audio features and capturing temporal dependencies, they leverage the advantages of skip connections to identify complex audio patterns, enhancing performance of audio-centric deep learning models.

Being highly successful in various computer vision tasks [242, 213, 85], CNNs inherit limitations of their basic blocks, namely convolutional layers. The predefined architecture limits the model’s ability to learn abstract features, and it may amplify struggling with geometric transformations [213]. To address these issues, one can consider using more advanced—often attention-based—architectures such as Vision Transformer (Section 2.1.3) or its descendants.

2.1.3 Vision Transformer

The Vision Transformer (ViT) [50] architecture is a recent development that adapts the Transformer model to image-based tasks. It has shown impressive performance in computer vision tasks, challenging the dominance of convolutional neural networks.

The regular Transformer architecture was introduced as a sequence-to-sequence model for NLP tasks [253]. It consists of two main components: the encoder and the decoder. The encoder processes the input sequence by applying self-attention mechanisms to capture global dependencies and then feeding the resulting representations through feed-forward neural networks. The decoder generates the output sequence by attending to the encoder’s output and using self-attention and cross-attention mechanisms.

Vision Transformer tailors the Transformer architecture to handle images. Rather than treating an image as a sequence of pixels, ViT divides the image into patches and flattens them into a sequence of vectors. Figure 2.7 summarizes a typical ViT-workflow. The obtained patch embeddings are then processed by the Transformer encoder, which enables the model to capture long-range dependencies in the image. However, the

standard Transformer struggles to capture local spatial information due to the absence of convolutional operations.

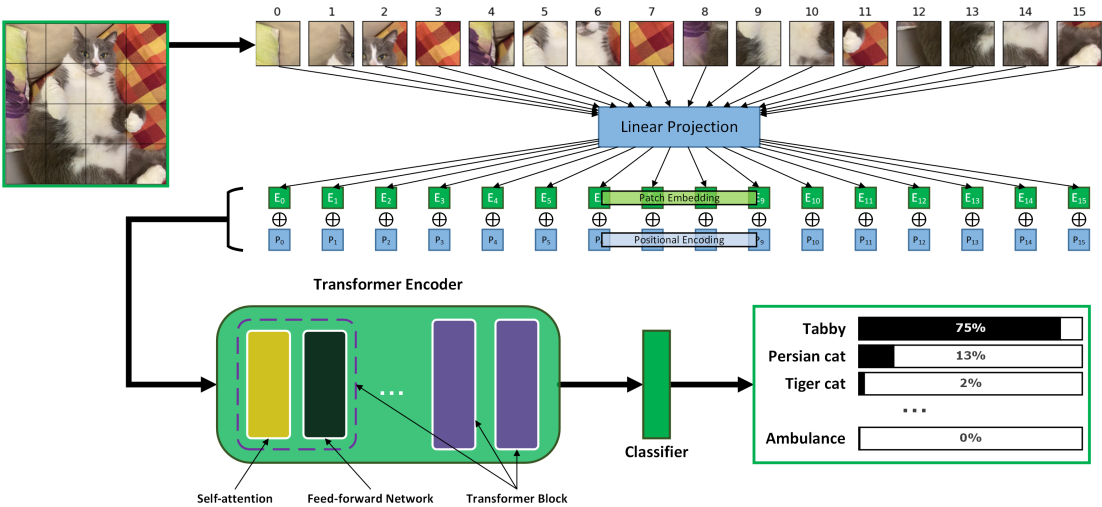


Fig. 2.7: Sample workflow of a Vision Transformer.

To address this limitation, ViT introduces a positional embedding to represent the spatial information of the patches and adds a learnable classification token to the sequence. During training, a pretraining phase is often employed using a large dataset, such as ImageNet [45], where the model learns to predict the correct class labels. This pretraining is typically followed by fine-tuning on specific downstream tasks.

The Vision Transformer has demonstrated competitive performance on various image classification benchmarks [50], showcasing its ability to learn meaningful representations from image patches. It offers advantages such as enormous model capacity [112], inherent sequence processing, and a great potential for generalization to other modalities [68, 7].

To known limitations of ViT-based models one can refer to the quadratic computation complexity concerning sequence length, which is inherited from the attention mechanism described in Section 2.1.1. At the same time, the basic design of ViT lacks inductive biases for spatial hierarchies that convolutional layers naturally exploit, potentially making it less effective in capturing local spatial patterns [50]. To overcome this limitation, hybrid architectures enhance the vanilla ViT so that it enables to mimic behavior of convolutional layers [42] or perform a multiscale processing of the input [279].

In the context of audio analysis in multimodal learning, the Vision Transformer can be extended to process audio spectrograms or other audio representations, applying similar principles as in image processing [68]. By leveraging the attention mechanisms, the Vision Transformer can capture global dependencies and relationships in audio

signals, enabling it to extract high-level audio features and improve the performance of audio-centric deep learning models.

2.1.4 Complex-valued Neural Networks

Complex-valued neural networks (CVNNs) are neural network models that operate on complex numbers instead of real numbers. These networks have gained attention in various domains, including audio analysis [134], where complex-valued representations are commonly used.

In complex-valued neural networks, the weights, biases, and activations are complex numbers. This allows the network to capture and process both magnitude and phase information, which is important in many signal processing tasks [136].

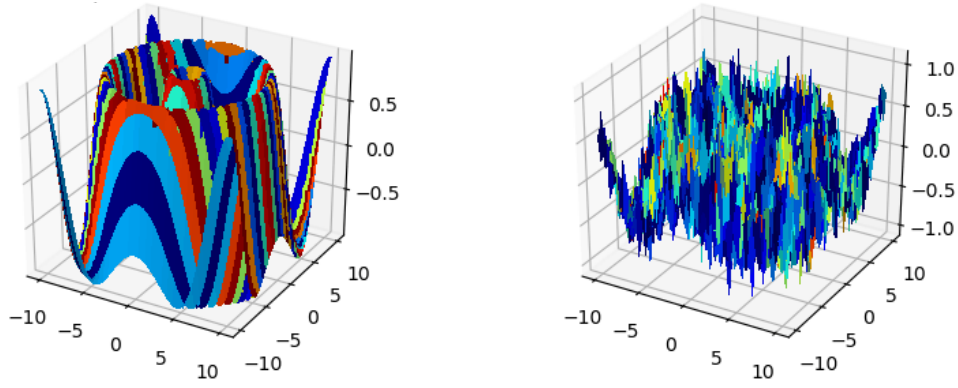
To train complex-valued neural networks, the backpropagation algorithm needs to be adapted. One approach is to use Wirtinger derivatives, which extend the concept of gradients to complex-valued functions $f : \mathbb{C} \rightarrow \mathbb{C}$. The Wirtinger derivative treats the complex variable $z \in \mathbb{C}$ and its conjugate $\bar{z} \in \mathbb{C}$ as independent variables and calculates the derivatives with respect to both of them, as shown in Equation 2.23a and Equation 2.23b. This allows the gradients for capturing the information from both the real and imaginary parts of the complex-valued activations.

$$\frac{\partial}{\partial z} = \frac{1}{2} \cdot \left(\frac{\partial}{\partial \Re(z)} - i \cdot \frac{\partial}{\partial \Im(z)} \right) \quad (2.23a)$$

$$\frac{\partial}{\partial \bar{z}} = \frac{1}{2} \cdot \left(\frac{\partial}{\partial \Re(z)} + i \cdot \frac{\partial}{\partial \Im(z)} \right) \quad (2.23b)$$

By using complex-valued neural networks, one can benefit from applying enhanced representation power and modeling capabilities provided by corresponding layers. These networks can natively extract phase-conditioned relationships and dynamics present in the data, thus allowing for improved generalization [74] and faster convergence (e.g., Figure 2.8).

While CVNNs have the potential to represent and process both amplitude and phase information, the increased computational demands arise from the need to handle complex-valued weights and activations, leading to higher memory requirements ($\times 2$) and computational overhead ($\times 4$) [21]. However, specialized hardware, tailored to efficiently handle complex number computations, can significantly enhance the performance of complex-valued networks [290].



(a) Convergence of a $\mathbb{C}$ -valued neural network. Mean squared error is 9.128×10^{-10} .
 (b) Convergence of an $\mathbb{R}$ -valued neural network. Mean squared error is 1.411×10^{-1} .

Fig. 2.8: Example of fitting to a $\mathbb{C}$ -valued function for 300 epochs.

This thesis presents a trainable time-frequency transformation—an fbsp-layer—that is cognate to linear filtering techniques, allowing for effective noise reduction and enhancement of desired signal components, and provide a more complete and accurate description of the audio signal in the time-frequency domain, as described in Section 4.2.3 and Section 5.2.2.

2.1.5 Knowledge Distillation

Knowledge distillation [91] is a technique used in machine learning to transfer the knowledge learned by a large, complex model (known as the teacher model) to a smaller, more efficient model (known as the student model). The goal of knowledge distillation is to improve the performance of the student model by leveraging the rich information captured by the teacher model.

The process of knowledge distillation involves training the student model to mimic the outputs of the teacher model, as shown in Figure 2.9. Instead of directly learning from the training data, the student model learns from the softened probabilities produced by the teacher model [91]. This softening is typically achieved by applying a temperature parameter to the logits (pre-softmax outputs) of the teacher model, resulting in a smoothed probability distribution. The student model is then trained to minimize the difference between its own predictions and the softened targets provided by the teacher model.

Knowledge distillation offers several benefits in the context of audio analysis in multi-modal learning. First, it enables the transfer of knowledge from a larger, more accurate

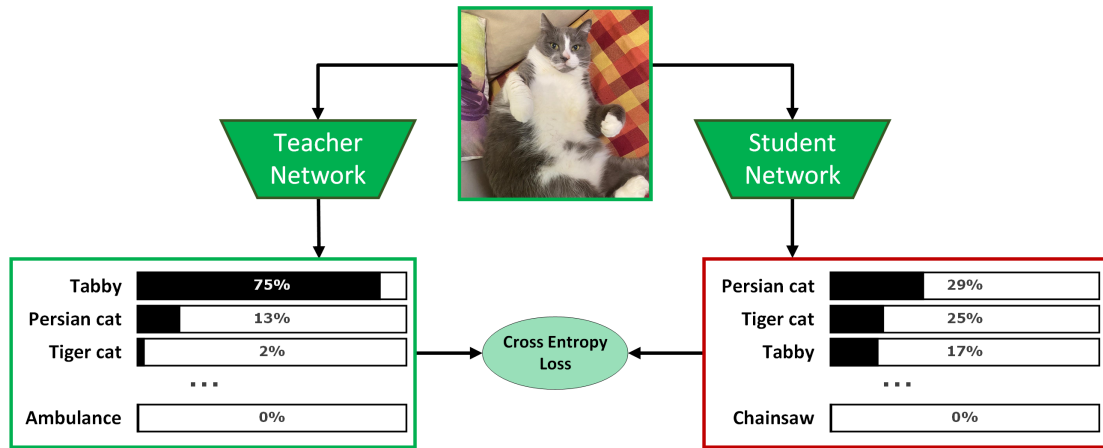


Fig. 2.9: Knowledge distillation workflow.

model to a smaller model, allowing for more efficient deployment and reduced computational resources [258]. This is particularly advantageous in scenarios where real-time or resource-constrained audio processing is required. Second, knowledge distillation can improve the generalization performance of the student model [258]. By learning from the soft targets provided by the teacher model, the student model can learn to generalize beyond the training data and capture the underlying patterns and relationships in the audio data more effectively [61]. Finally, knowledge distillation can also facilitate the incorporation of external knowledge sources [287]. The teacher model can be pre-trained on a large dataset or trained with additional domain-specific knowledge, and this knowledge can be transferred to the student model through distillation. This allows the student model for benefiting from the collective knowledge acquired by the teacher model, as shown in Chapter 5.

At the same time, knowledge distillation imposes challenges related to its relative novelty. Specifically, relationships between the rate of knowledge transfer and the architectures of the teacher and student models are not yet clear, as well as modeling the knowledge acquired from different layers of the teacher model—whether it involves logits or deep features—can be non-trivial due to the intricate interplay of information at various levels [71]. These limitations necessitate careful consideration and experimentation to effectively leverage knowledge distillation across diverse model architectures [92, 149] and achieve optimal knowledge transfer [73].

2.1.6 Siamese Networks

The Siamese architecture [120] is a neural network design that involves sharing weights and parameters between two or more identical subnetworks, commonly referred to as

Siamese twins. This architecture is particularly useful for tasks that involve comparing or measuring similarity between pairs of inputs, as shown in Figure 2.10.

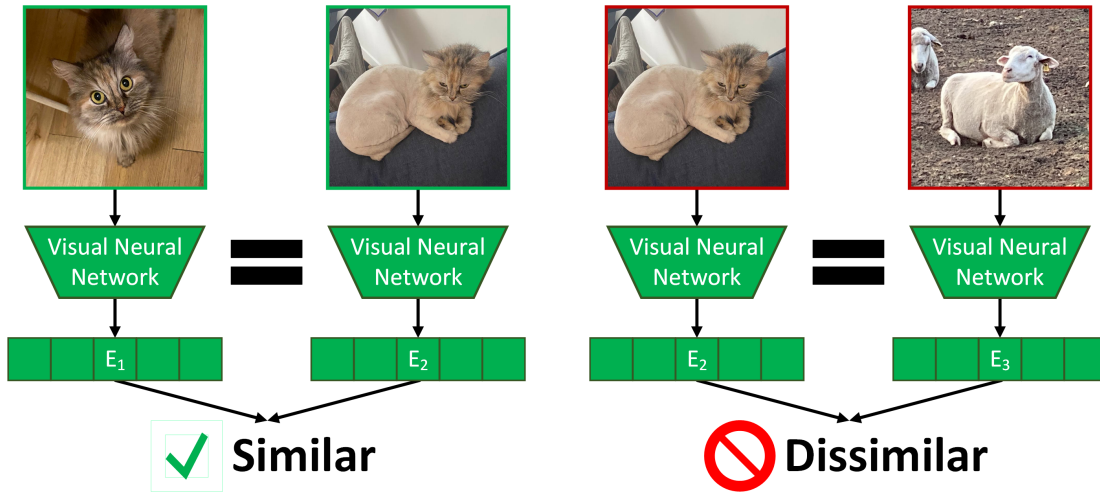


Fig. 2.10: Workflow of a sample visual Siamese network.

In the Siamese architecture, each input in a pair is processed independently by its respective subnetwork sharing the same architecture and weights, which allows them for extracting and learning representations that capture relevant patterns.

After processing the inputs through their respective subnetworks, the outputs are compared using a similarity measure or distance metric, such as Euclidean or cosine distance. The objective is to minimize distance—or maximize similarity—between similar inputs while maximizing distance between dissimilar inputs. Slightly different approach involves the use of a triplet loss [49] that focuses on identifying relationships among training instances by contrasting the example with positive and negative samples.

The versatility of the Siamese networks in capturing and comparing complex patterns makes them suitable for a wide range of tasks that involve learning similarity or dissimilarity relationships between pairs of data. Some notable applications include image retrieval [163], object tracking [84], and face [120] and gesture [280] recognition. This thesis leverages the Siamese architectural approach to incorporate multichannel processing into an audio-centric model, as described in Section 4.2.3.

2.1.7 Contrastive Learning

Contrastive learning [31] is an approach used in machine learning to train models leveraging representations that capture underlying structures and patterns within individual samples, and identify relationships between them.

Contrastive learning can be seen as a broader framework that encompasses various methods, including Siamese networks. In contrastive learning, the goal is to maximize the agreement between representations of similar instances while minimizing the agreement between representations of dissimilar instances, usually through the use of a contrastive loss function, as defined in Listing 2.1.

```

1  # image.shape == (batch_size, num_channels, height, width)
2  # text.shape == (batch_size, token_dim)
3
4  # features_*.shape == (batch_size, feat_dim)
5  features_img = clip.encode_image(image)
6  features_txt = clip.encode_text(text)
7
8  features_img = features_img / torch.linalg.norm(
9      features_img, dim=-1, keepdim=True
10 )
11 features_txt = features_txt / torch.linalg.norm(
12     features_txt, dim=-1, keepdim=True
13 )
14
15 # scaled pairwise cosine similarities
16 # logits_img_txt.shape == (batch_size, batch_size)
17 logit_scale = torch.exp(clip.logit_temperature)
18 logit_scale = torch.clamp(logit_scale, max=100)
19 logits_img_txt = logit_scale * features_img @ features_txt.T
20
21 reference = torch.arange(batch_size)
22 loss_img = torch.nn.functional.cross_entropy(
23     logits_img_txt, reference
24 )
25 loss_txt = torch.nn.functional.cross_entropy(
26     logits_img_txt.T, reference
27 )
28 loss = 0.5 * (loss_img + loss_txt)

```

Listing 2.1: PyTorch [196] code to compute symmetric cross entropy loss. Adapted from [206].

In Siamese networks, pairs of inputs are used by comparing them through applying shared weights and parameters, to explicitly learn similarities. The Siamese architecture focuses on learning representations that emphasize the differences between individual inputs, enabling tasks such as similarity measurement or identification, as described in Section 2.1.6.

Contrastive learning, while a powerful approach to self-supervised learning, is computationally intensive, requiring pairwise comparisons—thus $\mathcal{O}(N_{samples}^2)$ complexity—that may limit scalability. Additionally, it tends to focus on global similarities and capture

high-level semantic information, potentially limiting the model’s generalization ability for downstream tasks [156]. Although models exploiting this property are quite successful in global discrimination tasks [12, 16], considering local similarities improves performance in more fine-grained applications [271, 276].

A notable architecture within the field of multimodal learning is the CLIP model (Contrastive Language-Image Pretraining) [206]. CLIP is a powerful model that jointly learns visual and textual representations through the contrastive learning mechanism. It leverages a large-scale dataset of images and their associated text descriptions to learn a shared embedding space where images and text can be directly compared.

CLIP is built upon the Transformer architecture and trains the model to maximize the similarity between corresponding image-text pairs while minimizing the similarity between non-matching pairs [206], as depicted in Figure 2.11. By learning these joint representations, CLIP enables cross-modal retrieval, where queries can be performed in either modality to retrieve relevant instances from the other modality.

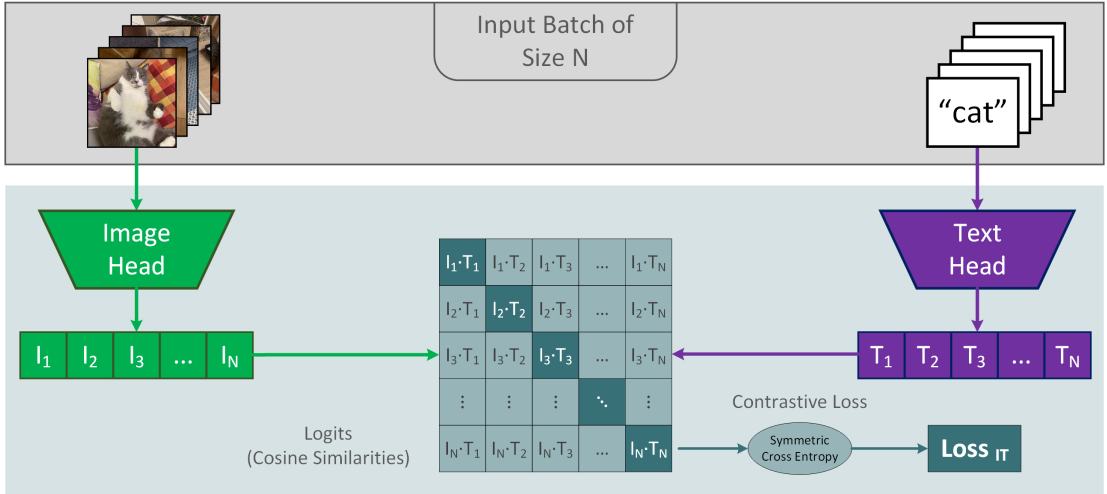


Fig. 2.11: Architecture of the CLIP model.

This thesis extends the CLIP framework to the audible modality and demonstrates in Chapter 5 the capabilities of the proposed AudioCLIP model [77] in various multimodal tasks, including single- and multilabel sound classification, multimodal retrieval and zero-shot environmental sound classification.

2.1.8 Transfer Learning

Transfer learning is a machine learning technique where knowledge gained from solving one task is applied to improve performance on a different but related task [183]. It

involves leveraging pre-trained models or features from a source domain and adapting them to a target domain, as depicted in Figure 2.12.

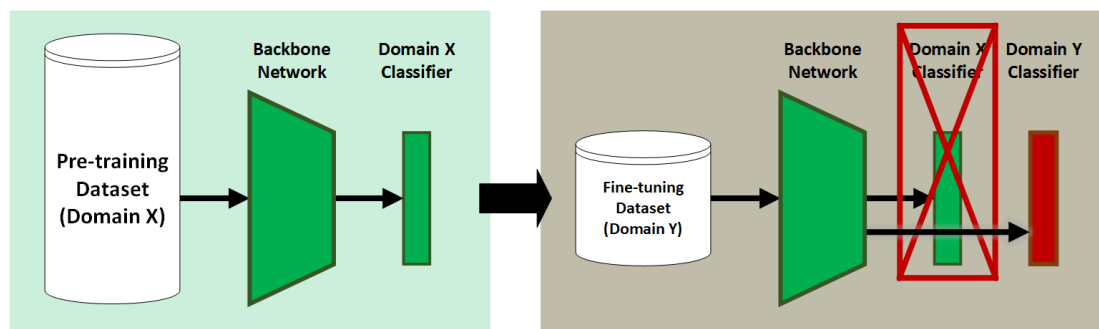


Fig. 2.12: Transfer learning workflow.

Transfer learning and multimodal knowledge distillation share similarities in their goal of transferring semantical information from one model or domain to another. Both approaches aim to improve the performance of a target model by utilizing knowledge obtained from a different model or domain.

In transfer learning, a pre-trained model, often trained on a large-scale dataset, is used as a starting point. The knowledge encoded in the model's parameters and learned representations is transferred to a target task or domain by fine-tuning the model on a usually smaller target dataset. This enables the model to learn task-specific features while benefiting from the general ones captured by the pre-trained model. However, success of the model fine-tuning in the transfer learning setup depends on the similarities and differences between the source and target domains [257]. If the dissimilarity between the domains is substantial, transfer learning may not yield significant benefits, so addressing this limitation often involves careful consideration of domain alignment [191, 57].

Multimodal knowledge distillation, in turn, focuses on transferring knowledge between different modalities. It involves training a student model to mimic the predictions or behavior of a teacher model that is proficient in multiple modalities. The teacher model's knowledge, which captures the relationships and dependencies between modalities, is distilled into the student model, enabling it to effectively handle multimodal inputs.

The AudioCLIP model exemplifies the combination of transfer learning and multimodal knowledge distillation. Parts of AudioCLIP—i.e., CLIP: text- and image-head—were trained on a large-scale dataset containing images and their associated text descriptions. By jointly training the model on these multimodal inputs, CLIP learned a shared embedding space where images and text can be directly compared.

In AudioCLIP, transfer learning is utilized by leveraging a pre-trained vision model and a language model to provide a cross-modal supervision when training an audio-head. The knowledge captured by these models is distilled into the audio-centric part of the AudioCLIP model through the contrastive learning framework, enabling it to perform multimodal tasks such as classification and audio/image/text retrieval in various combinations, as described in Section 5.3.

2.1.9 Zero-shot Learning

Zero-shot learning is a machine learning approach where a model is trained to recognize and classify objects or concepts belonging to categories that it has never encountered during training. It enables the model to generalize to new classes by leveraging auxiliary information or attributes associated with those classes.

In zero-shot learning, the model is trained on a set of seen classes for which labeled data is available. Additionally, the model is provided with semantic information or attributes that describe the unseen classes, as shown in Figure 2.13. These attributes capture high-level characteristics or properties of the classes, such as textual descriptions or modality-specific features.

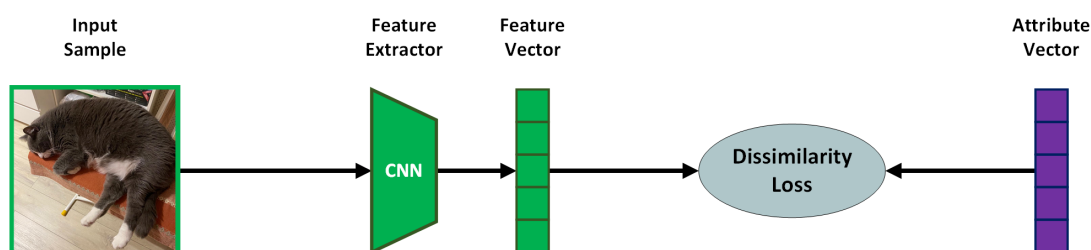


Fig. 2.13: Workflow of a sample zero-shot learning setup.

During inference, the model can recognize and classify instances from both seen and unseen classes by leveraging the learned relationships between the semantic attributes and the identified features. Zero-shot learning assumes inherently that a model can generalize to unseen classes based on knowledge learned from seen classes. However, it may not hold in cases of significant domain shifts or dissimilarities between seen and unseen classes [121]. Recent research has actively addressed the domain shift problem in zero-shot learning, with a common approach involving the application of generative models to bridge the gap between seen and unseen domains [113, 62, 282].

The CLIP model demonstrates the concept of zero-shot learning in a text-visual setting, as a result of a contrastive cross-modal training on a large-scale dataset of images with accompanying textual descriptions. The AudioCLIP model further developed the ideas of

CLIP by extending the set of modalities to the audible one. By learning a joint embedding space for audio, images and text, AudioCLIP is able to associate audio representations with textual attributes and visual concepts [77].

AudioCLIP exhibits zero-shot capabilities by recognizing and classifying sound categories that were not present in the training data. This is achieved by leveraging the textual descriptions and visual attributes associated with these unseen classes. By generalizing from the learned relationships between audible and visual/textual information, AudioCLIP can make accurate predictions for novel classes, as demonstrated in Section 5.4.

2.2 Time-frequency Analysis of Audio

The analysis of audio signals in the time domain provides information about the signal's temporal evolution. However, this information might not be sufficient for tasks that require more detailed information about the frequency-domain content of the signal. In such cases, the time-frequency representation of audio signals is essential. Time-frequency analysis is a fundamental signal processing technique that provides a joint representation of the temporal and spectral information of a signal. This section introduces basic concepts of time-frequency analysis, including windowing, the Fourier transform, and the Short-Time Fourier Transform. Additionally, it covers advanced techniques such as the Continuous Wavelet Transform and particular mother wavelets.

Window Function

In time-frequency analysis, windowing refers to the process of multiplying a signal by a window function, which is a mathematical function that goes to zero at its edges. The purpose of windowing is to mitigate the effect of spectral leakage, which can occur when the signal is not periodic and its Fourier transform is taken over a finite time window. Spectral leakage causes the energy of the signal to leak into adjacent frequency bins, leading to ambiguous frequency estimates. By multiplying the signal by a window function, which is not trivial (as, for example, the boxcar window that is depicted by the red line in Figure 2.14), the edges of the signal are tapered to zero, reducing the amount of leakage and providing a more accurate frequency representation of the signal. Windowing is commonly used in Fourier-based analysis techniques such as short-time Fourier transform and spectrogram analysis.

The Blackman-Harris window [83], which is given by Equation 2.24 (black line in Figure 2.14), is a general-purpose window that has good performance characteristics in terms of its main lobe width, sidelobe attenuation, and stopband attenuation [86]. Specifically, the Blackman-Harris window has a narrow main lobe width, which provides good frequency resolution, and its sidelobe attenuation is high, which contributes to the reduction of spectral leakage.

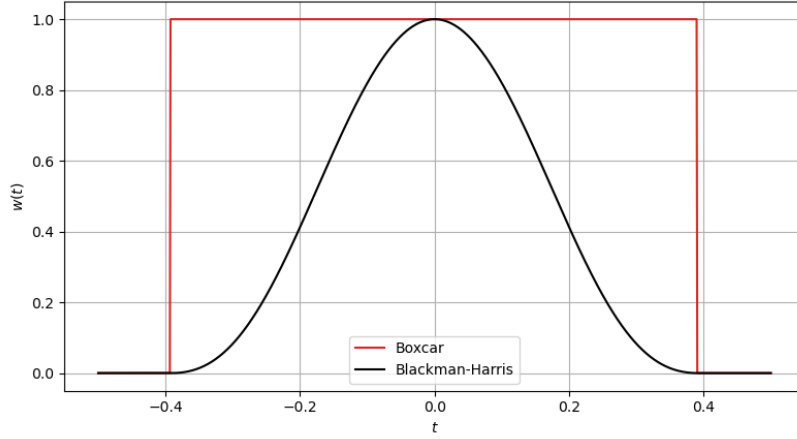


Fig. 2.14: Sample window functions.

$$w[k] = a_0 - a_1 \cos\left(\frac{2\pi k}{N}\right) + a_2 \cos\left(\frac{4\pi k}{N}\right) - a_3 \cos\left(\frac{6\pi k}{N}\right) \quad (2.24)$$

$$a_0 = 0.35875; a_1 = 0.48829; a_2 = 0.14128; a_3 = 0.01168$$

As the stopband attenuation of the Blackman-Harris window is better than many other commonly used window functions, it becomes particularly useful for applications that require unambiguous spectral estimates and low noise [86]. Overall, it makes the minimum four-term Blackman-Harris window a reliable choice of a window function for time-frequency analysis due to its characteristics resulting in ability to obtain high-quality signal representation in a variety of applications.

2.2.1 Fourier Analysis

Fourier analysis is a mathematical technique that decomposes a complex signal into its constituent frequencies. In the context of audio signals, Fourier analysis is applied to convert a time-domain audio signal into its frequency-domain representation. This is done using the Fourier transform, which is a mathematical technique for expressing a

signal as a sum of sine and cosine waves of different frequencies, amplitudes, and phases (Equation 2.25). By applying the Fourier transform to audio, one can analyze its spectral content, which provides the information about the elementary sound components present in the signal.

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(t) \cdot e^{-i\omega t} dt \quad (2.25)$$

Here, the transform of a continuous time-domain function $f(t)$ at angular frequency ω is denoted by the complex number $\hat{f}(\omega)$.

Discrete Fourier Transform

The Discrete Fourier Transform (DFT) is a discretized in time version of the Fourier transform, which is a mathematical technique for transforming signals between the time domain and the frequency domain. The DFT is a way to perform Fourier analysis on a discrete set of data points, which is the case for digital signals. The DFT is an algorithm that takes a finite sequence of equally spaced samples of a function x_n , as shown in Equation 2.26, and transforms it into a sequence of complex numbers that represent the signal in the frequency domain, using a DFT-kernel defined by Equation 2.27. The main difference between the regular Fourier transform and the DFT is that the regular Fourier transform produces a continuous spectrum, while the DFT produces a discrete spectrum of frequencies [186]. Time- and frequency-domain representations of sample Fourier filters are shown in Figure 2.15.

$$X(f_c) = \sum_{n=0}^{N-1} x_n \cdot K_{DFT}^{f_c}(n) \quad (2.26)$$

where

$$K_{DFT}^{f_c}(n) = e^{-i\frac{2\pi}{N}f_c n} \quad (2.27)$$

Here, the transform of a discrete sequence $x_n \in \mathbb{C}$ at central frequency f_c is denoted by another sequence $X(f_c) \in \mathbb{C}$ of length N .

The DFT is commonly used in audio signal processing for various purposes that encompass spectral analysis to identify the frequency components of a signal [11], filtering to remove or modify certain components [17], compression to reduce the size of a file [110], watermarking to embed imperceptible and robust information into a sound

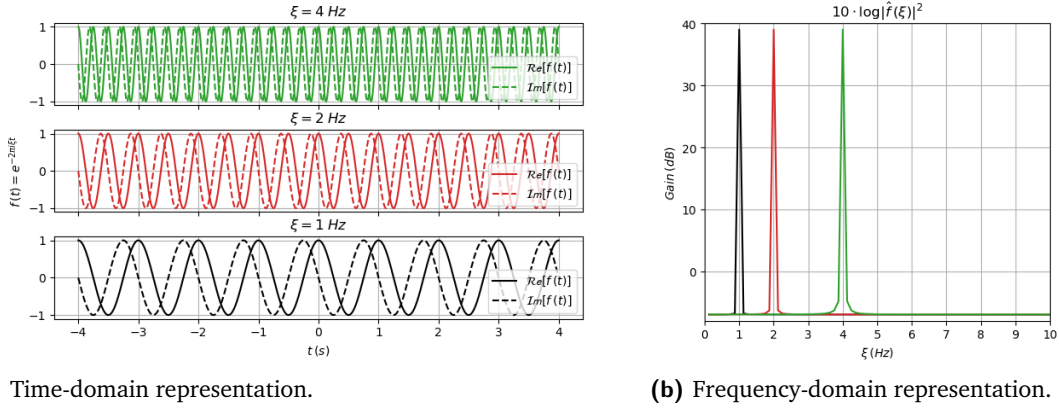


Fig. 2.15: Sample filters of a Fourier filter bank.

recording [225], and synthesis to generate new audio signals from existing ones by manipulating their frequency components [217].

Short-time Fourier Transform

The Short-Time Fourier Transform [10] is a technique used in time-frequency analysis to determine the frequency content of a signal over time. It allows observing how frequencies comprising the signal change over time by breaking the signal down into small overlapping segments, or windows, and performing a Fourier transform on each window separately. This is done by multiplying the signal fragments by a window function and then taking the Fourier transform of the resulting windowed sequences. In practice, to compute Equation 2.28, one applies the fast Fourier transform [40] to each frame of the windowed signal.

$$X(\tau, f_c) = \sum_{n=0}^{N-1} x[n]w[n - \tau]K_{DFT}^{f_c}(n) \quad (2.28)$$

The resulting frequency domain representation of windowed segments is then stacked to form a two-dimensional matrix called spectrogram (Figure 2.16), where the horizontal axis refers to time and the vertical axis denotes frequencies. The STFT is commonly used in audio signal processing for tasks such as spectral analysis [19], feature extraction [211, 52], and time-frequency visualization [218, 275] by determining magnitude and phase of basis sinusoidal frequencies f_c at different time points τ in a time-domain signal x .

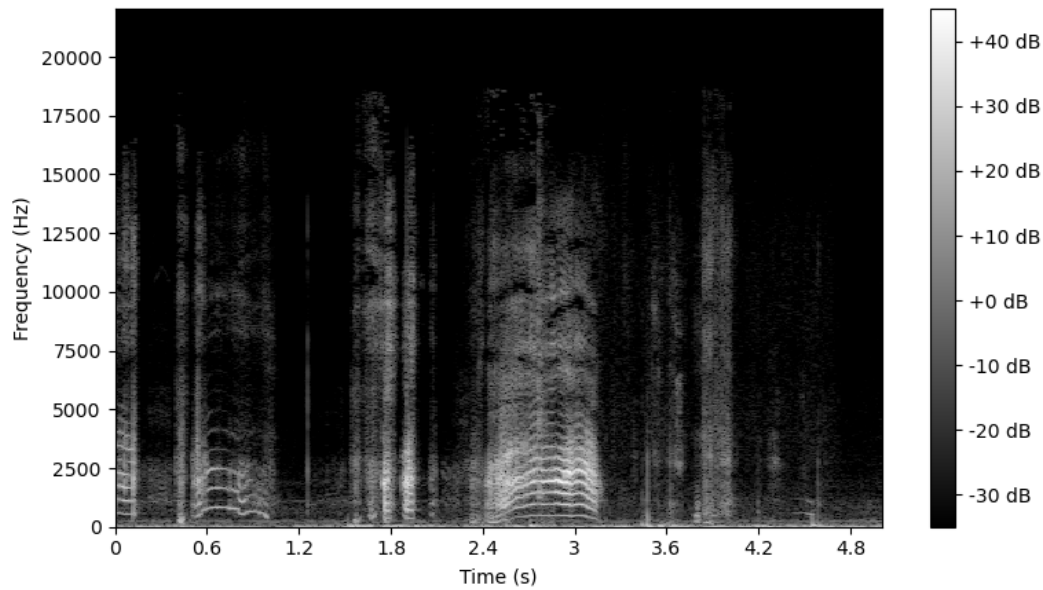


Fig. 2.16: Sample STFT-spectrogram.

2.2.2 Wavelet Analysis

Wavelet analysis is an advanced mathematical tool used for time-frequency analysis. It involves decomposing a signal into a set of wavelets, which are small waves that are localized in both time and frequency [43]. Wavelets can capture both high- and low-frequency components of a signal and thus, can provide a more detailed and accurate representation of a signal's time-varying frequency content than other traditional Fourier-based methods [245].

In audio signal processing, wavelet analysis can be applied to detect and analyze the temporal variations in the spectral content of a signal [250]. It also allows for extracting information from different regions of the frequency spectrum, making it useful for tasks such as audio compression [234], noise reduction [9], and feature extraction for machine learning applications [128, 296, 109].

The process of applying wavelet analysis to an audio signal involves decomposing the signal into a series of wavelet coefficients at different scales and positions in time. This is achieved through a process called wavelet transform. The wavelet transform can be performed using different types of wavelet functions, each with different properties that make them suitable for different types of signals or analysis purposes [180]. This thesis proposes in Chapter 4 a learnable time-frequency transformation that relies on the fb-sp-wavelet discussed later in the current section.

Continuous Wavelet Transform

The continuous wavelet transform [223] is a time-frequency analysis technique that differs from the discrete Fourier transform and the short-time Fourier transform, which use fixed basis functions, by using a family of functions called wavelets. Wavelets are small, localized functions that can be used to analyze signals at different scales and frequencies, as shown in Equation 2.29.

$$\hat{f}_\psi(a, b) = \frac{1}{|a|^{\frac{1}{2}}} \int_{-\infty}^{\infty} f(t) \bar{\psi}\left(\frac{t-b}{a}\right) dt \quad (2.29)$$

In the CWT, this is done by convolving a time-domain signal $f(t)$ with the complex-conjugate wavelet $\bar{\psi}$ that is scaled by the factor a and translated in time by the factor b . The result is a time-frequency representation $\hat{f}_\psi(a, b)$ of the signal, where the frequency resolution is high at high frequencies and low at low frequencies. The advantage of the CWT is that it can capture both time and frequency information at different scales and resolutions, allowing for a more detailed analysis of non-stationary signals such as audio.

On the other hand, the CWT is computationally expensive and can take a significant amount of time to compute, especially for long signals [216]. Additionally, since the continuous wavelet transform applies a continuously varying window to a finite-length signal, it can introduce edge effects or boundary effects [237, 67], where the analysis is affected by the beginning and end of the signal. Another limitation is that the selection of the wavelet function and its parameters can significantly influence the analysis [180]. Therefore, careful selection and tuning of the wavelet function and its parameters are necessary for obtaining accurate results [171].

Shannon Wavelet

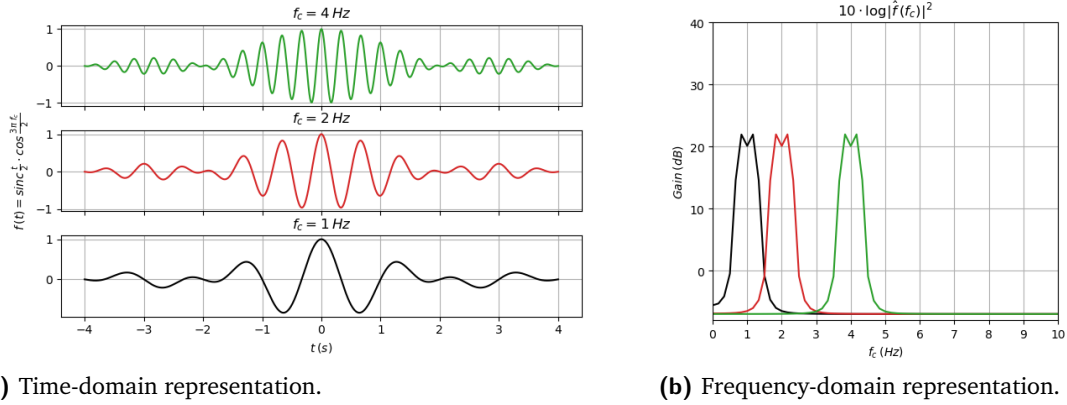
The Shannon wavelet is a type of wavelet that is widely used in signal processing [36] and for solving integrodifferential equations [24]. Similarly to sinusoidal functions used in the Fourier transform, the Shannon wavelet is a noncompact waveform that has compact support in the frequency domain (Equation 2.30). The Shannon wavelet is obtained by modulating a cosine function with a sinc window (Equation 2.31).

$$K_{Shannon}(t) = \text{sinc} \frac{t}{2} \cdot \cos \frac{3\pi}{2} \quad (2.30)$$

where

$$\text{sinc } t = \frac{\sin \pi t}{\pi t} \quad (2.31)$$

The sinc window is a bell-shaped curve that infinitely tapers off the cosine function. The product of the cosine function and the sinc window results in a waveform that is concentrated in both time- and frequency-domain, as shown in Figure 2.17.



(a) Time-domain representation.

(b) Frequency-domain representation.

Fig. 2.17: Sample Shannon wavelets.

In the Fourier transform, the analysis is performed using a fixed set of sinusoidal functions that are defined for the entire duration of the signal. This results in a global analysis that may not capture changes of the signal properties in time. Using the Shannon wavelet, on the other hand, results in a time-frequency representation that is well-suited for analyzing nonstationary signals [142].

Complex Frequency B-Spline Wavelet

The complex frequency B-spline (fbsp) wavelet [245] is a type of wavelet used in time-frequency analysis, which is described in terms of order m , bandwidth f_b and central frequency f_c (Equation 2.32). In comparison to the Shannon wavelet, which is a real-valued windowed function, the fbsp-wavelet is a complex-valued function that is modulated by a base sinusoid. Figure 2.18 demonstrates sample fbsp-filters given a fixed central frequency f_c and order m .

$$K_{fbsp}^{m-f_b-f_c}(n) = \sqrt{f_b} \cdot \left(\text{sinc} \frac{f_b n}{Nm} \right)^m \cdot e^{i \frac{2\pi}{N} f_c n} \quad (2.32)$$

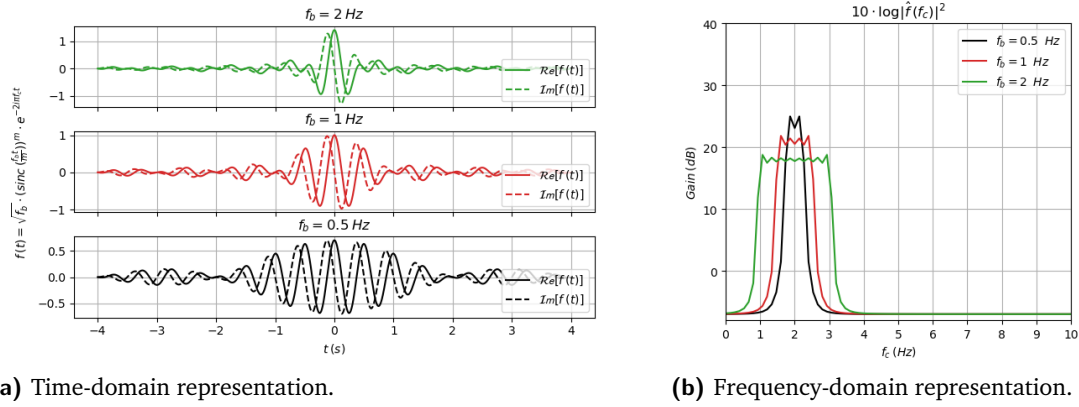


Fig. 2.18: Sample fbsp-wavelets at $f_c = 2$ Hz and $m = 1$.

The main difference between the fbsp- and the Shannon wavelet is that the former can achieve higher time-frequency resolution due to its ability to tune its bandwidth and phase shift, i.e. it can be optimized to specific task requirements [245]. This richer representation helps boosting the performance of audio-visual models by preserving fine-grained temporal and spectral details, aiding in tasks such as sound localization [188], source separation [115], or audio-visual speech enhancement [2] where accurate phase information is crucial for effective integration of audible and visual cues.

Concluding Remarks

This chapter has offered a comprehensive bottom-up overview of deep learning methods and signal processing techniques, providing a foundation for understanding the nuances of audio-centric multimodal learning. The knowledge acquired here will serve as an essential underpinning for the subsequent chapters. Moving forward, the focus of the next chapter will shift towards a survey on audio-centric tasks, delving into the specific challenges and applications within the multimodal domain. This survey will not only expand understanding of existing areas of research but will also establish a methodological basis for the forthcoming experimental evaluation.

Knowledge Framework and Conceptualization

This chapter serves as a pivotal bridge between the theoretical foundations of audio-visual deep learning models and the experimental evaluation of approaches proposed later in this thesis. By delving into the intricate details of the existing literature, this chapter lays the groundwork for establishing independent model classification criteria, namely modeling, modality, and supervision. The exploration of these criteria enables a comprehensive understanding of the various approaches, techniques, and methodologies employed in the field. Moreover, it provides a conceptual framework that guides the narrative in the exploration of novel challenges and advancements in audio analysis for multimodal learning. Through this integrative approach, the chapter contributes not only by elucidating the state of the field but also by establishing a clear and coherent trajectory for the empirical contributions presented in the subsequent chapters.

First, the chapter provides an overview of the deep learning landscape related to audio-visual learning, which encompasses a wide range of topics, covering representation learning, correspondence learning, audio-visual object localization, sound source separation, and audio-visual generation. Second, the chapter identifies and explores learning objectives related to audio analysis, such as audio-visual multimodal classification, regular and zero-shot environmental sound classification, and multimodal retrieval. The selection of these specific objectives is based on their importance in the field and their potential for addressing real-world problems. Thus, the chapter offers a comprehensive overview of state-of-the-art approaches in audio-visual learning, while also aiding in the identification of gaps and opportunities for future research.

3.1 Multimodal Learning with Audio Signals

The widespread application of neural architectures and supporting technologies has enabled their use in various fields and tasks, such as entertainment, education, and

healthcare [205, 14]. As neural network-based methods continue to develop, their complexity and scale have been growing exponentially. This trend is exemplified by a shift from simple single-modality classification tasks to more complex learning objectives that involve the use of multiple modalities simultaneously.

Integrating information from different sensory modalities is a crucial aspect of human perception, which includes visual, auditory, olfactory, and tactile cues [195]. Multimodal machine learning (ML) has emerged as an active research field aimed at developing algorithms and models that can combine information from multiple modalities to tackle challenging learning tasks. Using multiple modalities helps mitigate the limitations and uncertainties associated with a single modality, resulting in more accurate and robust models. For instance, in video analysis, incorporating audio modality provides complementary temporal information that is useful in aligning different parts of the video sequence and enhancing the accuracy of analysis [150]. Therefore, sound plays a crucial role in complementing other modalities, especially vision, and can significantly improve the performance of the model [77].

The realm of multimodal machine learning itself is full of details and demands proficiency across diverse domains such as computer vision, natural language processing, and signal and audio processing. Navigating this interdisciplinary landscape is further complicated by the continuous evolution of new architectures and applications, necessitating researchers to remain abreast of the latest advancements. One of the principal challenges in evaluating multimodal machine learning models lies in the subjectivity and elusive nature of certain criteria used for assessment. As a prime example, evaluating the aesthetic quality of generated videos and sounds involves subjective judgments, and, hence, poses a significant challenge due to the personal and context-dependent nature of aesthetic experiences. Nevertheless, multimodal machine learning holds significant promise due to its potential to address complex learning tasks that align with the inherently multimodal nature of human perception and cognition. This stems from the unique ability of multimodal approaches to emulate and excel at tasks closely tied to the way humans naturally perceive and learn [80].

The objective of this thesis is to investigate how neural network architectures can be designed and improved for solving audio-centric multimodal tasks. The focus of the research lies on five challenging multimodal problems that rely on sound signals. These tasks encompass representation learning, correspondence learning, audio-visual object localization, sound source separation, and audio-visual generation.

Given the similarities between the aforementioned tasks, it is important to highlight the main differences of model architectures that have been proposed. To this effect, there was defined a taxonomy of audio-visual learning systems that distinguishes between

three main axes: modeling, modality, and supervision. This choice of classification is not common, but it allows for a versatile examination of current multimodal learning methods that incorporate sound. The mutual independence of the aspects means that they can vary without impacting each other, thereby providing a solid basis for an architecture-agnostic and task-agnostic analysis based on a thorough review of the current literature.

Representation learning is a type of machine learning that focuses on identifying important features within data to solve a specific task [22]. This process involves exploring the internal structure of the data to identify these key features. It is considered a low-level task, as it involves the automatic exploration of input features. Furthermore, the subsequent categories of tasks discussed in this thesis rely on representation learning either directly or indirectly.

Moving beyond representation learning, *correspondence learning* is a more complex task that involves identifying semantic relationships between disparate inputs [304]. This type of task often includes recognition, retrieval, and detection as common examples. Recognition involves identifying learned features directly, while retrieval takes into account the relevance of the output given a query. Detection goes even further by providing spatial information about recognized entities in addition to their identities.

Audio-visual object localization [304] is a more specific and fine-grained task compared to object detection. It requires constraints on the number of objects and their specific classes. The goal is to determine the precise location of an object in both the audio and visual domains, by exploiting the complementary information from both modalities. This task is particularly useful in scenarios where visual information alone is insufficient, such as in low-light conditions or when the object of interest is occluded.

Compared to the tasks discussed earlier, a *sound source separation* [166] model is special in that it actively modifies its input to produce isolated audio streams, rather than passively predicting class labels, coordinates, or timings. This fundamental difference in the model's behavior leads to a significant increase in complexity.

The task of *generation* [69] involves creating output within a specific domain, such as visual [29] or audible [177], based on given initial conditions. This task is already challenging in a unimodal setting, and becomes even more difficult when multiple modalities are involved. Despite these challenges, recent progress in generative models suggests that video generation is likely to receive increased attention in the near future.

3.1.1 Taxonomy of Audio-visual Learning Systems

The area of deep learning models for multimodal analysis has been gaining popularity lately, particularly with regards to audio-visual learning. Various researchers are exploring innovative approaches to tackle the challenges that arise in this field, including cross-modal representation learning [124, 12, 155], multimodal fusion [175, 94], and temporal alignment [95, 94]. Recent survey papers provide an overview of the latest state-of-the-art methods in audio-visual learning. Typically, authors of these reviews consider several aspects that are commonly found in such studies.

Two common aspects considered by authors of reviews on multimodal learning are the model *architecture* and *learning objective*. These two aspects are often closely interrelated because the choice of a specific architecture design depends significantly on the task at hand. For instance, when training a model to perform a specific task, one typically chooses a suitable tool, such as a convolutional neural network (CNN) or a Generative Adversarial Network (GAN) [69], depending on the task requirements.

Although the learning objective aspect may seem simplistic, recent review papers utilize it mainly to assist in constructing a comprehensive and coherent representation of the field [20, 304, 44, 55, 143, 229, 305], rather than using it as an independent classification aspect [166, 262]. However, some authors have concentrated specifically on the learning objective aspect by proposing a high-level classification of the frequently appearing tasks [262]. In this thesis, the same approach was adopted by using several machine learning tasks as benchmarks to describe a range of models on both task-specific and complexity development levels.

The architecture of deep learning models is often used to characterize them in the context of multimodal learning [111, 60, 107, 240, 166, 304, 143, 44, 207]. However, concentrating too much on technical details and ignoring the broader context and applications of multimodal learning could limit the scope of relevant aspects to be addressed.

Many survey papers have identified modality as an important aspect in multimodal learning [20, 229, 305]. While some researchers use it for narrative purposes, the aspect of modality can also be used to classify machine learning models independently of their architecture or task [240, 207]. When combined with *supervision* [207], it provides a basis for “orthogonal” classification of models, which can be further extended by adding the third independent aspect, modeling, as done in this thesis.

Recent review papers [166, 60] mention the *fusion strategy* aspect, which is not as commonly discussed as other aspects. It describes how different modalities are combined

in a model, such as early fusion at the input level or late fusion at a later stage in the model. However, by focusing on implementation details, the aspect of fusion strategy may neglect providing a high-level overview of the field.

The aspect of *model performance*, much like the fusion strategy, has limited use in the systematic overview of the field of multimodal learning [240, 107]. While it allows for the comparison of various architectures, the challenge lies in ensuring the same or comparable conditions for model evaluation (e.g., dataset, task), which makes it difficult to use when comparing a broad range of models.

The objective of the current section is to propose a method for categorizing models based on three independent aspects: *modeling*, *modality*, and *supervision*, as shown in Figure 3.1. The use of modeling and supervision to classify methods in contrastive representation learning is demonstrated in [111]. These aspects were selected because of their independence from each other, as well as from modality. The level of supervision in model training is independent of the modalities used or the position of the model on the spectrum of “discriminative – generative”. Similarly, the other two aspects are task-invariant and independent of the level of supervision, as well as mutually, thus providing a basis for model classification.

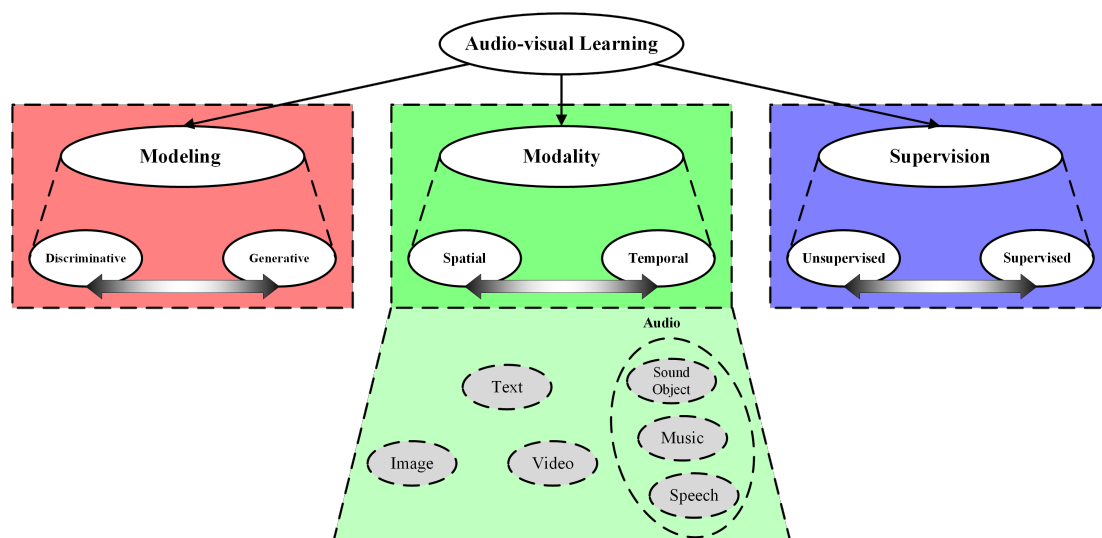


Fig. 3.1: Proposed model description basis.

The *modeling* aspect pertains to the capability of a model to produce its output in a non-deterministic manner, and it distinguishes between the discriminative and generative poles of the spectrum, which are different from those defined through conditional probability in [179]. Just like the other aspects mentioned before, the modeling aspect is essential in describing a wide range of models that address a variety of tasks. In this

work, the extent of a model's output produced in a (non-)deterministic way defines the balance between the discriminative and generative poles of the spectrum.

The *modality* aspect refers to the type of signals that a model processes. These signals are represented according to their modality, which determines to how they are carried out or perceived [20]. Describing the modality aspect is fundamental for characterizing an ML-model, and it can be applied to a diverse range of architectures, tasks, datasets, etc.

This thesis distinguishes between two opposing poles of modality: spatial and temporal. The placement of a model on the modality axis depends on the specific input types that it is designed to handle, such as images or speech. Multimodal models typically focus on a single primary modality or treat all inputs equally. Audio processing models typically use audio or video as the primary input, depending on the task. The former approach is commonly used for representation and correspondence learning, audio-visual localization, sound source separation, and sound generation tasks, while the latter is more relevant for video generation, which prioritizes the visual modality.

The aspect of *supervision* in machine learning models corresponds to the requirement of labeling training data, which reflects the real-world knowledge being modeled. This labeling is necessary to manage the modeling process, and every model training procedure can be described in terms of supervision, making it a universal criterion for assessing machine learning methods. In this thesis, the models were described based on their architectural features, such as the use of contrastive learning, and the extent to which they required semi-automatic or manual modification of existing datasets or the creation of new ones.

3.1.2 Representation Learning

The primary goal of representation learning is to develop more efficient and powerful methods for machine learning that aid in identifying, extracting, and learning relevant features and patterns from input data. The development of high-quality representations is critical to building downstream task models with optimal performance, such as classifiers and segmentation models [22].

ML-models process information differently than humans do, and there is no universal way to encode real-world data into algorithms, resulting in various ways of representing information [304]. Therefore, each model aims to learn its own way of representation, and the efficacy of downstream tasks performed by the model largely depends on the comprehensiveness, clarity, and expressiveness of the learned representations.

Audio plays a crucial role in multimodal representation learning because it can capture unique temporal aspects of the environment that are difficult to obtain through other modalities. This includes events such as footsteps, rustling leaves, and traffic, which can provide important contextual information. In addition, sounds can convey semantic information such as speech, music, and environmental sounds, which are challenging to extract from visual cues. Furthermore, audio can be particularly useful when visual signals are limited or noisy, such as in low-light or partially occluded scenes. This can provide important hints that help identify the object or event of interest, leading to improved performance in multimodal representation learning. Furthermore, audio can reveal critical temporal information that is absent from static visual inputs. Audio signals contain rich time-varying patterns that allow for capturing dynamics of events, such as the rhythm of music or the cadence of speech, which is essential for tasks such as action recognition or event detection.

Representation learning has been increasingly combined with other tasks in recent years, although it remains a fundamental component. As a result, there are fewer models exclusively dedicated to this task. Nonetheless, there are some recent models that focus primarily on learning representations, as outlined in Table 3.1.

Tab. 3.1: Models learning representations.

Year	Model	Source	Modeling ^A		Modality ^B		Supervision ^C	
			D	G	S	T	U	S
Unimodal								
2017	Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification	[224]						
2017	Sequence to sequence autoencoders for unsupervised representation learning from audio	[14]						
2017	Unsupervised learning of semantic audio representations	[102]						
2018	auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks	[58]						
2019	An unsupervised autoregressive model for speech representation learning	[37]						
2019	Unsupervised speech representation learning using wavenet autoencoders	[35]						
Multimodal								
2018	Cooperative learning of audio and video models from self-supervised synchronization	[124]						
2019	Zero-shot audio classification based on class label embeddings	[272]						
2020	Active contrastive learning of audio-visual video representations	[155]						
2020	Multi-modal self-supervision from generalized data transformations [197]							
2020	Self-supervised learning by cross-modal audio-video clustering	[12]						
2020	Self-supervised multimodal versatile networks	[8]						
2021	AudioCLIP: Extending CLIP to Image, Text and Audio	[77]						
2021	Audio-visual instance discrimination with cross-modal agreement	[173]						
2021	Enhancing audio-visual association with self-supervised curriculum learning	[291]						
2021	TriBERT: Full-body Human-centric Audio-visual Representation Learning for Visual Sound Separation	[209]						
2021	Zero-shot audio classification via semantic embeddings	[273]						

^A Modeling: Discriminative, Generative.

^B Modality: Spatial, Temporal.

^C Supervision: Unsupervised, Supervised.

Unimodal Representation Learning

Unimodal representation learning models unify both the discriminative and generative approaches among [224, 102, 58, 37] as well as within models [14, 35]. Being a basic

task, unimodal representation learning can be addressed in numerous ways, as can be observed in Table 3.1.

This class of models focuses on the audible signal, which is inherently temporal. All of the described models center their architectures around time-dependent input, thus contributing to the homogeneity among the approaches aiming to learn representations in a unimodal manner.

Additionally, most of the models either exploit unsupervised learning methods [102, 58, 37, 35] or combine them with supervised approaches [224, 14].

Multimodal Representation Learning

When examining the aspect axes in Figure 3.1, it becomes apparent that generative approaches are not widely used among researchers working with these models. This is likely due to the straightforward nature of the task, in contrast to more complex objectives, which are discussed in Section 3.1.5 and Section 3.1.6.

In terms of modality, many multimodal approaches that use audio tend to prioritize the temporal aspect. Some methods focus primarily on the time-domain signals, as seen in works such as [272, 197, 273]. On the other hand, fewer methods place emphasis on the spatial component, as described in [77].

Various specific problem formulations may serve as a starting point for solving representation learning due to its fundamental nature in relation to other tasks. Supervised [272, 155, 173, 209, 273] and self-supervised [124, 197, 12, 8, 77, 291] approaches are commonly used, with a slight preference for the latter, to solve this problem. Self-supervised approaches are used more often in more challenging tasks, while less supervised paradigms are followed in such scenarios, as discussed in Section 3.1.5 and Section 3.1.6.

3.1.3 Correspondence Learning

Models that learn correspondences aim to discover semantic correlations between different inputs at a higher level of abstraction, similar to models learning representations. Instead of focusing on features useful for a particular input or modality, these models find useful relationships within the data. As shown in Table 3.2, for multimodal spatio-temporal setups, such models require a fine-grained time-domain alignment between the modalities involved and a rather coarse spatial cross-modal matching.

Tab. 3.2: Models learning correspondences.

Year	Model	Source	Modeling		Modality		Supervision	
			D	G	S	T	U	S
Recognition								
2020	Audiovisual slowfast networks for video recognition	[271]						
2020	Labelling unlabelled videos from scratch with multi-modal self-supervision	[16]						
2020	Multi-modal self-supervision from generalized data transformations	[197]						
2021	Attention bottlenecks for multimodal fusion	[175]						
2021	Cross-modal spectrum transformation network for acoustic scene classification	[146]						
2021	Distilling audio-visual knowledge by compositional contrastive learning	[32]						
2021	Localizing visual sounds the hard way	[27]						
2022	Audio-visual scene classification via contrastive event-object alignment and semantic-based fusion	[94]						
Retrieval								
2019	Jointly learning of visual and auditory: A new approach for RS image and audio cross-modal retrieval	[289]						
2020	Deep triplet neural networks with cluster-CCA for audio-visual cross-modal retrieval	[75]						
2020	AVLNet: Learning audio-visual language representations from instructional videos	[219]						
2020	Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos	[194]						
2021	Audio retrieval with natural language queries	[184]						
2021	AudioCLIP: Extending CLIP to Image, Text and Audio	[77]						
2021	AVGZSLNet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings	[162]						
2021	Distilling audio-visual knowledge by compositional contrastive learning	[32]						
2021	HiT: Hierarchical transformer with momentum contrast for video-text retrieval	[144]						
2021	T2VLAD: global-local sequence alignment for text-video retrieval	[260]						
2022	ECLIPSE: Efficient Long-range Video Retrieval using Sight and Sound	[140]						
Detection								
2020	Emotions don't lie: An audio-visual deepfake detection method using affective cues	[170]						
2020	Audiovisual slowfast networks for video recognition	[271]						
2020	Self-supervised learning of audio-visual objects from video	[3]						
2021	Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection	[244]						
2021	There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge	[252]						
2022	Self-supervised object detection from audio-visual correspondence	[1]						

Aligning audio and visual signals is a major hurdle in correspondence learning due to differences in sampling rates, frame rates, and processing delays [271, 260]. Moreover, variations in audio-visual correspondence across different contexts and environments pose a challenge to developing a generalized model that can function effectively in different situations.

When dealing with correspondence learning, the audible modality plays a significant role in effectively aligning and fusing different modalities [124]. The audible information provides additional hints that aid in accurately synchronizing the audio and visual inputs. As the audio stream is generally synchronized with visual events—for instance, the sound of a basketball hitting the ground—using audible cues improves the alignment of the audible and visual streams, even when there is noise or other distortions [124].

This work regards correspondence learning as a machine learning task that is more abstract than representation learning. This implies that models learning correspondences also learn representations. The term “correspondence learning” encompasses a wide range of applications, and the chapter concentrates on three specific topics to investigate the subject from different viewpoints.

While there is a conceptual difference between models that learn correspondences and those that learn representations, models that learn correspondences can be seen as a progressive development towards solving more high-level problems. This implies that in terms of architecture, training, and application, both types of models share similar characteristics.

The following sections identify three distinct tasks within the domain of correspondence learning: *recognition*, *retrieval*, and *detection*. Recognition entails identifying and categorizing specific objects, actions, or events within an audio-visual scene, while detection involves identifying whether a particular object or event is present or absent. Finally, retrieval focuses on finding relevant information within a large dataset.

Recognition in Correspondence Learning

The field of recognition within correspondence learning is focused on designing models that can directly identify patterns in the data. Unlike unimodal recognition tasks, multimodal recognition tasks utilize multiple input streams at once, which provide complementary information and give the model a stronger semantic understanding of the data. This is particularly useful when applied to various types of audio-visual recognition tasks such as object recognition [271], video labeling [16], or acoustic scene classification [146, 94].

The effective alignment and fusion of different modalities present a significant challenge in the multimodal recognition task. The variations in correspondence across diverse environments and contexts make it difficult to create a generalizable model that can perform well in different setups [27].

Recognition models presented in Table 3.2 share a similar modeling pattern to those shown in Table 3.1, with only one exception [146]. In their work, the authors of [146] incorporate an adversarial pre-training step in addition to the discriminative paradigm used by other approaches, which leads to better performance on downstream tasks. However, most models in both correspondence and representation learning employ a discriminative architecture design and do not make use of generative methods.

A comparable tendency towards the spatial and temporal features can be seen in correspondence learning models, with a slight emphasis on the temporal aspect. This parallels the approaches observed in representation learning models, as demonstrated in Table 3.1. Such similarity arises from the requirement for models to account for not only the content of video frames but also the synchronization of inter-frame and inter-modal data.

The recognition task offers an opportunity for researchers to move away from supervised learning, and many recognition models employ self-supervision to learn [16, 197, 32, 27, 94], with only a few heavily relying on labeled datasets [271, 175, 146].

Retrieval in Correspondence Learning

Retrieval is a learning task that goes beyond the recognition of known concepts in the data by incorporating a user-defined query and ranking matching findings accordingly. Initially, retrieval models were developed for a single modality, with either text or image serving as the source of queries [169]. However, with the emergence of multimodal learning, text inputs are now treated equally to other modalities, enabling queries based on images, audio recordings, videos, etc. [75, 77, 140]

The main difficulty in retrieval models is handling the semantic gap between the low-level features of various modalities and the high-level concepts that users intend to find. Furthermore, the effectiveness of the retrieval model may be affected by variations in the user's query, such as different languages, accents, or dialects.

Retrieval models concentrate on developing discriminative features, which is akin to recognition models and follows the same trend in correspondence learning (Table 3.2). However, similarly to recognition models, an alternative approach to the model design also exists. For instance, in [162], the authors introduce an auxiliary label reconstruction task to improve the quality of modality-specific embeddings in the intermediate stage.

On average, multimodal retrieval models include processing of both audio and video, similar to the tasks discussed earlier. However, unlike the models described in Section 3.1.3, there is no consistent balance between the intra- and inter-frame focus. Instead, some models prioritize the temporal aspect [289, 77], while others focus more on the spatial characteristics [184].

Modern retrieval models require significant supervision for training, and there seems to be a trend towards supervised approaches [289, 219, 184, 162, 144, 260, 140], although it was common previously to rely on the use of un- or self-supervised techniques [289, 194, 77, 32].

Detection in Correspondence Learning

Audio-visual detection models are designed to identify and locate specific concepts in a signal, such as a video. In multimodal scenarios, these models often prioritize one modality while using others as supplementary cues. Unlike some other tasks, this one

typically has no limit on the number of objects that can be detected, but usually restricts the set of possible classes to a predefined list.

In the case of multimodal recognition, the complexity and variability of the data are major challenges. This includes issues such as lighting, camera angles, and noise which can have an impact on detection accuracy. Even though detecting objects in such conditions can be difficult, some researchers have taken on the challenge of audio-visual tracking of detected objects, as seen in [1].

The audio-visual detection task does not require the use of the generative design paradigm, which is consistent with the architectural patterns observed in representation learning (Section 3.1.2), as well as recognition and retrieval (Table 3.2) models.

The distribution of attention between spatial and temporal components is not evenly balanced in multimodal detection models, with a greater emphasis placed on the latter, as can be seen by comparing models by modality.

Lastly, it is worth noting that multimodal detection models can be trained using both unsupervised/self-supervised [3, 252, 1] and supervised [170, 271, 244] methods, as demonstrated by several studies.

3.1.4 Audio-visual Object Localization

Audio-visual object localization is a task that involves identifying a specific target within a video and determining its spatial and temporal location, similar to the detection task. However, unlike detection models, localization models usually deal with a single object rather than enumerating all target instances. The need for locating the object makes the task both more difficult and more fine-grained.

Accurately synchronizing the audible and visual modalities is a significant challenge in the field of audio-visual object localization. To address it, researchers rely on a variety of temporal alignment techniques, such as cross-correlation estimation [201] or dynamic time warping [81]. In addition to synchronization, there is a need to account for the different sound propagation properties of various environments, including room acoustics and outdoor spaces, which necessitates sophisticated modeling techniques to capture the complex interactions between the audio signal and the physical environment [268]. Moreover, the presence of multiple sound sources and occlusions in the visual scene can pose additional challenges in accurately localizing the sound source [193].

Localization models for audio-visual object detection often rely on learning discriminative features exclusively, as presented in Table 3.3. Nonetheless, a generative method can also

be advantageous, as shown by the authors of [228]. In their work, the authors improved the model performance on the task by incorporating an acoustic image synthesizer into the network’s architecture.

Tab. 3.3: Models localizing audio-visual objects.

Year	Model	Source	Modeling		Modality		Supervision	
			D	G	S	T	U	S
2017	Audio-visual object localization and separation using low-rank and sparsity	[204]						
2018	Audio-visual event localization in unconstrained videos	[246]						
2019	Identify, locate and separate: Audio-visual object extraction in large video collections using weak supervision	[193]						
2019	Weakly supervised representation learning for audio-visual scene analysis	[192]						
2020	Audiovisual transformer with instance attention for audio-visual event localization	[141]						
2020	Curriculum audiovisual learning	[96]						
2020	Discriminative sounding objects localization via self-supervised audiovisual matching	[95]						
2020	Learning differentiable sparse and low rank networks for audio-visual object localization	[203]						
2020	Self-supervised learning of audio-visual objects from video	[3]						
2020	Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling	[161]						
2021	Audio-visual localization by synthetic acoustic image generation	[228]						
2021	Binaural Audio-Visual Localization	[268]						
2021	Class-aware sounding objects localization via audiovisual correspondence	[97]						
2021	Dual Normalization Multitasking for Audio-Visual Sounding Object Localization	[182]						
2021	Space-Time Memory Network for Sounding Object Localization in Videos	[137]						
2021	Audio-Visual Segmentation	[300]						
2022	Unsupervised Sounding Object Localization with Bottom-Up and Top-Down Attention	[232]						

The trend observed in the modality aspect pattern is consistent with that seen in Section 3.1.2 and Section 3.1.3. Most models, with few exceptions, rely on temporal information to a greater extent than spatial information [300] (Table 3.3).

Audio-visual localization models have been developed using unsupervised, self-supervised, and supervised approaches with similar frequency. In the past, weak supervision during training was more prevalent [204, 193, 192, 141, 95, 203, 3, 161] than regular supervised learning [246, 96]. However, recent advancements in the field rely more on labeled data than on unsupervised learning methods. These developments have been reflected in the more recent literature, which places a greater emphasis on supervised learning [268, 97, 182, 137, 300] than on unsupervised approaches [228, 232].

3.1.5 Sound Source Separation

The sound source separation task is more complex than the previous ones discussed. In this task, a model needs to process a mixed audio signal containing different sources and provide multiple isolated signals that are perfectly clean. This requirement significantly affects the design of the model, which often incorporates generative components in the architecture, as shown in Table 3.4.

The task of sound source separation presents several difficulties, mainly due to the intricate interplay between different sound sources, which can produce overlapping and interdependent signals [54, 59]. To overcome these issues, advanced modeling techniques that can effectively distinguish between sources based on their spectral and

Tab. 3.4: Models separating sound sources.

Year	Model	Source	Modeling		Modality		Supervision	
			D	G	S	T	U	S
2018	Listen and look: Audio-visual matching assisted speech source separation	[154]						
2018	Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation	[54]						
2018	Seeing through noise: Visually driven speaker separation and enhancement	[59]						
2019	Audio-visual deep clustering for speech separation	[153]						
2019	End-to-end sound source separation conditioned on instrument labels	[236]						
2019	Self-supervised audio-visual co-segmentation	[220]						
2019	Time domain audio visual speech separation	[266]						
2019	Two stage audio-video speech separation using multimodal convolutional neural networks	[270]						
2020	Deep audio-visual speech separation with attention mechanism	[135]						
2020	FaceFilter: Audio-visual speech separation using still images	[38]						
2020	Into the wild with AudioScope: Unsupervised audio-visual separation of on-screen sounds	[251]						
2020	Multi-modal multi-channel target speech separation	[72]						
2020	Sep-stereo: Visually guided stereophonic audio generation by associating source separation	[299]						
2021	Deep Audio-Visual Speech Separation Based on Facial Motion	[215]						
2021	Looking into your speech: Learning cross-modal affinity for audio-visual speech separation	[132]						
2021	Move2hear: Active audio-visual source separation	[158]						
2021	Visual scene graphs for audio source separation	[26]						
2021	VisualVoice: Audio-visual speech separation with cross-modal consistency	[63]						
2021	Weakly-supervised audio-visual sound source detection and separation	[208]						
2022	Audio-Visual Based Online Multi-Source Separation	[185]						
2022	Learning Audio-Visual Dynamics Using Scene Graphs for Audio Source Separation	[25]						
2022	Multi-Modal Multi-Correlation Learning for Audio-Visual Speech Separation	[261]						
2022	Time-Domain Audio-Visual Speech Separation on Low Quality Videos	[269]						

temporal properties are required. Moreover, challenges can arise when dealing with diverse sound types, such as speech, music, and environmental noise, which exhibit distinct spectral and temporal characteristics.

In contrast to the models discussed earlier, sound source separation methods heavily rely on generative approaches. This typically involves generating a clean audio signal from a source, although specific implementations may differ, such as generating separation masks [54, 220, 251, 185] or producing an intermediate non-audible representation [236].

Compared to audio-visual object localization models, sound source separation models require more detailed visual information. Therefore, the relative importance of temporal and spatial aspects of a signal shifts towards the latter in this class of models. However, in the past, it was common for model design to focus more on time-related information, as seen in some previous publications [59, 153, 236, 220, 266], but recent works have shown a shift in this trend [185].

The amount of supervision required for different tasks decreases gradually from representation and correspondence learning to object localization and sound source separation [236, 266, 135, 72]. Self-supervised sound separation methods have been dominant in the past [154, 54, 59, 220, 270, 299], but weak supervision is more common in recent research [158, 63, 208, 185, 261, 269] than self-supervised methods [26, 25], with a little exception in the earlier years [153, 38, 251].

3.1.6 Audio-visual Generation

In contrast to the previous sections, where the use of generative models was limited or unnecessary, the current task requires the development of models that follow the generative architectural paradigm, as shown in Table 3.5. This approach typically involves employing an “encoder – decoder” model setup and adversarial training techniques, which will be explored in detail in this section.

Tab. 3.5: Models generating (spatio-)temporal signal.

Year	Model	Source	Modeling		Modality		Supervision	
			D	G	S	T	U	S
Sound generation								
2018	Lip2AudSpec: Speech reconstruction from silent lip movements video	[6]						
2018	MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment	[48]						
2018	Self-supervised generation of spatial audio for 360 video	[172]						
2018	Visual to sound: Generating natural sound for videos in the wild	[301]						
2018	Visually indicated sound generation by perceptually optimized classification	[28]						
2019	DurIAN: Duration informed attention network for multimodal synthesis	[286]						
2020	Audeo: Audio generation for a silent performance video	[238]						
2020	Generating visually aligned sound from videos	[30]						
2021	FoleyGAN: Visually Guided Generative Adversarial Network-Based Synchronous Sound Generation in Silent Videos	[202]						
2021	How Does it Sound? Generation of Rhythmic Soundtracks for Human Movement Videos	[239]						
2021	Taming visually guided sound generation	[98]						
2022	Audio-Visual Speech Codecs: Rethinking Audio-Visual Speech Enhancement by Re-Synthesis	[281]						
2022	VarietySound: Timbre-Controllable Video to Sound Generation via Unsupervised Information Disentanglement	[41]						
Video generation								
2019	Talking face generation by adversarially disentangled audio-visual representation	[298]						
2019	You said that?: Synthesising talking faces from audio	[101]						
2020	Audio-driven talking face video generation with learning-based personalized head pose	[284]						
2020	Multi Modal Adaptive Normalization for Audio to Video Generation	[127]						
2020	Photorealistic audio-driven video portraits	[264]						
2020	Robust one shot audio to video generation	[126]						
2021	AD-NeRF: Audio driven neural radiance fields for talking head synthesis	[76]						
2021	Audio-driven emotional video portraits	[104]						
2021	Let's Play Music: Audio-Driven Performance Video Generation	[303]						
2022	Audio-driven talking face video generation with dynamic convolution kernels	[283]						
2022	Audio-visual face reenactment	[4]						
2022	Semantic-aware implicit neural audio-driven video portrait generation	[145]						
2022	Sound-Guided Semantic Video Generation	[133]						
2022	StyleTalker: One-shot Style-based Audio-driven Talking Head Video Generation	[167]						

The use of the audible modality enables producing more realistic and immersive outputs. To illustrate, by modeling distinct characteristics of the audio signal such as pitch, duration, and timbre, sound generation models generate more expressive and controllable sounds [41]. Likewise, joint modeling of the audible and visual modalities in video generation models allows for producing video content that is semantically meaningful and interpretable [238].

This section focuses on two specific types of generative models: sound and video generation. These tasks represent a shift in learning paradigms from purely discriminative architectures (as discussed in Section 3.1.2) to those that involve synthesis (as discussed in Section 3.1.5), and eventually to temporal (as discussed in Section 3.1.6) and spatio-temporal generative models (as discussed in Section 3.1.6).

Sound Generation

Sound generation in a multimodal setup pertains to the process of generating audio based on a visual input that varies over time, such as video. These models generate synthetic signals by conditioning on both visual and textual inputs. Unlike sound source separation models, which aim to reconstruct a clean signal from a noisy one, sound generation models generate sound in a probabilistic way.

Generating high-quality audio representing diverse and naturally sounding scenes that encompass various categories, including speech, music, and environmental sounds, requires the development of specialized models that can capture the unique features of each sound type. Finally, evaluating the quality of generated audio can be a difficult task, as subjective measures such as naturalness, clarity, and fidelity are challenging to quantify and standardize.

There are two primary design patterns that sound generation models follow: the “encoder – decoder” approach and GAN. In recent years, the popularity of the GAN approach has been on the rise among researchers [28, 48, 30, 202, 98, 41], but non-adversarial methods continue to be of use as well [6, 172, 301, 286, 281]. In addition to generating audible sounds, there is also an alternative method that involves generating a non-audible intermediate representation suitable for music generation [238, 239], similar to the approach used in sound source separation models described earlier in Section 3.1.5.

Regarding the modality perspective, sound generation models usually prioritize temporal aspects over spatial ones, although there are some rare cases where the visual component is more critical [286]. However, there has been a recent increase in focus on time-domain characteristics in sound generation models [238, 239, 41].

It is apparent from Table 3.5 that the level of supervision in the training process is consistently minimal for sound generation models and appears to fluctuate slightly depending on the modeling approach.

Video Generation

Video generation models are designed to solve a rather new task that has emerged from the fusion of sound and image generation tasks. Several recent models have been published that generate synthetic video with a particular focus on realistic face rendering, which is achieved by conditioning the model on speech [283, 4, 145, 167].

Capturing non-trivial temporal dynamics of the video sequence is one of the major obstacles in video generation. This necessitates the use of sophisticated architectures that can account for long-term dependencies and create plausible motion patterns. Furthermore, generating coherent and consistent video sequences may be difficult owing to variations in the visual appearance and context of the scene [256].

Video generation models, like sound generation models, are mainly represented by GANs, which have become dominant in the field (Table 3.5). Unlike audio-centric architectures, video generation models seldom use a pure “encoder – decoder” architecture [101, 104]. Nevertheless, there is no clear distinction between the applicability of the two paradigms, as demonstrated by the authors of [167], who use both architectural patterns within the same model.

Regarding the modality, video generation models typically strive to balance the spatial and temporal aspects of the signal [298, 101, 284, 127, 76, 303, 133, 167, 283]. However, there is a recent increase in interest within the research community for high-quality, particularly visual, generated videos [264, 126, 104, 4, 145].

The level of supervision during training in video generation models is consistent with that of sound generation models, despite the different dimensionality of the signals they aim to generate. This similarity can be attributed to the fact that both types of models tend to follow similar architectural patterns, leading to comparable levels of required supervision during training.

3.2 Optimization Objectives for Audio-visual Learning

The landscape of audio-visual learning objectives includes various tasks that differ in their complexity and abstractness from low-level representation learning to high-level audio-visual generation, while in between there are the others discussed so far. Altogether, they build a spectrum of challenges that reflect the diverse facets of multimodal machine learning.

Among these tasks, representation learning and correspondence learning are the most basic concepts that are reflected in other, more specific ones. By automatically learning structures within raw data, representation learning enables more abstract tasks to operate on informative and compact representations, thus building the foundation for many machine learning setups. Correspondence learning, in turn, while being associated with specific tasks such as cross-modal retrieval, plays a fundamental role in aligning and integrating information across different domains or modalities. By establishing

correspondences between data points, it enables exchange and alignment of information, facilitating tasks that require interactions between signals of different kinds.

However, temporal variability and complex relationships between frequency, time, and semantics pose difficult challenges in the audio-centric tasks, and addressing them is easier on a well-defined, concrete task, in comparison to a more abstract or complex problem. A narrower problem scope allows for concentrating the efforts on addressing the specific challenges associated with this task, as a clear understanding of the requirements allows for developing specialized models that better fit to the data and address the challenges encountered. Takeaways gained from overcoming these challenges and leveraged in more general settings contribute to the advancement of the broader field of deep learning.

3.2.1 Audio-visual Multilabel Classification

Audio-visual multilabel classification is a task that involves the simultaneous assignment of multiple labels to audio-visual data. In this task, the objective is to recognize and categorize audio and visual components within a given input, enabling comprehensive understanding and interpretation. By associating multiple labels with the audio and visual elements, the model aims to capture the rich semantics and diverse characteristics present in the data.

One way to view audio-visual multilabel classification is as a form of correspondence learning, where the model focuses on establishing relationships and alignments between different modalities, such as audio and visual data. In audio-visual multilabel classification, a neural network aims to learn the correspondence between the audible and visual components of the input to assign appropriate labels to each modality. By capturing semantic similarities, it leverages the inherent associations between audio and visual information, enhancing the network's ability to recognize and classify diverse audio-visual patterns.

The task of audio-visual multilabel classification poses several challenges, one of which is the multilevel structure within the target categories. Often, the labels in this task are organized in a hierarchical manner, where some labels may be subclasses of others. For instance, in a taxonomy of human-related sounds [65], the label “Singing” may be hierarchically dependent on the label “Music.” Such a structure introduces complexities in the classification process, as the prediction of one label can influence the prediction of its parent or child labels. Identifying and modeling these hierarchical dependencies accurately becomes a key requirement to ensure coherent and consistent label assignments throughout the classification process [103].

It is important to note that audio-visual multilabel classification can be viewed as both a high-level task, such as multimodal retrieval, as well as a more specific task, such as environmental sound classification. In the context of multimodal retrieval, audio-visual multilabel classification serves as a foundational step, allowing for effective retrieval and matching of multimedia content based on their associated labels. On the other hand, in the realm of environmental sound classification, audio-visual multilabel classification narrows its focus specifically on recognizing and categorizing sounds within the environmental context. This varying formulation of the task emphasizes its versatility and applicability in different domains, highlighting its significance in both broader multimodal scenarios and specific sound classification contexts.

3.2.2 Environmental Sound Classification

Environmental sound classification (ESC) is concerned with correctly differentiating between sound classes that one experiences in the everyday environment (e.g., “baby crying”, “car honking”, “children playing”, “dog barking”, “siren”, “snoring”, “street music” [200]). This means that the models need to be able to handle a wide range of acoustic environments, including variations in background noise, acoustic distortion, and other factors that can affect the quality of the sound recordings. The environmental sound classification task, being a highly specific and well-defined formulation of the broader audio classification task, serves as an perfect choice for addressing challenges specific to the realm of audio-centric analysis, as environmental sounds often contain complex and diverse spectrotemporal patterns that can be difficult to capture with standard feature extraction methods [230]. Therefore, environmental sound classification requires sophisticated feature extraction techniques and powerful classification models that are able to identify patterns in the data that are not immediately apparent to the human ear [199], which makes it a good fit for advancing audio-dedicated models.

Although ESC has many potential application areas, one of the most obvious ones is multimedia retrieval, in which ESC can be employed to improve performance of video retrieval systems by making better use of the audible modality [106]. Another application area is the automated analysis of urban sounds, for example to offer more detailed insights for high noise levels [210].

While environmental sound classification is a comparably young field, a lot of progress has been made after significant datasets such as the ESC-50 [200] and UrbanSound8K (US8K) [227] found wide acceptance in the community. Initially, the general trend in the ESC community was to design audio-domain-specific architectures and combine them with specially engineered features [230]. On one side, this approach makes more

difficult to benefit from advances made in other fields, such as the computer vision community. On the other, this scenario sparked an interest to investigate how well a current state-of-the-art approach from the image domain would perform on ESC [79].

In the last years, however, the focus has shifted to the use of common techniques from other domains, such as the visual one, while both directions are being combined usually with either a raw signal or a pre-computed time-frequency transformation, which is more common. Section 5.2 provides a more detailed overview of the evolution of audio-centric deep learning models. Specifically, one can observe that learning of a time-frequency transformation in an end-to-end fashion is a rare exception that, although it provides an increase of accuracy [224]. Also, the usage of a weight initialization obtained on large-scale datasets is in alignment with recent trends. Such a weight transfer is performed usually in either cross- or intra-domain manner only. Thus, the field of ESC was lacking studies on the assessment of effects of a two-stage domain adaptation using large-scale audio datasets. Besides that, a typical accuracy evaluation of models is being accomplished in “ideal” conditions. Measuring of the influence of a perturbed signal on the predictions of best performing models is not quite usual.

Although the environmental sound classification task is atomic in its formulation, focusing specifically on the recognition and categorization of sounds within specific contexts, the insights and results obtained from this task can be generalized and applied to less constrained sound-related tasks. By understanding the nuances of environmental sound classification, such as capturing nonstationary sound characteristics and addressing temporal dynamics, one can develop methodologies and models that effectively handle more difficult and complex audio-centric scenarios, expanding the applicability and impact of the findings.

3.2.3 Multimodal Retrieval

Multimodal retrieval is a task that involves retrieving and matching multimedia content across different modalities, such as text, images, and audio. The goal of multimodal retrieval is to establish associations and identify relevant instances within a multimodal database based on a query from one or more modalities. This task enables efficient searching and retrieval of information, leveraging the complementary nature of different modalities to enhance the retrieval process [75].

One way to view multimodal retrieval is as a more general kind of classification. In classification tasks, the goal is to assign labels or categories to instances based on their features. In multimodal retrieval, instead of assigning a single label, the objective is to

match and retrieve instances that are relevant to a given query from one modality or a combination of those. By considering multiple modalities simultaneously, multimodal retrieval expands upon the classification paradigm and allows for more comprehensive information retrieval and matching across different representation aspects.

Among others, the task of multimodal retrieval poses a challenge of integrating information from multiple modalities in a meaningful and effective manner, which is essential for obtaining good retrieval results [194]. Combining features and representations from different modalities, such as text, images, and audio, while capturing the relevant information and preserving the discriminative power of each modality, is a non-trivial task. Selecting appropriate fusion techniques, such as early fusion, late fusion, or cross-modal embeddings, plays a crucial role in achieving optimal retrieval performance [175].

Another challenge in multimodal retrieval is temporal alignment [260]. Different modalities are often scaled differently so that they need to be aligned to ensure consistent associations [144]. For example, in video retrieval, the alignment of audio and visual frames is necessary to capture the correspondence between the audible and visual content [140]. Aligning different modalities in terms of temporal dynamics and synchronizing their representations is essential for accurate and coherent matching during the retrieval process.

Overcoming challenges in multimodal fusion and temporal alignment requires advanced techniques and high-performance models that can effectively exploit the complementary information across modalities. Cutting edge deep learning methods such as attention mechanisms [144], cross-modal embeddings [194, 162], and contrastive learning techniques [32] have contributed to addressing these challenges and improving the performance of multimodal retrieval systems. By addressing the complexities of multimodal fusion and temporal alignment, one can leverage the power of modern deep learning architectures enabling the models to expand their capabilities to new modalities as well as improving the retrieval performance on the already known ones [77].

3.2.4 Zero-shot Sound Classification

Zero-shot sound classification is a task that involves classifying sound data into predefined categories, even when the model has not been exposed to examples from those categories during training [129]. In this task, the goal is to generalize the learned representations and knowledge to accurately classify sounds from novel classes. The key idea behind zero-shot sound classification is to leverage semantic representations and knowledge transfer to bridge the gap between seen and unseen concepts [272].

Zero-shot sound classification can be viewed as an intermediate step between multimodal retrieval and regular environmental sound classification. While multimodal retrieval focuses on matching and retrieving instances across modalities, zero-shot sound classification takes a step further by assigning categorical labels to sound data directly [272]. It capitalizes on the semantic associations and shared information between known and unknown classes [273], allowing the model to generalize its understanding of sound across a broader range of classes.

The task of zero-shot sound classification poses several challenges, with one of the main difficulties being the extension to new unseen categories. Unlike regular sound classification, where the model is trained on a specific set of known classes, zero-shot sound classification requires the model to make predictions for labels that were never encountered during training. This necessitates the utilization of external information sources, such as semantic embeddings or class descriptions [273], to facilitate the generalization to new classes. Incorporating and aligning these external sources of information in a form of the knowledge distillation “teacher – student” setup adds an additional layer of complexity, which, in turn, pays off in the extension of model capabilities [77].

3.2.5 Foundations for Experimental Exploration

The exploration of the vast landscape of audio-visual learning has provided invaluable insights into the fundamental tasks that underpin this field. Through careful analysis, there has been identified a number of basic problems that form building blocks of audio-centric deep learning, recognition, classification, and retrieval of audio and visual information.

To delve deeper into the intricacies of audio-centric deep learning, this thesis focuses on these problems and articulates them in a form of specific tasks that encompass audio-visual multilabel classification, environmental sound classification, multimodal retrieval, and zero-shot sound classification, thus representing distinct facets of audio-centric analysis. They serve as pivotal points of exploration, each posing unique challenges and offering particular insights into the field.

In addition, the four selected tasks form a hierarchy, with each one offering a more specialized perspective while building upon the knowledge acquired from its predecessor. Starting with audio-visual multilabel classification, which involves assigning multiple labels to audio-visual data, the hierarchy progresses to environmental sound classification, focusing on the recognition and categorization of sounds in specific contexts. Taking step

towards a more high-level setup, multimodal retrieval broadens the scope, encompassing the matching and retrieval of multimedia content across modalities. Finally, zero-shot sound classification expands the boundaries by exploring the ability to classify sounds from novel or unseen classes.

These observations serve as a foundation for the next chapter, where, by formulating hypotheses and proposing design of experiments, this thesis builds a set of methods to overcome the challenges identified within this hierarchy of tasks. Subsequent experimental evaluation of the proposed methods allows for establishing audio-centric approaches that advance multimodal deep neural network architectures and pave the way for improved performance and understanding. Through this iterative process of hypothesis formulation, experimentation, and validation, this thesis contributes to the evolution of the field of artificial intelligence by pushing the boundaries of audio-visual deep learning and addressing the exciting challenges presented by each task.

Methods and Applications

This chapter delves into the core techniques and practical applications employed in this thesis. It serves the purpose of providing a comprehensive understanding of the research objectives that arise from the questions of the thesis and investigates the applied methods in detail, emphasizing the technical intricacies involved.

Specifically, each of the learning objectives is described through the prism of associated challenges and equipped with a number of hypotheses that aim to overcome the identified issues. By thoroughly discussing these objectives and the corresponding approaches, this chapter builds a holistic picture of technical aspects related to the practical implementations arising in the context of audio analysis for multimodal learning. This leads to establishing connections between the tasks and approaches to introduce useful modifications and forms a foundation for the subsequent experimental evaluation of the proposed hypotheses, as discussed later in Chapter 5.

4.1 Task Analysis

This section delves into a detailed analysis of audio-centric tasks, namely audio-visual multimodal classification, environmental sound classification, multimodal retrieval, and zero-shot sound classification. This analysis allows for identifying the key challenges associated with each task and formulating hypotheses on how to overcome these challenges. By carefully examining each task, underlying factors that contribute to their difficulty are identified and strategies to address them proposed. Through this rigorous exploration, the thesis provides insights into the specific challenges faced in audio-centric tasks and proposes solutions to enhance the performance and capabilities of models in these domains.

4.1.1 Audio-visual Multilabel Classification

Audio-visual multilabel classification poses a challenging hierarchy of target categories, where labels are organized in a multilevel manner, with certain labels being hierarchically dependent on others, as represented by the AudioSet [65] dataset. This hierarchical structure introduces complexities in the classification process, as the accurate prediction of one label may rely on the correct prediction of its parent or child labels. Capturing and modeling these dependencies accurately becomes essential for obtaining high-performance models. To enhance deep learning models solving the audio-visual multilabel classification task, changes can be introduced into both the training setup and the model architecture.

$\mathcal{H}^{(1.1)}$: The use of transfer learning from the visual domain improves model performance in audio-centric multilabel classification.

One approach to improve the performance of models in audio-centric multilabel classification lies in leveraging transfer learning from the visual domain. Transfer learning involves utilizing knowledge learned from a source domain, such as image classification on the ImageNet [45] dataset, and applying it to a target domain, in this case, audio classification, as described in Section 2.1.8. By fine-tuning the pre-trained model on audio data, one can benefit from the knowledge and representations learned from the image data, which, in turn, leads to the significant improvement of performance in audio classification.

$\mathcal{H}^{(1.2)}$: The use of trainable time-frequency transformation of audio improves model performance in audio-centric multilabel classification.

In audio-centric multilabel classification, one key aspect that greatly influences model performance is the representation of audio data. Traditionally, audio signals are transformed into a time-frequency representation, usually regular or Mel-scaled spectrograms [151], which capture both the temporal and spectral characteristics of the audio. However, these transformations are typically fixed and do not adapt to the specific characteristics of the given dataset. To address this limitation, this thesis introduces a trainable time-frequency transformation of audio. It involves fitting this mapping directly from the audio data within the neural network. By making the time-frequency transformation trainable, the model leverages representations that align with the specific audio data and the task at hand.

$\mathcal{H}^{(1.3)}$: The use of an advanced backbone architecture improves model performance in environmental sound classification.

In environmental sound classification, the choice of the underlying backbone architecture plays a crucial role in determining the model's ability to learn discriminative features present in the audio data. While traditional backbone architectures like ResNet [85] have been widely used and proven effective in various computer vision tasks, leveraging more advanced architectures such as ResNeXt [274] is a way to further enhance the performance of models in environmental sound classification. By exploiting grouped convolutions, ResNeXt is able to capture a richer set of features by simultaneously considering multiple paths of the information flow [274]. This enhanced modeling capability is particularly beneficial for handling complex and diverse patterns within the data.

$\mathcal{H}^{(1.4)}$: The use of knowledge-distillation-based correspondence learning improves model performance in audio-visual multilabel classification.

Knowledge distillation involves transferring knowledge from a large, well-performing model (teacher) to a smaller model (student), as described in more detail in Section 2.1.5. In the context of audio-visual multilabel classification, the teacher model possesses a deeper understanding of the correspondence between modalities, effectively capturing the intricate dependencies and correlations. By distilling the knowledge, the student model can learn from the teacher's insights and exploit the existing relationships between audio and visual features. This enables the student model to better capture and leverage the multimodal cues present in the audio and visual data, resulting in improved performance in multilabel classification [147].

$\mathcal{H}^{(1.5)}$: The full model fine-tuning in the knowledge-distillation-based correspondence learning setup improves model performance in audio-visual multilabel classification.

By the joint fine-tuning of both parts of the knowledge distillation setup, the correspondence learning process becomes more comprehensive and mutually beneficial. Specifically, the full model fine-tuning allows both the teacher and student models to adapt and refine their representations and predictions by exchanging the knowledge. The teacher model, with its deep knowledge and insights, can guide the student model to better capture the important multimodal cues and dependencies, while the student's ability to learn from the teacher's knowledge can provide additional perspectives and refinements to the teacher's understanding [116].

4.1.2 Environmental Sound Classification

In the task of environmental sound classification, model overfitting and the lack of specialized architectures present significant challenges [79]. Overfitting occurs when the model becomes overly tailored to the training data and fails to generalize well to unseen examples [263]. Given the diversity and complexity of environmental sounds, it is essential to address this issue to ensure robust performance. Additionally, the absence of specialized architectures designed for ESC may limit the model's ability to capture the distinctive features and patterns inherent in this domain. To attenuate the influence of these challenges, both the training setup and the model architecture need to be modified, thus resulting in better generalization ability and higher classification accuracy.

$\mathcal{H}^{(2.1)}$: The use of transfer learning from the visual domain improves model performance in environmental sound classification.

When solving the environmental sound classification task, one can often notice the limited amount of labeled data in the absence of audio-centric models specifically designed for the task. By utilizing transfer learning from the visual domain, where large-scale labeled datasets and powerful models are readily available, one can tap into the wealth of knowledge learned from visual tasks and apply it to the downstream ESC task, represented by the ESC-50/-10 [200] and UrbanSound8K [227] datasets.

$\mathcal{H}^{(2.2)}$: The use of an advanced backbone architecture improves model performance in environmental sound classification.

In environmental sound classification, the choice of the underlying backbone architecture determines the model's ability to represent descriptive audio features. By utilizing ResNeXt [274] as a backbone architecture, the model gains an increased capacity, leading to improved discriminative power. The better expressiveness of ResNeXt enables the network to model nonlinear relationships between audio features, thus expanding its capability to distinguish between different sound classes [78].

$\mathcal{H}^{(2.3)}$: The use of knowledge-distillation-based correspondence learning improves model performance in environmental sound classification.

In the context of ESC, knowledge distillation-based correspondence learning enables the student to leverage the rich insights and learned representations of the teacher, which has a broader and deeper understanding of the audio-visual data. Being guided by this knowledge, the student model benefits from the teacher's expertise and enhances its own ability in capturing discriminative audio features [77].

$\mathcal{H}^{(2.4)}$: The use of more than one audio channel improves model performance in environmental sound classification.

By leveraging multiple audio streams, such as in stereo or multi-channel audio, the model can better exploit the temporal characteristics of the sound sources, which results in improved classification performance. As each audio channel may capture unique temporal patterns and dynamics, their joint use enriches the model's understanding of the sound events and facilitates extraction of temporal features from different perspectives. By simultaneously considering the temporal information from multiple channels, the network exploits variations in channel-wise sound characteristics, which leads to improved discriminative power and more accurate classification [79].

$\mathcal{H}^{(2.5)}$: The use of spectrotemporal attention improves model performance in environmental sound classification.

The ESC task requires models to effectively learn time-frequency patterns present in the audio data. Adding spectrotemporal attention layers into the network architecture allows the model for gaining a capability to dynamically allocate its attention to specific regions in the input spectrogram and low-level feature maps, thus exploiting more informative parts of the signal. Precisely, this attention mechanism supports the model by selectively focusing on relevant frequency and time segments, thus effectively highlighting features that significantly contribute to producing a correct decision [79].

$\mathcal{H}^{(2.6)}$: The use of the intermediate pre-training in the audio-domain improves model performance in environmental sound classification.

When tackling environmental sound classification, leveraging transfer learning from the visual domain provides valuable insights and features learned from abundant visual data. However, to further optimize model performance specifically for audio tasks, employing intermediate pre-training in the audio domain after transfer learning is a way to improve results. After initially transferring knowledge from the visual domain, the model's representation may still be biased towards visual patterns rather than audio-specific characteristics [47]. By introducing an intermediate pre-training in the audio domain, the model undergoes additional fitting to the audio data, allowing it for adapting and refining learned representations to better align with the target domain [78].

$\mathcal{H}^{(2.7)}$: The use of trainable time-frequency transformation of audio improves model performance in environmental sound classification.

By introducing trainable time-frequency transformations, the model gains flexibility to adapt and learn representations that are specifically optimized for the ESC task.

Instead of relying on the STFT that fixed in both temporal and frequency resolution, such transformations enable the network fitting to spectrotemporal patterns and variations by learning filters—or kernels—that are tailored to the specific dataset. Such adaptability allows the network for processing fine-grained details, distinguishing subtle variations, and focusing on the most informative aspects of the audio signals, which leads to improved classification performance [78].

$\mathcal{H}^{(2.8)}$: The use of trainable time-frequency transformation of audio improves model robustness against additive white Gaussian noise.

Learnable time-frequency transformations allow the model for adapting and learning representations that are specifically tailored to the training data. By selectively amplifying or attenuating certain frequency bands, the trainable transformation layer effectively filters out undesirable components and enhances the Signal-to-Noise Ratio (SNR). This helps the network to robustly recognize underlying audio patterns despite the presence of noise [78].

$\mathcal{H}^{(2.9)}$: The use of trainable time-frequency transformation of audio improves model robustness against sample rate reduction.

When dealing with environmental sound classification, the sample rate of audio signals can vary due to external conditions such as aggressive compression or bandwidth limitations, and the reduction of the sample rate leads to the model performance degradation. However, extending the network architecture with a trainable time-frequency transformation layer introduces a certain level of robustness against sample rate reduction. Adaptability of the model to the dataset distribution enables it to handle sample rate reduction more effectively, mitigating the potential loss of information caused by signal disturbances due to downsampling [78].

4.1.3 Multimodal Retrieval

The field of multimodal retrieval has seen substantial research focused on extracting visual information, but there has been relatively less emphasis on audio-centric aspects. This lack of interest has resulted in a limited understanding of challenges and potential solutions specific to audio in the context of multimodal retrieval. However, recent advancements in audio-centric model design, visual contrastive learning, and knowledge-distillation-based correspondence learning offer a promising avenue to bridge this gap and enable audio-centric capabilities in multimodal retrieval, as was discussed earlier in Section 3.1.3.

$\mathcal{H}^{(3.1)}$: The use of knowledge-distillation-based correspondence learning enables audio-centric capabilities in multimodal retrieval.

The task of multimodal retrieval poses cross-modal alignment challenges [75], as incorporating visual information alongside visual data is complicated due to the inherent differences between modalities, as discussed in Section 3.1.3. However, audio-centric capabilities can be effectively integrated into the multimodal retrieval framework by leveraging knowledge-distillation-based correspondence learning [157]. The knowledge distillation approach helps establish meaningful connections between audio and visual modalities by aligning their representations in a shared embedding space [32]. By distilling the non-audio knowledge, the audio-centric model can better capture the semantic relationships and associations between audio and visual features, thus bringing the audio-specific capabilities into the multimodal retrieval framework [77].

$\mathcal{H}^{(3.2)}$: The full model fine-tuning in the knowledge-distillation-based correspondence learning setup improves model performance in multimodal retrieval.

Model performance in a multimodal setup benefits from both unidirectional [32] and mutual [116]—i.e. both teacher and student exchange the knowledge—distillation. The teacher model, which has been pre-trained on large-scale text-visual data, possesses a strong understanding of the semantic relationships between text and visual features [148]. This knowledge is then distilled into the student model, which specializes in audio analysis. The full model fine-tuning step further refines the student by adjusting its parameters using the knowledge distilled from the teacher model, and the other way around, thus enhancing the network’s retrieval capabilities and enabling it to effectively exploit the complementary nature of audio and visual information in multimodal retrieval scenarios [77].

4.1.4 Zero-shot Sound Classification

The field of zero-shot sound classification has received relatively limited attention compared to other audio-centric tasks, resulting in a dearth of research and understanding in this area. However, the integration of audio-specific knowledge into the zero-shot classification framework helps the network fit to characteristics of audio [117], which, in turn, empowers the model to generalize well to new, unseen classes and make accurate predictions. Hence, application of knowledge-distillation-based learning methods provides a reliable basis for advancing the state-of-the-art in audio-centric zero-shot learning tasks [249, 165].

$\mathcal{H}^{(4.1)}$: The use of knowledge-distillation-based correspondence learning enables model’s zero-shot sound classification capabilities.

Zero-shot sound classification is a challenging task where a model is required to recognize and classify samples belonging to categories that it has not been explicitly trained on [190]. By simultaneously leveraging visual, textual, and audible representations obtained from AudioSet [65], the knowledge-distillation-based training enables the student (that focuses on audio) to learn semantics of the multimodal data under the joint cross-modal supervision [77]. Hence, the audio-dedicated network obtains capabilities to classify sound samples according to previously undefined class labels.

$\mathcal{H}^{(4.2)}$: The full model fine-tuning in the knowledge-distillation-based correspondence learning setup improves model performance in zero-shot sound classification.

Effectively tackling the challenges of zero-shot sound classification involves a comprehensive fine-tuning process. Joint adaptation of the teacher and student model through mutual knowledge distillation allows the network parameters to be adjusted so that both cross-modal semantics [148] and audio-specific characteristics of acoustic classes are taken into account [77]. Similarly to multimodal retrieval, this synergistic approach enhances the student’s ability to generalize to unseen classes, ensuring accurate predictions in the zero-shot classification setup, as further presented in Section 5.4.

Experimental Essentials

In the preceding sections, inherent challenges for audio-centric tasks have been outlined and discussed. Recognizing them is crucial in establishing strategies to overcome these limitations and to improve the performance of models in audio-centric tasks. By thoroughly examining these challenges, prerequisites have been identified, which are essential for the experimental design and downstream analysis of the experimental results. As the formulated hypotheses $\mathcal{H}^{(1.1)} - \mathcal{H}^{(4.2)}$ build a basis for a systematic evaluation of the enhancement methods for audio-centric multimodal models, there is a natural need for verifying these conjectures in a series of experiments. The next section gathers the requirements on the experimental design and architectural modifications and discusses them in more detail.

4.2 Experimental Design

The task analysis allowed to discover the components that are necessary to address the audio-centric challenges related to the audio-visual learning objectives. At a high-level, these components can be split into the data sources, training workflow modifications, and model architecture enhancements. Since the data is the most basic component in any machine learning setup, first, audio- and visual-centric datasets were identified, those list includes ImageNet [45], AudioSet [65], ESC-50/-10 [200], and UrbanSound8K [227].

Adjustment of the training workflow can be performed in numerous ways, however, this thesis focuses on two specific setups: transfer learning from the visual domain and multimodal correspondence learning through knowledge distillation. Also, the architectural enhancements can be categorized into main and supporting ones. The auxiliary actions that include *static* time-frequency representation of audio and adjustment of such input to a specific training workflow are considered to be necessary, although they do not contribute to the model advancement directly. In contrast, incorporation of spectrotemporal attention, multichannel audio processing, and *trainable* time-frequency representation provides a valuable increase in performance, thus these changes attributed to main enhancements.

4.2.1 Datasets

This section introduces datasets relevant to the tasks of audio-visual multimodal classification, regular and zero-shot environmental sound classification, and multimodal retrieval. These datasets have been widely used in the deep learning community for training and evaluating models for various audio-related tasks [199, 226, 88, 64]. A brief overview of each is provided, including size, class distribution, and notable features. Understanding properties of these datasets is required for developing effective deep learning models that can perform well in the aforementioned scenarios. The following sections delve deeper into the dataset details and discuss how they can be employed for specific tasks.

ImageNet

ImageNet is a large-scale visual recognition dataset that has been widely used for training and evaluating deep learning models for image classification tasks [45]. It contains more than 1 million images, with each image being labeled with one of 1000

possible object categories. The images in the dataset cover a wide range of visual concepts, such as animals, plants, vehicles, and common household objects. The dataset has become a benchmark for evaluating the performance of deep learning models for image classification, and it has been instrumental in the development of advanced deep learning architectures such as ResNet [85] and Inception [242].

One of the key characteristics of ImageNet is its large size, which allows deep learning models for effectively learning discriminative visual features from a massive amount of training data, despite a slight class imbalance. The dataset is also highly diverse, containing images captured from a wide range of sources, including professional photography, social media, and user-generated content [45]. As a result, ImageNet has become a critical resource for the computer vision community and has contributed significantly to the development of state-of-the-art deep neural networks for image classification. This thesis investigates effects of transfer learning using ImageNet-initialized models, as discussed later in this chapter and in Chapter 5.

AudioSet

AudioSet [65] was created and released in 2017 by Google Research. The dataset was designed to facilitate development of machine learning models for various audio-related tasks such as sound event detection [233], sound classification [122], and audio tagging [123]. The amount of training data allows for using AudioSet as an initialization basis for audio-centric deep neural networks, thus providing a better starting point for fine-tuning to more fine-grained sound classification tasks on the ESC-50 (Section 4.2.1) and UrbanSound8K (Section 4.2.1) datasets.

It consists of ~ 1.8 M video clips, each 10 s long and labeled with one or more diverse sound event classes out of 527. AudioSet categories build a hierarchy, which, if disregarded, can aggravate the issue of class imbalance present in the labels, necessitating the use of a weighted loss function. The labeled data covers a wide range of everyday sounds and includes but not limited to human actions, animal sounds, and natural phenomena, as well as more specific sounds like musical instruments and transportation noises [65].

ESC-50/-10

The ESC-50 [200] dataset consists of 2000 monaural samples belonging to 50 classes that can be divided into 5 groups, such as *animal* sounds, *natural and water* sounds, *non-speech human* sounds, *interior* and *exterior* sounds. Samples are distributed equally

among classes, and thus each category consists of 40 recordings. Each track has a length of 5 seconds, the native sample rate is 44.1 kHz. The dataset was divided by the author into 5 folds that were used in this thesis to perform the performance evaluation in environmental sound classification, retrieval, and zero-shot ESC. The ESC-10 dataset is a subset of the ESC-50 dataset. It consists of only 10 classes that are restricted to the following categories: *transient/percussive* sounds with temporal patterns, sounds with strong *harmonic* content, and *noise/soundscapes*. All other characteristics of the ESC-10 dataset are equal to those of the ESC-50 dataset.

UrbanSound8K

The UrbanSound8K [227] dataset consists of 8732 samples (both mono and stereo) belonging to 10 classes: “air conditioner”, “car horn”, “children playing”, “dog bark”, “drilling”, “engine idling”, “gun shot”, “jackhammer”, “siren”, and “street music”. The classes are not balanced in terms of overall recording lengths per class. Each track has a variable length of up to 4 seconds, the native sample rate varies from 16 kHz to 48 kHz. The dataset was divided by its authors into 10 folds that were used in this thesis to perform the performance evaluation in environmental sound classification, retrieval, and zero-shot ESC.

4.2.2 Transfer Learning from the Visual Domain

This training workflow modification appears to be basic yet powerful method to boost the model performance. Although it can be applied in both intra- and interdomain manner, one of the important limitations of this approach corresponds to the binding to a specific model (e.g., ResNet-50 [85] or ViT-B32 [50]). The typical experimental workflow that involves testing of the influence of the transfer learning from the visual domain assumes the performance comparison of at least two variants of a same model: with and without pretraining on the visual domain data. Figure 4.1 provides a high-level architecture overview for the ESResNe(X)t model [79, 78] that inherently supports such experimental setup.

Using visual-domain-models for solving audio-specific tasks provides a way to leverage the vast amount of labeled data available in the visual domain, which helps in solving the lack of labeled data problem in the audio domain. This can be particularly useful in the case of audio-visual multilabel classification, where the model needs to classify an input sample into multiple classes using both visual and audio information. Also, visual-domain-models can capture semantic information that is relevant to both visual and

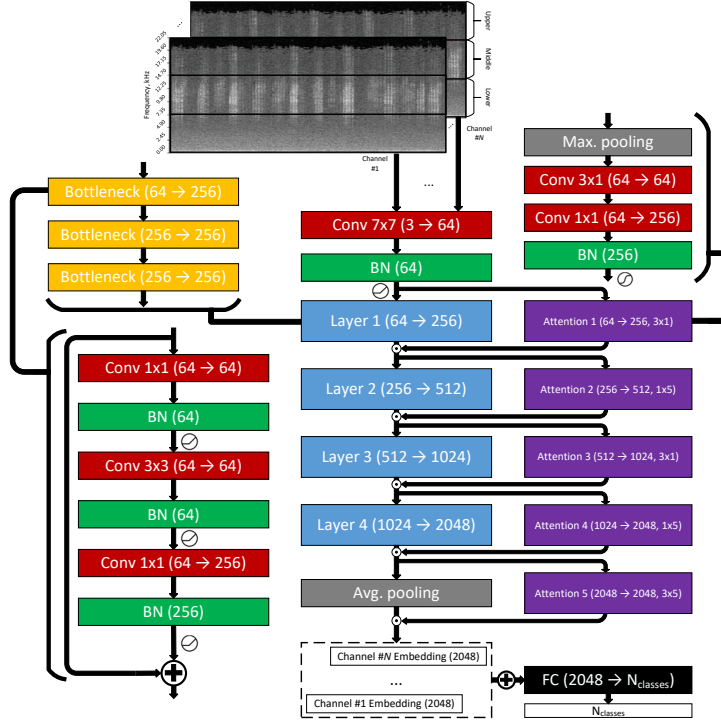


Fig. 4.1: Architecture of ESResNe(X)t.

audio modalities, and using them helps in better cross-modal mapping of information between the two domains. This improves the performance of the model in tasks such as environmental sound classification, where the model has to classify an input audio into a fixed set of predefined categories based on the underlying environmental sounds present in the sample [79]. Finally, using models from the visual domain helps in achieving better generalization and transfer learning performance [77], as visual features tend to capture high-level abstract representations that are transferable across different tasks and domains [297].

Spectrogram

This modification refers to the supporting changes in the architecture. A spectrogram is an image-like representation of the spectrum of frequencies varying with time. In relation to digital signal processing, there are several ways to obtain a spectrogram. It can be generated using filter banks, Fourier (or more generally wavelet) transform, etc. Here, the log-power spectrogram S is computed from the STFT (see Section 2.2.1 for details) of an audio signal $X(\tau, \omega)$ (Equation 4.1), which is, as a trade-off between time and frequency resolution, split into 37.5 ms-frames. The corresponding overlap between

subsequent frames follows the recommendations for the Blackman-Harris window [86] and is 66.1 % (24.8 ms).

$$S = 10 \cdot \log_{10}|X(\tau, \omega)|^2 \quad (4.1)$$

Spectrochromatic Mapping

Similarly to the use of time-frequency representation via STFT-spectrograms, this modification refers to the supporting changes in the architecture. For image classification models, a usual way to represent input data is an RGB model with three input channels (red, green, and blue). However, the spectrogram-based representation only provides input in the form of a single-channel (grayscale values). One way to tackle this issue is to replicate the spectrogram to other channels or to pass zeros instead. The major drawback of this solution lies, however, in either unnecessary redundancy and increased computational costs, or in the loss of information.

In order to overcome this limitation, there was proposed to map the spectrogram along its frequency axis onto the three input channels, so it is split into three frequency bands (Figure 4.2): lower (0.00 – 7.35 kHz), middle (7.35 – 14.70 kHz) and upper (14.7 – 22.05 kHz).

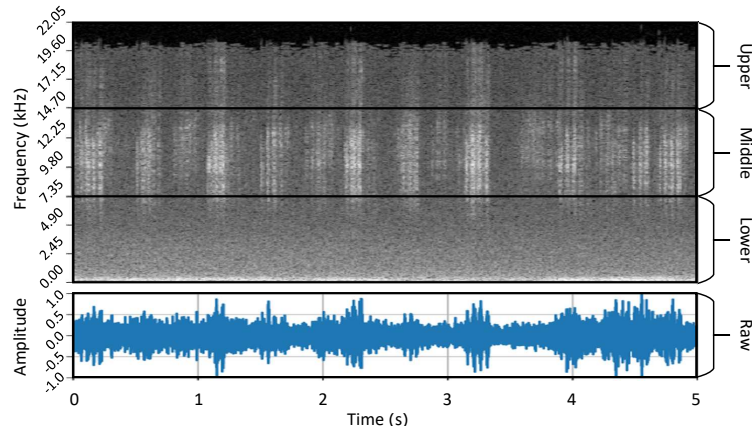


Fig. 4.2: Example of spectrochromatic mapping.

Visual Domain Initialization

This setup corresponds to the training workflow adjustments and becomes especially important in the case of the ESC-50 dataset as it provides the challenging task to distinguish between 50 classes using only 1600 training samples [200].

Modern deep neural networks consist of millions of parameters, which is substantially more than the number of training samples in common datasets, even those considered large-scale [45, 65]. For rather small datasets, this can lead to overfitting when trained from scratch. One way to mitigate such issue is to perform fine-tuning. For the ESResNet model, the decision was to employ a model that was trained from scratch on the ImageNet dataset [45]. As it will be shown in Chapter 5, the initialization of weights based on a pre-training on the ImageNet image classification task is beneficial for the environmental sound classification.

4.2.3 Audio-specific Architecture Enhancement

In the past years, the research community is witnessing a shift from computationally expensive, over-parametrized end-to-end models to more lightweight audio-domain-specific models that leverage efficient signal processing techniques [214]. This transition is motivated by the desire to address computational costs and enhance model interpretability in audio applications. Indeed, modifying convolutional neural networks for audio by using adaptive filters [288, 78] to represent spectral features of audio signals helps the model to capture relevant patterns in the input signal and learn more meaningful representations of the input data [56]. Using audio-specific architecture enhancements in deep learning models offers not only improvement of the model performance on audio-related tasks [224] but also reduction of the number of parameters needed to train the model [247, 248] and better interpretability of the learned representations [212].

Backbone Architecture

This architectural modification is one of the most influencing ones and refers to the group of main changes. The ESResNet model was proposed in [79] and combined commonly used visual domain techniques such as a ESResNet-based [85] backbone, Siamese-like [120] multichannel processing, and depth-wise separable convolutions [34] together with the computation of log-power spectrograms obtained using Short-Time Fourier Transform and cross-domain transfer learning in application to the ESC task. The chosen method allowed for achieving state-of-the-art results on the ESC-50 [200] and UrbanSound8K [227] datasets at the time of publishing of [78]. An overview of the ESResNet's processing pipeline is given by Figure 4.1. The further development of residual CNNs reflected in the one named ResNeXt, which served as a base for the improved sound classification model ESResNeXt [78].

Spectrotemporal Attention

Unlike the backbone architecture, which is a general-purpose component, the use of a spectrotemporal attention is an audio-specific enhancement that refers to the main modifications. The attention mechanism was presented initially for the use in conjunction with recurrent models, in particular, in sequence modelling tasks [253]. The initial main goal of it was to highlight relevant parts of a long sequence and to get rid of irrelevant ones. However, in the visual domain, one uses attention blocks in order to produce weighting for spatial regions of the input signal. Usually, there are several attention sub-branches consisting of one or many convolutional layers that process feature maps in parallel with the main branch.

For the environmental sound classification task, the main purpose of the attention blocks is to focus the model on the most important information in both the time and frequency domain. To implement the attention mechanism, the base ESResNet model (inspired by [293]) was extended by adding a stack of attention blocks in parallel to the main branch, as shown in Figure 4.1. Each block among the first four handles either frequency- or time-related information. For instance, the first attention block A_1 receives the same input as the first layer L_1 , then it processes the signal x using frequency-dedicated convolutional filters and provides an output of the same shape as the one provided by the L_1 . Finally, the input of the second layer L_2 is constructed by the element-wise multiplication of outputs of L_1 and A_1 blocks (Equation 4.2).

$$L_i^{att.}(x) = L_i(x) \odot A_i(x) \quad (4.2)$$

The last attention block handles the joint time-frequency representation. The core of the attention block is the depth-wise separable convolution [34], while the output of each attention block is given by the logistic function.

Siamese Multichannel Audio Processing

Changing the model architecture towards Siamese-like [120] processing of multichannel input is another example of a main audio-centric architectural enhancement. As the way how humans perceive audible information is inherently stereo, Siamese neural networks for audio mimic it by exploiting (dis-)similarity between two input channels [79]. As Figure 4.3 illustrates, a two-channel audio input followed by log-power spectrogram computation (via STFT) is taken and passed through the layers in a channel-wise manner. After obtaining the network's outputs, they are fused by an element-wise addition, and

the resulting embedding is passed through the last fully-connected layer that performs the final classification.

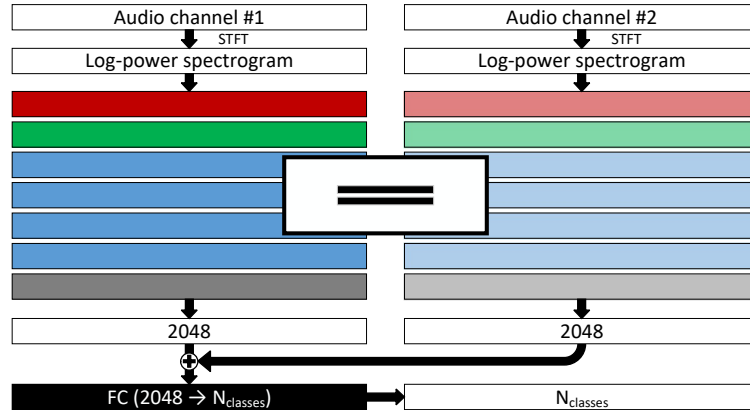


Fig. 4.3: Schema of processing the stereo input.

Trainable fbasp-layer

This modification from the group of main enhancements is the most audio-centric among those described in this thesis.

As mentioned in Section 4.2.2, in the ESResNet model [79], the short-time Fourier transform was applied to transform the input audio to the corresponding time-frequency representation. This operation can be decomposed into two independent steps: signal framing and a subsequent discrete Fourier transform of each frame. STFT (Equation 2.28) is used to determine magnitude and phase of basis sinusoidal frequencies f_c at different time points τ in a time-domain signal x using a kernel function $K(n)$ (Equation 2.27), thus obtaining a *static* time-frequency representation of audio.

Complex Frequency B-Spline Wavelets (fbasp)

The *adaptive* time-frequency transformation performed by the proposed fbasp-layer is defined by Equation 4.3.

$$X_{fbasp}(\tau, m, f_b, f_c) = \sum_{n=0}^{N-1} x[n]w[n - \tau] \cdot K_{fbasp}^{m-f_b-f_c}(n) \quad (4.3)$$

However, one can notice that the DFT-kernel defined by Equation 2.27 represents an inverted in time (i.e. complex conjugate) limiting case of the fbasp-kernel (Equation 2.32). Thus, assigning the fbasp-layer's parameters $m = 0$ and $f_b = 1$, its initialization becomes

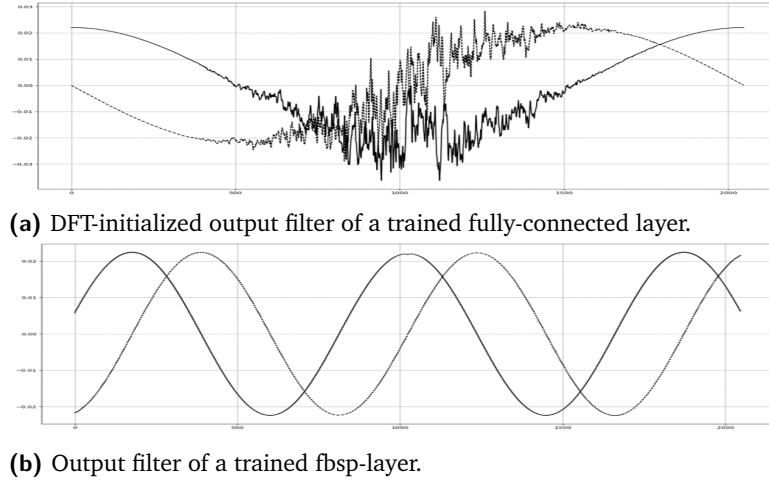


Fig. 4.4: Example of a trained filter from a linear and fbsp-layer.

identical to the inverse DFT, which makes it possible to start the network's training from a good enough state.

A regular linear layer can be regarded as a collection of output filters that transform inputs through the computation of dot products. However, it often suffers from the issue of coefficients in the output filters being excessively influenced by the noise present in the input data during training, as shown in Figure 4.4a. This sensitivity to noise can lead to a lack of robustness in the layer's performance. To address this limitation, an alternative approach is to algorithmically derive the coefficients instead of explicitly learning them, as done for the fbsp-layer. This involves the use of a kernel (Equation 2.32) that depends on a small number of parameters and, thus, results in a smooth time-frequency transformation, as shown in Figure 4.4b. By learning these parameters, the coefficients of the output filters can be obtained in the runtime, enabling the dot product computation with the input. This modified layer exhibits reduced sensitivity to high-frequency noise, resulting in improved non-adversarial robustness of the associated models, as demonstrated in Section 5.2.

fbsp-specific Loss Term

In order to regularize weights of the fbsp-filter bank and to preserve the overall signal's energy, an additional loss term should be introduced that is specific to the fbsp-layer and, for the N-point fbsp-transform, is described by Equation 4.4.

$$\mathcal{L}_{fbsp} = \frac{1}{N} \sum_{f_c=0}^N (\|K_{fbsp}^{m-f_b-f_c}\|^2 - 1)^2 \quad (4.4)$$

4.2.4 Audio-visual Correspondence Learning

Moving from the transfer-learning-based setup to the audio-visual correspondence learning is a major modification that turns an audio-centric model into a multimodal one, thus providing new opportunities for intra- and intermodal signal processing. First, it allows the transfer of knowledge from multiple modalities to improve the classification accuracy of sound events [32, 77]. Second, it enables the incorporation of the knowledge learned from different modalities to enhance the understanding of the semantic relationships between sounds and their corresponding visual cues, which can lead to more robust classification models. Third, knowledge distillation helps reduce the size and complexity of the model, which is crucial for the deployment of the model in resource-constrained environments. Finally, using a knowledge distillation approach can facilitate the integration of new modalities, thus enhancing the overall performance of the model without the need for extensive retraining. Therefore, the benefits of using a knowledge distillation approach in sound classification tasks lie in its ability to leverage the knowledge learned from multiple modalities and to enhance the robustness and generalization capabilities of the model.

CLIP and Beyond

This thesis proposes a novel audio-visual model named AudioCLIP [77], whose architecture is shown in Figure 4.5.

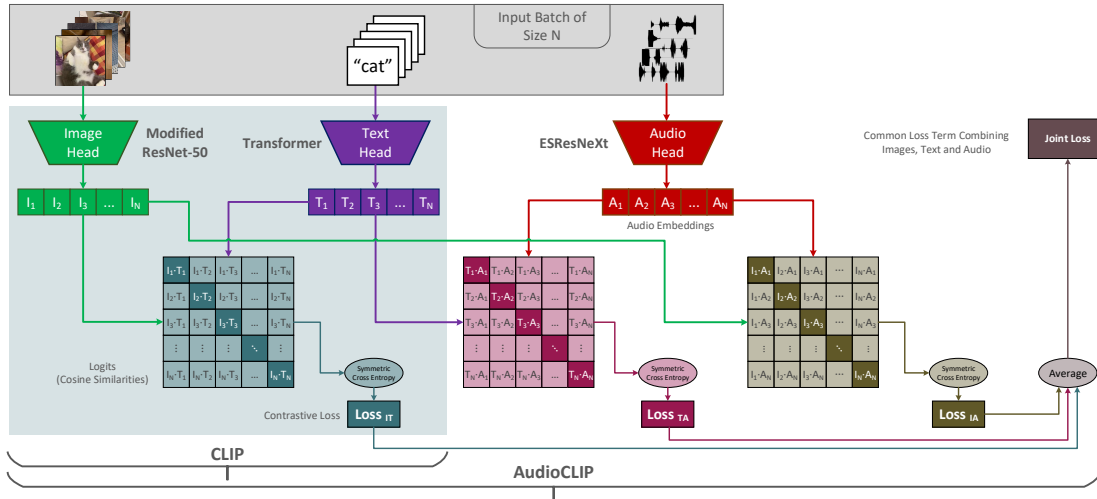


Fig. 4.5: Architecture of the AudioCLIP model.

From the network design point of view, the proposed model inherits from the CLIP framework and naturally extends the existing architecture by introducing an additional

audible modality. AudioCLIP combines a ResNet-based CLIP model [206] for visual and textual modalities and ESResNeXt [78] for audible modality, thus the proposed architecture consists of image-, text- and audio-heads. The choice of the CLIP’s variant in favor of the one based on ResNet-50 was conditioned by the lower computational complexity of the latter, in comparison to the ViT-based one [206]. The newly added modality is considered as equally important as the originally present ones.

Similar to CLIP, AudioCLIP learns to align different representations of an input sample in a pair-wise manner across modalities. Such an alignment is done by generating M -dimensional vector embeddings for each modality ($M = 1\,024$) and the subsequent computing of pair-wise cosine similarities across samples, where the latter serve as inputs for the symmetric cross entropy loss function [206].

The use of the third (audible) modality also introduces two new cross-modal alignment terms: text-audio (TA) and image-audio (IA), as shown in Figure 4.5. Going further, one can calculate the number of cross-modal loss terms as the number of 2-combinations within $N_{Modalities}$ modality-specific loss terms, as specified by Equation 4.5. Thus, AudioCLIP fuses the original CLIP’s image-text contrastive learning setup (filled rectangle, on the left of Figure 4.5) and its audible modality-based extensions in order to enable the simultaneous processing of the three basic modalities in arbitrary combinations.

$$N_{\text{Loss Terms}} = \frac{N_{Modalities}!}{2 \cdot (N_{Modalities} - 2)!} \quad (4.5)$$

Multistage Training

Similarly to [206], training of AudioCLIP relies on the contrastive learning approach that is based on the maximization of cosine similarities between multimodal embeddings using symmetric cross-entropy loss; it can consist of up to three stages.

For each experiment, modality-specific heads of AudioCLIP need to be pre-initialized (**Stage 1**). Image- and text-heads, which build up CLIP, obtain vanilla weights acquired on the CLIP dataset [65]. The audio-head, namely ESResNeXt, in all cases uses pre-loaded ImageNet-weights to be then fitted on AudioSet, as it was shown to be beneficial [78].

Due to its nature, AudioSet allows for putting different modalities together so that audio recordings, their textual labels, and the corresponding video frames serve as inputs for the audio-, text- and image-heads of AudioCLIP, respectively. In particular, audio tracks and the respective class labels allow for performing a joint image-to-audio transfer learning of ESResNeXt (**Stage 2**), and then, the extracted frames, in addition to audio

and class names, serve as an input for AudioCLIP. During training, one out of 10 video frames has to be randomly picked ($\sim \mathcal{U}$) and passed through the image-head. In the evaluation phase, the same extraction procedure holds, with the difference that only a central frame should be presented to the model.

For the multimodal training based on AudioSet (**Stage 2**), two different paths are available: either to keep the shared embedding space compatible to the one of vanilla CLIP (filled rectangle on the left of Figure 4.5) or to fit AudioCLIP better to the distribution of audio. The first option implies training of the audio-head while keeping image- and text-heads unchanged, resulting in audio embeddings trying to align to the frozen CLIP embeddings. The second option allows the full AudioCLIP for following the audio distribution closer, as in this case, a mutual knowledge distillation involving all three heads is performed (Chapter 5 presents the evaluation results).

Independent of the chosen AudioSet training strategy, one can also introduce a final fine-tuning (**Stage 3**, optional) to fit AudioCLIP to the target datasets, namely UrbanSound8K and ESC-50. The AudioCLIP design allows for performing zero-shot inference on these datasets without the fine-tuning step, although it is still beneficial. The latter, if needed, has to involve the audio-head exclusively, under the supervision of the rest of the AudioCLIP model.

When trained on AudioSet, the audio recordings, their textual labels, and the corresponding video frames are used as inputs for the audio-, text- and image-heads of AudioCLIP, respectively. Freezing of the image- and text-heads implies exclusion of the corresponding network parameters from the backpropagation during the optimization step, thus keeping the values of these parameters equal to the initialization. This makes possible the comparison of the performance of AudioCLIP relying on a vanilla CLIP initialization to the audio domain-adapted one, as shown later in Chapter 5.

Concluding Remarks

This chapter has delved into the core techniques and practical applications of audio analysis within the realm of multimodal learning. By exploring the research objectives, associated challenges, and proposed hypotheses, a comprehensive understanding of the technical aspects and practical implementations in the field of audio analysis for multimodal learning has been established. These insights pave the way for the subsequent experimental evaluation of the proposed hypotheses, where their effectiveness and potential enhancements can be rigorously tested. The next chapter reveals the efficacy of the proposed approaches and provides insights for further advancements in audio-centric multimodal learning.

Results and Discussion

In this chapter, the benefits of using visual-domain-models for solving audio-specific tasks are demonstrated. It is also shown how audio-specific architecture enhancements improve the performance and robustness of the models. Finally, the benefits of going from the transfer-learning-based setup to the audio-visual correspondence learning are discussed, along with the results obtained from the related experiments.

These results cover audio-related tasks, such as audio-visual multi-label classification on AudioSet [65], environmental sound classification on ESC-50/-10 [200] and UrbanSound8K [227], cross-modal retrieval of images or audio by textual labels and *vice versa* on ImageNet [45], AudioSet, ESC-50, and UrbanSound8K, as well as findings in zero-shot environmental sound classification on ESC-50 and UrbanSound8K. The chapter presents evaluation of the enhanced and baseline network architectures and analyzes the impact of different modifications on the performance of the models, thus providing insights into the mechanisms behind the improvements.

5.1 Audio-visual Multilabel Classification

Audio-visual multilabel classification allows for assessing the proposed model enhancement approaches to quantify their influence on the resulting performance. As AudioSet [65] consists of both audible and visual counterparts of the samples, it served as a unified evaluation basis for comparing different stages of the architecture’s development. Specifically, Table 5.1 covers audio-only models ESResNet [79] and ESResNeXt [78], as well as their further evolution towards the multimodal learning, namely AudioCLIP [77].

The evaluation of ESResNet equipped with both *static* (STFT) and *adaptive* (fbasp) time-frequency transformation is demonstrating the performance growth in the range $18.92\% \rightarrow \mathbf{25.14\%}$ and $23.94\% \rightarrow \mathbf{26.16\%}$, respectively. This clearly shows the advantages of using the visual-domain-based initialization for solving the purely audio task, which supports the hypothesis $\mathcal{H}^{(1.1)}$. Additionally, for the base time-frequency transformation that relies on the *static* short-time Fourier transform, the change of

Tab. 5.1: Mean average precision (mAP, %) on AudioSet.

Model	Notes	Modality		Audio Input	ImageNet Initialized	mAP
		Image	Audio			
ESResNet			✓	STFT	✓	18.92
				fbasp	✓	25.14
						23.94
						26.16
ESResNeXt			✓	STFT	✓	25.14
				fbasp	✓	28.17
						8.93
						14.82
AudioCLIP	* IH, TH	✓	✓	fbasp	AH	25.85
						28.36
						25.11
						32.38

* – frozen weights.

Abbreviations: AH: Audio-Head a.k.a. ESResNeXt; IH: Image-Head; TH: Text-Head;
STFT: Short-Time Fourier Transform; fbasp: (complex) Frequency B-Spline (wavelets).

the model backbone from ResNet-50 to ResNeXt-50 is advantageous in terms of the multilabel classification accuracy (18.92 % → **23.94** %).

The next aspect of the model evaluation refers to the audio-specific architecture enhancement, namely the use of the trainable fbasp-layer instead of the STFT. Here, both variants of the base model—ESResNet and ESResNeXt—gain additional mAP-score, and the latter model demonstrates a larger increase in the performance (25.14 % → **26.16** % and 25.14 % → **28.17** %, respectively), thus supporting the hypotheses $\mathcal{H}^{(1.2)}$ and $\mathcal{H}^{(1.3)}$.

Finally, the change from the transfer-learning-based setup to the knowledge-distillation-based correspondence learning allows for leveraging the multimodal nature of AudioSet by evaluating the peak of the evolution of the proposed architectures on two input modalities in both setups, separately and simultaneously, which is in agreement with the hypothesis $\mathcal{H}^{(1.4)}$.

In all three input layouts—image, or audio, or both—the full model fine-tuning allows for achieving higher mean average precision—8.93 % → **14.82** %, 25.85 % → **28.36** %, and 25.11 % → **32.38** %, respectively—, thus suggesting that the fitting of all modality-specific branches plays an important role even if the particular modality is not directly related to the task. At the same time, the simultaneous processing of both visual and audible modalities increases the value of the evaluation metric to $mAP = 32.38$ %, in

comparison to using images (14.82 %) or audio (28.36 %) exclusively. These evidences are clearly consistent with the hypothesis $\mathcal{H}^{(1.5)}$.

Therefore, the experimental results obtained on the audio-visual multilabel classification task provide empirical evidences that transfer learning from the visual domain, audio-dedicated trainable time-frequency transformation, and the use of a knowledge-distillation-based correspondence learning setup, as well as combination of the methods contribute to the increased model performance.

5.2 Environmental Sound Classification

As the environmental sound classification is a more fine-grained instance of the multilabel audio classification, it provides a perfect test ground for new methods to improve performance of the deep learning models. Moreover, due to the conciseness of the task formulation, one can go beyond measuring accuracies towards assessing other properties of models, such as robustness to input distortions.

5.2.1 Performance Evaluation

The release of two key environmental sound classification datasets—ESC-50/-10 [200] and UrbanSound8K [227]—matched the dawn of deep learning, which set the direction for the development of audio-centric models. This section provides a thorough performance evaluation of the proposed models along with the retrospective on the progress of the field.

Table 5.2 outlines the evolution of audio classification methods that were aiming to reach state-of-the-art performance at the time of publishing. Although the majority of the approaches centered around image-like representations—first of all, Mel-scaled [151] spectrograms—other avenues to handle the sound within deep learning models exist as well and include but are not limited to the use of raw audio or learning the time-frequency representation as a part of the model.

Since the aforementioned representations are based on spectrograms and imply a nonlinear downscaling of the input, it was not a surprise that pure full-size spectrograms allowed a powerful CNN leveraging the benefits of transfer learning, thus out-performing its competitors [79], as assumed by the hypothesis $\mathcal{H}^{(2.1)}$. The use of an enhanced architecture—in both the backbone and time-frequency representation aspects—provided an additional improvement over other methods and yielded a special

Tab. 5.2: Accuracy (%) on ESC-50 and UrbanSound8K.

Year	Model	Source	Audio Input	ImageNet Initialized	AudioSet Pre-Trained	ESC-10	ESC-50	US8K
2015	<i>Human</i>	[200]	N/A	N/A	N/A	95.70	81.30	–
Raw audio and 1D-CNN								
2107	EnvNet	[247]	raw			88.10	74.10	71.10
2017	EnvNet v2	[248]	raw			91.30	84.70	78.30
2018	Multiresolution 1D-CNN	[302]	raw			–	75.10	–
Learnable filter bank and 2D-CNN								
2017	Piczak-CNN + ConvRBM	[224]	FBE			–	86.50	–
Spectrogram and 2D-CNN								
2015	Piczak-CNN	[199]	Mel-spec			90.20	64.50	73.70
2017	SB-CNN	[226]	Mel-spec			–	–	79.00
2017	GoogLeNet	[23]	Mel-spec, MFCC, CRP			86.00	73.00	–
2017	Piczak-CNN	[5]	(TEO-)GT-spec			–	81.95	–
2017	Piczak-CNN	[243]	(PE)FBE			–	84.15	–
2018	VGG-like CNN + mix-up	[292]	Mel-, GT-spec			91.70	83.90	83.70
2019	VGG-like CNN + Bi-GRU + att.	[293]	GT-spec			94.20	86.50	–
2020	CNN10	[15]	Mel-spec	✓		–	90.00	86.10
2020	WEANET N^0	[125]	Mel-spec	✓		–	92.60	–
2020	WEANET N^4	[125]	Mel-spec	✓		–	94.10	–
2020	DenseNet-201	[189]	Mel-spec			–	72.50	76.32
2020	DenseNet-201	[189]	Mel-spec	✓		–	91.16	85.14
2020	DenseNet-201 × 5 (ensemble)	[189]	Mel-spec	✓		–	92.89	87.42
2021	AST	[68]	Mel-spec	✓	✓	–	95.60	–
2021	ERANN	[254]	Mel-spec		✓	–	96.10	–
2020	ESResNet	[79]	STFT-spec	✓		97.00	91.50	85.42
2021	ESResNeXt	[78]	fbsp-spec	✓	✓	–	95.20	89.14
2022	AudioCLIP	[77]	fbsp-spec	✓	✓	–	97.15	90.07

Abbreviations: FBE: Filter Bank Energies [224]; spec: spectrogram; MFCC: Mel-Frequency Cepstral Coefficients [105]; CRP: Cross Recurrence Plot [160]; TEO: Teager's Energy Operator [108]; GT: GammaTone [235]; (PE)FBE: (Phase-Encoded) Filter Bank Energies [243]; STFT: Short-Time Fourier Transform; fbsp: (complex) Frequency B-Spline (wavelets).

gain in robustness [78] (discussed later in Section 5.2.2), being in alignment with the hypothesis $\mathcal{H}^{(2.2)}$. Finally, the audible and visual modality used jointly with architectural enhancements for the knowledge-distillation-based model training led to a further increase of accuracy, thus supporting the point that the aforementioned methods contribute to the performance of deep learning models (hypothesis $\mathcal{H}^{(2.3)}$).

Multichannel Processing

Since the use of more than one audio channel was not common among existing environmental sound classification models, such an architectural enhancement was considered for a more detailed investigation. To not make the model dependent on a particular configuration of audio channels, a Siamese-like modification was done for the base ESResNet architecture. As the altered model is able to process both mono and stereo input, the influence of considering additional audio channels was studied.

Specifically, the between-channel difference provided useful information that allowed the model out-performing previous state-of-the-art results on UrbanSound8K, as shown in Table 5.3. The performance gain kept positive independently from the use of spectrotemporal attention or ImageNet initialization and varied from 84.21 % $\rightarrow$ **85.42 %** (1.21 p.p.,

attention-boosted + initialization from ImageNet) to 81.00 % $\rightarrow$ **82.76 %** (1.76 p.p., attention-boosted + training from scratch). Thus, the Siamese-like architecture enhancement for the multichannel audio processing was shown to be advantageous for the model performance in the environmental sound classification task, which supports the hypothesis $\mathcal{H}^{(2.4)}$.

Spectrotemporal Attention

In the context of audio-centric architecture enhancements, it has been shown [293] that the parallel stacking of auxiliary attention layers positively influences on the model classification abilities. Therefore, a lightweight spectrotemporal attention layer made its place within the structure of ESResNe(X)t. Such an attention block contributed to the performance improvements by providing a steady accuracy increase in different scenarios, as presented in Table 5.3.

Interestingly, the higher performance gain is consistently observed for the training from scratch (81.15 % $\rightarrow$ **83.15 %**, ESC-50), in comparison to the fine-tuning of the ImageNet-initialized model (90.80 % $\rightarrow$ **91.50 %**, ESC-50), which suggests that the introduced attention blocks help the model learn features that have greater discriminative power. To conclude, the spectrotemporal attention blocks support the idea of improving the model accuracy by incorporating audio-centric architecture modifications, thus supporting the hypothesis $\mathcal{H}^{(2.5)}$.

Tab. 5.3: Influence of architecture on the accuracy (%).

Model	Notes	Audio Input	ImageNet Initialized	ESC-50	UrbanSound8K	
				mono	mono	stereo
ESResNet		STFT	✓	81.15	79.91	81.31
				90.80	83.59	84.90
			✓	83.15	81.00	82.76
				91.50	84.21	85.42

Abbreviations: STFT: Short-Time Fourier Transform.

Trainable fbasp-layer

Changing from the regular short-time Fourier transform to the trainable complex-frequency-B-spline-wavelet-layer allows the model to learn adaptive representations tailored to the specific characteristics of the input data. As Table 5.4 demonstrates, such replacement leads, in general, to improved performance.

Tab. 5.4: Influence of transformation on the accuracy (%).

Model	Audio Input	ImageNet Initialized	AudioSet Pre-Trained	ESC	US8K
ESResNet	STFT	✓	✓	83.15	82.76
				91.50	85.42
				92.45	87.74
	fbsp	✓	✓	93.35	88.03
				86.25	83.20
				91.25	85.92
ESResNeXt	STFT	✓	✓	92.40	88.47
				93.80	88.38
				91.60	86.02
	fbsp	✓	✓	95.00	89.02
				91.30	85.47
				95.20	89.14

Abbreviations: ESC: ESC-50 [200]; US8K: UrbanSound8K [227]; STFT: Short-Time Fourier Transform; fbsp: (complex) Frequency B-Spline (wavelets).

For the randomly initialized ESResNet model, the fbsp-layer provides the increased accuracy on both datasets. However, using only ImageNet initialization or pre-training on AudioSet does not contribute to the classification results, as expected. Specifically, one can observe a slight degradation for ESC-50 while the accuracy UrbanSound8K still grows with the change from the STFT to the fbsp-layer. The joint two-stage model initialization by using both pre-training datasets—ImageNet and AudioSet—is beneficial for both kinds of the time-frequency transformation (which is consistent with the hypothesis $\mathcal{H}^{(2.6)}$), and the fbsp-layer offers an improvement over the former transformation on ESC-50, thus being in alignment with the hypothesis $\mathcal{H}^{(2.7)}$. At the same time, one can notice a minor deterioration of the performance obtained from a two-stage initialized ESResNet that was fine-tuned to UrbanSound8K, which can hint at the exhaustion of the effective model capacity.

The incorporation of ResNeXt, an advanced model backbone, and its experimental evaluation produces results that align with the performance observations made with the base ESResNet architecture. This finding implies that a seamless transition between datasets using the two-stage pre-training approach is beneficial for harnessing the advantages of the adaptive time-frequency transformation, particularly the fbsp-layer. By introducing a trainable fbsp-layer into the model architecture, it offers an additional avenue for enhancing the performance of audio-centric deep learning models in the task of environmental sound classification, regardless of the underlying model architecture. This modification allows the model for adapting and refining its representation of audio signals, improving its ability to capture relevant features and discriminate between different sound classes. Thus, the integration of the powerful backbone and the trainable

fbsp-layer demonstrates a promising direction for advancing the performance of audio-centric deep learning models in environmental sound classification tasks.

5.2.2 Robustness Evaluation

Apart from the model comparison, there was also evaluated the robustness against two types of signal perturbations: additive white Gaussian noise and reduction of an effective sample rate.

Additive White Gaussian Noise

Additive White Gaussian Noise, a type of background noise commonly found in audio recordings, can significantly affect the model's performance. The evaluation of model performance on the ESC-50 and UrbanSound8K datasets given different values of SNR shows clearly that the use of the fbsp-layer improves the model's robustness to the presence of additive white Gaussian noise, which attributes to the hypothesis $\mathcal{H}^{(2.8)}$. The fbsp-equipped models provide higher classification accuracy given decreasing SNR on both datasets, ESC-50 and UrbanSound8K, as shown in Figure 5.1 and Figure 5.2, respectively, hence demonstrating robustness of audio-centric deep learning models to the additive noise in a non-adversarial scenario.

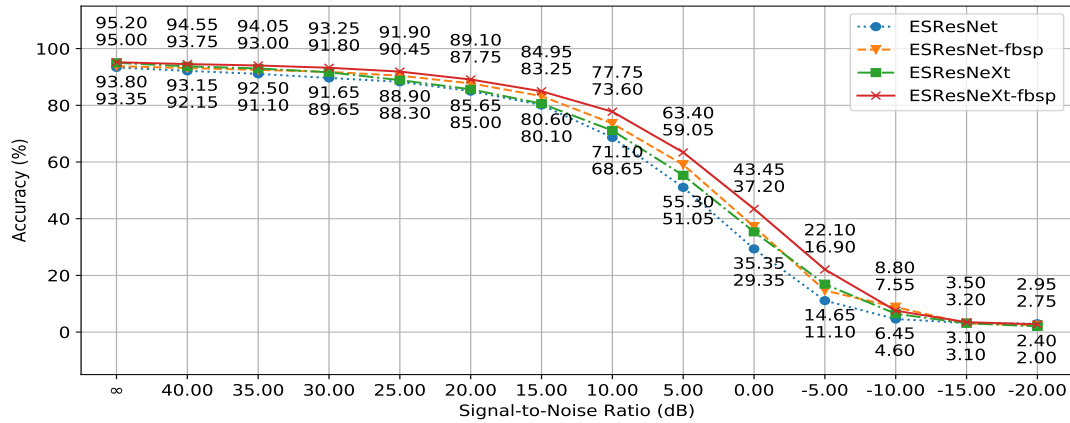


Fig. 5.1: Accuracy (%) at different SNRs (dB) on ESC-50.

Sample Rate Reduction

Decreasing the number of samples in an audio signal can impact the quality and fidelity of the audio. In order to quantify the influence of the effective sample rate reduction

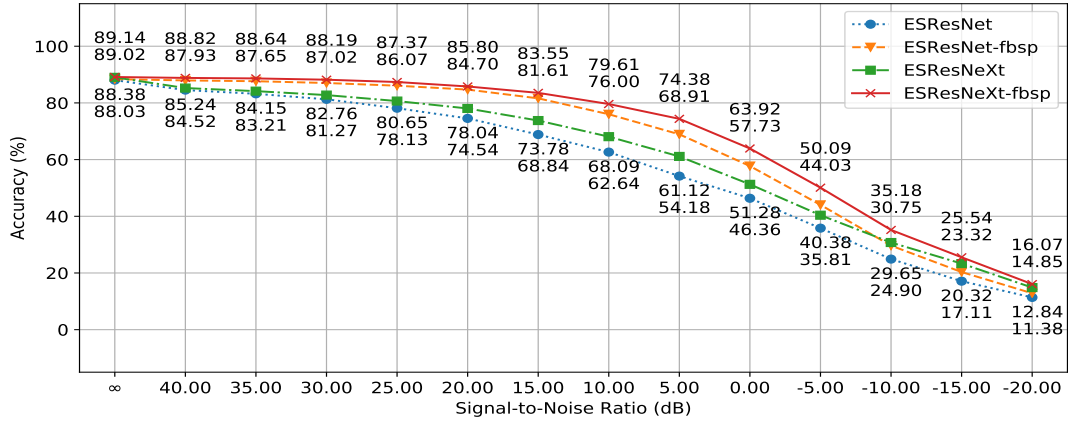


Fig. 5.2: Accuracy (%) at different SNRs (dB) on UrbanSound8K.

on the model accuracy, several experiments were performed. The obtained results support the point that fbsp-equipped models are able to extract the information that is relevant for classification more effectively. In particular, the significant performance drop at the cutoff frequency 4 kHz occurred in the case of the STFT-based models on both datasets ESC-50 (Figure 5.3) and UrbanSound8K (Figure 5.4) indicates that the use of the fbsp-equipped models is beneficial for the applications that imply the use of low-bandwidth channels.

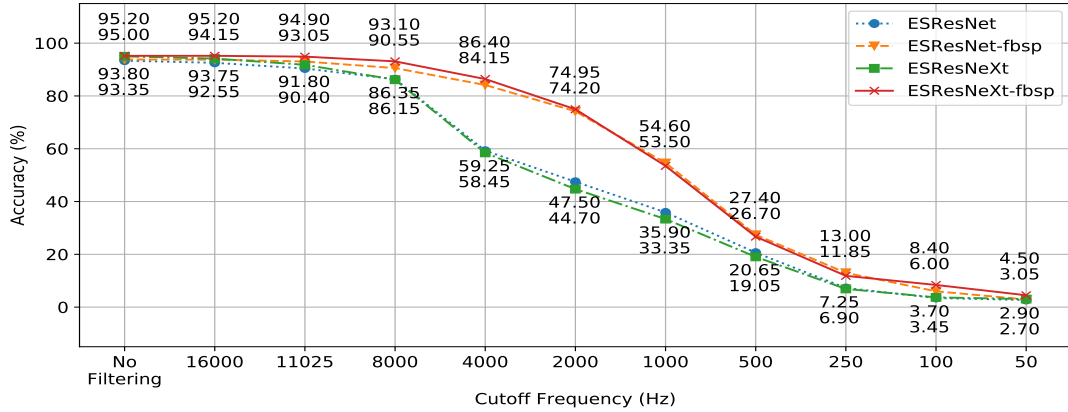


Fig. 5.3: Accuracy (%) at different sample rates (Hz) on ESC-50.

Also, the ResNet backbone demonstrated less sensitivity to lower frequency cutoffs, in comparison to ResNeXt. Thus, modifying the audio-dedicated ESResNe(X)t model with the trainable fbsp-layer contributes to its robustness against the degradation of the signal quality, thus supporting the hypothesis $\mathcal{H}^{(2.9)}$.

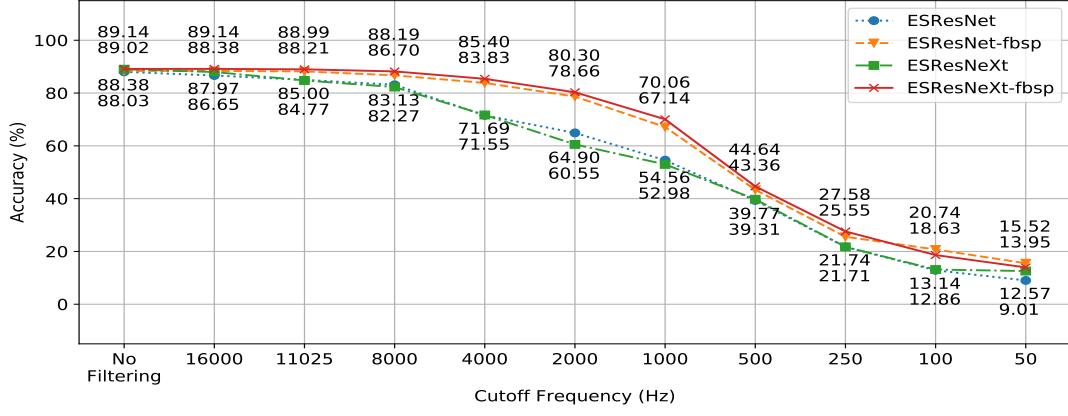


Fig. 5.4: Accuracy (%) at different sample rates (Hz) on UrbanSound8K.

Qualitative Results

Apart from this, there was also computed frequency response—dependency of the output gain on the frequencies of the input signal—of fbsp-layers trained on AudioSet, which was then compared to those computed for the STFT-filter bank.

The frequency responses of the trained fbsp-filter banks (Figure 5.5a – Figure 5.5c) consist of distinguishable peaks and valleys in the frequency domain. This fact may indicate that the input signal is redundant w.r.t. sample rate, i.e. the networks learned some sort of signal decimation. Also, the trained fbsp-filter banks obtain stopbands in the higher frequency band, which suggests that the filter banks are able to suppress high-frequency components of an input signal’s noise.

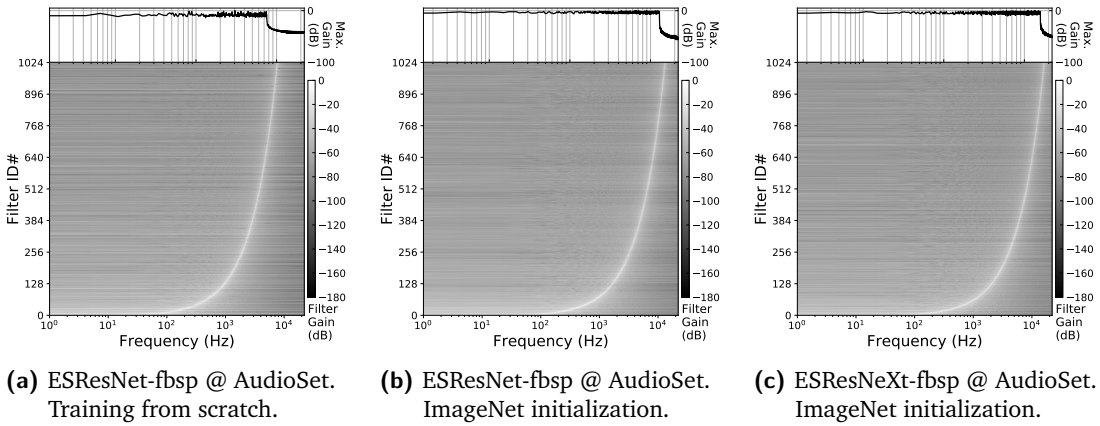


Fig. 5.5: Frequency response of learned filter banks.

In contrast, for the DFT-matrix of the STFT-filter bank, one can observe in Figure 5.6 that the gain is almost flat for the entire frequency spectrum, up to the Nyquist frequency,

meaning that the STFT-based time-frequency transformation does not alter the input of the model, keeping both signal and noise unaffected.

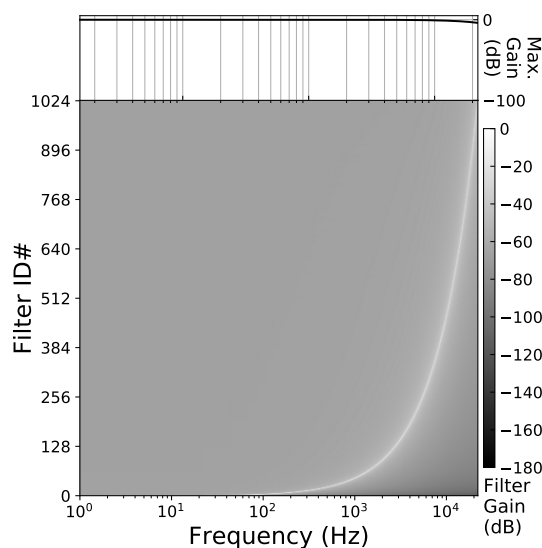


Fig. 5.6: Frequency response of DFT filter bank.

Insights and Implications

In the case of audio-centric models, incorporating the trainable time-frequency transformation via the fbsp-layer is advantageous to maintain prediction accuracy in the presence of common environmental factors. By mitigating the influence of audio quality degradation factors, such as the additive white Gaussian noise and sample rate reduction, audio-centric deep learning models provide more reliable predictions in real-world audio scenarios, improving their practical applicability and performance in various audio analysis tasks. Thus, enhancing ESResNe(X)t by using the fbsp-layer is beneficial for increasing the non-adversarial robustness of the audio-centric models.

5.3 Multimodal Retrieval

The CLIP model, which stands for Contrastive Language-Image Pretraining, introduced the ability to perform retrieval given either an image or a text query, which CLIP then maps into the joint embedding space. Then, by calculating distances or similarities in this space, one can identify the closest matching images or texts. This retrieval capability enables tasks like image search based on text descriptions or finding relevant text descriptions for a given image.

AudioCLIP, in turn, extends CLIP’s multimodal capabilities to audio. As AudioSet provides sound recordings and accompanying textual tags and visual frames, AudioCLIP trained on a large corpus of grouped audio, frames, and their corresponding text captions, is able to identify associations between different representations of the same concept. Thus, the model learns to encode its inputs into a shared latent space, where objects having similar semantics are closer to each other. This joint embedding space allows AudioCLIP for catching relationships between audio, images and texts.

Text $\Rightarrow$ Image

In this setup, the text input is encoded into a vector representation using the language encoder of the AudioCLIP model. Similarly, the image is encoded into a vector representation using the vision encoder. These two vectors are then compared in the joint embedding space to find the closest matches.

The complete training on the AudioSet dataset results in an improvement in the mAP performance score. Nonetheless, this training approach attenuates the querying performance on ImageNet, as its underlying distribution should be substantially different from that of AudioSet.

Text $\Rightarrow$ Audio

This schema involves encoding the text query and the audio sample separately using their respective heads. The text encoder transforms the query into a numerical representation in the model’s latent space, while the audio encoder projects the audio sample into the same space. The AudioCLIP model then compares the text embedding to the audio one by measuring the cosine similarity. The results presented in Table 5.5 indicate the model’s ability to perform text-based retrieval of audio, which is aligned with the hypothesis $\mathcal{H}^{(3.1)}$.

On the AudioSet and UrbanSound8K datasets, training the models using the full training set results in improved querying performance. In contrast, for the ESC-50 dataset, the impact on the querying performance is not that significant, with only a minimal difference observed.

Tab. 5.5: Precision, recall, and mAP (%) of AudioCLIP.

Dataset	Modality		Audio-Head Only			Full Model		
	Query	Target	P@1	R@1	mAP	P@1	R@1	mAP
ImageNet	text	image	5.42	84.15	52.91	1.61	89.00	33.13
AudioSet	text	image	0.81	46.79	9.51	1.31	74.95	17.22
	text	audio	2.51	84.38	23.54	5.33	76.13	30.79
	audio	image	0.62	56.39	5.45	1.03	52.12	7.22
	image	audio	0.61	61.94	4.86	1.20	53.15	6.86
ESC-50	text	audio	51.25	85.20	77.20	51.78	80.40	76.97
UrbanSound8K			40.81	47.69	77.43	42.28	48.18	80.04

Abbreviations: P@1: Top-1 Precision (%); R@1: Top-1 Recall (%); mAP: Mean Average Precision (%).

Image $\Leftrightarrow$ Audio

AudioCLIP enables trully multimodal retrieval, allowing for querying of audio by image and *vice versa*. In the case of querying audio by image, AudioCLIP processes the image input by extracting visual features and projects it into the embedding space. The visual embedding is compared then to the audio embeddings to identify relevant sound recordings that correspond to the queried image, thus supporting the hypothesis $\mathcal{H}^{(3.1)}$. Similarly, when querying an image by audio, the audio input is processed to obtain an audio embedding, which is compared to the image embeddings to find semantically relevant images. The complete training of the AudioCLIP model yields improved querying performance—in terms of mAP score—for both types of queries: audio by image and image by audio.

[Audio $\vee$ Image] $\Rightarrow$ Text

Retrieving textual labels by images or audio can be considered equivalent to performing a regular classification task. In both cases, the goal is to assign a label or a set of labels to the input data based on its content. The underlying process involves mapping the input (image or audio) to a set of predefined classes or categories. The retrieval task is essentially a form of classification, where the model aims to associate the most relevant labels with the given input. This thesis summarizes the model performance in retrieving textual labels given an image or audio in Section 5.2.

Insights and Implications

The incorporation of the audible modality into the CLIP model by using knowledge-distillation-based correspondence learning enables it to effectively handle audio, expand-

ing its capabilities beyond images and text, which supports the hypothesis $\mathcal{H}^{(3.1)}$. By simultaneously handling these three modalities, AudioCLIP becomes a powerful tool for performing cross-modal retrieval, allowing for seamless retrieval of information across different representations. Additionally, the full model fine-tuning to the audio-centric data further increases the retrieval performance, which aligns with the hypothesis $\mathcal{H}^{(3.2)}$. At the same time, it is important to mention the negligible top-1 precision values obtained on AudioSet. Such a low performance strays from the mainstream results and can be possibly attributed to the challenging hierarchical structure of the dataset’s targets. Nevertheless, the integration of multiple modalities and the ability to leverage joint semantical space have significant implications for improving audio-centric deep learning models by boosting their performance and building an architectural basis for further advancements.

5.4 Zero-shot Sound Classification

Zero-shot classification is a technique in machine learning where a model is trained to attribute input samples to categories that it has never seen during training. By leveraging the joint embedding space, AudioCLIP is able to perform zero-shot classification by projecting both the sample (audio) and the target (text label) into the shared space, which provides supporting evidences for the hypothesis $\mathcal{H}^{(4.1)}$. The model then computes the similarity or distance between the sample and target, enabling the classification of the sample into the appropriate class.

Table 5.6 presents zero-shot classification results. Being fitted to the audible counterpart of AudioSet—i.e., image- and text-related weights were frozen—the model was able to out-perform baseline methods on ESC-50 by a wide margin (33.00 % $\rightarrow$ **69.40 %**). As on the UrbanSound8K there existed no zero-shot classification baseline, the obtained accuracy (65.31 %) could be considered as a first of the kind.

The full fitting of the AudioCLIP parameters to AudioSet was shown to be beneficial in the regular sound classification setup. Similarly, the performance improvement on UrbanSound8K (65.31 % $\rightarrow$ **68.78 %**) follows the trend. Interestingly, the results on the ESC-50 dataset experienced a slight drop (**69.40 %** $\rightarrow$ 68.60 %), which might suggest that the model has ceased to distinguish certain fine-grained differences between sound classes. Such controversial results do not allow to confidently accept or reject the hypothesis $\mathcal{H}^{(4.2)}$.

Thus, enabling zero-shot capabilities in audio-centric deep learning models provides distinct advantages. The enhanced flexibility and adaptability in audio analysis tasks

Tab. 5.6: Zero-shot accuracy (%) on ESC-50 and UrbanSound8K.

Year	Model	Notes	Source	Zero-Shot Inference	Target	
					ESC	US8K
2019	VGGish		[272]	✓	26.00	–
2021			[273]	✓	33.00	–
2022	AudioCLIP	* IH, TH	[77]	✓	69.40	65.31
					96.65	89.95
				✓	68.60	68.78
					97.15	90.07

* – frozen weights.

Abbreviations: ESC: ESC-50 [200]; US8K: UrbanSound8K [227]; IH: Image-Head; TH: Text-Head.

empowers audio-centric models to generalize their knowledge and apply it to novel audio data, enabling more efficient and effective analysis in the environmental sound classification and associated applications. This makes zero-shot classification a powerful tool for advancing performance and versatility of audio-centric deep learning models.

5.5 Experimental Findings

In this chapter, there have been presented and discussed the results of the experimental work on improving deep learning models for audio analysis through various architectural enhancements. In Chapter 1, the main research question was decomposed into three subquestions, and this chapter has shown that addressing each of them leads to improvements in model performance and functionality. First, it was shown that using visual-domain models for solving audio-specific tasks contributes to the increase of performance. Second, the use of audio-specific architecture enhancements improves model performance and—as shown for fbsp-layer—provides non-adversarial robustness. Finally, it was demonstrated that moving from the transfer-learning-based setup to the audio-visual correspondence learning improves model performance and enables cross-modal retrieval capabilities, including zero-shot classification.

The obtained results indicate that these enhancements are advantageous in addressing the research questions $\mathcal{Q}^{(1)} - \mathcal{Q}^{(3)}$ and provide insights into how deep learning models can be improved for purposes of audio analysis. Specifically, the use of visual-domain models, audio-specific architecture enhancements, and knowledge distillation-based multimodal setup leads to better model performance and functionality, making it possible to solve challenging audio analysis tasks more effectively.

Conclusion

This thesis presents a comprehensive analysis of various methods and approaches for improving the performance of deep learning models in audio analysis, with a particular focus on multimodal learning.

In particular, it gives an overview of the deep learning landscape in the field of audio-visual learning. This overview depicts key aspects of learning objectives, including representation learning, correspondence learning, audio-visual object localization, sound source separation, and audio-visual generation. A selection of audio-related tasks includes audio-visual multimodal classification, regular and zero-shot environmental sound classification, and multimodal retrieval, and is naturally connected to the aforementioned objectives, thus building a foundation for understanding the current state-of-the-art in audio-visual deep learning.

With this knowledge, one could focus on specific techniques and methods to tackle the selected tasks, utilizing the insights gained from observing the state of the research field. The main research questions are addressed by exploring the advantages of using visual-domain models for solving audio-specific tasks, introducing audio-specific architecture enhancements of deep learning models, and shifting from the transfer-learning-based setup to the audio-visual contrastive learning. The experimental results demonstrate that these methods and approaches lead to significant improvements in performance and robustness.

As such, this work makes a diverse contribution to the field of audio analysis in multimodal learning. In this concluding chapter, the main findings of the study are summarized, and future research directions are outlined.

6.1 Transfer Learning from the Visual Domain

Transfer learning has emerged as a powerful technique in deep learning and has been widely used in various fields. In this thesis, the tasks of audio-visual multilabel classification and environmental sound classification have been used to evaluate the effect of using visual-domain models for solving audio-specific tasks. The experimental results

showcased significant improvements when integrating visual information into deep learning models processing audio. This performance increase became possible by leveraging representative features learned by visual-domain models, which had been pre-trained on the large-scale visual classification task. However, it is important to consider the potential limitations arising from the domain and task mismatch, which corresponds to the degree of similarity between the source (visual) and target (audio) setup. Bearing this in mind, techniques such as domain adaptation or domain-specific fine-tuning are to be further explored to bridge the domain gap and enhance the transfer of knowledge between modalities in a more effective manner.

Hence, for the first subquestion

$\mathcal{Q}^{(1)}$: “How does leveraging models from the visual domain contribute to performance in audio-specific tasks?”

the answer is

$\mathcal{A}^{(1)}$: “Leveraging models from the visual domain enhances performance in audio-specific tasks by using transfer learning mechanism.”

6.2 Audio-specific Architecture Enhancement

Audio-specific architecture enhancement is a special branch of audio-centric deep learning. By comparing the performance of audio-domain neural networks and standard models, the benefits of using audio-specific architecture enhancements become identifiable and measurable. Specifically, techniques such as learnable adaptive time-frequency transformation and attention mechanisms have shown excellent results in improving the performance on audio-specific tasks. The results of these comparisons provide insights into the design of future models for audio-specific tasks. In particular, tailoring a time-frequency transformation layer to the data increases accuracy in the ESC task, which suggests that focusing on improving robustness has a strong potential for contributing to the model performance. One important aspect to consider in these enhancements, however, is the computational complexity associated with calculating complex-valued gradients for the trainable time-frequency transformation by using real-valued automatic differentiation instead of applying the $\mathbb{C}$ -arithmetic. To soften the influence of this bottleneck, one has to consider implementing low-level complex-valued primitives or exploring real-valued transformation kernels. Overall, increased model performance and non-adversarial robustness are the main improvements introduced by the incorporation of audio-specific enhancements into the model architecture.

Thus, the second subquestion

$\mathcal{Q}^{(2)}$: “How can one leverage audio-specific modules within existing deep learning backbones to obtain more accurate and robust models?”

becomes the answer

$\mathcal{A}^{(2)}$: “Audio-specific architectural enhancements improve model performance and enable a non-adversarial robustness of multimodal models by tailoring the architecture to capture relevant audio features and patterns.”

6.3 Audio-visual Correspondence Learning

Solving a range of tasks in audio analysis, such as audio-visual multilabel classification, environmental sound classification, zero-shot sound classification, and multimodal retrieval, reveals benefits of shifting from transfer-learning-based to audio-visual correspondence learning. First, comparing the performance of transfer-learning- and knowledge-distillation-based models demonstrate the effectiveness of knowledge distillation in enhancing the performance of multimodal learning systems. Second, exploring the model performance in the zero-shot sound classification task allows for identifying generalization capabilities of knowledge-distillation-based models and their effectiveness in handling new and unseen data. Finally, evaluating the model performance in the multimodal retrieval task demonstrates the practical applications of the learned joint embedding space for tasks beyond classification. Although the sensitivity of models to the audio-visual alignment errors appears to be a significant limitation in specific setups [94], the use of advanced alignment refinement techniques can provide a powerful tool to mitigate it. The results obtained for the described tasks clearly outline that improved model performance and ability to perform cross-modal retrieval (including zero-shot setup) are the major benefits of making a multimodal model design follow the correspondence learning paradigm.

So, to the third subquestion

$\mathcal{Q}^{(3)}$: “Are there better training strategies beyond transfer learning for multimodal models?”

one can answer that

$\mathcal{A}^{(3)}$: “Knowledge distillation proves to be a superior training strategy compared to transfer learning, effectively enhancing the performance of multimodal models and enabling cross-modal as well as zero-shot retrieval capabilities.”

6.4 Research Outcome

The incorporation of audio-specific and audio-agnostic modifications into deep learning models leads to several improvements, answering the question of how to enhance these models. First, audio-centric models significantly benefit from advancements from the visual domain through transfer learning. Additionally, audio-specific enhancements introduce increased model performance and non-adversarial robustness to the model architecture. The results of tasks performed on these models show that the benefits of following the correspondence learning paradigm include improved model performance and the ability to perform cross-modal retrieval, including zero-shot setups.

Finally, the main question

$\mathcal{Q}$: “Can the accuracy and robustness of current multimodal models be improved by combining methods that are unique to the audio domain with existing audio-agnostic techniques?”

is answered as follows:

$\mathcal{A}$: “Combining audio-specific and audio-agnostic approaches improves performance of multimodal models, introduces non-adversarial robustness, and enables cross-modal and zero-shot retrieval capabilities.”

6.5 Outlook

While the current work focused on a clearly defined set of objectives, the related field of audio-centric models, including multimodal ones, is incredibly broad. All the discussed tasks and methods are worth the further research. However, the most prominent future directions need to be enumerated.

One potential direction is the improvement of the proposed fbsp-layer to make it a memory-efficient alternative to a fully-connected or convolutional layer. This, in turn,

allows for more efficient and accurate models that are better suited for audio as well as multimodal analysis tasks.

Another potential area of research is modifying the Transformer architecture to apply the compressed layer as a part of the attention mechanism, which has been shown to be highly effective for audio, image, and text data [7]. Such modification focuses on optimizing memory consumption and computational complexity of attention, thereby lightening multimodal neural networks and offering a way to tackle the challenge of rapid growth in the scale of models and tasks to be solved.

From the application point of view, employing the Transformer architecture to process audio data without an intermediate time-frequency transformation is another promising direction for future research [255]. This will require development of novel audio-specific techniques for representing and analyzing audio data, but opens up new perspectives for highly effective models for audio-centric tasks.

Finally, extending the proposed AudioCLIP model to videos is a further prominent area for future research. Although preserving cross-modal temporal alignment poses non-trivial challenges [81, 260], the expected performance gain as well as qualitative extension of the model capabilities are a strong argument in favor of such a modification.

In conclusion, there are many promising avenues for follow up developments in neural networks for multimodal and audio analysis, and continued advancements are expected in this exciting and rapidly evolving field [214]. This thesis was devoted to investigating methods to enhance the training workflow and refine the model architecture for audio analysis in uni- and multimodal scenarios, which is a small but important part of the challenges of conducting research in the area of audio-centric multimodal deep learning, and has been written with the aim of providing a solid building block for future studies.

Bibliography

- [1] Triantafyllos Afouras, Yuki M Asano, Francois Fagan, Andrea Vedaldi, and Florian Metze. “Self-supervised object detection from audio-visual correspondence”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10575–10586.
- [2] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. “The conversation: Deep audio-visual speech enhancement”. In: *arXiv preprint arXiv:1804.04121* (2018).
- [3] Triantafyllos Afouras, Andrew Owens, Joon Son Chung, and Andrew Zisserman. “Self-supervised learning of audio-visual objects from video”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 208–224.
- [4] Madhav Agarwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. “Audio-visual face reenactment”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 5178–5187.
- [5] Dharmesh M Agrawal, Hardik B Sailor, Meet H Soni, and Hemant A Patil. “Novel TEO-based Gammatone features for environmental sound classification”. In: *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE. 2017, pp. 1809–1813.
- [6] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani. “Lip2AudSpec: Speech reconstruction from silent lip movements video”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 2516–2520.
- [7] Hassan Akbari, Linagzhe Yuan, Rui Qian, et al. “VATT: Transformers for multimodal self-supervised learning from raw video, audio and text”. In: *arXiv preprint arXiv:2104.11178* (2021).
- [8] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, et al. “Self-supervised multi-modal versatile networks”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 25–37.
- [9] MA Ali and PM Shemi. “An improved method of audio denoising based on wavelet transform”. In: *2015 international conference on Power, Instrumentation, Control and Computing (PICC)*. IEEE. 2015, pp. 1–6.
- [10] J. Allen. “Short term spectral analysis, synthesis, and modification by discrete Fourier transform”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 25.3 (June 1977), pp. 235–238.
- [11] Jonathan Allen. “Short term spectral analysis, synthesis, and modification by discrete Fourier transform”. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 25.3 (1977), pp. 235–238.

- [12]Humam Alwassel, Dhruv Mahajan, Bruno Korbar, et al. “Self-supervised learning by cross-modal audio-video clustering”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9758–9770.
- [13]Shahin Amiriparian. “Deep representation learning techniques for audio signal processing”. PhD thesis. Technische Universität München, 2019.
- [14]Shahin Amiriparian, Michael Freitag, Nicholas Cummins, and Björn Schuller. “Sequence to sequence autoencoders for unsupervised representation learning from audio”. In: *DCASE*. 2017, pp. 17–21.
- [15]Augustin Arnault, Baptiste Hanssens, and Nicolas Riche. *Urban Sound Classification: striving towards a fair comparison*. 2020. arXiv: 2010.11805 [cs.SD].
- [16]Yuki Asano, Mandela Patrick, Christian Rupprecht, and Andrea Vedaldi. “Labelling unlabelled videos from scratch with multi-modal self-supervision”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 4660–4671.
- [17]Yekutiel Avargel and Israel Cohen. “System identification in the short-time Fourier transform domain with crossband filtering”. In: *IEEE transactions on Audio, Speech, and Language processing* 15.4 (2007), pp. 1305–1319.
- [18]Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. “Layer normalization”. In: *arXiv preprint arXiv:1607.06450* (2016).
- [19]Tatsuro Baba. “Time-frequency analysis using short time Fourier transform”. In: *The Open Acoustics Journal* 5.1 (2012).
- [20]Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. “Multimodal machine learning: A survey and taxonomy”. In: *IEEE transactions on pattern analysis and machine intelligence* 41.2 (2018), pp. 423–443.
- [21]Joshua Bassey, Lijun Qian, and Xianfang Li. “A survey of complex-valued neural networks”. In: *arXiv preprint arXiv:2101.12249* (2021).
- [22]Yoshua Bengio, Aaron Courville, and Pascal Vincent. “Representation learning: A review and new perspectives”. In: *IEEE transactions on pattern analysis and machine intelligence* 35.8 (2013), pp. 1798–1828.
- [23]Venkatesh Boddapati, Andrej Petef, Jim Rasmusson, and Lars Lundberg. “Classifying environmental sounds using image recognition networks”. In: *Procedia computer science* 112 (2017), pp. 2048–2056.
- [24]Carlo Cattani et al. “Shannon wavelets for the solution of integrodifferential equations”. In: *Mathematical Problems in Engineering* 2010 (2010).
- [25]Maitreya Chatterjee, Narendra Ahuja, and Anoop Cherian. “Learning Audio-Visual Dynamics Using Scene Graphs for Audio Source Separation”. In: *arXiv preprint arXiv:2210.16472* (2022).
- [26]Maitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. “Visual scene graphs for audio source separation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 1204–1213.

- [27]Honglie Chen, Weidi Xie, Triantafyllos Afouras, et al. “Localizing visual sounds the hard way”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 16867–16876.
- [28]Kan Chen, Chuanxi Zhang, Chen Fang, et al. “Visually indicated sound generation by perceptually optimized classification”. In: *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 2018.
- [29]Lele Chen, Guofeng Cui, Ziyi Kou, Haitian Zheng, and Chenliang Xu. “What comprises a good talking-head video generation?: A survey and benchmark”. In: *arXiv preprint arXiv:2005.03201* (2020).
- [30]Peihao Chen, Yang Zhang, Mingkui Tan, et al. “Generating visually aligned sound from videos”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 8292–8302.
- [31]Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. “A simple framework for contrastive learning of visual representations”. In: *International conference on machine learning*. PMLR. 2020, pp. 1597–1607.
- [32]Yanbei Chen, Yongqin Xian, A Koepke, Ying Shan, and Zeynep Akata. “Distilling audio-visual knowledge by compositional contrastive learning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 7016–7025.
- [33]Benjamin Chidester, Minh N Do, and Jian Ma. “Rotation equivariance and invariance in convolutional neural networks”. In: *arXiv preprint arXiv:1805.12301* (2018).
- [34]François Chollet. “Xception: Deep learning with depthwise separable convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 1251–1258.
- [35]Jan Chorowski, Ron J Weiss, Samy Bengio, and Aäron Van Den Oord. “Unsupervised speech representation learning using wavenet autoencoders”. In: *IEEE/ACM transactions on audio, speech, and language processing* 27.12 (2019), pp. 2041–2053.
- [36]Charles K Chui and Jianzhong Wang. “Shannon wavelet approach to sub-band coding”. In: *International Journal of Wavelets, Multiresolution and Information Processing* 1.02 (2003), pp. 233–242.
- [37]Yu-An Chung, Wei-Ning Hsu, Hao Tang, and James Glass. “An unsupervised autoregressive model for speech representation learning”. In: *arXiv preprint arXiv:1904.03240* (2019).
- [38]Soo-Whan Chung, Soyeon Choe, Joon Son Chung, and Hong-Goo Kang. “FaceFilter: Audio-visual speech separation using still images”. In: *arXiv preprint arXiv:2005.07074* (2020).
- [39]Tim Cooijmans, Nicolas Ballas, César Laurent, Çağlar Gülçehre, and Aaron Courville. “Recurrent batch normalization”. In: *arXiv preprint arXiv:1603.09025* (2016).
- [40]James W Cooley and John W Tukey. “An algorithm for the machine calculation of complex Fourier series”. In: *Mathematics of computation* 19.90 (1965), pp. 297–301.
- [41]Chenye Cui, Yi Ren, Jinglin Liu, Rongjie Huang, and Zhou Zhao. “VarietySound: Timbre-Controllable Video to Sound Generation via Unsupervised Information Disentanglement”. In: *arXiv preprint arXiv:2211.10666* (2022).

- [42]Stéphane d'Ascoli, Hugo Touvron, Matthew L Leavitt, et al. "ConViT: Improving vision transformers with soft convolutional inductive biases". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 2286–2296.
- [43]Ingrid Daubechies. "The wavelet transform, time-frequency localization and signal analysis". In: *IEEE transactions on information theory* 36.5 (1990), pp. 961–1005.
- [44]Shohreh Deldari, Hao Xue, Aaqib Saeed, et al. "Beyond Just Vision: A Review on Self-Supervised Representation Learning on Multimodal and Temporal Data". In: *arXiv preprint arXiv:2206.02353* (2022).
- [45]J. Deng, W. Dong, R. Socher, et al. "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR09*. 2009.
- [46]Li Deng, Geoffrey Hinton, and Brian Kingsbury. "New types of deep neural network learning for speech recognition and related applications: An overview". In: *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE. 2013, pp. 8599–8603.
- [47]Xiang Deng and Zhongfei Zhang. "Comprehensive knowledge distillation with causal intervention". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 22158–22170.
- [48]Hao-Wen Dong, Wen-Yi Hsiao, Li-Chia Yang, and Yi-Hsuan Yang. "MuseGAN: Multi-track sequential generative adversarial networks for symbolic music generation and accompaniment". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.
- [49]Xingping Dong and Jianbing Shen. "Triplet loss in siamese network for object tracking". In: *Proceedings of the European conference on computer vision (ECCV)*. 2018, pp. 459–474.
- [50]Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).
- [51]Timothy Dozat. "Incorporating Nesterov Momentum into Adam". In: *Proceedings of the 4th International Conference on Learning Representations*. 2016, pp. 1–4.
- [52]Ahmet Elbir, Hamza Osman İlhan, Gorkem Serbes, and Nizamettin Aydın. "Short Time Fourier Transform based music genre classification". In: *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*. IEEE. 2018, pp. 1–4.
- [53]Dennis Elbrächter, Dmytro Perekrestenko, Philipp Grohs, and Helmut Bölcskei. "Deep neural network approximation theory". In: *IEEE Transactions on Information Theory* 67.5 (2021), pp. 2581–2623.
- [54]Ariel Ephrat, Inbar Mosseri, Oran Lang, et al. "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation". In: *arXiv preprint arXiv:1804.03619* (2018).
- [55]Linus Ericsson, Henry Gouk, Chen Change Loy, and Timothy M Hospedales. "Self-Supervised Representation Learning: Introduction, advances, and challenges". In: *IEEE Signal Processing Magazine* 39.3 (2022), pp. 42–62.

- [56]Xavier Favory, Konstantinos Drossos, Tuomas Virtanen, and Xavier Serra. “COALA: Co-aligned autoencoders for learning semantically enriched audio representations”. In: *arXiv preprint arXiv:2006.08386* (2020).
- [57]Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. “Unsupervised visual domain adaptation using subspace alignment”. In: *Proceedings of the IEEE international conference on computer vision*. 2013, pp. 2960–2967.
- [58]Michael Freitag, Shahin Amiriparian, Sergey Pugachevskiy, Nicholas Cummins, and Björn Schuller. “auDeep: Unsupervised learning of representations from audio with deep recurrent neural networks”. In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 6340–6344.
- [59]Aviv Gabbay, Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. “Seeing through noise: Visually driven speaker separation and enhancement”. In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, pp. 3051–3055.
- [60]Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. “A survey on deep learning for multimodal data fusion”. In: *Neural Computation* 32.5 (2020), pp. 829–864.
- [61]Liang Gao, Kele Xu, Huaimin Wang, and Yuxing Peng. “Multi-representation knowledge distillation for audio classification”. In: *Multimedia Tools and Applications* 81.4 (2022), pp. 5089–5112.
- [62]Rui Gao, Xingsong Hou, Jie Qin, et al. “Zero-VAE-GAN: Generating unseen features for generalized and transductive zero-shot learning”. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 3665–3680.
- [63]Ruohan Gao and Kristen Grauman. “VisualVoice: Audio-visual speech separation with cross-modal consistency”. In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2021, pp. 15490–15500.
- [64]Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. “Listen to look: Action recognition by previewing audio”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 10457–10467.
- [65]Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, et al. “Audio Set: An ontology and human-labeled dataset for audio events”. In: *Proc. IEEE ICASSP 2017*. New Orleans, LA, 2017.
- [66]Hossein Gholamalinezhad and Hossein Khosravi. “Pooling methods in deep neural networks, a review”. In: *arXiv preprint arXiv:2009.07485* (2020).
- [67]Damian Gogolewski. “Influence of the edge effect on the wavelet analysis process”. In: *Measurement* 152 (2020), p. 107314.
- [68]Yuan Gong, Yu-An Chung, and James Glass. “AST: Audio Spectrogram Transformer”. In: *arXiv preprint arXiv:2104.01778* (2021). arXiv: 2104.01778 [cs.LG].
- [69]Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, et al. “Generative adversarial networks”. In: *Communications of the ACM* 63.11 (2020), pp. 139–144.

- [70]Alexander N Gorban and Ivan Yu Tyukin. “Blessing of dimensionality: mathematical foundations of the statistical physics of data”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2118 (2018), p. 20170237.
- [71]Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. “Knowledge distillation: A survey”. In: *International Journal of Computer Vision* 129 (2021), pp. 1789–1819.
- [72]Rongzhi Gu, Shi-Xiong Zhang, Yong Xu, et al. “Multi-modal multi-channel target speech separation”. In: *IEEE Journal of Selected Topics in Signal Processing* 14.3 (2020), pp. 530–541.
- [73]Yushuo Guan, Pengyu Zhao, Bingxuan Wang, et al. “Differentiable feature aggregation search for knowledge distillation”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII* 16. Springer. 2020, pp. 469–484.
- [74]Nitzan Guberman. “On complex valued convolutional neural networks”. In: *arXiv preprint arXiv:1602.09046* (2016).
- [75]Mao Guo, Chenghu Zhou, and Jiahang Liu. “Jointly learning of visual and auditory: A new approach for RS image and audio cross-modal retrieval”. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 12.11 (2019), pp. 4644–4654.
- [76]Yudong Guo, Keyu Chen, Sen Liang, et al. “AD-NeRF: Audio driven neural radiance fields for talking head synthesis”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 5784–5794.
- [77]Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. “AudioCLIP: Extending CLIP to Image, Text and Audio”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 976–980.
- [78]Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. “ESResNe(X)t-fb: Learning Robust Time-Frequency Transformation of Audio”. In: *2021 International Joint Conference on Neural Networks (IJCNN)*. 2021.
- [79]Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. “ESResNet: Environmental Sound Classification Based on Visual Domain Models”. In: *25th International Conference on Pattern Recognition (ICPR)*. Milan, Italy, Jan. 2021, pp. 4933–4940.
- [80]Erik E. Guzik, Christian Byrge, and Christian Gilde. “The originality of machines: AI takes the Torrance Test”. In: *Journal of Creativity* 33.3 (2023), p. 100065.
- [81]Tavi Halperin, Ariel Ephrat, and Shmuel Peleg. “Dynamic temporal alignment of speech to lips”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 3980–3984.
- [82]Jiequn Han, Arnulf Jentzen, and Weinan E. “Solving high-dimensional partial differential equations using deep learning”. In: *Proceedings of the National Academy of Sciences* 115.34 (2018), pp. 8505–8510.
- [83]Fredric J Harris. “On the use of windows for harmonic analysis with the discrete Fourier transform”. In: *Proceedings of the IEEE* 66.1 (1978), pp. 51–83.

- [84]Anfeng He, Chong Luo, Xinmei Tian, and Wenjun Zeng. “A twofold siamese network for real-time object tracking”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 4834–4843.
- [85]Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep Residual Learning for Image Recognition”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016.
- [86]Gerhard Heinzel, Albrecht Rüdiger, and Roland Schilling. *Spectrum and spectral density estimation by the Discrete Fourier transform (DFT), including a comprehensive list of window functions and some new flat-top windows*. Tech. rep. Max-Planck-Institut für Gravitationsphysik, 2002.
- [87]Dan Hendrycks and Kevin Gimpel. “Gaussian error linear units (gelus)”. In: *arXiv preprint arXiv:1606.08415* (2016).
- [88]Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, et al. “CNN architectures for large-scale audio classification”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 131–135.
- [89]Geoffrey Hinton. “The forward-forward algorithm: Some preliminary investigations”. In: *arXiv preprint arXiv:2212.13345* (2022).
- [90]Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. “Neural networks for machine learning lecture 6a overview of mini-batch gradient descent”. In: (2012).
- [91]Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. “Distilling the knowledge in a neural network”. In: *arXiv preprint arXiv:1503.02531* (2015).
- [92]Sebastian Hofstätter, Sophia Althammer, Michael Schröder, Mete Sertkan, and Allan Hanbury. “Improving efficient neural ranking models with cross-architecture knowledge distillation”. In: *arXiv preprint arXiv:2010.02666* (2020).
- [93]Kurt Hornik, Maxwell Stinchcombe, and Halbert White. “Multilayer feedforward networks are universal approximators”. In: *Neural networks* 2.5 (1989), pp. 359–366.
- [94]Yuanbo Hou, Bo Kang, and Dick Botteldooren. “Audio-visual scene classification via contrastive event-object alignment and semantic-based fusion”. In: *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*. IEEE. 2022, pp. 1–6.
- [95]Di Hu, Rui Qian, Minyue Jiang, et al. “Discriminative sounding objects localization via self-supervised audiovisual matching”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 10077–10087.
- [96]Di Hu, Zheng Wang, Haoyi Xiong, et al. “Curriculum audiovisual learning”. In: *arXiv preprint arXiv:2001.09414* (2020).
- [97]Di Hu, Yake Wei, Rui Qian, et al. “Class-aware sounding objects localization via audiovisual correspondence”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2021).
- [98]Vladimir Iashin and Esa Rahtu. “Taming visually guided sound generation”. In: *arXiv preprint arXiv:2110.08791* (2021).

- [99]Hidenori Ide and Takio Kurita. “Improvement of learning for CNN with ReLU activation by sparse regularization”. In: *2017 international joint conference on neural networks (IJCNN)*. IEEE. 2017, pp. 2684–2691.
- [100]Sergey Ioffe and Christian Szegedy. “Batch normalization: Accelerating deep network training by reducing internal covariate shift”. In: *International conference on machine learning*. pmlr. 2015, pp. 448–456.
- [101]Amir Jamaludin, Joon Son Chung, and Andrew Zisserman. “You said that?: Synthesising talking faces from audio”. In: *International Journal of Computer Vision* 127.11 (2019), pp. 1767–1779.
- [102]Aren Jansen, Manoj Plakal, Ratheet Pandya, et al. “Unsupervised learning of semantic audio representations”. In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2018, pp. 126–130.
- [103]Arindam Jati, Naveen Kumar, Ruxin Chen, and Panayiotis Georgiou. “Hierarchy-aware loss function on a tree structured label space for audio event detection”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 6–10.
- [104]Xinya Ji, Hang Zhou, Kaisiyuan Wang, et al. “Audio-driven emotional video portraits”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 14080–14089.
- [105]Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai. “Music type classification by spectral contrast feature”. In: *Proceedings. IEEE International Conference on Multimedia and Expo*. Vol. 1. IEEE. 2002, pp. 113–116.
- [106]Qin Jin and Junwei Liang. “Video Description Generation Using Audio and Visual Cues”. In: *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ICMR '16. New York, New York, USA: Association for Computing Machinery, 2016, pp. 239–242.
- [107]Longlong Jing and Yingli Tian. “Self-supervised visual feature learning with deep neural networks: A survey”. In: *IEEE transactions on pattern analysis and machine intelligence* 43.11 (2020), pp. 4037–4058.
- [108]J. F. Kaiser. “Some useful properties of Teager’s energy operators”. In: *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vol. 3. 1993, 149–152 vol.3.
- [109]M Davidson Kamaladas and M Maxina Dialin. “Fingerprint extraction of audio signal using wavelet transform”. In: *2013 International Conference on Signal Processing, Image Processing & Pattern Recognition*. IEEE. 2013, pp. 308–312.
- [110]Samsul Ariffin Abdul Karim, Mohd Hafizi Kamarudin, Bakri Abdul Karim, Mohammad Khatim Hasan, and Jumat Sulaiman. “Wavelet Transform and Fast Fourier Transform for signal compression: A comparative study”. In: *2011 International Conference on Electronic Devices, Systems and Applications (ICEDSA)*. IEEE. 2011, pp. 280–285.
- [111]Phuc H Le-Khac, Graham Healy, and Alan F Smeaton. “Contrastive representation learning: A framework and review”. In: *IEEE Access* 8 (2020), pp. 193907–193934.

- [112]Salman Khan, Muzammal Naseer, Munawar Hayat, et al. “Transformers in vision: A survey”. In: *ACM computing surveys (CSUR)* 54.10s (2022), pp. 1–41.
- [113]Varun Khare, Divyat Mahajan, Homanga Bharadhwaj, Vinay Kumar Verma, and Piyush Rai. “A generative framework for zero shot learning with adversarial domain adaptation”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, pp. 3101–3110.
- [114]Jack Kiefer and Jacob Wolfowitz. “Stochastic estimation of the maximum of a regression function”. In: *The Annals of Mathematical Statistics* (1952), pp. 462–466.
- [115]Chanwoo Kim, Anjali Menon, Michiel Bacchiani, and Richard Stern. “Sound source separation using phase difference and reliable mask selection selection”. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2018, pp. 5559–5563.
- [116]Jangho Kim, Minsung Hyun, Inseop Chung, and Nojun Kwak. “Feature fusion for on-line mutual knowledge distillation”. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 4619–4625.
- [117]Sunwoo Kim and Minje Kim. “Test-time adaptation toward personalized speech enhancement: Zero-shot learning with knowledge distillation”. In: *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 2021, pp. 176–180.
- [118]Taesup Kim, Inchul Song, and Yoshua Bengio. “Dynamic layer normalization for adaptive neural acoustic modeling in speech recognition”. In: *arXiv preprint arXiv:1707.06065* (2017).
- [119]Diederik P Kingma and Jimmy Ba. “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980* (2014).
- [120]Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. “Siamese neural networks for one-shot image recognition”. In: *ICML deep learning workshop*. Vol. 2. Lille. 2015.
- [121]Elyor Kodirov, Tao Xiang, Zhenyong Fu, and Shaogang Gong. “Unsupervised domain adaptation for zero-shot learning”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2452–2460.
- [122]Qiuqiang Kong, Yin Cao, Turab Iqbal, et al. “PANNs: Large-scale pretrained audio neural networks for audio pattern recognition”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020), pp. 2880–2894.
- [123]Qiuqiang Kong, Changsong Yu, Yong Xu, et al. “Weakly labelled audioset tagging with attention neural networks”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.11 (2019), pp. 1791–1802.
- [124]Bruno Korbar, Du Tran, and Lorenzo Torresani. “Cooperative learning of audio and video models from self-supervised synchronization”. In: *Advances in Neural Information Processing Systems* 31 (2018).
- [125]Anurag Kumar and Vamsi Ithapu. “A Sequential Self Teaching Approach for Improving Generalization in Sound Event Recognition”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 5447–5457.

- [126]Neeraj Kumar, Srishti Goel, Ankur Narang, and Mujtaba Hasan. “Robust one shot audio to video generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, pp. 770–771.
- [127]Neeraj Kumar, Srishti Goel, Ankur Narang, and Brejesh Lall. “Multi Modal Adaptive Normalization for Audio to Video Generation”. In: *arXiv preprint arXiv:2012.07304* (2020).
- [128]Tryphon Lambrou, Panos Kudumakis, R Speller, M Sandler, and A Linney. “Classification of audio signals using statistical features on time and wavelet transform domains”. In: *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP’98 (Cat. No. 98CH36181)*. Vol. 6. IEEE. 1998, pp. 3621–3624.
- [129]Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. “Attribute-based classification for zero-shot visual object categorization”. In: *IEEE transactions on pattern analysis and machine intelligence* 36.3 (2013), pp. 453–465.
- [130]Yann LeCun, Bernhard Boser, John S Denker, et al. “Backpropagation applied to handwritten zip code recognition”. In: *Neural computation* 1.4 (1989), pp. 541–551.
- [131]Yann LeCun, Léon Bottou, Genevieve B Orr, and Klaus-Robert Müller. “Efficient backprop”. In: *Neural networks: Tricks of the trade*. Springer, 2002, pp. 9–50.
- [132]Jiyoung Lee, Soo-Whan Chung, Sunok Kim, Hong-Goo Kang, and Kwanghoon Sohn. “Looking into your speech: Learning cross-modal affinity for audio-visual speech separation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 1336–1345.
- [133]Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, et al. “Sound-Guided Semantic Video Generation”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 34–50.
- [134]Yuan-Shan Lee, Chien-Yao Wang, Shu-Fan Wang, Jia-Ching Wang, and Chung-Hsien Wu. “Fully complex deep neural network for phase-incorporating monaural source separation”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 281–285.
- [135]Chenda Li and Yanmin Qian. “Deep audio-visual speech separation with attention mechanism”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 7314–7318.
- [136]Hualiang Li and Tülay Adalı. “Complex-valued adaptive signal processing using nonlinear functions”. In: *EURASIP Journal on Advances in Signal Processing* 2008 (2008), pp. 1–9.
- [137]Sizhe Li, Yapeng Tian, and Chenliang Xu. “Space-Time Memory Network for Sounding Object Localization in Videos”. In: *arXiv preprint arXiv:2111.05526* (2021).
- [138]Zhiyuan Li and Sanjeev Arora. “An exponential learning rate schedule for deep learning”. In: *arXiv preprint arXiv:1910.07454* (2019).
- [139]Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. “Focal loss for dense object detection”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 2980–2988.
- [140]Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. “ECLIPSE: Efficient Long-range Video Retrieval using Sight and Sound”. In: *arXiv preprint arXiv:2204.02874* (2022).

- [141]Yan-Bo Lin and Yu-Chiang Frank Wang. “Audiovisual transformer with instance attention for audio-visual event localization”. In: *Proceedings of the Asian Conference on Computer Vision*. 2020.
- [142]Jie Liu. “Shannon wavelet spectrum analysis on truncated vibration signals for machine incipient fault detection”. In: *Measurement Science and Technology* 23.5 (2012), p. 055604.
- [143]Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, et al. “Audio self-supervised learning: A survey”. In: *Patterns* 3.12 (2022), p. 100616.
- [144]Song Liu, Haoqi Fan, Shengsheng Qian, et al. “HiT: Hierarchical transformer with momentum contrast for video-text retrieval”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 11915–11925.
- [145]Xian Liu, Yinghao Xu, Qianyi Wu, et al. “Semantic-aware implicit neural audio-driven video portrait generation”. In: *arXiv preprint arXiv:2201.07786* (2022).
- [146]Yang Liu, Alexandras Neophytou, Sunando Sengupta, and Eric Sommerlade. “Cross-modal spectrum transformation network for acoustic scene classification”. In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2021, pp. 830–834.
- [147]Yongcheng Liu, Lu Sheng, Jing Shao, et al. “Multi-label image classification via knowledge distillation from weakly-supervised detection”. In: *Proceedings of the 26th ACM international conference on Multimedia*. 2018, pp. 700–708.
- [148]Yongfei Liu, Chenfei Wu, Shao-yen Tseng, et al. “KD-VLP: Improving end-to-end vision-and-language pretraining with object knowledge distillation”. In: *arXiv preprint arXiv:2109.10504* (2021).
- [149]Yufan Liu, Jiajiong Cao, Bing Li, et al. “Cross-architecture knowledge distillation”. In: *Proceedings of the Asian Conference on Computer Vision*. 2022, pp. 3396–3411.
- [150]Anna Llagostera Casanovas and Andrea Cavallaro. “Audio-visual events for multi-camera synchronization”. In: *Multimedia Tools and Applications* 74.4 (2015), pp. 1317–1340.
- [151]Beth Logan et al. “Mel frequency cepstral coefficients for music modeling.” In: *Proceeding of the International Symposium on Music Information Retrieval (ISMIR)*. Plymouth, USA, Oct. 2000.
- [152]Ilya Loshchilov and Frank Hutter. “Decoupled weight decay regularization”. In: *arXiv preprint arXiv:1711.05101* (2017).
- [153]Rui Lu, Zhiyao Duan, and Changshui Zhang. “Audio–visual deep clustering for speech separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.11 (2019), pp. 1697–1712.
- [154]Rui Lu, Zhiyao Duan, and Changshui Zhang. “Listen and look: Audio–visual matching assisted speech source separation”. In: *IEEE Signal Processing Letters* 25.9 (2018), pp. 1315–1319.
- [155]Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. “Active contrastive learning of audio-visual video representations”. In: *arXiv preprint arXiv:2009.09805* (2020).

- [156] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. “Contrastive learning of global-local video representations”. In: *arXiv preprint arXiv:2104.05418* (2021).
- [157] Wentao Ma, Qingchao Chen, Tongqing Zhou, Shan Zhao, and Zhiping Cai. “Using Multi-modal Contrastive Knowledge Distillation for Video-Text Retrieval”. In: *IEEE Transactions on Circuits and Systems for Video Technology* (2023).
- [158] Sagnik Majumder, Ziad Al-Halah, and Kristen Grauman. “Move2hear: Active audio-visual source separation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 275–285.
- [159] Pedro Henrique Martins, Zita Marinho, and André FT Martins. “ ∞ -former: Infinite Memory Transformer”. In: *arXiv preprint arXiv:2109.00301* (2021).
- [160] Norbert Marwan, Niels Wessel, Udo Meyerfeldt, Alexander Schirdewan, and Jürgen Kurths. “Recurrence-plot-based measures of complexity and their application to heart-rate-variability data”. In: *Physical review E* 66.2 (2002), p. 026702.
- [161] Yoshiki Masuyama, Yoshiaki Bando, Kohei Yatabe, et al. “Self-supervised neural audio-visual sound source localization via probabilistic spatial modeling”. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2020, pp. 4848–4854.
- [162] Pratik Mazumder, Pravendra Singh, Kranti Kumar Parida, and Vinay P Namboodiri. “AVGZSLNet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021, pp. 3090–3099.
- [163] Iaroslav Melekhov, Juho Kannala, and Esa Rahtu. “Siamese network features for image matching”. In: *2016 23rd international conference on pattern recognition (ICPR)*. IEEE. 2016, pp. 378–383.
- [164] Anil Menon, Kishan Mehrotra, Chilukuri K Mohan, and Sanjay Ranka. “Characterization of a class of sigmoid functions with applications to neural networks”. In: *Neural networks* 9.5 (1996), pp. 819–835.
- [165] Otniel-Bogdan Mercea, Lukas Riesch, A Koepke, and Zeynep Akata. “Audio-visual generalised zero-shot learning with cross-modal attention and language”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 10553–10563.
- [166] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, et al. “An overview of deep-learning-based audio-visual speech enhancement and separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 1368–1396.
- [167] Dongchan Min, Minyoung Song, and Sung Ju Hwang. “StyleTalker: One-shot Style-based Audio-driven Talking Head Video Generation”. In: *arXiv preprint arXiv:2208.10922* (2022).
- [168] Benjamin R Mitchell et al. “The spatial inductive bias of deep learning”. PhD thesis. Johns Hopkins University, 2017.
- [169] Mandar Mitra and BB Chaudhuri. “Information retrieval from documents: A survey”. In: *Information retrieval* 2 (2000), pp. 141–163.

- [170] Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. “Emotions don’t lie: An audio-visual deepfake detection method using affective cues”. In: *Proceedings of the 28th ACM international conference on multimedia*. 2020, pp. 2823–2832.
- [171] Aleksandra Mojsilovic, Miodrag V Popovic, and Dejan M Rackov. “On the selection of an optimal wavelet basis for texture characterization”. In: *IEEE Transactions on image processing* 9.12 (2000), pp. 2043–2050.
- [172] Pedro Morgado, Nuno Nvasconcelos, Timothy Langlois, and Oliver Wang. “Self-supervised generation of spatial audio for 360 video”. In: *Advances in neural information processing systems* 31 (2018).
- [173] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. “Audio-visual instance discrimination with cross-modal agreement”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 12475–12486.
- [174] Xiaomin Mou. “Artificial intelligence: Investment trends and selected industry uses”. In: *International Finance Corporation* 8 (2019).
- [175] Arsha Nagrani, Shan Yang, Anurag Arnab, et al. “Attention bottlenecks for multimodal fusion”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 14200–14213.
- [176] Vinod Nair and Geoffrey E Hinton. “Rectified linear units improve restricted boltzmann machines”. In: *Proceedings of the 27th international conference on machine learning (ICML-10)*. 2010, pp. 807–814.
- [177] Anastasia Natsiou and Seán O’Leary. “Audio representations for deep learning in sound synthesis: A review”. In: *2021 IEEE/ACS 18th International Conference on Computer Systems and Applications (AICCSA)*. IEEE. 2021, pp. 1–8.
- [178] Yu Nesterov. “A method of solving a convex programming problem with convergence rate $O(1/k^2)$ ”. In: *Sov. Math. Dokl.* Vol. 27. 2. 1983.
- [179] Andrew Ng and Michael Jordan. “On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes”. In: *Advances in neural information processing systems* 14 (2001).
- [180] Wai Keng Ngui, M Salman Leong, Lim Meng Hee, and Ahmed M Abdelrhman. “Wavelet analysis: mother wavelet selection methods”. In: *Applied mechanics and materials* 393 (2013), pp. 953–958.
- [181] Anh Nguyen, Khoa Pham, Dat Ngo, Thanh Ngo, and Lam Pham. “An analysis of state-of-the-art activation functions for supervised deep neural network”. In: *2021 International Conference on System Science and Engineering (ICSSE)*. IEEE. 2021, pp. 215–220.
- [182] Tokuhiro Nishikawa, Daiki Shimada, and Jerry Jun Yokono. “Dual Normalization Multitasking for Audio-Visual Sounding Object Localization”. In: *arXiv preprint arXiv:2106.00180* (2021).
- [183] Emilio Soria Olivas, Jos David Mart Guerrero, Marcelino Martinez-Sober, Jose Rafael Magdalena-Benedito, L Serrano, et al. *Handbook of research on machine learning applications and trends: Algorithms, methods, and techniques: Algorithms, methods, and techniques*. IGI global, 2009, pp. 242–264.

- [184] Andreea-Maria Oncescu, A Koepke, Joao F Henriques, Zeynep Akata, and Samuel Albanie. “Audio retrieval with natural language queries”. In: *arXiv preprint arXiv:2105.02192* (2021).
- [185] Jonah Ong, Ba Tuong Vo, Sven Nordholm, et al. “Audio-Visual Based Online Multi-Source Separation”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 30 (2022), pp. 1219–1234.
- [186] Alan V Oppenheim. *Discrete-time signal processing*. Pearson Education India, 1999.
- [187] Rasmus Pagh. “Compressed matrix multiplication”. In: *ACM Transactions on Computation Theory (TOCT)* 5.3 (2013), pp. 1–17.
- [188] Junhyeong Pak and Jong Won Shin. “Sound localization based on phase difference enhancement using deep neural networks”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.8 (2019), pp. 1335–1345.
- [189] Kamalesh Palanisamy, Dipika Singhania, and Angela Yao. *Rethinking CNN Models for Audio Classification*. 2020. arXiv: 2007.11154 [cs.CV].
- [190] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. “Zero-shot learning with semantic output codes”. In: *Advances in neural information processing systems* 22 (2009).
- [191] Sinno Jialin Pan, Xiaochuan Ni, Jian-Tao Sun, Qiang Yang, and Zheng Chen. “Cross-domain sentiment classification via spectral feature alignment”. In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 751–760.
- [192] Sanjeel Parekh, Slim Essid, Alexey Ozerov, et al. “Weakly supervised representation learning for audio-visual scene analysis”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2019), pp. 416–428.
- [193] Sanjeel Parekh, Alexey Ozerov, Slim Essid, et al. “Identify, locate and separate: Audio-visual object extraction in large video collections using weak supervision”. In: *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 2019, pp. 268–272.
- [194] Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. “Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2020, pp. 3251–3260.
- [195] Cesare V Parise and Marc O Ernst. “Correlation detection as a general mechanism for multisensory integration”. In: *Nature communications* 7.1 (2016), p. 11543.
- [196] Adam Paszke, Sam Gross, Francisco Massa, et al. “PyTorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32 (2019).
- [197] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, et al. “Multi-modal self-supervision from generalized data transformations”. In: *arXiv preprint arXiv:2003.04298* (2020).

- [198]Geoffroy Peeters. “Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal”. In: *IEEE Transactions on Audio, Speech, and Language Processing* 19.5 (2010), pp. 1242–1252.
- [199]K. J. Piczak. “Environmental sound classification with convolutional neural networks”. In: *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*. Sept. 2015, pp. 1–6.
- [200]Karol J. Piczak. “ESC: Dataset for Environmental Sound Classification”. In: *Proceedings of the 23rd ACM International Conference on Multimedia*. MM ’15. Brisbane, Australia: Association for Computing Machinery, 2015, pp. 1015–1018.
- [201]R Gnana Praveen, Eric Granger, and Patrick Cardinal. “Cross attentional audio-visual fusion for dimensional emotion recognition”. In: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE. 2021, pp. 1–8.
- [202]John J Prevost and Sanchita Ghose. “FoleyGAN: Visually Guided Generative Adversarial Network-Based Synchronous Sound Generation in Silent Videos”. In: *IEEE Transactions on Multimedia* (2022).
- [203]Jie Pu, Yannis Panagakis, and Maja Pantic. “Learning differentiable sparse and low rank networks for audio-visual object localization”. In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2020, pp. 8668–8672.
- [204]Jie Pu, Yannis Panagakis, Stavros Petridis, and Maja Pantic. “Audio-visual object localization and separation using low-rank and sparsity”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2017, pp. 2901–2905.
- [205]Hendrik Purwins, Bo Li, Tuomas Virtanen, et al. “Deep learning for audio signal processing”. In: *IEEE Journal of Selected Topics in Signal Processing* 13.2 (2019), pp. 206–219.
- [206]Alec Radford, Jong Wook Kim, Chris Hallacy, et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV].
- [207]Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. “Multimodal co-learning: challenges, applications with datasets, recent advances and future directions”. In: *Information Fusion* 81 (2022), pp. 203–239.
- [208]Tanzila Rahman and Leonid Sigal. “Weakly-supervised audio-visual sound source detection and separation”. In: *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE. 2021, pp. 1–6.
- [209]Tanzila Rahman, Mengyu Yang, and Leonid Sigal. “TriBERT: Full-body Human-centric Audio-visual Representation Learning for Visual Sound Separation”. In: *arXiv preprint arXiv:2110.13412* (2021).
- [210]Manon Raimbault and Danièle Dubois. “Urban soundscapes: Experiences and knowledge”. In: *Cities* 22.5 (2005), pp. 339–350.
- [211]Arunan Ramalingam and Sridhar Krishnan. “Gaussian mixture modeling of short-time Fourier transform features for audio fingerprinting”. In: *IEEE Transactions on Information Forensics and Security* 1.4 (2006), pp. 457–463.

- [212]Mirco Ravanelli and Yoshua Bengio. “Interpretable convolutional filters with sincnet”. In: *arXiv preprint arXiv:1811.09725* (2018).
- [213]Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. “You only look once: Unified, real-time object detection”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 779–788.
- [214]Gaël Richard, Paris Smaragdis, Sharon Gannot, et al. “Audio Signal Processing in the 21st Century”. In: *IEEE Signal Processing Magazine* (2023).
- [215]Remi Rigal, Jacques Chodorowski, and Benoit Zerr. “Deep Audio-Visual Speech Separation Based on Facial Motion”. In: *Interspeech*. 2021, pp. 3540–3544.
- [216]Olivier Rioul and Pierre Duhamel. “Fast algorithms for discrete and continuous wavelet transforms”. In: *IEEE transactions on information theory* 38.2 (1992), pp. 569–586.
- [217]Xavier Rodet and Philippe Depalle. “Spectral envelopes and inverse FFT synthesis”. In: *Audio Engineering Society Convention 93*. Audio Engineering Society. 1992.
- [218]EJ Rothwell, KM Chen, and DP Nyquist. “An adaptive-window-width short-time Fourier transform for visualization of radar target substructure resonances”. In: *IEEE Transactions on Antennas and Propagation* 46.9 (1998), pp. 1393–1395.
- [219]Andrew Rouditchenko, Angie Boggust, David Harwath, et al. “AVLNet: Learning audio-visual language representations from instructional videos”. In: *arXiv preprint arXiv:2006.09199* (2020).
- [220]Andrew Rouditchenko, Hang Zhao, Chuang Gan, Josh McDermott, and Antonio Torralba. “Self-supervised audio-visual co-segmentation”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 2357–2361.
- [221]Sebastian Ruder. “An overview of gradient descent optimization algorithms”. In: *arXiv preprint arXiv:1609.04747* (2016).
- [222]David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. “Learning representations by back-propagating errors”. In: *nature* 323.6088 (1986), pp. 533–536.
- [223]John Sadowsky. “Investigation of signal characteristics using the continuous wavelet transform”. In: *johns hopkins apl technical digest* 17.3 (1996), pp. 258–269.
- [224]Hardik B Sailor, Dharmesh M Agrawal, and Hemant A Patil. “Unsupervised Filterbank Learning Using Convolutional Restricted Boltzmann Machine for Environmental Sound Classification”. In: *INTERSPEECH*. 2017, pp. 3107–3111.
- [225]Euschi Salah, Khaldi Amine, Kafi Redouane, and Kahlessenane Fares. “A Fourier transform based audio watermarking algorithm”. In: *Applied Acoustics* 172 (2021), p. 107652.
- [226]Justin Salamon and Juan Pablo Bello. “Deep convolutional neural networks and data augmentation for environmental sound classification”. In: *IEEE Signal Processing Letters* 24.3 (2017), pp. 279–283.

- [227]Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. “A Dataset and Taxonomy for Urban Sound Research”. In: *Proceedings of the 22nd ACM International Conference on Multimedia*. MM ’14. Orlando, Florida, USA: Association for Computing Machinery, 2014, pp. 1041–1044.
- [228]Valentina Sanguineti, Pietro Morerio, Alessio Del Bue, and Vittorio Murino. “Audio-visual localization by synthetic acoustic image generation”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 3. 2021, pp. 2523–2531.
- [229]Madeline C Schiappa, Yogesh S Rawat, and Mubarak Shah. “Self-supervised learning for videos: A survey”. In: *ACM Computing Surveys* (2022).
- [230]Romain Serizel, Victor Bisot, Slim ESSID, and Gaël Richard. “Acoustic features for environmental sound analysis”. In: *Computational analysis of sound scenes and events* (2018), pp. 71–101.
- [231]Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Grégoire Mesnil. “Learning semantic representations using convolutional neural networks for web search”. In: *Proceedings of the 23rd international conference on world wide web*. 2014, pp. 373–374.
- [232]Jiayin Shi and Chao Ma. “Unsupervised Sounding Object Localization with Bottom-Up and Top-Down Attention”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2022, pp. 1737–1746.
- [233]Ruibo Shi, Raymond WM Ng, and Pawel Swietojanski. “Teacher-student training for acoustic event detection using audioset”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 875–879.
- [234]Deepen Sinha and Ahmed H Tewfik. “Low bit rate transparent audio compression using adapted wavelets”. In: *IEEE Transactions on signal processing* 41.12 (1993), pp. 3463–3479.
- [235]Malcolm Slaney et al. “An efficient implementation of the Patterson-Holdsworth auditory filter bank”. In: *Apple Computer, Perception Group, Tech. Rep* 35.8 (1993).
- [236]Olga Slizovskaia, Leo Kim, Gloria Haro, and Emilia Gomez. “End-to-end sound source separation conditioned on instrument labels”. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2019, pp. 306–310.
- [237]HANG Su, QUAN Liu, and JINGSONG Li. “Boundary effects reduction in wavelet transform for time-frequency analysis”. In: *WSEAS transactions on signal processing* 8.4 (2012), pp. 169–179.
- [238]Kun Su, Xiulong Liu, and Eli Shlizerman. “Audeo: Audio generation for a silent performance video”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 3325–3337.
- [239]Kun Su, Xiulong Liu, and Eli Shlizerman. “How Does it Sound? Generation of Rhythmic Soundtracks for Human Movement Videos”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 29258–29273.
- [240]Zehua Sun, Qiuhong Ke, Hossein Rahmani, et al. “Human action recognition from various data modalities: A review”. In: *arXiv preprint arXiv:2012.11866* (2020).
- [241]Grzegorz Swirszcz, Wojciech Marian Czarnecki, and Razvan Pascanu. “Local minima in training of neural networks”. In: *arXiv preprint arXiv:1611.06310* (2016).

- [242]Christian Szegedy, Wei Liu, Yangqing Jia, et al. “Going deeper with convolutions”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 1–9.
- [243]Rishabh N Tak, Dharmesh M Agrawal, and Hemant A Patil. “Novel phase encoded mel filterbank energies for environmental sound classification”. In: *International Conference on Pattern Recognition and Machine Intelligence*. Springer. 2017, pp. 317–325.
- [244]Ruijie Tao, Zexu Pan, Rohan Kumar Das, et al. “Is someone speaking? Exploring long-term temporal features for audio-visual active speaker detection”. In: *Proceedings of the 29th ACM International Conference on Multimedia*. 2021, pp. 3927–3935.
- [245]Anthony Teolis and John J Benedetto. *Computational signal processing with wavelets*. Vol. 182. Springer, 1998.
- [246]Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. “Audio-visual event localization in unconstrained videos”. In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 247–263.
- [247]Y. Tokozume and T. Harada. “Learning environmental sounds with end-to-end convolutional neural network”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mar. 2017, pp. 2721–2725.
- [248]Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. *Learning from Between-class Examples for Deep Sound Recognition*. 2017. arXiv: 1711.10282 [cs.LG].
- [249]Achyut Mani Tripathi and Om Jee Pandey. “Divide and distill: new outlooks on knowledge distillation for environmental sound classification”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 1100–1113.
- [250]George Tzanetakis, Georg Essl, and Perry Cook. “Audio analysis using the discrete wavelet transform”. In: *Proc. conf. in acoustics and music theory applications*. Vol. 66. Citeseer. 2001.
- [251]Efthymios Tzinis, Scott Wisdom, Aren Jansen, et al. “Into the wild with AudioScope: Unsupervised audio-visual separation of on-screen sounds”. In: *arXiv preprint arXiv:2011.01143* (2020).
- [252]Francisco Rivera Valverde, Juana Valeria Hurtado, and Abhinav Valada. “There is more than meets the eye: Self-supervised multi-object detection and tracking with sound by distilling multimodal knowledge”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 11612–11621.
- [253]Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. “Attention is all you need”. In: *Advances in neural information processing systems*. 2017, pp. 5998–6008.
- [254]Sergey Verbitskiy and Viacheslav Vyshegorodtsev. “ERANNs: Efficient Residual Audio Neural Networks for Audio Pattern Recognition”. In: *arXiv preprint arXiv:2106.01621* (2021).
- [255]Prateek Verma and Jonathan Berger. “Audio transformers: Transformer architectures for large scale audio understanding. adieu convolutions”. In: *arXiv preprint arXiv:2105.00335* (2021).
- [256]Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. “Generating videos with scene dynamics”. In: *Advances in neural information processing systems* 29 (2016).

- [257] Boyu Wang, Jorge Mendez, Mingbo Cai, and Eric Eaton. “Transfer learning via minimizing the performance gap between domains”. In: *Advances in neural information processing systems* 32 (2019).
- [258] Lin Wang and Kuk-Jin Yoon. “Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks”. In: *IEEE transactions on pattern analysis and machine intelligence* 44.6 (2021), pp. 3048–3068.
- [259] Sinong Wang, Belinda Z Li, Madian Khabisa, Han Fang, and Hao Ma. “LinFormer: Self-attention with linear complexity”. In: *arXiv preprint arXiv:2006.04768* (2020).
- [260] Xiaohan Wang, Linchao Zhu, and Yi Yang. “T2VLAD: global-local sequence alignment for text-video retrieval”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 5079–5088.
- [261] Xiaoyu Wang, Xiangyu Kong, Xiulian Peng, and Yan Lu. “Multi-Modal Multi-Correlation Learning for Audio-Visual Speech Separation”. In: *arXiv preprint arXiv:2207.01197* (2022).
- [262] Yake Wei, Di Hu, Yapeng Tian, and Xuelong Li. “Learning in audio-visual context: A review, analysis, and new perspective”. In: *arXiv preprint arXiv:2208.09579* (2022).
- [263] Andreas S Weigend. “On overfitting and the effective number of hidden units”. In: *Proceedings of the 1993 connectionist models summer school*. Psychology Press. 2014, pp. 335–342.
- [264] Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu. “Photorealistic audio-driven video portraits”. In: *IEEE Transactions on Visualization and Computer Graphics* 26.12 (2020), pp. 3457–3466.
- [265] Steven Euijong Whang, Yuji Roh, Hwanjun Song, and Jae-Gil Lee. “Data collection and quality challenges in deep learning: A data-centric ai perspective”. In: *The VLDB Journal* 32.4 (2023), pp. 791–813.
- [266] Jian Wu, Yong Xu, Shi-Xiong Zhang, et al. “Time domain audio visual speech separation”. In: *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*. IEEE. 2019, pp. 667–673.
- [267] Lei Wu. “Learning a single neuron for non-monotonic activation functions”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 4178–4197.
- [268] Xinyi Wu, Zhenyao Wu, Lili Ju, and Song Wang. “Binaural Audio-Visual Localization”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 4. 2021, pp. 2961–2968.
- [269] Yifei Wu, Chenda Li, Jinfeng Bai, Zhongqin Wu, and Yanmin Qian. “Time-Domain Audio-Visual Speech Separation on Low Quality Videos”. In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 256–260.
- [270] Yang Xian, Yang Sun, Wenwu Wang, and Syed Mohsen Naqvi. “Two stage audio-video speech separation using multimodal convolutional neural networks”. In: *2019 Sensor Signal Processing for Defence Conference (SSPD)*. IEEE. 2019, pp. 1–5.

- [271]Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. “Audiovisual slowfast networks for video recognition”. In: *arXiv preprint arXiv:2001.08740* (2020).
- [272]Huang Xie and Tuomas Virtanen. “Zero-shot audio classification based on class label embeddings”. In: *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE. 2019, pp. 264–267.
- [273]Huang Xie and Tuomas Virtanen. “Zero-shot audio classification via semantic embeddings”. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), pp. 1233–1242.
- [274]Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. *Aggregated Residual Transformations for Deep Neural Networks*. 2017. arXiv: 1611.05431 [cs.CV].
- [275]Fangjian Xing, Hongwei Chen, Shizhong Xie, and Jianping Yao. “Ultrafast three-dimensional surface imaging based on short-time Fourier transform”. In: *IEEE Photonics Technology Letters* 27.21 (2015), pp. 2264–2267.
- [276]Yuwen Xiong, Mengye Ren, and Raquel Urtasun. “LoCo: Local contrastive representation learning”. In: *Advances in neural information processing systems* 33 (2020), pp. 11142–11153.
- [277]Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and Junyang Lin. “Understanding and improving layer normalization”. In: *Advances in Neural Information Processing Systems* 32 (2019).
- [278]Yichong Xu, Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, and Zheng Zhang. “Scale-invariant convolutional neural networks”. In: *arXiv preprint arXiv:1411.6369* (2014).
- [279]Yufei Xu, Qiming Zhang, Jing Zhang, and Dacheng Tao. “ViTAE: Vision transformer advanced by exploring intrinsic inductive bias”. In: *Advances in neural information processing systems* 34 (2021), pp. 28522–28535.
- [280]Jianfei Yang, Han Zou, Yuxun Zhou, and Lihua Xie. “Learning gestures from WiFi: A Siamese recurrent convolutional architecture”. In: *IEEE Internet of Things Journal* 6.6 (2019), pp. 10763–10772.
- [281]Karren Yang, Dejan Marković, Steven Krenn, Vasu Agrawal, and Alexander Richard. “Audio-Visual Speech Codecs: Rethinking Audio-Visual Speech Enhancement by Re-Synthesis”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 8227–8237.
- [282]Yalan Ye, Yukun He, Tongjie Pan, Jingjing Li, and Heng Tao Shen. “Alleviating domain shift via discriminative learning for generalized zero-shot learning”. In: *IEEE Transactions on Multimedia* 24 (2021), pp. 1325–1337.
- [283]Zipeng Ye, Mengfei Xia, Ran Yi, et al. “Audio-driven talking face video generation with dynamic convolution kernels”. In: *IEEE Transactions on Multimedia* (2022).
- [284]Ran Yi, Zipeng Ye, Juyong Zhang, Hujun Bao, and Yong-Jin Liu. “Audio-driven talking face video generation with learning-based personalized head pose”. In: *arXiv preprint arXiv:2002.10137* (2020).

- [285]Hongwei Yong, Jianqiang Huang, Deyu Meng, Xiansheng Hua, and Lei Zhang. “Momentum batch normalization for deep learning with small batch size”. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII* 16. Springer. 2020, pp. 224–240.
- [286]Chengzhu Yu, Heng Lu, Na Hu, et al. “DurIAN: Duration informed attention network for multimodal synthesis”. In: *arXiv preprint arXiv:1909.01700* (2019).
- [287]Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. “Visual relationship detection with internal and external linguistic knowledge distillation”. In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 1974–1982.
- [288]Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi. “LEAF: A learnable frontend for audio classification”. In: *arXiv preprint arXiv:2101.08596* (2021).
- [289]Donghuo Zeng, Yi Yu, and Keizo Oyama. “Deep triplet neural networks with cluster-CCA for audio-visual cross-modal retrieval”. In: *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16.3 (2020), pp. 1–23.
- [290]Hui Zhang, Mile Gu, XD Jiang, et al. “An optical neural chip for implementing complex-valued neural network”. In: *Nature communications* 12.1 (2021), p. 457.
- [291]Jingran Zhang, Xing Xu, Fumin Shen, et al. “Enhancing Audio-Visual Association with Self-Supervised Curriculum Learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 4. 2021, pp. 3351–3359.
- [292]Zhichao Zhang, Shugong Xu, Shan Cao, and Shunqing Zhang. “Deep convolutional neural network with mixup for environmental sound classification”. In: *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*. Springer. 2018, pp. 356–367.
- [293]Zhichao Zhang, Shugong Xu, Shunqing Zhang, Tianhao Qiao, and Shan Cao. “Learning Attentive Representations for Environmental Sound Classification”. In: *IEEE Access* 7 (2019), pp. 130327–130339.
- [294]Zhilu Zhang and Mert Sabuncu. “Generalized cross entropy loss for training deep neural networks with noisy labels”. In: *Advances in neural information processing systems* 31 (2018).
- [295]Zijun Zhang. “Improved Adam optimizer for deep neural networks”. In: *2018 IEEE/ACM 26th international symposium on quality of service (IWQoS)*. Ieee. 2018, pp. 1–2.
- [296]Xiaolan Zhao, Zuguo Wu, Jiren Xu, Keren Wang, and Jihai Niu. “Speech signal feature extraction based on wavelet transform”. In: *2011 International Conference on Intelligent Computation and Bio-Medical Instrumentation*. IEEE. 2011, pp. 179–182.
- [297]Liangli Zhen, Peng Hu, Xi Peng, Rick Siow Mong Goh, and Joey Tianyi Zhou. “Deep multimodal transfer learning for cross-modal retrieval”. In: *IEEE Transactions on Neural Networks and Learning Systems* 33.2 (2020), pp. 798–810.
- [298]Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. “Talking face generation by adversarially disentangled audio-visual representation”. In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01. 2019, pp. 9299–9306.

- [299]Hang Zhou, Xudong Xu, Dahua Lin, Xiaogang Wang, and Ziwei Liu. “Sep-stereo: Visually guided stereophonic audio generation by associating source separation”. In: *European Conference on Computer Vision*. Springer. 2020, pp. 52–69.
- [300]Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, et al. “Audio–Visual Segmentation”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 386–403.
- [301]Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. “Visual to sound: Generating natural sound for videos in the wild”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 3550–3558.
- [302]Boqing Zhu, Kele Xu, Dezhi Wang, et al. “Environmental sound classification based on multi-temporal resolution convolutional neural network combining with multi-level features”. In: *Pacific Rim Conference on Multimedia*. Springer. 2018, pp. 528–537.
- [303]Hao Zhu, Yi Li, Feixia Zhu, Aihua Zheng, and Ran He. “Let’s Play Music: Audio-Driven Performance Video Generation”. In: *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE. 2021, pp. 3574–3581.
- [304]Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. “Deep audio-visual learning: A survey”. In: *International Journal of Automation and Computing* 18.3 (2021), pp. 351–376.
- [305]Ye Zhu, Yu Wu, Nicu Sebe, and Yan Yan. “Vision+X: A Survey on Multimodal Learning in the Light of Data”. In: *arXiv preprint arXiv:2210.02884* (2022).

List of Figures

2.1	Sample smooth function and its derivative.	10
2.2	Sample activation functions and the corresponding derivatives.	15
2.3	Example of a two-dimensional convolution.	18
2.4	Example of the distribution normalization.	19
2.5	Example of attention matrices.	21
2.6	Architecture of a residual CNN (ResNet-50).	22
2.7	Sample workflow of a Vision Transformer.	24
2.8	Example of fitting to a $\mathbb{C}$ -valued function for 300 epochs.	26
2.9	Knowledge distillation workflow.	27
2.10	Workflow of a sample visual Siamese network.	28
2.11	Architecture of the CLIP model.	30
2.12	Transfer learning workflow.	31
2.13	Workflow of a sample zero-shot learning setup.	32
2.14	Sample window functions.	34
2.15	Sample filters of a Fourier filter bank.	36
2.16	Sample STFT-spectrogram.	37
2.17	Sample Shannon wavelets.	39
2.18	Sample fbsp-wavelets at $f_c = 2$ Hz and $m = 1$	40
3.1	Proposed model description basis.	45
4.1	Architecture of ESResNe(X)t.	76
4.2	Example of spectrochromatic mapping.	77
4.3	Schema of processing the stereo input.	80
4.4	Example of a trained filter from a linear and fbsp-layer.	81
4.5	Architecture of the AudioCLIP model.	82
5.1	Accuracy (%) at different SNRs (dB) on ESC-50.	91
5.2	Accuracy (%) at different SNRs (dB) on UrbanSound8K.	92
5.3	Accuracy (%) at different sample rates (Hz) on ESC-50.	92
5.4	Accuracy (%) at different sample rates (Hz) on UrbanSound8K.	93
5.5	Frequency response of learned filter banks.	93
5.6	Frequency response of DFT filter bank.	94

List of Tables

3.1	Models learning representations.	47
3.2	Models learning correspondences.	49
3.3	Models localizing audio-visual objects.	53
3.4	Models separating sound sources.	54
3.5	Models generating (spatio-)temporal signal.	55
5.1	Mean average precision (mAP, %) on AudioSet.	86
5.2	Accuracy (%) on ESC-50 and UrbanSound8K.	88
5.3	Influence of architecture on the accuracy (%).	89
5.4	Influence of transformation on the accuracy (%).	90
5.5	Precision, recall, and mAP (%) of AudioCLIP.	96
5.6	Zero-shot accuracy (%) on ESC-50 and UrbanSound8K.	98

List of Listings

2.1 PyTorch [196] code to compute symmetric cross entropy loss. Adapted
 from [206]. 29



A.Guzhov@proton.me



AndreyGuzhov

Andrey Guzhov

Highlights

Key Skills	Deep learning, digital signal processing, multimedia and audio analysis
Technologies	Python, PyTorch, scikit-learn, Git, MLflow, SQL, Java, Slurm, bash
Languages	English, German, Russian

Experience

2017 – Present	AI Researcher Deutsches Forschungszentrum für Künstliche Intelligenz GmbH – Kaiserslautern • Focus on deep learning for audio analysis • Enhancements of neural network architectures • Data acquisition and data mining • Master's thesis supervision
2010 – 2016	Software Developer ArtService Ltd. – Moscow • Taopix Photobook Software Platform development and maintenance • PHP, MySQL, JavaScript/jQuery development • WordPress plugin development
2015	Software Developer PARAMITEC Ltd. – Moscow • Enterprise information systems development • Information security management • Java, PostgreSQL development

Education

2017 – 2024	PhD in Computer Science University of Kaiserslautern-Landau – Kaiserslautern • Research focusing on multimodal deep learning • Publications and talks at top computer science conferences • Taught university hands-on course on Symbolic AI • Thesis: Audio analysis for multimodal learning
-------------	---

- 2010 – 2016 **Graduate Engineer**
 Bauman Moscow State Technical University – Moscow
 Majors: artificial intelligence, neural networks, natural language processing
 Thesis: Methods of detection of elementary (complex) events in the stream of text information
- 2010 – 2014 **Bachelor in Computer Science and Engineering**
 Bauman Moscow State Technical University – Moscow
 Majors: automated information processing, databases, computer networks
 Thesis: Subsystem of the distribution of payments from the students of paid form of tuition between departments

Publications & Conferences

- 2022 A. Guzhov, F. Raue, J. Hees and A. Dengel, "AudioCLIP: Extending CLIP to Image, Text and Audio," ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, Singapore, 2022, pp. 976-980, doi: 10.1109/ICASSP43922.2022.9747631.
- 2021 A. Guzhov, F. Raue, J. Hees and A. Dengel, "ESResNe(X)t-fbsp: Learning Robust Time-Frequency Transformation of Audio," 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 2021, pp. 1-8, doi: 10.1109/IJCNN52387.2021.9533654.
- 2020 A. Guzhov, F. Raue, J. Hees and A. Dengel, "ESResNet: Environmental Sound Classification Based on Visual Domain Models," 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 2021, pp. 4933-4940, doi: 10.1109/ICPR48806.2021.9413035.

Colophon

This thesis was typeset with $\text{\LaTeX}$ 2 ϵ . It uses the *Clean Thesis* style developed by Ricardo Langner and modified by Andrey Guzhov.

